UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Head Office: Università degli Studi di Padova

Department of Chemical Sciences

GRADUATE SCHOOL IN: Molecular Sciences

CURRICULUM: Pharmaceutical Sciences

CYCLE XXXV

TACKLING THE COVID-19 PANDEMIC THROUGH THE COMBINATION OF ESTABLISHED AND NOVEL COMPUTER-AIDED DRUG DISCOVERY WORKFLOWS

Coordinator: Ch.mo Prof. Leonard Jan Prins

Supervisor: Ch.mo Prof. Stefano Moro

Ph.D. student: Matteo Pavan

Abstract

Since its outbreak in December 2019, the COVID-19 pandemic has caused the death of more than 6.5 million people around the world to date. The high transmissibility of its causative agent, the SARS-CoV-2 virus, coupled with its potentially lethal outcome, provoked a profound global economic and social crisis. The urgency of finding suitable pharmacological tools to tame the pandemic shed light on the ever-increasing importance of computer simulations in rationalizing and speeding up the design of new drugs, further stressing the need for developing quick and reliable methods to identify novel active molecules and characterize their mechanism of action.

In the present work, several existing computer-aided drug discovery tools are successfully exploited for deciphering the recognition process between active molecules and crucial targets for the SARS-CoV-2 reproductive cycle such as the 3CL protease and the spike protein. Incidentally, the knowledge acquired upon working on SARS-CoV-2 has been successfully applied also to investigate other pharmaceutically relevant non-viral targets, including casein kinase 1δ and adenosine receptors.

Finally, two novel, in-house developed, methodologies for the characterization of binding processes between biological entities are presented, with the first one being the application of Supervised Molecular Dynamics (SuMD) to the study of RNA-protein complexes formation, while the second one being Thermal Titration Molecular Dynamics (TTMD), a brand new protocol for unbinding kinetics estimation.

Sommario

Sin dal suo scoppio nel dicembre 2019, la pandemia da COVID-19 ha causato fino ad oggi la morte di oltre 6,5 milioni di persone in tutto il mondo. L'elevata trasmissibilità del suo agente eziologico, il virus SARS-CoV-2, combinata con il suo esito potenzialmente letale, hanno provocato una profonda crisi economica e sociale a livello globale. L'urgenza di trovare strumenti farmacologici adeguati a contrastare la pandemia ha fatto luce sulla sempre crescente importanza dei metodi computazionali nel razionalizzare e accelerare la progettazione di nuovi farmaci, sottolineando ulteriormente la necessità di sviluppare metodi rapidi e affidabili per l'identificazione di nuove molecole attive e la caratterizzazione del loro meccanismo d'azione.

Nel presente lavoro di tesi, diverse metodologie computazionali atte alla scoperta di nuovi farmaci sono state impiegate con successo per decifrare il processo di riconoscimento fra le molecole attive e alcuni bersagli molecolari cruciali per il ciclo riproduttivo di SARS-CoV-2, quali la proteasi principale $3CL^{pro}$ e la proteina spike. Incidentalmente, le conoscenze acquisite lavorando su SARS-CoV-2 sono state applicate con successo anche per studiare altri bersagli non virali farmaceuticamente rilevanti, tra cui la caseina chinasi 1 e i recettori dell'adenosina.

Infine, vengono presentate due nuove metodologie, sviluppate internamente, per la caratterizzazione dei processi di legame tra entità biologiche, con la prima che consiste nell'applicazione della Dinamica Molecolare Supervisionata (SuMD) allo studio della formazione di complessi tra RNA e proteine, mentre la seconda consiste nella Titolazione Termica mediante Dinamica Molecolare (TTMD), un nuovo protocollo per la stima della cinetica dei processi di dissociazione.

Contents

INTRODUCTION

The COVID-19 pandemic

In December 2019, a cluster of pneumonia cases of unknown etiology emerged in the Chinese city of Wuhan[1]. Soon after, analyses of the patient's lung fluid, blood, and throat swab reconducted this outbreak to a newly identified virus, tentatively named 2019-new coronavirus (2019-nCoV)[2].

Phylogenetic analyses performed on viral genomes isolated from patients' samples revealed a close relationship between the new virus with several bat coronaviruses isolated in China (>90%). A lower degree of similarity was also found with SARS-CoV (80%) and MERS-CoV (50%), the causative agents of two recent coronavirus-related epidemics[3]. Based on phylogeny, taxonomy, and established practice, the virus was renamed SARS-CoV-2[4], while the associated illness was defined as COVID-19 by the World Health Organization (WHO)[5].

The striking similarity between the SARS-CoV-2 genome and several bat coronaviruses led to the hypothesis that bats could be the animal reservoir for SARS-CoV-2, with pangolins or other mammals acting as the intermediate host before human transmission[6]. The assumption that bats could be the animal reservoir of SARS-CoV-2 was further reaffirmed at a later stage by the work of Temmam et al., which identified in the caverns of North Laos a series of bat coronaviruses that share a high level of sequence similarity (96%) with the SARS-CoV-2 genome[7].

From a clinical perspective, the spectrum of COVID-19 manifestation is broad, ranging from asymptomatic infections to severe viral pneumonia with respiratory failure and even death[8]. The most common symptoms, similar to influenza, are related to mild upper respiratory tract affection, such as fever, cough, myalgia, and headache[9]. Less common but still relevant ones include gastrointestinal manifestations, such as diarrhea, more severe respiratory illnesses, such as dyspnea, and multi-organ failure[10].

The long incubation time compared to similar infections[11], the capability of asymptomatic[12] or paucisymptomatic[13] patients to transmit the virus even before the eventual symptoms' manifestation, and the aerial transmission modality[14,15] all concurred to determine a higher transmissibility index (estimated between 2.5 and 3.0) for the SARS-CoV-2 virus, compared to similar viral infections[16]. These factors contributed to the rapid spread of SARS-CoV-2

worldwide, resulting in more than 650 million cases and more than 6.5 million deaths globally[17].

In the first stages of the COVID-19 pandemic, extraordinary sanitary measures, such as physical and social distancing, wearing face masks and eye protection devices[18,19] were adopted to prevent the collapse of the public healthcare system[20], due to the unbalance between the high demand and the low availability of critical supplies[21,22]. Although this short-term plan has proven helpful in gaining time[23,24], more sustainable and long-term oriented strategies were needed to better cope with the socio-economic[25] and psychological[26] consequences of the pandemic, other than ensuring a fair and efficient resource management[27].

Drug repurposing

Considering that bringing a brand-new drug on the market is usually a very long and expensive process[28], the so-called "drug repurposing" was the first approach to finding suitable therapeutic options for COVID-19 patients[29,30]. This strategy extends the applicability domain of already marketed drugs for treating diseases other than the one it was conceived for[31]. This approach is appealing because it involves using de-risked compounds, with potentially lower overall development costs and shorter development timelines[32]. Unfortunately, despite all the promising premises[33], this approach was largely unsuccessful[34]. Indeed, several investigated drugs showed little to no efficacy in randomized clinical trials[34]. The few successful cases were primarily symptomatic treatments, mostly limited to hospital usage for the most severe cases due to the therapy's high costs or route of administration[35].

Failure of the drug repurposing strategy against COVID-19 can be mostly reconducted to the very first stages of the pandemic, where few clinical pieces of evidence were available for the rational elaboration of therapy plans. For example, the combination of HIV protease inhibitors Lopinavir and Ritonavir was examined[36], despite a suboptimal predicted recognition pattern towards the SARS-CoV-2 main protease (M^{pro}) compared to other compounds of the same class[37]. Another example is the combined use of an antimalaria drug (hydroxychloroquine) and an antibiotic (azithromycin) despite no clear indication of the possible mechanism of action[38,39].

With more and more clinical observations becoming available, more fine-tuned treatments, especially symptomatologic ones, were adopted. This is the case, for example, of corticosteroids such as dexamethasone[40], employed to tame the inflammatory response associated with severe COVID-19 cases, and low molecular weight heparins[41], used to prevent or treat thrombo-embolic events associated caused by interference with the cardiocirculatory system.

A group of anti-arthritis drugs represents another successful example of drug repurposing to their ability to modulate the immune response[42] and cytokine storm[43] caused by severe SARS-CoV-2 infection. This family includes the monoclonal antibodies Tocilizumab[44] and Sarilumab[45], which both inhibit Interleukin-6 (IL-6) signaling, Anakinra[46], that interferes instead with IL-1 signaling, and the Janus Kinase (JAK) inhibitor Baricitinib[47], alone or in conjunction with Remdesivir[48], with the latest representing maybe the most successful example of drug repurposing against COVID-19 being the first approved drug against this illness[49].

Originally designed against Ebola Virus, Remdesivir is a nucleotide analog prodrug that acts as a viral polymerase inhibitor[50] and is efficient in shortening the recovery time in hospitalized adult patients affected by COVID-19[51]. Unfortunately, as previously mentioned, Remdesivir and the other repurposed drugs need parenteral administration, thereby limiting their massive-scale adoption as pharmacological treatments against COVID-19[35].

Convalescent plasma and monoclonal antibodies

With the first round of spontaneously healed patients, doctors started flanking standard treatment with the use of convalescent plasma (CP), i.e., the plasma derived from recently recovered donors with a sufficiently high neutralizing antibody titer[52]. A similar protocol was previously adopted to face Ebola[53] and MERS[54] outbreaks, justifying its emergency use in the first stages of the COVID-19 pandemic. Unfortunately, despite promising observational data from the first studies performed on small-size patients' cohorts[55], more thorough investigations from more extensive clinical trials demonstrated the inefficacy of this treatment[56,57], leading to its dismission from routine clinical practices. Despite this failure, CP inspired the design of safer and more targeted immunological treatments in the form of monoclonal antibodies (mAbs)[58,59]. Since the beginning of the pandemic, several mAbs

directed against COVID-19 have been developed, with some obtaining approval from regulatory agencies[60]. Multiple of these mAbs are often used in conjunction to combine their neutralizing power and boost their therapeutic efficiency, exploiting their ability to bind at different epitopes[61].

The list of approved ones contains the therapeutic combinations of casirivimab and imdevimab (Regeneron/Roche), redanvimab (Celltrion Healthcare), sotrovimab (GSK), and the combination of tixagevimab and cilgavimab[62,63]. Furthermore, the association of bamlanivimab and etesevimab is nearly approved, despite the previous failure of trials investigating bamlanivimab on its own[63].

Vaccines

As seen in the case of CP and mAbs, a targeted immune response against SARS-CoV-2 can be a beneficial treatment for patients[64]. While immunoglobulins are limited to treating ongoing infections in hospital settings due to the high costs and the parenteral administration route, a more economical and scalable approach would be instructing the human body to produce this type of response without needing external intervention[65]. Based on this assumption and parallel to the drug repurposing approach, the industry and academia spent a consistent joint effort on developing preventive tools to avoid the infection in the first place or at least mitigate the most detrimental effects of the illness. This endeavor resulted in the quick approval by regulatory agencies of several vaccines[66].

Three different classes of these therapeutic entities can be recognized[67]. The first one, related to inactivated virus vaccines, comprises the Chinese CoronaVac (Sinovac) and the Russian CoviVac. The second group is formed by adenovirus vector vaccines such as Vaxzevria/ChAdOx1-S (AstraZeneca), Sputnik V/Gam-COVID-Vac, and Jcovden/Ad26.COV2.S (Janssen). Finally, the third one is composed of mRNA-based vaccines, including Comirnaty/BNT162b2 (Pfizer-BioNTech) and Spikevax/mRNA-1273 (Moderna)[68,69].

Despite the poor performances of the first class of vaccines[70,71], several independent studies have asserted worldwide the efficacy of vaccination campaigns based on the other two types of vaccines, particularly in the case of mRNA-based ones[72,73].

Spike protein

The ability of the SARS-CoV-2 virus to infect human cells heavily depends on a surface glycoprotein known as the S/spike protein[74], named after its peculiar shape[75]. For this reason, both mRNA vaccines and mAbs are designed to target this protein and prevent the virus entrance within the cell, thereby limiting its replication[76].

Concerning these, although different pathways for SARS-CoV-2 cell entry are possible[77,78], the principal and better-characterized one involves binding to the human ACE2 receptor (hACE2)[79], a membrane-anchored metallopeptidase that is abundantly present in various districts of the human body, from the vascular endothelium to the epithelia of lungs and small intestine[80]. On its own, host cell receptor binding is not sufficient to ensure entrance within host cells. Priming and activating the S protein by host proteases is required to enhance its cell-cell and virus-cell fusion processes and increase viral shielding from neutralizing antibodies[79,81]. The list of priming proteases included, but is not limited to, TMPRSS2, a transmembrane serine protease that is often co-expressed with ACE2 in SARS-CoV-2 target cells, Furin and cathepsin B/L[79,82,83]. The priming process entails the exposure of a lipophilic fusion peptide (FP), which penetrates the host cell membrane triggering the viral fusion[84] thanks to its strong membrane-perturbing capacities[85]

From a structural perspective, the spike is a trimeric transmembrane glycoprotein composed of 1273 amino acids organized in two main subunits, S1 and S2, and several functional domains[86].

The S1 subunit comprises two main domains, specifically the N-terminal and C-terminal domains (NTD and CTD, respectively), which are both involved in the binding to host cell receptors[86]. The CTD contains the receptor-binding domain (RBD, residues 319–541), consisting of two motifs. Firstly, a core structure formed by a twisted five-stranded anti-parallel β sheet (β1, β2, β3, β4, and β7), with three short helices (α1, α2, and α3). Secondly, an extended loop (receptor binding motif, RBM), formed by a two-stranded β sheet (β5 and β6), lying at one edge of the core and containing most of the residues involved in binding to hACE2[87].

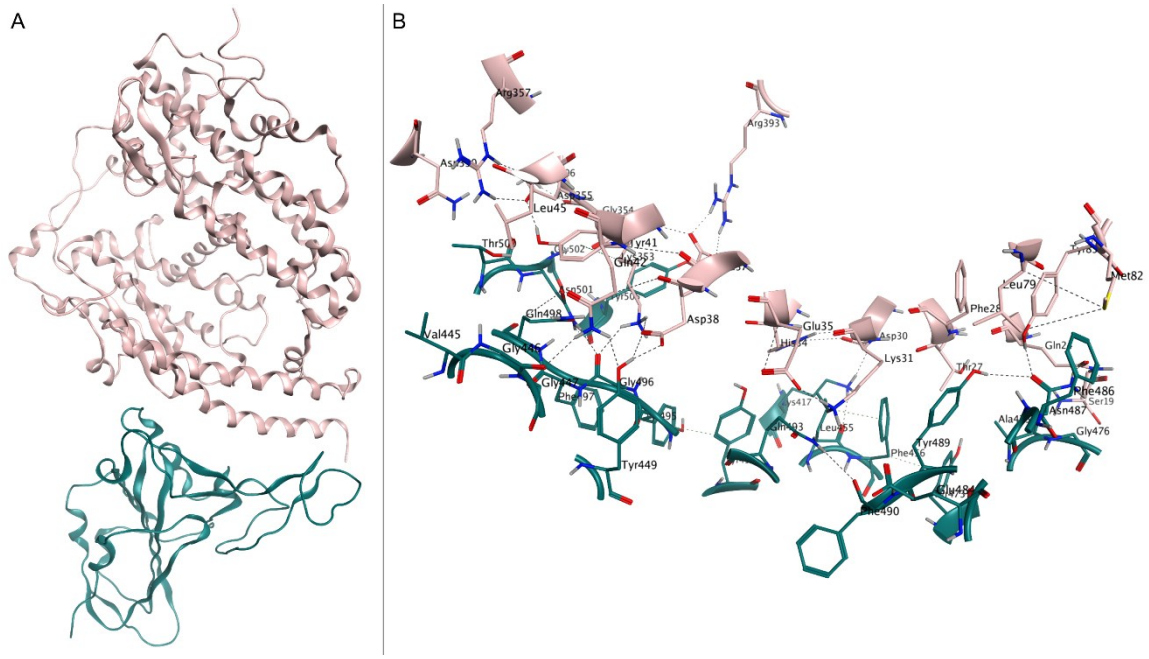


Figure 1. a) crystal structure of spike RBD (pink) in complex with hACE2 (teal), deposited in the Protein Data Bank with accession code 6M0J. b) close-up view of interface contacts between the spike RBD and hACE2: hydrogen bonds are represented as black dashed lines.

The S2 subdomain has significant roles in spike protein trimerization and in mediating the virion entry into the host cell once the molecular contacts have been established[88]. It is formed by relevant subdomains such as the transmembrane domain (TD) (residues 1296–1317), which exerts both the spike anchoring to the outer side of the viral membrane and the maintenance of the trimeric quaternary structure[89,90] and a cytoplasm domain (CD) (residues 1318–1353), which mediates viral assembly and cell-cell fusion[91]. Furthermore, the previously mentioned fusion peptide, a cleavage S2' site (residues 815/816), and two heptad-repeat domains (HR1/HR2) (residues 984–1104/1246–1295) are also part of S2[92].

Viral variants

Due to its exposition on the external surface of the SARS-CoV-2 membrane and its pivotal role in the virus' ability to infect host cells, the spike protein is often subjected to mutations that alter the virus's infectivity and antigenicity[93,94]. Therefore, since the spreading of the original viral strain (Wuhan-Hu-1) began, several viral variants appeared on the scene[95], particularly in the third-world nations where collective sanitary practices like social and physical distancing[96] or wearing face masks in public places [18] were hardly implementable[97].

The insurgence of novel viral strains with different susceptibility to the protective effect of vaccines[98] demands a periodical update of their original formulations coupled with multiple booster shots to maintain their efficacy[99], thus hampering the management of the pandemic based on massive vaccination of the world population[100,101].

Among the large pool of SARS-CoV-2 mutations[102], some gathered the scientific community's attention due to their increased fitness, gaining the "variant of concern" (VOC) status[103].

The first ever SARS-CoV-2 VOC was the B.1.1.7 variant, more commonly referred to as the "Alpha" or "English" variant due to being first identified in November 2020 in the Kent region of the United Kingdom[104,105]. Despite worries about the higher transmissibility compared to other circulating variants at the time[106,107], clinical studies appointed how mAbs, CP, and especially vaccines, were still able to confer protection against B.1.1.7[108–110], containing its impact on the sanitary system[111].

Unfortunately, soon after the emergence of the Alpha variant, a more threatening VOC arose. The B.1.617.2 variant, commonly known as the "Delta" or "Indian" variant due to being first identified in India in late 2020, quickly overthrew the B.1.1.7 one thanks to its strikingly increased transmissibility[105]. The advent of the Delta variant was associated with the first signs of reduced protection provided by mAbs, CP, and most importantly, vaccines[112–114], thanks to its increased immune system evasion capability[115], posing a heavier workload on the sanitary system[116].

The latest hallmark in the history of SARS-CoV-2 variants is represented by the B.1.1.529 variant, first detected in South Africa and more often recalled as the Omicron variant[117]. The combination of increased transmissibility[118] and immune system evasion[119] conferred this variant a net selective advantage in bypassing the protection provided by the complete primary vaccination cycle and a variety of clinically utilized mAbs[120–122] compared to other circulating strains. The ground-breaking impact the Omicron variant had on the worldwide spread of SARS-CoV-2 even led to the introduction of the "booster dose" to compensate for the reduced coverage of the primary vaccine cycle[98,123].

Lately, several subvariants germinated from the original Omicron strain (also labeled as BA.1), namely BA.2, BA.3, BA.4, and BA.5[124–126]. Although different studies indicated how the first identified Omicron subvariants (BA.2 and BA.3) were similarly susceptible to existing treatments despite their increased transmissibility[127–129], it also emerged how the most recently identified ones (BA.4 and BA.5) are significantly more efficient in evading the immune response[130–132].

These findings indicate that SARS-CoV-2 continued to evolve by increasing its immune-evasion capability rather than counting on sheer higher transmissibility, sustaining the virus spread even in populations with high vaccination frequency and recovery rates[130–132].

Main protease ($3CL^{pro}$)

Considering the uncertainty about the efficacy of existing treatments[133] and booster vaccinations[134] against present and future Omicron subvariants, the need to find more reliable and variant-agnostic therapeutic tools against COVID-19 is emerging more and more. The previously mentioned issues with the continuously mutating spike protein, which affects most present gold-standard COVID-19 treatments, indicate that different viral targets should be explored for developing novel antiviral drugs[135]. Generally speaking, an ideal target would have to play a pivotal role in the virus replication cycle and be highly conserved across different viral strains[136]. Within SARS-CoV-2, this role is portrayed by its main protease [137] (M^{pro}, or 3C-like protease / $3CL^{pro}$ due to similarities with the picornavirus 3C protease[138]), thanks to its conserved fold across different coronaviruses[138–141](including SARS-CoV[142]) and essentiality for the replication of this viruses' subfamily[143].

SARS-CoV-2 M^{pro}, also called nsp5, is a cysteine protease composed of 306 residues[144] that steers the maturation of two partially overlapping polyproteins (pp1a and pp1ab) into individual mature nonstructural proteins (including M^{pro} itself) through their proteolytic cleavage[145].

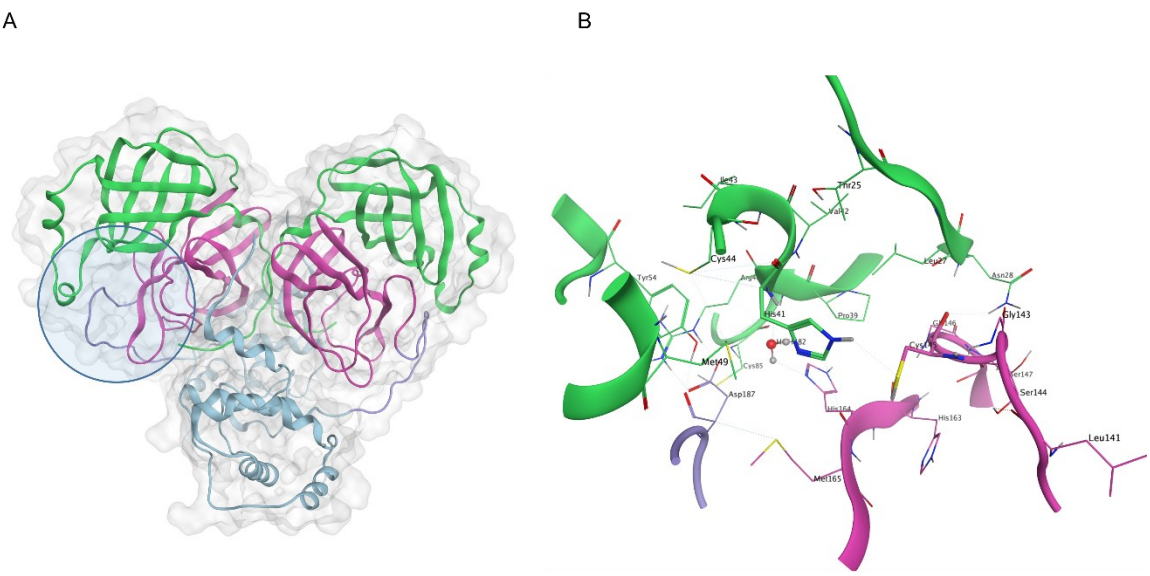


Figure 2. a) crystal structure of SARS-CoV-2 M^{pro} (PDB ID: 6Y2E): within each protomer, domain I is colored in green, domain II is colored in magenta, the 16-residue flexible loop is colored in violet, domain III is colored in light blue, while the active site position is highlighted by a blue circle. b) close-up view of the catalytic site: the H41-C145 dyad is highlighted, alongside the conserved water molecule that substitutes the third member of the canonical catalytic triad diffused in several cysteine proteases.

Functionally speaking, M^{pro} exists in equilibrium between a monomeric and a homodimer form[146–148]. This dimerization directly influences the shape of the catalytic site[147], thus altering the enzymatic activity[138] and playing an indirect regulatory role during the virus replication cycle[149,150].

Within the M^{pro} functional dimer, each protomer is composed of three structural domains. The chymotrypsin-like fold, including β-barrel domain I (residues 1-99) and II (residues 100-182), hosts the active site and thus has direct control over the catalytic event[138,147], while the α-helical domain III (residues 198-306) is mainly involved in the direct regulation of dimerization, exerting only a secondary and indirect role on regulating M^{pro}'s enzymatic activity[151]. Between the second and third domains lies a flexible 16-residue loop (residues 183-197)[152].

As anticipated, the catalytic site is located between domains I and II, bordered by the N-terminal domain I of the second protomer in the dimer. Notably, the N-finger (residues 1-7) interacts with the binding site through a salt bridge between the positively charged end of

Ser1 and the negatively charged end of Glu166[153]. The latest is also involved in forming a hydrogen bond with His172, an essential interaction for the enzyme's proteolytic activity[154]. These interactions are so crucial in stabilizing the catalytic site[155] that N-finger deletion impairs dimerization and abolishes the protease's enzymatic activity[156].

M^{pro}'s shallow, plastic, and solvent-exposed active site[152,157] comprises several subpockets (ranging from S6 to S3'), hosting the corresponding substrate residues (which vary from P6 to P3')[139]. Speaking of substrates, the SARS-CoV-2 M^{pro} cleaves peptide bonds at the C-terminus end of a glutamine residue (P1)[137], which is conserved across different SARS-CoV-2, SARS-CoV, and even MERS-CoV substrate sequences[152].

SARS-CoV-2 M^{pro} recognizes sequences as long as ten residues (P6–P5–P4–P3–P2–P1↓P1'–P2'–P3' P4', where ↓ indicates the scissile bond[139]), but only shows remarkable selectivity at four subsites: S4, S2, S1, and S1'[158]. On the contrary, prime recognition subsites located at the C-terminus of the conserved P2 (Leu/Val/Phe), P1 (Gln) ↓-P1' (Ser/Ala) sequence are not conserved and show remarkable plasticity[152,159]. Furthermore, the main structural alterations of the binding site derive from flexibility at residues that line the S_1 subpocket and segments incorporating methionine 49 and glutamine 189[152,160].

Differently from many other chymotrypsin-like proteases, M^{pro} exerts its enzymatic functions through a catalytic dyad instead of the usual triad, where His41 and Cys145 are flanked by a conserved water molecule that substitutes the sidechain of the third component (usually an aspartate or an asparagine)[138,161].

Aside from the catalytic dyad, another vital component of the catalytic machinery is represented by a set of conserved residues contouring the S1 subpocket known as the oxyanion loop (138-145)[152,162]. Notably, the correct conformation[87,163,164] of the oxyanion hole (Gly143-Ser144-Cys145) is required for stabilizing the tetrahedral transition state through a coordinated series of hydrogen bonds involving the backbone amides[138,155,165]. Accordingly, alternative oxyanion loop conformations are associated with catalytically incompetent/inactive proteases[140,152,154,166,167].

Rational design of COVID-19 drugs

Several characteristics of the viral proteases family, including SARS-CoV-2 M^{pro}, make them an attractive target for the rational development of tailored drugs against COVID-19. First, the low sequence identity with human proteases coupled with distinct cleavage-site specificities reduces the possibility of off-target/side effects associated with the therapy[168]. Second, the striking conservation of protein fold and structural organization of the active site among different members of the same family leads to the possibility of developing pan-coronaviral drugs[169]. Third, the abundance of structural data about the SARS-CoV-2 main protease (659 structures have been deposited in the Protein Data Bank[170] to date [27th December 2022]) makes it possible to exploit the state-of-the-art structure-based approaches in drug design[171]. Furthermore, a similar strategy has already proved successful in finding efficient treatments against the hepatitis C virus[172,173] and human immunodeficiency virus (HIV)[174,175]. Finally, the experience acquired studying the original SARS-CoV protease[176], in conjunction with the rapid release to the scientific community of the SARS-CoV-2 protease[164], certainly played a major role in determining its prominent place within most COVID-19 drug discovery campaigns.

The first attempts at finding SARS-CoV-2 M^{pro} inhibitors involved the repurposing of existing protease inhibitors. Particularly, the hepatitis C protease inhibitor Boceprevir[177,178] and the feline coronavirus $3CL^{pro}$ inhibitor GC373 (derived from its prodrug GC376) [179] were found to be active in the low µM potency range against M^{pro} [180], with the latter being particularly interesting due to its promiscuous anti-coronaviral activity[181]. Both candidate drugs share a similar peptidomimetic scaffold, which entails the most prominent interaction features of the first identified ones[164].

Although these primary hit compounds present a good binding pattern, their evolution towards clinical candidates and drugs is prevented by two main factors: first, covalent inhibitors are usually associated with selectivity problems, due to their ability to react promiscuously with a plethora of nucleophile moieties[182], second peptidomimetic scaffold are usually associated with suboptimal pharmacokinetic properties that affect the preferred route of administration[183].

With this regard, a step forward was obtained when the first SARS-CoV-2 M^{pro} inhibitors were able to reach clinical stage experimentation, namely PF-07304814 (lately renamed as Lufotrelvir), a prodrug for the active principle PF-00835231, and PF-07321332 (Nirmatrelvir).

Lufotrelvir was originally developed by Pfizer in 2002-2003 towards the SARS-CoV virus and later repurposed against the SARS-CoV-2 due to the high similarities between the two proteases[184]. Due to its efficacy against several viral strains in preclinical studies[185,186], it was advanced to the clinical stages of experimentation, albeit quickly overcome by Nirmatrelvir thanks to its more favorable pharmacokinetic profile[187].

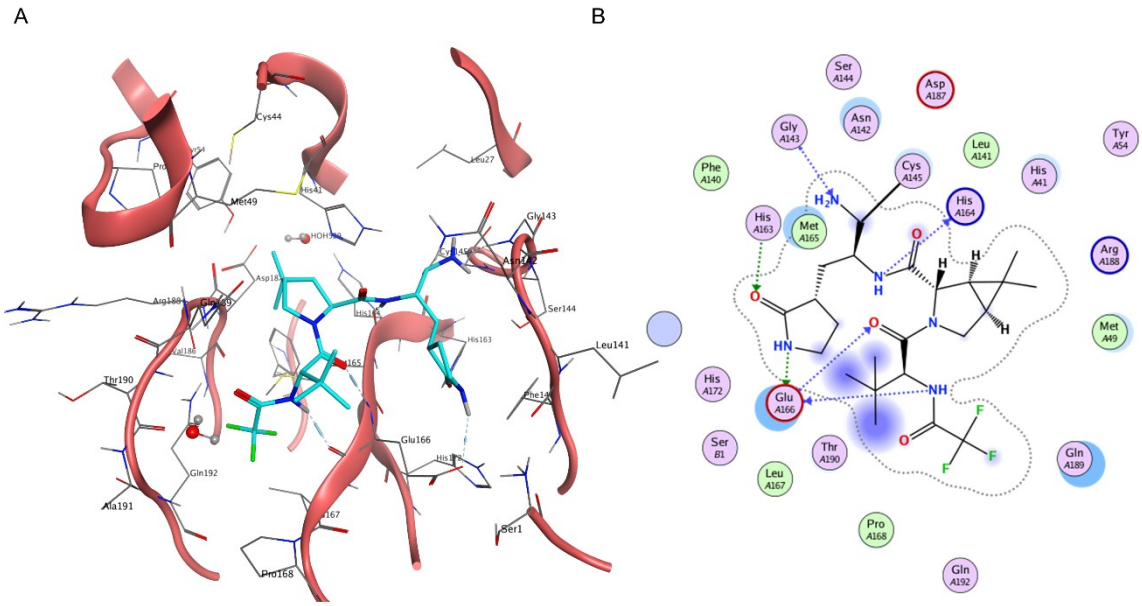


Figure 3. a) three-dimensional depiction of Nirmatrelvir orientation within the catalytic site of SARS-CoV-2 M^{pro} (PDB ID: 7RFW). b). bidimensional representation of intermolecular interactions of Nirmatrelvir – SARS-CoV-2 M^{pro} 7RFW complex.

Contrary to Lufotrelvir, which like Remdesivir requires parenteral administration, Nirmatrelvir can be administered orally[188], a must-have characteristic for the widespread adoption of drugs[189,190]. Designed by Pfizer amid the pandemic through the rational modification of Lufotrelvir[191], the structure of Nirmatrelvir was officially presented to the general audience on April 6th at the American Chemical Society Spring 2021 meeting[192], only one year after the official start of its development process[191].

This peptidomimetic inhibitor, which is administered in association with the pharmacokinetic enhancer Ritonavir and sold under the commercial name of Paxlovid, represents a hallmark in the history of both the COVID-19 pandemic and structure-based drug discovery, due to the groundbreaking speed of its discovery campaign[193]. Although clinical studies highlighted a remarkable therapeutic efficacy of Paxlovid in preventing the most severe COVID-19 cases[194], its effectiveness on more mild infections remains unclear[195]. Furthermore, the impact of viral mutations on present and future protease inhibitors has yet to be disclosed[196,197], thus justifying the current effort to find novel and diverse drugs that can enlarge the pool of pharmacological tools available against COVID-19.

An important step in this direction is represented by the development of Ensitrelvir (formerly known as S-217622), the first non-covalent, non-peptidomimetic, orally available M^{pro} inhibitor to reach clinical stage experimentation[198]. This compound has successfully reached the third and final stage of clinical experimentation, thanks to its proven efficacy against mild-to-moderate or even asymptomatic infections[199,200]. Possible approval of this active principle by regulatory agencies would provide an additional and orthogonal therapeutic tool to Nirmatrelvir in the treatment of COVID-19 cases, thus reducing the impact of resistance mechanisms associated with the emergence of mutated viral strains [196,197].

Potential targets of interest

Although targeting the SARS-CoV-2 main protease was successful in individuating several clinical candidate drugs, and even led to the first approval of COVID-19 specifically designed drugs, other drug discovery campaigns aimed at different viral targets are needed for therapy diversification, potential combined and synergic treatment, and resistance prevention[201–203].

Altogether, the SARS-CoV-2 genome encodes four major structural proteins, including nucleocapsid (N), membrane (M), envelope (E), and the spike as mentioned earlier (S), plus 16 nonstructural proteins, encompassing the previously mentioned main protease[204].

Although M^{pro} plays a pivotal role in processing the SARS-CoV-2 viral polyproteins, it is not the only component of the functional replicase complex that is required for the viral spread process[205]. Alongside her, a secondary but still relevant enzyme operates, namely the papain-like protease (PL^{pro}, the catalytic domain of protein nsp3)[206]. Despite being a cysteine protease

like M^{pro}, PL^{pro} exerts its enzymatic functions through a catalytic triad composed of Cys111, His272, and Asp286[207]. Further, PL^{pro} processes peptide bonds located at the C-terminal end of LXGG motifs[208]. Functionally speaking, this 343 residues segment which is part of the multidomain nsp3 protein is responsible for cleaving the SARS-CoV-2 polyproteins at three different sites, resulting in the liberation of nsp1, nsp2, and nsp3 proteins[209]. Moreover, PL^{pro} is also responsible for cleaving post-translational modifications on known regulators of host innate immune response[210].

As demonstrated by the approval of Remdesivir by regulatory agencies, another valuable target for the development of COVID-19 drugs is represented by the RNA-dependent RNA polymerase (RdRp)[49]. This complex machinery comprises four subunits, including one nsp12, responsible for the catalytic activity of the assembly, one nsp7, and two nsp8, with the latest two acting as cofactors[211]. The assembled holoenzyme presides RNA replication, a process that results in the formation of nine subgenomic RNAs[212]. The active site of nsp12 resides in its C-terminal RdRp domain and includes residues spanning from Thr611 to Met626, which are involved in binding one turn of double-stranded RNA, while residues D760 and D761 are required for recognition of the 3' end and essential for RNA synthesis[213,214]. Remdesivir binds within the active site, forming direct contact with residues K545, R553, D623, S682, T687, N691, S759, D760, and D761 and blocking the catalytic machinery by delaying the chain termination process[215,216].

During the RNA synthesis process, the RdRp interoperates also with nsp13 (helicase)[217], an enzyme involved in unwinding the RNA secondary structure of the 5' untranslated section of viral genome[218] to increase the efficiency of the copy process[219,220]. From a structural perspective, the nsp13 is a 596 residue, triangular pyramid-shaped helicase, which exploits its function thanks to the energy provided by its NTPase domain composed of six conserved residues (K288, S289, D374, E375, Q404, R567)[221]. Adding to its helicase activity, the nsp13 active site also exerts RNA 5' triphosphatase activity, further highlighting its importance in the maturation process of the viral mRNA[222].

The 5' end of the newly synthesized mRNA is then subjected to post-translational modifications to boost both its stability (preventing cleavage from exonucleases), protein

translation, and viral immune escape[223]. This activity is sequentially carried out by two S-adenosyl-L-methionine-dependent methyltransferases, namely nsp14 and nsp16[224].

Specifically, the 527 residue nsp14 encompasses both a proofreading exoribonuclease (ExoN) and an N7-methyltransferase enzymatic activity [225]. Furthermore, it has recently been suggested that it could encompass also a third, essential function for the viral replication cycle, based on the fact that SARS-CoV-2 ExoN knockout mutants are nonviable despite the 95% sequence identity with SARS-CoV [226] and the conservation of important active site amino acids including both the cap-binding residues (N306, C309, R310, W385, N386, N422, and F426) and the S-adenosyl methionine (SAM) binding residues (D352, Q354, F367, Y368, and W385)[227,228].

After its cleavage by the M^{pro}, evidence suggests that it forms a binary complex with nsp10 which cooperatively exerts the proofreading activity on fresh RNAs produced by the RdRp machinery [229,230]. Although the binary complex theory is the most prominent one, an alternative hypothesis based on the formation of a ternary nsp10-nsp14-nsp16 has been proposed due to the flexibility of the lid subdomain of nsp14 and the fact that nsp10 also forms a heterocomplex with nsp16 [230].

Particularly, the nsp16-nsp10 heterodimer is responsible for the 2' O-methyltransferase activity that is required to complete the cap-0 → cap-1 conversion of mRNA that is initiated by nsp14[224]. While the catalytic activity entirely resides on nsp16, nsp10 portrays a support role, aiding the recruitment of both the m7GppA-RNA substrate (which happens at a binding site defined by residues K24, C25, L27, Y30, K46, Y132, K137, K170, T172, E173, H174, S201, and S202) and the SAM cofactor (which binds in a pocket defined by N43, G71, G73, G81, D99, D114, C115, D130, and M131), thus enhancing nsp16's catalytic activity[231–233].

Lastly, another essential target for coronavirus biology is represented by nsp15, a uridine-specific endoribonuclease (NendoU)[234]. The active form of this enzyme is a dimer of trimers, with each monomer composed of 345 residues organized in three different domains: N-terminal, middle, and C-terminal NendoU, where the catalytic activity resides[235].

The active site contains six conserved residues: His250, His250, and Lys290, which compose the catalytic triad, Thr341, Tyr343, and Ser294, with the latest associated with selectivity in substrate recognition[236]. Due to their localization within the hexamer, cooperativity or anti-cooperativity between different binding sites is possible[237]. Nsp15 enzymatic activity involves the cleavage of both single- and double-stranded RNA at uridine sites producing 2',3'-cyclic phosphodiester, and 5'-hydroxyl termini[238].

Functionally speaking, Nsp15 seems to directly participate in viral replication through interference with the innate immune response[236]. Indeed, to evade host pattern recognition receptor MDA5 responsible for activating the host defenses, the Nsp15 cleaves the 5'-polyuridine tracts in (-) sense viral RNAs[239], though it has also been suggested that Nsp15 degrades viral RNA to hide it from the host defenses[237].

Computer simulations for rational drug design

For most of its existence, the human genre has exploited natural products such as leaves, seeds, roots, barks, and flowers as medicines, based on empirical observations purely based on symptom relief[240,241].

Nevertheless, throughout the latest two centuries, the process of drug discovery has evolved rapidly from the serendipitous discovery of novel active principles derived from or inspired by natural compounds[242,243] to the rational design of brand-new chemical entities[244].

The major turning point in the history of modern drug discovery can be traced back to the 1980s when experimentally solved macromolecular structures become routinely available[245]. The enhanced accessibility of structural data about biological targets reflected in a rapid interest in the development of computational methods that could valorize this information and aid medicinal chemists' work[246].

Today, computer simulations are a staple point of drug discovery campaigns, thanks to their ability to streamline and reduce their attrition rate[247]. From a functional perspective, computer-aided drug discovery (CADD) techniques are employed in the earliest stages of the pipeline for hit identification, hit-to-lead optimization, and pharmacokinetic evaluations[248].

CADD methodologies can either fall into one of two subgroups, based on the rationale behind them: the first group is represented by ligand-based (LBDD) approaches, while the second one includes structure-based (SBDD) methods[249]. The main difference between these two orthogonal and complementary approaches is that the first one does not exploit any information about the target macromolecule structure (e.g., a protein or a nucleic acid), while the second one does[250].

Nowadays, with the advent of cryo-electron microscopy (cryo-EM)[251] and groundbreaking tools for de novo prediction of protein structures such as AlphaFold[252], the second approach has become the gold-standard[253].

CADD strategies against COVID-19

The starting point of every SBDD campaign is the identification of a target macromolecule (a protein or a nucleic acid) that is involved in the etiology and or pathogenesis of a disease of

interest, whose function can be opportunely modulated through a specifically designed ligand, usually a small organic molecule[171].

Once the target has been identified, its structure must be retrieved, either through experimental methods such as X-Ray crystallography (XRC, the gold standard)[254], nuclear magnetic resonance (NMR)[255], and cryo-EM[256] or hypothesized through homology modeling or de novo prediction[257].

Homology modeling involves the use of a homologous protein with high primary sequence identity with the target as a template for the construction of its three-dimensional model[258,259]. De novo prediction, instead, does not rely on any information about other protein's structures and outputs a structural hypothesis that is solely based on the primary sequence of the target of interests[260].

While the second approach has gained a lot of momentum during the last two years, thanks to its unprecedentedly high accuracy[261,262], the first one is still relevant in those cases where important structural rearrangements occur between different states of the target functional cycle, other than predicting ligand-bound conformations[263,264].

In the context of the COVID-19 pandemic, where the extraordinary effort promoted by the scientific community quickly made several experimentally determined structures available, the relevance of structural modeling was highlighted by the ability to keep up with the high mutation rate of the virus[135,206], other than providing useful starting points for drug discovery campaigns for targets whose structure had yet to be elucidated[265,266]. For example, several studies were conducted to investigate the impact of mutations found in both the spike protein[135,267–271] and the main protease[135,197,272,273] of emerging strains on viral fitness and resistance to existing therapies. These studies showed that relatively inexpensive approaches such as homology modeling and positional scanning can be reliable tools to rationalize the origin of the virus[272,274–276], quickly track the evolution of the original strain[135,277,278], predict the impact of future possible mutations[268,270] and adjust existing therapeutics tools accordingly[197,279].

The huge amount of structural information available on several SARS-CoV-2 druggable targets was fertile terrain for various COVID-19 SBDD campaigns [280,281], both in academia and in the

industry, with most effort aimed at hitting well-characterized and pivotal viral targets such as M^{pro} or spike [282,283].

A remarkable example is represented by the COVID Moonshot Consortium, a drug discovery campaign driven by a collaborative effort among different research groups across the world aimed at targeting the SARS-CoV-2 main protease. This project led to the advancement of novel noncovalent orally available nanomolar M^{pro} inhibitors to clinical stage sperimentation[284].

The Swiss knife of SBDD: molecular docking

Within every SBDD campaign, available information about the target structure is exploited to fetch molecules able to recognize it selectively and potently[285]. Usually, this involves the identification of molecules that have good steric and electrostatic complementarity with the active site[286]. Depending on the steric and volumetric features of the binding site, the ligand type can be chosen accordingly, with small organic molecules being a better solution for buried cavities[287] and peptides, aptamers, or antibodies a better one for larger, flatter, and solvent-exposed interaction surfaces[288].

To narrow down the list of potentially active molecules to experimentally test to a feasible number, and to avoid wasting resources on compounds that do not possess the appropriate features to interact with the target, most SBDD campaigns start with a virtual screening process (SBVS)[289]. The most widely and successfully adopted method for SBVS is molecular docking, a computational protocol developed in the 1980s by Kuntz et al.[290] for predicting the preferred orientation of a certain ligand within the active site of a receptor[291].

Each docking program has two major components, which cooperate to find the solution to the protein-ligand docking problem[292]. The first part is the search algorithm (SA), which explores the ligand degrees of freedom within a user-defined search space centered around the active site of the protein[293]. The SA generates several ligand conformations (poses) that are fed to the second element of the program, i.e., the scoring function (SF), which qualitatively evaluates subsisting protein-ligand interaction features[294].

In the context of the COVID-19 pandemic, docking was also the king of computational methods used for drug discovery, thanks to the combination between its accuracy[295] and rapidity, which allows it to virtually screen billions of compounds in just a few days[296–298].

For example, Corona et al. reported the discovery of four low micromolar nsp13 inhibitors through a virtual screening carried out with the LiGen[299] docking program on an in-house natural compounds' library[300].

Kolarič et al. identified two micromolar SARS-CoV-2 cell-entry inhibitors, that act by binding human neuropilin-1 (nrp-1) and preventing its interaction with the spike protein, by performing a virtual screening with the GOLD[301] program on a library of commercially available compounds[302].

Vatansever et al. performed a virtual screening based on the Autodock[303] program on a library of drugs approved by the Food and Drug Administration and by the European Medical Agency (EMA) to discover six micromolar M^{pro} inhibitors[304].

Kao et al. reported the discovery of three sub-micromolar, synergistic nsp1 inhibitors identified through two independently executed virtual screenings with ICM[305,306] and Vina[307] software on a library of FDA-approved drugs[308].

Zhang et al. identified 11 natural compound M^{pro} inhibitors active in the low micromolar range through a virtual screening purely based on the commercial software Glide[309], developed by Schrödinger[310]. Another strategic use of docking-based virtual screening based on the Glide program is portrayed by the work of Huff et al., which designed six mixed covalent and non-covalent nanomolar M^{pro} inhibitors[311]. Another Glide-based virtual screening performed by Liu et al. led to the repurposing of histone deacetylase (HDAC) inhibitors as SARS-CoV-2 cell entry inhibitors through allosteric modulation of ACE2 and alteration of its ability to recognize the spike protein[312].

Wang et al. used LibDock[313] to perform a virtual screening on a library composed of FDA-approved peptides which led to the identification of a nanomolar SARS-CoV-2 cell entry inhibitor that exerts its effect by binding the human ACE2 receptor[314].

A remarkable result was obtained by Luttens et al., which identified 8 M^{pro} inhibitors (including a nanomolar compound with pan coronaviral activity) by combining fragment-based drug design with ultra-large virtual screening based on the DOCK[290] program[315].

Welker et al. exploited the molecular docking pipeline of the LeadIT[316] program to repurpose previously identified SARS-CoV PL^{pro} inhibitors towards its SARS-CoV-2 homolog, demonstrating their activity on viral replication in cell-based assays[317].

Otava et al. utilized docking calculations with the GOLD software to rationalize the structure-activity relationship of a series of rationally designed S-adenosyl-L-homocysteine derivatives, some of which showed inhibitory activity towards SARS-CoV-2 nsp14 in the low nanomolar potency range[318].

Similarly, Wang et al. exploited docking with Vina to rationalize the SAR of a series of rationally designed phenanthridine nucleocapsid protein (NPro) inhibitors, including two compounds showing low micromolar inhibitory activity[319].

Complementary strategies to address docking limitations

Although a very efficient and useful tool, molecular docking is rarely used on its own within SBDD campaigns and, indeed, is most often coupled with other methods to compensate for its weak points, such as neglecting receptor flexibility or the role of solvent[320], thus increasing the virtual screening success rate[321]. Another major limitation is represented by the poor ranking capabilities of classical scoring functions[322], which is the main cause of the high false positive rate of docking-based virtual screenings[323].

A possible solution to the limited physical description of the protein-ligand binding event of docking is to couple it with molecular dynamics (MD) simulations[292,324]. Molecular dynamics is a computational technique that allows investigating the time-dependent evolution of biological systems following the rules of molecular mechanics, i.e., determining the atomic trajectories by numerically solving Newton's equation of motion, where forces between the particles and their potential energies are calculated according to molecular mechanical force fields[325]. Due to the heavy computational workload required to run these types of simulations, MD is rarely used for screening purposes, while it is more frequently exploited for the refinement of docking results, i.e., evaluating the pose stability or optimizing the

protein-ligand complex geometry for a more accurate estimation of the free binding energy[326,327].

Regarding the pitfalls of the scoring component of docking programs, one possible strategy is to apply some form of knowledge-based filter upon docking results, in a similar fashion to what would happen if each pose were visually inspected[328]. For example, experimental information about critical protein-ligand interactions required for binding can be encoded within a pharmacophore filter or an interaction fingerprint, both of which can be used as constraints in the pose selection process[329]. In the case of pharmacophore filters, poses are filtered based on their ability to place a given functional group within a defined volume[330,331], while in the case of protein-ligand interaction fingerprint, the selection is usually based on the similarity between the reference and the query vector, representing the interaction features of the reference compound (a true active) and the investigated molecule respectively[332,333].

For instance, Wang et al. used a combination of structure-based pharmacophore screening, docking (both performed with the appropriate tools of the Molecular Operating Environment suite), and post-docking molecular dynamics refinement to identify a set of four sub-micromolar M^{pro} inhibitors among a database of in-house compounds[334].

The same protocol was successfully exploited by Tian et al. to identify four sub-micromolar PL^{pro} inhibitors in the same in-house library[335].

Furthermore, a slight variation of the protocol was also employed by Yin et al. to discover a non-covalent cyclic peptide that simultaneously inhibits both SARS-CoV-2 M^{pro} and nrp-1 with an activity in the low nanomolar range[336]. Within this scientific work, pharmacophore constraints were used for scoring peptide poses on M^{pro}, while traditional docking scores were used for the nrp-1 screening.

A remarkable joint computational work by Gossen et al. led to the molecular dynamics-driven design of a structure-based pharmacophore filter, which was then exploited to identify two nanomolar M^{pro} inhibitors among a library of publicly available compounds[337].

A similar approach was exploited by Hu et al., which exploited the combination between MD-based pharmacophore filtering, docking-based virtual screening within the Molecular

Operating Environment suite, and MD-based post-docking refinement to identify micromolar SARS-CoV-2 cell entry inhibitors targeting the FP of the spike protein[338].

Jang et al. used protein-ligand interaction fingerprint similarity as a post-docking filter for their double virtual screening on both M^{pro} and RdRp with the Vina program to identify seven compounds inhibiting SARS-CoV-2 replication in cell-based assays among a library of approved drugs[339].

Due to the static nature of molecular docking, which does not consider receptor flexibility, the choice of the input structure is vital for the success rate of a virtual screening[340]. Although molecular dynamics can be a useful "a posteriori" refinement of poses, a wrong input conformation of the target macromolecule could prevent the sampling of native-like poses for active compounds, leading to a reduced hit-finding rate[341]. For this reason, multiple conformations of the same receptor derived from MD simulations or experimentally solved in different conditions can be used in parallel in a process defined as ensemble docking (ED)[342]. When this approach is used, docking calculations are independently run on each structure, with virtual hit compounds being identified either through a consensus scoring or a consensus ranking approach[343,344]. In the case of consensus scoring, the docking score of the same molecule is averaged across the different virtual screenings, with the final ranking based on the consensus score[345]. Differently, consensus ranking involves the selection of top-ranking hit compounds across different virtual screenings, regardless of congruence between scores[346]. A consensus approach can also be utilized to rank molecules based on virtual screening executed on the same receptor structures with different docking protocols[347].

For example, Gimeno et al. applied a consensus scoring approach to three independently executed virtual screening through Glide, FRED[348,] and Vina software to identify two M^{pro} micromolar inhibitors within the Drugbank database, a library which includes all approved drugs by the Food and Drug Administration (FDA)[349].

Yang et al., instead, employed an ensemble docking approach with the Glide docking software to identify six M^{pro} inhibitors among a library of commercially available peptidomimetic compounds, two of which demonstrated sub-micromolar potency[350].

Rubio-Martinez et al. used a combination of ensemble docking based on QVina2[351] and post-docking molecular dynamics refinement to identify five M^{pro} micromolar inhibitors within a library of commercially available natural compounds[352].

A mixture of the previous two approaches was exploited by Clyde et al. for their High-Throughput Virtual Screening (HTVS), based on both ensemble docking and consensus scoring between the FRED and Vina docking programs, that led to the discovery 7 micromolar M^{pro} inhibitors among a set of commercially available compounds[353].

Further, a combination of consensus ranking among Autodock, Hybrid, and FlexX and post-docking molecular dynamics refinement was utilized by Glaab et al. to virtually screen a library of commercially available compounds and identify two micromolar M^{pro} inhibitors[354].

Similarly, Ghahremanpour et al. applied consensus ranking among three independent virtual screenings performed with the Glide, Autodock, and Vina software and post-docking molecular dynamics refinement to identify 14 micromolar M^{pro} inhibitors within the Drugbank database[355].

Another possible solution to cope with inaccuracy in free binding energy determination by traditional scoring functions is to rescore docking poses using more computationally intensive and accurate methods such as Free Energy Perturbation (FEP)[356] or MMGBSA/MMPBSA[357]. The first approach relies on performing a series of alchemical transformations across a set of ligands that need to be evaluated. This conversion cycle allows calculating relative differences in the free binding energy that can be used for a more accurate ranking of hit compounds derived from a virtual screening[358]. The second approach relies instead on correcting the gas phase interaction energy calculated according to the molecular mechanics force field with a term accounting for the desolvation free energy, where the polar component is estimated either by numerically solving the Poisson-Boltzmann equation (MMPBSA) or through the Generalized-Born method (MMGBSA)[359].

Intriguingly, one of the hit compounds identified in the work of Ghahremanpour et al. was then used by Zhang et al. for the FEP-driven design of multiple nanomolar M^{pro} inhibitors[360].

A similar combination of Glide docking and FEP to determine the absolute binding free energy was also employed by Li et al. to identify 15 micromolar M^{pro} inhibitors within the Drugbank

database[361]. The efficacy of FEP in estimating the binding energy of potential M^{pro} inhibitors was also highlighted by a retrospective study by Ngo et al[362].

A multistep virtual screening involving semiflexible docking with Glide, Schrödinger induced-fit docking[363], MD-based post docking refinement, and binding free energy estimation with the MMGBSA[364] protocol was exploited by Ibrahim et al. to identify one low micromolar nsp15 inhibitor[365].

Although the estimation of thermodynamic properties such as the free binding energy has been a staple point of drug discovery campaigns, both from a computational and an experimental perspective, lately there has been a major interest shift towards the determination of kinetic parameters since they better correlate with *in vivo* efficacy[366]. Specifically, several MD-based methods have been developed throughout the years to rank compounds based on their predicted residence time, i.e., the time that the ligand spends in the receptor-bound state[367]. Among those, Pavan et al. developed Thermal Titration Molecular Dynamics (TTMD), a new method for qualitative estimation of protein-ligand complex stability, that was successfully applied for correctly discriminating tight, low nanomolar, binders from weak, micromolar, SARS-CoV-2 M^{pro} inhibitors[368].

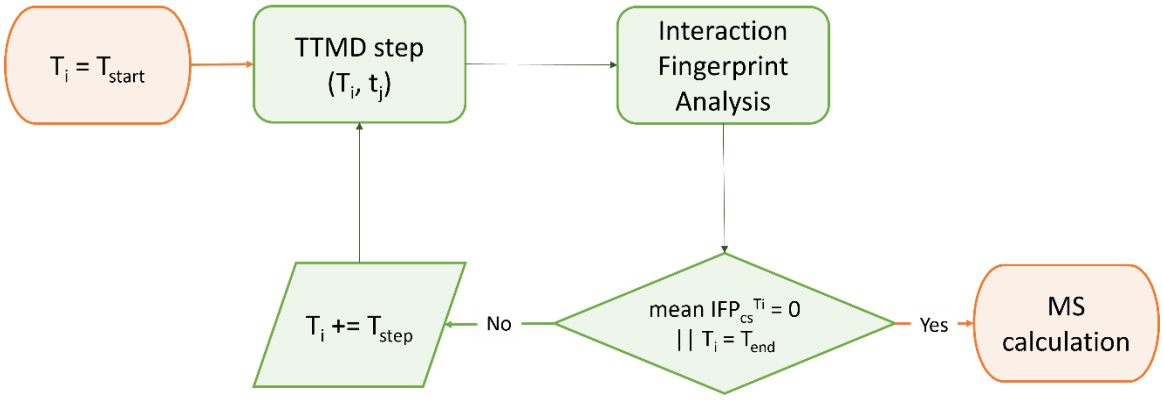


Figure 4. Workflow of a Thermal Titration Molecular Dynamics (TTMD) simulation. The time-dependent conservation of the native binding mode within a protein-ligand complex of interest is monitored with a scoring function based on interaction fingerprint through a series of short molecular dynamics simulations performed at progressively increasing temperatures. The simulation is carried out until the target temperature is reached or the dissociation process is completed. A coefficient called MS is then calculated and used to rank ligands based on the persistence of their native binding mode.

Beyond protein-ligand docking: alternative strategies for rational drug development

Despite the indisputable relevance of molecular docking within most SARS-CoV-2 drug discovery campaigns, other approaches were successfully implemented, especially for projects which deviate from the design of standard small molecule non-covalent binder.

For example, Zaidman et al. developed *Covalentizer*, an automated pipeline for the conversion of non-covalent binders to irreversible ones, which was successfully applied to the conversion of a SARS-CoV M^{pro} reversible inhibitor to a sub-micromolar SARS-CoV-2 M^{pro} irreversible one[369].

Valiente et al. reported the discovery of D-peptides that bind the spike RBD with low nanomolar affinity, hence blocking SARS-CoV-2 infection in cell-based assays These ACE2-mimicking peptides were selected within the starting library through a combination of structural alignment, MD-based post docking refinement, and binding free energy estimation[370].

Similarly, a series of peptides mimicking the HR2 domain of the spike protein able to prevent SARS-CoV-2 infection in cell-base assays with low micromolar potency were designed through the combination between structural alignment, mutational scanning with the BeAtMuSiC[371] tool and MD-based post docking refinement[372].

Jeong et al. used Rosetta[373] to rationally design a mAb that recognizes a conserved surface on the spike RBD of various coronaviruses with picomolar binding affinities, thereby strongly inhibiting SARS-CoV-2 replication in cell-based assay[374].

A similar strategy was exploited by Miao et al., which employed Rosetta docking and MD-based post-docking refinement to design an RNA aptamer that binds with picomolar affinity to the spike RBD and inhibits SARS-CoV-2 replication with sub-micromolar potency in cell-based assay[375].

Further, Cao et al. utilized a combination of modeling with Rosetta and docking with RifDock[376] to design ten mini proteins which bind with picomolar affinity to the spike RBD thus inhibiting SARS-CoV-2 infection within cell-based assays[377].

Moreover, Glasgow et al. combined modeling with Rosetta and computational alanine scanning with Robetta[378,379] to rationally design "ACE2 receptor traps", i.e., engineered proteins that bind the spike RBD with high affinity and neutralize SARS-CoV-2 infection as effectively as clinically used mAbs[380].

As thoroughly discussed in previous paragraphs, many SARS-CoV-2 drug discovery campaigns favored static, time-independent approaches such as docking or structural alignment, over time-dependent methods such as molecular dynamics. This can be attributed to the long calculation times, the reduced conformational sampling capabilities, and the lower accessibility of MD simulations to the general medicinal chemistry audience[327,381]. Despite these issues, several works demonstrated the potential of using full-fledged MD-based drug discovery pipelines, especially, when smart enhanced-sampling strategies are employed[381].

For example, Bissaro et al. showed how high-throughput supervised molecular dynamics (HT-SuMD)[382], a virtual screening platform based on an enhanced sampling MD protocol, could be successfully exploited for docking fragments to the active site of SARS-CoV-2 M^{pro}, overcoming accuracy limitations of most docking protocols[383] in identifying the native-like binding mode for frag-like compounds[384].

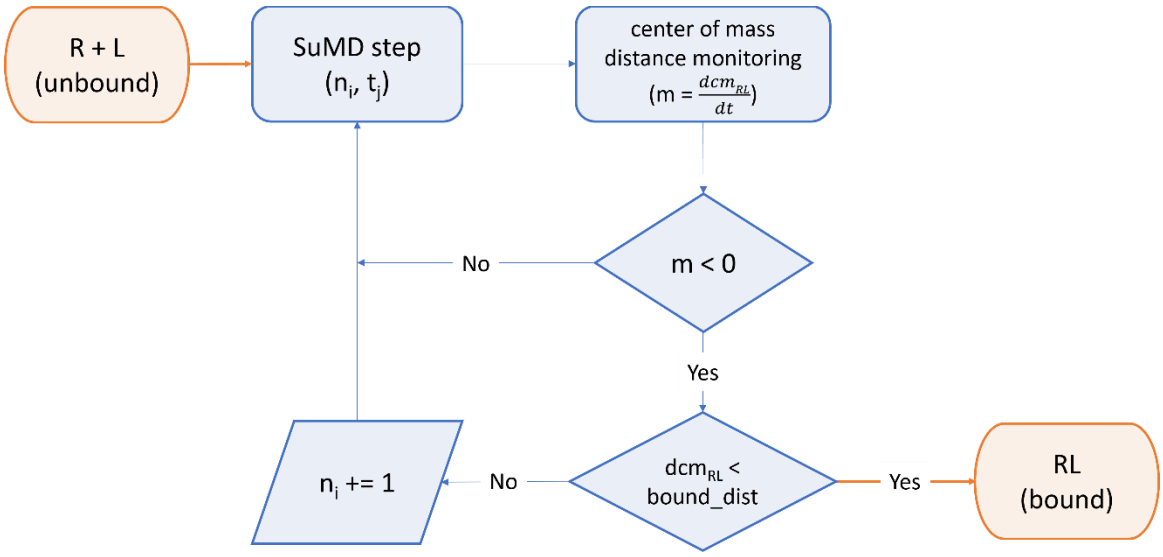


Figure 5. Workflow of a Supervised Molecular Dynamics (SuMD) simulation. The ligand is dynamically docked within a user-defined binding site through a series of short, unbiased molecular dynamics simulations. At the end of each step, the distance of mass between the ligand and the receptor binding site is computed for each trajectory frame and is fed to a tabu-like algorithm. If the slope of the straight line which interpolates the data is negative, indicating the ligand is approaching the binding site, the step is retained, and the simulation continues with the next "SuMD-step". If not, the step is discarded and repeated, randomly reassigning particles'

velocities through the Langevin thermostat. This cycle is repeated until a threshold distance is reached or other user-defined termination criteria are met.

Furthermore, the SuMD[385,386] algorithm was successfully exploited by Pavan et al. to decipher details about the recognition mechanism of Nirmatrelvir upon the SARS-CoV-2 M^{pro} catalytic site before any structural detail was revealed by the drug developer, successive structural[188] and molecular medicine[197] studies confirming the prediction validity[387].

Moreover, an evolved version of the SuMD protocol was developed by Pavan et al. and successfully applied to the study of the recognition mechanism between RNA aptamers and proteins, including an RNA-aptamer that binds to the spike RBD with picomolar affinity thus preventing the viral infection of host cells[388].

References

1. Lu, H., Stratton, C. W. & Tang, Y. W. Outbreak of pneumonia of unknown etiology in Wuhan, China: The mystery and the miracle. *J Med Virol* **92**, 401 (2020).

2. Guarner, J. Three Emerging Coronaviruses in Two Decades: The Story of SARS, MERS, and Now COVID-19. *American Journal of Clinical Pathology* vol. 153 420–421 Preprint at https://doi.org/10.1093/ajcp/aqaa029 (2020).

3. Lu, R. *et al.* Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *The Lancet* **395**, 565–574 (2020).

4. Gorbalenya, A. E. *et al.* The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. *Nature Microbiology 2020 5:4* **5**, 536–544 (2020).

5. WHO Director-General's remarks at the media briefing on 2019-nCoV on 11 February 2020. https://www.who.int/director-general/speeches/detail/who-director-general-s-remarks-at-the-media-briefing-on-2019-ncov-on-11-february-2020.

6. Andersen, K. G., Rambaut, A., Lipkin, W. I., Holmes, E. C. & Garry, R. F. The proximal origin of SARS-CoV-2. *Nature Medicine 2020 26:4* **26**, 450–452 (2020).

7. Temmam, S. *et al.* Bat coronaviruses related to SARS-CoV-2 and infectious for human cells. *Nature* (2022) doi:10.1038/s41586-022-04532-4.

8. Zhou, F. *et al.* Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *The Lancet* **395**, 1054–1062 (2020).

9. Guan, W. *et al.* Clinical Characteristics of Coronavirus Disease 2019 in China. *New England Journal of Medicine* **382**, 1708–1720 (2020).

10. Huang, C. *et al.* Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The Lancet* **395**, 497–506 (2020).

11. Lauer, S. A. *et al.* The incubation period of coronavirus disease 2019 (CoVID-19) from publicly reported confirmed cases: Estimation and application. *Ann Intern Med* **172**, 577–582 (2020).

12. Bai, Y. *et al.* Presumed Asymptomatic Carrier Transmission of COVID-19. *JAMA* **323**, 1406–1407 (2020).

13. He, X. *et al.* Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nat Med* **26**, 672–675 (2020).

14. Rothan, H. A. & Byrareddy, S. N. The epidemiology and pathogenesis of coronavirus disease (COVID-19) outbreak. *J Autoimmun* **109**, 102433 (2020).

15. Wiersinga, W. J., Rhodes, A., Cheng, A. C., Peacock, S. J. & Prescott, H. C. Pathophysiology, Transmission, Diagnosis, and Treatment of Coronavirus Disease 2019 (COVID-19): A Review. *JAMA - Journal of the American Medical Association* **324**, 782–793 (2020).

16. Petersen, E. *et al.* Comparing SARS-CoV-2 with SARS-CoV and influenza pandemics. *Lancet Infect Dis* **20**, e238–e244 (2020).

17. COVID Live - Coronavirus Statistics - Worldometer. https://www.worldometers.info/coronavirus/.

18. Chu, D. K. *et al.* Physical distancing, face masks, and eye protection to prevent person-to-person transmission of SARS-CoV-2 and COVID-19: a systematic review and meta-analysis. *The Lancet* **395**, 1973–1987 (2020).

19. Wilder-Smith, A. & Freedman, D. O. Isolation, quarantine, social distancing and community containment: pivotal role for old-style public health measures in the novel coronavirus (2019-nCoV) outbreak. *J Travel Med* **27**, (2020).

20. Remuzzi, A. & Remuzzi, G. COVID-19 and Italy: what next? *The Lancet* **395**, 1225–1228 (2020).

21. Ranney, M. L., Griffeth, V. & Jha, A. K. Critical Supply Shortages — The Need for Ventilators and Personal Protective Equipment during the Covid-19 Pandemic. *New England Journal of Medicine* **382**, e41 (2020).

22. Grasselli, G., Pesenti, A. & Cecconi, M. Critical Care Utilization for the COVID-19 Outbreak in Lombardy, Italy: Early Experience and Forecast During an Emergency Response. *JAMA* **323**, 1545–1546 (2020).

23. Pan, A. *et al.* Association of Public Health Interventions With the Epidemiology of the COVID-19 Outbreak in Wuhan, China. *JAMA* **323**, 1915–1923 (2020).

24. Giordano, G. *et al.* Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy. *Nature Medicine 2020 26:6* **26**, 855–860 (2020).

25. Nicola, M. *et al.* The socio-economic implications of the coronavirus pandemic (COVID-19): A review. *International Journal of Surgery* **78**, 185–193 (2020).

26. Rajkumar, R. P. COVID-19 and mental health: A review of the existing literature. *Asian J Psychiatr* **52**, 102066 (2020).

27. Emanuel, E. J. *et al.* Fair Allocation of Scarce Medical Resources in the Time of Covid-19. *New England Journal of Medicine* **382**, 2049–2055 (2020).

28. Schlander, M., Hernandez-Villafuerte, K., Cheng, C. Y., Mestre-Ferrandiz, J. & Baumann, M. How Much Does It Cost to Research and Develop a New Drug? A Systematic Review and Assessment. *Pharmacoeconomics* **39**, 1243–1269 (2021).

29. Liu, C. *et al.* Research and Development on Therapeutic Agents and Vaccines for COVID-19 and Related Human Coronavirus Diseases. *ACS Cent Sci* **6**, 315–331 (2020).

30. Tu, Y. F. *et al.* A review of sars-cov-2 and the ongoing clinical trials. *Int J Mol Sci* **21**, (2020).

31. Ashburn, T. T. & Thor, K. B. Drug repositioning: identifying and developing new uses for existing drugs. *Nature Reviews Drug Discovery 2004 3:8* **3**, 673–683 (2004).

32. Pushpakom, S. *et al.* Drug repurposing: Progress, challenges and recommendations. *Nat Rev Drug Discov* **18**, 41–58 (2018).

33. Mani, D., Wadhwani, A. & Krishnamurthy, P. T. Drug Repurposing in Antiviral Research: A Current Scenario. *Journal of Young Pharmacists* **11**, 117–121 (2019).

34. Viveiros Rosa, S. G. & Santos, W. C. Clinical trials on drug repositioning for COVID-19 treatment. *Revista Panamericana de Salud Publica/Pan American Journal of Public Health* **44**, e40 (2020).

35. TREATMENTS TO BE USED IN COVID-19 PATIENTS IN A HOSPITAL SETTING.

36. Cao, B. *et al.* A Trial of Lopinavir–Ritonavir in Adults Hospitalized with Severe Covid-19. *New England Journal of Medicine* **382**, 1787–1799 (2020).

37. Bolcato, G., Bissaro, M., Pavan, M., Sturlese, M. & Moro, S. Targeting the coronavirus SARS-CoV-2: computational insights into the mechanism of action of the protease inhibitors lopinavir, ritonavir and nelfinavir. *Sci Rep* **10**, 20927 (2020).

38. Gautret, P. *et al.* Hydroxychloroquine and azithromycin as a treatment of COVID-19: results of an open-label non-randomized clinical trial. *Int J Antimicrob Agents* **56**, (2020).

39. Arshad, S. *et al.* Treatment with hydroxychloroquine, azithromycin, and combination in patients hospitalized with COVID-19. *International Journal of Infectious Diseases* **97**, 396–403 (2020).

40. Dexamethasone in Hospitalized Patients with Covid-19. *New England Journal of Medicine* **384**, 693–704 (2021).

41. Tang, N. *et al.* Anticoagulant treatment is associated with decreased mortality in severe coronavirus disease 2019 patients with coagulopathy. *Journal of Thrombosis and Haemostasis* **18**, 1094–1099 (2020).

42. Giamarellos-Bourboulis, E. J. *et al.* Complex Immune Dysregulation in COVID-19 Patients with Severe Respiratory Failure. *Cell Host Microbe* **27**, 992-1000.e3 (2020).

43. Moore, J. B. & June, C. H. Cytokine release syndrome in severe COVID-19. *Science (1979)* **368**, 473–474 (2020).

44. Xu, X. *et al.* Effective treatment of severe COVID-19 patients with tocilizumab. *Proceedings of the National Academy of Sciences* **117**, 10970–10975 (2020).

45. Interleukin-6 Receptor Antagonists in Critically Ill Patients with Covid-19. *New England Journal of Medicine* **384**, 1491–1502 (2021).

46. Cavalli, G. *et al.* Interleukin-1 blockade with high-dose anakinra in patients with COVID-19, acute respiratory distress syndrome, and hyperinflammation: a retrospective cohort study. *Lancet Rheumatol* **2**, e325–e331 (2020).

47. Marconi, V. C. *et al.* Efficacy and safety of baricitinib for the treatment of hospitalised adults with COVID-19 (COV-BARRIER): a randomised, double-blind, parallel-group, placebo-controlled phase 3 trial. *Lancet Respir Med* **9**, 1407–1418 (2021).

48. Kalil, A. C. *et al.* Baricitinib plus Remdesivir for Hospitalized Adults with Covid-19. *New England Journal of Medicine* **384**, 795–807 (2021).

49. Rubin, D., Chan-Tack, K., Farley, J. & Sherwat, A. FDA Approval of Remdesivir — A Step in the Right Direction. *New England Journal of Medicine* **383**, 2598–2600 (2020).

50. Malin, J. J., Suárez, I., Priesner, V., Fätkenheuer, G. & Rybniker, J. Remdesivir against COVID-19 and other viral diseases. *Clin Microbiol Rev* **34**, 1–21 (2021).

51. Beigel, J. H. *et al.* Remdesivir for the Treatment of Covid-19 — Final Report. *New England Journal of Medicine* **383**, 1813–1826 (2020).

52. Chen, L., Xiong, J., Bao, L. & Shi, Y. Convalescent plasma as a potential therapy for COVID-19. *Lancet Infect Dis* **20**, 398–400 (2020).

53. Use of convalescent whole blood or plasma collected from patients recovered from Ebola virus disease for transfusion, as an empirical treatment during outbreaks: interim guidance for national health authorities and blood transfusion services. https://apps.who.int/iris/handle/10665/135591.

54. Arabi, Y. *et al.* Feasibility, safety, clinical, and laboratory effects of convalescent plasma therapy for patients with Middle East respiratory syndrome coronavirus infection: a study protocol. *Springerplus* **4**, 1–8 (2015).

55. Duan, K. *et al.* Effectiveness of convalescent plasma therapy in severe COVID-19 patients. *Proc Natl Acad Sci U S A* **117**, 9490–9496 (2020).

56. Li, L. *et al.* Effect of Convalescent Plasma Therapy on Time to Clinical Improvement in Patients with Severe and Life-threatening COVID-19: A Randomized Clinical Trial. *JAMA - Journal of the American Medical Association* **324**, 460–470 (2020).

57. Simonovich, V. A. *et al.* A Randomized Trial of Convalescent Plasma in Covid-19 Severe Pneumonia. *New England Journal of Medicine* **384**, 619–629 (2021).

58. Pinto, D. *et al.* Cross-neutralization of SARS-CoV-2 by a human monoclonal SARS-CoV antibody. *Nature* **583**, 290–295 (2020).

59. Tian, X. *et al.* Potent binding of 2019 novel coronavirus spike protein by a SARS coronavirus-specific human monoclonal antibody. *Emerg Microbes Infect* **9**, 382–385 (2020).

60. Use of monoclonal antibodies for COVID-19 | Italian Medicines Agency. https://www.aifa.gov.it/en/uso-degli-anticorpi-monoclonali.

61. Corti, D., Purcell, L. A., Snell, G. & Veesler, D. Tackling COVID-19 with neutralizing monoclonal antibodies. *Cell* **184**, 3086–3108 (2021).

62. Kaplon, H. & Reichert, J. M. Antibodies to watch in 2021. *MAbs* **13**, (2021).

63. Kaplon, H., Chenoweth, A., Crescioli, S. & Reichert, J. M. Antibodies to watch in 2022. *MAbs* **14**, (2022).

64. Zost, S. J. *et al.* Potently neutralizing and protective human antibodies against SARS-CoV-2. *Nature* **584**, 443–449 (2020).

65. Thanh Le, T. *et al.* The COVID-19 vaccine development landscape. *Nat Rev Drug Discov* **19**, 305–306 (2020).

66. Kashte, S., Gulbake, A., El-Amin, S. F. & Gupta, A. COVID-19 vaccines: rapid development, implications, challenges and future prospects. *Hum Cell* **34**, 711–733 (2021).

67. Pollard, A. J. & Bijker, E. M. A guide to vaccinology: from basic principles to new developments. *Nat Rev Immunol* **21**, 83–100 (2021).

68. Mascellino, M. T., di Timoteo, F., de Angelis, M. & Oliva, A. Overview of the main anti-sars-cov-2 vaccines: Mechanism of action, efficacy and safety. *Infect Drug Resist* **14**, 3459–3476 (2021).

69. Doroftei, B., Ciobica, A., Ilie, O. D., Maftei, R. & Ilea, C. Mini-review discussing the reliability and efficiency of covid-19 vaccines. *Diagnostics* **11**, (2021).

70. Jara, A. *et al.* Effectiveness of an Inactivated SARS-CoV-2 Vaccine in Chile. *New England Journal of Medicine* **385**, 875–884 (2021).

71. Tanriover, M. D. *et al.* Efficacy and safety of an inactivated whole-virion SARS-CoV-2 vaccine (CoronaVac): interim results of a double-blind, randomised, placebo-controlled, phase 3 trial in Turkey. *The Lancet* **398**, 213–222 (2021).

72. Polack, F. P. *et al.* Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine. *New England Journal of Medicine* **383**, 2603–2615 (2020).

73. Baden, L. R. *et al.* Efficacy and Safety of the mRNA-1273 SARS-CoV-2 Vaccine. *New England Journal of Medicine* **384**, 403–416 (2021).

74. Shang, J. *et al.* Cell entry mechanisms of SARS-CoV-2. *Proc Natl Acad Sci U S A* **117**, (2020).

75. Lwoff, A. & Tournier, P. The Classification of Viruses. *https://doi.org/10.1146/annurev.mi.20.100166.000401* **20**, 45–74 (2003).

76. Walls, A. C. *et al.* Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. *Cell* **181**, 281-292.e6 (2020).

77. Gadanec, L. K. *et al.* Can SARS-CoV-2 virus use multiple receptors to enter host cells? *Int J Mol Sci* **22**, 1–35 (2021).

78. Shen, S. *et al.* SARS-CoV-2 interacts with platelets and megakaryocytes via ACE2-independent mechanism. *J Hematol Oncol* **14**, (2021).

79. Hoffmann, M. *et al.* SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell* **181**, 271-280.e8 (2020).

80. Hamming, I. *et al.* Tissue distribution of ACE2 protein, the functional receptor for SARS coronavirus. A first step in understanding SARS pathogenesis. *J Pathol* **203**, 631–637 (2004).

81. Glowacka, I. *et al.* Evidence that TMPRSS2 Activates the Severe Acute Respiratory Syndrome Coronavirus Spike Protein for Membrane Fusion and Reduces Viral Control by the Humoral Immune Response. *J Virol* **85**, 4122–4134 (2011).

82. Belouzard, S., Millet, J. K., Licitra, B. N. & Whittaker, G. R. Mechanisms of coronavirus cell entry mediated by the viral spike protein. *Viruses* **4**, 1011–1033 (2012).

83. Hoffmann, M., Kleine-Weber, H. & Pöhlmann, S. A Multibasic Cleavage Site in the Spike Protein of SARS-CoV-2 Is Essential for Infection of Human Lung Cells. *Mol Cell* **78**, 779-784.e5 (2020).

84. Wrapp, D. *et al.* Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science (1979)* **367**, 1260–1263 (2020).

85. Guillén, J., Pérez-Berná, A. J., Moreno, M. R. & Villalaín, J. Identification of the membrane-active regions of the severe acute respiratory syndrome coronavirus spike membrane glycoprotein using a 16/18-mer peptide scan: implications for the viral fusion mechanism. *J Virol* **79**, 1743–1752 (2005).

86. Zhu, C. *et al.* Molecular biology of the SARs-CoV-2 spike protein: A review of current knowledge. *J Med Virol* **93**, 5729–5741 (2021).

87. Yan, R. *et al.* Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. *Science (1979)* **367**, 1444–1448 (2020).

88. Huang, Y., Yang, C., Xu, X. feng, Xu, W. & Liu, S. wen. Structural and functional properties of SARS-CoV-2 spike protein: potential antivirus drug development for COVID-19. *Acta Pharmacol Sin* **41**, 1141–1149 (2020).

89. Schroth-Diez, B. *et al.* The role of the transmembrane and of the intraviral domain of glycoproteins in membrane fusion of enveloped viruses. *Biosci Rep* **20**, 571–595 (2000).

90. Reuven, E. M. *et al.* HIV-1 gp41 transmembrane domain interacts with the fusion peptide: implication in lipid mixing and inhibition of virus-cell fusion. *Biochemistry* **51**, 2867–2878 (2012).

91. Petit, C. M. *et al.* Genetic analysis of the SARS-coronavirus spike glycoprotein functional domains involved in cell-surface expression and cell-to-cell fusion. *Virology* **341**, 215–230 (2005).

92. Li, F. Structure, Function, and Evolution of Coronavirus Spike Proteins. *https://doi.org/10.1146/annurev-virology-110615-042301* **3**, 237–261 (2016).

93. Li, Q. *et al.* The Impact of Mutations in SARS-CoV-2 Spike on Viral Infectivity and Antigenicity. *Cell* **182**, 1284 (2020).

94. Harvey, W. T. *et al.* SARS-CoV-2 variants, spike mutations and immune escape. *Nat Rev Microbiol* **19**, 409–424 (2021).

95. Forster, P., Forster, L., Renfrew, C. & Forster, M. Phylogenetic network analysis of SARS-CoV-2 genomes. *Proc Natl Acad Sci U S A* **117**, 9241–9243 (2020).

96. Oran, D. P. & Topol, E. J. Prevalence of Asymptomatic SARS-CoV-2 Infection. *https://doi.org/10.7326/M20-3012* **173**, 362–368 (2020).

97. Hellewell, J. *et al.* Feasibility of controlling COVID-19 outbreaks by isolation of cases and contacts. *Lancet Glob Health* **8**, e488–e496 (2020).

98. Garcia-Beltran, W. F. *et al.* Multiple SARS-CoV-2 variants escape neutralization by vaccine-induced humoral immunity. *Cell* **184**, 2372-2383.e9 (2021).

99. Fiolet, T., Kherabi, Y., MacDonald, C. J., Ghosn, J. & Peiffer-Smadja, N. Comparing COVID-19 vaccines for their characteristics, efficacy and effectiveness against SARS-CoV-2 and variants of concern: a narrative review. *Clinical Microbiology and Infection* **28**, 202–221 (2022).

100. Randolph, H. E. & Barreiro, L. B. Herd Immunity: Understanding COVID-19. *Immunity* **52**, 737–741 (2020).

101. Mathieu, E. *et al.* A global database of COVID-19 vaccinations. *Nat Hum Behav* **5**, 947–953 (2021).

102. Aksamentov, I., Roemer, C., Hodcroft, E. B. & Neher, R. A. Nextclade: clade assignment, mutation calling and quality control for viral genomes. *J Open Source Softw* **6**, 3773 (2021).

103. Tracking SARS-CoV-2 variants. https://www.who.int/activities/tracking-SARS-CoV-2-variants.

104. Frampton, D. *et al.* Genomic characteristics and clinical effect of the emergent SARS-CoV-2 B.1.1.7 lineage in London, UK: a whole-genome sequencing and hospital-based cohort study. *Lancet Infect Dis* **21**, 1246 (2021).

105. Campbell, F. *et al.* Increased transmissibility and global spread of SARSCoV- 2 variants of concern as at June 2021. *Eurosurveillance* **26**, 1–6 (2021).

106. Davies, N. G. *et al.* Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. *Science (1979)* **372**, (2021).

107. Volz, E. *et al.* Assessing transmissibility of SARS-CoV-2 lineage B.1.1.7 in England. *Nature* **593**, 266–269 (2021).

108. Collier, D. A. *et al.* Sensitivity of SARS-CoV-2 B.1.1.7 to mRNA vaccine-elicited antibodies. *Nature* **593**, 136–141 (2021).

109. Chen, R. E. *et al.* Resistance of SARS-CoV-2 variants to neutralization by monoclonal and serum-derived polyclonal antibodies. *Nat Med* **27**, 717–726 (2021).

110. Wang, P. *et al.* Antibody resistance of SARS-CoV-2 variants B.1.351 and B.1.1.7. *Nature* **593**, 130–135 (2021).

111. Haas, E. J. *et al.* Impact and effectiveness of mRNA BNT162b2 vaccine against SARS-CoV-2 infections and COVID-19 cases, hospitalisations, and deaths following a nationwide vaccination campaign in Israel: an observational study using national surveillance data. *The Lancet* **397**, 1819–1829 (2021).

112. Lopez Bernal, J. *et al.* Effectiveness of Covid-19 Vaccines against the B.1.617.2 (Delta) Variant. *New England Journal of Medicine* **385**, 585–594 (2021).

113. Planas, D. *et al.* Reduced sensitivity of SARS-CoV-2 variant Delta to antibody neutralization. *Nature* **596**, 276–280 (2021).

114. Liu, C. *et al.* Reduced neutralization of SARS-CoV-2 B.1.617 by vaccine and convalescent serum. *Cell* **184**, 4220-4236.e13 (2021).

115. Mlcochova, P. *et al.* SARS-CoV-2 B.1.617.2 Delta variant replication and immune evasion. *Nature* (2021) doi:10.1038/S41586-021-03944-Y.

116. Twohig, K. A. *et al.* Hospital admission and emergency care attendance risk for SARS-CoV-2 delta (B.1.617.2) compared with alpha (B.1.1.7) variants of concern: a cohort study. *Lancet Infect Dis* **22**, 35–42 (2022).

117. Gao, S. J., Guo, H. & Luo, G. Omicron variant (B.1.1.529) of SARS-CoV-2, a global urgent public health alert! *J Med Virol* **94**, 1255–1256 (2022).

118. Araf, Y. *et al.* Omicron variant of SARS-CoV-2: Genomics, transmissibility, and responses to current COVID-19 vaccines. *J Med Virol* **94**, 1825–1832 (2022).

119. Liu, L. *et al.* Striking antibody evasion manifested by the Omicron variant of SARS-CoV-2. *Nature* **602**, 676–681 (2022).

120. Dejnirattisai, W. *et al.* Reduced neutralisation of SARS-CoV-2 omicron B.1.1.529 variant by post-immunisation serum. *The Lancet* **399**, 234–236 (2022).

121. Cao, Y. *et al.* Omicron escapes the majority of existing SARS-CoV-2 neutralizing antibodies. *Nature* **602**, 657–663 (2022).

122. Hoffmann, M. *et al.* The Omicron variant is highly resistant against antibody-mediated neutralization: Implications for control of the COVID-19 pandemic. *Cell* **185**, 447-456.e11 (2022).

123. Nemet, I. *et al.* Third BNT162b2 Vaccination Neutralization of SARS-CoV-2 Omicron Infection. *New England Journal of Medicine* **386**, 492–494 (2022).

124. Dhawan, M. *et al.* Omicron variant (B.1.1.529) and its sublineages: What do we know so far amid the emergence of recombinant variants of SARS-CoV-2? *Biomedicine and Pharmacotherapy* **154**, (2022).

125. Ou, J. *et al.* Tracking SARS-CoV-2 Omicron diverse spike gene mutations identifies multiple inter-variant recombination events. *Signal Transduct Target Ther* **7**, (2022).

126. Kimura, I. *et al.* Virological characteristics of the SARS-CoV-2 Omicron BA.2 subvariants, including BA.4 and BA.5. *Cell* **185**, 3992-4007.e16 (2022).

127. Wilhelm, A. *et al.* Limited neutralisation of the SARS-CoV-2 Omicron subvariants BA.1 and BA.2 by convalescent and vaccine serum and monoclonal antibodies. *EBioMedicine* **82**, (2022).

128. Arora, P. *et al.* Comparable neutralisation evasion of SARS-CoV-2 omicron subvariants BA.1, BA.2, and BA.3. *Lancet Infect Dis* **22**, 766–767 (2022).

129. Evans, J. P. *et al.* Neutralization of SARS-CoV-2 Omicron sub-lineages BA.1, BA.1.1, and BA.2. *Cell Host Microbe* **30**, 1093-1102.e3 (2022).

130. Hachmann, N. P. *et al.* Neutralization Escape by SARS-CoV-2 Omicron Subvariants BA.2.12.1, BA.4, and BA.5. *New England Journal of Medicine* **387**, 86–88 (2022).

131. Arora, P. *et al.* Augmented neutralisation resistance of emerging omicron subvariants BA.2.12.1, BA.4, and BA.5. *Lancet Infect Dis* **22**, 1117–1118 (2022).

132. Wang, Q. *et al.* Antibody evasion by SARS-CoV-2 Omicron subvariants BA.2.12.1, BA.4 and BA.5. *Nature 2022 608:7923* **608**, 603–608 (2022).

133. Takashita, E. *et al.* Efficacy of Antibodies and Antiviral Drugs against Omicron BA.2.12.1, BA.4, and BA.5 Subvariants. *New England Journal of Medicine* **387**, 468–470 (2022).

134. Qu, P. *et al.* Neutralization of the SARS-CoV-2 Omicron BA.4/5 and BA.2.12.1 Subvariants. *New England Journal of Medicine* **386**, 2526–2528 (2022).

135. Pavan, M., Bassani, D., Sturlese, M. & Moro, S. From the Wuhan-Hu-1 strain to the XD and XE variants: is targeting the SARS-CoV-2 spike protein still a pharmaceutically relevant option against COVID-19? *J Enzyme Inhib Med Chem* **37**, 1704–1714 (2022).

136. V'kovski, P., Kratzel, A., Steiner, S., Stalder, H. & Thiel, V. Coronavirus biology and replication: implications for SARS-CoV-2. *Nat Rev Microbiol* **19**, 155–170 (2021).

137. Ullrich, S. & Nitsche, C. The SARS-CoV-2 main protease as drug target. *Bioorg Med Chem Lett* **30**, 127377 (2020).

138. Anand, K. *et al.* Structure of coronavirus main proteinase reveals combination of a chymotrypsin fold with an extra alpha-helical domain. *EMBO J* **21**, 3213–3224 (2002).

139. Anand, K., Ziebuhr, J., Wadhwani, P., Mesters, J. R. & Hilgenfeld, R. Coronavirus main proteinase (3CLpro) structure: basis for design of anti-SARS drugs. *Science* **300**, 1763–1767 (2003).

140. Xue, X. *et al.* Structures of two coronavirus main proteases: implications for substrate binding and antiviral drug design. *J Virol* **82**, 2515–2527 (2008).

141. Ho, B. L. *et al.* Critical Assessment of the Important Residues Involved in the Dimerization and Catalysis of MERS Coronavirus Main Protease. *PLoS One* **10**, (2015).

142. Tan, J. *et al.* pH-dependent conformational flexibility of the SARS-CoV main proteinase (M(pro)) dimer: molecular dynamics simulations and multiple X-ray structure analyses. *J Mol Biol* **354**, 25–40 (2005).

143. Xia, B. & Kang, X. Activation and maturation of SARS-CoV main protease. *Protein Cell* **2**, 282–290 (2011).

144. Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature 2020 579:7798* **579**, 265–269 (2020).

145. Snijder, E. J., Decroly, E. & Ziebuhr, J. The Nonstructural Proteins Directing Coronavirus RNA Synthesis and Processing. *Adv Virus Res* **96**, 59–126 (2016).

146. Chen, H. *et al.* Only one protomer is active in the dimer of SARS 3C-like proteinase. *Journal of Biological Chemistry* **281**, 13894–13898 (2006).

147. Zhang, L. *et al.* Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved α-ketoamide inhibitors. *Science (1979)* **368**, 409–412 (2020).

148. El-Baba, T. J. *et al.* Allosteric Inhibition of the SARS-CoV-2 Main Protease: Insights from Mass Spectrometry Based Assays*. *Angew Chem Int Ed Engl* **59**, 23544–23548 (2020).

149. Hsu, M. F. *et al.* Mechanism of the maturation process of SARS-CoV 3CL protease. *J Biol Chem* **280**, 31257–31266 (2005).

150. Li, C. *et al.* Conformational Flexibility of a Short Loop near the Active Site of the SARS-3CLpro is Essential to Maintain Catalytic Activity. *Sci Rep* **6**, (2016).

151. Shi, J. & Song, J. The catalysis of the SARS 3C-like protease is under extensive regulation by its extra domain. *FEBS J* **273**, 1035–1045 (2006).

152. Fornasier, E. *et al.* A new inactive conformation of SARS-CoV-2 main protease. *Acta Crystallogr D Struct Biol* **78**, 363–378 (2022).

153. Cheng, S. C., Chang, G. G. & Chou, C. Y. Mutation of glu-166 blocks the substrate-induced dimerization of SARS coronavirus main protease. *Biophys J* **98**, 1327–1336 (2010).

154. Yang, H. *et al.* The crystal structures of severe acute respiratory syndrome virus main protease and its complex with an inhibitor. *Proc Natl Acad Sci U S A* **100**, 13190–13195 (2003).

155. Verschueren, K. H. G. *et al.* A structural view of the inactivation of the SARS coronavirus main proteinase by benzotriazole esters. *Chem Biol* **15**, 597–606 (2008).

156. Zhong, N. *et al.* Without Its N-Finger, the Main Protease of Severe Acute Respiratory Syndrome Coronavirus Can Form a Novel Dimer through Its C-Terminal Domain. *J Virol* **82**, 4227–4234 (2008).

157. Kneller, D. W. *et al.* Malleability of the SARS-CoV-2 3CL Mpro Active-Site Cavity Facilitates Binding of Clinical Antivirals. *Structure* **28**, 1313-1320.e3 (2020).

158. Zhao, Y. *et al.* Structural basis for replicase polyprotein cleavage and substrate specificity of main protease from SARS-CoV-2. *Proc Natl Acad Sci U S A* **119**, (2022).

159. Lee, J. *et al.* X-ray crystallographic characterization of the SARS-CoV-2 main protease polyprotein cleavage sites essential for viral processing and maturation. *Nat Commun* **13**, (2022).

160. Behnam, M. A. M. Protein structural heterogeneity: A hypothesis for the basis of proteolytic recognition by the main protease of SARS-CoV and SARS-CoV-2. *Biochimie* **182**, 177–184 (2021).

161. Gorbalenya, A. E., Donchenko, A. P., Blinov, V. M. & Koonin, E. v. Cysteine proteases of positive strand RNA viruses and chymotrypsin-like serine proteases. A distinct protein superfamily with a common structural fold. *FEBS Lett* **243**, 103–114 (1989).

162. Perry A. Frey, A. D. H. *Enzymatic Reaction Mechanisms*. (Oxford University Press, 2007).

163. Douangamath, A. *et al.* Crystallographic and electrophilic fragment screening of the SARS-CoV-2 main protease. *Nat Commun* **11**, (2020).

164. Jin, Z. *et al.* Structure of Mpro from SARS-CoV-2 and discovery of its inhibitors. *Nature* **582**, 289–293 (2020).

165. Lee, J. *et al.* Crystallographic structure of wild-type SARS-CoV-2 main protease acyl-enzyme intermediate with physiological C-terminal autoprocessing site. *Nat Commun* **11**, (2020).

166. Shi, J., Sivaraman, J. & Song, J. Mechanism for controlling the dimer-monomer switch and coupling dimerization to catalysis of the severe acute respiratory syndrome coronavirus 3C-like protease. *J Virol* **82**, 4620–4629 (2008).

167. Allaire, M., Chernaia, M. M., Malcolm, B. A. & James, M. N. G. Picornaviral 3C cysteine proteinases have a fold similar to chymotrypsin-like serine proteinases. *Nature 1994 369:6475* **369**, 72–76 (1994).

168. de Clercq, E. Strategies in the design of antiviral drugs. *Nature Reviews Drug Discovery 2002 1:1* **1**, 13–25 (2002).

169. Kilianski, A. & Baker, S. C. Cell-based antiviral screening against coronaviruses: Developing virus-specific and broad-spectrum inhibitors. *Antiviral Res* **101**, 105–112 (2014).

170. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res* **28**, 235–242 (2000).

171. Anderson, A. C. The Process of Structure-Based Drug Design. *Chem Biol* **10**, 787–797 (2003).

172. Njoroge, F. G., Chen, K. X., Shih, N. Y. & Piwinski, J. J. Challenges in modern drug discovery: A case study of boceprevir, an HCV protease inhibitor for the treatment of hepatitis C virus infection. *Acc Chem Res* **41**, 50–59 (2008).

173. Pawlotsky, J. M., Feld, J. J., Zeuzem, S. & Hoofnagle, J. H. From non-A, non-B hepatitis to hepatitis C virus cure. *J Hepatol* **62**, S87–S99 (2015).

174. Wlodawer, A. & Vondrasek, J. Inhibitors of HIV-1 protease: A major success of structure-assisted drug design. *Annu Rev Biophys Biomol Struct* **27**, 249–284 (1998).

175. Palella, F. J. *et al.* Declining Morbidity and Mortality among Patients with Advanced Human Immunodeficiency Virus Infection. *New England Journal of Medicine* **338**, 853–860 (1998).

176. Wu, C. Y. *et al.* Small molecules targeting severe acute respiratory syndrome human coronavirus. *Proc Natl Acad Sci U S A* **101**, 10012–10017 (2004).

177. Poordad, F. *et al.* Boceprevir for Untreated Chronic HCV Genotype 1 Infection. *New England Journal of Medicine* **364**, 1195–1206 (2011).

178. Bacon, B. R. *et al.* Boceprevir for Previously Treated Chronic HCV Genotype 1 Infection. *New England Journal of Medicine* **364**, 1207–1217 (2011).

179. Pedersen, N. C. *et al.* Efficacy of a 3C-like protease inhibitor in treating various forms of acquired feline infectious peritonitis. *J Feline Med Surg* **20**, 378 (2018).

180. Ma, C. *et al.* Boceprevir, GC-376, and calpain inhibitors II, XII inhibit SARS-CoV-2 viral replication by targeting the viral main protease. *Cell Research 2020 30:8* **30**, 678–692 (2020).

181. Vuong, W. *et al.* Feline coronavirus drug inhibits the main protease of SARS-CoV-2 and blocks virus replication. *Nat Commun* **11**, (2020).

182. Singh, J., Petter, R. C., Baillie, T. A. & Whitty, A. The resurgence of covalent drugs. *Nat Rev Drug Discov* **10**, 307–317 (2011).

183. Pillaiyar, T., Manickam, M., Namasivayam, V., Hayashi, Y. & Jung, S. H. An overview of severe acute respiratory syndrome-coronavirus (SARS-CoV) 3CL protease inhibitors: Peptidomimetics and small molecule chemotherapy. *J Med Chem* **59**, 6595–6628 (2016).

184. Hoffman, R. L. *et al.* Discovery of Ketone-Based Covalent Inhibitors of Coronavirus 3CL Proteases for the Potential Therapeutic Treatment of COVID-19. *J Med Chem* **63**, 12725–12747 (2020).

185. Boras, B. *et al.* Preclinical characterization of an intravenous coronavirus 3CL protease inhibitor for the potential treatment of COVID19. *Nat Commun* **12**, (2021).

186. Vries, M. de *et al.* A Comparative Analysis of SARS-CoV-2 Antivirals Characterizes 3CLpro Inhibitor PF-00835231 as a Potential New Treatment for COVID-19. *J Virol* **95**, (2021).

187. Vandyck, K. & Deval, J. Considerations for the discovery and development of 3-chymotrypsin-like cysteine protease inhibitors targeting SARS-CoV-2 infection. *Curr Opin Virol* **49**, 36 (2021).

188. Owen, D. R. *et al.* An oral SARS-CoV-2 M pro inhibitor clinical candidate for the treatment of COVID-19. *Science (1979)* **374**, 1586–1593 (2021).

189. Lipinski, C. A., Lombardo, F., Dominy, B. W. & Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Deliv Rev* **46**, 3–26 (2001).

190. Lipinski, C. A. Lead- and drug-like compounds: the rule-of-five revolution. *Drug Discov Today Technol* **1**, 337–341 (2004).

191. How Pfizer scientists transformed an old drug lead into a COVID-19 antiviral. https://cen.acs.org/pharmaceuticals/drug-discovery/How-Pfizer-scientists-transformed-an-old-drug-lead-into-a-COVID-19-antiviral/100/i3.

192. Pfizer unveils its oral SARS-CoV-2 inhibitor. https://cen.acs.org/acs-news/acs-meeting-news/Pfizer-unveils-oral-SARS-CoV/99/i13.

193. Lamb, Y. N. Nirmatrelvir Plus Ritonavir: First Approval. *Drugs 2022 82:5* **82**, 585–591 (2022).

194. Mahase, E. Covid-19: Pfizer's paxlovid is 89% effective in patients at risk of serious illness, company reports. *BMJ* **375**, n2713 (2021).

195. Hung, Y. P. *et al.* Oral Nirmatrelvir/Ritonavir Therapy for COVID-19: The Dawn in the Dark? *Antibiotics* **11**, (2022).

196. Ullrich, S., Ekanayake, K. B., Otting, G. & Nitsche, C. Main protease mutants of SARS-CoV-2 variants remain susceptible to nirmatrelvir. *Bioorg Med Chem Lett* **62**, (2022).

197. Heilmann, E. *et al.* SARS-CoV-2 3CL ^{pro} mutations selected in a VSV-based system confer resistance to nirmatrelvir, ensitrelvir, and GC376. *Sci Transl Med* (2022) doi:10.1126/scitranslmed.abq7360.

198. McCarthy, M. W. Ensitrelvir as a potential treatment for COVID-19. *Expert Opin Pharmacother* 1–4 (2022) doi:10.1080/14656566.2022.2146493.

199. Shimizu, R. *et al.* Safety, Tolerability, and Pharmacokinetics of the Novel Antiviral Agent Ensitrelvir Fumaric Acid, a SARS-CoV-2 3CL Protease Inhibitor, in Healthy Adults. *Antimicrob Agents Chemother* **66**, (2022).

200. Mukae, H. *et al.* A Randomized Phase 2/3 Study of Ensitrelvir, a Novel Oral SARS-CoV-2 3C-Like Protease Inhibitor, in Japanese Patients with Mild-to-Moderate COVID-19 or Asymptomatic SARS-CoV-2 Infection: Results of the Phase 2a Part. *Antimicrob Agents Chemother* **66**, (2022).

201. Anighoro, A., Bajorath, J. & Rastelli, G. Polypharmacology: Challenges and opportunities in drug discovery. *J Med Chem* **57**, 7874–7887 (2014).

202. Peters, J. U. Polypharmacology - Foe or friend? *J Med Chem* **56**, 8955–8971 (2013).

203. Reddy, A. S. & Zhang, S. Polypharmacology: Drug discovery for the future. *Expert Rev Clin Pharmacol* **6**, 41–47 (2013).

204. Nasab, M. G., Saghazadeh, A. & Rezaei, N. SARS-CoV-2-A Tough Opponent for the Immune System. *Arch Med Res* **51**, 589–592 (2020).

205. Gordon, D. E. *et al.* A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature 2020 583:7816* **583**, 459–468 (2020).

206. Lubin, J. H. *et al.* Evolution of the SARS-CoV-2 proteome in three dimensions (3D) during the first 6 months of the COVID-19 pandemic. *Proteins: Structure, Function, and Bioinformatics* **90**, 1054–1080 (2022).

207. Osipiuk, J. *et al.* Structure of papain-like protease from SARS-CoV-2 and its complexes with non-covalent inhibitors. *Nat Commun* **12**, (2021).

208. Rut, W. *et al.* Activity profiling and crystal structures of inhibitor-bound SARS-CoV-2 papain-like protease: A framework for anti–COVID-19 drug design. *Sci Adv* **6**, 4596–4612 (2020).

209. Klemm, T. *et al.* Mechanism and inhibition of the papain-like protease, PLpro, of SARS-CoV-2. *EMBO J* **39**, e106275 (2020).

210. Shin, D. *et al.* Papain-like protease regulates SARS-CoV-2 viral spread and innate immunity. *Nature 2020 587:7835* **587**, 657–662 (2020).

211. Gao, Y. *et al.* Structure of the RNA-dependent RNA polymerase from COVID-19 virus. *Science (1979)* **368**, 779–782 (2020).

212. Kim, D. *et al.* The Architecture of SARS-CoV-2 Transcriptome. *Cell* **181**, 914-921.e10 (2020).

213. Hillen, H. S. *et al.* Structure of replicating SARS-CoV-2 polymerase. *Nature* **584**, 154–156 (2020).

214. Wang, Q. *et al.* Structural Basis for RNA Replication by the SARS-CoV-2 Polymerase. *Cell* **182**, 417-428.e13 (2020).

215. Yin, W. *et al.* Structural basis for inhibition of the RNA-dependent RNA polymerase from SARS-CoV-2 by remdesivir. *Science (1979)* **368**, 1499–1504 (2020).

216. Kokic, G. *et al.* Mechanism of SARS-CoV-2 polymerase stalling by remdesivir. *Nat Commun* **12**, 1–7 (2021).

217. Chen, J. *et al.* Structural Basis for Helicase-Polymerase Coupling in the SARS-CoV-2 Replication-Transcription Complex. *Cell* **182**, 1560-1573.e13 (2020).

218. Awan, F. M. *et al.* The emerging role and significance of circular RNAs in viral infections and antiviral immune responses: possible implication as theranostic agents. *RNA Biol* **18**, (2021).

219. Adedeji, A. O. *et al.* Mechanism of nucleic acid unwinding by SARS-CoV helicase. *PLoS One* **7**, (2012).

220. Mickolajczyk, K. J. *et al.* Force-dependent stimulation of RNA unwinding by SARS-CoV-2 nsp13 helicase. *Biophys J* **120**, 1020–1030 (2021).

221. Newman, J. A. *et al.* Structure, mechanism and crystallographic fragment screening of the SARS-CoV-2 NSP13 helicase. *Nature Communications 2021 12:1* **12**, 1–11 (2021).

222. Ivanov, K. A. *et al.* Multiple Enzymatic Activities Associated with Severe Acute Respiratory Syndrome Coronavirus Helicase. *J Virol* **78**, 5619–5632 (2004).

223. Chen, Y. *et al.* Biochemical and structural insights into the mechanisms of SARS coronavirus RNA ribose 2'-O-methylation by nsp16/nsp10 protein complex. *PLoS Pathog* **7**, (2011).

224. Romano, M., Ruggiero, A., Squeglia, F., Maga, G. & Berisio, R. A Structural View of SARS-CoV-2 RNA Replication Machinery: RNA Synthesis, Proofreading and Final Capping. *Cells 2020, Vol. 9, Page 1267* **9**, 1267 (2020).

225. Khailany, R. A., Safdar, M. & Ozaslan, M. Genomic characterization of a novel SARS-CoV-2. *Gene Rep* **19**, (2020).

226. Ogando, N. S. *et al.* The Enzymatic Activity of the nsp14 Exoribonuclease Is Critical for Replication of MERS-CoV and SARS-CoV-2. *J Virol* **94**, (2020).

227. Ma, Y. *et al.* Structural basis and functional analysis of the SARS coronavirus nsp14-nsp10 complex. *Proc Natl Acad Sci U S A* **112**, 9436–9441 (2015).

228. Lin, S. *et al.* Crystal structure of SARS-CoV-2 nsp10 bound to nsp14-ExoN domain reveals an exoribonuclease with both structural and functional integrity. *Nucleic Acids Res* **49**, 5382–5392 (2021).

229. Eckerle, L. D. *et al.* Infidelity of SARS-CoV Nsp14-exonuclease mutant virus replication is revealed by complete genome sequencing. *PLoS Pathog* **6**, 1–15 (2010).

230. Czarna, A. *et al.* Refolding of lid subdomain of SARS-CoV-2 nsp14 upon nsp10 interaction releases exonuclease activity. *Structure* **30**, 1050-1054.e2 (2022).

231. Wilamowski, M. *et al.* 2'-O methylation of RNA cap in SARS-CoV-2 captured by serial crystallography. *Proc Natl Acad Sci U S A* **118**, e2100170118 (2021).

232. Lin, S. *et al.* Crystal structure of SARS-CoV-2 nsp10/nsp16 2'-O-methylase and its implication on antiviral drug design. *Signal Transduction and Targeted Therapy 2020 5:1* **5**, 1–4 (2020).

233. Rosas-Lemus, M. *et al.* High-resolution structures of the SARS-CoV-2 2'-O-methyltransferase reveal strategies for structure-based inhibitor design. *Sci Signal* **13**, 1202 (2020).

234. Yoshimoto, F. K. The Proteins of Severe Acute Respiratory Syndrome Coronavirus-2 (SARS CoV-2 or n-COV19), the Cause of COVID-19. *Protein Journal* **39**, 198–216 (2020).

235. Frazier, M. N. *et al.* Flipped over U: structural basis for dsRNA cleavage by the SARS-CoV-2 endoribonuclease. *Nucleic Acids Res* **50**, 8290–8301 (2022).

236. Pillon, M. C. *et al.* Cryo-EM structures of the SARS-CoV-2 endoribonuclease Nsp15 reveal insight into nuclease specificity and dynamics. *Nature Communications 2021 12:1* **12**, 1–12 (2021).

237. Kim, Y. *et al.* Crystal structure of Nsp15 endoribonuclease NendoU from SARS-CoV-2. *Protein Science* **29**, 1596–1605 (2020).

238. Frazier, M. N. *et al.* Characterization of SARS2 Nsp15 nuclease activity reveals it's mad about U. *Nucleic Acids Res* **49**, 10136–10149 (2021).

239. Kim, Y. *et al.* Tipiracil binds to uridine site and inhibits Nsp15 endoribonuclease NendoU from SARS-CoV-2. *Communications Biology 2021 4:1* **4**, 1–11 (2021).

240. Cragg, G. M. & Newman, D. J. Natural products: A continuing source of novel drug leads. *Biochim Biophys Acta Gen Subj* **1830**, 3670–3695 (2013).

241. Gurib-Fakim, A. Medicinal plants: Traditions of yesterday and drugs of tomorrow. *Mol Aspects Med* **27**, 1–93 (2006).

242. Harvey, A. L. Natural products in drug discovery. *Drug Discov Today* **13**, 894–901 (2008).

243. Ban, T. A. The role of serendipity in drug discovery. *Dialogues Clin Neurosci* **8**, 335–344 (2006).

244. Morphy, R., Kay, C. & Rankovic, Z. From magic bullets to designed multiple ligands. *Drug Discov Today* **9**, 641–651 (2004).

245. Jaskolski, M., Dauter, Z. & Wlodawer, A. A brief history of macromolecular crystallography, illustrated by a family tree and its Nobel fruits. *FEBS J* **281**, 3985–4009 (2014).

246. Sliwoski, G., Kothiwale, S., Meiler, J. & Lowe, E. W. Computational Methods in Drug Discovery. *Pharmacol Rev* **66**, 334–395 (2014).

247. Leelananda, S. P. & Lindert, S. Computational methods in drug discovery. *Beilstein Journal of Organic Chemistry* **12**, 2694–2718 (2016).

248. Kapetanovic, I. M. Computer-aided drug discovery and development (CADDD): In silico-chemico-biological approach. *Chem Biol Interact* **171**, 165–176 (2008).

249. Macalino, S. J. Y., Gosu, V., Hong, S. & Choi, S. Role of computer-aided drug design in modern drug discovery. *Arch Pharm Res* **38**, 1686–1701 (2015).

250. Yu, W. & Mackerell, A. D. Computer-aided drug design methods. *Methods in Molecular Biology* **1520**, 85–106 (2017).

251. Bai, X. chen, McMullan, G. & Scheres, S. H. W. How cryo-EM is revolutionizing structural biology. *Trends Biochem Sci* **40**, 49–57 (2015).

252. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. **596**, 583–589 (2021).

253. Anderson, A. C. The Process of Structure-Based Drug Design. *Chem Biol* **10**, 787–797 (2003).

254. Smyth, M. S. & Martin, J. H. J. x Ray crystallography. *Molecular Pathology* **53**, 8 (2000).

255. Markwick, P. R. L., Malliavin, T. & Nilges, M. Structural Biology by NMR: Structure, Dynamics, and Interactions. *PLoS Comput Biol* **4**, (2008).

256. Nwanochie, E. & Uversky, V. N. Structure Determination by Single-Particle Cryo-Electron Microscopy: Only the Sky (and Intrinsic Disorder) is the Limit. *International Journal of Molecular Sciences 2019, Vol. 20, Page 4186* **20**, 4186 (2019).

257. Kuhlman, B. & Bradley, P. Advances in protein structure prediction and design. *Nature Reviews Molecular Cell Biology 2019 20:11* **20**, 681–697 (2019).

258. Martí-Renom, M. A. *et al.* Comparative protein structure modeling of genes and genomes. *Annu Rev Biophys Biomol Struct* **29**, 291–325 (2000).

259. Cavasotto, C. N. & Phatak, S. S. Homology modeling in drug discovery: current trends and applications. *Drug Discov Today* **14**, 676–683 (2009).

260. Bonneau, R. *et al.* De Novo Prediction of Three-dimensional Structures for Major Protein Families. doi:10.1016/S0022-2836(02)00698-8.

261. Senior, A. W. *et al.* Protein structure prediction using multiple deep neural networks in the 13th Critical Assessment of Protein Structure Prediction (CASP13). *Proteins: Structure, Function, and Bioinformatics* **87**, 1141–1148 (2019).

262. Pereira, J. *et al.* High-accuracy protein structure prediction in CASP14. *Proteins* **89**, 1687–1699 (2021).

263. Jumper, J. *et al.* Applying and improving AlphaFold at CASP14. *Proteins: Structure, Function and Bioinformatics* **89**, 1711–1721 (2021).

264. Kwon, S., Won, J., Kryshtafovych, A. & Seok, C. Assessment of protein model structure accuracy estimation in CASP14: Old and new challenges. *Proteins: Structure, Function and Bioinformatics* **89**, 1940–1948 (2021).

265. Dong, S. *et al.* A guideline for homology modeling of the proteins from newly discovered betacoronavirus, 2019 novel coronavirus (2019-nCoV). *J Med Virol* **92**, 1542–1548 (2020).

266. Wu, C. *et al.* Analysis of therapeutic targets for SARS-CoV-2 and discovery of potential drugs by computational methods. *Acta Pharm Sin B* **10**, 766–788 (2020).

267. Bassani, D., Ragazzi, E., Lapolla, A., Sartore, G. & Moro, S. Omicron Variant of SARS-CoV-2 Virus: In Silico Evaluation of the Possible Impact on People Affected by Diabetes Mellitus. *Front Endocrinol (Lausanne)* **13**, 284 (2022).

268. Gan, H. H., Twaddle, A., Marchand, B. & Gunsalus, K. C. Structural Modeling of the SARS-CoV-2 Spike/Human ACE2 Complex Interface can Identify High-Affinity Variants Associated with Increased Transmissibility. *J Mol Biol* **433**, (2021).

269. Zhao, P. *et al.* Virus-Receptor Interactions of Glycosylated SARS-CoV-2 Spike and Human ACE2 Receptor. *Cell Host Microbe* **28**, 586-601.e6 (2020).

270. Bai, C. *et al.* Predicting Mutational Effects on Receptor Binding of the Spike Protein of SARS-CoV-2 Variants. *J Am Chem Soc* **143**, 17646–17654 (2021).

271. Shahhosseini, N., Babuadze, G., Wong, G. & Kobinger, G. P. Mutation signatures and in silico docking of novel sars-cov-2 variants of concern. *Microorganisms* **9**, (2021).

272. Pavan, M., Bassani, D., Sturlese, M. & Moro, S. Bat coronaviruses related to SARS-CoV-2: what about their 3CL proteases (MPro)? *https://doi.org/10.1080/14756366.2022.2062336* **37**, 1077–1082 (2022).

273. Martin, R. W. *et al.* Sequence characterization and molecular modeling of clinically relevant variants of the SARS-CoV-2 main protease. *Biochemistry* **59**, 3741–3756 (2020).

274. Huang, X., Zhang, C., Pearce, R., Omenn, G. S. & Zhang, Y. Identifying the Zoonotic Origin of SARS-CoV-2 by Modeling the Binding Affinity between the Spike Receptor-Binding Domain and Host ACE2. *J Proteome Res* **19**, 4844–4856 (2020).

275. Rodrigues, J. P. G. L. M. *et al.* Insights on cross-species transmission of SARS-CoV-2 from structural modeling. *PLoS Comput Biol* **16**, (2020).

276. Piplani, S., Singh, P. K., Winkler, D. A. & Petrovsky, N. In silico comparison of SARS-CoV-2 spike protein-ACE2 binding affinities across species and implications for virus origin. *Sci Rep* **11**, (2021).

277. Sharma, P. *et al.* Genomic and structural mechanistic insight to reveal the differential infectivity of omicron and other variants of concern. *Comput Biol Med* **150**, (2022).

278. Jacob, J. J. *et al.* Evolutionary Tracking of SARS-CoV-2 Genetic Variants Highlights an Intricate Balance of Stabilizing and Destabilizing Mutations. *mBio* **12**, (2021).

279. Luo, R., Delaunay-Moisan, A., Timmis, K. & Danchin, A. SARS-CoV-2 biology and variants: anticipation of viral evolution and what needs to be done. *Environ Microbiol* **23**, 2339–2363 (2021).

280. Ghosh, A. K., Brindisi, M., Shahabi, D., Chapman, M. E. & Mesecar, A. D. Drug Development and Medicinal Chemistry Efforts toward SARS-Coronavirus and Covid-19 Therapeutics. *ChemMedChem* **15**, 907–932 (2020).

281. Ghosh, A. K., Mishevich, J. L., Mesecar, A. & Mitsuya, H. Recent Drug Development and Medicinal Chemistry Approaches for the Treatment of SARS-CoV-2 Infection and COVID-19. *ChemMedChem* **17**, e202200440 (2022).

282. Tiwari, V., Beer, J. C., Sankaranarayanan, N. V., Swanson-Mungerson, M. & Desai, U. R. Discovering small-molecule therapeutics against SARS-CoV-2. *Drug Discov Today* **25**, 1535–1544 (2020).

283. Adamson, C. S. *et al.* Antiviral drug discovery: preparing for the next pandemic. *Chem Soc Rev* **50**, 3647–3655 (2021).

284. Consortium, T. C. M. *et al.* Open Science Discovery of Oral Non-Covalent SARS-CoV-2 Main Protease Inhibitor Therapeutics. *bioRxiv* 2020.10.29.339317 (2022) doi:10.1101/2020.10.29.339317.

285. Kuntz, I. D. Structure-Based Strategies for Drug Design and Discovery. *Science (1979)* **257**, 1078–1082 (1992).

286. Forli, S. Charting a path to success in virtual screening. *Molecules* **20**, 18732–18758 (2015).

287. Halgren, T. A. Identifying and characterizing binding sites and assessing druggability. *J Chem Inf Model* **49**, 377–389 (2009).

288. Liang, J., Edelsbrunner, H. & Woodward, C. Anatomy of protein pockets and cavities: Measurement of binding site geometry and implications for ligand design. *Protein Science* **7**, 1884–1897 (1998).

289. Kitchen, D. B., Decornez, H., Furr, J. R. & Bajorath, J. Docking and scoring in virtual screening for drug discovery: Methods and applications. *Nat Rev Drug Discov* **3**, 935–949 (2004).

290. Kuntz, I. D., Blaney, J. M., Oatley, S. J., Langridge, R. & Ferrin, T. E. A geometric approach to macromolecule-ligand interactions. *J Mol Biol* **161**, 269–288 (1982).

291. Meng, X.-Y., Zhang, H.-X., Mezei, M. & Cui, M. Molecular Docking: A Powerful Approach for Structure-Based Drug Discovery. *Current Computer Aided-Drug Design* **7**, 146–157 (2012).

292. Salmaso, V. & Moro, S. Bridging molecular docking to molecular dynamics in exploring ligand-protein recognition process: An overview. *Front Pharmacol* **9**, 1–16 (2018).

293. Halperin, I., Ma, B., Wolfson, H. & Nussinov, R. Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins: Structure, Function and Genetics* **47**, 409–443 (2002).

294. Warren, G. L. *et al.* A critical assessment of docking programs and scoring functions. *J Med Chem* **49**, 5912–5931 (2006).

295. Bassani, D., Pavan, M., Bolcato, G., Sturlese, M. & Moro, S. Re-Exploring the Ability of Common Docking Programs to Correctly Reproduce the Binding Modes of Non-Covalent Inhibitors of SARS-CoV-2 Protease Mpro. *Pharmaceuticals* **15**, 180 (2022).

296. Ton, A. T., Gentile, F., Hsing, M., Ban, F. & Cherkasov, A. Rapid Identification of Potential Inhibitors of SARS-CoV-2 Main Protease by Deep Docking of 1.3 Billion Compounds. *Mol Inform* **39**, 2000028 (2020).

297. Acharya, A. *et al.* Supercomputer-Based Ensemble Docking Drug Discovery Pipeline with Application to Covid-19. *J Chem Inf Model* **60**, 5832–5852 (2020).

298. Gorgulla, C. *et al.* A multi-pronged approach targeting SARS-CoV-2 proteins using ultra-large virtual screening. *iScience* **24**, (2021).

299. Manelfi, C. *et al.* Combining Different Docking Engines and Consensus Strategies to Design and Validate Optimized Virtual Screening Protocols for the SARS-CoV-2 3CL Protease. *Molecules 2021, Vol. 26, Page 797* **26**, 797 (2021).

300. Corona, A. *et al.* Natural Compounds Inhibit SARS-CoV-2 nsp13 Unwinding and ATPase Enzyme Activities. *ACS Pharmacol Transl Sci* **5**, 226–239 (2022).

301. Verdonk, M. L., Cole, J. C., Hartshorn, M. J., Murray, C. W. & Taylor, R. D. Improved protein-ligand docking using GOLD. *Proteins* **52**, 609–623 (2003).

302. Kolarič, A., Jukič, M. & Bren, U. Novel Small-Molecule Inhibitors of the SARS-CoV-2 Spike Protein Binding to Neuropilin 1. *Pharmaceuticals* **15**, 165 (2022).

303. Morris, G. M. *et al.* AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility. *J Comput Chem* **30**, 2785 (2009).

304. Vatansever, E. C. *et al.* Bepridil is potent against SARS-CoV-2 in vitro. *Proc Natl Acad Sci U S A* **118**, e2012201118 (2021).

305. Neves, M. A. C., Totrov, M. & Abagyan, R. Docking and scoring with ICM: the benchmarking results and strategies for improvement. *J Comput Aided Mol Des* **26**, 675–686 (2012).

306. Lam, P. C. H., Abagyan, R. & Totrov, M. Ligand-biased ensemble receptor docking (LigBEnD): a hybrid ligand/receptor structure-based approach. *J Comput Aided Mol Des* **32**, 187–198 (2018).

307. Trott, O. & Olson, A. J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading. *J Comput Chem* **31**, 455 (2010).

308. Kao, H. T., Orry, A., Palfreyman, M. G. & Porton, B. Synergistic interactions of repurposed drugs that inhibit Nsp1, a major virulence factor for COVID-19. *Scientific Reports 2022 12:1* **12**, 1–15 (2022).

309. Friesner, R. A. *et al.* Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J Med Chem* **47**, 1739–1749 (2004).

310. Zhang, Y. *et al.* Structure-Based Discovery and Structural Basis of a Novel Broad-Spectrum Natural Product against the Main Protease of Coronavirus. *J Virol* **96**, 1253–1274 (2022).

311. Huff, S. *et al.* Discovery and Mechanism of SARS-CoV-2 Main Protease Inhibitors. *J Med Chem* **65**, 2866–2879 (2022).

312. Liu, K. *et al.* Clinical HDAC Inhibitors Are Effective Drugs to Prevent the Entry of SARS-CoV2. *ACS Pharmacol Transl Sci* **3**, 1361–1370 (2020).

313. Rao, S. N., Head, M. S., Kulkarni, A. & LaLonde, J. M. Validation studies of the site-directed docking program LibDock. *J Chem Inf Model* **47**, 2159–2171 (2007).

314. Wang, G. *et al.* Dalbavancin binds ACE2 to block its interaction with SARS-CoV-2 spike protein and is effective in inhibiting SARS-CoV-2 infection in animal models. *Cell Research 2020 31:1* **31**, 17–24 (2020).

315. Luttens, A. *et al.* Ultralarge Virtual Screening Identifies SARS-CoV-2 Main Protease Inhibitors with Broad-Spectrum Activity against Coronaviruses. *J Am Chem Soc* **144**, 2905–2920 (2022).

316. Cross, S. S. J. Improved FlexX docking using FlexS-determined base fragment placement. *J Chem Inf Model* **45**, 993–1001 (2005).

317. Welker, A. *et al.* Structure-Activity Relationships of Benzamides and Isoindolines Designed as SARS-CoV Protease Inhibitors Effective against SARS-CoV-2. *ChemMedChem* **16**, 340–354 (2021).

318. Otava, T. *et al.* The Structure-Based Design of SARS-CoV-2 nsp14 Methyltransferase Ligands Yields Nanomolar Inhibitors. *ACS Infect Dis* **7**, 2214–2220 (2021).

319. Wang, Y. T. *et al.* Novel nucleocapsid protein-targeting phenanthridine inhibitors of SARS-CoV-2. *Eur J Med Chem* **227**, 113966 (2022).

320. Chen, Y. C. Beware of docking! *Trends Pharmacol Sci* **36**, 78–95 (2015).

321. Llanos, M. A. *et al.* Strengths and Weaknesses of Docking Simulations in the SARS-CoV-2 Era: The Main Protease (Mpro) Case Study. *J Chem Inf Model* **61**, 3758–3770 (2021).

322. Chaput, L. & Mouawad, L. Efficient conformational sampling and weak scoring in docking programs? Strategy of the wisdom of crowds. *J Cheminform* **9**, 37 (2017).

323. Neves, B. J. *et al.* Best Practices for Docking-Based Virtual Screening. *Molecular Docking for Computer-Aided Drug Design: Fundamentals, Techniques, Resources and Applications* 75–98 (2021) doi:10.1016/B978-0-12-822312-3.00001-1.

324. Alonso, H., Bliznyuk, A. A. & Gready, J. E. Combining docking and molecular dynamic simulations in drug design. *Med Res Rev* **26**, 531–568 (2006).

325. Hollingsworth, S. A. & Dror, R. O. Molecular dynamics simulation for all. *Neuron* **99**, 1129 (2018).

326. Karplus, M. & McCammon, J. A. Molecular dynamics simulations of biomolecules. *Nature Structural Biology 2002 9:9* **9**, 646–652 (2002).

327. Durrant, J. D. & McCammon, J. A. Molecular dynamics simulations and drug discovery. *BMC Biol* **9**, 1–9 (2011).

328. Ferreira, L. G., dos Santos, R. N., Oliva, G. & Andricopulo, A. D. Molecular Docking and Structure-Based Drug Design Strategies. *Molecules 2015, Vol. 20, Pages 13384-13421* **20**, 13384–13421 (2015).

329. Tan, L., Batista, J. & Bajorath, J. Computational methodologies for compound database searching that utilize experimental protein-ligand interaction information. *Chem Biol Drug Des* **76**, 191–200 (2010).

330. Peach, M. L. & Nicklaus, M. C. Combining docking with pharmacophore filtering for improved virtual screening. *J Cheminform* **1**, (2009).

331. Muthas, D., Sabnis, Y. A., Lundborg, M. & Karlén, A. Is it possible to increase hit rates in structure-based virtual screening by pharmacophore filtering? An investigation of the advantages and pitfalls of post-filtering. *J Mol Graph Model* **26**, 1237–1251 (2008).

332. Rácz, A., Bajusz, D. & Héberger, K. Life beyond the Tanimoto coefficient: Similarity measures for interaction fingerprints. *J Cheminform* **10**, (2018).

333. Pavan, M., Menin, S., Bassani, D., Sturlese, M. & Moro, S. Implementing a Scoring Function Based on Interaction Fingerprint for Autogrow4: Protein Kinase CK1δ as a Case Study. *Front Mol Biosci* **0**, 629 (2022).

334. Wang, H. *et al.* Identification of highly effective inhibitors against SARS-CoV-2 main protease: From virtual screening to in vitro study. *Front Pharmacol* **13**, 4934 (2022).

335. Tian, X. *et al.* Discovery of Novel and Highly Potent Inhibitors of SARS CoV-2 Papain-Like Protease Through Structure-Based Pharmacophore Modeling, Virtual Screening, Molecular Docking, Molecular Dynamics Simulations, and Biological Evaluation. *Front Pharmacol* **13**, 16 (2022).

336. Yin, S. *et al.* Non-covalent cyclic peptides simultaneously targeting Mpro and NRP1 are highly effective against Omicron BA.2.75. *Front Pharmacol* **13**, 4723 (2022).

337. Gossen, J. *et al.* A Blueprint for High Affinity SARS-CoV-2 Mpro Inhibitors from Activity-Based Compound Library Screening Guided by Analysis of Protein Dynamics. *ACS Pharmacol Transl Sci* **4**, 1079–1095 (2021).

338. Hu, X. *et al.* Discovery of Small Molecule Entry Inhibitors Targeting the Fusion Peptide of SARS-CoV-2 Spike Protein. *ACS Med Chem Lett* **12**, 1267–1274 (2021).

339. Jang, W. D., Jeon, S., Kim, S. & Lee, S. Y. Drugs repurposed for COVID-19 by virtual screening of 6,218 drugs and cell-based assay. *Proc Natl Acad Sci U S A* **118**, e2024302118 (2021).

340. McGovern, S. L. & Shoichet, B. K. Information decay in molecular docking screens against holo, apo, and modeled conformations of enzymes. *J Med Chem* **46**, 2895–2907 (2003).

341. Salmaso, V., Sturlese, M., Cuzzolin, A. & Moro, S. Combining self- and cross-docking as benchmark tools: the performance of DockBench in the D3R Grand Challenge 2. *J Comput Aided Mol Des* **32**, 251–264 (2018).

342. Korb, O. *et al.* Potential and limitations of ensemble docking. *J Chem Inf Model* **52**, 1262–1274 (2012).

343. Knegtel, R. M. A., Kuntz, I. D. & Oshiro, C. M. Molecular docking to ensembles of protein structures. *J Mol Biol* **266**, 424–440 (1997).

344. Huang, S. Y. & Zou, X. Ensemble docking of multiple protein structures: Considering protein structural variations in molecular docking. *Proteins: Structure, Function and Genetics* **66**, 399–421 (2007).

345. Wang, R. & Wang, S. How Does Consensus Scoring Work for Virtual Library Screening? An Idealized Computer Experiment. *J Chem Inf Comput Sci* **41**, 1422–1426 (2001).

346. Charifson, P. S., Corkery, J. J., Murcko, M. A. & Walters, W. P. Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *J Med Chem* **42**, 5100–5109 (1999).

347. Bissantz, C., Folkers, G. & Rognan, D. Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. *J Med Chem* **43**, 4759–4767 (2000).

348. McGann, M. FRED and HYBRID docking performance on standardized datasets. *J Comput Aided Mol Des* **26**, 897–906 (2012).

349. Gimeno, A. *et al.* Prediction of Novel Inhibitors of the Main Protease (M-pro) of SARS-CoV-2 through Consensus Docking and Drug Reposition. *International Journal of Molecular Sciences 2020, Vol. 21, Page 3793* **21**, 3793 (2020).

350. Yang, J. *et al.* Structure-Based Discovery of Novel Nonpeptide Inhibitors Targeting SARS-CoV-2 Mpro. *J Chem Inf Model* **61**, 3917–3926 (2021).

351. Alhossary, A., Handoko, S. D., Mu, Y. & Kwoh, C. K. Fast, accurate, and reliable molecular docking with QuickVina 2. *Bioinformatics* **31**, 2214–2216 (2015).

352. Rubio-Martínez, J. *et al.* Discovery of Diverse Natural Products as Inhibitors of SARS-CoV-2 MproProtease through Virtual Screening. *J Chem Inf Model* **61**, 6094–6106 (2021).

353. Clyde, A. *et al.* High-Throughput Virtual Screening and Validation of a SARS-CoV-2 Main Protease Noncovalent Inhibitor. *J Chem Inf Model* **62**, 116–128 (2022).

354. Glaab, E., Manoharan, G. B. & Abankwa, D. Pharmacophore Model for SARS-CoV-2 3CLpro Small-Molecule Inhibitors and in Vitro Experimental Validation of Computationally Screened Inhibitors. *J Chem Inf Model* **61**, 4082–4096 (2021).

355. Ghahremanpour, M. M. *et al.* Identification of 14 Known Drugs as Inhibitors of the Main Protease of SARS-CoV-2. *ACS Med Chem Lett* **11**, 2526–2533 (2020).

356. Wang, L. *et al.* Accurate and reliable prediction of relative ligand binding potency in prospective drug discovery by way of a modern free-energy calculation protocol and force field. *J Am Chem Soc* **137**, 2695–2703 (2015).

357. Rastelli, G., Degliesposti, G., del Rio, A. & Sgobba, M. Binding estimation after refinement, a new automated procedure for the refinement and rescoring of docked ligands in virtual screening. *Chem Biol Drug Des* **73**, 283–286 (2009).

358. Jespers, W., Åqvist, J. & Gutiérrez-de-Terán, H. Free Energy Calculations for Protein–Ligand Binding Prediction. *Methods in Molecular Biology* **2266**, 203–226 (2021).

359. Hou, T., Wang, J., Li, Y. & Wang, W. Assessing the performance of the molecular mechanics/Poisson Boltzmann surface area and molecular mechanics/generalized Born surface area methods. II. The accuracy of ranking poses generated from docking. *J Comput Chem* **32**, 866–877 (2011).

360. Zhang, C. H. *et al.* Potent Noncovalent Inhibitors of the Main Protease of SARS-CoV-2 from Molecular Sculpting of the Drug Perampanel Guided by Free Energy Perturbation Calculations. *ACS Cent Sci* **7**, 467–475 (2021).

361. Li, Z. *et al.* Identify potent SARS-CoV-2 main protease inhibitors via accelerated free energy perturbation-based virtual screening of existing drugs. *Proc Natl Acad Sci U S A* **117**, 27381–27387 (2020).

362. Ngo, S. T., Tam, N. M., Pham, M. Q. & Nguyen, T. H. Benchmark of Popular Free Energy Approaches Revealing the Inhibitors Binding to SARS-CoV-2 Mpro. *J Chem Inf Model* **61**, 2302–2312 (2021).

363. Sherman, W., Beard, H. S. & Farid, R. Use of an induced fit receptor structure in virtual screening. *Chem Biol Drug Des* **67**, 83–84 (2006).

364. Genheden, S. & Ryde, U. The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opin Drug Discov* **10**, 449 (2015).

365. Ibrahim, I. M., Elfiky, A. A., Fathy, M. M., Mahmoud, S. H. & ElHefnawi, M. Targeting SARS-CoV-2 endoribonuclease: a structure-based virtual screening supported by in vitro analysis. *Scientific Reports 2022 12:1* **12**, 1–11 (2022).

366. Copeland, R. A., Pompliano, D. L. & Meek, T. D. Drug-target residence time and its implications for lead optimization. *Nat Rev Drug Discov* **5**, 730–739 (2006).

367. Bernetti, M., Masetti, M., Rocchia, W. & Cavalli, A. Kinetics of Drug Binding and Residence Time. *https://doi.org/10.1146/annurev-physchem-042018-052340* **70**, 143–171 (2019).

368. Pavan, M., Menin, S., Bassani, D., Sturlese, M. & Moro, S. Qualitative Estimation of Protein–Ligand Complex Stability through Thermal Titration Molecular Dynamics Simulations. *J Chem Inf Model* (2022) doi:10.1021/ACS.JCIM.2C00995.

369. Zaidman, D. *et al.* An automatic pipeline for the design of irreversible derivatives identifies a potent SARS-CoV-2 Mpro inhibitor. *Cell Chem Biol* **28**, 1795-1806.e5 (2021).

370. Valiente, P. A. *et al.* Computational Design of Potent D-Peptide Inhibitors of SARS-CoV-2. *J Med Chem* **64**, 14955–14967 (2021).

371. Dehouck, Y., Kwasigroch, J. M., Rooman, M. & Gilis, D. BeAtMuSiC: Prediction of changes in protein-protein binding affinity on mutations. *Nucleic Acids Res* **41**, (2013).

372. Kandeel, M. *et al.* Discovery of New Fusion Inhibitor Peptides against SARS-CoV-2 by Targeting the Spike S2 Subunit. *Biomol Ther (Seoul)* **29**, 282–289 (2021).

373. Leman, J. K. *et al.* Macromolecular modeling and design in Rosetta: recent methods and frameworks. *Nat Methods* **17**, 665–680 (2020).

374. Jeong, B. S. *et al.* Computational design of a neutralizing antibody with picomolar binding affinity for all concerning SARS-CoV-2 variants. *MAbs* **14**, (2022).

375. Sun, M. *et al.* Aptamer Blocking Strategy Inhibits SARS-CoV-2 Virus Infection. *Angewandte Chemie International Edition* **60**, 10266–10272 (2021).

376. Dou, J. *et al.* De novo design of a fluorescence-activating β-barrel. *Nature* **561**, 485–491 (2018).

377. Cao, L. *et al.* De novo design of picomolar SARS-CoV-2 miniprotein inhibitors. *Science (1979)* **370**, (2020).

378. Kortemme, T., Kim, D. E. & Baker, D. Computational alanine scanning of protein-protein interfaces. *Sci STKE* **2004**, (2004).

379. Kortemme, T. & Baker, D. A simple physical model for binding energy hot spots in protein-protein complexes. *Proc Natl Acad Sci U S A* **99**, 14116–14121 (2002).

380. Glasgow, A. *et al.* Engineered ACE2 receptor traps potently neutralize SARS-CoV-2. *Proc Natl Acad Sci U S A* **117**, 28046–28055 (2020).

381. de Vivo, M., Masetti, M., Bottegoni, G. & Cavalli, A. Role of Molecular Dynamics and Related Methods in Drug Discovery. *J Med Chem* **59**, (2016).

382. Ferrari, F. *et al.* HT-SuMD: Making Molecular Dynamics Simulations Suitable for Fragment-Based Screening. a Comparative Study with NMR. (2020) doi:10.26434/CHEMRXIV.12582662.V1.

383. Verdonk, M. L. *et al.* Docking performance of fragments and druglike compounds. *J Med Chem* **54**, 5422–5431 (2011).

384. Bissaro, M. *et al.* Inspecting the Mechanism of Fragment Hits Binding on SARS-CoV-2 M ^{pro} by Using Supervised Molecular Dynamics (SuMD) Simulations. *ChemMedChem* **16**, 2075–2081 (2021).

385. Sabbadin, D. & Moro, S. Supervised molecular dynamics (SuMD) as a helpful tool to depict GPCR-ligand recognition pathway in a nanosecond time scale. *J Chem Inf Model* **54**, 372–376 (2014).

386. Cuzzolin, A. *et al.* Deciphering the Complexity of Ligand-Protein Recognition Pathways Using Supervised Molecular Dynamics (SuMD) Simulations. *J Chem Inf Model* **56**, 687–705 (2016).

387. Pavan, M., Bolcato, G., Bassani, D., Sturlese, M. & Moro, S. Supervised Molecular Dynamics (SuMD) Insights into the mechanism of action of SARS-CoV-2 main protease inhibitor PF-07321332. *J Enzyme Inhib Med Chem* **36**, 1646–1650 (2021).

388. Pavan, M., Bassani, D., Sturlese, M. & Moro, S. Investigating RNA–protein recognition mechanisms through supervised molecular dynamics (SuMD) simulations. *NAR Genom Bioinform* **4**, (2022).

AIM OF THE WORK

As stated in the introduction, the COVID-19 outbreak in December 2019 drastically changed the lives of many people around the world. The social and economic crisis generated by this pandemic had such drastic consequences that many members of the scientific community decided to devoid most of their efforts into increasing the knowledge about this topic. This situation resulted in unprecedented cooperative action in an attempt to put an end to this unpleasant situation.

The present work was conceived within this framework, aiming to evaluate the usefulness of existing structure-based computer-aided drug discovery (CADD) strategies within a pandemic scenario, where reduced timescales are required compared to traditional drug discovery campaigns. Our effort served two different purposes: the first and most immediate one was to provide helpful information for developing novel candidate drugs against COVID-19, especially concerning inhibitors of the SARS-CoV-2 main protease. Instead, the second and more long-term oriented one was to assess the helpfulness and reliability of CADD techniques in facing future similar scenarios. Indeed, a study published in PNAS[1] estimated that the likelihood of a highly infectious disease epidemic could double in the coming decades. So, the resurgence of a COVID-19-like scenario is a question of when rather than whether, thus requiring adequate preparation by everyone.

Furthermore, another critical part of the present work was dedicated to promoting methodological advancement, either by extending the applicability domain of existing techniques to new fields or developing completely novel methodologies, aiming to compensate for the weakness of existing ones.

The development of RNA-based therapeutics is becoming more and more relevant these days, as demonstrated by the pivotal role played by mRNA vaccines against COVID-19. We present a novel implementation of Supervised Molecular Dynamics (SuMD)[2], an enhanced sampling approach for studying molecular recognition processes, applied to RNA-protein complexes. Our SuMD-based pipeline will provide the community with a valuable tool to help rationalize RNA-based therapeutics design, which will soon play a prominent role in tackling untreated diseases.

Finally, one major issue that all drug discovery campaigns face, regardless of the nature of investigated molecules, is related to the accurate prediction of drug efficacy. This parameter has been increasingly linked to kinetic properties concerning the drug-target interaction, generating high interest in developing novel methods for predicting such properties. In this work, we developed an entirely new platform named Thermal Titration Molecular Dynamics (TTMD)[3], an enhanced sampling technique for unbinding kinetics estimation. Thanks to its swiftness and ease of use, TTMD will hopefully increase the adoption of CADD techniques in the field of drug-target kinetic investigation.

1. Marani, M., Katul, G. G., Pan, W. K. & Parolari, A. J. Intensity and frequency of extreme novel epidemics. *Proc Natl Acad Sci U S A* **118**, e2105482118 (2021).

2. Pavan, M., Bassani, D., Sturlese, M. & Moro, S. Investigating RNA–protein recognition mechanisms through supervised molecular dynamics (SuMD) simulations. *NAR Genom Bioinform* **4**, (2022).

3. Pavan, M., Menin, S., Bassani, D., Sturlese, M. & Moro, S. Qualitative Estimation of Protein–Ligand Complex Stability through Thermal Titration Molecular Dynamics Simulations. *J Chem Inf Model* (2022) doi:10.1021/ACS.JCIM.2C00995.

METHODS OVERVIEW

During the present work, several different computational methods belonging to the field of structure-based drug discovery were used. This section briefly summarizes the various approaches employed. A more thorough and referenced overview of the topic can be found in the “Introduction” as well as within each publication where these methods have been used, while information about the techniques used within each work can be found in the “Articles overview” section.

- **Homology modeling / de novo prediction**: in the absence of an experimentally determined three-dimensional structure for a biological target of interest, it can be predicted with various degrees of accuracy either through homology modeling or de novo prediction. In the case of homology modeling, the structure of the target macromolecule is modeled using as a template a protein with conserved function and high sequence identity to the target, provided that its structure has already been determined experimentally. This method assumes that proteins with high sequence identity and similar functions will fold similarly. In the case of de novo prediction, the protein fold is guessed solely based on the primary sequence of the target macromolecule, with limited to no exploitation of the knowledge about the structure of homologous macromolecules.
- **Molecular docking**: one of the crucial tasks in the rational design of drugs is being able to predict how a certain compound (the ligand) will bind to the active site of a target macromolecule (the receptor). This is usually accomplished through molecular docking, a method for determining the most plausible orientation of a ligand within a binding site that operates through the combination of a search algorithm and a scoring function. The search algorithm generates a series of possible ligand conformations within a user-defined search space, usually a sphere or a rectangular prism box including the binding site. The scoring function, instead, evaluates the quality of protein-ligand interactions for each pose and attributes a score which is used to rank the conformations generated by the search algorithm. At the end of each docking run, the user retrieves a series of poses (ligand conformations within the binding sites), ranked based on the docking score.

- **Pharmacophore model**: the knowledge about the interaction features required for selective and potent binding to a specific biological target can be encoded into a three-dimensional model, i.e. a pharmacophore. A pharmacophore model is a set of spheres that defines the position in the space of each functional group that is needed for defining the minimum set of intermolecular interactions necessary for a compound to be active towards a target of interest. Once built, this model can then be used to perform virtual screening, i.e. selecting compounds within a virtual library based on the possibility to portray this set of features, or to filter docking poses, retaining only those poses which match the pharmacophore model.
- **Protein-Ligand Interaction Fingerprint**: another way to encode information about binding features of active compounds against specific biological targets of interest is to construct protein-ligand interaction fingerprints (PLIFs). Various types of PLIFs exist, depending on the way the information is encoded, but generally speaking, protein-ligand interactions are translated into numeric vectors/matrices where each cell accounts for a different type of interaction and the numeric value represents its frequency. PLIFs can be compared using different similarity metrics, producing scores that can be used to rank compounds based on the congruence between their binding pattern.
- **Molecular Dynamics**: a pivotal aspect when working with biological systems is being able to study their time-dependent evolution and related properties. This computationally daunting task is performed by using molecular dynamics. In molecular dynamics simulations, molecular mechanics is used to discretize the description of biological systems in such a way as to make it feasible to compute the time-dependent evolution of a system at the atomic level for simulation times in the order of microseconds/milliseconds. Practically speaking, the presence of electrons is neglected, with their distribution being approximated by the attribution of punctiform partial charges on each atom, which is described as a sphere connected to other atoms by springs representing each covalent bond. The evolution of atomic coordinates follows the gradient of the potential energy of the system, calculated through a series of parametric equations known as a force field,

where different contributions to the potential energy of the system are weighed to accurately reproduce experimental data or calculations performed at a higher level of theory (quantum mechanics).

- S**upervised Molecular Dynamics**: the real-world time required to observe infrequent events such as association or dissociation processes between biological entities is usually too high to rely on classic, unbiased, molecular dynamics simulations. For this reason, several enhanced-sampling strategies have been developed. One of these algorithms, known as Supervised Molecular Dynamics, relies on performing a series of short molecular dynamics simulations (SuMD-steps) where the distance between the center of mass of the ligand and of the receptor is monitored and linearly interpolated: if the angular coefficient is negative, indicating the ligand is approaching the binding site in a given simulation window, the SuMD-step is retained for the generation of the final trajectory, if it is positive, the SuMD-step is discarded and the simulation is repeated until a negative coefficient is obtained. This method is particularly useful for studying association processes between a ligand and a receptor.
- **Thermal Titration Molecular Dynamics**: another enhanced sampling method for molecular dynamics simulations, which is useful for investigating dissociation processes, is Thermal Titration Molecular Dynamics. This technique monitors the conservation of the native binding mode for a certain receptor-ligand complex with a scoring function based on protein-ligand interaction fingerprints through a series of short molecular dynamics simulations performed at progressively increasing temperatures. The temperature increase is used to augment the kinetic energy of the system, to facilitate overcoming the energy barrier between the bound and unbound states.

SCIENTIFIC PUBLICATIONS

Articles overview

The present work is divided into three main sections. The first and most conspicuous part is dedicated to COVID-19-related scientific production, with a particular focus on the study of the SARS-CoV-2 main protease (M^{pro}/$3CL^{pro}$). The central portion is a collection of ancillary works, both in the field of neurodegeneration and cancer. The third and final section focuses, instead, on methodological projects. The pivotal points of the obtained research results are summarized in the current paragraph and detailed afterward through the presentation of the already published scientific publications.

Section 1: COVID-19

The first eight articles are all related to COVID-19, mainly focusing on the SARS-CoV-2 main protease (M^{pro}). As recalled in the introduction, M^{pro} is a validated target for the development of effective and safe COVID-19 pharmacological treatments, so an essential part of the present work has been dedicated to the study of recognition processes between small molecule ligands and the catalytic site of the enzyme through a combination of both time-dependent and time-independent structure-based approaches.

1. **Pavan, M., Bassani, D., Sturlese, M. & Moro, S. Bat coronaviruses related to SARS-CoV-2: what about their 3CL proteases (MPro)? *https://doi.org/10.1080/14756366.2022.2062336* 37, 1077–1082 (2022).**

   Temmam et al. reported the discovery of bat coronaviruses found in North Laos caverns that present a high degree of genome similarity with SARS-CoV-2 and are as much or even more infective than their human counterpart. This news raised important concerns in the general population about the possibility of a spillover and the rise of novel pandemic waves caused by these viruses or closely related ones. Complementing the original work, which fully focuses on the analysis of the spike protein, which mediates the capability of such viruses to infect human cells, in the present work, we used homology modeling to analyze the similarities and differences between the main proteases of these newly reported viruses and SARS-CoV-2, discussing their relevance relative to the efficacy of existing therapeutic approaches against COVID-19, particularly concerning the first orally available anti-SARS-CoV-2 drug (Paxlovid), and the development of future ones.

2. **Pavan, M., Bassani, D., Sturlese, M. & Moro, S. From the Wuhan-Hu-1 strain to the XD and XE variants: is targeting the SARS-CoV-2 spike protein still a pharmaceutically relevant option against COVID-19? *J Enzyme Inhib Med Chem* 37, 1704–1714 (2022).**

   Since the outbreak of the COVID-19 pandemic in December 2019, the SARS-CoV-2 genome has undergone several mutations, resulting in the rise of a multitude of viral variants, some of which have shown an increased pathogenic potential. The continuous rise of novel variants provoked skepticism in the general population regarding the efficiency and trustworthiness of existing treatments (vaccines, in particular) and the ones in development. In the present work, homology modeling has been extensively exploited to characterize the structural features of a set of variants of concerns, analyzing the different impact that the viral evolution process had on the two most important targets for the development of anti-SARS-CoV-2 treatments, namely the spike protein and the main protease.

3. **Heilmann, E. *et al.* SARS-CoV-2 3CL ^{pro} mutations selected in a VSV-based system confer resistance to nirmatrelvir, ensitrelvir, and GC376. *Sci Transl Med* (2022) doi:10.1126/scitranslmed.abq7360.**

   The emergence of novel SARS-CoV-2 strains threatens the efficacy of existing COVID-19 treatments, due to their possible antiviral resistance. To identify mutations that confer resistance to existing inhibitors of SARS-CoV-2 main protease, Heilmann et al. engineered a chimeric vesicular stomatitis virus (VSV) whose replication cycle was dependent on the autocatalytic processing of its precursor polyprotein by the main protease itself. This platform was then used to generate resistant mutants (some of which already circulating), particularly towards Nirmatrelvir, the active principle of the first approved orally available anti-COVID-19 drug (Paxlovid). In the present work, we exploited homology modeling and in silico alanine and resistance scanning to investigate the mechanistic rationale behind the Nirmatrelvir resistance.

4. **Fornasier, E. *et al.* A new inactive conformation of SARS-CoV-2 main protease. *Acta Crystallogr D Struct Biol* 78, 363–378 (2022).**

   Thanks to the early release to the scientific community of its three-dimensional structure and its pivotal role in the SARS-CoV-2 replication cycle, the main protease (M^{pro}) has become the most investigated target for developing drugs against COVID-19. During a screening campaign, a novel conformation of M^{pro} was observed for the first time. This structure adopts a new conformation of the oxyanion loop, a short motif involved in substrate recognition and enzymatic activity. Many fundamental interactions that stabilize the active site are lost compared to the canonical M^{pro} structure, resulting in the abolition of the catalytic activity. In this work, molecular dynamics simulations were employed to complement and support experimental observations about the novel conformation of the SARS-CoV-2 main protease.

5. **Bassani, D., Pavan, M., Bolcato, G., Sturlese, M. & Moro, S. Re-Exploring the Ability of Common Docking Programs to Correctly Reproduce the Binding Modes of Non-Covalent Inhibitors of SARS-CoV-2 Protease Mpro. *Pharmaceuticals* 15, 180 (2022).**

   Computational methodologies portray an essential role in the early stages of modern drug discovery campaigns, especially for identifying hit compounds and steering their development into mature leads. The remarkable amount of structural information about the SARS-CoV-2 main protease made it possible to exploit several structure-based approaches, such as molecular docking, for finding candidate inhibitors. Despite the good success rate of docking-based campaigns both for drug repurposing and for discovering novel active molecules against M^{pro}, an article by Zev et al. harshly critiqued the usefulness and reliability of docking runs performed on this target. In this work, we countered their observations by demonstrating how several commonly used docking algorithms are perfectly capable of reproducing most experimentally determined binding poses on the SARS-CoV-2 M^{pro}, except for very unstable and solvent-exposed ones.

6. **Bissaro, M. *et al.* Inspecting the Mechanism of Fragment Hits Binding on SARS-CoV-2 M ^{pro} by Using Supervised Molecular Dynamics (SuMD) Simulations. *ChemMedChem* 16, 2075–2081 (2021).**

   Computational methodologies which can characterize the binding mode of fragments are valuable assets for the success of fragment-based drug discovery campaigns. Concerning this, molecular docking represents a state-of-the-art technique. The low molecular complexity of fragments, combined with the important role of solvent in mediating their interactions with the target, impair the predictive power of docking, issuing the need for novel, more accurate, protocols. In this work, we showed how a refined version of the in-house developed High-Throughput Supervised Molecular Dynamics platform (HT-SuMD) can be successfully employed to rapidly and accurately determine the binding mode of fragment compounds to the catalytic site of the SARS-CoV-2 main protease (M^{pro}).

7. **Bolcato, G., Bissaro, M., Pavan, M., Sturlese, M. & Moro, S. Targeting the coronavirus SARS-CoV-2: computational insights into the mechanism of action of the protease inhibitors lopinavir, ritonavir and nelfinavir. *Sci Rep* 10, 20927 (2020).**

   In the early stages of the COVID-19 pandemic, the rapid spread of the virus combined with the lack of efficient pharmacological treatment, lead to several attempts at repurposing existing drugs. Particularly, the therapeutic combination of two HIV protease inhibitors, lopinavir, and ritonavir, was investigated in clinical trials, regardless of the lack of mechanistic insights regarding their anti-SARS-CoV-2 action at the time. To overcome this lack of information, and predict the validity of this therapeutic approach, we carried out Supervised Molecular Dynamics Simulations (SuMD) to study the binding mechanism of lopinavir and ritonavir to the SARS-CoV-2 main protease (M^{pro}), comparing their binding features with the ones of nelfinavir, an analog compound which showed promising in vitro activity. Follow-up experimental evidence supported our computational prediction, which speculated how nelfinavir would outperform both lopinavir and ritonavir, despite not being good enough for clinical use.

8. **Pavan, M., Bolcato, G., Bassani, D., Sturlese, M. & Moro, S. Supervised Molecular Dynamics (SuMD) Insights into the mechanism of action of SARS-CoV-2 main protease inhibitor PF-07321332. *J Enzyme Inhib Med Chem* 36, 1646–1650 (2021).**

   In April 2021, the chemical structure of PF-07321332 (Nirmatrelvir), the first orally available drug specifically designed against COVID-19 to reach the clinical stages of experimentation, was revealed to the general audience. No further information was provided about the interaction pattern between PF-07321332, which would eventually become the active principle of Paxlovid, the first anti-COVID-19 drug, and its biomolecular counterpart, the SARS-CoV-2 main protease (M^{pro}). In this work, we compensated for this shortage by elucidating the binding pattern of PF-07321332 to M^{pro} through Supervised Molecular Dynamics (SuMD) simulations.

   Successive experimental evidence confirmed the SuMD-predicted binding mode, other than indirectly confirming certain details about the whole recognition process.

Section 2: neurodegeneration and cancer

The second part collects seven articles in which the expertise acquired through the production of COVID-19-related work has been applied to different side-projects, ranging from the field of neurodegeneration to the one of cancer. Concerning neurodegeneration, efforts have been mainly devolved to the implementation of a computational pipeline for the fragment-based development of novel inhibitors of casein kinase 1, with a final glimpse into future therapeutic perspectives regarding the relatively unexplored world of GPCRs. Particularly, we concentrated on adenosine receptors, a class of GPCRs that are involved in various diseases, involving both neurodegeneration and cancer. Regarding cancer, methodological and applicative works are presented, involving both a classic typology of targets (GPCRs) and a novel type, i.e., the mitochondrial ribosome.

9. **Pavan, M. *et al.* Computational Strategies to Identify New Drug Candidates against Neuroinflammation. *Curr Med Chem* 29, 4756–4775 (2022).**

   Within the frame of contemporary drug discovery campaigns, computational approaches have been increasingly and successfully adopted. In the field of neuroinflammation, some intrinsic issues of the task, such as the peculiar anatomical localization and the presence of the blood-brain barrier, require the adoption of measures to fine-tune the physicochemical properties of candidate drugs from the early stages of the process. In this literature review, we provide an overview of the start-of-the-art computational strategies that can be exploited for the rational design of novel small molecules controlling neuroinflammation, particularly focusing on those which exploit information about the three-dimensional structure of the biological target of interest.

10. **Bolcato, G. *et al.* A Computational Workflow for the Identification of Novel Fragments Acting as Inhibitors of the Activity of Protein Kinase CK1δ. *Int J Mol Sci* 22, 9741 (2021).**

    In the last twenty years, Fragments-Based Drug Discovery has become a prominent paradigm for the rational design of novel active molecules, especially towards targets difficult to address with traditional pipelines. One example of successful implementation of this strategy is represented by kinases, as represented by the approval of the BRAF inhibitor vemurafenib. In the Kinase family, protein kinase CK1 isoform δ (CK1δ) has become a promising target in the treatment of different neurodegenerative diseases such as Alzheimer's disease, Parkinson's disease, and amyotrophic lateral sclerosis. In the present work, we set up a computational workflow based on the combination of molecular docking, pharmacophore filter, and molecular dynamics simulations for the identification of putative fragment binders in large virtual databases. The method was successfully applied to the screening of several libraries of commercially available compounds, leading to the identification of 7 novel fragment inhibitors of CK1δ.

11. **Pavan, M., Menin, S., Bassani, D., Sturlese, M. & Moro, S. Implementing a Scoring Function Based on Interaction Fingerprint for Autogrow4: Protein Kinase CK1δ as a Case Study. *Front Mol Biosci* 0, 629 (2022).**

    Fragment-Based Drug Discovery is an attractive approach for the design of novel active molecules against difficult targets. While several strategies, both computational and experimental, are readily available for the initial screening campaign, leading to the identification of fragment hits, novel pipelines for the rational hit-to-lead optimization are needed, to render the process smoother and more time and cost-efficient. One example is represented by Autogrow, a semi-automated computational protocol that exploits a combination between a genetic algorithm and molecular docking for de novo drug design and lead optimization. In the present work, we present a customized version of this software which implement a new scoring function based on protein-ligand interaction fingerprints. The protocol was applied to the generation of novel fragment-based CK1δ potential inhibitors with high similarity of shape and electrostatic features to known inhibitors used as reference.

12. **Bassani, D. *et al.* The Multifaceted Role of GPCRs in Amyotrophic Lateral Sclerosis: A New Therapeutic Perspective? *International Journal of Molecular Sciences 2022, Vol. 23, Page 4504* 23, 4504 (2022).**

    Amyotrophic lateral sclerosis (ALS) is a degenerating disease involving the motor neurons, which causes a progressive loss of movement ability, usually leading to death within 2 to 5 years from the diagnosis. Much effort has been put into research for an effective therapy for its eradication, but still, no cure is available. The only two drugs approved for this pathology, Riluzole, and Edaravone, are only able to slow down the inevitable disease progression. As assessed in the literature, drug targets such as protein kinases have already been extensively examined as potential drug targets for ALS, with some molecules already in clinical trials. In this literature review, we provide an overview of the involvement of another very important and studied class of biological entities, G protein-coupled receptors (GPCRs), in the onset and progression of ALS.

13. **Bassani, D., Pavan, M., Sturlese, M. & Moro, S. Sodium or Not Sodium: Should Its Presence Affect the Accuracy of Pose Prediction in Docking GPCR Antagonists? *Pharmaceuticals* 15, 346 (2022).**

    The presence of a sodium ion that acts as an allosteric modulator to stabilize the inactive conformation of class A GPCRs has been a staple point of pharmacology for a while. Several studies highlighted how this ion is essential for the binding of ligands acting as antagonists at the orthosteric site. From a structural point of view, not all experimentally determined class A GPCRs structures present this ion in its usual location. This usually requires manual editing of the structure by molecular modelers, by inserting the missing ion at its usual location. After examining in the past the influence of the allosteric sodium in both docking and molecular dynamics simulations of adenosine receptors, in the present work, we decided to extend our investigation to all other class A GPCRs whose structures in complex with an antagonist have been experimentally solved and deposited in the Protein Data Bank.

14. **Bolcato, G., Pavan, M., Bassani, D., Sturlese, M. & Moro, S. Ribose and Non-Ribose A2A Adenosine Receptor Agonists: Do They Share the Same Receptor Recognition Mechanism? *Biomedicines 2022, Vol. 10, Page 515* 10, 515 (2022).**

    Adenosine receptors have been a promising class of targets for the development of new therapies for several diseases, including neuro. In recent years, a renewed interest in this field has risen, thanks to the implementation of a novel class of agonists that lack the ribose moiety, once considered essential for the agonistic profile. Recently, an X-ray crystal structure of the A_{2A} adenosine receptor has been solved, providing insights about receptor activation from this novel class of agonists. Starting from this structural information, we have performed supervised molecular dynamics (SuMD) simulations to investigate the binding pathway of a non-nucleoside adenosine receptor agonist as well as one of three classic agonists. Furthermore, we analyzed the possible role of water molecules in receptor activation.

15. **Sighel, D. *et al.* Streptogramin A derivatives as mitochondrial translation inhibitors to suppress glioblastoma stem cell growth. *Eur J Med Chem* 114979 (2022) doi:10.1016/J.EJMECH.2022.114979**

    New therapeutic strategies for glioblastoma treatment, especially tackling the tumor's glioblastoma stem cell (GSC) component, are an urgent medical need. Recently, mitochondrial translation inhibition has been shown to affect GSC growth, clonogenicity, and self-renewal capability, therefore becoming an attractive therapeutic target. The combination of streptogramins B and A antibiotics quinupristin/dalfopristin (Q/D), which inhibits mitochondrial ribosome function, affects GSCs more effectively *in vitro* than the standard of care temozolomide. Here, we performed docking calculations based on the cryo-EM structure of the Q/D-bound mitochondrial ribosome to develop a series of streptogramin A derivatives. A couple of the designed compounds resulted more potent and more able to penetrate cancer cells compared to the parent compounds, thus justifying their election for further evaluation *in vivo* as antineoplastic agents.

Section 3: method development

The third and final section of this work focuses on two methodological papers, in which a new implementation of an in-house developed technique and a brand-new methodology are presented. Both methodological works are an attempt to overcome some of the limitations found in the employment of the classical computational approaches in the development of candidate drugs against COVID-19. Particularly, in the first article Supervised Molecular Dynamics (SuMD) has been applied for the first time to the study of RNA-protein complexes, a necessity that was issued by the rising interest in developing RNA-based therapeutics after the success of RNA vaccines against COVID-19. The development of Thermal Titration Molecular Dynamics, instead, was inspired by the need to have a quick and robust screening tool to distinguish strong and weak binders in the initial stages of a drug discovery campaign, such was the case for the development of candidate inhibitors of SARS-CoV-2 main protease.

16. **Pavan, M., Bassani, D., Sturlese, M. & Moro, S. Investigating RNA–protein recognition mechanisms through supervised molecular dynamics (SuMD) simulations. *NAR Genom Bioinform* 4, (2022).**

    Ribonucleic acid (RNA) plays a key regulatory role within the cell, cooperating with proteins to control the genome expression and several biological processes. Thanks to its structural plasticity, this polymer can mold itself into different three-dimensional structures able to recognize target biomolecules with high affinity and specificity, as demonstrated by aptamers, thereby attracting the interest of drug developers and medicinal chemists. In this scientific work, we present the first application of Supervised Molecular Dynamics (SuMD), an enhanced sampling Molecular Dynamics-based method for the study of receptor-ligand association processes in the nanoseconds timescale, to the study of recognition pathways between RNA aptamers and proteins, elucidating the main advantages and limitations of the technique while discussing its possible role in the rational design of RNA-based therapeutics.

17. **Pavan, M., Menin, S., Bassani, D., Sturlese, M. & Moro, S. Qualitative Estimation of Protein–Ligand Complex Stability through Thermal Titration Molecular Dynamics Simulations. *J Chem Inf Model* (2022) doi:10.1021/ACS.JCIM.2C00995**

    Recently, there has been an increased interest in the prediction and determination of binding kinetic properties, since they better correlate with ligand efficacy compared to thermodynamic properties such as the equilibrium dissociation constant. In the present work, we present Thermal Titration Molecular Dynamics (TTMD), an alternative computational method that combines a series of molecular dynamics simulations performed at progressively increasing temperatures with a scoring function based on protein-ligand interaction fingerprints for the qualitative estimation of protein-ligand binding stability. The protocol has been applied to four different pharmaceutically relevant test cases, on a variety of ligands with different sizes, structures, and experimentally determined affinity values. In all four cases, TTMD was able to distinguish between high-affinity compounds (low nanomolar range) and low-affinity ones (micromolar), proving to be a useful screening tool for the prioritization of compounds in a drug discovery campaign.

Bat coronaviruses related to SARS-CoV-2: what about their 3CL Proteases (M^{pro})?

Matteo Pavan, Davide Bassani, Mattia Sturlese, and Stefano Moro

Pavan, M., Bassani, D., Sturlese, M. & Moro, S. Bat coronaviruses related to SARS-CoV-2: what about their 3CL proteases (MPro)? *https://doi.org/10.1080/14756366.2022.2062336* **37**, 1077–1082 (2022).

Abstract

Despite a huge effort by the scientific community to determine the animal reservoir of SARS-CoV-2, which led to the identification of several SARS-CoV-2-related viruses both in bats and in pangolins, the origin of SARS-CoV-2 is still not clear. Recently, Temmam et al. reported the discovery of bat coronaviruses with a high degree of genome similarity with SARS-CoV-2, especially concerning the RBDs of the S protein, which mediates the capability of such viruses to enter and therefore infect human cells through a hACE2-dependent pathway. These viruses, especially the one named BANAL-236, showed a higher affinity for the hACE2 compared to the original strain of SARS-CoV-2. In the present work, we analyze the similarities and differences between the 3CL protease (main protease, M^{pro}) of these newly reported viruses and SARS-CoV-2, discussing their relevance relative to the efficacy of existing therapeutic approaches against COVID-19, particularly concerning the recently approved orally available Paxlovid, and the development of future ones.

Introduction

Since its outbreak in December 2019, the COVID-19 pandemic has caused to date the death of almost 6 million people all around the world[1,2]. This worldwide-spread disease is caused by a betacoronavirus known as SARS-CoV-2, which infects the respiratory system of the host organism compromising its health status[3]. The symptoms of this illness range from the ones typical of influenza (cough, fever, and headache) to very serious complications such as breathing difficulty, pneumonia, and hypoxia, eventually leading to respiratory failure and death[4]. The high transmissibility of the SARS-CoV-2 virus allowed its fast diffusion all around the world, rapidly attracting the interest of experts in the medical, biological, and pharmaceutical environments, who have extensively worked and

are still putting relevant efforts into the elaboration of proper solutions to fight this pathogen.

The first approach to finding viable therapeutic options was the so-called “drug-repurposing”, i.e. the use of drugs that are already marketed for the treatment of different diseases to cure COVID-19 patients. Concerning this, particular attention was directed towards HIV protease inhibitors such as Kaletra (therapeutic combination of Lopinavir and Ritonavir)[5] and antimalarial drug Plaquenil (commercial name of hydroxychloroquine)[6]. Unfortunately, despite the promising premises (especially from a timescale perspective[7]), this approach was unsuccessful, with investigated drugs showing little to no efficacy in randomized clinical trials[8].

Parallel to the first approach, a considerable amount of labor by both the industry and academia has been spent on developing tools that prevent the detrimental effect of the pathology and has resulted in the approval by the Food and Drug Administration (FDA) of several vaccines[9]. These therapeutic entities can be divided into three different classes[10]: the first one is composed of the inactivated virus vaccines, such as Chinese CoronaVac and the Russian CoviVac, the second family is formed by the ones based on adenovirus vectors, like Vaxzevria, Sputnik V, and the Janssen COVID-19 vaccine, while the third and final family consists of the mRNA-based ones such as the Pfizer-BioNTech “Comirnaty” and the Moderna “Spikevax”.

While vaccines based on inactivated viruses have given poor results, several studies have proven the efficacy of vaccination campaigns with the other two classes of vaccines (especially m-RNA ones) all around the world[11,12]. Despite the success of said vaccines, the SARS-CoV-2 Spike protein is often subjected to immune system-escaping mutations which lead to the development of new viral variants[13], obliging the vaccines to be periodically updated to maintain their efficacy.

The high variability of the Spike protein among different coronavirus strains, which threatens the efficacy of already approved vaccines in the long period, led the scientific community to join forces to identify effective treatments for ongoing infections and to prevent future pandemic waves. Regarding this, a remarkable example is portrayed by the

COVID Moonshot consortium, a collaborative project that involved scientists from all over the world in an attempt to design and develop an orally available drug against COVID-19[14,15]. COVID Moonshot aside, the great amount of knowledge accumulated on the target since the SARS-CoV epidemic in 2002/2003 rapidly resulted in the approval of the first COVID-19 specific treatments.

The first drug to be approved was Remdesivir, a polymerase inhibitor that was initially designed against the Ebola Virus and has then been repositioned against COVID-19. This drug, unfortunately, has an unfavorable pharmacokinetic profile, which limits its administration to the intravenous route in a hospital setting[16,17]. Tocilizumab, an interleukin-6 receptor monoclonal antibody originally developed to cure rheumatoid arthritis, obtained the emergency use authorization (EUA) for the treatment of COVID-19 in the United States in June 2021[18]. The oral RNA-polymerase inhibitor Favipiravir has also been approved for marketing in countries such as Japan, China, India, Saudi Arabia, and the United Arab Emirates, but is still under examination from the FDA[19].

An important milestone has been achieved at the end of 2021 with the FDA approval of the therapy based on the SARS-CoV-2 main protease (M^{pro}) inhibitor Nirmatrelvir (also known as PF-07321332) in combination with Ritonavir, sold under the commercial name "Paxlovid" (which is available also in Europe since the end of January 2022)[20]. Thanks to its efficient, reversibly covalent inhibition of M^{pro}, the Nirmatrelvir-based therapy demonstrated to lower of 89% the risk of severe complications after COVID-19 infection in symptomatic, unvaccinated, non-hospitalized adults[21].

A recent scientific work by Temmam et al. reported the discovery of a high level of sequence similarity between the SARS-CoV-2 genome and that of other coronavirus species infecting cave bats living in North Laos[22], raising serious concerns about the potential threat to public health that these coronaviruses could portray. Despite giving an in-depth analysis on the similarities and differences between the S protein of these newly reported viruses, no consideration is reported in the original work about their main proteases. For this reason, in order to assess the impact that these bat coronaviruses could have on public health, we performed a computational analysis to shed light on similarities and differences between the main protease of SARS-CoV-2 and that of these

newly discovered bat coronaviruses, discussing the role that these alterations could have on the efficacy of existing therapies (Paxlovid, in particular) and the development of future ones.

Materials and Methods

Organism	Isolate	Accession Code	Product	Protein ID	Residues
SARS-CoV-2	"Wuhan-Hu-1"	NC_045512.2	ORF1ab polyprotein	YP_009724389.1	S3264-Q3569
Bat coronavirus	"BANAL-20-52/Laos/2020"	MZ937000.1	ORF1ab polyprotein	UAY13216.1	S3255-Q3560
Bat coronavirus	"BANAL-20-103/Laos/2020"	MZ937001.1	ORF1ab polyprotein	UAY13228.1	S3256-Q3561
Bat coronavirus	"BANAL-20-236/Laos/2020"	MZ937003.2	ORF1ab polyprotein	UAY13252.1	S3256-Q3561
Bat coronavirus RaTG13	"RaTG13"	MN996532.2	ORF1ab polyprotein	QHR63299.2	S3263-Q3568

Table 1. This table reports the protein sequences used in this work and their origin.

The genome sequence for SARS-CoV-2, BANAL-52, BANAL-103, BANAL-236, and RaTG13 was obtained through GenBank. Table 1 reports the accession codes for each of the considered genomes. The protein sequence associated with the 3CL protease (main protease, M^{pro}), was extracted, aligned using the appropriate tool from MOE 2019.01[23], and used for the generation of the correspondent homology model (except for SARS-CoV-2, for which several crystal structures are available in the Protein Data Bank).

The structure of SARS-CoV-2 main protease in its unliganded state was retrieved from the Protein Data Bank (PDB ID: 6Y2E[24]) and prepared using MOE 2019.01. At first, the functional dimer was restored applying the symmetric crystallographic transformation to each asymmetric unit. Secondly, residues with fractional occupancy values were assigned to the most probable state. Then, missing hydrogen atoms were added to the system according to the most probable protonation state at pH 7.4 for each titratable residue

exploiting the “Protonate 3D” tool. Afterward, hydrogen atoms coordinates were energy minimized according to the AMBER10: EHT force field until a gradient of 0.1 Kcal mol^{-1} $Å^{-2}$ was reached. Finally, crystallographic water molecules were removed.

Four different homology models were generated exploiting the “Homology Model” tool, one for each bat coronavirus considered in the present work. The sequences used for the generation of homology models are reported in Table 1, while the structure 6Y2E, prepared as described before, was used as a template for the model generation.

Results

In order to compare similarities and dissimilarities between the SARS-CoV-2 M^{pro} and correspondent proteases in the most closely related bat coronaviruses, four different homology models (one for each different virus considered in this work) were generated, as reported in Materials and Methods. Due to the high degree of sequence identity (99,7% for BANAL-52, BANAL-103, and BANAL-236, 99,4% for RaTG13) between considered bat coronaviruses and SARS-CoV-2 M^{pro}, homology modeling is expected to give a representative result, very closely related to the experimental data.

As illustrated by Figure 1, there are only two differences in the primary sequences of considered viruses. These small changes to the amino acid sequences led to the generation of homology models that are practically superimposable to the reference structure (6Y2E), as is also depicted in Figure 1.

Figure 2, instead, reports a comparison between the four homology models and SARS-CoV-2 M^{pro} from structure 6Y2E, mapping the differences between various proteases onto their three-dimensional structure.

The first difference is related to residue 96, which in the case of SARS-CoV-2 is a proline. This residue is conserved in each BANAL coronavirus reported by Temmam et al. but is not conserved in RaTG13, which was previously considered to be the most closely related bat coronavirus to SARS-CoV-2. Instead of a proline, RaTG13 presents a threonine residue at the 96 position, which is expected to increase both the flexibility and the hydrophilicity of the surroundings.

The second structural alteration is referred to residue 180, which in the case of SARS-CoV-2 is an asparagine. In this case, there is a higher variability between different coronavirus strains, with each BANAL virus presenting a hydrophobic residue (alanine, in the case of BANAL-52 and BANAL-236, isoleucine in the case of BANAL-103), while RaTG13 once again differentiate from both BANAL viruses and SARS-CoV-2 presenting a hydrophilic threonine residue.

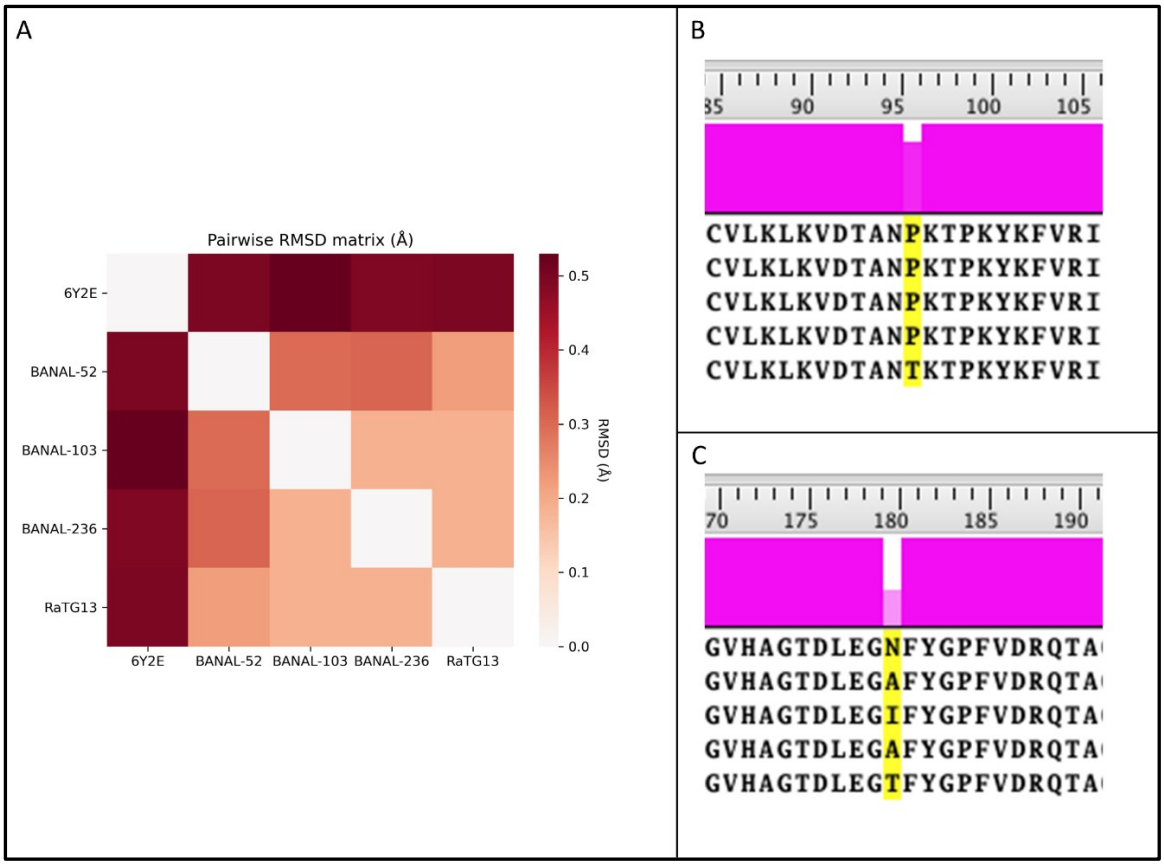


Figure 1. Comparison between SARS-CoV-2 3CL protease (M^{pro}) from crystal structure 6Y2E (blue) and homology models of M^{pro} from four different bat coronaviruses, reported in Table 1. In Panel A, the pairwise RMSD matrix derived from the superposition of each bat coronavirus homology model to the template structure 6Y2E is reported. Panel B and C summarize the differences in the primary sequence between SARS-CoV-2 and bat coronaviruses M^{pro}.

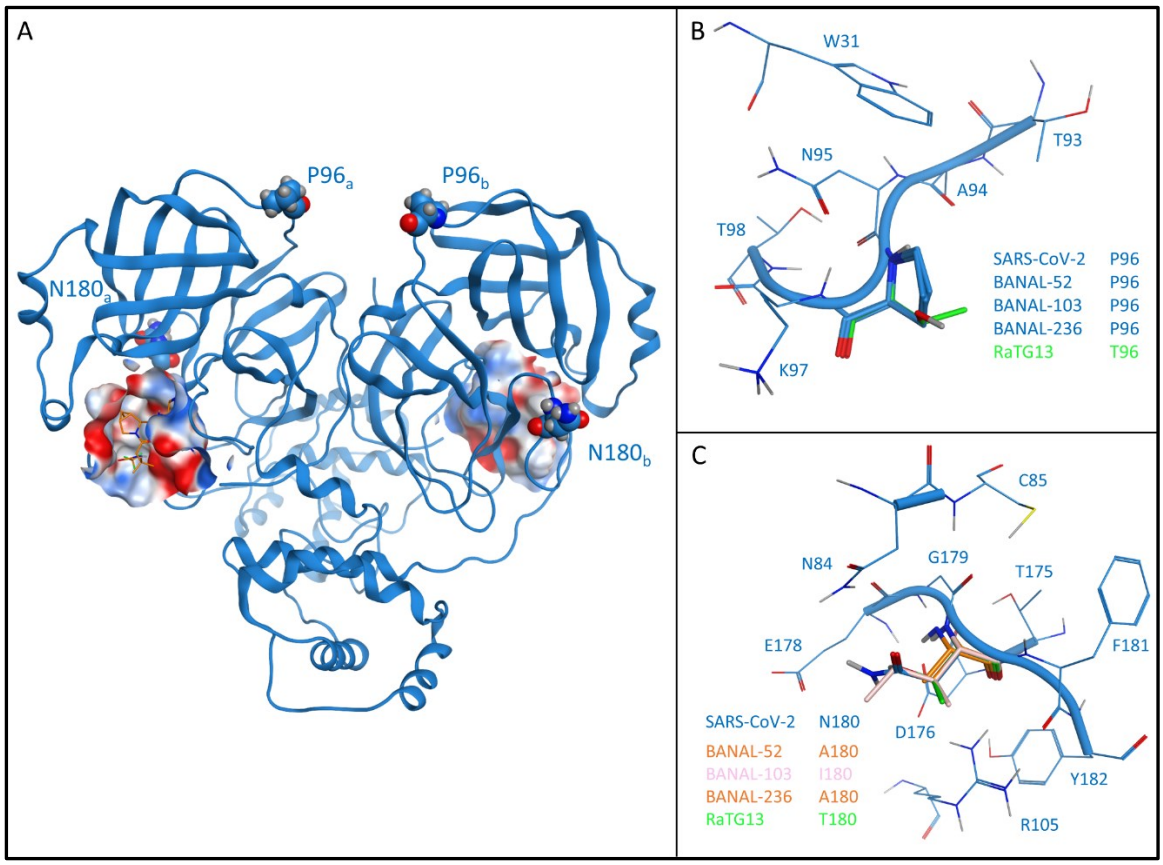


Figure 2. Comparison between SARS-CoV-2 3CL protease (M^{pro}) from crystal structure 6Y2E (blue) and homology models of M^{pro} from four different bat coronaviruses, reported in Table 1. Panel A reports the structure of SARS-CoV-2 M^{pro} (PDB ID: 6Y2E) in its free form. The protein is depicted in blue ribbons, while mutated residues (namely, P96 and N180) in comparison with bat coronaviruses are highlighted and depicted as CPK models. For visual reference, Nirmatrelvir (also known as PF-07321332, commercial name Paxlovid) from structure 7RFS is also shown in the picture, alongside the binding site surface colored according to electrostatic properties. Panel B highlights the comparison between residue 96 of SARS-CoV-2 M^{pro} and homology models of bat coronaviruses M^{pro}. Panel C reports a comparison between residue 180 of SARS-CoV-2 M^{pro} and homology models of bat coronaviruses M^{pro}.

Discussion

The comparison between the crystal structure of SARS-CoV-2 and homology models of bat coronaviruses M^{pro} showed that there are two main structural differences, both of which do not involve the catalytic site.

In the native SARS-CoV-2 structure, Phe96 is involved in a series of hydrophobic contacts with the side chain of Trp31, Thr93, and Lys97 through its pyrrolidine core. In the case of RaTG13, the only bat coronavirus that presents an alteration at this position, the presence of a threonine causes a reduction of possible hydrophobic contacts with the surrounding aminoacids but does not cause the loss of any crucial interaction for structural integrity. Moreover, this residue is located in a solvent-exposed flexible loop region that connects between two beta-sheets, a further indication that this substitution should not compromise the structural integrity of the protease.

Concerning the second structural alteration, in the native SARS-CoV-2 structure Asn180 is involved in a double interaction with the sidechain of two charged residues, namely Asp176 and Arg105. Both of these interactions happen with the backbone of Asn180 and do not involve its sidechain, which is stretched towards the solvent. Intriguingly, in this case, the newly discovered bat coronaviruses all present a hydrophobic residue at position 180: in all these cases, no loss of native interaction happens, coherently with the fact that they do not involve the sidechain of residue 180 and only occur through its backbone. Once again, RaTG13 is the most diverse one, being the only analyzed bat coronavirus that presents a polar aminoacid (a threonine) at this position. As previously mentioned, the sidechain of residue 180 is not involved in any structurally relevant interaction, and therefore the presence of a hydroxyethyl sidechain does not give a particular advantage to this virus strain. Furthermore, as is the case for Pro96, this structural modification is also located in a solvent-exposed, non-structured loop region, indicating that no critical harm to the protease integrity should be provoked by this alteration.

Altogether, our structural analysis reveals that neither of these two structural differences between SARS-CoV-2 and bat coronaviruses M^{pro} should determine any relevant structural alteration of the main protease. Notably, this observation is in agreement with a recent article that characterized the effect of each possible M^{pro} mutation on its functionality: both Pro96 and Asn180 are marked as highly tolerant to mutations[25].

Concerning the implications of these two mutations on the efficacy of M^{pro} inhibitors, several elements point to the conclusion that neither mutation should have a relevant effect. As can be seen from Figure 2, which gives a depiction of the localization of these

two mutations relative to the position of the catalytic site (which is also the binding site of most protease inhibitors, including PF-07321332, the active principle of Paxlovid) shows that both these mutations are not directly linked to the catalytic site, indicating that the binding cleft that harbors PF-07321332 should not be altered. Moreover, as thoroughly assessed in a previous scientific work from our laboratory, neither of these two residues is in any way involved in the recognition process of PF-07321332[26], complementing the structural information provided by crystal structures 7VH8[27], 7RFS, and 7RFW which clearly show how none of these two residues contributes to the interaction with PF-07321332 in the final bound state.

The fact that the SARS-CoV-2 main protease and the one from closely related bat coronaviruses are very similar and practically identical at the catalytic site supports the idea that targeting this protease is still a viable therapeutic option not only for the present but also for the prevention of future pandemic waves. To date, several studies have contributed to thoroughly characterizing the nature of the shallow and solvent-exposed catalytic site of the SARS-CoV-2[28], which has proven to be readily investigable with both time-dependent and time-independent structure based-approaches such as molecular docking[29] and molecular dynamics[30], leading to the development of compounds with affinities in the low nanomolar range[31,32].

All these factors, combined with the fact that striking 3D structure similarities exists also with other coronaviral M^{pro} such as the one from Porcine transmissible Gastroenteritis virus (TGEV)[33], Human coronavirus strain 229E (HCoV)[34], Infectious bronchitis virus (IBV)[24] and MERS-CoV[35], validate the pursue of novel M^{pro} inhibitors that could increase the pool of available treatment for COVID-19 and also for future coronavirus-related diseases, acting as pan-coronaviral drugs.

Conclusions

Recently, a scientific work by Temmam et al. reported the discovery of bat coronaviruses closely related to SARS-CoV-2 that can infect human cells. This scientific work raised the attention of both the scientific community and the general audience to the possible threat to public health that these newly discovered coronaviruses could represent. Despite a

thorough characterization of Spike protein of these bat coronaviruses, no information was given in the original work about their main proteases, which is considered the main target for the development of COVID-19 specific active principles.

In the present scientific work, we performed a computational analysis to characterize structural similarities and differences between the main proteases of SARS-CoV-2 and closely related bat coronaviruses. A comparison between the crystal structure of SARS-CoV-2 M^{pro} and homology models of bat coronavirus M^{pro} shows that two main differences exist, involving the mutation of Pro96 and Asn180. None of these structural alterations are predicted to have an impact on the protease structural integrity, functionality, or affinity for existing inhibitors (especially the recently approved orally available Paxlovid), nor towards the development of novel protease inhibitors. Furthermore, the high degree of structural conservation among main proteases from different coronaviruses suggests that Mpro is not only a valid target for the treatment of COVID-19, but that the knowledge acquired on this target could be useful in the identification and development of pan-coronaviral drugs that can cure different diseases and prevent future pandemic waves.

References

1. Guarner, J. Three Emerging Coronaviruses in Two Decades: The Story of SARS, MERS, and Now COVID-19. *American Journal of Clinical Pathology* vol. 153 420–421 Preprint at https://doi.org/10.1093/ajcp/aqaa029 (2020).

2. COVID Live - Coronavirus Statistics - Worldometer. https://www.worldometers.info/coronavirus/.

3. Zhou, F. *et al.* Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *The Lancet* **395**, 1054–1062 (2020).

4. https://www.cdc.gov/coronavirus/2019-ncov/hcp/clinical-guidance-management-patients.html. Centers for Disease Control and Prevention, 'Interim Clinical Guidance for Management of Patients with Confirmed Coronavirus Disease (COVID-19)'.

5. Bolcato, G., Bissaro, M., Pavan, M., Sturlese, M. & Moro, S. Targeting the coronavirus SARS-CoV-2: computational insights into the mechanism of action of the protease inhibitors lopinavir, ritonavir and nelfinavir. *Sci Rep* **10**, 20927 (2020).

6. Gautret, P. *et al.* Hydroxychloroquine and azithromycin as a treatment of COVID-19: results of an open-label non-randomized clinical trial. *International Journal of Antimicrobial Agents* **56**, (2020).

7. Mani, D., Wadhwani, A. & Krishnamurthy, P. T. Drug Repurposing in Antiviral Research: A Current Scenario. *Journal of Young Pharmacists* **11**, 117–121 (2019).

8. Viveiros Rosa, S. G. & Santos, W. C. Clinical trials on drug repositioning for COVID-19 treatment. *Revista Panamericana de Salud Publica/Pan American Journal of Public Health* **44**, e40 (2020).

9. Fiolet, T., Kherabi, Y., MacDonald, C. J., Ghosn, J. & Peiffer-Smadja, N. Comparing COVID-19 vaccines for their characteristics, efficacy and effectiveness against SARS-CoV-2 and variants of concern: a narrative review. *Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases* **28**, 202–221 (2022).

10. Pollard, A. J. & Bijker, E. M. A guide to vaccinology: from basic principles to new developments. *Nature Reviews Immunology* **21**, 83–100 (2021).

11. Moghadas, S. M. *et al.* The Impact of Vaccination on Coronavirus Disease 2019 (COVID-19) Outbreaks in the United States. *Clinical Infectious Diseases* **73**, 2257–2264 (2021).

12. Rinott, E., Youngster, I. & Lewis, Y. E. Reduction in COVID-19 Patients Requiring Mechanical Ventilation Following Implementation of a National COVID-19 Vaccination Program — Israel, December 2020–February 2021. *MMWR. Morbidity and Mortality Weekly Report* **70**, 326–328 (2021).

13. Harvey, W. T. *et al.* SARS-CoV-2 variants, spike mutations and immune escape. *Nature Reviews Microbiology* **19**, 409–424 (2021).

14. Consortium, T. C. M. *et al.* Open Science Discovery of Oral Non-Covalent SARS-CoV-2 Main Protease Inhibitor Therapeutics. *bioRxiv* 2020.10.29.339317 (2022) doi:10.1101/2020.10.29.339317.

15. Morris, A. *et al.* Discovery of SARS-CoV-2 main protease inhibitors using a synthesis-directed de novo design model. *Chemical communications (Cambridge, England)* **57**, 5909–5912 (2021).

16. Kokic, G. *et al.* Mechanism of SARS-CoV-2 polymerase stalling by remdesivir. *Nature Communications* **12**, 1–7 (2021).

17. Beigel, J. H. *et al.* Remdesivir for the Treatment of Covid-19 — Final Report. *New England Journal of Medicine* **383**, 1813–1826 (2020).

18. Xu, X. *et al.* Effective treatment of severe COVID-19 patients with tocilizumab. *Proceedings of the National Academy of Sciences* **117**, 10970–10975 (2020).

19. Manabe, T., Kambayashi, D., Akatsu, H. & Kudo, K. Favipiravir for the treatment of patients with COVID-19: a systematic review and meta-analysis. *BMC Infectious Diseases* **21**, 1–13 (2021).

20. Owen, D. R. *et al.* An oral SARS-CoV-2 M pro inhibitor clinical candidate for the treatment of COVID-19. *Science (1979)* **374**, 1586–1593 (2021).

21. Hammond, J. *et al.* Oral Nirmatrelvir for High-Risk, Nonhospitalized Adults with Covid-19. *New England Journal of Medicine* (2022) doi:10.1056/NEJMoa2118542.

22. Temmam, S. *et al.* Bat coronaviruses related to SARS-CoV-2 and infectious for human cells. *Nature* (2022) doi:10.1038/s41586-022-04532-4.

23. Molecular Operating Environment (MOE), 2019.01; Chemical Computing Group ULC, 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, 2021. https://www.chemcomp.com/Research-Citing_MOE.htm.

24. Xue, X. *et al.* Production of authentic SARS-CoV M(pro) with enhanced activity: application as a novel tag-cleavage endopeptidase for protein overproduction. *Journal of molecular biology* **366**, 965–975 (2007).

25. Flynn, J. M. *et al.* Comprehensive fitness landscape of SARS-CoV-2 M^{pro} reveals insights into viral resistance mechanisms. *bioRxiv* 2022.01.26.477860 (2022) doi:10.1101/2022.01.26.477860.

26. Pavan, M., Bolcato, G., Bassani, D., Sturlese, M. & Moro, S. Supervised Molecular Dynamics (SuMD) Insights into the mechanism of action of SARS-CoV-2 main protease inhibitor PF-07321332. *J Enzyme Inhib Med Chem* **36**, 1646–1650 (2021).

27. Zhao, Y. *et al.* Crystal structure of SARS-CoV-2 main protease in complex with protease inhibitor PF-07321332. *Protein & Cell* (2021) doi:10.1007/s13238-021-00883-2.

28. Fornasier, E. *et al.* A new inactive conformation of SARS-CoV-2 main protease. *Acta Crystallographica Section D Structural Biology* **78**, (2022).

29. Bassani, D., Pavan, M., Bolcato, G., Sturlese, M. & Moro, S. Re-Exploring the Ability of Common Docking Programs to Correctly Reproduce the Binding Modes of Non-Covalent Inhibitors of SARS-CoV-2 Protease Mpro. *Pharmaceuticals* **15**, 180 (2022).

30. Bissaro, M. *et al.* Inspecting the mechanism of fragment hit binding on SARS-CoV-2 Mpro by using supervised molecular dynamics (SuMD) simulations. *ChemMedChem* (2021) doi:10.1002/cmdc.202100156.

31. Zhang, C. H. *et al.* Potent Noncovalent Inhibitors of the Main Protease of SARS-CoV-2 from Molecular Sculpting of the Drug Perampanel Guided by Free Energy Perturbation Calculations. *ACS Central Science* **7**, 467–475 (2021).

32. Luttens, A. *et al.* Ultralarge Virtual Screening Identifies SARS-CoV-2 Main Protease Inhibitors with Broad-Spectrum Activity against Coronaviruses. *J Am Chem Soc* **144**, 2905–2920 (2022).

33. Anand, K. *et al.* Structure of coronavirus main proteinase reveals combination of a chymotrypsin fold with an extra alpha-helical domain. *The EMBO journal* **21**, 3213–3224 (2002).

34. Anand, K., Ziebuhr, J., Wadhwani, P., Mesters, J. R. & Hilgenfeld, R. Coronavirus main proteinase (3CLpro) structure: basis for design of anti-SARS drugs. *Science (New York, N.Y.)* **300**, 1763–1767 (2003).

35. Ho, B. L. *et al.* Critical Assessment of the Important Residues Involved in the Dimerization and Catalysis of MERS Coronavirus Main Protease. *PloS one* **10**, (2015).

From the Wuhan-Hu-1 strain to the XD and XE variants: is targeting the SARS-CoV-2 Spike protein still a pharmaceutically relevant option against COVID-19?

Matteo Pavan, Davide Bassani, Mattia Sturlese, and Stefano Moro

Pavan, M., Bassani, D., Sturlese, M. & Moro, S. From the Wuhan-Hu-1 strain to the XD and XE variants: is targeting the SARS-CoV-2 spike protein still a pharmaceutically relevant option against COVID-19? *J Enzyme Inhib Med Chem* **37**, 1704–1714 (2022).

Abstract

Since the outbreak of the COVID-19 pandemic in December 2019, the SARS-CoV-2 genome has undergone several mutations, resulting in the rise of a multitude of viral variants, some of which have shown an increased pathogenic potential. The emergence of such variants has resulted in multiple pandemic waves, contributing to sustaining to date the number of infections, hospitalizations, and deaths despite the swift development of vaccines, since most of these mutations are concentrated on the Spike protein, a viral surface glycoprotein that is the main target for most vaccines. A milestone in the fight against the COVID-19 pandemic has been represented by the development of Paxlovid, the first orally available drug against COVID-19, which acts instead on a different viral target, i.e., the Main Protease (M^{pro}). In this article, we analyze the structural features of both the Spike protein and the Main Protease of the recently reported SARS-CoV-2 variant XE, as well as the closely related XD and XF ones, discussing their impact on the efficacy of existing treatments against COVID-19 and on the development of future ones.

Introduction

More than two years have now passed since the beginning of the COVID-19 pandemic, back in December 2019[1,2]. Caused by a betacoronavirus known as SARS-CoV-2 and characterized by para-influenzal symptoms such as fever, cough, and dyspnea, this worldwide-spread disease has resulted in the death of more than 6 million people around the world, becoming one of the deadliest illnesses in human history[3,4].

The SARS-CoV-2 virus was first identified in the Chinese city of Wuhan, where the pandemic was firstly spotted[5]. The genomic sequence of this virus (named Wuhan-Hu-1 from now on in the article) is 80% identical to the one of the SARS-CoV virus[6,7], which was responsible for the Severe Acute Respiratory Syndrome (SARS) that stroke the South East of Asia in 2002/2003, causing the death of about 800 patients over 9000 cases (10 % death rate)[8,9]. The exact origin of the SARS-CoV-2 virus is still to this date unknown, however several pieces of evidence point out bat coronaviruses as closely related ancestors and to the pangolin as the intermediate host before the human spillover[10–13].

Soon after the original virus started spreading all over the world, several viral variants began to emerge[14,15], especially in the poorest countries where public health measures such as social distancing and wearing surgical masks in public places were difficult to implement[16–18]. Most of the genome mutations that characterized these variants were concentrated in the S gene[19], which encodes for the Spike protein, a surface glycoprotein that mediates the virus entry within the human cell through an interaction with the human ACE2 receptor[20]. Some of these mutations gathered the attention of the scientific community due to the selective advantage that they provided to the correspondent viral variants, regarding both the virus' ability to infect human cells and to escape the immune system response[21], gaining for these reasons the status of "variant of concern" (VOC).

The first SARS-CoV-2 variant to be labeled as VOC was the so-called Alpha variant (B.1.1.7). First identified in November 2020 in the Kent region of the United Kingdom and for this reason also known as the "English variant", B.1.1.7 was estimated to be 29% more transmissible than the original virus[22,23]. Despite being more transmissible than other circulating viral strains[24,25], and despite showing the first signs of reduced protection provided by vaccines, monoclonal antibodies, and convalescent sera[26–28], the indication from clinical studies showed that the vaccine coverage (especially in those who had already completed the vaccination cycle) was still able to contain the impact of this variant on the sanitary system[29]. Soon after the identification of the Alpha variant, a second VOC arose: the Delta variant (B.1.617.2), also known as the "Indian variant" due to being first detected in India in late 2020, rapidly overcame the Alpha variant becoming the dominant strain in the world, thanks to being 97% more transmissible than the original Wuhan

virus[23]. The replacement of the less threatening Alpha variant with the Delta marked a significant change of pace in the pandemic trend, signing an increased burden for the health system[30] and posing for the first time a serious threat to the protection provided by vaccines, convalescent sera, and monoclonal antibodies[31–33] due to its increased ability to evade the immune system response[34]. From November 2021 onwards, the Delta variant has been flanked by another VOC firstly identified in South Africa and defined as the Omicron variant (B.1.1.529)[35]. The rise of the Omicron variant, fuelled by a contemporary increase in transmissibility[36] and immune evasion[37], resulted in an unprecedented diffusion of the SARS-CoV-2 virus all over the world, being able to overcome even the protection provided by the full primary vaccination cycle and by most neutralizing antibodies used in therapy[38–40], thus inducing the introduction of a "booster dose" to bring the protective effect of vaccines back to adequate levels[41,42].

In the face of this increasingly troublesome variant landscape, characterized by a progressive reduction of the efficacy of existing therapeutic options against COVID-19, a light at the end of the tunnel is possibly represented by the development and release on the market of Paxlovid. This therapeutic combination between the active principle Nirmatrelvir (also known as PF-07321332) and the pharmacokinetic enhancer Ritonavir, represents the first orally available drug specifically designed against SARS-CoV-2 virus[43]. Instead of targeting the Spike protein, this peptidomimetic entity is designed to inhibit the SARS-CoV-2 Main Protease (M^{pro}) by covalently binding to Cysteine 145, one of the two components of the protease's catalytic diad[44]. Clinical studies showed a remarkable therapeutic efficacy of this novel treatment, with Paxlovid being able to lower by 89% the risk of severe complications associated with COVID-19 infection in symptomatic, non-vaccinated, and non-hospitalized adult patients[45].

Recently, three novel recombinant SARS-CoV-2 variants were identified in the United Kingdom: Xd, Xe, and Xf[46]. These variants are derived from the combination of the genomes of other major circulating variants, namely Delta, Omicron, and Omicron 2[46,47]. Among these three novel viral strains, particular worry is related to the XE variant, which derives from the recombination between two VOCs, Omicron and Omicron 2, and is supposed to be 13-20 % more transmissible than the Omicron 2 variant[46].

The rise of novel SARS-CoV-2 variants derived from the recombination of other threatening and heavily diffused ones poses a serious challenge in the fight against the COVID-19 pandemic, since it could contribute to render existing therapeutic options inefficient or practically useless. To evaluate the impact that these recently reported recombinant variants could have on the efficacy of existing vaccines and treatments (Paxlovid, in particular), we performed a computational analysis to shed light on the key structural features that characterize both the Spike glycoprotein and the Main Protease of these novel viral strains. Moreover, we analyzed the structural evolution of these two viral proteins throughout the pandemic, discussing the impact that mutations found on these strains had and will have on the efficacy of existing therapeutic options against COVID-19 and the development of future ones.

Materials and Methods

The genome sequence for the SARS-CoV-2 virus and its variants, namely Delta, Omicron, XD, XE and XF, was obtained through GenBank[48]. Accession codes for each of the considered genomes are reported in Table 1. In the case of newly discovered variants XD, XE, XF, the sequence was chosen according to the one reported by the Nextclade project[49].

Organism	**Isolate**	**Accession Code**
SARS-CoV-2	"Wuhan-Hu-1"	NC_045512.2
SARS-CoV-2 "Delta"	"SARS-CoV-2/human/JPN/SARS-CoV-2"	OK091006.1
SARS-CoV-2 "Omicron"	"SARS-CoV-2/human/NLD/EMC-Omicron-1/2021"	OM287553.1
SARS-CoV-2 "XD"	"SARS-CoV-2/human/FRA/IHUCOVID-64762/2022"	OM990851.1
SARS-CoV-2 "XE"	/	OW018845.1
SARS-CoV-2 "XF"	/	OV940149.1

Table 1. The genome sequences used in this work and their origin.

All the basic molecular modeling operations have been executed with the Molecular Operating Environment (MOE) suite (version 2019.01)[50].

For what concerns the Spike protein, the approach chosen depended on the variant considered. For the wild-type (WT) Spike, the three-dimensional structure was retrieved from the Protein Data Bank (PDB code: 6ZDH[51], method: cryo-EM, resolution: 3.70 Å), as well as for the Delta (PDB code: 7W9E[52], method: cryo-EM, resolution: 3.10 Å), and the Omicron (PDB code: 7WPD[53], method: cryo-EM, 3.18 Å) variants. The cited structures were all subjected to the same preparation procedure for molecular modeling.

After being downloaded, the *Structure Preparation* tool implemented in MOE was applied in order to rebuild the missing loops in the structures, the proper protonation state was assigned to each amino acid with the MOE "*Protonate 3D"* application, and finally, the added hydrogens were minimized under the AMBER10:EHT[54] force field implemented in MOE. Since experimentally resolved structures for the XD, XE, and XF variants are not available in public databases, the models considered for our study were created starting from the WT SARS-CoV-2 Spike coming from the PDB code 6ZDH by manually mutating the residues, exploiting the MOE *"Protein builder"* tool, and subjecting each protein to the preparation procedure reported above. For the realization of the video reported in the Supplementary Materials, the program VMD 1.9.2[55] (Visual Molecular Dynamics) was used.

For what concerns M^{pro}, the protein sequences corresponding to the main protease were extracted from the whole genome sequence and aligned to the reference sequence (Wuhan-Hu-1) making use of the appropriate tool of MOE 2019.01. Subsequently, homology models for each variant were created making use of the "*Homology Model*" tool of MOE 2019.01, using the structure deposited in the Protein Data Bank with accession code 6Y2E (Crystal structure of the free enzyme of the SARS-CoV-2 main protease) as a template for model generation.

Results and Discussion

Structural analysis of Spike glycoprotein mutations found in SARS-CoV-2 XD, XE, and XF variants and their impact on hACE2 binding

The SARS-CoV-2 Spike protein (S) consists of a large biological entity formed by 1273 amino acids organized in different functional domains. The main role of the Spike protein is to mediate the virus entry into the host cell, with the principal and better-characterized mechanism being the pathway involving the binding to the human ACE2 receptor (hACE2)[56], a membrane-bound enzyme that is widely expressed in various districts of the human body (from the endothelial cells of the blood vessels to kidneys, liver, intestine, lungs[57], and cells of bronchial and nasal epithelia[58]).

The S protein, which works in a trimeric organization, is divided into two main subunits, S1 and S2. The second of these has very important roles in spike protein trimerization and in mediating the virion entry into the host cell once the molecular contacts have been established. It is formed by relevant subdomains such as the fusion peptide (FP, residues 943-982, crucial for viral fusion to the host cell membrane), the transmembrane domain (TM, composed of 24 aminoacids and deputed both to the anchoring of S protein to the viral membrane and the maintenance of the Spike trimeric organization), and the cytoplasmatic fusion domain (CT, mediating virus-cell fusion).

The S1 subunit, instead, contains both the N-terminal and the C-terminal domains (NTD and CTD, respectively), which are involved in the binding to host cell receptors. Specifically, the CTD contains the receptor-binding domain (RBD, aminoacids 319-541), the region deputed to the binding with hACE2. This function is more precisely carried out by a particular RDB subdomain, called receptor-binding motif (RBM), which is formed by two beta-sheets (b5 and b6) composed of those residues which are in close contact with hACE2 (from 438 to 506[59])[60].

Looking at all the S proteins of the different SARS-CoV-2 relevant variants, the RBD contains the highest “single-point mutations/sequence length” ratio in all cases. Examining the different SARS-CoV-2 variants discovered up to date, the Spike protein is surely the viral entity that has mutated the most in the evolutionary process of the virus[61].

Its exposition on the viral surface and its crucial function in viral cell entry make this protein the eligible target for the host immune system[62].

The SARS-CoV-2 S protein has experienced several mutations in the past two years[63], as reported in Tables 1 and 2 for the variants considered in our study. As can be noticed, variants such as Delta (but also Alpha and Beta, not specifically treated in this article) showed few mutations in the overall viral genome, and Spike protein displayed never more than a tenth of single-point changes. The game-changing event was the advent of the Omicron variant, much different from its previous analogs, with 30-single nucleotide mutations involving the S protein only. Many of these, such as K417N, T478K, and P614G were inherited from the previous lineages (mainly Beta and Delta), but other mutations were completely new, such as G339D, G446S, or E484A.

Delta variant	**Omicron variant**	**XD variant**	**XE variant**	**XF variant**
T19R		T19R	T19R	
		A27S	A27S	
	A67V			A67V
	T95I	T95I		T95I
	G142D	G142D	G142D	
				Y145D
R158G		R158G		
	L212I	L212I		L212I
			V213G	
	G339D	G339D	G339D	G339D
	S371L	S371L	S371L	S371L
	S373P	S373P	S373P	S373P
	S375F	S375F	S375F	S375F
			T376A	
			D405N	
			R408S	
	K417N	K417N	K417N	K417N
	N440K	N440K	N440K	N440K

	G446S	G446S		G446S
L452R				
	S477N	S477N	S477N	S477N
T478K	T478K	T478K	T478K	T478K
	E484A	E484A	E484A	E484A
	Q493R	Q493R	Q493R	Q493R
	G496S	G496S		G496S
	Q498R	Q498R	Q498R	Q498R
	N501Y	N501Y	N501Y	N501Y
	Y505H	Y505H	Y505H	Y505H
	T547K	T547K		T547K
D614G	D614G	D614G	D614G	D614G
	H655Y	H655Y	H655Y	H655Y
	N679K	N679K	N679K	N679K
P681R	P681H	P681H	P681H	P681H
	N764K	N764K	N764K	N764K
	D796Y	D796Y	D796Y	D796Y
	N856K	N856K		N856K
	Q954H	Q954H	Q954H	Q954H
	N969K	N969K	N969K	N969K
	L981F	L981F		L981F

Table 2. List of all the single-point mutations affecting the SARS-CoV-2 Spike protein for all the variants considered in our study (Delta, Omicron, XE, XD, and XF). The mutations have been aligned to give a better perspective of the ones which have been conserved through the evolutionary process. The mutations involving the RBD have been highlighted in green, while the ones involving the RBM are colored in cyan.

Most of the mutations listed in Table 2 have been related to higher infectivity, mainly due to a consequent gain of affinity for hACE2 or improved shielding from the immune cells. As evidence of this, the most successful vaccination campaigns for COVID-19 always involved the forced recognition of Spike protein from the patients' immune cells[64].

Specifically, mutations such as S371L, K417N, and Q493R were related to a diminished binding to the anti-coronavirus monoclonal antibody Casivirimab, while mutations like

N440K and G446S confer resistance towards the antibody Imdevimab[65]. The combination of Casivirimab and Imdevimab has been used to treat COVID-19 patients but has demonstrated to be ineffective against the Omicron variant[66]. Other mutations external to the RBD have been linked to different outcomes, such as increased viral replication (Δ69-70[67] and D614G[68]) and higher viral resistance (G339D and N440K[69]). In other scenarios, mutations have been reported to influence tropism of the S1/S2 cleavage, as in the cases of N679K and P681H[70].

The majority of the mutations highlighted up to date on the SARS-CoV-2 S protein impact the binding with hACE2, as in the cases of S477N, Q498R, and N501Y[71]. These last mutations, as can be seen from Table 2, have been conserved from all the variants following Omicron, assessing their importance for the viral evolutionary process.

As can be seen in Figures 1 to 5 and video.mp4 (Supplementary Materials), the highest "number of mutations/sequence length" ratio is owned by the RDB, as previously mentioned. Indeed, taking Omicron as an example, among the 30 mutations in the overall 1273-residues structure, 15 are located just in the 222 residues-sequence forming the RBD. The insertions and the deletions (summarized in Table 3), on the other hand, are located far outside the hACE2-binding domain in all the variants examined, allowing us to assert that these mutations should not impact at all the host cell recognition process.

Delta variant	Omicron variant	XD variant	XE variant	XF variant
			Δ24-26	
	Δ69-70			Δ69-70
				Δ142-144
	Δ143-145			
Δ156-157		Δ156-157		
	Δ211	Δ211		Δ211
	ins214EPE	ins214EPE		ins214EPE

Table 3. List of all the insertions and deletions operated by SARS-CoV-2 Spike protein for all the variants considered in our study (Delta, Omicron, XD, XE, and XF). The mutations have been aligned to give a better perspective of the ones which have been conserved through the evolutionary process.

Interestingly, as depicted in the aforementioned figures, single-point mutations that are present in the RBD for all considered variants tend to progressively increase the positively-

charged character of this protein region. Moreover, of all the changes operated by the evolutionary process of Spike protein, the very few mutations which transform a residue into a negatively-charged one (Asp or Glu) are always located away from the RBD (except for G339D, which is located in the posterior part of the RBD, away from the RBM that contacts hACE2). Indeed, in this region, the changes from polar aminoacids to positively charged ones are abundant (N440K, T478K, Q498R, Y505H), and there are also cases in which non-polar residues transform into polar ones (e.g., G446S and G496S, which are conserved in all examined post-Omicron variants except for the XE one). Another conserved structural feature across all post-Omicron variants is also the fact that E484, located in the RBM, mutates into an alanine, while a peculiar mutation exclusive to the XE variant is represented by the transformation of D405 into an asparagine. Taken together, all these pieces of evidence converge in assessing that an increase in the polar characteristics of the RBD (more specifically, the RBM), with particular relevance to an increase in the number of positively charged aminoacids, could be the mechanism adopted by SARS-CoV-2 to continuously increase its infectivity through an increase in the interaction with hACE2.

To further support this evidence, in Figure 6 we reported the electrostatic surface of hACE2 complexed with WT-Spike RBD highlighting the prevalence of negative charge on the surface facing Spike RBM (colored in green in the image). Coherently, the only mutation present in the RBD in which a positively-charged residue shifts into a polar one (K417N) has been reported to reduce the affinity with hACE2[75]. It is worth noting that other hACE2-independent entry routes for SARS-CoV-2 have been described in literature[76,77], but the lack of reliable structural information about the interaction with the target at the present date hampers and limits the possibility to analyze and discuss the impact that these mutations could have on them in a meaningful way. However, it cannot be excluded that this mutation pattern and other future Spike mutations could also impact these other entry pathways, contributing to making them more relevant for the ability of SARS-CoV-2 to infect human cells and increasing its overall infectivity.

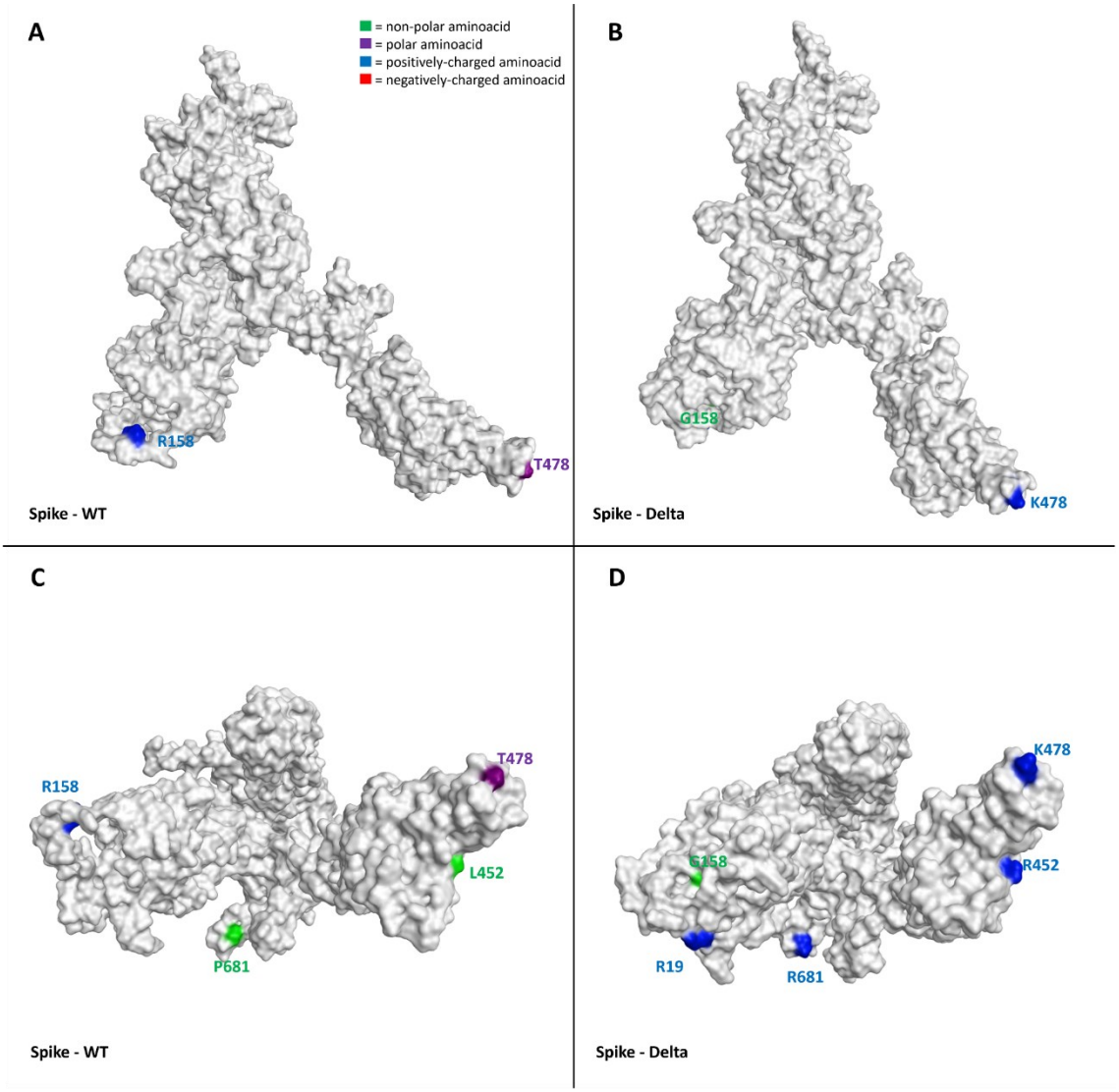


Figure 1. Representation of structural differences between the WT (taken from the Protein Data Bank[72], PDB code: 6ZDH) and the Delta variant (retrieved from the PDB, code 7W9E[73]) of SARS-CoV-2 Spike protein. Panels A and B offer a front view of the comparison between the structures, while panels C and D shift the point of view to the bottom of the proteins. To give a clearer view of the mutations, only one monomer was considered to create the image. The aminoacids involved in mutations are labeled in the figure and are colored based on their kind, following the legend reported in the panel A. Specifically, Gly, Ala, Val, Leu, Ile, Pro, Cys, Met, Phe and Trp are considered non-polar aminoacids (green), Asp and Glu represent the negatively-charged aminoacids (red), and Lys, Arg, and His form the positively-charged aminoacid group (blue). Finally, Ser, Thr, Asn, Gln, and Tyr are all considered polar aminoacids (purple). All images were created and rendered using the Molecular Operating Environment (MOE) suite.

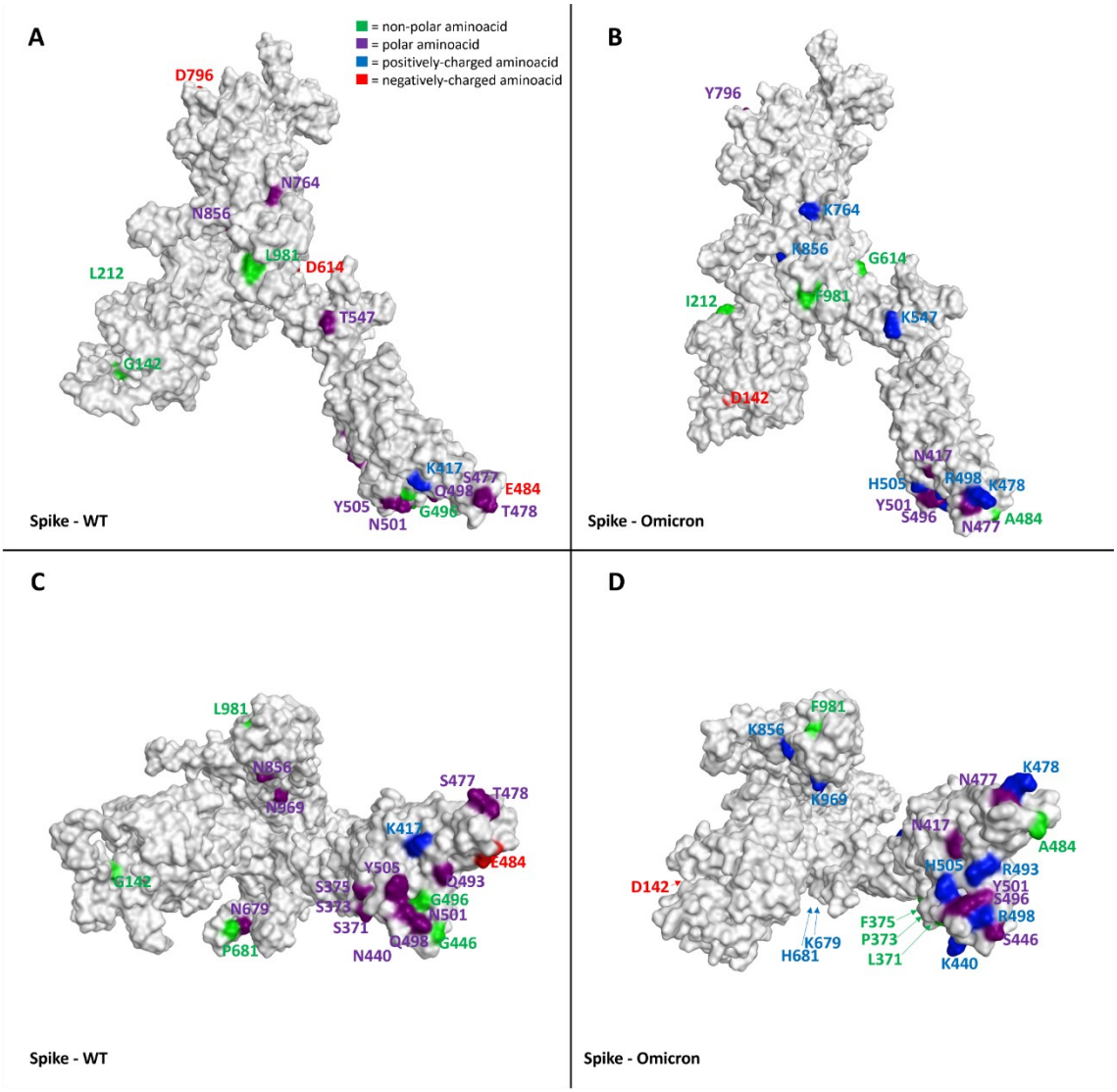


Figure 2. Representation of the differences between the WT (taken from PDB code: 6ZDH) and the Omicron variant (retrieved from the PDB, code 7WPD) of SARS-CoV-2 Spike protein. Panels A and B offer a front view of the comparison between the structures, while panels C and D shift the point of view to the bottom of the proteins. To give a clearer view of the mutations, only one monomer was considered to create the image. The aminoacids involved in mutations are labeled in the figure and are colored based on their kind, following the legend reported in the panel A. Specifically, Gly, Ala, Val, Leu, Ile, Pro, Cys, Met, Phe and Trp are considered non-polar aminoacids (green), Asp and Glu represent the negatively-charged aminoacids (red), and Lys, Arg, and His form the positively-charged aminoacid group (blue). Finally, Ser, Thr, Asn, Gln, and Tyr are all considered polar aminoacids (purple). All images were created and rendered using the Molecular Operating Environment (MOE) suite.

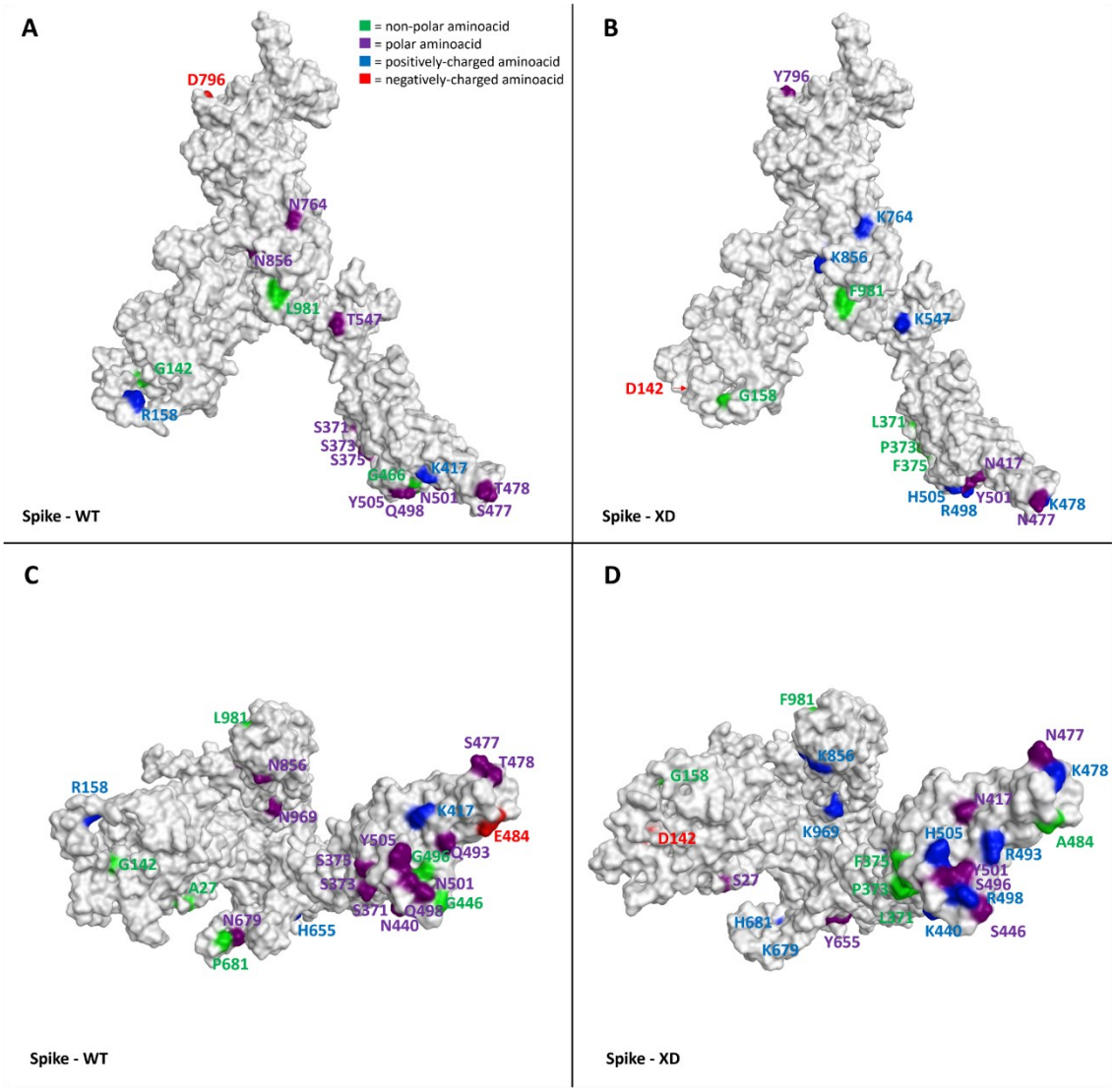


Figure 3. Representation of the differences between the WT (taken from PDB code: 6ZDH) and the XD variant of SARS-CoV-2 Spike protein. Due to the lack of experimentally resolved structure of SARS-CoV-2 XD variant, the three-dimensional structure represented was obtained from the wild-type S protein coming from PDB code 6DZH, and then manually mutating the residues involved in the mutations (exploiting the MOE *"Protein builder"* tool). Panels A and B offer a front view of the comparison between the structures, while panels C and D shift the point of view to the bottom of the proteins. To give a clearer view of the mutations, only one monomer was considered to create the image. The aminoacids involved in mutations are labeled in the figure and are colored based on their kind, following the legend reported in the panel A. Specifically, Gly, Ala, Val, Leu, Ile, Pro, Cys, Met, Phe and Trp are considered non-polar aminoacids (green), Asp and Glu represent the negatively-charged aminoacids (red), and Lys, Arg, and His form the positively-charged aminoacid group (blue). Finally, Ser, Thr, Asn, Gln, and Tyr are all considered polar aminoacids (purple). All images were created and rendered using the Molecular Operating Environment (MOE) suite.

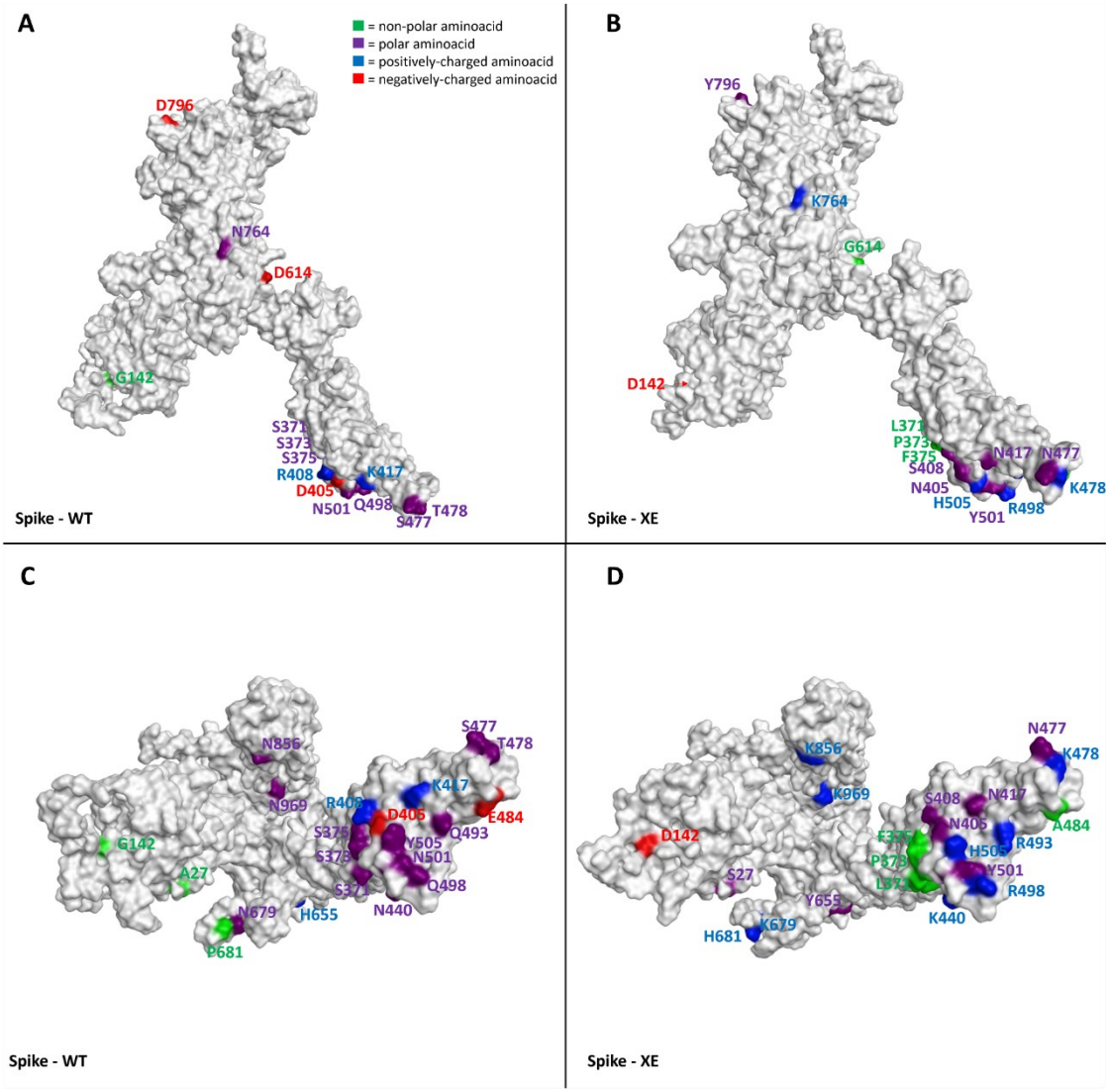


Figure 4. Representation of the differences between the WT (taken from PDB code: 6ZDH) and the XE variant of SARS-CoV-2 Spike protein. Due to the lack of experimentally resolved structure of SARS-CoV-2 XE variant, the three-dimensional structure represented was obtained from the wild-type S protein coming from PDB code 6DZH, and then manually mutating the residues involved in the mutations (exploiting the MOE *"Protein builder"* tool). Panels A and B offer a front view of the comparison between the structures, while panels C and D shift the point of view to the bottom of the proteins. To give a clearer view of the mutations, only one monomer was considered to create the image. The aminoacids involved in mutations are labeled in the figure and are colored based on their kind, following the legend reported in the panel A. Specifically, Gly, Ala, Val, Leu, Ile, Pro, Cys, Met, Phe and Trp are considered non-polar aminoacids (green), Asp and Glu represent the negatively-charged aminoacids (red), and Lys, Arg, and His form the positively-charged aminoacid group (blue). Finally, Ser, Thr, Asn, Gln, and Tyr are all considered polar aminoacids (purple). All images were created and rendered using the Molecular Operating Environment (MOE) suite.

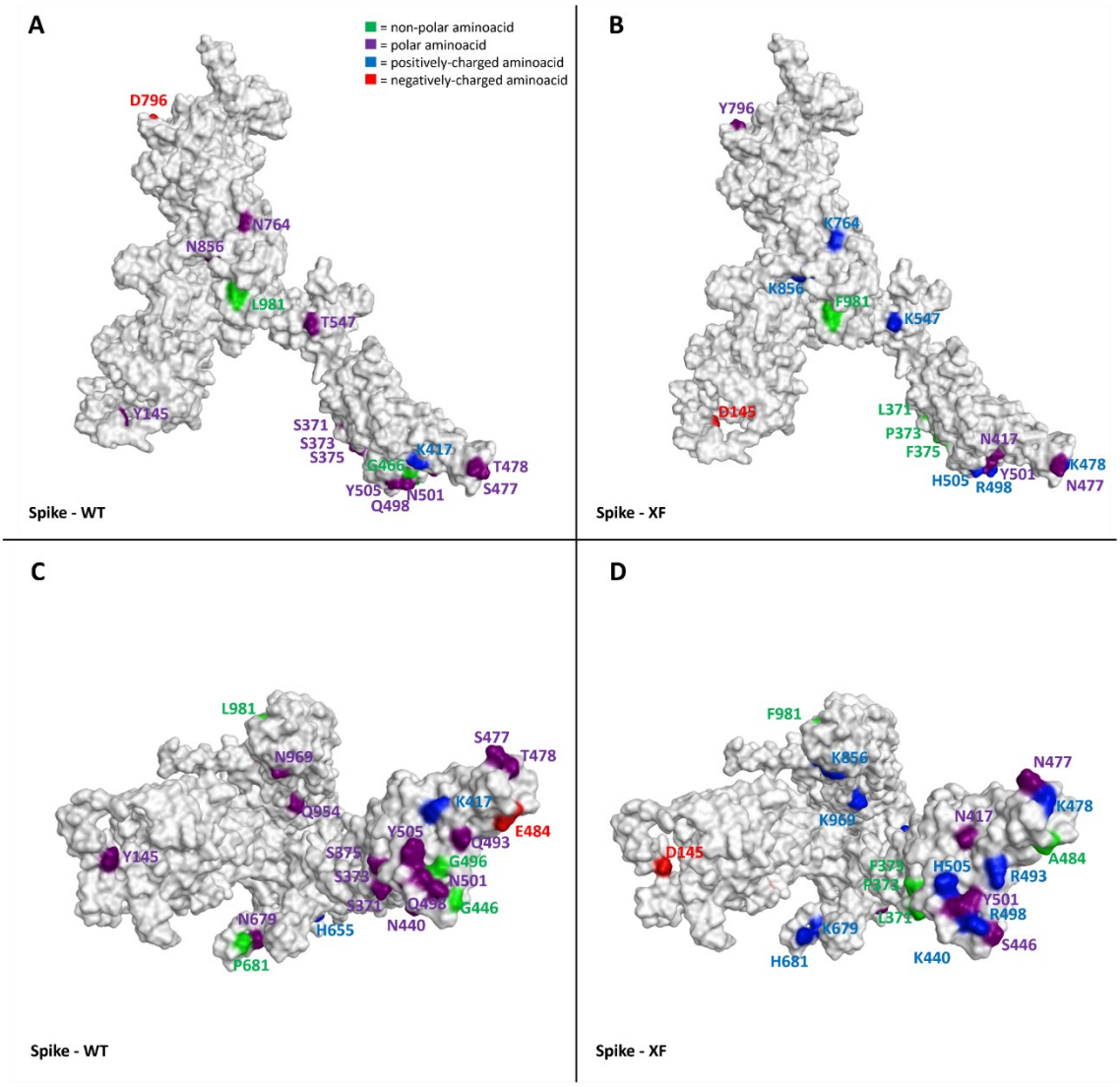


Figure 5. Representation of the differences between the WT (taken from PDB code: 6ZDH) and the XF variant of SARS-CoV-2 Spike protein. Due to the lack of experimentally resolved structure of SARS-CoV-2 XF variant, the three-dimensional structure represented was obtained from the wild-type S protein coming from PDB code 6DZH, and then manually mutating the residues involved in the mutations (exploiting the MOE *"Protein builder"* tool). Panels A and B offer a front view of the comparison between the structures, while panels C and D shift the point of view to the bottom of the proteins. To give a clearer view of the mutations, only one monomer was considered to create the image. The aminoacids involved in mutations are labeled in the figure and are colored based on their kind, following the legend reported in the panel A. Specifically, Gly, Ala, Val, Leu, Ile, Pro, Cys, Met, Phe and Trp are considered non-polar aminoacids (green), Asp and Glu represent the negatively-charged aminoacids (red), and Lys, Arg, and His form the positively-charged aminoacid group (blue). Finally, Ser, Thr, Asn, Gln, and Tyr are all considered polar aminoacids (purple). All images were created and rendered using the Molecular Operating Environment (MOE) suite.

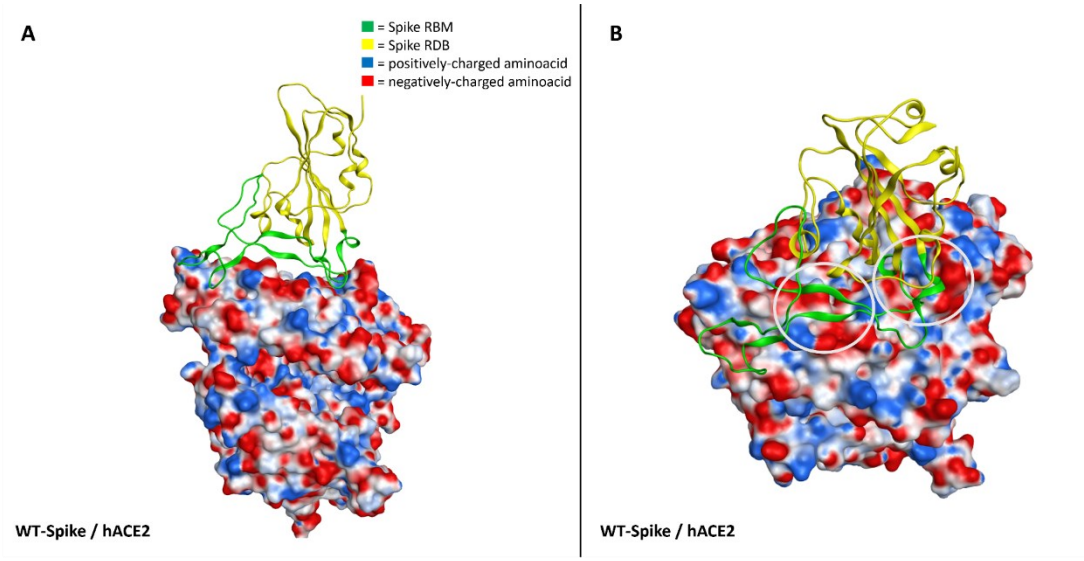


Figure 6. Representation of the interaction between WT-Spike receptor-binding domain (RBD) and hACE2 (coming from PDB code: 6M0J[74]). The Spike RBD is colored in yellow, while the receptor-binding motif (RBM) is colored in green. The hACE2 surface is colored according to the electrostatic properties of underlying residues (blue, positively-charged regions, red, negatively-charged regions, white, neutral regions). Panel A offers a lateral view of the complex, while panel B focuses the attention on a top-lateral perspective. As can be seen from panel B, the hACE2 regions in contact with Spike-RBM are prevalently negatively-charged (red color): concerning this, for visualization purpose, the most extended negative regions at the Spike-hACE2 interface have also been highlighted with grey circles.

Structural analysis of Main Protease mutations found in SARS-CoV-2 XD, XE and XF variants and their impact on the recognition of known inhibitors

The main protease M^{pro}, also known as 3C-like protease or $3CL^{pro}$, is a cysteine peptidase that is essential for the replication cycle of SARS-CoV-2[78,79]. Its catalytic activity revolves around the processing of two overlapping polyproteins, namely pp1a and pp1ab, which leads to the formation of 16 mature non-structural proteins (NSPs)[80]. Composed of 306 aminoacids, the SARS-CoV-2 M^{pro} shares 96% sequence identity and a highly conserved three-dimensional structure with the SARS-CoV M^{pro} (0.53 Å R.M.S.D. between PDB entries 6Y2E and 2BX4)[81,82]. Although a dynamic equilibrium between a monomeric and a dimeric form exists, only the dimer is responsible for the protease's enzymatic activity[83,84]. Each protomer composing the catalytically active dimer is composed of three different domains: the chymotrypsin-like β-barrel domains I (residues 1-99) and II (residues 100-182), which comprehend the substrate binding site and directly control the catalytic

event, and the extra α-helical domain III (residues 198-306), which is connected to the remaining domains through a 16 residues loop and is involved in the dimerization process, thus playing an indirect role in the regulation of M^{pro} catalytic activity[83,85].

The catalytic site is a shallow, solvent-exposed cavity which is formed by several sub-pockets that are responsible for the recognition of various residues composing the substrate peptide sequences[81,85]. Concerning this, particularly important is the conserved sequence Gln↓-Ser, where Gln↓- indicates the glutamine residue that precedes the cleavage site[86].

Despite its peculiar structural features, which makes it a difficult target to drug, rational Structure-Based approaches such as Molecular Docking[87] and Molecular Dynamics[88] have proven to be useful tools in the identification and characterization of M^{pro} small molecule inhibitors, leading to the discovery of both covalent and non-covalent lead compounds[89,90]. Further reinforcing the importance of the Main Protease as a key drug target against COVID-19, is the discovery and approval by regulatory agencies of Nirmatrelvir, the first drug specifically designed against SARS-CoV-2 to enter the market[43].

Due to its pivotal role for the virus replication cycle, the Main Protease is, on the contrary of Spike, particularly conserved in its primary sequence and its threedimensional structural features among different viral strains. Taking a closer look at the main proteases from previously mentioned SARS-CoV-2 variants, only one out of 306 aminoacids is mutated compared to the reference sequence, precisely residue 132, which is a proline in the case of Wuhan-Hu-1, Delta and XD viral strains, while it is mutated to a histidine in the case of Omicron, XE and XF variants.

As can be seen in Figure 7, this mutated residue is located outside the substrate binding site, specifically in a turn that precedes the sequence leading to the oxyanion loop (residues 138-145), which is a vital part of the catalytic machinery that is responsible for the processing of substrate peptides[81]. Although the position of such mutation could suggest a possible destabilization of the catalytic site related to an alteration of the enzymatic activity of the protease, a visual inspection of the surroundings of residue 132 suggests that this mutation should not affect in any way the stability of the

threedimensional structure of the protease, thereby not harming its ability to correctly process the substrate.

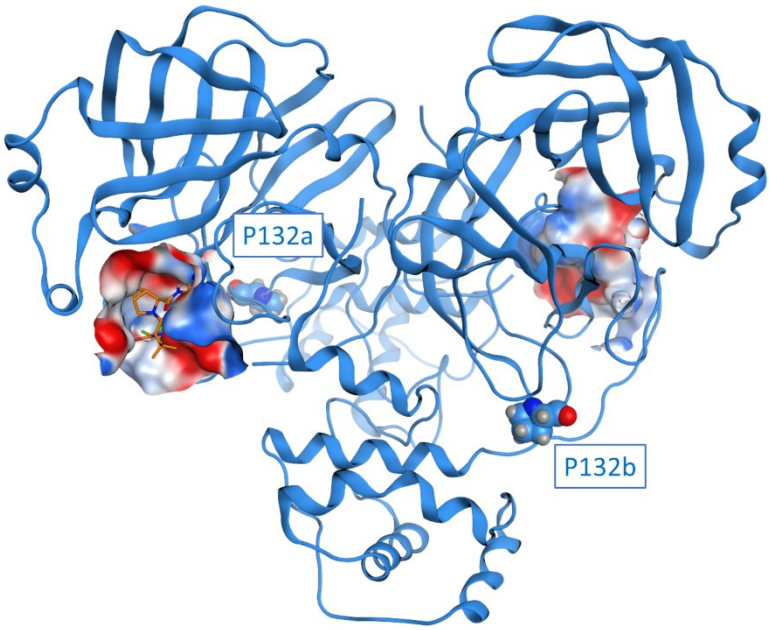


Figure 7. The structure of SARS-CoV-2 M^{pro} (PDB ID: 6Y2E) in its free form. The protein is depicted in blue ribbons, while the mutated residue P132 in comparison with considered SARS-CoV-2 variants (Delta, Omicron, XD, XE and XF) is highlighted and depicted as a CPK model. For visual reference, Nirmatrelvir (also known as PF-07321332, commercial name Paxlovid) from structure 7RFS is also shown in the picture, alongside the binding site surface coloured according to electrostatic properties.

As can be seen in Figure 8, indeed, the proline residue is not involved in any intermolecular interaction relevant for the structural stability of the protease, suggesting that its only role could be limited to a joint between more relevant residues such as R131, which mediates several interactions through its sidechain guanidium group (specifically, a salt bridge with both D289 and D197, and a hydrogen bond with the backbone of T198) and its backbone (a hydrogen bond between its backbone amide proton and the amide carbonyl oxygen of T135 and another one between its carbonyl oxygen and the amide proton of F134), and N133, which is itself involved in a network of intermolecular interactions with both its backbone (hydrogen bond between its amide proton and the carboxyl oxygen of D197) and its sidechain (the amide proton donates to the carbonyl oxygen of G195 while the carbonyl oxygen receives from the hydroxyl group of T135).

These structural insights are confirmed also by a functional screening performed by Flynn et al., which showed that mutations at this position, especially the P132H found in these viral variants, are generally well tolerated, while mutations of both R131 and N133 drastically reduce or even abolish the catalytic activity of the protease[91].

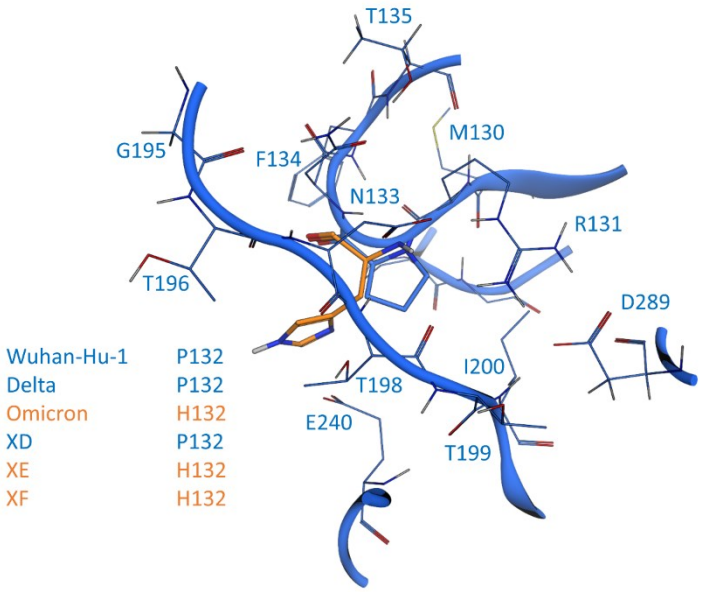


Figure 8. Comparison between SARS-CoV-2 3CL protease (M^{pro}) from crystal structure 6Y2E (blue) and homology models of M^{pro} from five different SARS-CoV-2 variants, reported in Table 1: focus is on residue 132 (either a proline or a histidine) of SARS-CoV-2 M^{pro} and homology models of Delta, Omicron, XD, XE and XF variants M^{pro}.

Concerning the relevance of this mutation for the efficacy of existing therapeutic treatments and the development of future ones, a recent study from Greasley et al. reported the crystal structure of Nirmatrelvir in complex with the main protease from three different viral variants that presented a mutation on M^{pro}[92]. The analyzed mutations included the P132H, which characterizes both the Omicron SARS-CoV-2 variant and the recently found XE and XF. Greasley and collaborators established that the P132H mutation does not affect the affinity of Nirmatrelvir for the main protease catalytic site, thereby indicating the same data would be extendable also to XE and XF variants considered that they share the same P132H mutation as the Omicron variant.

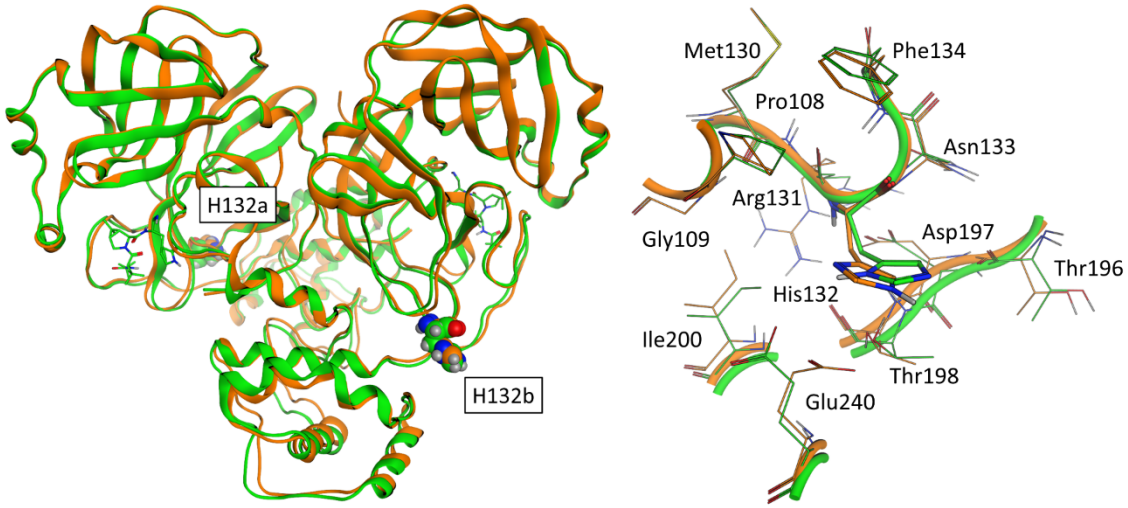


Figure 9. This panel reports the structural superposition between the crystal structure of SARS-CoV-2 Omicron variant M^{pro} in complex with Nirmatrelvir (PDB ID: 7TLL, green) and the homology model of the XE/XF variant based on the unliganded state of the original SARS-CoV-2 virus (orange). In panel A, the whole protease structure is shown in ribbon: for visual reference, both Nirmatrelvir and H132 are reported in licorice and CPK models respectively. In panel B, a focused view of residue 132 and its surrounding residues is reported.

As can be seen in Figure 9, despite the fact that our homology model of the XE/XF variant is based on the structure 6Y2E, which represents the SARS-CoV-2 main protease in its free/unliganded form, there is an almost perfect structural superposition between our homology model and the experimentally resolved structure of the complex between the M^{pro} from the Omicron variant and Nirmatrelvir (PDB ID: 7TLL), as is also quantitatively assessed by the 0.67 Å R.M.S.D. between the two structures after optimal superimposition of the backbone. The congruence between our structural prediction and the experimental data supports the idea that the overall fold of the main protease is conserved across several variants and that the structural effect that residue mutations could have on the effectiveness of main protease inhibitor could be accurately predicted through the combination of computational techniques such as homology modelling, molecular docking and molecular dynamics. Moreover, based on available structural information, the high degree of structural similarity between the main proteases is not only shared by variants of the SARS-CoV-2 virus but also by other coronaviruses such as

bat coronavirus[13], the Porcine transmissible Gastroenteritis virus (TGEV)[83], Human coronavirus strain 229E (HCoV)[85], Infectious bronchitis virus (IBV)[93] and MERS-CoV[94], thereby validating the pursue of novel M^{pro} inhibitors that could act as pan-coronaviral drugs and help preventing future coronavirus associated pandemics.

Conclusions

The recent emergence of novel recombinant SARS-CoV-2 variants, namely XD, XE and XF, poses a serious threat to the efficacy of existing therapeutic options against COVID-19. In the face of a continuous evolution of the SARS-CoV-2 genome under an evolutionary pressure opposed by the development of vaccines and by the natural immunity induced by infections, the more recent viral variants have increased both their infectiveness and their ability to escape the immune response. The structural analysis reported in this article depicts a scenario where the Spike protein, which is responsible for the ability of the virus to infect human cells by interaction with the hACE2 receptor, is the viral entity that is accumulating the highest number of mutations in an attempt to increase its affinity towards the hACE2 and decrease the one towards antibodies, while the main protease M^{pro}, a key enzyme for the virus replication cycle, is still practically identical to the wild-type virus. The difference behaviour of these two proteins in response to SARS-CoV-2 genome evolution could be vital not only for the development of efficient therapies against COVID-19 but, considering the striking structural similarities between the main protease from different viruses, also in the development of pan-coronaviral drugs that could prevent the development of future coronavirus-associated pandemics.

References

1. Lu, H., Stratton, C. W. & Tang, Y. W. Outbreak of pneumonia of unknown etiology in Wuhan, China: The mystery and the miracle. *Journal of Medical Virology* **92**, 401 (2020).

2. Heymann, D. L. & Shindo, N. COVID-19: what is next for public health? *The Lancet* vol. 395 542–545 Preprint at https://doi.org/10.1016/S0140-6736(20)30374-3 (2020).

3. COVID Live - Coronavirus Statistics - Worldometer. https://www.worldometers.info/coronavirus/.

4. Guarner, J. Three Emerging Coronaviruses in Two Decades: The Story of SARS, MERS, and Now COVID-19. *American Journal of Clinical Pathology* vol. 153 420–421 Preprint at https://doi.org/10.1093/ajcp/aqaa029 (2020).

5. Zhou, F. *et al.* Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *The Lancet* **395**, 1054–1062 (2020).

6. Bolcato, G., Bissaro, M., Pavan, M., Sturlese, M. & Moro, S. Targeting the coronavirus SARS-CoV-2: computational insights into the mechanism of action of the protease inhibitors lopinavir, ritonavir and nelfinavir. *Sci Rep* **10**, 20927 (2020).

7. Gorbalenya, A. E. *et al.* The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. *Nature Microbiology 2020 5:4* **5**, 536–544 (2020).

8. Ksiazek, T. G. *et al.* A Novel Coronavirus Associated with Severe Acute Respiratory Syndrome. *New England Journal of Medicine* **348**, 1953–1966 (2003).

9. Peiris, J. S. M. *et al.* Clinical progression and viral load in a community outbreak of coronavirus-associated SARS pneumonia: a prospective study. *The Lancet* **361**, 1767–1772 (2003).

10. Li, W. *et al.* Bats are natural reservoirs of SARS-like coronaviruses. *Science* **310**, 676–679 (2005).

11. Andersen, K. G., Rambaut, A., Lipkin, W. I., Holmes, E. C. & Garry, R. F. The proximal origin of SARS-CoV-2. *Nature Medicine 2020 26:4* **26**, 450–452 (2020).

12. Temmam, S. *et al.* Bat coronaviruses related to SARS-CoV-2 and infectious for human cells. *Nature* (2022) doi:10.1038/s41586-022-04532-4.

13. Pavan, M., Bassani, D., Sturlese, M. & Moro, S. Bat coronaviruses related to SARS-CoV-2: what about their 3CL proteases (MPro)? *https://doi.org/10.1080/14756366.2022.2062336* **37**, 1077–1082 (2022).

14. Forster, P., Forster, L., Renfrew, C. & Forster, M. Phylogenetic network analysis of SARS-CoV-2 genomes. *Proceedings of the National Academy of Sciences of the United States of America* **117**, 9241–9243 (2020).

15. Harvey, W. T. *et al.* SARS-CoV-2 variants, spike mutations and immune escape. *Nature Reviews Microbiology* **19**, 409–424 (2021).

16. Chu, D. K. *et al.* Physical distancing, face masks, and eye protection to prevent person-to-person transmission of SARS-CoV-2 and COVID-19: a systematic review and meta-analysis. *The Lancet* **395**, 1973–1987 (2020).

17. Hellewell, J. *et al.* Feasibility of controlling COVID-19 outbreaks by isolation of cases and contacts. *The Lancet Global Health* **8**, e488–e496 (2020).

18. Oran, D. P. & Topol, E. J. Prevalence of Asymptomatic SARS-CoV-2 Infection. *https://doi.org/10.7326/M20-3012* **173**, 362–368 (2020).

19. Garcia-Beltran, W. F. *et al.* Multiple SARS-CoV-2 variants escape neutralization by vaccine-induced humoral immunity. *Cell* **184**, 2372-2383.e9 (2021).

20. Hoffmann, M. *et al.* SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell* **181**, 271-280.e8 (2020).

21. Li, Q. *et al.* The Impact of Mutations in SARS-CoV-2 Spike on Viral Infectivity and Antigenicity. *Cell* **182**, 1284-1294.e9 (2020).

22. Frampton, D. *et al.* Genomic characteristics and clinical effect of the emergent SARS-CoV-2 B.1.1.7 lineage in London, UK: a whole-genome sequencing and hospital-based cohort study. *The Lancet. Infectious Diseases* **21**, 1246 (2021).

23. Campbell, F. *et al.* Increased transmissibility and global spread of SARSCoV- 2 variants of concern as at June 2021. *Eurosurveillance* **26**, 1–6 (2021).

24. Davies, N. G. *et al.* Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. *Science* **372**, (2021).

25. Volz, E. *et al.* Assessing transmissibility of SARS-CoV-2 lineage B.1.1.7 in England. *Nature* **593**, 266–269 (2021).

26. Collier, D. A. *et al.* Sensitivity of SARS-CoV-2 B.1.1.7 to mRNA vaccine-elicited antibodies. *Nature* **593**, 136–141 (2021).

27. Chen, R. E. *et al.* Resistance of SARS-CoV-2 variants to neutralization by monoclonal and serum-derived polyclonal antibodies. *Nature Medicine* **27**, 717–726 (2021).

28. Wang, P. *et al.* Antibody resistance of SARS-CoV-2 variants B.1.351 and B.1.1.7. *Nature* **593**, 130–135 (2021).

29. Haas, E. J. *et al.* Impact and effectiveness of mRNA BNT162b2 vaccine against SARS-CoV-2 infections and COVID-19 cases, hospitalisations, and deaths following a nationwide vaccination campaign in Israel: an observational study using national surveillance data. *The Lancet* **397**, 1819–1829 (2021).

30. Twohig, K. A. *et al.* Hospital admission and emergency care attendance risk for SARS-CoV-2 delta (B.1.617.2) compared with alpha (B.1.1.7) variants of concern: a cohort study. *The Lancet Infectious Diseases* **22**, 35–42 (2022).

31. Lopez Bernal, J. *et al.* Effectiveness of Covid-19 Vaccines against the B.1.617.2 (Delta) Variant. *New England Journal of Medicine* **385**, 585–594 (2021).

32. Planas, D. *et al.* Reduced sensitivity of SARS-CoV-2 variant Delta to antibody neutralization. *Nature* **596**, 276–280 (2021).

33. Liu, C. *et al.* Reduced neutralization of SARS-CoV-2 B.1.617 by vaccine and convalescent serum. *Cell* **184**, 4220-4236.e13 (2021).

34. Mlcochova, P. *et al.* SARS-CoV-2 B.1.617.2 Delta variant replication and immune evasion. *Nature* (2021) doi:10.1038/S41586-021-03944-Y.

35. Gao, S. J., Guo, H. & Luo, G. Omicron variant (B.1.1.529) of SARS-CoV-2, a global urgent public health alert! *Journal of Medical Virology* **94**, 1255–1256 (2022).

36. Araf, Y. *et al.* Omicron variant of SARS-CoV-2: Genomics, transmissibility, and responses to current COVID-19 vaccines. *Journal of Medical Virology* **94**, 1825–1832 (2022).

37. Liu, L. *et al.* Striking antibody evasion manifested by the Omicron variant of SARS-CoV-2. *Nature* **602**, 676–681 (2022).

38. Dejnirattisai, W. *et al.* Reduced neutralisation of SARS-CoV-2 omicron B.1.1.529 variant by post-immunisation serum. *The Lancet* **399**, 234–236 (2022).

39. Cao, Y. *et al.* Omicron escapes the majority of existing SARS-CoV-2 neutralizing antibodies. *Nature* **602**, 657–663 (2022).

40. Hoffmann, M. *et al.* The Omicron variant is highly resistant against antibody-mediated neutralization: Implications for control of the COVID-19 pandemic. *Cell* **185**, 447-456.e11 (2022).

41. Nemet, I. *et al.* Third BNT162b2 Vaccination Neutralization of SARS-CoV-2 Omicron Infection. *New England Journal of Medicine* **386**, 492–494 (2022).

42. Garcia-Beltran, W. F. *et al.* mRNA-based COVID-19 vaccine boosters induce neutralizing immunity against SARS-CoV-2 Omicron variant. *Cell* **185**, 457-466.e4 (2022).

43. Owen, D. R. *et al.* An oral SARS-CoV-2 M pro inhibitor clinical candidate for the treatment of COVID-19. *Science (1979)* **374**, 1586–1593 (2021).

44. Pavan, M., Bolcato, G., Bassani, D., Sturlese, M. & Moro, S. Supervised Molecular Dynamics (SuMD) Insights into the mechanism of action of SARS-CoV-2 main protease inhibitor PF-07321332. *J Enzyme Inhib Med Chem* **36**, 1646–1650 (2021).

45. Hammond, J. *et al.* Oral Nirmatrelvir for High-Risk, Nonhospitalized Adults with Covid-19. *New England Journal of Medicine* (2022) doi:10.1056/NEJMoa2118542.

46. Investigation of SARS-CoV-2 variants: technical briefings - GOV.UK. https://www.gov.uk/government/publications/investigation-of-sars-cov-2-variants-technical-briefings.

47. Lacek, K. A. *et al.* Identification of a Novel SARS-CoV-2 Delta-Omicron Recombinant Virus in the United States. *bioRxiv* 2022.03.19.484981 (2022) doi:10.1101/2022.03.19.484981.

48. Aksamentov, I., Roemer, C., Hodcroft, E. & Neher, R. Nextclade: clade assignment, mutation calling and quality control for viral genomes. *Journal of Open Source Software* **6**, 3773 (2021).

49. Aksamentov, I., Roemer, C., Hodcroft, E. & Neher, R. Nextclade: clade assignment, mutation calling and quality control for viral genomes. *Journal of Open Source Software* **6**, 3773 (2021).

50. Molecular Operating Environment (MOE), 2019.01; Chemical Computing Group ULC, 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, 2021. https://www.chemcomp.com/Research-Citing_MOE.htm.

51. Zhou, D. *et al.* Structural basis for the neutralization of SARS-CoV-2 by an antibody from a convalescent patient. *Nature Structural & Molecular Biology* **27**, 950–958 (2020).

52. Wang, Y. *et al.* Structural basis for SARS-CoV-2 Delta variant recognition of ACE2 receptor and broadly neutralizing antibodies. *Nature Communications* **13**, 871 (2022).

53. Yin, W. *et al.* Structures of the Omicron spike trimer with ACE2 and an anti-Omicron antibody. *Science* **375**, 1048–1053 (2022).

54. D.A. Case et al. Amber 10, University of California, San Francisco.

55. Humphrey, W., Dalke, A. & Schulten, K. VMD: Visual molecular dynamics. *J Mol Graph* **14**, 33–38 (1996).

56. Huang, Y., Yang, C., Xu, X., Xu, W. & Liu, S. Structural and functional properties of SARS-CoV-2 spike protein: potential antivirus drug development for COVID-19. *Acta Pharmacologica Sinica* **41**, 1141–1149 (2020).

57. Li, M.-Y., Li, L., Zhang, Y. & Wang, X.-S. Expression of the SARS-CoV-2 cell receptor gene ACE2 in a wide variety of human tissues. *Infectious Diseases of Poverty* **9**, 45 (2020).

58. Jackson, C. B., Farzan, M., Chen, B. & Choe, H. Mechanisms of SARS-CoV-2 entry into cells. *Nature Reviews Molecular Cell Biology* **23**, 3–20 (2022).

59. Lan, J. *et al.* Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature* **581**, 215–220 (2020).

60. Zhu, C. *et al.* Molecular biology of the SARs-CoV-2 spike protein: A review of current knowledge. *Journal of Medical Virology* **93**, 5729–5741 (2021).

61. Turkahia, Y. *et al.* Pandemic-Scale Phylogenomics Reveals Elevated Recombination Rates in the SARS-CoV-2 Spike Region. *bioRxiv* 2021.08.04.455157 (2021) doi:10.1101/2021.08.04.455157.

62. Zhang, B. *et al.* Mining of epitopes on spike protein of SARS-CoV-2 from COVID-19 patients. *Cell Research* **30**, 702–704 (2020).

63. Aksamentov, I., Roemer, C., Hodcroft, E. & Neher, R. Nextclade: clade assignment, mutation calling and quality control for viral genomes. *Journal of Open Source Software* **6**, 3773 (2021).

64. Hadj Hassine, I. Covid-19 vaccines and variants of concern: A review. *Reviews in Medical Virology* (2021) doi:10.1002/rmv.2313.

65. University, S. VOC - Omicron.

66. VanBlargan, L. A. *et al.* An infectious SARS-CoV-2 B.1.1.529 Omicron virus escapes neutralization by therapeutic monoclonal antibodies. *Nature Medicine* **28**, 490–495 (2022).

67. Meng, B. *et al.* Recurrent emergence of SARS-CoV-2 spike deletion H69/V70 and its role in the Alpha variant B.1.1.7. *Cell Reports* **35**, 109292 (2021).

68. Hou, Y. J. *et al.* SARS-CoV-2 D614G variant exhibits efficient replication ex vivo and transmission in vivo. *Science* **370**, 1464–1468 (2020).

69. Weisblum, Y. *et al.* Escape from neutralizing antibodies 1 by SARS-CoV-2 spike protein variants. *eLife* **9**, 1 (2020).

70. Lubinski, B. *et al.* Functional evaluation of the P681H mutation on the proteolytic activation of the SARS-CoV-2 variant B.1.1.7 (Alpha) spike. *iScience* **25**, 103589 (2022).

71. Zahradník, J. *et al.* SARS-CoV-2 variant prediction and antiviral drug design are enabled by RBD in vitro evolution. *Nature Microbiology* **6**, 1188–1198 (2021).

72. Berman, H. M. The Protein Data Bank. *Nucleic Acids Research* **28**, 235–242 (2000).

73. Wang, Y. *et al.* Structural basis for SARS-CoV-2 Delta variant recognition of ACE2 receptor and broadly neutralizing antibodies. *Nature Communications* **13**, 871 (2022).

74. Lan, J. *et al.* Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature* **581**, 215–220 (2020).

75. Collier, D. A. *et al.* Sensitivity of SARS-CoV-2 B.1.1.7 to mRNA vaccine-elicited antibodies. *Nature* **593**, 136–141 (2021).

76. Masre, S. F., Jufri, N. F., Ibrahim, F. W. & Abdul Raub, S. H. Classical and alternative receptors for SARS-CoV-2 therapeutic strategy. *Reviews in Medical Virology* **31**, 1–9 (2021).

77. Sartore, G. *et al.* In silico evaluation of the interaction between ACE2 and SARS-CoV-2 Spike protein in a hyperglycemic environment. *Scientific Reports* **11**, 22860 (2021).

78. Jin, Z. *et al.* Structure of Mpro from SARS-CoV-2 and discovery of its inhibitors. *Nature 2020 582:7811* **582**, 289–293 (2020).

79. Xia, B. & Kang, X. Activation and maturation of SARS-CoV main protease. *Protein & cell* **2**, 282–290 (2011).

80. Snijder, E. J., Decroly, E. & Ziebuhr, J. The Nonstructural Proteins Directing Coronavirus RNA Synthesis and Processing. *Advances in Virus Research* **96**, 59–126 (2016).

81. Fornasier, E. *et al.* A new inactive conformation of SARS-CoV-2 main protease. *Acta Crystallogr D Struct Biol* **78**, 363–378 (2022).

82. Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature 2020 579:7798* **579**, 265–269 (2020).

83. Anand, K. *et al.* Structure of coronavirus main proteinase reveals combination of a chymotrypsin fold with an extra alpha-helical domain. *EMBO J* **21**, 3213–3224 (2002).

84. Chen, H. *et al.* Only one protomer is active in the dimer of SARS 3C-like proteinase. *Journal of Biological Chemistry* **281**, 13894–13898 (2006).

85. Anand, K., Ziebuhr, J., Wadhwani, P., Mesters, J. R. & Hilgenfeld, R. Coronavirus main proteinase (3CLpro) structure: basis for design of anti-SARS drugs. *Science* **300**, 1763–1767 (2003).

86. Ullrich, S. & Nitsche, C. The SARS-CoV-2 main protease as drug target. *Bioorg Med Chem Lett* **30**, 127377 (2020).

87. Bassani, D., Pavan, M., Bolcato, G., Sturlese, M. & Moro, S. Re-Exploring the Ability of Common Docking Programs to Correctly Reproduce the Binding Modes of Non-Covalent Inhibitors of SARS-CoV-2 Protease Mpro. *Pharmaceuticals* **15**, 180 (2022).

88. Bissaro, M. *et al.* Inspecting the Mechanism of Fragment Hits Binding on SARS-CoV-2 M ^{pro} by Using Supervised Molecular Dynamics (SuMD) Simulations. *ChemMedChem* **16**, 2075–2081 (2021).

89. Luttens, A. *et al.* Ultralarge Virtual Screening Identifies SARS-CoV-2 Main Protease Inhibitors with Broad-Spectrum Activity against Coronaviruses. *J Am Chem Soc* **144**, 2905–2920 (2022).

90. Zhang, C. H. *et al.* Potent Noncovalent Inhibitors of the Main Protease of SARS-CoV-2 from Molecular Sculpting of the Drug Perampanel Guided by Free Energy Perturbation Calculations. *ACS Cent Sci* **7**, 467–475 (2021).

91. Flynn, J. M. *et al.* Comprehensive fitness landscape of SARS-CoV-2 M^{pro} reveals insights into viral resistance mechanisms. *bioRxiv* 2022.01.26.477860 (2022) doi:10.1101/2022.01.26.477860.

92. Greasley, S. E. *et al.* Structural basis for Nirmatrelvir in vitro efficacy against SARS-CoV-2 variants. *bioRxiv* 2022.01.17.476556 (2022) doi:10.1101/2022.01.17.476556.

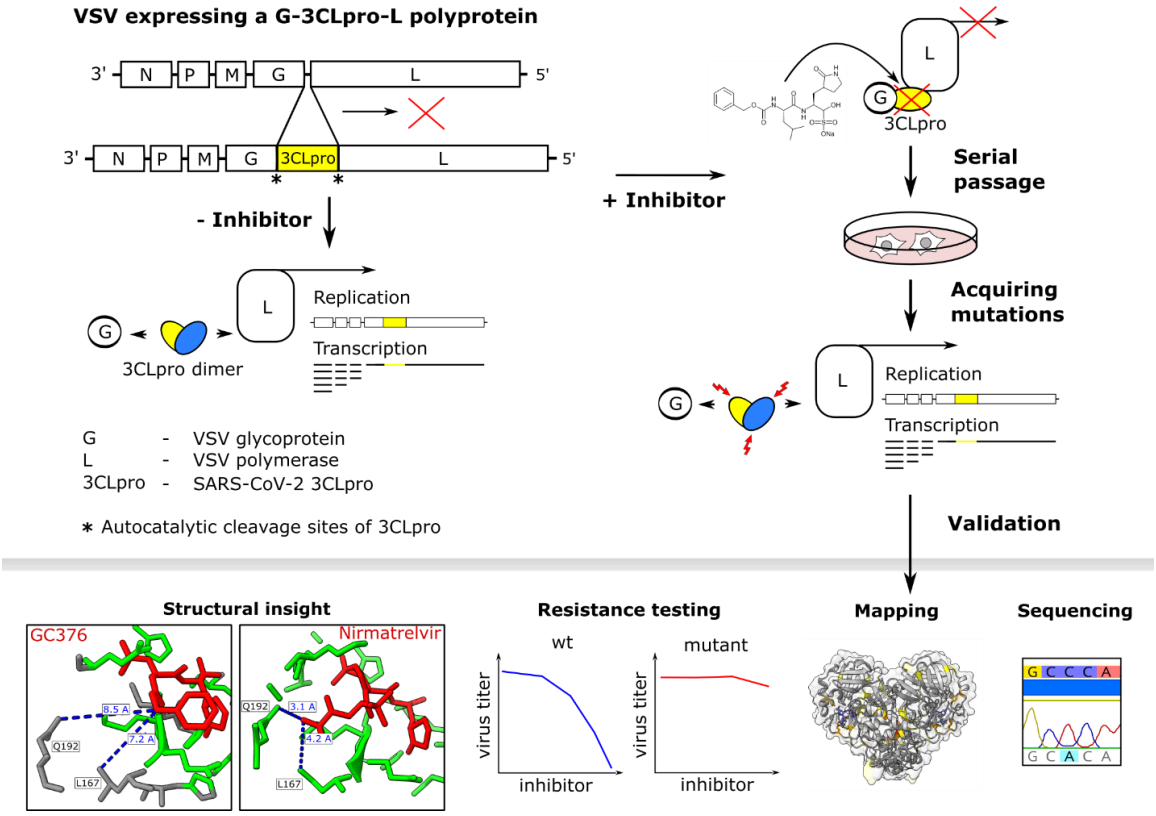
93. Xue, X. *et al.* Production of authentic SARS-CoV M(pro) with enhanced activity: application as a novel tag-cleavage endopeptidase for protein overproduction. *J Mol Biol* **366**, 965–975 (2007).

94. Ho, B. L. *et al.* Critical Assessment of the Important Residues Involved in the Dimerization and Catalysis of MERS Coronavirus Main Protease. *PLoS One* **10**, (2015).

SARS-CoV-2 3CL^{pro} mutations selected in a VSV-based system confer resistance to nirmatrelvir, ensitrelvir, and GC376

Emmanuel Heilmann, Francesco Costacurta, Seyed Arad Moghadasi, Chengjin Ye, **Matteo Pavan**, Davide Bassani, Andre Volland, Claudia Ascher, Alexander Kurt Hermann Weiss, David Bante, Reuben Stewart Harris, Stefano Moro, Bernhard Rupp, Luis Martinez-Sobrido and Dorothee von Laer

Heilmann, E. *et al.* SARS-CoV-2 3CL ^{pro} mutations selected in a VSV-based system confer resistance to nirmatrelvir, ensitrelvir, and GC376. *Sci Transl Med* (2022) doi:10.1126/scitranslmed.abq7360.



Abstract

Protease inhibitors are among the most powerful antiviral drugs. Nirmatrelvir is the first protease inhibitor against the SARS-CoV-2 protease 3CL^{pro} that has been licensed for clinical use. To identify mutations that confer resistance to this protease-inhibitor, we engineered a chimeric vesicular stomatitis virus (VSV) that expressed a polyprotein composed of the VSV glycoprotein G, the SARS-CoV-2 3CL^{pro}, and the VSV polymerase L.

Viral replication was thus dependent on the autocatalytic processing of this precursor protein by $3CL^{pro}$ and release of the functional viral polymerase L, and replication of this chimeric VSV was effectively inhibited by nirmatrelvir.

Using this system, we applied nirmatrelvir to select for resistance mutations. Resistance was confirmed by retesting nirmatrelvir against the selected mutations in an additional VSV-based systems, in an independently developed cellular system, in a biochemical assay, and in a recombinant SARS-CoV-2 system. We demonstrate that some mutants are cross-resistant to ensitrelvir and GC376, whereas others are less resistant to these compounds. Furthermore, we found that most of these resistance mutations already existed in SARS-CoV-2 sequences that have been deposited in the NCBI and GISAID databases, indicating that these mutations were present in circulating SARS-CoV-2 strains.

Introduction

In late 2019, the zoonotic transmission of a new coronavirus, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), into the human population[1], has led to worldwide efforts to find effective treatments against the various pathologies caused by the virus. Inhibitors of viral enzymes, such as proteases, have proven to be highly potent drugs in the treatment of HIV and Hepatitis C virus infections. However, resistant viruses rapidly emerge unless the protease inhibitors are given in combination with other directly acting antivirals[2,3]. SARS-CoV-2 encodes two proteases. The 3-Chymo-trypsin-like protease ($3CL^{pro}$) cleaves 11 sites in the viral polyproteins pp1a and pp1ab and is also referred to as the main protease (M^{pro}) or non-structural protein 5 (nsp5), indicating that it cleaves more sites than the second protease and its location within the polyproteins, respectively[4]. The second viral protease, Papain-like protease (PL^{pro}) cleaves three additional sites in pp1a and pp1ab[5]. Thus, both proteases are essential for viral replication and therefore interesting drug targets.

Recently, the $3CL^{pro}$ inhibitor nirmatrelvir was approved for clinical use. Nirmatrelvir acts as a peptidomimetic, covalent inhibitor binding to the catalytic site cysteine (C145), thereby blocking its function[6–8]. Nirmatrelvir has been authorized in combination with ritonavir by the U.S. food and drug administration (FDA) for emergency use in high-risk

SARS-CoV-2-infected individuals under the trade name Paxlovid (EUA 105 Pfizer Paxlovid, 22.12.2021). In the studies leading to the Paxlovid (nirmatrelvir/ritonavir) emergency use authorization (EUA), mouse hepatitis virus (MHV) $3CL^{pro}$ was used as a surrogate for SARS-CoV-2 $3CL^{pro}$ to generate resistance data, which may not be the ideal system. In addition, very recently several preprints have described nirmatrelvir resistance mutations in authentic SARS-CoV-2, either generated de novo[9,10], found in isolates[11,12], or modelled in silico[13]. Working with SARS-CoV-2 requires biosafety level 3 (BSL- 3) installations due to its virulence[14]. Even more so, performing SARS-CoV-2 antibody or antiviral resistance studies demand utmost caution to avoid biosafety breaches and subsequent spread of mutant variants. To address these caveats, we describe in this study a BSL-2 system based on VSV that allows the selection of resistance mutations in the SARS-CoV-2 $3CL^{pro}$. Several mutations identified were validated in cell-based, biochemical, and recombinant SARS-CoV-2 assays and two mutations were found to be identical to (L167F) or at the same residue (Q192) as those described in the other manuscripts characterizing resistance in authentic SARS-CoV-2. We further-more showed that some mutations selected by one $3CL^{pro}$ inhibitor can confer cross-resistance to other inhibitors. In contrast, other mutations appeared more inhibitor-specific. We attributed these effects to the distinct chemical structures of the inhibitors and occupation within the active site of the $3CL^{pro}$. Lastly, we modelled catalytic site mutations with Robetta[15] and Molecular Operating Environment[16] to elucidate their mechanism of resistance.

Results

MHV $3CL^{pro}$ is less sensitive to nirmatrelvir than SARS-CoV-2 $3CL^{pro}$

We compared the sensitivity of SARS-CoV-2 and MHV $3CL^{pro}$ to the active component of Paxlovid, nirmatrelvir, using the gain-of- signal variant of a VSV-based $3CL^{pro}$ measurement assay shown in fig. S1A and B and described in detail recently[17]. In brief, the coronavirus $3CL^{pro}$ proteases flanked by autocatalytic sites were cloned into the P protein of a red fluorescent protein (dsRed) ex-pressing VSV. The P protein with the internal $3CL^{pro}$ is functional and essential for viral genome replication and dsRed expression. In the absence of protease inhibitor, the P:$3CL^{pro}$ protein is autocatalytically cleaved and dsRed is not expressed. In the presence of a protease inhibitor, the P protein

is functional, the VSV genome replicates and dsRed is expressed. Using this system, we found that MHV 3CL^{pro} showed a weaker response to nirmatrelvir than to SARS-CoV-2 3CL^{pro} (Fig. 1A). The sequence identity of the two proteins is only 50% (fig. S2), whereas the structures of the SARS-CoV-2 and MHV 3CL^{pro} enzymes are strongly conserved. However, the interaction site of nirmatrelvir (a distance of 5 Å or less from the compound) shows seven amino acid differences between the two enzymes, namely H164 - Q, M165 - L, P168 - S, V186 - R, R188 - A, T190 - V and A191 - V (counting from the first residue (serine) after the glutamine of the N-terminal cleavage site). We therefore suggest that MHV 3CL^{pro} is not an optimal proxy to study SARS resistance mutations.

A VSV-based non-gain-of-function system was generated to predict SARS-CoV-2 3CL^{pro} mutations

To generate a safer alternative to selection of drug resistant SARS-CoV-2 for studying mutants, we engineered a chimeric VSV variant, where the intergenic region between the glycoprotein (G) and the polymerase (L) was replaced by the 3CL^{pro} of SARS-CoV-2 (fig. S3A). Upon translation, G, 3CL^{pro}, and L form a surrogate polyprotein, which must be processed by 3CL^{pro} to generate the functional viral proteins G and L. This surrogate polyprotein mimics the polyprotein that is produced by SARS-CoV-2 as dimerization of 3CL^{pro} is obligate for its function[18,19] and cleavage of the cognate 3CL^{pro} N- and C-terminal motifs must occur for successful VSV replication. By applying an appropriate protease inhibitor (+PI), this processing is disturbed and therefore viral replication cannot occur (fig. S3B). Through passaging the chimeric VSV variant in presence of suboptimal concentrations of a protease inhibitor, 3CL^{pro} mutations that are generated by the error-prone viral polymerase[20,21] are selected for resistance to the inhibitor (fig. S3C). In a first proof-of-concept study, we selected a mutant against the inhibitor GC376, which acquired the amino acid change in the 3CL^{pro} from phenylalanine to leucine at position 305 (F305L) in the autocatalytic cleavage motif at the C terminus of the protease. This virus gained a mildly faster replication kinetic and produced higher titers in the presence of GC376 and nirmatrelvir compared to the parental virus (Fig. 1B and C). Related coronaviruses have leucine at position 305 as a preferred cleavage motif (Fig. 1D); therefore, the likely mechanism of the selection of F305L is auto-cleavage site optimization. We used the wild-

type VSV-3CL^{pro} for subsequent mutation selection studies with nirmatrelvir. We also included the F305L as parental virus for further selection experiments, because the F305L mutation has been found in regional out-breaks (mostly in England) and has been deposited in the global initiative on sharing avian flu data (GISAID) database[22–24] with three different codon usages to obtain leucine instead of phenylalanine (fig. S4A). The mutants were variants of Delta, mainly the sublineage AY.4 (fig. S4B). We therefore assumed F305L could be an advantageous mutation that, in combination with further mutations, may give rise to protease inhibitor resistant lineages.

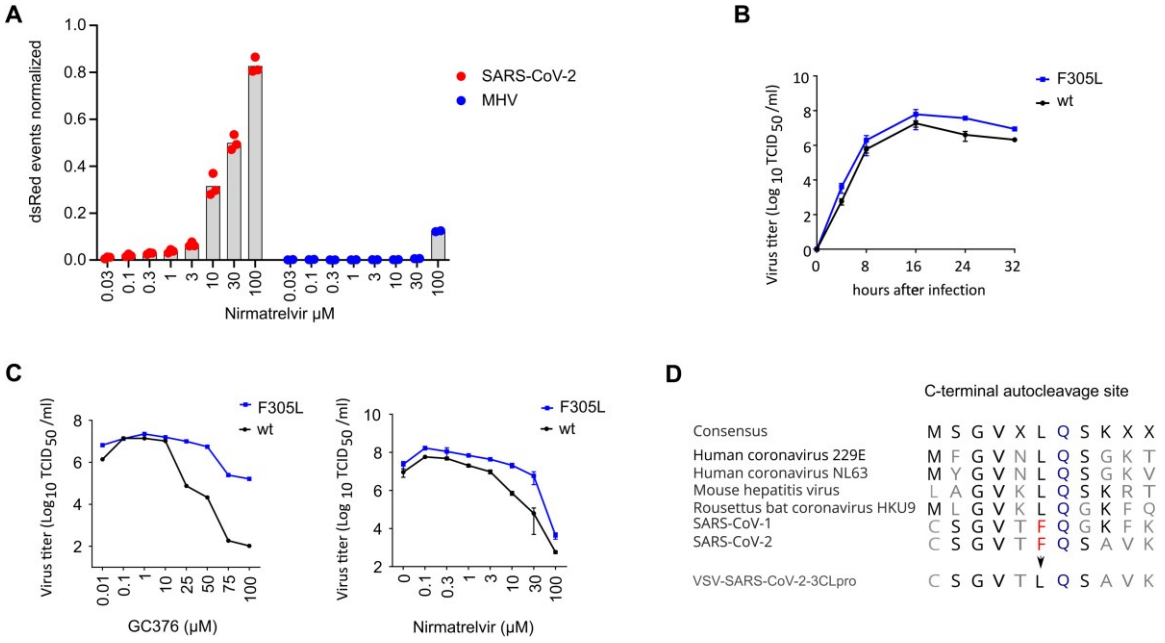


Figure 1. A VSV-based non-gain-of-function system was developed to predict SARS-CoV-2 3CL^{pro} mutations. (A) 3CL^{pro} from SARS-CoV-2 and mouse hepatitis virus (MHV) were tested in a gain-of-signal assay. Data are presented as individual points of n = 3 biologically independent replicates per condition for SARS-CoV-2 3CL^{pro} and n = 2 for MHV 3CL^{pro}, average values are represented by histogram bars. (B) Replication kinetics are shown for wild-type (wt) VSV-G-3CL^{pro}-L and GC376-selected F305L variant. Data are presented as SD of n = 2 biologically independent replicates per condition. (C) GC376 and nirmatrelvir dose responses are shown for wild-type (wt) VSV-G-3CL^{pro}-L and GC376-selected F305L variant. Data are presented as means of n = 2 (GC376) and n = 3 (Nirmatrelvir) biologically independent replicates per condition. (D) Sequence alignment of C-terminal autocleavage sites is shown for SARS-CoV-2 3CL^{pro} and related coronaviruses.

Nirmatrelvir resistant 3CL^{pro} mutants were selected for in the VSV-3CL^{pro} system

We next used the wild-type and F305L mutant viruses to select for nirmatrelvir resistant 3CL^{pro}. BHK-21 cells in a 96-well plate were infected at a low multiplicity of infection (MOI; 0.01). Where cytopathic effects were visible in the first passage (25 out of 48 wells from parental wild-type and 17 out of 48 wells from parental F305L), supernatants were used for passaging individual wells with increasing concentrations (wild-type initial infection: 30 µM, second round: 40 µM, and third round: 50 µM; F305L initial infection: 50 µM, second round: 75 µM, and third round: 100 µM) of nirmatrelvir. At every passage, where cytopathic effects were observed again, supernatants were collected from the cell culture of individual 96-wells and transferred to individual new wells of a 96-well plate. At every passage, each well was sequenced individually, the target region being 3CL^{pro} and adjacent parts of G and L. We only counted mutants from un-ambiguous chromatogram peaks (as exemplified in fig. S3C). If, in the first or second passage there were still overlapping peaks, we sequenced the well again after the next passage. By this continuous selection pressure, the fittest mutant virus variant became dominant over the wild-type (and potential other mutants) in each well and made up the entirety of the genomic RNA, cDNA, and subsequent polymerase chain reaction (PCR) fragment. To finally exclude minority mutant populations that were not visible in a Sanger sequencing chromatogram, but could contribute to the resistance phenotype, mutations were later re-introduced into 3CL^{pro} measurement systems individually. We found 39 distinct mutations within 3CL^{pro} by Sanger sequencing. Viruses carried from one dominant mutation up to four mutations. The mutations were distributed over the entire sequence of 3CL^{pro} (Fig. 2A, table S1). We categorized them into catalytic site, near-catalytic site, dimerization interface and autocleavage site mutants. A fourth category for all mutations not fitting the first three was chosen as “allosteric” mutants. The mutants Y54C, L141F, L167F and Q192R occurred in residues in very close proximity (within 5 Å of PDB 3CL^{pro} structure 7vh8, Fig. 2B, table S2) to nirmatrelvir. We searched for the mutants in the National Center for Biotechnology Information (NCBI) Virus data base[25] and GISAID EpiCoV[22–24], and found most of the mutations, or at least the same residue with a different mutation, in deposited sequences with varying coverage (Fig. 2A, table S1). We further subdivided GISAID entries into de-

positions made before and after the emergency use authorization of Paxlovid (nirmatrelvir/ritonavir) on 22nd December 2021 (table S3). An update of the Paxlovid EUA (18th March 2022) included 3CL^{pro} mutants that were retrieved from patients treated with Paxlovid (nirmatrelvir/ritonavir) (Fig. 2A). The update stated that it was unclear whether these mutations had clinical relevance[26].

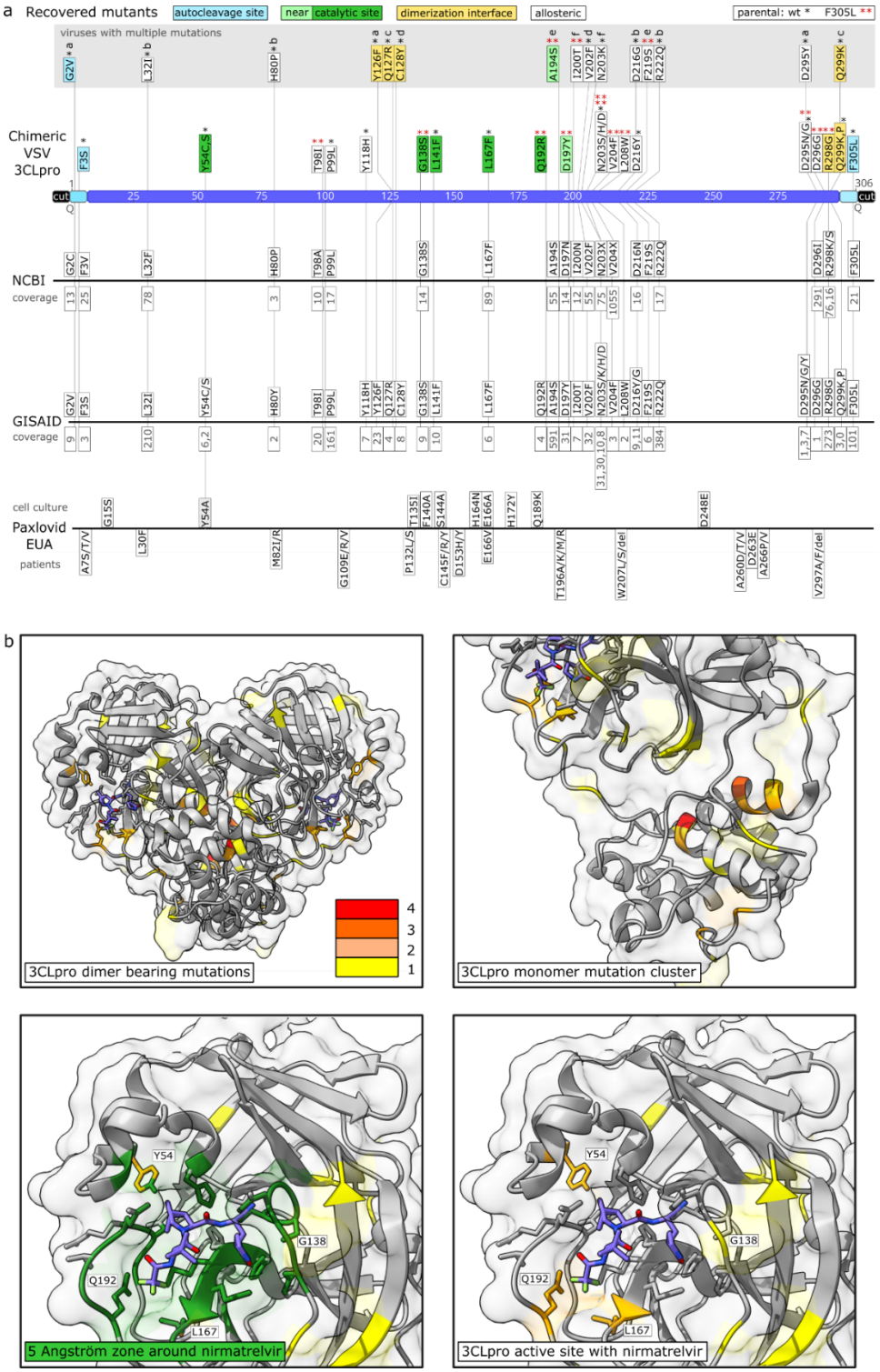


Figure 2. Sequencing of 3CLpro escape mutants and comparison to databases and Paxlovid EUA information. (A) Mutants were recovered from VSV-G-3CL^{pro}-L wild- type (*) and the F305L variant (red **). Autocleavage site mutants are colored in turquoise, catalytic site mutants in green, near catalytic site mutants in light green, dimerization interface mutants in yellow and "allosteric" mutants in white. Viruses with more than one mutation are displayed above in a gray box and named a to f. The number of mutated sequences in the databases from NCBI and GISAID are displayed below the mutations in gray. If specific mutations were not present in the database, the residue is displayed with any mutation that occurred at this position. Multiple such different amino acid changes that were not selected in our virus are displayed with X (N203X, V204X). Mutations from the Paxlovid EUA are divided into mutations found in cell culture and mutations sequenced from treated patients. The coverage of mutation entries was obtained on June second, 2022. (B) Visualizations of mutation-affected residues are shown. Residues that were mutated one time are highlighted in yellow, two times in light orange, three times dark orange, and four times in red. The 3CL^{pro} protease dimer with bound nirmatrelvir (blue) was visualized in ChimeraX from the Protein Data Bank structure 7vh8 (32). Catalytic center mutations are within a range of 5 Å as visualized in dark green.

Replications kinetics and dose response were analyzed for selected 3CL^{pro} mutants

To confirm these potential resistance mutations, we chose six virus samples to perform replication kinetics and dose response experiments. Mutants were selected for further testing based on two criteria. First, we chose virus variants with catalytic site mutations because alterations in drug binding residues (direct or indirect) are more likely to alter efficacy. Second, we chose the most frequently recovered mutant outside of the catalytic site. Four samples were derived from wild-type VSV-G-3CL^{pro}-L and two were derived from the F305L variant. Supernatants for the replication kinetic experiments were collected at indicated time points after infection, and supernatants for the virus nirmatrelvir dose response experiments were collected 24 hours after infection. The replication kinetics revealed that all variants were still capable of replicating to high titers, suggesting that resistance mutations did not result in a strong negative effect on 3CL^{pro} activity (Fig. 3). The dose responses showed that wild-type VSV-G-3CL^{pro}-L replication was inhibited 106-fold at 100 µM of nirmatrelvir, with a half maximal inhibitor concentration (IC_{50}) of about 185 nM (Fig. 3A). We tested two L167F variants, because this mutant arose twice independently. The similarity of their dose responses (Fig. 3B and C), as well as the low variation of the biological replicates, suggests that the differences in the degree of resistance we observed between the mutants were not artifacts. We tested additional

single mutants, namely the catalytic site mutant Y54C (Fig. 3D), a mutant from the mutation cluster shown in Fig. 2B, N203D (Fig. 3E) and the autocleavage site mutant F305L (Fig. 3F). To test if mutants that were selected from the F305L background had increased resistance, we also tested double mutants G138S/F305L (Fig. 3G) and Q192R/F305L (Fig. 3H). We observed the strongest resistance phenotype in the double mutant Q192R/F305L, replicating to high viral titers with a pronounced cytopathic effect (fig. S5) even in the presence of 100 µM nirmatrelvir.

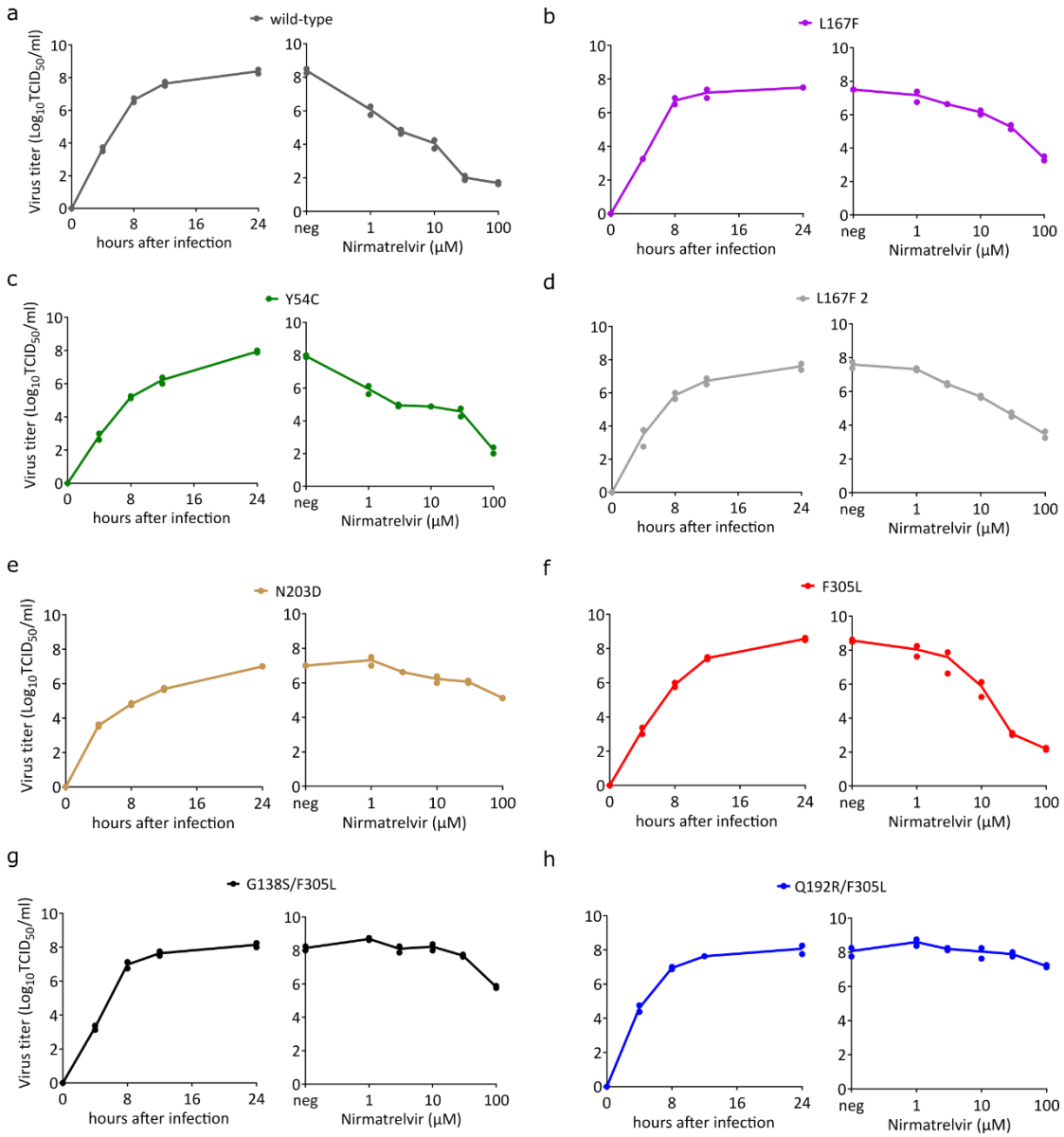


Figure 3. Replication kinetics and nirmatrelvir dose responses of parental VSV-G-3CL^{pro}-L and mutant variants. (A to H) Replication kinetics and dose responses are shown for wild-type (A), L167F (B), L167F-2 (C), Y54C (D), N203D (E), F305L (F), G138S/F305L (G) and Q192R/F305L (H) VSV-G-3CL^{pro}-L. Supernatants

for replication kinetics were collected at indicated time points. Supernatants for virus nirmatrelvir dose response were collected 24 hours after infection. (n = 2 biologically independent replicates per condition with individual data points shown and connecting lines of mean values). neg, without nirmatrelvir; $TCID_{50}$, 50% tissue culture infective dose.

Re-introduction of 3CL^{pro} mutations confirms their resistance phenotype

As shown in table S1, VSV-induced 3CL^{pro} mutations were observed after the first passage when nirmatrelvir was applied. To validate the resistance data of replication-competent VSV-3CL^{pro} and at the same time exclude the effects of potential additional mutations arising within the dose response experiment, we re-introduced some of the catalytic center mutations (Y54C, L167F, Q192R) into a recently developed protease activity measurement tool based on replication-incompetent VSV (17) (fig. S1A and B). In brief, the protease activity measurement tools comprise replication-incompetent VSV-dsRed variants missing either the viral phosphoprotein (ΔP) or polymerase (ΔL). These viruses are complemented with either an INTRAmolecular-3CL^{pro} tagged phosphoprotein or and INTERmolecular GFP-3CL^{pro}-L fusion protein, respectively. The P:3CL^{pro} or GFP-3CL^{pro}-L proteins are expressed in cells from transfected plasmids. The cells are then infected with the replication incompetent VSV-dsRed variant and treated with inhibitors. An intramolecular 3CL^{pro} tag in combination with VSV-ΔP-dsRed constitutes a gain-of-signal or “on”-assay. An intermolecular 3CL^{pro} tag in combination with VSV-ΔL-dsRed constitutes a loss-of-signal or “off”-assay. We found that the identified single catalytic center mutations indeed conferred partial resistance against nirmatrelvir of the 3CL^{pro} from the Wuhan-1 as well as the Omicron SARS-CoV-2 variant (Omicron signature mutation in 3CL^{pro} P132H) (Fig. 4A to D), which could be further enhanced by introduction of a second mutation in the autocleavage site (F305L, Fig. 4E to H). The mutation Q192R arose in the F305L parental virus. Introducing Q192R alone reduced 3CL^{pro} activity mildly, as we observed by increased values in 3CL^{pro}-On-Q192R at low nirmatrelvir concentrations. Adding F305L as second mutation, thereby restoring the original combination from the double mutant virus, rescued this phenotype (Fig. 4G). A randomly selected combination of catalytic center mutations led to a strong loss in enzymatic activity (fig. S6A and B). We further introduced two mutants (A194S, G138S) into the 3CL^{pro} measurement assays, which also conferred resistance to nirmatrelvir (Fig. 4I and J).

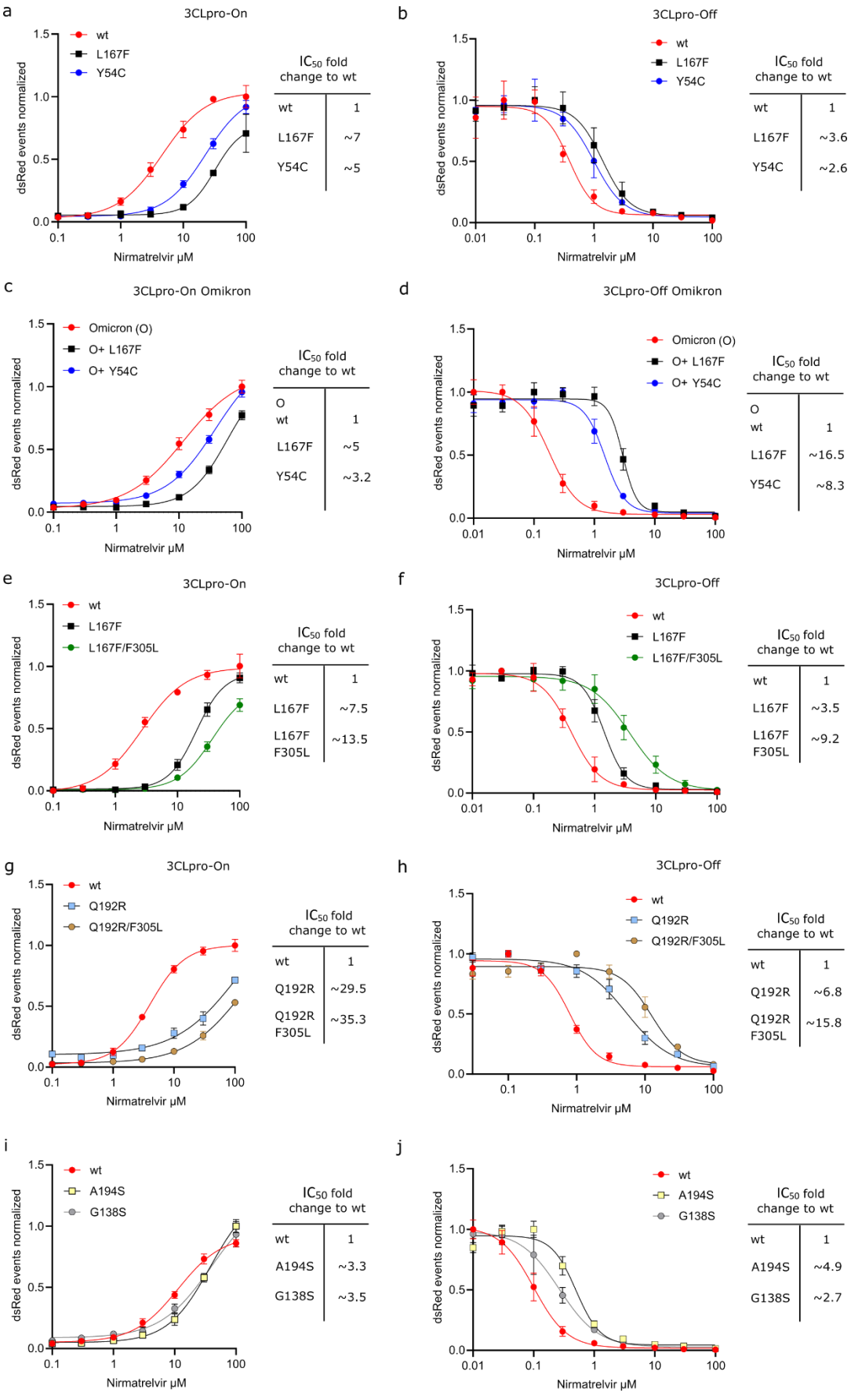

a
3CLpro-On
wt
L167F
Y54C
dsRed events normalized
Nirmatrelvir µM
IC50 fold change to wt
wt 1
L167F ~7
Y54C ~5
b
3CLpro-Off
wt
L167F
Y54C
dsRed events normalized
Nirmatrelvir µM
IC50 fold change to wt
wt 1
L167F ~3.6
Y54C ~2.6
c
3CLpro-On Omikron
Omicron (O)
O+ L167F
O+ Y54C
dsRed events normalized
Nirmatrelvir µM
IC50 fold change to wt
O
wt 1
L167F ~5
Y54C ~3.2
d
3CLpro-Off Omikron
Omicron (O)
O+ L167F
O+ Y54C
dsRed events normalized
Nirmatrelvir µM
IC50 fold change to wt
O
wt 1
L167F ~16.5
Y54C ~8.3
e
3CLpro-On
wt
L167F
L167F/F305L
dsRed events normalized
Nirmatrelvir µM
IC50 fold change to wt
wt 1
L167F ~7.5
L167F F305L ~13.5
f
3CLpro-Off
wt
L167F
L167F/F305L
dsRed events normalized
Nirmatrelvir µM
IC50 fold change to wt
wt 1
L167F ~3.5
L167F F305L ~9.2
g
3CLpro-On
wt
Q192R
Q192R/F305L
dsRed events normalized
Nirmatrelvir µM
IC50 fold change to wt
wt 1
Q192R ~29.5
Q192R F305L ~35.3
h
3CLpro-Off
wt
Q192R
Q192R/F305L
dsRed events normalized
Nirmatrelvir µM
IC50 fold change to wt
wt 1
Q192R ~6.8
Q192R F305L ~15.8
i
wt
A194S
G138S
dsRed events normalized
Nirmatrelvir µM
IC50 fold change to wt
wt 1
A194S ~3.3
G138S ~3.5
j
wt
A194S
G138S
dsRed events normalized
Nirmatrelvir µM
IC50 fold change to wt
wt 1
A194S ~4.9
G138S ~2.7

Figure 4. Re-introduction of individual or dual 3CL^{pro} mutations confirms their resistance phenotype. A graphic representation of the 3CL^{pro}-on and 3CL^{pro}-off system used to measure the inhibitory activity of the protease inhibitor against the different 3CL^{pro} mutants can be found in fig S3. (A) Gain-of-signal assay results are shown for single catalytic site mutations Y54C and L167F with nirmatrelvir. Data are presented as the standard deviation (SD) of n = 3 biologically independent replicates per condition. (B) Loss-of-signal assay results are shown for single catalytic site mutations Y54C and L167F with nirmatrelvir. Data are presented as the SD of n = 4 bio-logically independent replicates per condition. (C) Gain- of-signal assay results are shown for catalytic site mutations Y54C and L167F in combination with the Omicron 3CL^{pro} signature mutation P132H. Data are presented as the SD of n = 4 biologically independent replicates per condition. (D) Loss-of-signal assay results are shown for single catalytic site mutations Y54C and L167F in combination with the Omicron 3CL^{pro} signature mutation P132H. Data are presented as the SD of n = 4 biologically independent replicates per condition. (E) Gain-of-signal assay results are shown for double mutant L167F/F305L versus wild-type and single mutant L167F. Data are presented as the SD of n = 4 biologically independent replicates per condition. (F) Loss-of-signal assay results are shown for double mutant L167F/F305L versus wild-type and single mutant L167F. Data are presented as the SD of n = 4 biologically independent replicates per condition. (G) Gain-of-signal assay results are shown for double mutant Q192R-F305L versus wild-type and single mutant Q192R. Data are presented as the SD of n = 4 biologically independent replicates per condition. (H) Loss-of-signal assay results are shown for double mutant Q192R/F305L versus wild-type and single mutant Q192R. Data are presented as the SD of n = 4 biologically independent replicates per condition. (I) Gain-of-signal assay results are shown for mutants A194S and G138S versus wild-type. Data are presented as the SD of n = 4 biologically independent replicates per condition. (J) Loss-of-signal assay results are shown for mutants A194S and G138S versus wild-type. Data are presented as the SD of n = 4 biologically independent replicates per condition.

Nirmatrelvir and GC376 react differently to 3CL^{pro} mutants

Comparing GC376 to nirmatrelvir in the 3CL^{pro} Y54C and L167F mutants directly revealed that these mutants react differently to the compounds. Y54C confers a similar resistance to GC376 as to nirmatrelvir (Fig. 5A). GC376 and nirmatrelvir interact similarly with the residue Y54, whereas L167 and Q192 are distant to GC376 and close to nirmatrelvir (within 5 Å) (Fig. 5B). L167F and Q192R appeared to affect the activity of GC376 less than nirmatrelvir (Fig. 5C and D, tables S4 and S5). Nirmatrelvir IC50 values were especially high in the 3CL^{pro}-On construct. We sought to improve the assays sensitivity by changing the read-out method from a FluoroSpot to a flow cytometry-based readout. With this approach, we could decrease the IC50 of the wild-type 3CL^{pro}-On to 0.91 µM of nirmatrelvir (fig. S6C, table S6).

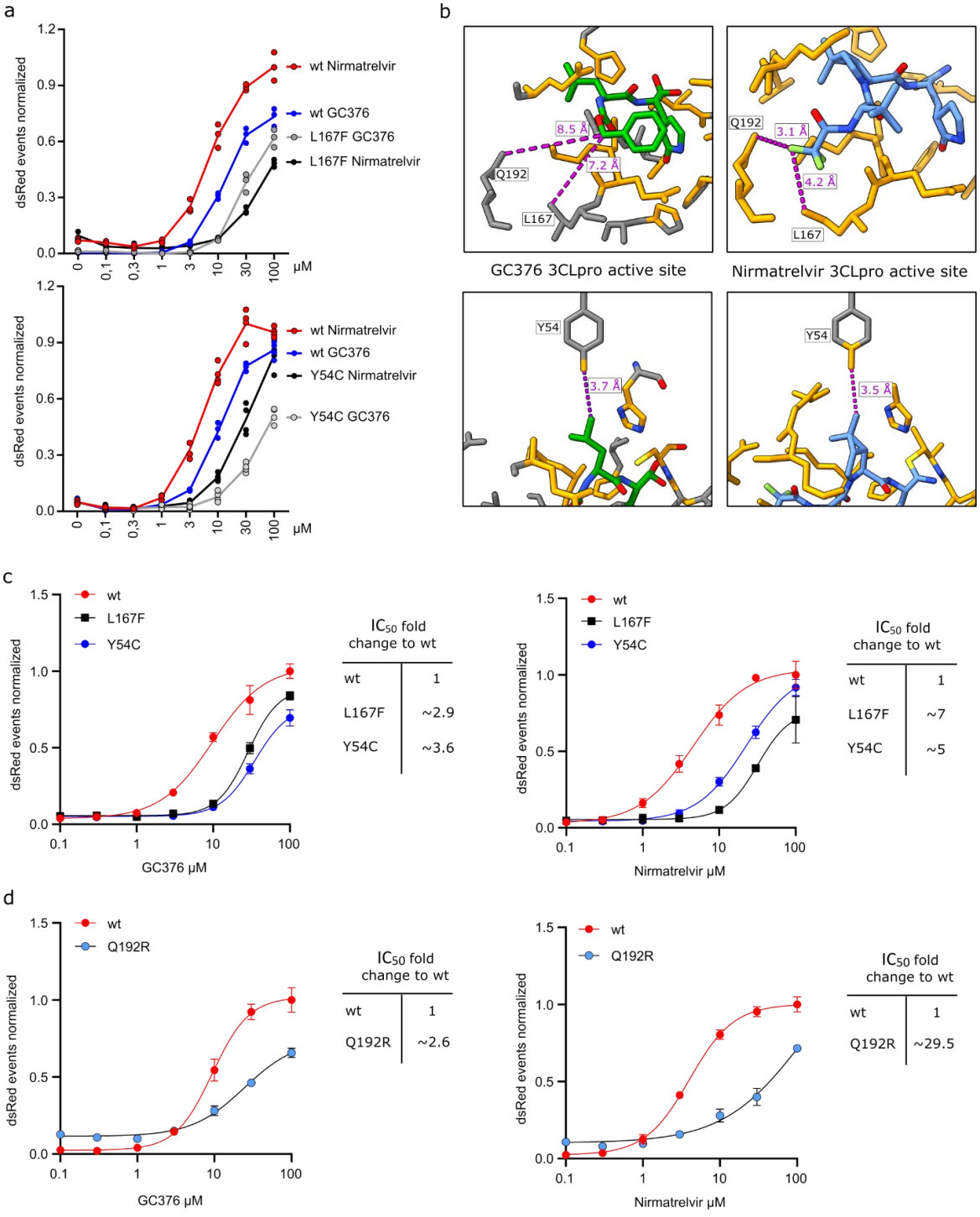


Figure 5. Nirmatrelvir and GC376 react differently to mutants. (A) Gain-of-signal assay results are shown for single mutants Y54C and L167F versus wild-type tested with GC376 and nirmatrelvir (Y54C: n = 4, L167F: n = 3 biologically independent replicates per condition). (B) GC376 (PDB: 7k0g) and nirmatrelvir (PDB: 7vh8) 3CL^{pro} crystal structures are shown with GC376 in green (and colored by heteroatom) and nirmatrelvir in light blue (and colored by heteroatom) and proximal residues in orange (within zone of 5 Å). Compound to residue distances are shown with dotted purple lines. (C) Fitting of gain-of-signal assay results are shown for single mutants Y54C and L167F versus wild-type tested with GC376 and nirmatrelvir. (D) Fitting of gain-

of-signal assay results are shown for single mutant Q192R versus wild-type tested with GC376 and nirmatrelvir. Data in (C and D) are presented as the SD of n = 4 biologically independent replicates per condition.

Confirmation of resistance mutations in a second cell-based assay system, biochemical assay, and with recombinant SARS-CoV-2

The resistance phenotype of L167F observed with gain- and loss-of- signal assays was confirmed using another recently published cellular system[27]. In this complementary assay, a polyprotein of Src, 3CL^{pro} with N-and C-terminal autocleavage sites, HIV Tat, and luciferase was used to repress transcription when 3CL^{pro} was active (fig. S7). Bona fide chemical inhibitors blocked 3CL^{pro} activity and restored luciferase signal in a dose-dependent manner (Fig. 6A to C). Similar to the results described above, L167F was more resistant to nirmatrelvir than GC376 (Fig. 6A and C, table S7). Furthermore, this mutant was most resistant to ensitrelvir, a recently developed compound in clinical trials in both the Src- 3CL^{pro}-Tat-Luc (Fig. 6B, table S7)[28] and our assay (fig. S8, table S8). For further confirmation of resistance phenotypes, we purified recombinant enzymes (fig. S9A to C). We tested catalytic activity with a substrate dose-response kinetic experiment with purified wild-type 3CL^{pro} versus mutants Y54C, L167F, and Q192R with the substrate Ac-Abu-Tle-Leu-Gln↓MCA releasing the fluorogenic molecule 7-amino-4-methylcoumarin (AMC) (fig. S10A). The ratio between the catalytic constant or turnover rate (kcat) and the Michaelis-Menten constant (Km), displayed as kcat/Km, showed that some of the mutants partially lose catalytic activity, most notable Q192R, which was in line with the cellular assays (Fig. 6D, table S9). For further resistance confirmation we applied a biochemical fluorescence resonance energy transfer (FRET) assay, which uses a quencher (DABCYL) and a fluorogenic substance (EDANS), which are connected by a peptide (KTSAVLQSGFRKME) that is recognized and cleaved by 3CL^{pro} (fig. S10B). Upon cleavage, fluorescence of EDANS increases. All three mutant 3CL^{pro} enzymes were more resistant to nirmatrelvir than the wild-type 3CL^{pro} (Fig. 6E, table S10). Lastly, we confirmed our findings in recombinant SARS-CoV-2 viruses expressing a reporter gene (fig. S10C)[29,30]. The recombinant SARS-CoV-2 variant expressing mCherry used for mutagenesis, aside from its transgene, was sequence identic to the Wuhan-1 variant[29,30] (Data file S1). Viruses carrying L167F alone and in combination with F305L were able to replicate but replicated

slower than wt virus and produced smaller plaques (Fig. 6F and G). The mutations introduced into recombinant SARS-CoV-2-mCherry had been found in clinical samples prior to our study (Data file S2). As expected, both L167F single and L167F/F305L double mutants were more resistant to nirmatrelvir than the wild-type (Fig. 6H, table S11).

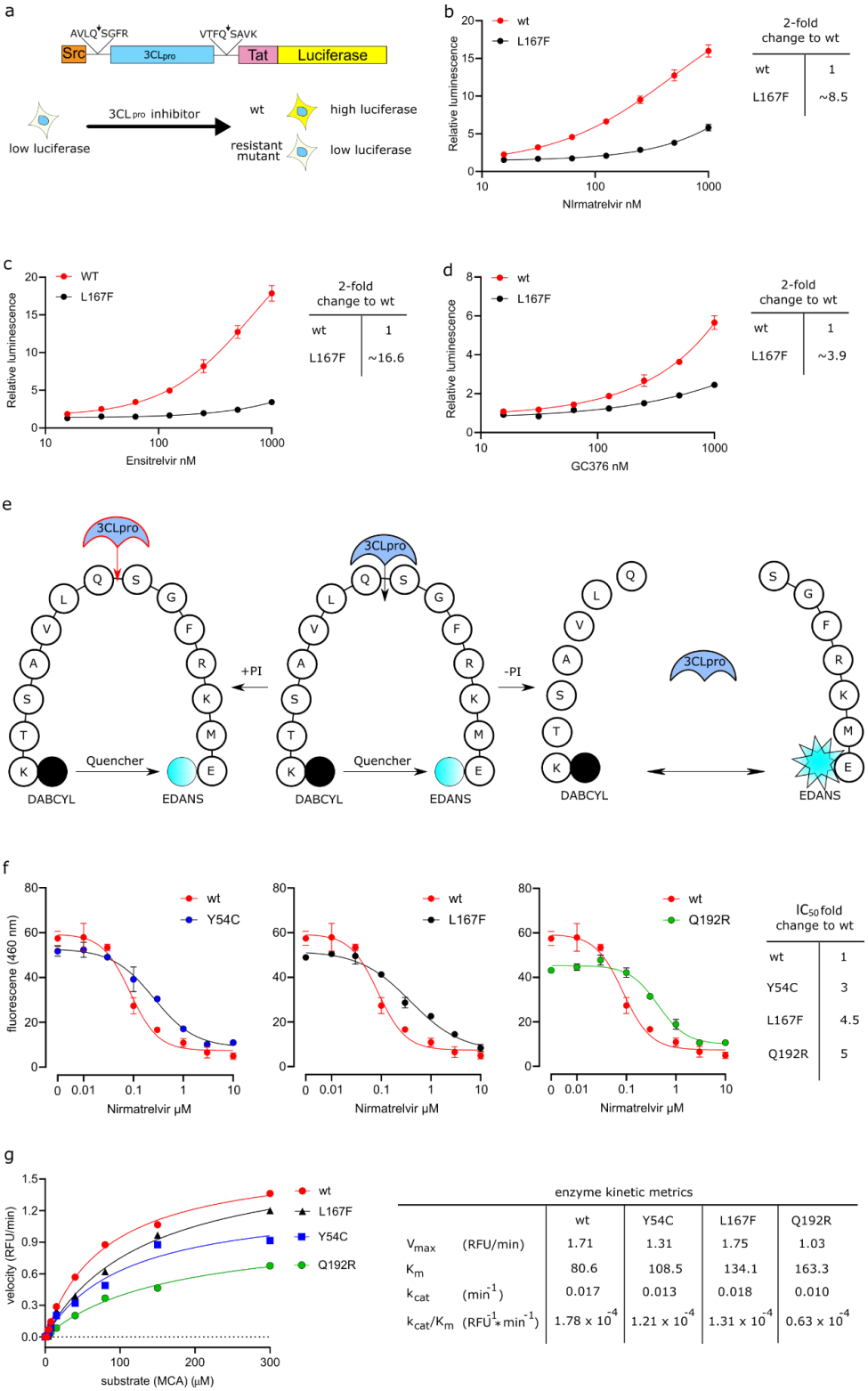


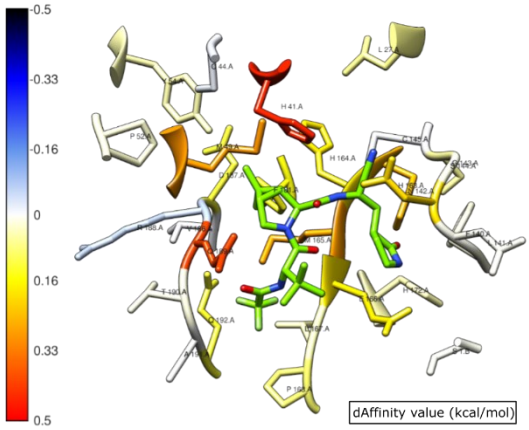
Figure 6. Cross-testing mutants and validation of enzyme kinetics. a: Scheme of cellular gain-of-signal assay based on Src-3CL^{pro}-Tat-Luc polyprotein used for cross-validation. b: Cross-validation of wild-type (wt) Src-3CL^{pro}-Tat-eGFP and L167F mutant with nirmatrelvir, ensitrelvir (c) and GC376 (d). e: Scheme of cellular biochemical assay from BPS Bioscience used for cross-validation. In presence of an appropriate protease inhibitor (+PI), 3CL^{pro} cannot cleave the substrate and fluorescence is low. Without inhibitor (-PI), 3CL^{pro} cleaves the substrate, quencher (DABCYL) and fluorogen (EDANS) are separated, fluorescence is high. f: Nirmatrelvir dose responses of wild-type (wt) vs. mutant 3CL^{pro} variants in biochemical assay and IC_{50} fold changes. g: Enzyme kinetics and metrics of wt 3CL^{pro} and mutants with the substrate Ac-Abu-Tle-Leu-Gln↓MCA releasing the fluorogenic molecule MCA.

Structural modelling of mutant 3CL^{pro} variants

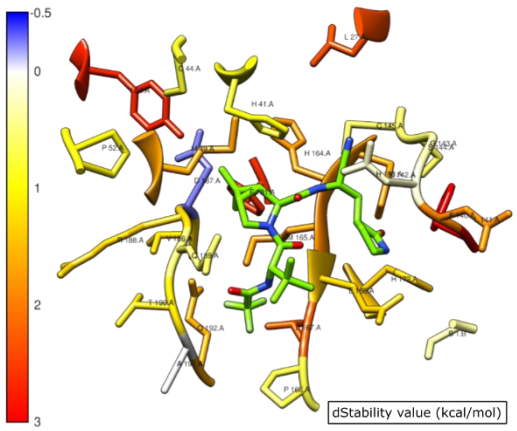
To explore potential mechanisms of resistance, we performed molecular modelling with an in-silico alanine mutation scanning as well as resistance mutation scanning with Molecular Operating Environment (MOE) suite[16] and the Robetta service[15]. MOE modelling was based on the PDB structure 7rfw[31], and Robetta modelling was based on 7vh8[32] (32), both of which are 3CL^{pro} structures with high nirmatrelvir occupancy. Experimental alanine scanning[33,34] as well as in silico alanine screening[35,36] is routinely utilized to evaluate the impact of single amino acid mutations on protein structure, and the models can provide plausible explanations for the structural basis of nirmatrelvir resistance. Fig. 7A shows that the most important losses of binding affinity are primarily related to mutation of residues whose sidechains directly contact the ligand, such as H41, M49, N142, H163, M165 and Q189, and secondarily to other residues lining the binding site like Y54, H164, E166, P168, D187, and Q192. Residues with hydrophobic sidechains, such as L27, Y54, F140, L167 and F181 seem to have a pivotal role in the structural integrity of the binding site (Fig. 7B), despite having a negligible impact on the variation of binding affinity. The primary effect of Y54C in these models is the disruption of a stabilizing inter-protein hydrogen bond to the backbone oxygen of D187 and disruption of additional weak but stabilizing interactions with surrounding hydrophobic residues such as a π-charge interaction between the phenyl ring of Y54 and the guanidinium group of R40. No major direct interaction to the 3.5 Å distant C20 methyl group of nirmatrelvir exists (Fig. 7C, fig. S11A). However, loss of the critical hydrogen bond between loop region 43 to 55 and the adjacent loop around D187 allows for a structural rearrangement destabilizing the distal part of the binding site, likely increasing the inherent plasticity of this protein region.

Residue G138 lies in a solvent accessible loop, with backbone torsion angles in the β sheet region. Replacing it with a polar serine (G138S) while maintaining the same backbone conformation led to the Cβ of S138 pointing into the protein interior, and all of the preferred rotamers led to unfavorable interactions or required a re-arrangement of the affected region. Formation of new hydrogen bonds, for example with a backbone hydrogen of F140 and the sulfur of C128 (Fig. 7D) likely led to a rearrangement of the S1 sub-pocket, which is responsible for hosting the terminal carboxamide moiety that mimics the P1 glutamine in natural peptide substrates. Supervised Molecular Dynamics simulations of the nirmatrelvir-3CL^{pro} recognition process revealed how L141 is one of the first residues that is contacted during the approach of nirmatrelvir into the binding site[37]. In the L167F mutant, the larger sidechain of phenylalanine cannot be accommodated without a structural rearrangement, which likely leads to repulsive interaction between the trifluoromethyl (CF3) moiety of nirmatrelvir and weakening its interactions with other proximal residues such as N142, which are thought to play a pivotal role in maneuvering the ligand entrance in the catalytic pocket[37]. As anticipated in the alanine scan, the L167F mutation seems to have an indirect effect on the binding affinity by alteration of the β sheet that constitutes the lower portion of the binding site, where a set of hydrogen bonds are established between nirmatrelvir and the backbone of both H164 and E166 (fig. S11B). A similar distortion of the binding pocket by the bulkier phenylalanine has also been described recently[10]. Finally, polar Q192 stabilizes a solvent exposed loop participating in hydrogen bonds to backbone oxygen and nitrogen of V186, backbone oxygen of R188, and a stabilizing contact to the CF3 group of nirmatrelvir. Replacement with positively charged R192 disrupted this network, which likely results in a structural rearrangement and altered binding to nirmatrelvir (Fig. 7E). The slight increase in binding affinity of Q192R concurrent with protease destabilization (Fig. 7A and B) could be explained by recontouring of the subpocket hosting the negatively polarized CF3 moiety of nirmatrelvir and interacting with positively charged R192 (Fig. 7E, fig. S11C). Despite the predicted, marginally more favorable interaction with nirmatrelvir, an overall unfavorable effect of the mutation could still be possible due to altered sequestration of nirmatrelvir and the destabilization of the loop region lining the S2 and S4 subpocket of the catalytic site, where important residues such as Q189 are located.

a



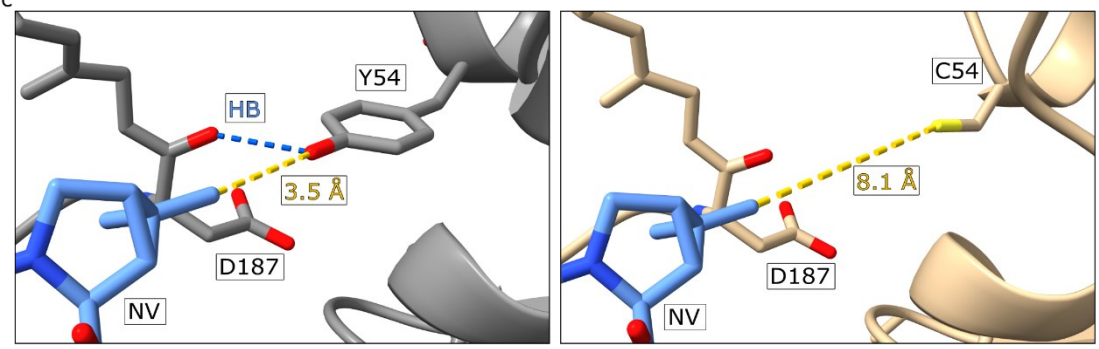
b



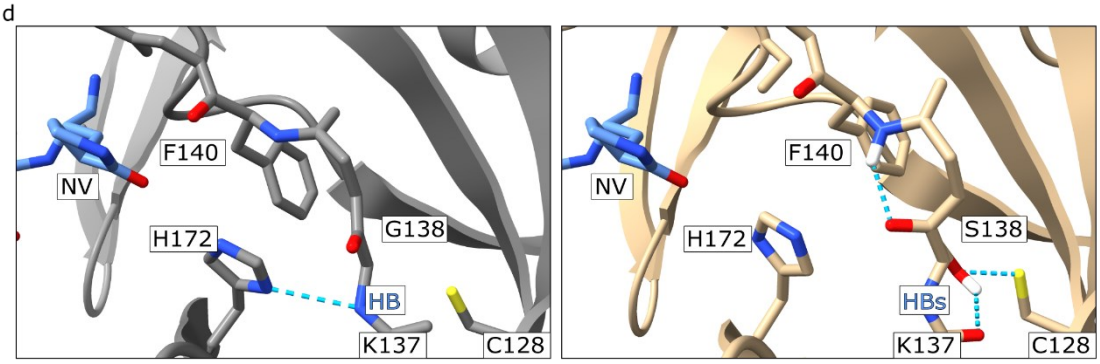
Mutation	Y54C	G138S	L141F	L167F	Q192R
dAffinity [mut-WT] (kcal/mol)	0	0.01	0.18	0.13	-0.12

Mutation	Y54C	G138S	L141F	L167F	Q192R
dStability [mut-WT] (kcal/mol)	2.82	1.06	0.58	0.81	0.53

c



d



e

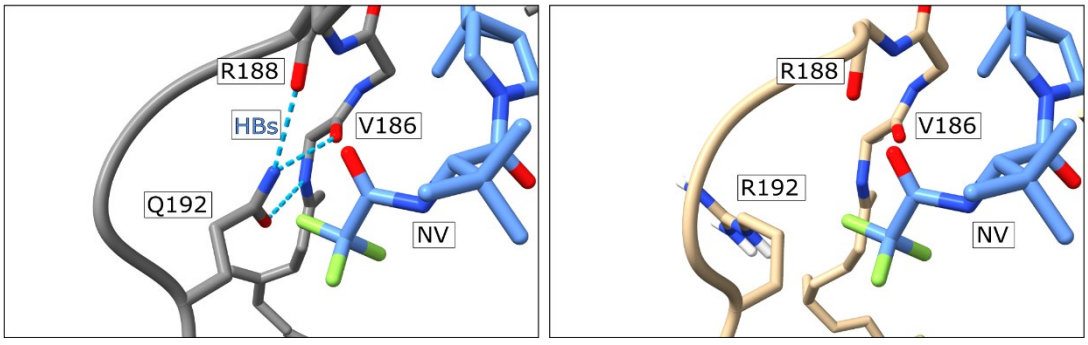


Figure 7. Structural modelling of mutant 3CL^{pro} variants. (A) Colorimetric mapping of the dAffinity value (kcal/mol) by virtual alanine scanning with MOE suite. Residues within 5 Å of the nirmatrelvir position are displayed. Colors range from blue (negative values, indicating increased protein-ligand affinity) to red (positive values, indicating decreased protein-ligand affinity). The nirmatrelvir (NV) structure is shown in light blue. (B) Colorimetric mapping of the dStability value (kcal/mol), computed as above for (A). Colors range from blue (negative values, indicating increased in the protein stability) to red (positive values, indicating decreased protein stability. (C) The catalytic center of 3CL^{pro} from PDB structure 7vh8 is shown with nirmatrelvir bound. Y54 (left) forms a strong hydrogen bond (HB, highlighted with a blue dashed line) with D187, whereas nirmatrelvir is at a distance of 3.5 Å (yellow dashed line). The exchange of Y54 with C (right) leads to a loss of the hydrogen bond to D187 and makes room in the nirmatrelvir binding pocket due to the smaller side- chain of cysteine versus tyrosine. (D) G138 (left) contacts H172 with a hydrogen bond. S138 (right) forms several new hydrogen bonds with the back-bone hydrogen of F140, backbone oxygen of K137 and the sulfur of C128. (E) Q192 (left) forms hydro-gen bonds with the oxygen and nitrogen of V186, the oxygen of R188, and stabilizes the polar contact to the CF3 group of nirmatrelvir. R192 (right) disrupts this hydrogen bond network; subsequent rearrangement could form additional interactions with the CF3 group.

Discussion

In our study, we selected mutations in the main protease 3CL^{pro} of SARS-CoV-2 against the protease inhibitor nirmatrelvir with a non- gain-of-function system based on VSV. The selected mutations were confirmed in two cellular assays and in one biochemical assay, along with confirmation using recombinant SARS-CoV-2. For the catalytic site mutations, a resistance mechanism was postulated based on mapping the mutations onto the co-crystal structure of 3CL^{pro}-nirmatrelvir and generating mutant models with Robetta[38]. In previous initial resistance studies leading to emergency use authorization of Paxlovid (nirmatrelvir/ritonavir), the 3CL^{pro} of a related coronavirus, MHV, was used to select for resistance mutations. The 3CL^{pro} of SARS-CoV-2 and MHV share 50% sequence identity. In this study, we compared the activity of 3CL^{pro} of SARS-CoV-2 and MHV and found that MHV 3CL^{pro} responded only mildly to nirmatrelvir in our gain-of-signal assay[17]. Although the structures of SARS-CoV-2 and MHV 3CL^{pro} are conserved, we propose that the low amino acid sequence identity alters the binding pocket affinity to nirmatrelvir sufficiently to reduce the sensitivity against the inhibitor. Key corresponding residues of the binding pocket (within 5 Å or less) are different, namely H164 - Q, M165 - L, P168 - S, V186 - R, R188 - A, T190 - V and A191 - V. Furthermore, amino acid changes that occurred in our

selection experiments, Y126F and F305L are already present in the MHV 3CL^{pro} sequence. Taken together, we argue that MHV 3CL^{pro} was not an optimal proxy for resistance studies. Recently, chimeric VSV variants with SARS-CoV-2 spike were used to predict spike protein immune escape mutations by selecting against neutralizing serum[39–41]. The fast occurrence of mutations was facilitated in those studies by the high error rate of the VSV polymerase[20,21]. In a similar approach, we exploited this high error rate in a recombinant VSV expressing 3CL^{pro} to select for 3CL^{pro} mutations that confer resistance against protease inhibitors. The 3CL^{pro} was used to replace the function of an intergenic region between the viral glycoprotein (G) and the polymerase (L). The intergenic regions of VSV are responsible for separate gene expression, which in other viruses is accomplished by a polyprotein and proteases. Although this polyprotein of VSV-G-3CL^{pro}-L is only a surrogate to the one in SARS-CoV-2, the cognate cleavage sites, the requirement for dimerization of the protease for proper function[18,19] and the context of a replicating virus in the cell make this approach an attractive proxy. Initially, we selected the 3CL^{pro} mutant F305L using GC376, which showed reduced sensitivity to GC376, as well as to nirmatrelvir. This mutation lies in the 3CL^{pro} cleavage site that flanks 3CL^{pro} at its C terminus. Interestingly, the LQ motif found in the mutated site is indeed known to be preferred over FQ as a target motif for 3CL^{pro}[42–44], which may explain the reduced sensitivity of the F305L mutant to the protease inhibitors. We then selected both wild-type and the F305L mutant against nirmatrelvir. We also used F305L as parental virus because we found that this variant existed already in regional outbreaks (mainly in England), underlining the viability of this mutation and its potential replicative advantage. These clusters were mainly of the Delta subvariant AY.4. Delta was replaced gradually by Omicron, which may have ended the spread of the Delta F305L. Nevertheless, we also found combinations of the Omicron signature mutation P132H with F305L. We were therefore interested in finding potential combinations of F305L with further protease inhibitor resistance mutations, assuming that the combination would show a higher degree of resistance than single mutations, which we did observe. Mutations from both wild-type and F305L were selected that ultimately allow the mutants to escape the inhibitor. Resistance phenotypes were confirmed by dose response experiments and re-introduction of mutations into recently developed protease activity measurement systems[17] as well as alternative methods such

as biochemical[45] and cellular assays[27]. We collected a total of 39 unique mutations, of which Y54C, L167F, N203D and D216Y occurred twice independently. F305L was selected for with both GC376 and nirmatrelvir. Six out of 39 occurred in the catalytic site, two near the catalytic site, seven at the dimer interface, three in the autocleavage sites and 21 in the rest of the $3CL^{pro}$ sequence (which we called "allosteric" mutations). First, we confirmed catalytic site mutants (Y54C, G138S, L167F and Q192R), where the resistance mechanism is likely straight forward: the steric disturbance of nirmatrelvir binding. Then, we tested the near-catalytic site substitution A194S, which is more prevalent than the previous catalytic site mutants in virus isolates. In GISAID, this mutation can be found in over 800 sequence depositions in the variants of concern Alpha, Gamma, Delta, Lambda, and Omicron. Changes of the residue A194 in general are frequent with over 3000 entries. Although it is not known if this mutant was selected for by the use of nirmatrelvir in patients, the fact that it is a resistant mutant and prevalent in virus sequences makes it a variant worth tracking. We further combined L167F and Q192R with the autocleavage site mutation F305L, which further increased the resistance. The combination of Y54C and L167F with the Omicron signature mutation P132H also conferred increased resistance, highlighting the potential relevance of these mutations for the Omicron variant. The substitution F305L was described as resistance mutation in this study. An adjacent mutant, T304I, was found in nirmatrelvir selection experiments with authentic SARS-CoV-2[9] and the suggested mechanism was auto-cleavage site optimization. Given F305L is also likely an autocleavage site optimizing mutant, we did not test it further in a biochemical assay, since such assays use mature protease in which autoprocessing does not play a role. Lacking an appropriate method, we therefore did not investigate the mechanism of F305L. Nevertheless, such mutants merit further study in assays that can elucidate the mechanism of action. One technical particularity in the $3CL^{pro}$-On construct is that the nirmatrelvir IC_{50} values are higher than generally reported in the literature. However, IC_{50} values are generally higher in cell-based assays than in biochemical assays as we described previously[17]. In brief, in the excess of an inhibitor and constant renewal of protease fusion protein, signals are expected to plateau later than in a biochemical assay with a fixed amount of enzyme. Furthermore, the screening method used in this study to assess mutants was originally developed as high-throughput screening tool for $3CL^{pro}$ inhibitors

using a FluoroSpot reader that allows fast sampling[17]. We im-proved the assays' sensitivity by changing the read-out method from FluoroSpot to flow cytometry-based sampling. Flow cytometry sampling is more sensitive, but also more time consuming. Flow cytometry read-outs captured milder degrees of inhibition and resulted in a more gradual signal increase; therefore, this resulted in lower IC_{50} values in $3CL^{pro}$-On assay (0.91 µM of nirmatrelvir, which is closer to the published range of 74.5 (66.5 to 83.4) nM)[31]. We cross-validated several of our mutants in different assays. We confirmed the resistance data of L167F with a previously published cellular assay[27] and the mutants Y54C, L167F and Q192R with a biochemical assay[46]. We showed also in a biochemical assay that the kinetic metrics of the mutants Y54C, L167F, and Q192R are attenuated to varying degrees. However, the VSV-chimeric viruses containing resistant 3CLpro showed little fitness loss. In three recent preprints, L167F as well as various mutants at Q192 were identified to be resistance mutations in authentic SARS-CoV-2[9,10,12]. Lastly, we confirmed the viability and resistance of the single mutant L167F and in combination with F305L in a previously published recombinant SARS-CoV-2 expressing a reporter gene[29,30], finally confirming the validity of our mutation prediction tool based on VSV as well as the resistance mutations identified. Genetically modifying highly pathogenic viruses such as SARS-CoV-2 can be considered as so-called gain-of-function experiments if the recombinant virus is more apt to cause disease, or if treatment is made more difficult than the wt variant. We therefore applied several safety measures such as using a virus for mutagenesis that was sequence identic to the Wuhan-1 variant. Therefore, compared to currently circulating viruses, it has not undergone extensive evolution and if set free, would unlikely be able to compete with current Omicron variants. Importantly, other previously described antivirals approved for use in humans have been shown to inhibit viral replication of this Wuhan-1 strain. Moreover, current vaccines used in humans to protect against SARS-CoV-2 have been developed based on the sequence of the Spike glycoprotein of this, or a similar, Wuhan-1 strain. Thus, neutralizing antibody responses induced by these vaccines will be able to protect against these recombinant viruses. Secondly, the plaques the mutant recombinant viruses formed in Vero E6 cells were found to be smaller than that of wild-type recombinant virus and replication kinetics also indicate mild attenuation. Third, the mutations introduced into recombinant SARS-CoV-2-mCherry had been found in clinical

samples already. In this study, we identified several mutations such as Y54C, G138S, L167F, Q192R, A194S and F305L in the SARS-CoV-2 3CL^{pro} that confer resistance to the 3CL^{pro} inhibitors nirmatrelvir and GC376. To understand these mutations in light of the Omicron variant, we combined two of our most intensively studied mutations, Y54C and L167F with the Omicron 3CL^{pro} signature mutation P132H. These results showed that the mutations are functional, thereby confirming their potential relevance in this context of the Omicron variant. Complementary structure modelling approaches based on Robetta and the Molecular Operating Environment (MOE) reveal potential effects of the catalytic site mutants Y54C, G138S, L141F, L167F and Q192R. Y54C and G138S seem not to directly affect nirmatrelvir binding, but may lead to a restructuring of the catalytic site, thereby indirectly affecting the binding site. L141F is likely impacting the early sampling of the catalytic pocket by nirmatrelvir. L167F may distort the distal region of the binding pocket. Lastly, Q192R could strengthen a polar interaction with nirmatrelvir, which may alter sequestration of the compound in an unfavorable position concurrent with destabilization of a loop containing important nirmatrelvir-catalytic site interaction partners such as Q189. Our study has limitations. The mutations generated in VSV occurred in an artificial polyprotein, which, like pp1a or pp1ab, comprises precursors for large protein subunits and requires 3CLpro dimerization for autocleavage. Nevertheless, the polyprotein structure is different, which could result in mutations not relevant in authentic SARS-CoV-2. Along the same line, in this artificial polyprotein, only autocleavage or cis-cleavage occurs, whereas in SARS-CoV-2, the mature 3CL^{pro} additionally cleaves distant or trans-cleavage sites. It has been shown that autocleavage of 3CL^{pro} in coronaviruses is a stepwise process with distinct N- and C-terminal autocleavage binding pocket confirmations[47,48], where the C-terminal autocleavage occurs after N-terminal auto-cleavage and might resemble a matured structure as in the trans- cleavage confirmation. Even so, this system could, in theory, disregard trans-cleavage specific mutants, if such exist. Finally, we did not elucidate the exact mode of resistance of the different mutants described in this study. Although we modelled catalytic site mutations and described a plausible mechanism for autocleavage site mutants, solving crystal structures was beyond the scope of this work and remains for future studies. In conclusion, our findings argue for a highly selective application of protease inhibitors to patients at increased risk of severe

disease, as extensive, unselective use is expected to rapidly lead to emergence of drug resistance. Furthermore, the combination of different drugs is a proven strategy to avoid resistance mutations, as has been shown for HIV[3] and HCV[2] therapy. As more compounds became available, combinations including classes of inhibitors targeting distinct viral functions, such as protease and polymerase inhibitors, may be an effective strategy. However, as we observed in this study, $3CL^{pro}$ mutants can react differently to specific compounds. Therefore, even the combination of different protease inhibitors could lower the risk of viral escape.

Materials and Methods

Study design

The overall rationale of the study was to develop a mutation selection tool based on VSV and to describe mutants as proof-of-concept for that tool. The study was performed on cell lines and in-silico, and no animal husbandry or human participants were involved. Human and monkey cell lines with replicating BSL-1, 2 and − 3 viruses were treated with protease inhibitors to observe resistance phenotypes in appropriate facilities. Viral titers were determined using $TCID_{50}$ and plaque assays. Measurement readouts were fluorescence-based, detected by flow cytometry, ELISpot and multi-well readers. Autofluorescent fibers were excluded automatically and manually from spot counting in the ELISpot readout. Experiments were neither blinded nor randomly distributed to experimenters. We chose sample sizes empirically based on experience from former studies. At least two and up to four biologically independent replicates were performed per condition. Biologically independent meant distinct wells with the same condition, not multiple measurement of the same wells (technical replicates). Resistance phenotypes were reproduced at least twice, usually more often and in different combinations (comparing single mutants to each other and the wild-type or wild-type to single and double mutants). Representative measurements were chosen to compile graphs and figures.

Cloning strategies

The chimeric VSV variant with 3CL^{pro} instead of the intergenic region between G and L was cloned by Gibson assembly (New England Biolabs, NEB)[49]. A VSV-G plasmid[50] was digested with KpnI and HpaI (NEB), removing a C-terminal part of G, the intergenic region and a small N-terminal part of L. Insert fragments were generated as follows. Missing C-terminal G with an additional overhang to the N-terminal cleavage site of 3CL^{pro} was amplified with primers 33n-before-KpnI-for and G-cut1-rev. 3CL^{pro} with its N- and C-terminal cleavage sites and a C-terminal overhang to L was amplified from Wuhan-1 (NCBI Reference Sequence: NC_045512.2) cDNA with primers cut1-for and cut2-L-rev. The N-terminal missing L sequence was amplified with primers cut2- L-for and 33n-after-HpaI-rev. For subsequent Gibson assembly, the fragments were ligated in a fusion PCR using the outer primers 33n-before-KpnI-for and 33n-after-HpaI-rev with all three fragments as templates. The cloning primers for VSV vectors are shown in table S12 and the annotated sequence is shown in data file S3.

Name	**Sequence (5'-3' direction)**
33n-before-KpnI-for	GAACCGGTCCTGCTTTCACC
G-cut1-rev	CATTTTTCTAAAACCACTCTGCAAAACAGCTGAGGTGATCTTTCCAAGTCGGTTC
cut1-for	ATCACCTCAGCTGTTTTGCAG
cut2-L-rev	GTCGGTCTCAAAATCGTGGACTTCCATGATTGTTCTTTTCACTGCACTTTG
cut2-L-for	AGTGCAGTGAAAAGAACAATCATGGAAGTCCACGATTTTGAG
33n-after-HpaI-rev	GATGTTGGGATGGGATTGGC

Table 1. Cloning primer for VSV vectors.

3CL^{pro}-Off and -On point mutants were generated by mutagenic Gibson assembly on parental plasmids (GenBank accession codes: 3CLpro-Off: 25684003; 3CLpro-On: 2568399). For 3CL^{pro}-Off mutants, a lentiviral expression plasmid expressing VSV L (identical sequence as blasticidin 3CL^{pro}-Off plasmid without GFP and 3CL^{pro}) was digested with HpaI, which removed the cPPT/CTS and CMV promoter sequences and a small N-terminal part of L. This missing sequence was replaced with the identical sequence from 3CL^{pro}-Off with the addition of the N-terminal 3CL^{pro} sequence up to the respective mutation site with primers blasticidin-for and 3CL^{pro}-*mut-x*-rev, where *mut-x* is the mutation of interest. The C-terminal part of 3CL^{pro} and the small missing fragment of L

were generated by PCRs on parental vectors with primers 3CL^{pro}-*mut-x*-for and 33n-after-HpaI-rev. For 3CL^{pro}-On mutants, a lentiviral hygromycin vector (modified from Addgene pLenti CMVie-IRES-BlastR accession: #119863) was digested with NheI and PacI. N-terminal 3CL^{pro} insert fragments with vector overhangs were generated with hygro-P-for and 3CL^{pro}-*mut-x*-rev. C-terminal 3CL^{pro} insert fragments with vector overhangs were generated with 3CL^{pro}-*mut-x*-for and P- hygro-rev. Double mutants were cloned by repeating the site directed mutagenesis with a second primer pair in combination with Gibson assembly on an already mutant-bearing plasmid. Cloning primers for 3CL^{pro}-Off and -On mutant variants are shown in table S13.

Name	**Sequence (5'-3' direction)**
blasticidin-for	CATTCGATTAGTGAACGGATCTC
3CL^{pro}-L167F-for	GCACCATATGGAATTTCCAACTG
3CL^{pro}-L167F-rev	CATGAACTCCAGTTGGAAATTCC
3CL^{pro}-Y54C-for	CATGCTTAACCCTAATTGTGAAGATTTACTC
3CL^{pro}-Y54C-rev	CTTACGAATGAGTAAATCTTCACAATTAGGG
3CL^{pro}-F305L-for	GCTCAGGTGTTACTCTCCAAAG
3CL^{pro}-F305L-rev	CACTGCACTTTGGAGAGTAACAC
3CL^{pro}-Q192R-for	GTTGACAGGCAAACAGCACGAGCAGCTG
3CL^{pro}-Q192R-rev	GTCCGTACCAGCTGCTCGTGCTGTTTG
Omicron-for	CAATGTGCTATGAGGCACAATTTCAC
Omicron-rev	CTTAATAGTGAAATTGTGCCTCATAGC
A194S-for	CAAACAGCACAAGCATCTGGTACG
A194S-rev	GTGTCCGTACCAGATGCTTGTGC
G138S-for	CCCAATTTCACTATTAAGAGTTCATTCCTTAATG
G138S-rev	ACCACATGAACCATTAAGGAATGAACTCTTAATA
hygro-P-for	CTGTTTTGACCTCCATAGAAGATTCTAGAGCTAGCATGGATAATCTCACAAAAGTTC
P-hygro-rev	GAGGGAGAGGGGCGGATCCCCTTAATTAACTACAGAGAATATTTGACTCTCGC

Table 2. Cloning primer for 3CLpro-Off and -On mutant variants.

Cell lines

BHK-21 cells (American Type Culture Collection, ATCC) were cultured in Glasgow Minimum Essential Medium (GMEM) (Lonza) supplemented with 10% fetal calf serum (FCS), 5% tryptose phosphate broth, and 100 units/ml penicillin and 0.1 mg/ml

streptomycin (P/S) (Gibco). 293 T cells (293tsA1609neo, ATCC), and 293- VSV (293 expressing N, P-GFP and L of VSV)[51] were cultured in Dulbecco's Modified Eagle Medium (DMEM) supplemented with 10% FCS, P/S, 2% glutamine, 1x sodium pyruvate and 1x non-essential amino acids (Gibco). Vero E6 (ATCC CRL-1586) were cultured in DMEM supplemented with 5% FCS (VWR) and 1% penicillin−streptomycin−glutamine (PSG) solution (Corning). A549-hACE2 (Biomedical Resource Ontology NR- 53821) were grown in DMEM supplemented with 4 mM L-gluta-mine, 4500 mg/l glucose, 1 mM sodium pyruvate, 1500 mg/l sodium bicarbonate, 10% FCS, 1x non-essential amino acid solution (Gibco) and 100 µg/ml blasticidin (Gibco).

Virus recovery

VSV-G-3CL^{pro}-L was rescued in 293 T cells by $CaPO_4$ transfection of whole-genome VSV plasmids together with T7-polymerase, N-, P-, M-, G- and L expression plasmids as helper plasmids[52]. Briefly, genome and helper plasmids were transfected into 293 T in the presence of 10 µM chloroquine to avoid lysosomal DNA degradation. After 6 to 16 hours, chloroquine was removed, and cells were cultured until cytopathic effects occurred. M and G proteins were used as helper plasmids; although these proteins are optional in the recovery of VSV, they were chosen here as a precaution to support the rescue of a potentially attenuated virus variant. After the rescue, viruses were passaged on 293-VSV cells and plaque purified twice on BHK-21 cells. ΔP and ΔL VSV variants expressing dsRed were produced on replication supporting 293-VSV cells. VSV-G-3CL^{pro}-L was fully replication competent and produced on BHK-21 cells.

Replication kinetics, $TCID_{50}$ assays, and dose responses

Initial replication kinetics (wild-type versus F305L) were performed as single-step kinetics. 105 BHK-21 cells per well were seeded in 24- well plates one day before infection. Cells were infected in duplicate with an MOI of 5 of VSV 3CL^{pro} wild-type or the F305L variant. One hour after infection, the medium was removed, cells were washed with phosphate-buffered saline (PBS), and fresh medium was added. Supernatant was collected at the indicated time points and stored at −80°C until further analysis. For quantification, $TCID_{50}$ assays were performed as described previously[53]. In short, 100 µl of serial dilutions of

virus were added in octuplicates to 10^3 BHK-21 cells seeded in a 96-well plate. Six days after infection, the $TCID_{50}$ values were read out and titers were calculated according to the Kaerber method[54]. For wild-type versus different mutants replication kinetics, multi-step growth kinetics were performed. 10^5 BHK-21 cells per well were seeded in 24-well plates one day before infection. Cells were infected in duplicates with an MOI of 0.5 of VSV $3CL^{pro}$ wild-type or mutant variants. For initial dose response experiments, 5 x 10^4 BHK-21 cells per well were seeded in 48-well plates one day before infection. Cells were infected in duplicates with a MOI of 0.05 of VSV $3CL^{pro}$ wild-type or mutant variants and indicated concentrations of nirmatrelvir were added to the wells. After 48 hours, supernatants were collected and titrated to determine the $TCID_{50}$. For mutant comparing dose response experiments, 5 x 10^4 BHK- 21 cells per well were seeded in 48-well plates one day before infection. Cells were infected in duplicates with an MOI of 0.05 of VSV $3CL^{pro}$ wild-type or mutant variants and indicated concentrations of nirmatrelvir added to the wells. To prevent initial escape or further mutation in wild-type or already mutation-bearing viruses ("intra-assay mutants"), respectively, supernatants of all viruses were collected after the first mutant (Q192R, F305L) showed a massive cytopathic effect at 100 µM nirmatrelvir (at about 24 hours after infection). Initial dose responses (wild-type versus F305L) were performed as described above, but the supernatant was collected after 48 hours.

Viral RNA isolation and 3CLpro sequencing

VSV-G-$3CL^{pro}$-L RNA was isolated with E.Z.N.A. Viral RNA Kit (Omega Bio-Tek Inc.) or NucleoSpin RNA Virus (Macherey- Nagel GmbH). BHK-21 cells were infected with VSV-G-$3CL^{pro}$-L wild-type and F305L ($3CL^{pro}$) mutant in 96-well plates. Virus-containing supernatants were collected from individual 96-wells and the RNA was purified from the supernatants according to manufacturers' instructions. Then, cDNA was synthesized from isolated viral RNA by RevertAid RT Reverse Transcription Kit (Thermo Fisher Scientific). $3CL^{pro}$ sequence was amplified by PCR with primers (for: CTCAGGTGTTCGAACATCCTCAC and rev: GAT GTTGGGATGGGATTGGC) and sent for sequencing (MicroSynth AG). Obtained sequences were mapped to the $3CL^{pro}$-wt (Wuhan- 1) reference sequence in Geneious Prime 2022.0.2 and examined for mutations.

Mutation selection assay

10^4 BHK-21 cells per well in a 96-well plate were seeded one day before infection with wild-type VSV-G-3CL^{pro}-L or VSV-G- 3CL^{pro}-L-F305L at an MOI of 0.01 and indicated nirmatrelvir doses. Each virus variant occupied 48 wells of the 96-well plate. Wells that displayed cytopathic effect after two days (25 out of 48 from parental wild-type and 17 out of 48 from parental F305L) were further passaged with increasing concentrations of nirmatrelvir with each passage (wild-type: 30, 40 and 50 µM; F305L: 50, 75 and 100 µM). Table S1 indicates at which passage a pure mutant virus could be distinguished by Sanger sequencing, such that only one base-pair peak appeared in the chromatogram instead of a mixture with the parental virus. Only pure mutants are displayed in Fig. 2 and table S1.

Expression and purification of his-tagged 3CL^{pro} and point mutations

Plasmids containing cDNA of SARS-CoV-2 main protease 3CL^{pro} (pMCSG92[55]) and mutants thereof were prepared as described in the following. Plasmids were cloned by site-directed mutagenesis with primers of table S13 on pMCSG92. 100 ng of each plasmid was applied to 50 µl of thawed competent BL21(DE3) TUNER E. coli (Merck) on ice in 1.5 ml tubes. Bacterial suspensions containing the plasmids were flicked and incubated for 30 minutes on ice. Subsequently, bacteria were heated to 42°C for 90 seconds in a thermo-mixer without shaking and put back on ice for 5 minutes. 400 µl of NZCYM Medium produced in-house (NZ amine (Art.-Nr. CP76.1, Roth) 10.0 g, NaCl 5.0 g, casamino acids (Gibco) 1.0 g, yeast extract (Art. Nr. 2363.2, Roth) 5.0 g, $MgSO_4$ x 7 H_2O 2.0 g, ddH_2O to 1 L, adjusted to pH 7.4) was added to each bacterial suspension, and the bacteria were amplified for 1 hour at 37°C in a bacterial shaker in 1.5 ml tubes. Meanwhile, LB agar plates containing selection antibiotics (ampicillin) were prepared and incubated at room temperature. 200 µl of bacterial culture was crossed out on individual plates and incubated overnight at 37°C. Single colony formation was observed the following day. Individual colonies were picked and placed in 5 ml of NZCYM medium supplemented with selection antibiotics, and amplified overnight at 37°C. The next day, overnight cultures were amplified in 1 L of NZCYM medium supplemented with selection antibiotics to an optical density (OD) of 0.2, after which protein expression was induced by applying 1 mM isopropyl-β-D-thiogalactopyranosid (IPTG). After 5 hours bacteria were harvested by

centrifugation and the supernatant discarded. Bacterial pellets were frozen at −20°C for further use. Pellets were suspended in 10 ml of Ni-NTA running buffer (20 mM Tris-HCl, 300 mM NaCl, 10 mM Imidazole, adjusted to pH 7.4) and transferred into 50 ml tubes. Bacteria were lysed using an ultra-sonic probe on ice. The homogenates were centrifuged at 10,000 x g for 10 minutes and the supernatant was filtered using 0.45 µM and 0.22 µM syringe filter units. After preparing a Ni-NTA agarose column (Invitrogen, Ni-NTA Agarose R90115) and washing with 30 ml Ni-NTA running buffer, the filtered homogenate was applied to the column, and the flow-through was collected. The column was again washed with 3 x 10 ml of Ni-NTA running buffer. The His-tagged protein bound to the Ni-NTA resin and was then eluted with 3 x 1 ml of Ni-NTA elution buffer (20 mM Tris-HCl, 300 mM NaCl, 200 mM imidazole, adjusted to pH 7.4). The obtained protein solutions were dialyzed at 4°C overnight against a storage buffer (Tris-HCl 1 mM, NaCl 4 mM, KCl 2.2 mM, TWEEN-20 0.04 vol-%, DTT 3 mM, glycerol 20.2 vol-%, adjusted to pH 8). Eluted protein samples were further purified using size exclusion chromatography (SEC) with fast protein liquid chromatography (FPLC) (ÄKTA Pure FPLC System, Superdex 200 10/300 GL). At each step of the protocol, samples for SDS-PAGE analysis were obtained, and the successful expression of the $3CL^{pro}$ proteins was monitored by SDS-PAGE and Coomassie R staining. The final degree of protein purity was estimated to be greater than 90% based on Coomassie R staining, similar among the different preparations of $3CL^{pro}$ wild-type and mutant forms.

Screening assay with Fluorospot read-out

3 x 10^5 cells were seeded per well in 6-well plates and transfected one day after seeding with $3CL^{pro}$ plasmids using TransIT-PRO (Mirus Bio LLC) and incubated overnight. Then, cells were seeded into a 96-well plate with 2 x 10^4 cells per well in 50 µl complete growth medium. Directly after seeding, compounds and virus (MOI 0.1) were added in 50 µl complete growth medium to wells. After 48 to 72 hours, supernatants were removed, and fluorescent spots counted in a Fluoro/ImmunoSpot counter (CTL Europe GmbH). Longer incubation times of 72 hours increased the overall signal and were chosen in order to achieve a clear signal of the more resistant double mutants, which as expected have a lower signal output in $3CL^{pro}$-On assays. The manufacturer-provided software CTL

switchboard 2.7.2. was used to scan 90% of each well area concentrically to exclude reflection from the well edges, and counts were normalized to the full area. Automatic fiber exclusion was applied while scanning. The excitation wavelength for dsRed was 570 nm, the D_F_R triple band filter was used to collect fluorescence. In addition, manual quality control for residual fibers was performed. To increase comparability between $3CL^{pro}$-On and -Off signals, we normalized dsRed events with the following strategies. In $3CL^{pro}$- On, the highest compound concentrations would not reach the same value due to the different response of each mutant. Therefore, we normalized to the highest mean of the experiment, which was the wild-type signal. In $3CL^{pro}$-Off, untreated wells reached the same signal yield in wild-type and mutants. Therefore, we normalized the signal to each individual highest mean of the construct.

Screening assay with flow cytometry read-out

3 x 10^5 cells were seeded per well in 6-well plates and transfected with $3CL^{pro}$ plasmids using TransIT-PRO (Mirus Bio LLC) and incubated overnight. Then, cells were seeded into a 96-well plate with twenty thousand cells per well in 50 µl. Compound and virus (MOI 0.1) were added in 50 µl to reach desired concentrations. After two days, cells were detached with 0.05% Trypsin-EDTA (Gibco) and transferred to a 96-well round-bottom plate (TPP Techno Plastic Products AG) for automatic sampling by flow cytometry using a BD FACSCanto II. Gates were set to distinguish live and dead cells and to exclude doublets. Singlet cells were divided into dsRed positive and negative based on reference to samples, which were infected, but not treated with inhibitor (17). Samples were analyzed using BD FACSDiva 8.0.1 (BD Biosciences).

Cross validation with orthologous cellular Src-3CLpro-tat- Luc assay

3x10^6 293 T were seeded per well in a 6-well dish. 24 hours later, they were transfected with 2 µg of the wild-type Src-$3CL^{pro}$-Tat- Luc or mutants thereof with TransIT-LT1 (Mirus, catalog number MIR 2304). Four hours after transfection, cells were washed with PBS, trypsinized, resuspended in medium and counted. 2x10^5 cell per well were seeded in 50 µl medium in a flat-bottom 96-well plate (Greiner). Inhibitor dilution series were added in 2-fold excess to required concentrations in 50 µl medium. After 44 hours, medium was

removed and 50 µl of Bright-Glo reagent (Promega) added to each well. Cells were incubated for five minutes in the dark and then transferred to a white flat 96-well plate (CLS3600, Corning) for measuring luminescence on a Synergy H1 plate reader (Agilent). The percent inhibition was calculated with the following formula.

$$\%\ inhibition = 100 - (\frac{100}{relative\ luminescense})$$

3CL^{pro} enzymatic activity

Wild type and variant proteases were produced in-house as described in 3CL^{pro} purification method section. Solution of wild type 3CL^{pro} and variants at 85 ng / 30 µl were prepared in appropriate buffer (20 mM Tris/HCl pH 8, 150 mM NaCl, 0.1 mg/ml BSA, 1 mM DTT) and these 30 µL were added to each well in a black 96- well plate (BPS Biosciences) to get a final concentration of 50 nM in 50 µl/well. The substrate Ac-Abu-Tle-Leu-Gln↓MCA (Acetyl-L-α- aminobutyroyl-L-tert-leucyl-L-leucyl-L-glutamine α-(4-methyl-coumaryl-7-amide)) was purchased from Peptide Inc., resuspended in dimethyl sulfoxide at 5 mM concentration. 20 µl buffer with diluted substrate was then added to the protein solutions at different concentrations. The plate was immediately placed inside a GloMax Explorer reader (Promega), and fluorescence emission measured with a substrate-appropriate filter. To determine enzymes initial velocities, we plotted RFU (relative fluorescent units) on the y-axis and time (min) on the x-axis. We performed a simple linear regression analysis. Fitting values from zero up to 60 minutes were used in a range that had a linear increase. The resulting slopes represented the initial velocity expressed as RFU/min for each protein variant at each substrate concentration. Slope values were plotted (y-axis) against the substrate concentration (x-axis). Finally, the obtained values were fitted using the "Michaelis-Menten" equation built-into GraphPad Prism 9 to extrapolate the kinetic parameters K_m and V_{max}.

$$Y = \frac{V_{max} \cdot X}{K_m + X}$$

k_{cat} was calculated dividing V_{max} by [ET], where [ET] is the give enzyme concentration. Wild type and variant 3CL^{pro} catalytic efficiencies were determined as $k_{cat}/K_{M.}$

Cross validation with biochemical 3CL^{pro} inhibition assay

The biochemical assay used to confirm mutations was based on the 3CL^{pro} activity assay from BPS Biosciences, catalogue number #78042–2. The 3CL^{pro} in the kit was replaced by an in-house produced 3CL^{pro} and mutants thereof, as described in 3CL^{pro} purification. Solutions of wild-type 3CL^{pro} and mutants at 5 ng/µl in 30 µl buffer (composition described above) were prepared according to the kit's manual. Ten µl of five-fold excess to tested nirmatrelvir concentrations were added to the 30 µl of 3CL^{pro} solution and incubated for 30 minutes. Then, 10 µl of fluorogenic substrate (DABCYL-KTSAVLQSGFRKME-EDANS) was added (generating a in total a 1:5 dilution of the excess nirmatrelvir and therefore final concentrations) and incubated for 4 hours. Fluorescence was induced with 365 nm UV-light and read at 460 nm in a Glomax Explorer (Promega). Blanks (assay buffer plus substrate) values were subtracted from sample values.

Replication kinetics with recombinant SARS-CoV-2 expressing mCherry

Monolayers of Vero E6 cells (6-well plate, 10^6 cells per well, triplicates) were infected with the indicated viruses at MOI 0.01. After viral absorption for 3 hours at 37°C, the supernatant was discarded, the cells were washed three times with PBS, and post-infection media (3 ml/well) was added. At the indicated time points, the supernatant (300 µl/well) was collected and titrated by plaque assay (29).

Cross-validation with recombinant SARS-CoV-2 (rWA1) expressing mCherry

A monolayer of A549-hACE2 cells was infected with 300 plaque- forming units (PFU) of indicated viruses in quadruplicates at 37°C. After viral adsorption for 1 hour, the supernatant was discarded, and the cells were washed twice with PBS. Then, phenol red-free post-infection medium (DMEM +2% fetal bovine serum +1% penicillin-streptomycin-glutamine (PSG)) containing the indicated concentrations of nirmatrelvir was added to each well. The mCherry intensity was determined at 48- and 72-hours post-infection under a Synergy LX Multimode Reader (Agilent). Wells without drug or virus were used as negative controls or baseline signal. Positive controls were wells with virus, but no drug. Infection percentages of wells with different amounts of inhibitor were calculated by subtracting the negative control (mean of wells without virus or drug) and then dividing

by the positive control (mean of wells with virus but without drug). Data were analyzed in GraphPad Prism 9 and IC_{50} values were calculated as the highest dilution of the nirmatrelvir-containing sample that prevents 50% plaque formation in infected cells, determined by a sigmoidal dose-response curve (see statistical analysis section).

IC_{50} and EC_{50} calculations

In this study, different assay systems were used to generate resistance data, namely VSV-based cellular assays with FluoroSpot and flow cytometry read-outs, an orthologous cell-based assay with a luciferase read-out, as well as a biochemical assay and SARS-CoV-2-mCherry assay with fluorescence read-outs. Al-though the magnitudes of resistance are different in these assays, the tendencies agree. We expected the dynamic range of the $3CL^{pro}$ cellular assays to be greater than in a biochemical assay, where there is a fixed amount of enzyme. In cells, the continuous renewal of protease-viral fusion proteins in an excess of inhibitor likely led to a later plateauing of the signal. At lower concentrations, compound molecules are depleted and the signal plateaus. In FluoroSpot read-outs, the $3CL^{pro}$-On assay data were normalized to the highest mean value in an experiment. $3CL^{pro}$-Off data were normalized to the highest value of each construct in an experiment. In the flow cytometry experiments, $3CL^{pro}$-On assay data were also normalized to the highest value of each construct in an experiment. For purified wild-type and mutant enzymes, IC_{50} values were determined using the biochemical assay “3CL Protease, Untagged (SARS-CoV-2)” from BPS Biosciences with the assay’s substrate DABCYL-KTSAVLQSGFRKME-EDANS. IC_{50} and EC_{50} calculations and statistical analysis for all assays were performed with GraphPad Prism 9 (see statistical analysis section).

Nanopore sequencing of recombinant SARS-CoV-2 (rWA1) expressing mCherry

To validate the sequence of the recombinant SARS-CoV-2 (rWA1) expressing mCherry, we used the Nanopore sequencing “Midnight protocol”, version 6[56]. Primer pools generating 1200 bp overlap-ping amplicons were purchased from Integrated DNA Technologies, as referenced in the abovementioned protocol. The sequencing reactions were prepared using the Rapid Barcoding Kit SQK-RBK110.96 (Oxford Nanopore Technologies) and were performed in a sequencer (MinION Mk1B) using a proprietary flowcell (R9.4.1, Oxford

Nanopore Technologies). Electrical signals are translated into nucleotide sequences (basecalling). Sequenced reads were sorted into separate files for each sample (de-multiplexing). Demultiplexing was done using the super high accuracy model in Guppy 6.1.5. Output sequences generated so called fastq files and sequences below 200 and above 1200 bp were removed. Sequences between 200 and 1200 bp were assembled with the algorithm epi2me-labs/wf-artic v0.3.18 in Nextflow 22.04.4. The SARS-CoV-2 lineage pangolin 4.1.1 was used to map the sequences. A visualization application (Nextclade 2.4.0) was used to check mutations.

Protein structure preparation for molecular modelling

The three-dimensional structure of the SARS-CoV-2 $3CL^{pro}$ complexed with nirmatrelvir was retrieved from the Protein Data Bank (PDB code: 7RFW, method: X-ray diffraction, resolution: 1.73 Å (31)) and prepared for molecular modeling evaluations exploiting several tools implemented in the Molecular Operating Environment (MOE) 2022.02 suite (16). Specifically, the “Structure Preparation” tool was used to assign each protein residue with alternative conformations to the one characterized by the highest occupancy value, and the “Protonate 3D” program was exploited to assign each titratable amino acid to the most appropriate protonation state at a pH of 7.4. Finally, the coordinates of hydrogen atoms were energy minimized using the AMBER10:EHT force field[57] until a gradient of 0.1 kcal mol^{-1} $Å^{-2}$ was reached.

In silico alanine and resistance mutation scanning

An in-silico evaluation of the impact of SARS-CoV-2 $3CL^{pro}$ mutations on both the stability of the protein and the affinity towards nirmatrelvir was conducted using the “Protein Design” module of MOE, using the previously mentioned complex structure. Particularly, the “Alanine Scan” and “Resistance Scan” tools were used to perform two virtual mutagenesis experiments. First, we applied the “Alanine Scan” interface, in which each of the 612 amino acids composing the dimeric SARS-CoV-2 $3CL^{pro}$ was mutated into an alanine residue, calculating at each given time the energy difference between the mutated protein and the wild-type form concerning both the potential energy of the protein itself (dStability) and the affinity towards nirmatrelvir (dAffinity). Values were

then color plotted on the crystal complex of nirmatrelvir and SARS-CoV-2 3CL^{pro} using UCSF Chimera[58]. Then we used the "Resistance Scan" interface to investigate the impact of a selected pool of mutations: Y54C, G138S, L141F, L167F, and Q192R. For both types of calculations, the conformational sampling was carried out through LowModeMD[59], using the AMBER10:EHT forced field coupled with the Generalized Born implicit solvent model[60]; the dAffinity value was determined through the GBVI/VSA[61] method.

Statistical analysis

Raw and normalized data are provided in Data file S4. Dose response data points of 3CL^{pro}-On, Off and biochemical assays were fitted using a four-parameter logistic regression (sigmoid, 4PL, X is concentration). IC_{50} values were extrapolated as the concentration value at which the signal was 50% between the top and bottom plateaus of each sub-dataset. Dose responses curves of the Src-3CL^{pro}-Tat-Luc-based assay were fitted with the same regression, setting X as 2 for IC_{50} extrapolation.

$$Y = Bottom + \frac{Top - Bottom}{1 + (\frac{IC_{50}}{X})^{HillSlope}}$$

Data obtained with flow cytometry were normalized and fit based using the non-linear regression function "[Agonist] vs. normalized response". The EC_{50} values were extrapolated as the medium value between the top and bottom plateaus of each sub-dataset.

$$Y = \frac{100 * X}{EC_{50} + X}$$

Nirmatrelvir dose response curves of recombinant SARS-CoV-2 expressing mCherry were normalized and fitted using the non-linear regression function "log(inhibitor) vs. normalized response - variable slope".

$$Y = \frac{100}{1 + 10^{\log(IC_{50} - X) * HillSlope}}$$

Kinetic parameters and catalytic activity of wild-type and mutant 3CL^{pro} enzymes were calculated as described in the corresponding method section. 95% confidence intervals were generated by the de-scribed fittings and are provided in supplementary tables

together with IC_{50} and EC_{50} values. All statistical analyses were performed with GraphPad Prism 9.

References

1. Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270–273 (2020).

2. Voelker, R. Combination Drug for HCV Infection. *JAMA* **318**, 790–790 (2017).

3. Harrington, M. & Carpenter, C. C. J. Hit HIV-1 hard, but only when necessary. *Lancet* **355**, 2147–2152 (2000).

4. Fan, K. *et al.* Biosynthesis, Purification, and Substrate Specificity of Severe Acute Respiratory Syndrome Coronavirus 3C-like Proteinase. *Journal of Biological Chemistry* **279**, 1637–1642 (2004).

5. Harcourt, B. H. *et al.* Identification of Severe Acute Respiratory Syndrome Coronavirus Replicase Products and Characterization of Papain-Like Protease Activity. *J Virol* **78**, 13600–13612 (2004).

6. Lamb, Y. N. Nirmatrelvir Plus Ritonavir: First Approval. *Drugs* **82**, 585–591 (2022).

7. Brogi, S. *et al.* Covalent Reversible Inhibitors of Cysteine Proteases Containing the Nitrile Warhead: Recent Advancement in the Field of Viral and Parasitic Diseases. *Molecules* **27**, (2022).

8. Greasley, S. E. *et al.* Structural basis for the in vitro efficacy of nirmatrelvir against SARS-CoV-2 variants. *J Biol Chem* **298**, 101972 (2022).

9. Iketani, S. *et al.* Multiple pathways for SARS-CoV-2 resistance to nirmatrelvir. *bioRxiv* 2022.08.07.499047 (2022) doi:10.1101/2022.08.07.499047.

10. Jochmans, D. *et al.* The substitutions L50F, E166A and L167F in SARS-CoV-2 3CLpro are selected by a protease inhibitor in vitro and confer resistance to nirmatrelvir. *bioRxiv* 2022.06.07.495116 (2022) doi:10.1101/2022.06.07.495116.

11. Moghadasi, S. A. *et al.* Transmissible SARS-CoV-2 variants with resistance to clinical protease inhibitors. *bioRxiv* 2022.08.07.503099 (2022) doi:10.1101/2022.08.07.503099.

12. Hu, Y. *et al.* Naturally occurring mutations of SARS-CoV-2 main protease confer drug resistance to nirmatrelvir. doi:10.1101/2022.06.28.497978.

13. Padhi, A. K. & Tripathi, T. Hotspot residues and resistance mutations in the nirmatrelvir-binding site of SARS-CoV-2 main protease: Design, identification, and correlation with globally circulating viral genomes. *Biochem Biophys Res Commun* **629**, 54–60 (2022).

14. Kaufer, A. M., Theis, T., Lau, K. A., Gray, J. L. & Rawlinson, W. D. Laboratory biosafety measures involving SARS-CoV-2 and the classification as a Risk Group 3 biological agent. *Pathology* **52**, 790 (2020).

15. Baek, M. *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).

16. Molecular Operating Environment (MOE), 2019.01; Chemical Computing Group ULC, 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, 2021. https://www.chemcomp.com/Research-Citing_MOE.htm.

17. Heilmann, E. *et al.* A VSV-based assay quantifies coronavirus Mpro/3CLpro/Nsp5 main protease activity and chemical inhibition. *Communications Biology 2022 5:1* **5**, 1–12 (2022).

18. Jaskolski, M. *et al.* Crystallographic models of SARS-CoV-2 3CLpro: in-depth assessment of structure quality and validation. *IUCrJ* **8**, 238–256 (2021).

19. Silvestrini, L. *et al.* The dimer-monomer equilibrium of SARS-CoV-2 main protease is affected by small molecule inhibitors. *Scientific Reports 2021 11:1* **11**, 1–16 (2021).

20. Drake, J. W. & Holland, J. J. Mutation rates among RNA viruses. *Proc Natl Acad Sci U S A* **96**, 13910–13913 (1999).

21. Steinhauer, D. A., Domingo, E. & Holland, J. J. Lack of evidence for proofreading mechanisms associated with an RNA virus polymerase. *Gene* **122**, 281–288 (1992).

22. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro Surveill* **22**, (2017).

23. Elbe, S. & Buckland-Merrett, G. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Glob Chall* **1**, 33–46 (2017).

24. Khare, S. *et al.* GISAID's Role in Pandemic Response. *China CDC Wkly* **3**, 1049 (2021).

25. Hatcher, E. L. *et al.* Virus Variation Resource - improved response to emergent viral outbreaks. *Nucleic Acids Res* **45**, D482–D490 (2017).

26. FACT SHEET FOR HEALTHCARE PROVIDERS: EMERGENCY USE AUTHORIZATION FOR PAXLOVID TM HIGHLIGHTS OF EMERGENCY USE AUTHORIZATION (EUA) These highlights of the EUA do not include all the information needed to use PAXLOVID TM under the EUA. See the FULL FACT SHEET FOR HEALTHCARE PROVIDERS for PAXLOVID.

27. Moghadasi, S. A. *et al.* Gain-of-Signal Assays for Probing Inhibition of SARS-CoV-2 Mpro/3CLpro in Living Cells. *mBio* **13**, (2022).

28. Mukae, H. *et al.* A Randomized Phase 2/3 Study of Ensitrelvir, a Novel Oral SARS-CoV-2 3C-Like Protease Inhibitor, in Japanese Patients with Mild-to-Moderate COVID-19 or Asymptomatic SARS-CoV-2 Infection: Results of the Phase 2a Part. *Antimicrob Agents Chemother* **66**, (2022).

29. Ye, C. *et al.* Analysis of SARS-CoV-2 infection dynamic in vivo using reporter-expressing viruses. *Proc Natl Acad Sci U S A* **118**, e2111593118 (2021).

30. Ye, C. *et al.* Rescue of SARS-CoV-2 from a single bacterial artificial chromosome. *mBio* **11**, 1–10 (2020).

31. Owen, D. R. *et al.* An oral SARS-CoV-2 M pro inhibitor clinical candidate for the treatment of COVID-19. *Science (1979)* **374**, 1586–1593 (2021).

32. Zhao, Y. *et al.* Crystal structure of SARS-CoV-2 main protease in complex with protease inhibitor PF-07321332. *Protein Cell* (2021) doi:10.1007/s13238-021-00883-2.

33. Morrison, K. L. & Weiss, G. A. Combinatorial alanine-scanning. *Curr Opin Chem Biol* **5**, 302–307 (2001).

34. Whittaker, J. *et al.* Alanine Scanning Mutagenesis of a Type 1 Insulin-like Growth Factor Receptor Ligand Binding Site. *Journal of Biological Chemistry* **276**, 43980–43986 (2001).

35. Liu, X., Peng, L., Zhou, Y., Zhang, Y. & Zhang, J. Z. H. Computational Alanine Scanning with Interaction Entropy for Protein–Ligand Binding Free Energies. *J Chem Theory Comput* **14**, 1772–1780 (2018).

36. Chandra, N., Anand, P., Nagarajan, D. & Mukherjee, S. ABS–Scan: In silico alanine scanning mutagenesis for binding site residues in protein–ligand complex. *F1000Res* **3**, (2014).

37. Pavan, M., Bolcato, G., Bassani, D., Sturlese, M. & Moro, S. Supervised Molecular Dynamics (SuMD) Insights into the mechanism of action of SARS-CoV-2 main protease inhibitor PF-07321332. *J Enzyme Inhib Med Chem* **36**, 1646–1650 (2021).

38. Kim, D. E., Chivian, D. & Baker, D. Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Res* **32**, (2004).

39. Liu, Z. *et al.* Identification of SARS-CoV-2 spike mutations that attenuate monoclonal and serum antibody neutralization. *Cell Host Microbe* **29**, 477-488.e4 (2021).

40. Schmidt, F. *et al.* High genetic barrier to SARS-CoV-2 polyclonal neutralizing antibody escape. *Nature 2021 600:7889* **600**, 512–516 (2021).

41. Yahalom-Ronen, Y. *et al.* Neutralization of SARS-CoV-2 Variants by rVSV-ΔG-Spike-Elicited Human Sera. *Vaccines (Basel)* **10**, (2022).

42. Ullrich, S. & Nitsche, C. The SARS-CoV-2 main protease as drug target. *Bioorg Med Chem Lett* **30**, 127377 (2020).

43. Ullrich, S., Ekanayake, K. B., Otting, G. & Nitsche, C. Main protease mutants of SARS-CoV-2 variants remain susceptible to nirmatrelvir. *Bioorg Med Chem Lett* **62**, (2022).

44. Rut, W. *et al.* SARS-CoV-2 Mpro inhibitors and activity-based probes for patient-sample imaging. *Nature Chemical Biology 2020 17:2* **17**, 222–228 (2020).

45. Manandhar, A. *et al.* Discovery of Novel Small-Molecule Inhibitors of SARS-CoV-2 Main Protease as Potential Leads for COVID-19 Treatment. *J Chem Inf Model* **61**, 4745–4757 (2021).

46. Manandhar, A. *et al.* Targeting SARS-CoV-2 M3CLpro by HCV NS3/4a Inhibitors: In Silico Modeling and In Vitro Screening. *J Chem Inf Model* **61**, 1020–1032 (2021).

47. Hsu, M. F. *et al.* Mechanism of the maturation process of SARS-CoV 3CL protease. *J Biol Chem* **280**, 31257–31266 (2005).

48. Muramatsu, T. *et al.* SARS-CoV 3CL protease cleaves its C-terminal autoprocessing site by novel subsite cooperativity. *Proc Natl Acad Sci U S A* **113**, 12997–13002 (2016).

49. Gibson, D. G. *et al.* Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nature Methods 2009 6:5* **6**, 343–345 (2009).

50. Schnell, M. J., Buonocore, L., Whitt, M. A. & Rose, J. K. The minimal conserved transcription stop-start signal promotes stable expression of a foreign gene in vesicular stomatitis virus. *J Virol* **70**, 2318–2323 (1996).

51. Panda, D., Dinh, P. X., Beura, L. K. & Pattnaik, A. K. Induction of Interferon and Interferon Signaling Pathways by Replication of Defective Interfering Particle RNA in Cells Constitutively Expressing Vesicular Stomatitis Virus Replication Proteins. *J Virol* **84**, 4826 (2010).

52. Witko, S. E. *et al.* An efficient helper-virus-free method for rescue of recombinant paramyxoviruses and rhadoviruses from a cell line suitable for vaccine development. *J Virol Methods* **135**, 91–101 (2006).

53. Dold, C. *et al.* Application of interferon modulators to overcome partial resistance of human ovarian cancers to VSV-GP oncolytic viral therapy. *Mol Ther Oncolytics* **3**, 16021 (2016).

54. Kärber, G. Beitrag zur kollektiven Behandlung pharmakologischer Reihenversuche. *Naunyn-Schmiedebergs Archiv für experimentelle Pathologie und Pharmakologie 1931 162:4* **162**, 480–483 (1931).

55. Zhang, L. *et al.* Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved α-ketoamide inhibitors. *Science (1979)* **368**, 409–412 (2020).

56. Freed, N. E., Vlková, M., Faisal, M. B. & Silander, O. K. Rapid and inexpensive whole-genome sequencing of SARS-CoV-2 using 1200 bp tiled amplicons and Oxford Nanopore Rapid Barcoding. *Biol Methods Protoc* **5**, (2020).

57. D.A. Case et al. Amber 10, University of California, San Francisco.

58. UCSF Chimera--a visualization system for exploratory research and analysis. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. J Comput Chem. 2004 Oct;25(13):1605-12. https://www.cgl.ucsf.edu/chimera/docs/credits.html.

59. Labute, P. LowModeMD—Implicit Low-Mode Velocity Filtering Applied to Conformational Search of Macrocycles and Protein Loops. *J Chem Inf Model* **50**, 792–800 (2010).

60. Onufriev, A., Case, D. A. & Bashford, D. Effective Born radii in the generalized Born approximation: The importance of being perfect. *J Comput Chem* **23**, 1297–1304 (2002).

61. Labute, P. The generalized Born/volume integral implicit solvent model: Estimation of the free energy of hydration using London dispersion instead of atomic surface area. *J Comput Chem* **29**, 1693–1698 (2008).

A new inactive conformation of SARS-CoV-2 main protease

Emanuele Fornasier, Maria Ludovica Macchia, Gabriele Giachin, Alice Sosic, **Matteo Pavan**, Mattia Sturlese, Cristiano Salata, Stefano Moro, Barbara Gatto, Massimo Bellanda and Roberto Battistutta

Fornasier, E. *et al.* A new inactive conformation of SARS-CoV-2 main protease. *Acta Crystallogr D Struct Biol* **78**, 363–378 (2022).

Abstract

The SARS-CoV-2 main protease (M^{pro}) has a pivotal role in mediating viral genome replication and transcription of the coronavirus, making it a promising target for drugs against the COVID-19 pandemic. Here, a crystal structure is presented in which M^{pro} adopts an inactive state that has never been observed before, called new-inactive. It is shown that the oxyanion loop, which is involved in substrate recognition and enzymatic activity, adopts a new catalytically incompetent conformation and that many of the key interactions of the active conformation of the enzyme around the active site are lost. Solvation/desolvation energetic contributions play an important role in the transition from the inactive to the active state, with Phe140 moving from an exposed to a buried environment and Asn 142 moving from a buried environment to an exposed environment. In new-inactive M^{pro} a new cavity is present near the S2' subsite, and the N-terminal and C-terminal tails, as well as the dimeric interface, are perturbed, with partial destabilization of the dimeric assembly. This novel conformation is relevant both for comprehension of the mechanism of action of M^{pro} within the catalytic cycle and for the successful structure-based drug design of antiviral drugs.

1. Introduction

To face the global COVID-19 pandemic, besides prevention via the use of vaccines, it is also essential to develop targeted therapeutic options for patients infected by the SARS-CoV-2 betacoronavirus. In general, one of the most promising classes of antiviral drug candidates are protease inhibitors, small molecules that are able to inhibit enzymes involved in virus replication within the cell. Very low sequence identity with human proteases and distinct cleavage-site specificities suggest that viral enzymes can be

inhibited with very low associated toxic effects (`off-target' effects), if any. Indeed, protease inhibitors have already been efficient in the treatment of viral pathogens such as hepatitis C virus[1] and human immunodeficiency virus (HIV)[2]. In coronaviruses, the main protease, M^{pro}, is a cysteine peptidase that is essential for the replication cycle of positive-sense, single-stranded RNA coronaviruses[3], including SARS-CoV-2. It is also known as 3C-like protease or $3CL^{pro}$ from the similarity of its active site and its substrate specificity to those of the picornavirus 3C protease[4]. M^{pro} is involved in the proteolytic processing of the two overlapping polyproteins pp1a and pp1ab, with the formation of individual mature nonstructural proteins[5], and as such it is a validated antiviral drug target[6–8]. Currently, there are at least two SARS-CoV-2 M^{pro} inhibitors in phase I clinical trials as candidates with potent antiviral activity: the orally administered PF-07321332[9] and the intravenously administered PF-00835231[10].

SARS-CoV-2 M^{pro} (nsp5), a 306-amino-acid polypeptide of molecular weight 33.8 kDa[11], shares 96% sequence identity and a very similar 3D structure with SARS-CoV M^{pro} [0.53 Å r.m.s.d. between PDB entries 6y2e[12] and 2bx4[13]]. Very similar 3D structures have also been found for other coronaviral M^{pro}s such as those from Porcine transmissible gastroenteritis virus (TGEV), which was the first structure of a coronaviral M^{pro} [4], Human coronavirus (HCoV) strain 229E[14], Infectious bronchitis virus (IBV)[15] and MERS-CoV[16]. This structural similarity, which is particularly relevant around the active site, leads to the possibility of the development of pan-coronaviral drugs. M^{pro} exists in an equilibrium between a monomer and a homodimer (with the two protomers roughly perpendicularly oriented; Fig. 1*a*), with an apparent K_d of between 0.8 and 14 µ*M* for the SARS-CoV enzyme, depending on the experimental conditions[17]. For SARS-CoV-2 M^{pro}, the K_d has been estimated to be 2.5 µ*M* by analytical ultracentrifugation[12] and 0.14 µ*M* by native mass spectrometry[18]. Unlike 3C protease, only the SARS-CoV M^{pro} dimer shows enzymatic activity[4] and the correct shape of the substrate-binding site, particularly of the S1 subsite; the correct conformation for productive catalytic events is linked to the dimerization process. It has been proposed that the dimerization process has a direct regulatory role of the activity of M^{pro} during the coronaviral replication process[19,20]. Given the high structural similarity, particularly at the dimeric interface, it was reasoned that

dimerization of the enzyme is also necessary for the catalytic activity of SARS-CoV-2 M^{pro} [12].

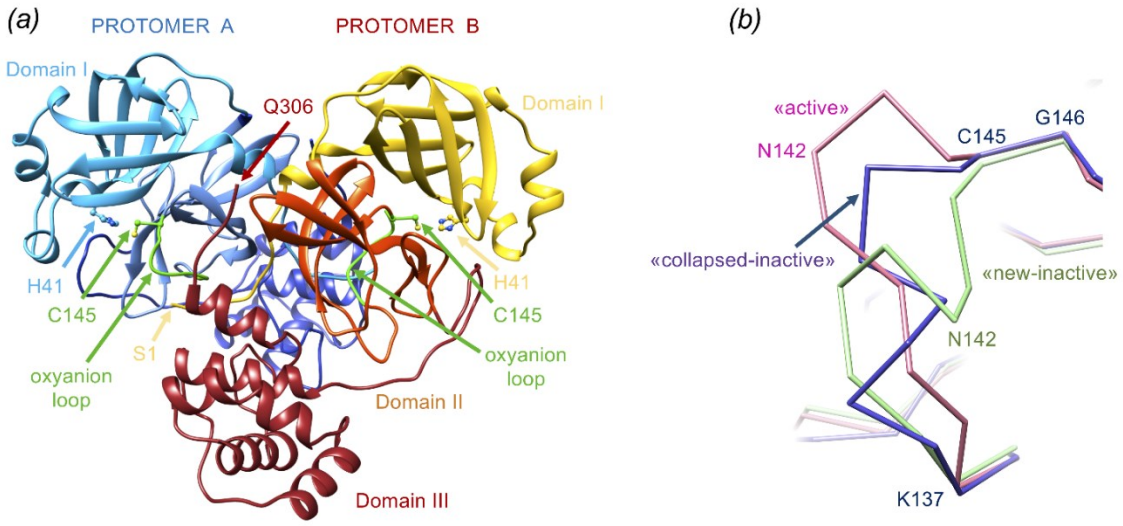


Figure 1. SARS-CoV-2 M^{pro} architecture, free form (PDB entry 6y2e). (*a*) Dimeric assembly of the protease with the main structural features discussed in the text highlighted. Protomer *A* is in blue-based colors and protomer *B* is in yellow/red-based colors. The two oxyanion loops and the two catalytic cysteines 145 are shown in green. (*b*) Comparison between different oxyanion-loop conformations of M^{pro}: active in SARS-CoV-2 M^{pro} (PDB entry 6y2e) in pink, collapsed-inactive in SARS-CoV M^{pro} (PDB entry 1uj1 chain *B*) in magenta and new-inactive in SARS-CoV-2 M^{pro} (this work) in green.

Each M^{pro} protomer is composed of three structural domains (Fig. 1*a*)[4]. The chymotrypsin-like and 3C protease-like β-barrel domains I (residues 1–99) and II (residues 100–182) directly control the catalytic event. The substrate-binding site is between these two domains and comprises several subsites for substrate binding (from S1 to S6 and from S1' to S3'), corresponding to the P1–P6 and P1'–P3' amino-acid positions of the substrates (according to the convention P6–P5–P4–P3–P2–P1↓P1'–P2'–P3', where ↓ indicates the hydrolyzed peptide bond)[14]. Enzymatic proteolysis by SARS-CoV-2 M^{pro} at the 11 cleavage sites on the viral polyprotein occurs on the C-terminal side of a conserved glutamine in position P1, with the most common consensus sequence being Leu-Gln↓(Ser/Ala), indicating that specificity is determined mostly by the P2, P1 and P1' positions[8]. Glutamine in position P1 is fully conserved not only for SARS-CoV-2 but also in substrates of SARS-CoV and MERS-CoV. Prime recognition sites at the C-terminus of P1' are not conserved.

M^{pro} subsites S4, S2, S1 and S1' have been identified as the most relevant subsites for substrate binding, with regions in the S5, S4 and S2 sites showing considerable conformational flexibility upon binding different chemical groups[21]. The chymotrypsin-like fold, including domains I and II, is connected by a 16-residue flexible loop to the extra α-helical domain III (residues 198–306; Fig. 1*a*). Domain III is absent in other RNA virus 3C-like proteases and plays a key role in enzyme dimerization and activity regulation of M^{pro} [4,22].

At variance with the classical catalytic triad of chymotrypsin-like proteases, coronaviral M^{pro} has a catalytic dyad, consisting of His41 and Cys145 in SARS-CoV-2 (Fig. 1*a*); a conserved water molecule occupies a position analogous to that of the side chain of the third member of the catalytic triad (for instance, aspartate in chymotrypsin and asparagine in papain) and forms hydrogen bonds to the side chains of His41, His164 and Asp187. It has been proposed that this conserved water is involved in the catalytic event[4].

A key role in the proper function of the enzyme is also played by the N-finger (residues 1–7) as the N-terminal tail of one protomer interacts and stabilizes the binding site (S1 subsite) of the other protomer[23]. Indeed, deletion of the N-finger hampers dimerization in solution and abolishes the proteolytic activity. Both the N-finger and the C-terminus are results of the autoproteolytic processing of M^{pro}. Accordingly, in the mature dimeric enzyme both termini of one protomer face the active site of the other.

The important conserved residues Phe140, Leu141, Asn142, Ser144 (SARS-CoV-2 numbering) are part of a structural element that is essential for a productive catalytic event, the so-called oxyanion loop comprising residues 138–145, which globally lines the binding site for glutamine P1. The central role of the oxyanion loop in the catalytic reaction mechanism of serine proteases and cysteine proteases has been extensively characterized[24]. The correct positioning of the oxyanion hole, which is part of the oxyanion loop (formed by the backbone of Gly143, Ser144 and Cys145 in SARS-CoV-2 M^{pro}), is essential for stabilization of the transient tetrahedral acyl (oxyanion) transition state via the hydrogen-bond donor properties of the amides[4,23,25]. In the known crystal structures of SARS-CoV and SARS-CoV-2 M^{pro}, the oxyanion loop adopts essentially the same `active'

conformation; here, we take PDB entry 6y2e as a reference for this conformation[12,26–28]. A specific conformation is defined to be active when the amino acids known to participate in the chemical reaction catalyzed by the enzyme are properly positioned and oriented for the reaction to proceed. We also term this conformation catalytically competent.

Variations from the active conformation of the oxyanion loop are found in a few forms of the enzyme, which were consequently considered to be inactive or catalytically incompetent, as in protomer *B* of SARS-CoV M^{pro} (PDB entries 1uj1 and 1uk2)[29], in the monomeric R298A mutant of SARS-CoV M^{pro} (PDB entry 2qcy)[30] and in the C172A mutant of $3C^{pro}$ from the picornavirus hepatitis A virus[31], as well as in IBV $3CL^{pro}$ (PDB entries 2q6f and 2q6d)[15]. In the inactive monomeric R298A mutant (PDB entry 2qcy), the region of the oxyanion loop, Ser139-Phe140-Leu141, is converted into a short 3_{10}-helix. In PDB entry 1uj1 (SARS-CoV M^{pro} crystallized at pH 6) the oxyanion loop of one of the two protomers exists in a `collapsed' conformation (similar to that found in PDB entry 2qcy), which is considered to be catalytically incompetent, in which the hydrogen bond between Glu166 and His172 that is important for activity is broken[29]. In the following, we will refer to these two inactive conformations with similar oxyanion-loop conformations as collapsed-inactive (Fig. 1*b*). In the vast majority of SARS-CoV and SARS-CoV-2 M^{pro} crystal structures, the dimer is crystallographic[32]; that is, there is only one molecule in the asymmetric unit and therefore the two protomers are perfectly identical. In the very few inactive structures, apart from the artificially induced monomeric forms, the dimer is formed by two different molecules present in the asymmetric unit, one of which is in the inactive state and the other of which is in the active state. Based on molecular-dynamics simulations coupled to activity data in solution, it was suggested that only one protomer at a time is active in the dimer[17].

Here, we describe a new inactive structure (called new-inactive) of the main protease of SARS-CoV-2 that is clearly distinct from both the active and the known collapsed-inactive structures, with an oxyanion-loop conformation that is very different from those previously described (Fig. 1*b*). In Section 4, we argue that this conformation has an important functional role as part of the catalytic cycle of coronaviral M^{pro}.

2. Materials and methods

2.1. Recombinant protein production and purification

The plasmid PGEX-6p-1 encoding SARS-CoV-2 M^{pro} [12] was a generous gift from Professor Rolf Hilgenfeld, University of Lübeck, Lübeck, Germany. Recombinant protein production and purification were adapted from Zhang *et al.*[12] (where the structure of M^{pro} in the active form was presented; PDB entry 6y2e). The expression plasmid was transformed into *Escherichia coli* strain BL21 (DE3) and then precultured in YT medium at 37°C (100 µg ml^{-1} ampicillin) overnight. The preculture was used to inoculate fresh YT medium supplemented with antibiotic and the cells were grown at 37°C to an OD_{600} of 0.6–0.8 before induction with 0.5 m*M* isopropyl β-D-1-thiogalactopyranoside (IPTG). After 5 h at 37°C, the cells were harvested by centrifugation (5000*g*, 4°C, 15 min) and frozen. The pellets were resuspended in buffer *A* (20 m*M* Tris, 150 m*M* NaCl pH 7.8) supplemented with lysozyme, DNase I and PMSF for lysis. The lysate was clarified by centrifugation at 12 000*g* at 4°C for 1 h and loaded onto a HisTrap HP column (GE Healthcare) equilibrated with 98% buffer *A*/2% buffer *B* (20 m*M* Tris, 150 m*M* NaCl, 500 m*M* imidazole pH 7.8). The column was washed with 95% buffer *A*/5% buffer *B*, and His-tagged M^{pro} was then eluted with a linear gradient of imidazole from 25 to 500 m*M*. Pooled fractions containing the target protein were subjected to buffer exchange with buffer *A* using a HiPrep 26/10 desalting column (GE Healthcare). Next, PreScission protease was added to remove the C-terminal His tag (20 µg of PreScission protease per milligram of target protein) at 12°C overnight. The protein solution was loaded onto a HisTrap HP column connected to a GSTrap FF column (GE Healthcare) equilibrated in buffer *A* to remove the GST-tagged Pre-Scission protease, the His tag and the uncleaved protein. M^{pro} was finally purified using a Superdex 75 prep-grade 16/60 SEC column (GE Healthcare) equilibrated with buffer *C* (20 m*M* Tris, 150 m*M* NaCl, 1 m*M* EDTA, 1 m*M* DTT pH 7.8). Fractions containing the target protein with high purity were pooled, concentrated to 25 mg ml^{-1} and flash-frozen in liquid nitrogen for storage in small aliquots at −80°C.

2.2. Protein characterization and enzymatic kinetics

The correctness of the M^{pro} DNA sequence was verified by sequencing the expression plasmid. The molecular mass was determined as follows: recombinant SARS-CoV-2 M^{pro}, diluted in 50% acetonitrile with 0.1% formic acid, was analyzed by direct infusion electrospray ionization (ESI) on a Xevo G2-XS QTOF mass spectrometer (Waters). The detected species displayed a mass of 33 796.64 Da, which very closely matches the value of 33 796.81 Da calculated from the theoretical full-length protein sequence (residues 1–306). A representative ESI-MS spectrum is shown in Supplementary Fig. S1. To characterize the enzymatic activity of our recombinant M^{pro}, we adopted a FRET-based assay using the substrate 5-FAM-AVLQ↓SGFRK(DABCYL)K (Proteogenix). The assay was performed by mixing 0.05 µ*M* M^{pro} with various concentrations of substrate (1–128 µ*M*) in a buffer composed of 20 m*M* Tris, 100 m*M* NaCl, 1 m*M* EDTA, 1 m*M* DTT pH 7.3. Fluorescence intensity (excitation at 485 nm and emission at 535 nm) was monitored at 37°C with a VictorIII microplate reader (Perkin Elmer). A calibration curve was created by measuring multiple concentrations (from 0.001 to 5 µ*M*) of free fluorescein in a final volume of 100 µl reaction buffer. Initial velocities were determined from the linear section of the curve, and the corresponding relative fluorescence units per time unit (ΔRFU s^{-1}) were converted to the amount of cleaved substrate per time unit (µ*M* s^{-1}) by fitting to the calibration curve of free fluorescein. The catalytic efficiency k_{cat}/K_m was 4819 ± 399 s^{-1} M^{-1}, which is in line with literature data[12,33].

2.3. Crystallization and data collection

A frozen aliquot of M^{pro} was thawed in ice, diluted in a 1:2 ratio with buffer *C* (20 m*M* Tris, 150 m*M* NaCl, 1 m*M* EDTA, 1 m*M* DTT pH 7.8) to a final concentration of 12.5 mg ml^{-1} and clarified by centrifugation at 16 000*g*. The inhibitors masitinib, manidipine, bedaquiline and boceprevir were dissolved in 100% DMSO to a concentration of 100 m*M*. The protein was crystallized both in the free form and in the presence of inhibitors by co-crystallization. In all cases, final crystal growth was obtained by microseeding starting from small crystals of the free enzyme. The protein in the free form was crystallized using the sitting-drop vapor-diffusion method at 18°C, mixing 1.0 µl M^{pro} solution with 1.0 µl precipitant solution [0.1 *M* MMT (DL-malic acid, MES and Tris base in a 1:2:2 molar ratio)

pH 7.0, 25% PEG 1500] and 0.2 µl seed stock (diluted 1:500, 1:1000 or 1:2000 with precipitant solution) and equilibrating against a 300 µl reservoir of precipitant solution. Crystals appeared overnight and grew for 48 h after the crystallization drops had been prepared. In the case of co-crystallization, M^{pro} was incubated for 16 h at 8°C with a 13-fold molar excess of inhibitor (final DMSO concentration 5%). After incubation with masitinib, manidipine or bedaquiline, a white precipitate appeared and the solutions were clarified by centrifugation at 16 000*g*; as the protein concentration was essentially unchanged after centrifugation, we concluded that the precipitate is composed of the inhibitors, which are poorly soluble in water. The fact that the protein was later crystallized under the same conditions as described for the free form further confirmed that its concentration was not altered by the centrifugation process. For data collections, crystals were fished from the drops, cryoprotected by a quick dip into 30% PEG 400 (with 5 m*M* inhibitor in the case of co-crystals) and flash-cooled in liquid nitrogen. The crystals were monoclinic (space group *C*2), isomorphous to the crystals of the free enzyme (PDB entry 6y2e), with one monomer in the asymmetric unit; the dimer is formed by the crystallographic twofold axis.

2.4. Structure determination, refinement and analysis

Data were collected on beamlines ID23-2 and ID23-1 at the ESRF. Diffraction data integration and scaling were performed with *XDS*[34] and data reduction and analysis were performed with *AIMLESS*[35]. Initially, structures were solved by molecular replacement (MR) with *Phaser*[36] in *Phenix*[37] using PDB entries 6y2e and 5rel (M^{pro} in complex with PCM-0102340)[26] as search models. To limit MR model bias in critical zones (namely residues 139–144, 1–3 and the side chain of His163) we then performed new MR runs using PDB entry 6y2e without residues 139–144 and 1–3, and with an alanine instead of a histidine at position 163, as the search model. Only for co-crystallization experiments with boceprevir was electron density for the ligand clearly visible from the beginning of the refinement (Supplementary Figs. S2 and S3), and the three final structures, modeled from residues 1 to 306 (compared with the `new' structure modeled to residue 301), are virtually identical to those deposited in the PDB[38]. In all of the other cases, no electron density indicating the presence of the inhibitors masitinib, manidipine or bedaquiline in

the active site (or elsewhere) was detectable. For four structures, it was possible to efficiently model residues 139–144, 1–3 and the side chain of His163 in `new' conformations. The final structures were obtained by alternating cycles of manual refinement with *Coot*[39] and automatic refinement with *phenix.refine*[40]. At the end, the model was submitted to *phenix.ensemble_refinement*[41] with default parameters. Data-collection and refinement statistics for the structure obtained by a co-crystallization experiment with masitinib (which was not visible in the final electron density) are reported in Table 1. Secondary-structure analysis was performed with *DSSP* [42,43]. Local energetic frustration analysis was performed with the *Frustratometer* server (http://frustratometer.qb.fcen.uba.ar)[44]. Interface analysis was performed using *PISA*[45].

Data collection	
X-ray source	ESRF ID23-2
Wavelength (Å)	0.873130
Space group	C2
Cell dimensions	
a, b, c (Å)	113.07, 54.71, 44.84
α, β, γ (°)	90.00, 101.30, 90.00
Resolution range (Å)	55.44 – 1.58 (1.61 – 1.58)
R_{merge}	0.070 (1.305)
R_{meas}	0.081 (1.505)
R_{pim}	0.040 (0.739)
Total number of observations	145297 (7276)
Total number unique	36653 (1847)
Mean(I)/σ(I)	9.2 (1.0)
$CC_{1/2}$ (%)	99.8 (35.7)
Completeness (%)	99.4 (99.5)
Multiplicity	4.0 (3.9)

Wilson B estimate (Å^2)	23.7
Refinement	
Resolution range (Å)	55.44 – 1.58
R_{work}/R_{free} (%)	17.71/20.31
Number of atoms	
Protein	2350
Water	218
B-factor	
Protein (Å^2)	32.6
Water (Å^2)	43.1
r.m.s.d.	
Bond lengths (Å)	0.008
Bond angles (°)	0.868
Coordinate error (maximum-likelihood based by Phenix) (Å)	0.21
Ramachandran statistics	
Favored (%)	97.99
Allowed (%)	2.01
Outliers (%)	0.00
PDB entry	7NIJ
Ensemble refinement	
Number of models	60
R_{work}/R_{free} (%)	15.47/20.80
Values in parentheses are for the highest-resolution shell	

Table 1. X-ray diffraction data processing and model refinement statistics

2.5. Molecular modeling

The majority of the computational work was performed on a Linux desktop workstation (Intel Xeon CPU E5-1620 3.60 GHz) running Ubuntu 16.04 LTS. Molecular-dynamics trajectories were collected on a heterogeneous Nvidia GPU cluster composed of 20 GPUs with models spanning from GTX1080 to RTX2080Ti. For structure preparation, coordinates of the active conformation of SARS-CoV-2 M^{pro} were retrieved from the Protein Data Bank (PDB entry 6y2e). Coordinates for both the active and the new-inactive conformation were processed with the aid of the *Molecular Operating Environment* (*MOE*) 2019.01 (Chemical Computing Group) structure-preparation tool. Initially, the functional unit of the protease (the dimeric form) was restored by applying a symmetric crystallographic transformation to each asymmetric unit. Residues with alternate conformations were assigned to the highest occupancy alternative. Moreover, missing residues that are present in the primary sequence were added using the *MOE* Loop Modeler tool. The *MOE* Protonate3D tool was used to assign the most probable protonation state to each residue (pH 7.4, T = 310 K, i.f. = 0.154). Partial charges were then assigned using the AMBER10 force field and H atoms were energy-minimized until the gradient was below 0.1 kcal mol^{-1} $Å^{-2}$. Finally, ions and all co-crystallized molecules except for water were removed before saving the structures. The system setup for the MD simulations was carried out using the *antechamber*, *parmchk* and *tleap* software implemented in the *AmberTools*14 suite[46]. AMBER ff14SB[47] was adopted for system parametrization and attribution of partial charges. Protein structures were explicitly solvated in a rectangular prismatic TIP3P[48] periodic water box with borders placed at a distance of 15 Å from any protein atom. Na^+ and Cl^- ions were added to neutralize the system until a salt concentration of 0.154 *M* was reached. MD simulations were then performed using *ACEMD*3[49], which is based upon an OpenMM 7.4.2 engine[50]. Initially, 1000 steps of energy minimization were executed using the conjugate-gradient algorithm. A two-step equilibration procedure was then carried out: the first step consisted of a 1 ns canonical ensemble (NVT) simulation with 5 kcal mol^{-1} $Å^{-2}$ harmonic positional constraints applied to each protein atom, while the second step consisted of a 1 ns isothermal–isobaric (NPT) simulation with 5 kcal mol^{-1} $Å^{-2}$ harmonic positional

constraints applied only to protein C^{α} atoms. The production phase consisted of three independent MD replicas for each protein conformation. Each simulation had a duration of 1 μs and was performed using the NVT ensemble at a constant temperature of 310 K with a timestep of 2 fs. For both the equilibration and the production stage, the temperature was maintained constant using a Langevin thermostat. During the second step of the equilibration stage, the pressure was maintained at a fixed value of 1 atm with a Monte Carlo barostat. MD trajectories were aligned using protein C^{α} atoms from the first trajectory frame as a reference, wrapped into an image of the system under periodic boundary conditions (PBC), and subsequently saved using a 200 ps interval between each frame and removing any ions and water molecules using *Visual Molecular Dynamics* 1.9.2 (*VMD*)[51]. The protein radius of gyration (R_g), the root-mean-square deviation (r.m.s.d.) and the root-mean-square fluctuation (r.m.s.f.) of atomic positions along the trajectory were calculated for protein C^{α} atoms exploiting the MDAnalysis[52,53] Python module. Secondary-structure analysis was carried out with the *STRIDE* package[54] as implemented in *VMD* 1.9.2. The collected data were then plotted using the Matplotlib Python library[55].

Furthermore, two classic MD simulations were performed on the complexes obtained by superposing the coordinates of peptide ligands from PDB entries 2q6g and 7khp on the new-inactive conformation of SARS-CoV-2 M^{pro} using *MOE* 2019.01. For each peptide–ligand complex, a two-stage equilibration protocol followed by a single productive simulation was carried out. The first equilibration step consisted of a 0.1 ns canonical ensemble (NVT) simulation with 5 kcal mol^{-1} $Å^{-2}$ harmonic positional constraints applied to each protein atom, while the second equilibration step consisted of a 0.5 ns isothermal–isobaric (NPT) simulation with 5 kcal mol^{-1} $Å^{-2}$ harmonic positional constraints applied only to protein C^{α} atoms. For both equilibration simulations, the temperature was maintained constant (T = 310 K) using a Langevin thermostat, while during the second equilibration stage the pressure was kept at a constant value of 1 atm using a Monte Carlo barostat. The productive simulation was carried out for 10 ns in the NVT ensemble (T = 310 K).

3. Results

3.1. Identification of a new-inactive conformation of M^{pro}

In a campaign to obtain structural insights into SARS-CoV-2 M^{pro}, we analyzed 27 different data sets to determine crystal structures of M^{pro} in complex with different inhibitors, among which were masitinib, manidipine and bedaquiline[56]. As `positive' controls (*i.e.,* structures that were already known), we considered ligand-free M^{pro} and M^{pro} in complex with the known α-ketoamide covalent reversible inhibitor boceprevir, an approved HCV drug that is also able to bind to SARS-CoV-2 M^{pro} [38]. M^{pro} samples were produced and crystallized in parallel, with very similar experimental procedures, analogous to those of the active enzyme (PDB entry 6y2e; see Section 2)[12]. Almost all tested crystals were monoclinic (space group *C*2, with unit-cell parameters $a \simeq 113.1$, $b \simeq 54.7$, $c \simeq 44.8$ Å, $\alpha = 90.0$, $\beta \simeq 101.3$, $\gamma = 90.0°$), isomorphous to the crystals of the free active enzyme (PDB entry 6y2e)[12] and to most of the deposited M^{pro} structures, signifying the same crystal contacts. After successful molecular replacement and a first round of refinement, in most cases (including the complex with boceprevir) electron density was clearly visible for the entire sequence, indicating a protein matrix with a very similar structure to the search models (PDB entries 6y2e and 5rel)[26]. However, there were a significant number of cases, around ten, in which the electron density was of much lower quality or was even absent in particular portions of the protein, namely residues 139–144 of the oxyanion loop, residues 1–3 of the N-finger and the side chain of His163 in the S1 specificity subsite, all of which are residues that are part of the active site. To cope with the known molecular-replacement bias problem and to correctly rebuild the ambiguous parts, we performed new MR runs using PDB entry 6y2e deprived of residues 139–144 and 1–3, and with an alanine instead of a histidine at position 163 (to remove the His side chain), as a search model. This allowed us to confirm perturbations in the conformation of the selected areas for ten structures, while clear electron density was visible for the remaining cases with the oxyanion loop unambiguously in the active conformation (Supplementary Figs. S2 and S3). In some cases, the electron density was so poor that the tracing of the chain was very problematic, and it was not possible to reliably rebuild the mobile zones entirely (Supplementary Fig. S2*b*). For four structures, it was possible to efficiently model residues

139–144, residues 1–3 and the side chain of His163 in `new' conformations (`new' because there are no equivalents in M^{pro} structures deposited in the PDB) that differ from the active conformations and also from the collapsed-inactive conformations, including PDB entry 2qcy, where the oxyanion loop adopts a 3_{10}-helix conformation (Supplementary Fig. S2*c*). In this regard, comprehensive analyses of the available SARS-CoV and SARS-CoV-2 M^{pro} crystal structures have recently appeared in the literature[32,57–59]. In no case was a conformation analogous to that presented here described, confirming our assessment of a new-inactive state. The most relevant structures discussed here are reported in Supplementary Table S1.

In summary, we found three different conformational states for the oxyanion loop: active (Supplementary Fig. S2*a*), flexible (*i.e.* with poor electron density; Supplementary Fig. S2*b*) and, strikingly, a new-inactive state (Supplementary Fig. S2*c*). A comparison of the known active and collapsed-inactive conformations with the new-inactive conformation presented here is shown in Fig. 1(*b*).

The new-inactive structures were derived solely from crystals obtained using M^{pro} pre-incubated with the inhibitors masitinib, manidipine or bedaquiline, but in no case was electron density indicating the presence of the inhibitors detected. This is explainable by the medium/high IC_{50} (in the range 2.5–19 µ*M*)[56,60] and the very low aqueous solubility of the molecules (when inhibitors in 100% DMSO were added to the protein solution, visible white precipitates appeared). It is tempting to speculate that the presence of these inhibitors in solution plays a role in favoring the selection of the new-inactive conformation by the crystallization process. Some structures of crystals from co-crystallization experiments with masitinib or manidipine, again without any evidence for the presence of the ligand in the binding site, show the oxyanion active conformation. This indicates that these molecules, although favoring the new state, are not strict determinants for its formation. In the free form of the enzyme (from crystallization experiments with no ligands), we obtained structures with very clear electron density for the oxyanion loop, as shown in Supplementary Fig. S2(*a*), with low local *B* factors in the refined model, but also structures with a very `destabilized', mobile oxyanion loop, as in Supplementary Fig. S2(*b*), with much higher *B* factors in the final model. This suggests that

the high flexibility of the oxyanion loop is an intrinsic property of the free enzyme and is not artificially induced by the presence of ligands in the crystallization experiments.

Here, we describe only one of the structures of M^{pro} determined in the new-inactive conformation, which was obtained by co-crystallization experiments with masitinib (no relevant differences exist among the four new-inactive M^{pro} structures). Data-collection and final model statistics are reported in Table 1; final electron densities for the most relevant regions discussed in the text are shown in Fig. 2. Unlike in other inactive structures of the enzyme, in which only one protomer adopts the inactive conformation, the dimeric arrangement of the new structure is due to a crystallographic symmetric axis, and the two subunits are therefore identical and both inactive.

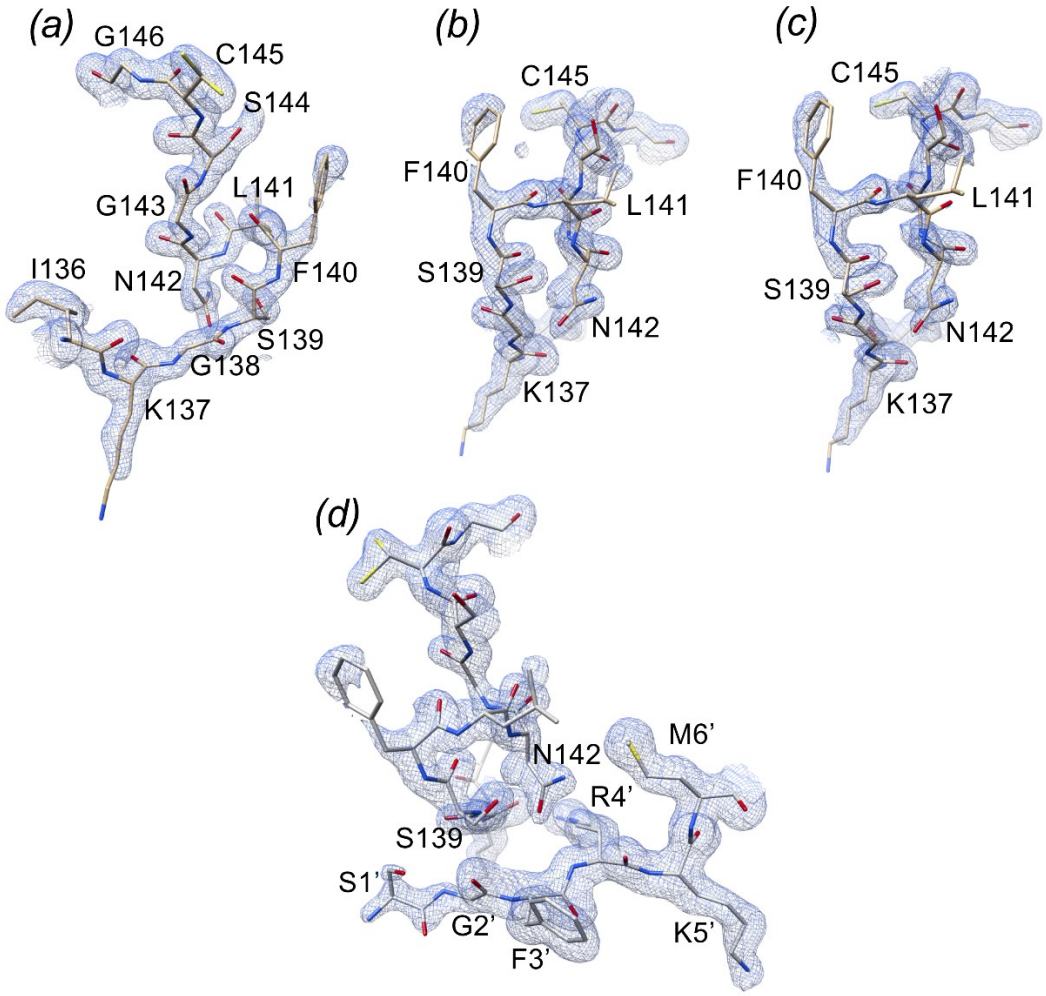


Figure 2. Final electron densities for the most relevant regions of new-inactive M^{pro}. $2F_O$-F_C maps contoured at 1.0 σ level are shown. *(a)* and *(b)* show two views of the final electron density for the oxyanion loop in the new conformation. Leu141 and the solvent-exposed Phe140 and Lys137 side chains have incomplete densities indicating various degrees of flexibility. *(c)*, simulated-annealing omit map (oxyanion loop residues 138-146 were omitted), viewed as in *(b)*. *(d)* Electron density in the inter-protomers (intra-dimer) interaction area between the oxyanion loop of one protomer and the N-finger of the other protomer (residues Ser1´-Met6´).

3.2. The oxyanion loop adopts a novel inactive conformation

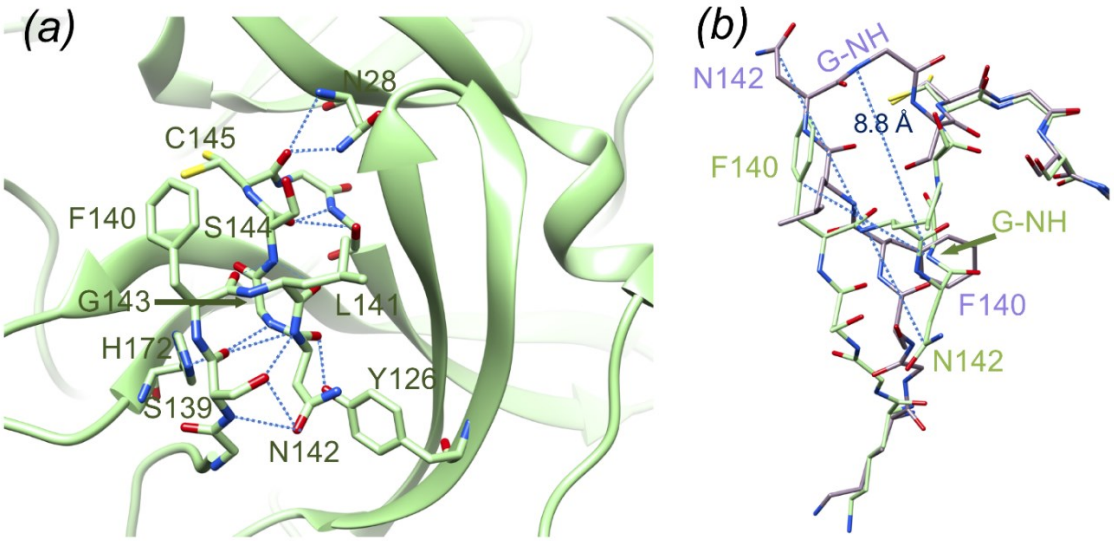


Figure 3. Details of the hydrogen-bond interactions in the oxyanion region of new-inactive M^{pro}. (*a*) The new conformation of the oxyanion loop is stabilized by several backbone hydrogen bonds (blue dashed lines) as described in the main text. The side chain of catalytic Cys145 has a double conformation. (*b*) Comparison between the new-inactive (green) and active (light magenta; PDB entry 6y2e) oxyanion loops. There are large movements (blue dashed lines) of the side chains of Asn142 and Phe140. In the new-inactive conformation, Asn142 moves from an exposed position with an ASA of 153.74 $Å^2$ to a buried position with an ASA of 49.00 $Å^2$ and Phe140 moves from a buried position with an ASA of 14.79 $Å^2$ to an exposed position with an ASA of 143.29 $Å^2$. Gly143 NH (G-NH) of the oxyanion hole, which is involved in the stabilization of the tetrahedral intermediate, moves 8.8 Å away.

The most striking property of the new structure is the significantly different conformational state of the oxyanion loop (Figs. 1 and 3), which is essential for stabilization of the tetrahedral acyl (oxyanion) transition state during the catalytic cycle. The loop backbone is stabilized by many hydrogen bonds in the new state (Fig. 3*a*).

According to the *DSSP* standardized secondary-structure assignment[42,43], in the new oxyanion loop there are two consecutive `3-turns' (β-turns) with hydrogen bonds between Leu141 CO and Ser144 NH and between Ser144 CO and Ser147 NH. This region is further stabilized by a `4-turn' (α-turn) with a hydrogen bond between Ser139 CO and Gly143 NH. *DSSP* does not recognize any 3_{10}-helical segments in the oxyanion loop (as present in the inactive PDB entry 2qcy).

There are other hydrogen bonds involving the backbone that stiffen the oxyanion loop: between Cys145 CO and Asn28 NH, between His163 CO and Gly146 NH and between Ser147 CO and His163 NH (Fig. 3*a*). As a result, the new conformation appears to be quite stable and rigid, as confirmed by the good quality of the local electron density (Fig. 2 and Supplementary Fig. S2*c*).

To analyze the energetics of the local contacts, we performed an energetic frustration analysis[44] on the active and new-inactive conformations. The concept of local frustration in protein structure refers to possible residual energetic conflicts in local interactions in folded proteins, using a `frustration index' that measures how favorable a particular contact is relative to the set of all possible contacts in that location[61]. The `principle of minimal frustration' assumes that proteins find their native state by minimizing the internal energetic conflicts within their polypeptide chain[62]. The degree of frustration is therefore dependent on the type of amino acids involved in the interaction. Local violations of this principle have been recognized to be important to exert the proper biological functions, specifically around the active sites of protein enzymes[63]. Analysis of the local configurational frustration of the most interesting contacts around the active site of active and new-inactive M^{pro} is shown in Supplementary Table S2. In both conformations, the catalytic Cys145 is a minimally frustrated `hub' (here we call a position with ≥10 minimally frustrated interactions a minimally frustrated hub), with a small prevalence of interactions in the active conformation. On the other hand, the difference for Phe140 is striking: eight minimally frustrated interactions are present in active M^{pro} (where it is buried in a hydrophobic pocket) as opposed to no interactions in new-inactive M^{pro} (where it is solvent-exposed). Differences between the two structures are also evident for other amino acids of the oxyanion loop, namely Leu141, Gly143 and Ser144,

indicating their diverse involvement in the local energetic contributions. The oxyanion loop of inactive M^{pro} has a larger number of minimally frustrated interactions with Cys117. This residue is a minimally frustrated hub in both conformations; however, given the higher number of minimally frustrated interactions in new-inactive M^{pro} (18 versus ten), Cys117 seems to play an important role in the stabilization of the new-inactive conformation. Internal to the oxyanion loop there is also a highly frustrated (unfavorable) interaction involving Leu141, with Ser139 in new-inactive M^{pro} and with Ser144 in active M^{pro}. This suggests that Leu141 may be important in switching between the two conformations.

3.3. Many key interactions of the active enzyme are lost in new-inactive M^{pro}

The correct location of Phe140, Leu141, Asn142, Ser144, Tyr161, His163, Met165, Glu166 and His172 (as seen in the active PDB entry 6y2e, for instance) is an absolute requirement for the reaction catalyzed by M^{pro} to properly proceed, with special reference to stabilization of the tetrahedral acyl-intermediate[4,23,25]. Notably, all of these residues are conserved among known coronaviral M^{pro}s, underlining their importance. In the new structure of M^{pro} most of these residues move away from the `active location': Phe140, Leu141, Asn142 and Ser144 because of displacement of the oxyanion loop (Fig. 3*b*) and His163 and His172 because of rotation of their side chains (Fig. 4).

Specifically, Asn142 C^{α} and the side chain of Phe140 are remarkably shifted from the active position by 9.8 and 7.5 Å, respectively (Fig. 3*b*). Phe140, which is buried in a hydrophobic cleft in active M^{pro} with as accessible surface area (ASA) of 14.79 $Å^2$), is now exposed to the solvent (ASA 143.29 $Å^2$), while Asn142, which is exposed in active M^{pro} (ASA 153.74 $Å^2$), is now buried (ASA 49.00 $Å^2$). The side chain of Asn142 is locked in the new position by hydrogen bonds to the side-chain O^{γ} and backbone NH of Ser139. Markedly, the oxyanion hole Gly143 NH, the correct positioning of which is essential for the stabilization of the tetrahedral oxyanion intermediate during catalysis, is moved 8.8 Å away.

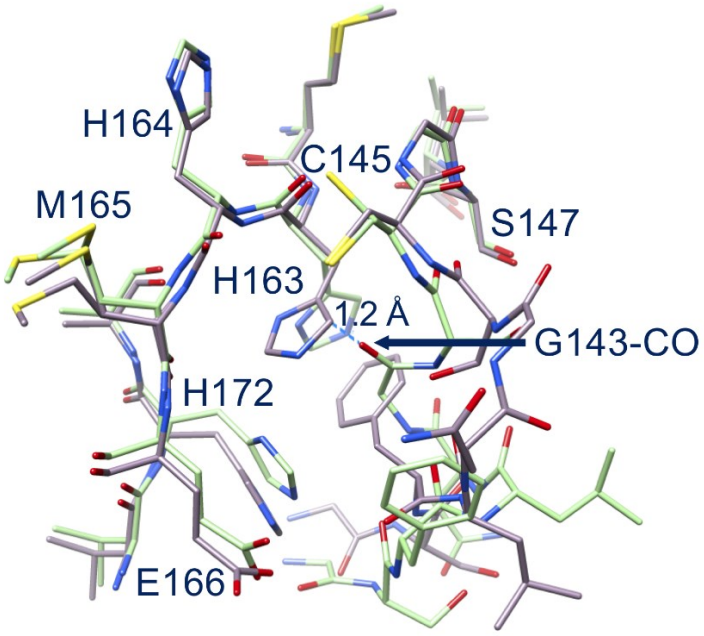


Figure 4. Comparison between new-inactive (green) and active (light magenta) M^{pro}. In the new structure the side chain of His163 rotates away to avoid steric clashes with the oxyanion loop: in the active conformation (PDB-ID 6Y2E) the His163 side chain would be at 1.2 Å from the new position of Gly143-CO. Note also the movement of His172.

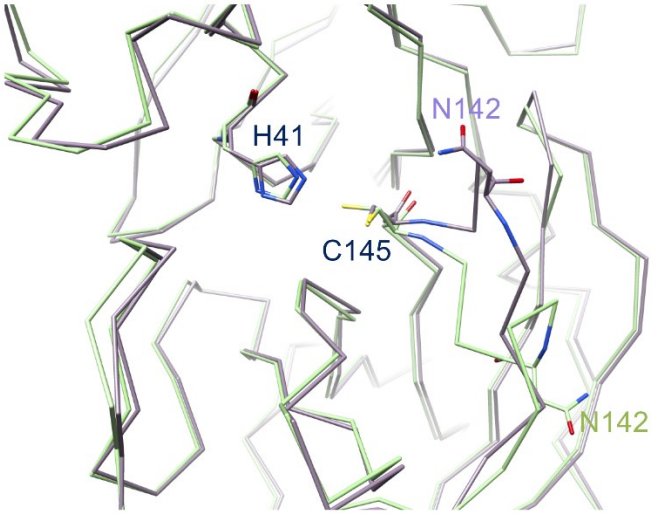


Figure 5. Catalytic dyad. In new-inactive M^{pro} (green) the position of the catalytic dyad His41 and Cys145 is similar to that in the active enzyme (6Y2E, light magenta), despite the large shift of residues 138-144. In new-inactive M^{pro} Cys145 adopts a double conformation.

As a consequence, many interactions that are recognized to be important for stabilization of the active conformation are lost, namely hydrogen bonds between Glu166 and His172 and between Tyr161 and His163, as well as the aromatic stacking between His163 and Phe140[64]. The rotation of the side chain of His163 (located at the very bottom of the S1 subsite), the hydrogen-bond properties of which seem to be very important in determining both substrate specificity and proper inhibitor binding[65], is a noteworthy characteristic of this new conformation of M^{pro}. His163 is no longer available for substrate binding as it rotates away to avoid steric clashes with Gly143 CO (Fig. 4). Its position is now `functionally' occupied by His172, which moves towards the S1 subsite (Fig. 4). The other three important residues, Tyr161, Met165 and Glu166, essentially maintain the same position as adopted in active M^{pro}. Despite the large displacement of the oxyanion loop, the position of the catalytic dyad His41 and Cys145 is not significantly altered, especially in the backbone, even though the Cys145 side chain now shows a double conformation (Fig. 5). The conserved water molecule near His41 is still present in the same position, making hydrogen bonds to the side chains of His41, His164 and Asp187 as in active SARS-CoV-2 M^{pro}.

3.4. The N-finger, the C-terminal tail and the dimeric interface are perturbed in new-inactive M^{pro}

In new-inactive M^{pro}, the dimeric interface is altered compared with that of the active conformation. *PISA* analysis of the interface shows that in new-inactive M^{pro} the interface area is reduced (from 1661 to 1273 $Å^2$), as are the number of hydrogen bonds (from 33 to six) and the number of salt bridges (from 12 to six). However, structural features that are important for stabilization of the dimeric form are essentially conserved, namely (i) the salt bridge between Glu290 of one protomer and Arg4' of the other[4], (ii) the hydrophobic aromatic interaction between Tyr126 and Met6' [66] and (iii) the interaction of Arg298 with the N-finger and the C-terminus[30]. This suggests that although new-inactive M^{pro} is still able to form dimers, the dimeric state is less stable compared with that of active M^{pro}.

At the dimeric interface, relevant changes in both the N- and C-termini are present. In active M^{pro}, the N-finger of one protomer interacts and stabilizes the S1 subsite of the

other protomer[23]. For instance, in active SARS-CoV-2 M^{pro} (PSB entry 6y2e) Ser1 of one protomer is hydrogen-bonded both to the carboxylate group of Glu166 and to the main chain of Phe140 of the other protomer. In the new-inactive structure, these interactions are lost as a consequence of the different oxyanion conformation of one protomer that `pushes away' residues 1–3 of the N-finger of the other protomer (Fig. 6), with Gly2′ CO now at 3.2 Å from Ser139 NH. The rearrangement of the oxyanion loop of one protomer also influences the C-terminal tail of the other protomer, the electron density of which is no longer visible from residue 301 onwards, indicating high flexibility (Figs. 6*b* and 7). Among the residues of the oxyanion loop, Leu141 shows major changes at the level of the dimeric interface (Fig. 7*b*), also causing rotation of the side chain of Tyr118 to avoid steric clashes, further supporting its possible central role in switching between the new-inactive and active conformations.

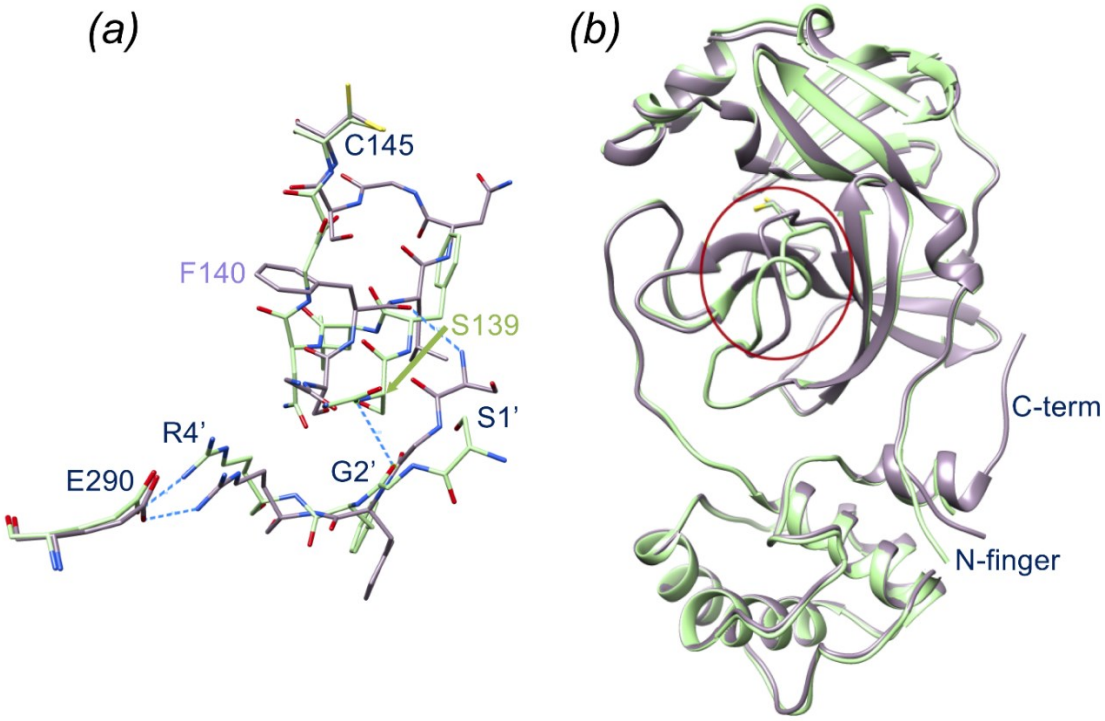


Figure 6. Displacements at the intra-protomer interface. New-inactive M^{pro} is in green and active M^{pro} is in magenta. (*a*) The new oxyanion loop of one protomer pushes away residues 1′–3′ of the other protomer; however, the key salt bridge between Arg4′ and Glu290, which is important for dimer stabilization, is conserved. (*b*) Overall superposition of active and new-inactive M^{pro} shows that besides those in the oxyanion loop (red ellipsoid), major differences are located in the N-finger and in the C-terminal tail, which is not visible in new-inactive M^{pro}.

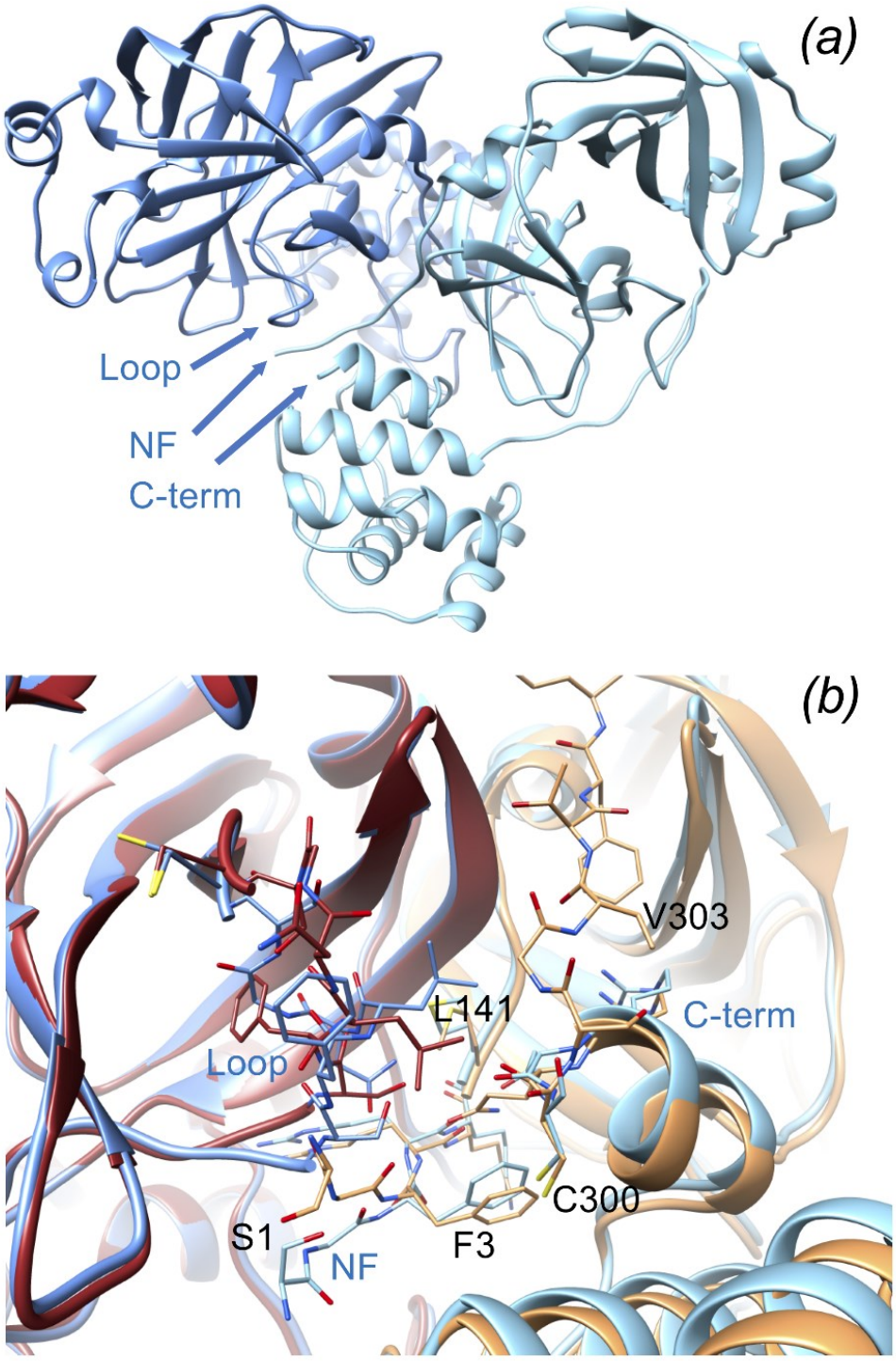


Figure 7. Dimeric architecture of new-inactive M^{pro}. (*a*) The new conformation of the oxyanion loop (labeled `loop') causes changes in the interface between protomer *A* (blue) and protomer *B* (light blue) at the level of the N-finger (labelled `NF') and the C-terminal tail (labeled `C-term'). (*b*) Local differences between the new structure [blue-based colors as in (*a*)] and the canonical structure (PDB entry 6y2e; brown-based colors, with intact C-terminus): the shift of the Leu141 side chain seems to have major effects in destabilizing the C-terminal tail of the new structure.

3.5. New-inactive M^{pro} can still bind substrates

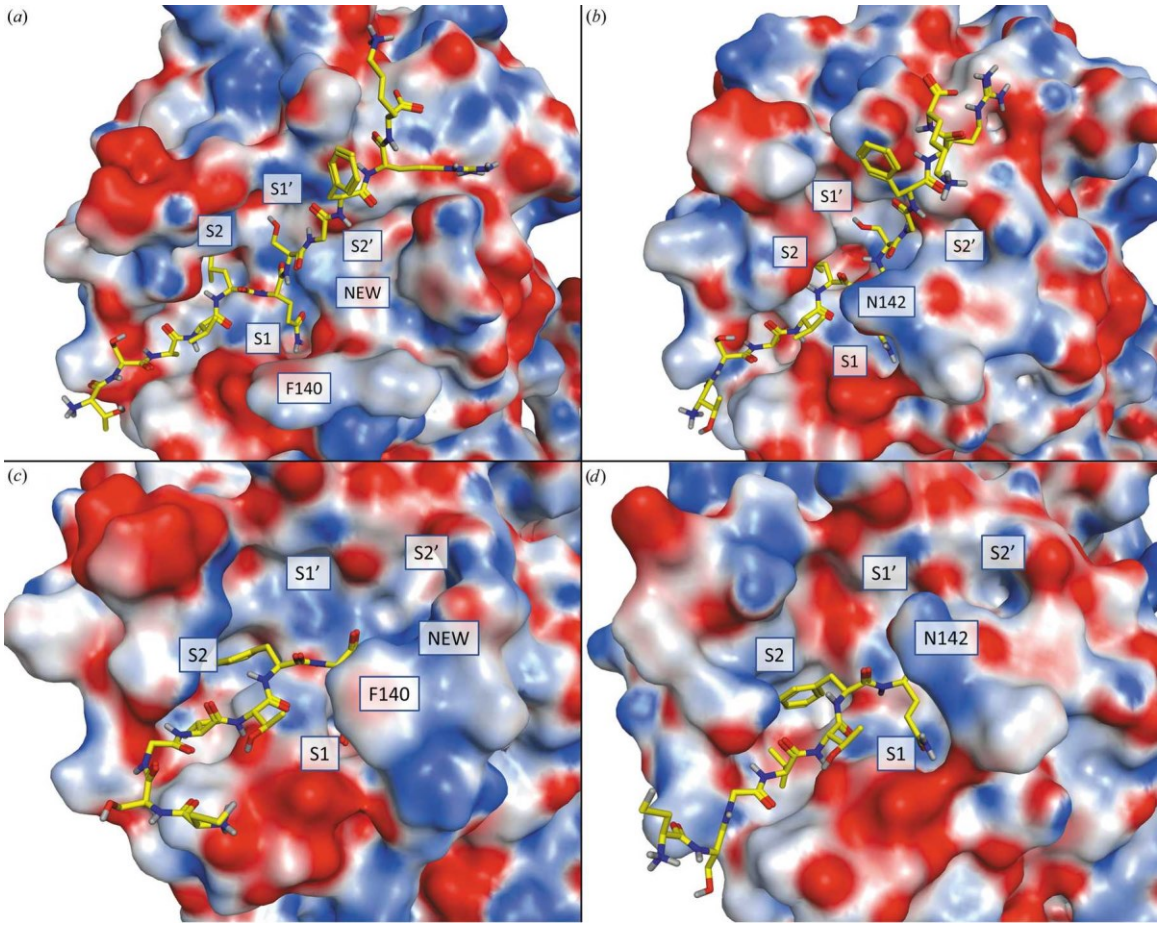


Figure 8. Reshaping of the S1 and S2′ subsites. Molecular-dynamics modeling of the hypothetical interaction of new-inactive M^{pro} with substrates is shown. Top, putative interaction with the 11-mer pseudo-substrate peptide from PDB entry 2q6g: (*a*) new-inactive M^{pro}, (*b*) SARS-CoV M^{pro} from PDB entry 2q6g. Bottom, putative interaction with the acyl-intermediate of the M^{pro} C-terminal autoprocessing site: (*c*) new-inactive M^{pro}, (*d*) M^{pro} in PDB entry 7khp. As a result of the rearrangement of the oxyanion loop, a new cavity near the S2′ site, labeled `NEW', is formed.

Having established that the new structure is catalytically incompetent, we tried to understand whether it is still able to bind natural substrates. Superposition of the new-inactive conformation with either the active conformation in complex with the C-terminal acyl-intermediate (PDB entry 7khp)[25] or the SARS-CoV M^{pro} active conformation in complex with its 11-mer substrate complex (PDB entry 2q6g)[15] does not show evident steric clashes for the substrate. This is also valid for superposition of the new-inactive conformation with two recent complexes between SARS-CoV-2 M^{pro} and two peptide substrates corresponding to the nsp4/5[67] and nsp8/9[68] cleavage sites. Additionally, a short

molecular-dynamics refinement of the complexes of the new-inactive conformation of SARS-CoV-2 M^{pro} with either the C-terminal acyl-intermediate or the 11-mer peptide substrate reveal compatible binding modes, with only minor side-chain rearrangements (Fig. 8). The reshaped S1 site of the new-inactive M^{pro} could still host a P1 glutamine, although the rearrangement causes the loss of its interactions with Glu166 O^{ϵ} and Phe140 CO in favor of a single hydrogen bond to Gly143 CO (Fig. 9). Aside from the alterations of the S1 subsite, which alter the recognition profile of the P1 glutamine, the other interaction features are retained, namely the hydrogen bonds to Glu166 and Gln189 and the hydrophobic interactions of the P2 phenylalanine within the S2 subpocket. This is a quite remarkable observation because it suggests that the new conformation could be inactive not necessarily because it is incapable of recognizing the substrate, but because the catalytic machinery is not properly organized for an efficient catalytic event, particularly in the oxyanion-hole region, and is unable to stabilize the tetrahedral acyl intermediate. The new conformation of the oxyanion loop generates a new cavity near position S2', as evident from comparison of the new structure with the SARS-CoV-2 acyl-enzyme (PDB entry 7khp)[25] and the SARS-CoV 11-mer substrate complex (PDB entry 2q6g)[15] (Fig. 8).

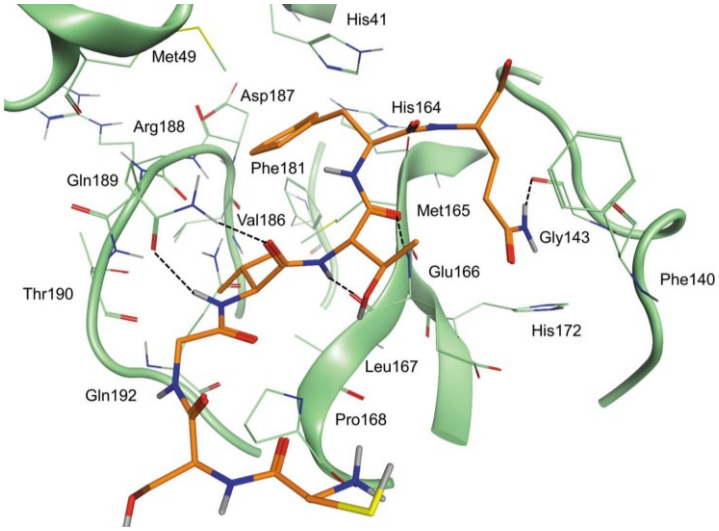


Figure 9. Details of the putative interaction between new-inactive M^{pro} (green) and the C-terminal acyl-intermediate peptide substrate from PDB entry 7khp (orange). Hydrogen bonds between the substrate and the binding site are depicted as dashed black lines. Aside from the P1 glutamine and its interactions with the P1 pocket, other common interaction features such as hydrogen bonds to Glu166 and Gln189 and hydrophobic interactions of the P2 phenylalanine side chain within the S2 subpocket are retained.

3.6. The new-inactive conformation is stable and in equilibrium with the active conformation in solution

For SARS-CoV M^{pro}, it has been shown that the active-site loops are very dynamic and sensitive to variations in the environmental conditions[13,15,29,69,70]. Similarly, the oxyanion loop of SARS-CoV-2 M^{pro} showed conformational flexibility as deduced from room-temperature X-ray crystallography[71,72]. To test the stability and to model the dynamics of new-inactive M^{pro}, specifically of the oxyanion loop and regions involved in substrate binding, we performed crystallographic ensemble refinement[41] and MD simulations.

The 60 structures generated by ensemble refinement of new-inactive M^{pro} compatible with the crystallographic restraints confirm the new conformation of the oxyanion loop and reveal that its flexibility is comparable to that of other portions of the substrate-binding region (residues 43–51 in domain I and residues 188–198 in the flexible linker connecting domains II and III; Fig. 10), as also found in the literature. In four out of 60 structures the oxyanion-loop conformation is similar to that in the active form, which is in line with the experimental observation of a residual electron density compatible with the presence of a small fraction of the oxyanion loop and of the side chain of His163 in the active conformation in the crystal state. In this respect, all structures determined here, including new-inactive M^{pro}, were obtained from batches of correctly autoprocessed protein (*i.e.* catalytically active towards itself at the N-terminus) which displayed normal catalytic activity in solution towards substrate peptides.

This strongly suggests the presence of a dynamic equilibrium in solution with the coexistence of different conformations, including inactive conformations. In other words, exhibition of the correct catalytic activity on the macroscopic level (with the full ensemble of conformational states available in solution for M^{pro}) does not contrast with the possibility of selection by the crystallization process (in this case probably favored by the presence of certain small molecules) of a subpopulation of a catalytically incompetent form of the enzyme as shown here and for the previous structure with PDB code 1uj1. The conclusion that the dynamic equilibrium in solution includes both the active and the new-inactive conformation is supported by comparing the results of ensemble refinement of

the structure in the free state with very poor electron density for the oxyanion loop (Supplementary Fig. S2*b*). The refined ensemble conformations show a highly dynamic oxyanion loop, with 20% of conformations similar to the active conformation, 23% of conformations similar to the inactive conformation and 57% of conformations in intermediate states.

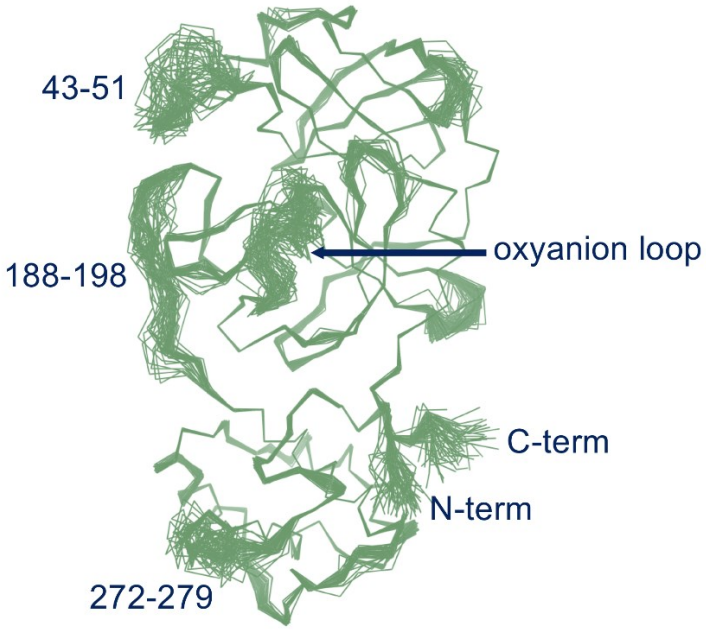


Figure 10. Ensemble refinement. The 60 structures generated by ensemble refinement highlight the mobile regions of new-inactive M^{pro}. The oxyanion loop, which is confirmed in the new conformation, has a flexibility similar to those of residues 43–51 and 188–198 involved in substrate recognition as the S3 and S4 sites.

To assess the structural stability of the new-inactive conformation of SARS-CoV-2 M^{pro} and to compare it with the active conformation, three independent 1 µs classical molecular-dynamics simulations were performed for both conformations. For the active state, PDB entry 6y2e was taken as a reference. As depicted in Fig. 11, which summarizes the principal geometric analysis performed along the MD trajectories, the two structures show a similar degree of stability. The backbone r.m.s.d. profile for PDB entry 7nij (Fig. 11*b*), representing the new-inactive conformation of M^{pro}, displays moderately higher fluctuations with respect to the active state (Fig. 11*a*). As can be seen in the per-residue r.m.s.f. plots (Figs. 11*c* and 11*d*), this difference can mainly be attributed to major

structural fluctuations in the same regions that were marked as flexible by the crystallographic data, namely the three flexible loops 43–51, 188–198 and 272–279 and the C-terminus (299–306), while the rest of the structure is quite stiff, as in the active state. Specifically, the C-terminus in the new-inactive conformation of M^{pro} shows the highest amplitude of movement, as denoted by the high r.m.s.f. values associated with these residues. This result agrees with the absence of electron density for residues 301–306, which indicates high flexibility of this region. Instead, the N-terminus (residues 1–4) shows more limited fluctuations for both M^{pro} conformations, which is in agreement with the presence of well-defined electron density in both structures. The overall structural stability of the new-inactive conformation of M^{pro} is also confirmed by the time-dependent evolution of both secondary-structure elements and the protein radius of gyration (R_g), with only minor oscillations, similar to those seen in the active conformation (Supplementary Figs. S4, S5 and S6). Despite the slightly higher fluctuations observed in the inactive conformation, no sufficient motions were observed to shed light on a possible transition mechanism between the two conformations. It is not surprising that such rearrangement was not sampled even on a 1 µs scale, since such collective motions in proteins usually involve longer timescales (*i.e.* millisecond to microsecond)[73].

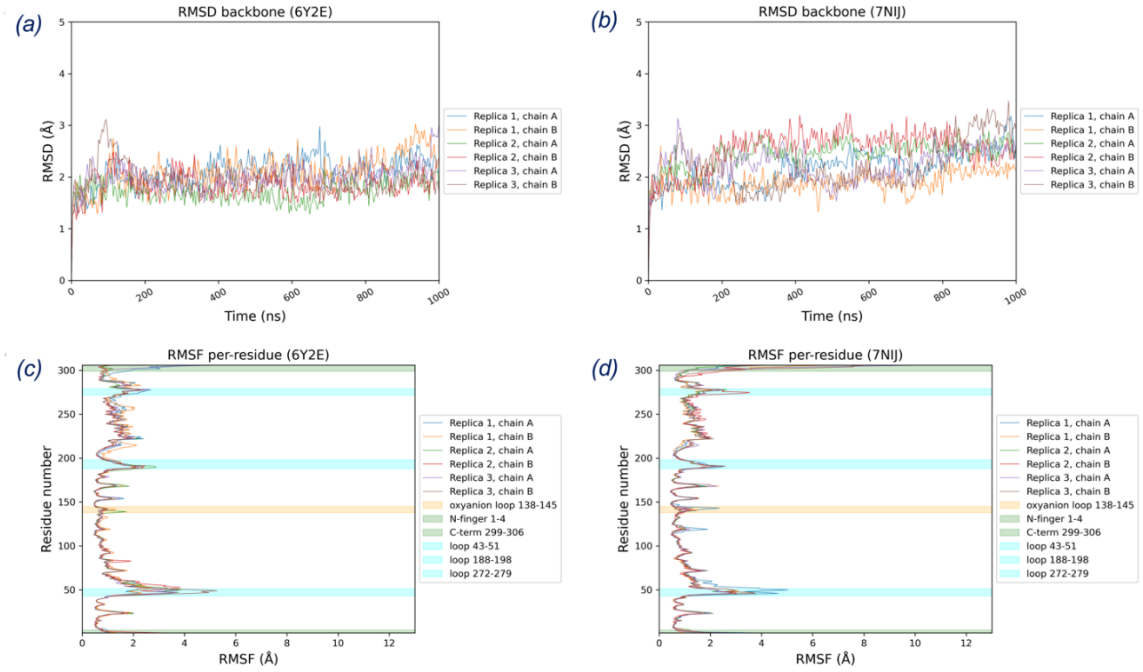


Figure 11. Results of MD simulations. Summary of the key geometric analysis performed along the MD trajectories for both the active (PDB entry 6y2e) and new-inactive (PDB entry 7nij) conformations of SARS-CoV-2 M^{pro}. (*a*) and (*b*) highlight the time-dependent variation of the protein root-mean-square deviation (r.m.s.d.) of C^{α} atomic positions for PDB entries 6y2e and 7nij, respectively. (*c*) and (*d*) summarize the per-residue mean root-mean-square fluctuation (r.m.s.f.) of atomic positions of protein C^{α} atoms for PDB entries 6y2e and 7nij, respectively. The most relevant regions of the protein are highlighted in the plot for visualization clarity as described in the legend. For both r.m.s.d. and r.m.s.f. analyses, each chain composing the crystallographic dimer is considered separately.

4. Discussion

We had the opportunity to capture a new and stable (as seen in MD simulations) inactive state of M^{pro}, called new-inactive, expanding the knowledge of the conformational space accessible to the enzyme. Altogether, the movements in the substrate-binding region and near the catalytic site result in a significant reshaping of the reaction center (Figs. 3, 4 and 8) that has never previously been observed and is much more pronounced than in the previously described collapsed-inactive M^{pro} conformation. The conformation adopted by residues 139–144 of the oxyanion loop is potentially catalytically incompetent. The backbones of key residues in the oxyanion hole are 8–10 Å away from the catalytically competent position. Fundamental interactions for the proper function of the enzyme are broken or absent, as illustrated in the previous section. Among the residues of the oxyanion loop, Phe140, Leu141 and Asn142 play a major role in the shift between the new-inactive and active conformations. The new state of the oxyanion loop of one protomer pushes the N-finger of the second protomer away from the position adopted in the active enzyme. The last six residues of the C-terminal tail are not visible in the electron-density map and were confirmed to be fully flexible by MD simulations. The novel conformations of the oxyanion loop and of the N- and C-termini result in a weakening of the dimeric architecture, as shown by decreases in the interaction surface area and in the number of inter-protomer interactions. Major variations in the dimeric interface are connected to Leu141 of the oxyanion loop.

This new structure is relevant for the analysis of the M^{pro} catalytic cycle, which was recently investigated using biodynamics theory under non-equilibrium conditions[74], using

the available crystal structures, which show M^{pro} in different conformational states[75]. This novel approach tries to mimic *in vivo* conditions, which depend on non-equilibrium structure–kinetics relationships. From this analysis a substrate-induced M^{pro} activation mechanism was developed, suggesting the existence of a complex substrate-binding activation mechanism in both SARS-CoV and SARS-CoV-2. The proposed catalytic cycle involves transition from the collapsed-inactive conformation of the oxyanion loop, represented by the free form of monomeric M^{pro} (PDB entry 2qcy), to the putative substrate-bound form of monomeric M^{pro}, represented by one monomer of PDB entry 2q6g (with an active oxyanion loop), and finally to the dimeric fully active state, represented by dimeric M^{pro} (PDB entry 6m03; very similar to PDB entry 6y2e). The new-inactive structure presented here shows a new conformational state with an accessible oxyanion loop, adding novel important pieces of information to the structural dynamics of the substrate-induced activation of M^{pro} in the context of its catalytic cycle. In the non-equilibrium model, it was hypothesized that transition of the oxyanion loop from the inactive to the active conformation is triggered mainly by solvation/desolvation effects. This also applies to transitions involving our new-inactive structure, where, for activation, Phe140 moves from an exposed position (with no minimally frustrated interactions) to a buried position (with eight minimally frustrated interactions), while Asn142 moves from a buried position to an exposed position. In the context of the conformational dynamics of M^{pro}, the intriguing possibility exists that the remodeling of the S2′ subsite can be correlated with the large amino-acid variation in position P2′ of SARS coronaviral nonstructural protein (nsp) cleavage sites, M^{pro} autoprocessing included. Despite being catalytically incompetent, this new state (with a novel cavity in position S2′) seems to be able to bind natural substrates of M^{pro} (see Figs. 8 and 9). Among the 11 substrates of SARS-CoV-2 M^{pro}, position P2′ is highly variable, hosting nine different amino acids with very different chemical and structural properties: small, such as Gly and Ala, bulky hydrophobic, such as Ile, Val and Leu, positively charged, such as Lys, negatively charged, such as Glu, and polar and hydrogen-bond donor/acceptor, such as Ser and Asn. It is conceivable that the flexibility of the oxyanion-loop conformation is correlated to this variability of the substrates, specifically in position P2′, and to the necessity to accommodate the different substrates during the maturation process of the pp1a and

pp1ab polypeptides, in the correct succession of proteolytic events. We suggest that this new conformational state is that preferred by the enzyme to efficiently host substrates with bulky hydrophobic residues in position P2', for instance for the processing of nsp7/8 (Ile), nsp12/13 (Val) and nsp14/15 (Leu) cleavage sites. According to the M^{pro} reaction scheme proposed by Wan *et al.*[75], the substrate-binding event triggers the conformational switch of the oxyanion loop, which adopts the necessary conformation for a productive catalytic event. Overall, the following scheme can be proposed: (i) for the initial binding, specific substrates (with bulky residues in position P2') select the new-inactive conformation among a complex ensemble of different conformations of M^{pro} in mutual equilibrium, (ii) the binding event causes conformational changes of the oxyanion loop and, mainly, of the side chains of Glu166, His172 and His163, (iii) the dimeric architecture is stabilized because of rearrangements of the N-finger and the C-terminus and (iv) the resulting activated enzyme is ready to properly hydrolyze the substrate.

The new-inactive structure is also important for the structure-based drug-discovery process that is currently being applied to M^{pro} [65]. The approach of `repurposing' already known drugs via classical docking methodologies on the 3D structure of the protein target is interesting because, methodologically, it is potentially fast and the safety profiles of the tested compounds are already known. This justifies the large amount of research devoted to repurposing known antiviral drugs against M^{pro} [76]. Obviously, the success rate of these campaigns would greatly benefit from the possibility of targeting significantly different, stable, conformations. In this respect, the discovery of the new stable inactive conformation of M^{pro} presented here, with the remodeling of the S1 subsite and the formation of the nearby new cavity near subsite S2' (poorly explored until now as known inhibitors usually span the enzyme S1–S4 subsites), offers solid attractive possibilities for the design of completely new classes of antiviral drugs targeting M^{pro}. Indeed, a putative binder of the new-inactive form could reduce the population of the active conformation by stabilizing the inactive conformation. Also, a ligand able to bind the novel, readapted site around the catalytic cysteine could sterically hamper the recognition of the substrate. In addition, the possibility of targeting a novel subpocket could increase the affinity by establishing novel contacts and interactions. Most of the more promising M^{pro} inhibitors

were developed by optimizing starting hits that were further decorated to explore the subpockets located around the catalytic center, following the classic route of fragment maturation in fragment-based lead discovery[77]. One notable example is represented by the optimization of portions of parampanel on S1 and S1' and its engagement of S3–S4, which lead to a fourfold boost in IC_{50} activity[78].

In conclusion, the new-inactive structure of M^{pro} is relevant for better understanding of the function and mechanism of action of this fundamental enzyme for SARS-CoV-2 replication in the cell, with a particular accent on the dynamics within the catalytic cycle of the enzyme, which explores different conformational states including that presented here for the first time. Further, the discovery of this unprecedented inactive conformation of M^{pro} provides a unique opportunity for the more successful design of antiviral drugs with improved pharmacological properties using both classical docking-based and innovative non-equilibrium-based approaches.

References

1. Pol, S. & Corouge, M. Treatment of hepatitis C: Perspectives. *Med Mal Infect* **44**, 449–454 (2014).

2. Skwarecki, A. S., Nowak, M. G. & Milewska, M. J. Amino Acid and Peptide-Based Antiviral Agents. *ChemMedChem* **16**, 3106–3135 (2021).

3. Xia, B. & Kang, X. Activation and maturation of SARS-CoV main protease. *Protein Cell* **2**, 282–290 (2011).

4. Anand, K. *et al.* Structure of coronavirus main proteinase reveals combination of a chymotrypsin fold with an extra alpha-helical domain. *EMBO J* **21**, 3213–3224 (2002).

5. Snijder, E. J., Decroly, E. & Ziebuhr, J. The Nonstructural Proteins Directing Coronavirus RNA Synthesis and Processing. *Adv Virus Res* **96**, 59–126 (2016).

6. Dai, W. *et al.* Structure-based design of antiviral drug candidates targeting the SARS-CoV-2 main protease. *Science (1979)* **368**, 1331–1335 (2020).

7. Günther, S. *et al.* X-ray screening identifies active site and allosteric inhibitors of SARS-CoV-2 main protease. *Science* **372**, 642–646 (2021).

8. Ullrich, S. & Nitsche, C. The SARS-CoV-2 main protease as drug target. *Bioorg Med Chem Lett* **30**, 127377 (2020).

9. Pavan, M., Bolcato, G., Bassani, D., Sturlese, M. & Moro, S. Supervised Molecular Dynamics (SuMD) Insights into the mechanism of action of SARS-CoV-2 main protease inhibitor PF-07321332. *J Enzyme Inhib Med Chem* **36**, 1646–1650 (2021).

10. Ahmad, B., Batool, M., Ain, Q. U., Kim, M. S. & Choi, S. Exploring the Binding Mechanism of PF-07321332 SARS-CoV-2 Protease Inhibitor through Molecular Dynamics and Binding Free Energy Simulations. *International Journal of Molecular Sciences 2021, Vol. 22, Page 9124* **22**, 9124 (2021).

11. Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature 2020 579:7798* **579**, 265–269 (2020).

12. Zhang, L. *et al.* Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved α-ketoamide inhibitors. *Science (1979)* **368**, 409–412 (2020).

13. Tan, J. *et al.* pH-dependent conformational flexibility of the SARS-CoV main proteinase (M(pro)) dimer: molecular dynamics simulations and multiple X-ray structure analyses. *J Mol Biol* **354**, 25–40 (2005).

14. Anand, K., Ziebuhr, J., Wadhwani, P., Mesters, J. R. & Hilgenfeld, R. Coronavirus main proteinase (3CLpro) structure: basis for design of anti-SARS drugs. *Science* **300**, 1763–1767 (2003).

15. Xue, X. *et al.* Structures of two coronavirus main proteases: implications for substrate binding and antiviral drug design. *J Virol* **82**, 2515–2527 (2008).

16. Ho, B. L. *et al.* Critical Assessment of the Important Residues Involved in the Dimerization and Catalysis of MERS Coronavirus Main Protease. *PLoS One* **10**, (2015).

17. Chen, H. *et al.* Only one protomer is active in the dimer of SARS 3C-like proteinase. *Journal of Biological Chemistry* **281**, 13894–13898 (2006).

18. El-Baba, T. J. *et al.* Allosteric Inhibition of the SARS-CoV-2 Main Protease: Insights from Mass Spectrometry Based Assays*. *Angew Chem Int Ed Engl* **59**, 23544–23548 (2020).

19. Hsu, M. F. *et al.* Mechanism of the maturation process of SARS-CoV 3CL protease. *J Biol Chem* **280**, 31257–31266 (2005).

20. Li, C. *et al.* Conformational Flexibility of a Short Loop near the Active Site of the SARS-3CLpro is Essential to Maintain Catalytic Activity. *Sci Rep* **6**, (2016).

21. Kneller, D. W. *et al.* Malleability of the SARS-CoV-2 3CL Mpro Active-Site Cavity Facilitates Binding of Clinical Antivirals. *Structure* **28**, 1313-1320.e3 (2020).

22. Shi, J. & Song, J. The catalysis of the SARS 3C-like protease is under extensive regulation by its extra domain. *FEBS J* **273**, 1035–1045 (2006).

23. Verschueren, K. H. G. *et al.* A structural view of the inactivation of the SARS coronavirus main proteinase by benzotriazole esters. *Chem Biol* **15**, 597–606 (2008).

24. Perry A. Frey, A. D. H. *Enzymatic Reaction Mechanisms*. (Oxford University Press, 2007).

25. Lee, J. *et al.* Crystallographic structure of wild-type SARS-CoV-2 main protease acyl-enzyme intermediate with physiological C-terminal autoprocessing site. *Nat Commun* **11**, (2020).

26. Douangamath, A. *et al.* Crystallographic and electrophilic fragment screening of the SARS-CoV-2 main protease. *Nat Commun* **11**, (2020).

27. Jin, Z. *et al.* Structure of Mpro from SARS-CoV-2 and discovery of its inhibitors. *Nature* **582**, 289–293 (2020).

28. Jin, Z. *et al.* Structural basis for the inhibition of SARS-CoV-2 main protease by antineoplastic drug carmofur. *Nat Struct Mol Biol* **27**, 529–532 (2020).

29. Yang, H. *et al.* The crystal structures of severe acute respiratory syndrome virus main protease and its complex with an inhibitor. *Proc Natl Acad Sci U S A* **100**, 13190–13195 (2003).

30. Shi, J., Sivaraman, J. & Song, J. Mechanism for controlling the dimer-monomer switch and coupling dimerization to catalysis of the severe acute respiratory syndrome coronavirus 3C-like protease. *J Virol* **82**, 4620–4629 (2008).

31. Allaire, M., Chernaia, M. M., Malcolm, B. A. & James, M. N. G. Picornaviral 3C cysteine proteinases have a fold similar to chymotrypsin-like serine proteinases. *Nature 1994 369:6475* **369**, 72–76 (1994).

32. Jaskolski, M. *et al.* Crystallographic models of SARS-CoV-2 3CLpro: in-depth assessment of structure quality and validation. *IUCrJ* **8**, 238–256 (2021).

33. Ma, C. *et al.* Boceprevir, GC-376, and calpain inhibitors II, XII inhibit SARS-CoV-2 viral replication by targeting the viral main protease. *Cell Research 2020 30:8* **30**, 678–692 (2020).

34. Kabsch, W. XDS. *urn:issn:0907-4449* **66**, 125–132 (2010).

35. Evans, P. R. & Murshudov, G. N. How good are my data and what is the resolution? *urn:issn:0907-4449* **69**, 1204–1214 (2013).

36. McCoy, A. J. *et al.* Phaser crystallographic software. *J Appl Crystallogr* **40**, 658–674 (2007).

37. Liebschner, D. *et al.* Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. *Acta Crystallogr D Struct Biol* **75**, 861–877 (2019).

38. Fu, L. *et al.* Both Boceprevir and GC376 efficaciously inhibit SARS-CoV-2 by targeting its main protease. *Nat Commun* **11**, (2020).

39. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr D Biol Crystallogr* **66**, 486–501 (2010).

40. Afonine, P. v. *et al.* Towards automated crystallographic structure refinement with phenix.refine. *Acta Crystallogr D Biol Crystallogr* **68**, 352–367 (2012).

41. Tom Burnley, B., Afonine, P. v., Adams, P. D. & Gros, P. Modelling dynamics in protein crystal structures by ensemble refinement. *Elife* **1**, (2012).

42. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637 (1983).

43. Touw, W. G. *et al.* A series of PDB-related databanks for everyday needs. *Nucleic Acids Res* **43**, D364–D368 (2015).

44. Parra, R. G. *et al.* Protein Frustratometer 2: a tool to localize energetic frustration in protein molecules, now with electrostatics. *Nucleic Acids Res* **44**, W356–W360 (2016).

45. Krissinel, E. & Henrick, K. Inference of macromolecular assemblies from crystalline state. *J Mol Biol* **372**, 774–797 (2007).

46. Case, D. A. *et al.* The Amber biomolecular simulation programs. *J Comput Chem* **26**, 1668–1688 (2005).

47. Maier, J. A. *et al.* ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J Chem Theory Comput* **11**, 3696–3713 (2015).

48. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J Chem Phys* **79**, 926–935 (1983).

49. Harvey, M. J., Giupponi, G. & de Fabritiis, G. ACEMD: Accelerating biomolecular dynamics in the microsecond time scale. *J Chem Theory Comput* **5**, 1632–1639 (2009).

50. Eastman, P. *et al.* OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Comput Biol* **13**, e1005659 (2017).

51. Humphrey, W., Dalke, A. & Schulten, K. VMD: Visual molecular dynamics. *J Mol Graph* **14**, 33–38 (1996).

52. Michaud-Agrawal, N., Denning, E. J., Woolf, T. B. & Beckstein, O. MDAnalysis: A toolkit for the analysis of molecular dynamics simulations. *J Comput Chem* **32**, 2319–2327 (2011).

53. Gowers, R. J. *et al.* MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations. *Proceedings of the 15th Python in Science Conference* 98–105 (2016) doi:10.25080/MAJORA-629E541A-00E.

54. Frishman, D. & Argos, P. Knowledge-based protein secondary structure assignment. *Proteins: Structure, Function, and Bioinformatics* **23**, 566–579 (1995).

55. Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Comput Sci Eng* **9**, 90–95 (2007).

56. Ghahremanpour, M. M. *et al.* Identification of 14 Known Drugs as Inhibitors of the Main Protease of SARS-CoV-2. *ACS Med Chem Lett* **11**, 2526–2533 (2020).

57. Behnam, M. A. M. Protein structural heterogeneity: A hypothesis for the basis of proteolytic recognition by the main protease of SARS-CoV and SARS-CoV-2. *Biochimie* **182**, 177–184 (2021).

58. Brzezinski, D. *et al.* Covid-19.bioreproducibility.org: A web resource for SARS-CoV-2-related structural models. *Protein Sci* **30**, 115–124 (2021).

59. Wlodawer, A. *et al.* Ligand-centered assessment of SARS-CoV-2 drug target models in the Protein Data Bank. *FEBS J* **287**, 3703–3718 (2020).

60. Drayman, N. *et al.* Masitinib is a broad coronavirus 3CL inhibitor that blocks replication of SARS-CoV-2. *Science* **373**, 931–936 (2021).

61. Chen, M. *et al.* Surveying biomolecular frustration at atomic resolution. *Nature Communications 2020 11:1* **11**, 1–9 (2020).

62. Bryngelson, J. D. & Wolynes, P. G. Spin glasses and the statistical mechanics of protein folding. *Proceedings of the National Academy of Sciences* **84**, 7524–7528 (1987).

63. Freiberger, M. I., Brenda Guzovsky, A., Wolynes, P. G., Gonzalo Parra, R. & Ferreiro, D. U. Local frustration around enzyme active sites. *Proc Natl Acad Sci U S A* **116**, 4037–4043 (2019).

64. Verma, N., Henderson, J. A. & Shen, J. Proton-Coupled Conformational Activation of SARS Coronavirus Main Proteases and Opportunity for Designing Small-Molecule Broad-Spectrum Targeted Covalent Inhibitors. *J Am Chem Soc* **142**, 21883–21890 (2020).

65. Deshmukh, M. G. *et al.* Structure-guided design of a perampanel-derived pharmacophore targeting the SARS-CoV-2 main protease. *Structure* **29**, 823-833.e5 (2021).

66. Wei, P. *et al.* The N-terminal octapeptide acts as a dimerization inhibitor of SARS coronavirus 3C-like proteinase. *Biochem Biophys Res Commun* **339**, 865–872 (2006).

67. Kneller, D. W. *et al.* Structural, Electronic, and Electrostatic Determinants for Inhibitor Binding to Subsites S1 and S2 in SARS-CoV-2 Main Protease. *J Med Chem* **64**, 17366–17383 (2021).

68. MacDonald, E. A. *et al.* Recognition of Divergent Viral Substrates by the SARS-CoV-2 Main Protease. *ACS Infect Dis* **7**, 2591–2595 (2021).

69. Lee, T. W. *et al.* Crystal structures of the main peptidase from the SARS coronavirus inhibited by a substrate-like aza-peptide epoxide. *J Mol Biol* **353**, 1137–1151 (2005).

70. Zheng, K. *et al.* Insight into the activity of SARS main protease: Molecular dynamics study of dimeric and monomeric form of enzyme. *Proteins* **66**, 467–479 (2007).

71. Kneller, D. W. *et al.* Unusual zwitterionic catalytic site of SARS-CoV-2 main protease revealed by neutron crystallography. *J Biol Chem* **295**, 17365–17373 (2020).

72. Kneller, D. W. *et al.* Structural plasticity of SARS-CoV-2 3CL Mpro active site cavity revealed by room temperature X-ray crystallography. *Nat Commun* **11**, (2020).

73. Orellana, L. Large-Scale Conformational Changes and Protein Function: Breaking the in silico Barrier. *Front Mol Biosci* **6**, (2019).

74. Selvaggio, G. & Pearlstein, R. A. Biodynamics: A novel quasi-first principles theory on the fundamental mechanisms of cellular function/dysfunction and the pharmacological modulation thereof. *PLoS One* **13**, (2018).

75. Wan, H., Aravamuthan, V. & Pearlstein, R. A. Probing the Dynamic Structure-Function and Structure-Free Energy Relationships of the Coronavirus Main Protease with Biodynamics Theory. *ACS Pharmacol Transl Sci* **3**, 1111–1143 (2020).

76. Cannalire, R., Barreca, M. L., Manfroni, G. & Cecchetti, V. A Journey around the Medicinal Chemistry of Hepatitis C Virus Inhibitors Targeting NS4B: From Target to Preclinical Drug Candidates. *J Med Chem* **59**, 16–41 (2016).

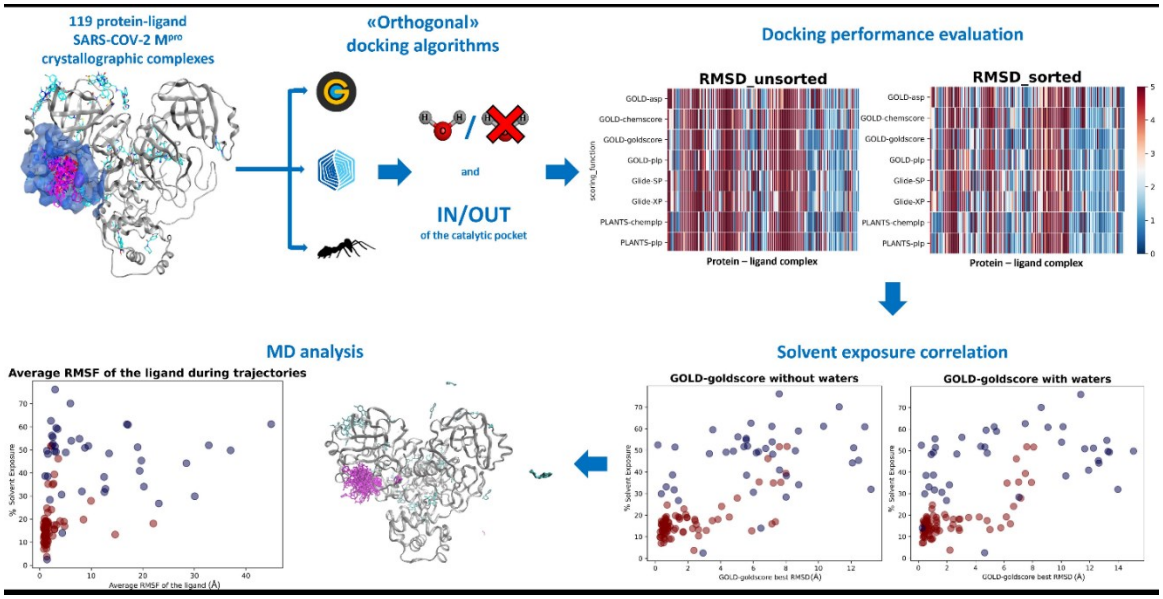
77. Yang, H. & Yang, J. A review of the latest research on Mpro targeting SARS-COV inhibitors. *RSC Med Chem* **12**, 1026–1036 (2021).

78. Zhang, C. H. *et al.* Potent Noncovalent Inhibitors of the Main Protease of SARS-CoV-2 from Molecular Sculpting of the Drug Perampanel Guided by Free Energy Perturbation Calculations. *ACS Cent Sci* **7**, 467–475 (2021).

Re-Exploring the Ability of Common Docking Programs to Correctly Reproduce the Binding Modes of Non-Covalent Inhibitors of SARS-CoV-2 Protease M^{pro}

Davide Bassani, **Matteo Pavan**, Giovanni Bolcato, Mattia Sturlese and Stefano Moro

Bassani, D., Pavan, M., Bolcato, G., Sturlese, M. & Moro, S. Re-Exploring the Ability of Common Docking Programs to Correctly Reproduce the Binding Modes of Non-Covalent Inhibitors of SARS-CoV-2 Protease Mpro. *Pharmaceuticals* **15**, 180 (2022).



Abstract

In the latest decades, molecular docking has imposed itself as one of the most used approaches to computational drug discovery. Several docking benchmarks have been published, comparing the performance of different algorithms in respect to a molecular target of interest, usually evaluating their ability in reproducing the experimental data, which in most cases comes from X-ray structures. In this study, we elucidated the variation of the performance of three docking algorithms, GOLD, Glide, and PLANTS, in replicating the coordinates of the crystallographic ligands of SARS-CoV-2 main protease (M^{pro}). Through the comparison of the data coming from docking experiments and the values derived from the calculation of the solvent exposure of the crystallographic ligands, we highlighted the importance of this last variable for docking performance. Indeed, we

underlined how an increase in the percentage of the ligand surface exposed to the solvent in a crystallographic complex makes it harder for the docking algorithms to reproduce its conformation. We further validated our hypothesis through molecular dynamics simulations, showing that the less stable protein-ligand complexes (in terms of root-mean-square deviation and root-mean-square fluctuation) tend to derive from the cases in which the solvent exposure of the ligand in the starting system is higher.

1. Introduction

In the 1980s, with the first study provided by Kuntz et. al[1], the computational technique of molecular docking had its birth. The efficiency, speed, and robustness of this method make its presence a constant in every structure-based drug discovery pipeline[2]. To give a brief explanation, molecular docking consists of a multi-step computational process that aims to find the best conformation of a molecule to bind to another to form a stable complex[3]. In the field of medicinal chemistry, as is deductible, its main application is finding the best molecules to bind in a firm way to the desired target (a protein, a nucleic acid, etc.). The algorithm starts with the exploration of the conformations space of the ligands (exploiting the so-called “search algorithm”). The conformations (called “poses”) are then classified by a “scoring function”, which attributes a numeric value to the goodness of the interaction according to energetical and/or geometrical function.

After a series of iterations, the final conformations are presented from the program to the user, ranked by the internal scoring function[4].

In the last 30 years, many docking programs have been developed. Among them, GOLD[5] (a genetic docking algorithm developed by the Cambridge Crystallographic Data Center - CCDC), Glide[6] (a systematic docking program released under license by Schrödinger), AutoDock[7] (a non-commercial genetic algorithm from The Scripps Research Institute), AutoDock VINA[8] (created by the same organization and released for non-commercial use), and PLANTS[9] (an algorithm based on an “Ant Colony Optimization” method) have gained popularity.

The performance of molecular docking programs can be measured by evaluating their ability to reproduce the experimental structural data, such as the crystallographic

coordinates of a ligand into its binding site[10]. This ability has been evaluated in several benchmarks[11,12] to rank the performance of different algorithms regarding a specific target, usually using as the key parameter the root-mean-square deviation (RMSD) between the coordinates of the different poses given by the program and the crystallographic ones.

The ability to reproduce a crystallographic conformation strongly relies on different factors. First, the geometrical characteristics of the binding site, like extension and shape play a very important role; it is known that the performance of the algorithms has been improved to dock molecules in “cavities” or “pockets” rather than surfaces of proteins[13]. Second, the nature and the dimensions of the ligand are also crucial. Indeed, very small ligands may explore different places in a binding site, and the interactions that they can establish are usually few, reducing the “synergism” which could induce a molecule to keep a peculiar shape into a pocket[14]. On the other hand, even if drug-like molecules generally have higher conformational freedom, their dimensions force them to be oriented into a site in a more conserved way, so they have less roto-translational freedom.

In this study, we examine the ability of three docking programs characterized by diverse conformational sampling algorithms to efficiently reproduce the crystallographic pose of different ligands bound in different sites of a protein. To accomplish this task, a target of which several crystal structures are solved with the ligands located in different sites of the macromolecule itself is needed. To this scope, we considered a very recent and relevant target in the current pharmaceutical scenario, such as the SARS-CoV-2 main protease (M^{pro}).

In the last couple of years, with the pandemic spreading of the SARS-CoV-2 virus, the world of medical sciences had found itself fighting a new and dangerous adversary[15,16]. This biological entity, which is part of the coronavirus family, has been demonstrated to cause a pulmonary infection which eventually leads to serious complications, as witnessed by the high number of deaths that have already been linked to it (more than 5 million, at the present day[17]). The replication cycle of this virus strongly relies on the activity of its main protease (known as M^{pro} or $3CL^{pro}$, a crystallographic structure example is reported in Figure 1)[18]. Indeed, this protein is responsible for the cleavage of the propeptide

transcribed by the viral genome. This way, the formation of all the functional proteins for the building of new virions takes place, and so the viral infection can proceed. Even if many molecules have been shown to bind to M^{pro} [19] and inhibit its activity, and even if a molecule is currently in phase III clinical trial for this purpose (PF-07321332, from Pfizer[20,21]) no drug has already been approved by the European Medicinal Agency for the treatment of COVID-19. Computational methods have already proven to be beneficial in the research for new potential inhibitors for M^{pro}, as literature witnesses[22,23]. In this work, we decided to implement a molecular docking-based approach relying on the programs GOLD, Glide, and PLANTS. These algorithms are considered "orthogonal" because they are characterized by diverse placing and scoring algorithms to obtain the best solution to the "protein-ligand posing problem". Each of these programs has been used to dock each of the different non-covalent ligands of the various crystal structures of M^{pro}, and this allowed us to evaluate the factors which influence the variability in reproducing the crystallographic poses. A self-docking protocol similar to the one herein reported had already been developed by our laboratory, with the name "DockBench". This program has been implemented with success in several workflows, as literature assesses[22,23]. In this study, a slightly modified version of that tool was used, which exploits only three docking programs at the present moment but can expand the analysis of the results obtained.

Looking at the docking benchmarking protocols on M^{pro}, a remarkable study has already been conducted and published by Zev et al.[26]. In that specific work, 6 different docking programs were evaluated in their performance in reproducing the M^{pro} non-covalent ligands' crystallographic poses, and 3 of those algorithms have also been compared in their ability to correctly place M^{pro} covalent ligands into their proper binding site. In our work, we decided to expand the considerations brought by that study, evaluating specifically how docking performance changes in respect of the crystallographic data that has to be reproduced.

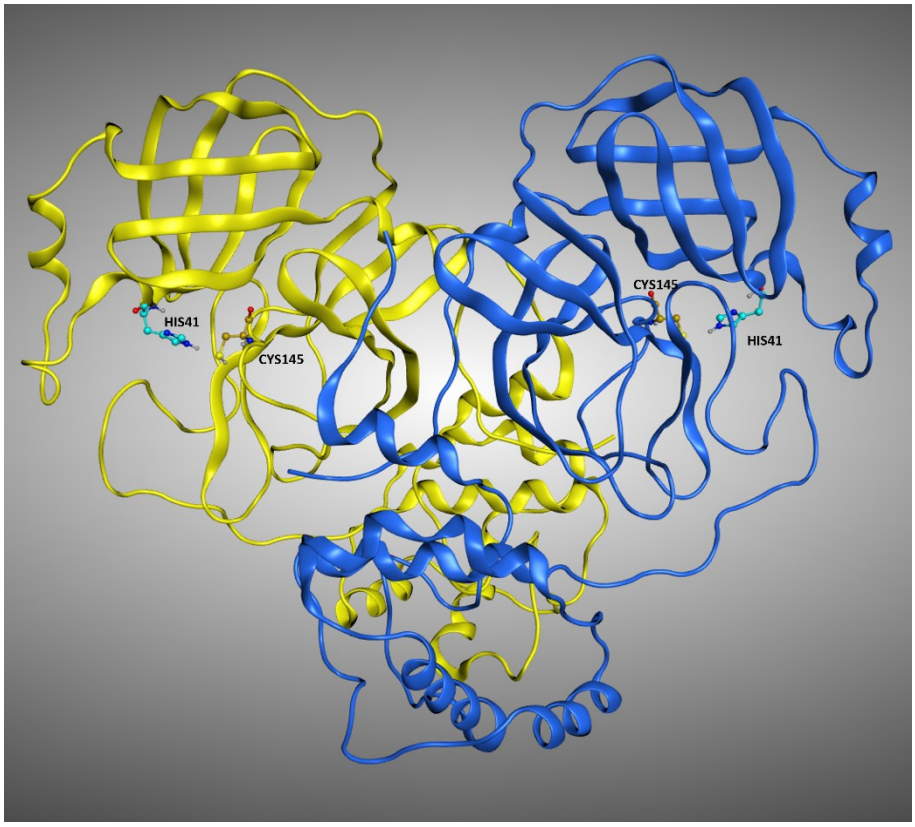


Figure 1. Representation of the crystal structure of M^{pro} (PDB: 7L10). The two monomers composing the protein are colored differently, while the residues of the catalytic dyad, Cys145, and His41, are labeled in each of the monomers.

Indeed, we considered in our calculations parameters like the solvent exposure of the ligand and the influence of the crystallographic water molecules in docking calculations, focusing our evaluations just on non-covalent M^{pro} ligands. We executed the experiment in two different scenarios, one which excluded the crystallographic waters from the calculation, and one which induced the docking programs to consider them. After that, we compared the docking results with the percentage of solvent exposure of the crystallographic pose of the ligand, successfully confirming that a higher solvent exposure tendentially reflects a worsening in the ability to reproduce the crystallographic pose by the algorithms (that, as already mentioned, are better trained for “cavities” rather than “surfaces”). To further investigate this aspect, we expanded our computational analysis by performing a molecular dynamics (MD) experiment, in which each crystallographic ligand was left free to move for 5 ns (3 replicas per system). This approach (known as “MD post-docking”) has already become part of our computational protocol[25,26] and is based

on the fact that the conformations of the ligands which are less prone to be displaced from their initial position during the simulation are related to higher stability and binding strength with the target. In the case presented, this principle was applied directly to the crystallographic conformations of the ligands rather than to docking poses. This was done because the goal wasn't to select the most promising molecules in binding to a specific region of the protein, but instead, the objective was to elucidate which are the features of the crystallographic ligands that tend to guarantee a tighter binding with the receptor. Our evaluation demonstrated that the molecules bound to the orthosteric pocket of M^{pro} keep their position much stronger than the molecules crystallized on other sites, further validating our solvent exposure-based theory. A representation of the M^{pro} ligands crystallized in the various sites of the protein is given in Figure 2.

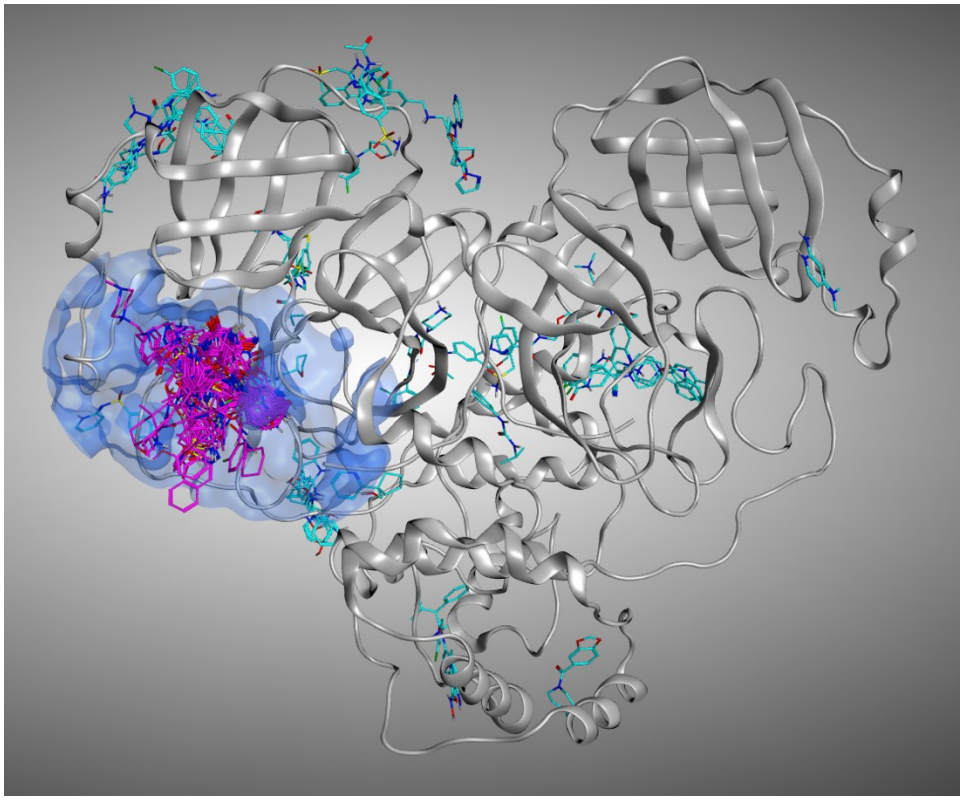


Figure 2. Representation of all the crystallographic ligands of M^{pro} superposed. To give a better view, just one protein structure is represented (the one coming from PDB: 7L10). The ligands which are crystallized inside the catalytic pocket are colored in magenta, while all the small molecules crystallized outside the orthosteric binding site are colored in cyan.

2. Results and Discussion

2.1. "Scenario 1" – docking calculations without considering the crystallographic water molecules

The results of our docking protocol for this section (which are numerically reported in the Supplementary materials file "Selfdocking_scenario1.csv") are graphically represented with colormaps. All the colormaps present in this study are based on a colorimetric scale delineating the RMSD values, starting from 0 Å, which corresponds to a molecular docking pose exactly superposable to the crystallographic one (maximum docking performance, represented by the dark blue color), and reaching values of 5 Å or higher (minimum docking performance, all represented by the dark red color), corresponding to a very low level of overlap between the coordinates of the pose produced and the ones of the crystallographic conformation. The colormaps in Figure 3 show the self-docking results obtained on the different M^{pro} crystal structures in the case in which water molecules are not considered in the calculation. As is depicted, the RMSD values were far lower for all the complexes in which the crystallographic ligand is located in the orthosteric pocket.

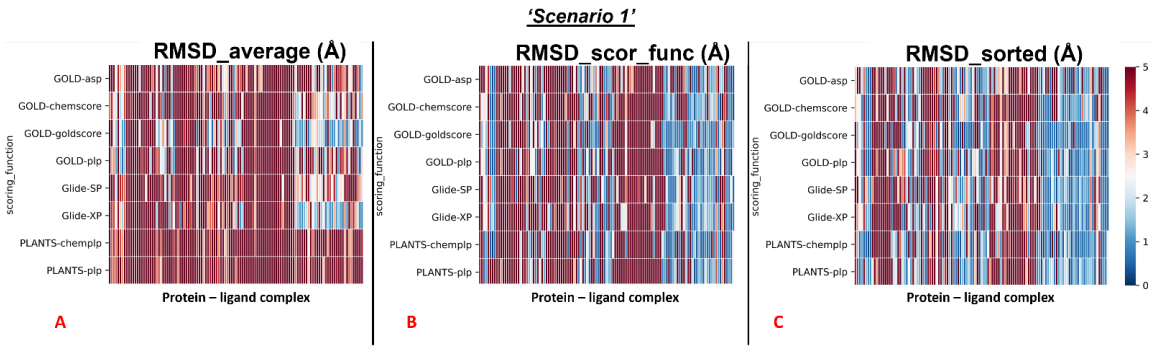


Figure 3. Colormaps represent the results of self-docking experiments in the case in which the crystallographic water molecules are not considered during the docking runs. The three plots illustrate respectively: A) the results coming from the average of the RMSDs of all the poses for each docking run; B) the results derived just from the RMSD between the crystallographic ligand coordinates and the pose classified as the best from the scoring function; C) the results of the self-docking experiments if just the pose showing the best RMSD value between its coordinates and the crystallographic ones are retained. The x-axis lists all the different protein-ligand complexes, which are plotted against the different pairs docking program-scoring function used for this study, reported in the y-axis.

To give a better resolution of this, we have separated each map into two different colormaps, one grouping all the 78 proteins in which the ligand is located into the catalytic pocket, and one including all other cases (41 complexes).

We analyzed the data coming from the calculations, and we computed that, looking at all the complexes with all the different couples docking program-scoring functions, the average values of all the RMSDs obtained was 5.76 Å ("RMSD_average"). Looking at the average of the RMSDs coming from the poses which were scored as the best ones from the algorithms' scoring functions ("RMSD_scor_func"), the value was 5.10 Å. If the lowest RMSD values only are taken into account for each docking run ("RMSD_sorted"), the mean of the values was 3.70 Å.

The average values have also been calculated separately for all the complexes in which the crystallographic ligand is located into the catalytic pocket, and for all other cases. The colormaps for these different conditions are reported in Figure 4 and Figure 5.

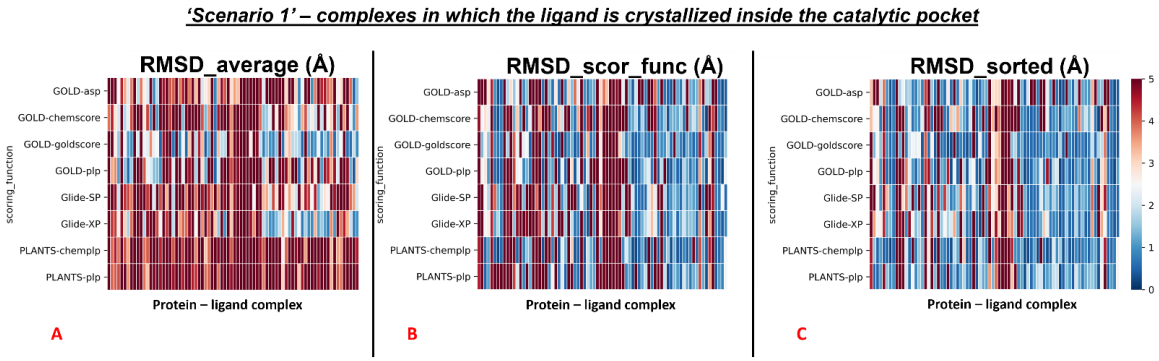


Figure 4. Colormaps represent the results of the self-docking experiments just for the ligands crystallized inside the orthosteric pocket when the crystallographic water molecules are not considered during the docking runs. The three plots illustrate respectively: (A) the results coming from the average of the RMSDs of all the poses for each docking run; (B) the results derived just from the RMSD between the crystallographic ligand coordinates and the pose classified as the best from the scoring function; (C) the results of the self-docking experiments if just the pose showing the best RMSD value between its coordinates and the crystallographic ones is retained. The x-axis lists all the different protein-ligand complexes, which are plotted against the different pairs docking program-scoring function used for this study, reported in the y-axis.

First, the analysis focused on the complexes having the crystallographic ligand located within the orthosteric pocket. For this set of systems, we calculated the average RMSD value of all the poses ("RMSD_average"), which was revealed to be 4.54 Å. Then, also the average of the RMSD values coming from the poses which were ranked with the best score from the scoring functions was computed ("RMSD_scor_func"), and its value was 3.43 Å. Finally, the average RMSD value of the poses with the lowest RMSD in each run was calculated ("RMSD_sorted"), and its measure was 2.45 Å.

Second, the same steps were done for the rest of the protein-ligand complexes, which are the ones in which the crystallographic ligand is located outside the orthosteric binding site. Also, in this case, the first passage involved the calculation of the average RMSD value of all the poses generated for these systems ("RMSD_average"), and its measure was 8.08 Å. Then, the mean of the RMSD values coming from the poses which received the highest rank from the scoring functions was calculated ("RMSD_scor_func") and is revealed to be 8.29 Å. In the end, the average value of the lowest RMSDs of each run was computed ("RMSD_sorted"), and its measure has shown to be 6.08 Å.

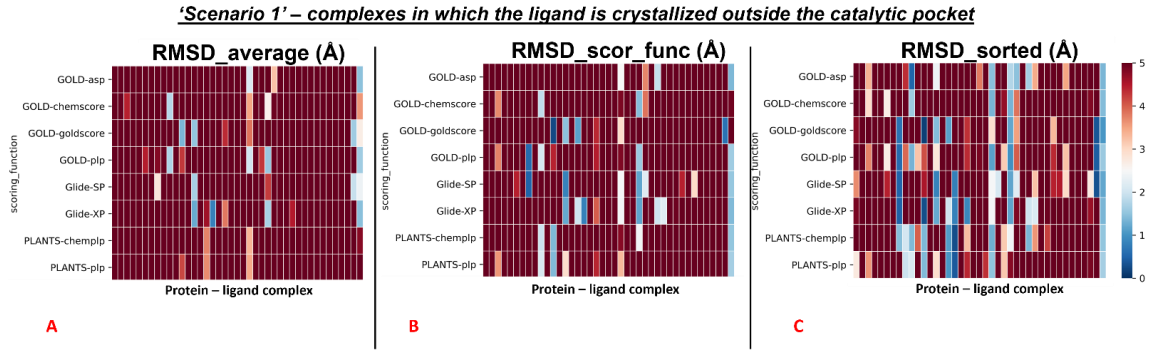


Figure 5. Colormaps represent the results of the self-docking experiments just for the ligands crystallized outside the orthosteric pocket in the case in which the crystallographic water molecules are not considered during the docking runs. The three plots illustrate respectively: (A) the results coming from the average of the RMSDs of all the poses for each docking run; (B) the results derived just from the RMSD between the crystallographic ligand coordinates and the pose classified as the best from the scoring function; (C) the results of the self-docking experiments if just the pose showing the best RMSD value between its coordinates and the crystallographic ones is retained. The x-axis lists all the different protein-ligand complexes, which are plotted against the different pairs docking program-scoring function used for this study, reported in the y-axis.

The results obtained for “scenario 1” are summarized in Table 1.

Results for “scenario 1” – no water molecules considered in docking calculations			
	RMSD_average (Å)	**RMSD_scor_func (Å)**	**RMSD_sorted (Å)**
All the 119 protein-ligand complexes	5.76	5.10	3.70
The 78 complexes with the ligand inside the catalytic pocket	4.54	3.43	2.45
The 41 complexes with the ligand outside the catalytic pocket	8.08	8.29	6.08

Table 1. Table representing the self-docking results obtained for “scenario 1”.

2.2. “Scenario 2” – docking calculations considering the crystallographic water molecules

The outcomes of our molecular docking experiment for this section (which are reported in the Supplementary materials file “Selfdocking_scenario2.csv”) are graphically represented with colormaps, which have been created with the same criteria listed in the previous paragraph. The results reported in the colormaps in Figures 6, 7, and 8 reveal the self-docking performance obtained on the different M^{pro} crystal structures in the case in which the crystallographic water molecules within 5 Å from the ligand were retained during the calculation. Also, in this case, it is easy to notice that the values result to be far better for the complexes in which the small molecule of interest is in the orthosteric binding site.

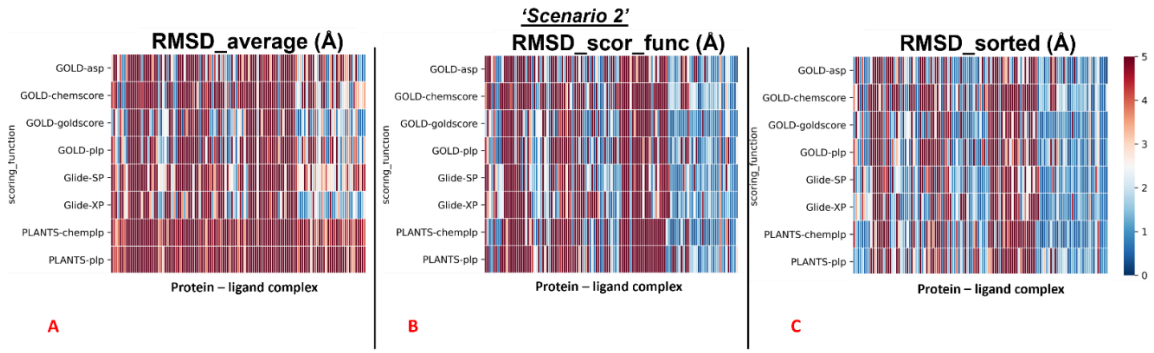


Figure 6. Colormaps represent the results of the self-docking experiments in the case in which the crystallographic water molecules at 5 Å or nearer to the ligand itself are considered during the docking runs. The three plots illustrate respectively: (A) the results coming from the average of the RMSDs of all the poses for each docking run; (B) the results derived just from the RMSD between the crystallographic ligand coordinates and the pose classified as the best from the scoring function; (C) the results of the self-docking experiments if just the pose showing the best RMSD value between its coordinates and the crystallographic ones is retained. The x-axis lists all the different protein-ligand complexes, which are plotted against the different pairs docking program-scoring function used for this study, reported in the y-axis.

Similar to the first scenario, we divided each colormap into two sets, one with the 78 proteins having the ligand located into the catalytic pocket, and the other including all the remaining cases (41 proteins). Considering all the protein-ligand complexes with all the different pairs docking program-scoring function, the mean values of all the RMSDs obtained ("RMSD_average") was 5.64 Å, but focusing only on the mean of the RMSDs derived from the poses which were given the highest rank from the algorithms ("RMSD_scor_func"), the value resulted to be 4.83 Å. Looking only at the best RMSDs for each docking run ("RMSD_sorted"), the average of the values was 3.68 Å.

As already done for "scenario 1", also in "scenario 2" the analysis was divided between the complexes in having the crystallographic ligand crystallized into the catalytic pocket, and for all other situations.

We reported the colormaps which resulted from this evaluation, and those are represented in Figures 7 and 8.

We started from the complexes in which the ligand is located inside the catalytic pocket in the crystal. For those systems, the mean of the RMSD values coming from all the poses("RMSD_average") resulted to be 4.22 Å. Then, the average of the RMSDs derived from the scoring function highest-ranked poses in all the docking runs ("RMSD_scor_func") was computed, and its value was 3.11 Å. In the end, also the average value between the lowest of the RMSDs in each docking run was calculated ("RMSD_sorted") and is revealed to be 2.26 Å.

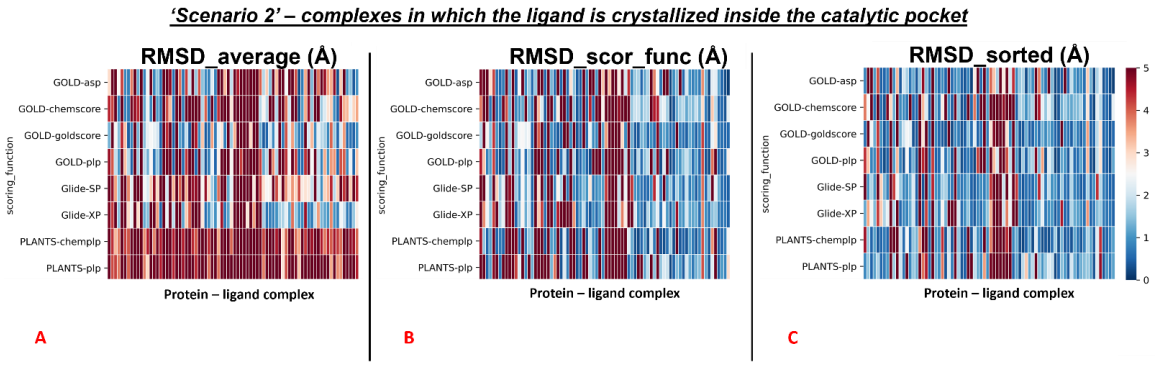


Figure 7. Colormaps represent the results of the self-docking experiments just for the ligands crystallized inside the orthosteric pocket when the crystallographic water molecules at 5 Å or nearer to the ligand itself are considered during the docking runs. The three plots illustrate respectively: (A) the results coming from the average of the RMSDs of all the poses for each docking run; (B) the results derived just from the RMSD between the crystallographic ligand coordinates and the once of the pose classified as the best from the scoring function; (C) the results of the self-docking experiments if just the pose showing the best RMSD value between its coordinates and the crystallographic ones are retained. The x-axis lists all the different protein-ligand complexes, which are plotted against the different pairs docking program-scoring function used for this study, reported in the y-axis.

Second, we repeated the analysis for all the complexes in which the crystallographic ligand is located outside the orthosteric pocket. For these systems, the average of the RMSD coming from all the poses collected in the docking runs ("RMSD_average") was calculated to be 8.32 Å. Next, we computed the mean of the RMSD values derived from the poses which received the highest score (from the scoring functions) in each run ("RMSD_scor_func"), and this value was 8.11 Å. Last, also the average value between the lowest of the RMSDs in each docking run was calculated ("RMSD_sorted"), giving 6.36 Å.

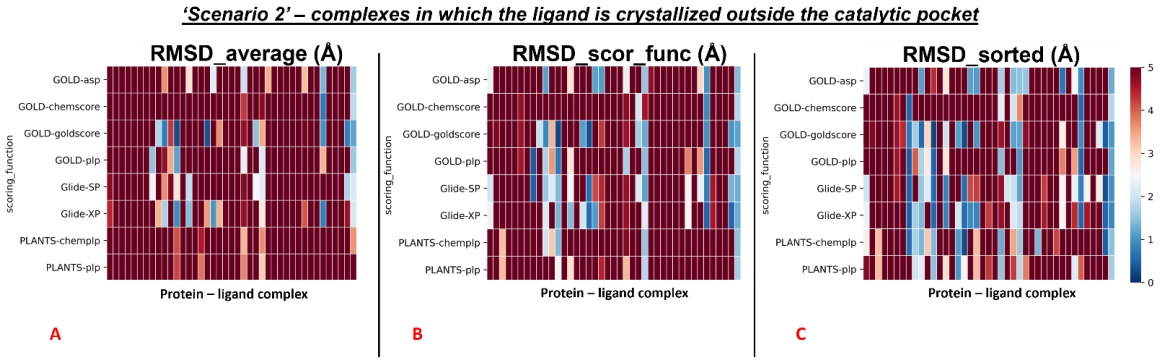


Figure 8. Colormaps represent the results of the self-docking experiments only for the ligands crystallized outside the orthosteric pocket when the crystallographic water molecules at 5 Å or nearer to the ligand itself are taken into account during the docking runs. The three plots illustrate respectively: (A) the results coming from the average of the RMSDs of all the poses for each docking run; (B) the results derived just from the RMSD between the crystallographic ligand coordinates and the once of the pose classified as the best from the scoring function; (C) the results of the self-docking experiments if just the pose showing the best RMSD value between its coordinates and the crystallographic ones are retained. The x-axis lists all the different protein-ligand complexes, which are plotted against the different pairs docking program-scoring function used for this study, reported in the y-axis.

The results obtained for "scenario 1" are summarized in Table 2.

Results for "scenario 2" – water molecules 5 Å or nearer to the ligand considered in docking calculations			
	RMSD_average (Å)	**RMSD_scor_func (Å)**	**RMSD_sorted (Å)**
All the 119 protein-ligand complexes	5.64	4.83	3.68
The 78 complexes with the ligand inside the catalytic pocket	4.22	3.11	2.26
The 41 complexes with the ligand outside the catalytic pocket	8.32	8.11	6.36

Table 2. Table representing the self-docking results obtained for "scenario 2".

Just analyzing the numbers coming from the average values, is depictable how the performance of the docking programs dramatically increases when the ligand is docked inside the catalytic pocket rather than on the surface of the protein, in line with the fact that the molecules have a limitation in the conformation that they can explore into a binding site. Together with this, the ligands can exploit their accessible surface area to interact with the protein more efficiently, following the principle of "complementarity"[27,28].

2.3. Solvent exposure analysis

The results of the docking calculations were then analyzed in light of the data coming from the solvent exposure analysis. For each docking program-scoring function pair, the best RMSDs given by the docking calculation were evaluated against the solvent exposure of the ligand in its crystallographic pose. The results were reported in different plots, one for each couple docking program-scoring function, also in this case dividing the graphs in respect to the "scenario" from which the docking result was coming. To give an example, we reported in this article the plots for the pair GOLD-goldscore for each of these cases (Figures 9 and 10).

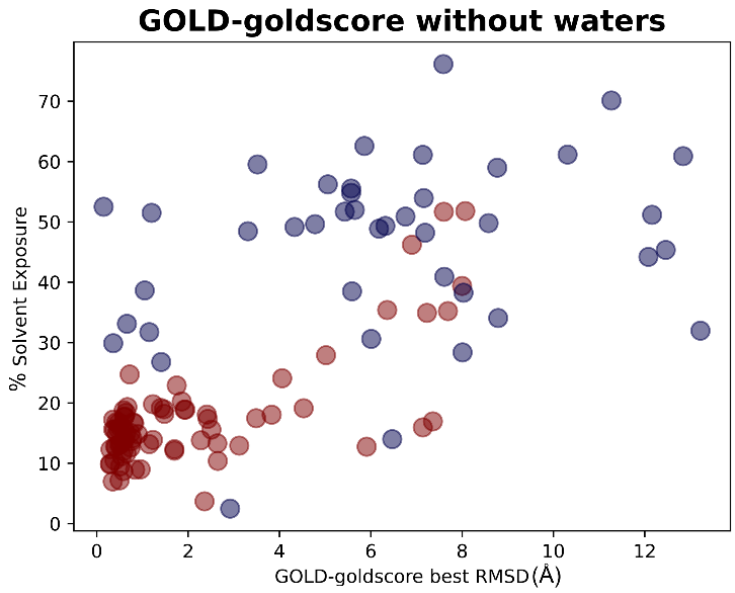


Figure 9. Scatter plots showing the different distribution of the RMSD values between the coordinates of the best pose from the GOLD-goldscore docking experiment in respect to the solvent exposure of the corresponding crystallographic ligands. The red dots represent the values having the ligand crystallized inside the catalytic pocket while the blue dots represent the ligands crystallized on the other side of M^{pro}. As can be noticed, the molecules showing the best values of RMSD are in most cases located inside the orthosteric pocket and characterized by low solvent exposure. This plot depicts part of the results of "scenario 1", and so the crystallographic water molecules are not considered in the docking runs of which the outcomes are here represented.

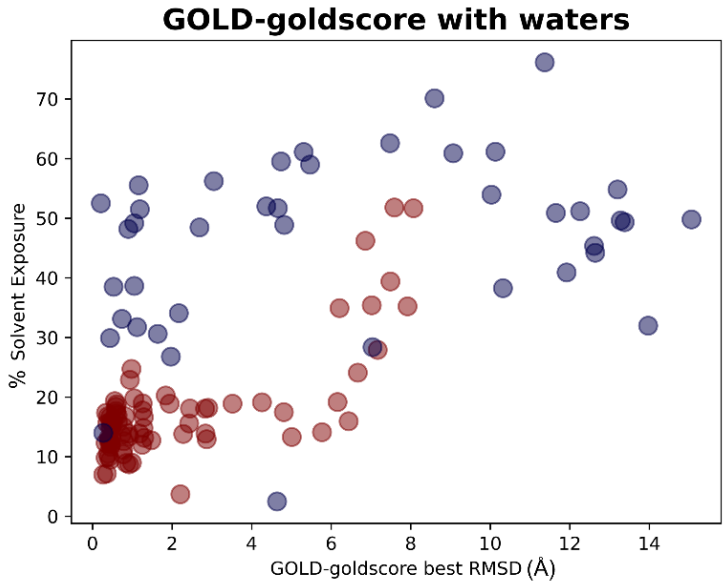


Figure 10. Scatter plots showing the different distribution of the RMSD values between the coordinates of the best pose from the GOLD-goldscore docking experiment in respect to the solvent exposure of the corresponding crystallographic ligands. The red dots represent the ligands that are originally crystallized inside the catalytic pocket, while the blue dots represent the ligands crystallized in the other parts of M^{pro}. As can be noticed, the molecules showing the best values of RMSD are in most cases located inside the orthosteric pocket and characterized by low solvent exposure. This plot depicts part of the results of "scenario 2", which means that the crystallographic water molecules at 5 Å or nearer to the ligand are also considered in the docking runs of which the outcomes are here represented.

The plots arising from all other docking program-scoring function pairs, both in "scenario 1" and "scenario 2", are reported in the supplementary material (Figure S1). From these graphs, it is easily depictable how the best RMSDs values tended to derive from protein-ligand complexes in which the solvent exposure of the ligand is low, and most of the time this means that the small molecule is crystallized in the binding pocket (indicated with the red dots in the plots). There are some cases in which the mean RMSD values were suboptimal also for this kind of ligands, and this can be due to several reasons. In some situations, of which the complexes 5REH, 5RE9, 5RGK (represented in Figure 11), and 7AVD are an example, the solvent exposure was tendentially higher in respect to the other orthosteric ligands, while in other cases the increase in RMSD can be attributable to the small dimensions of the ligand itself, which made harder for the docking algorithms to reproduce the reference pose in a pocket of such considerable volume (the complexes 5R82 and 5RG0 are an example for this)[31].

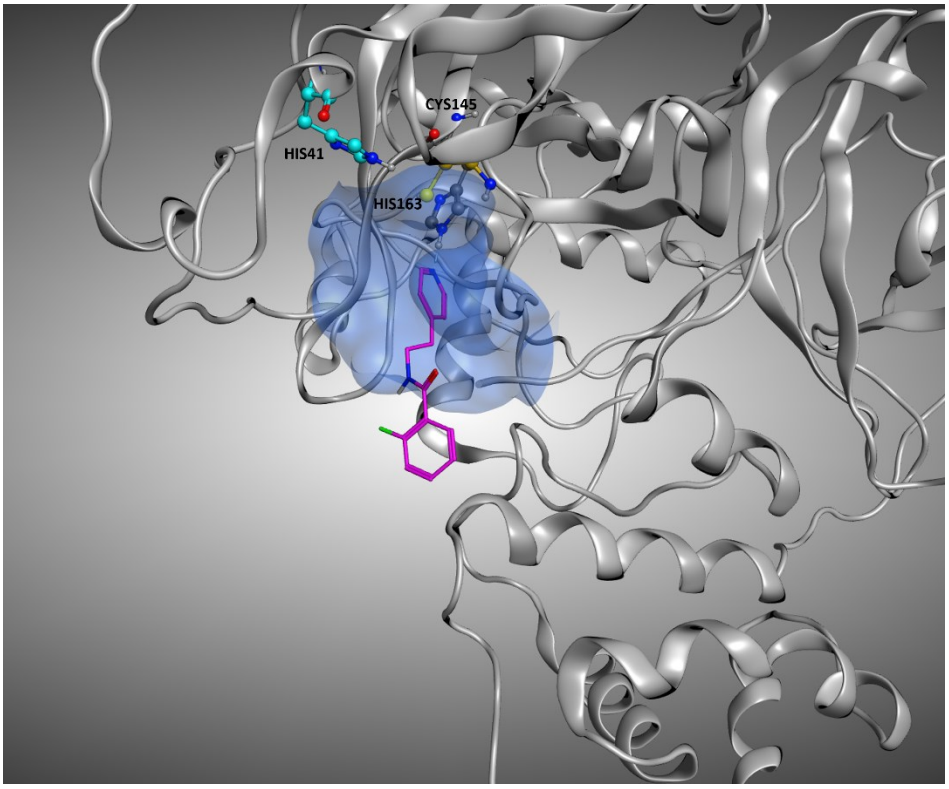


Figure 11. Representation of the crystallographic complex conformation of 5RGK, one of the protein-ligand complexes in which the crystallographic ligand is located inside the orthosteric binding site, but the docking calculation results in high RMSD values. This is mainly due to the high level of solvent exposure that characterizes this ligand, which locates just a small portion of its structure inside the pocket, leaving the rest in an outer zone. The ligand is represented with stick representation (C-atom are colored in magenta), the catalytic dyad (Cys145 and His41) is highlighted, as well as the His163, the binding site residue interacting with the ligand. To give a better representation, the surface of the protein in a 5 Å radius from the ligand is represented and colored in blue.

On the other hand, there are also some cases in which the best RMSD given by the protocol was pretty low even if the crystallographic ligand is not placed inside the orthosteric pocket. This is the case, for example, of 7LFP (the crystallographic pose is reported in Figure 12); the ligand is placed at the interface between the monomers, and so its solvent exposure and RMSDs values were low even if is marked to be “outside the catalytic pocket”. A similar situation is observed on 5RF0, where the ligand, even if not located into the orthosteric pocket, is neither situated in the peripheral part of the protease.

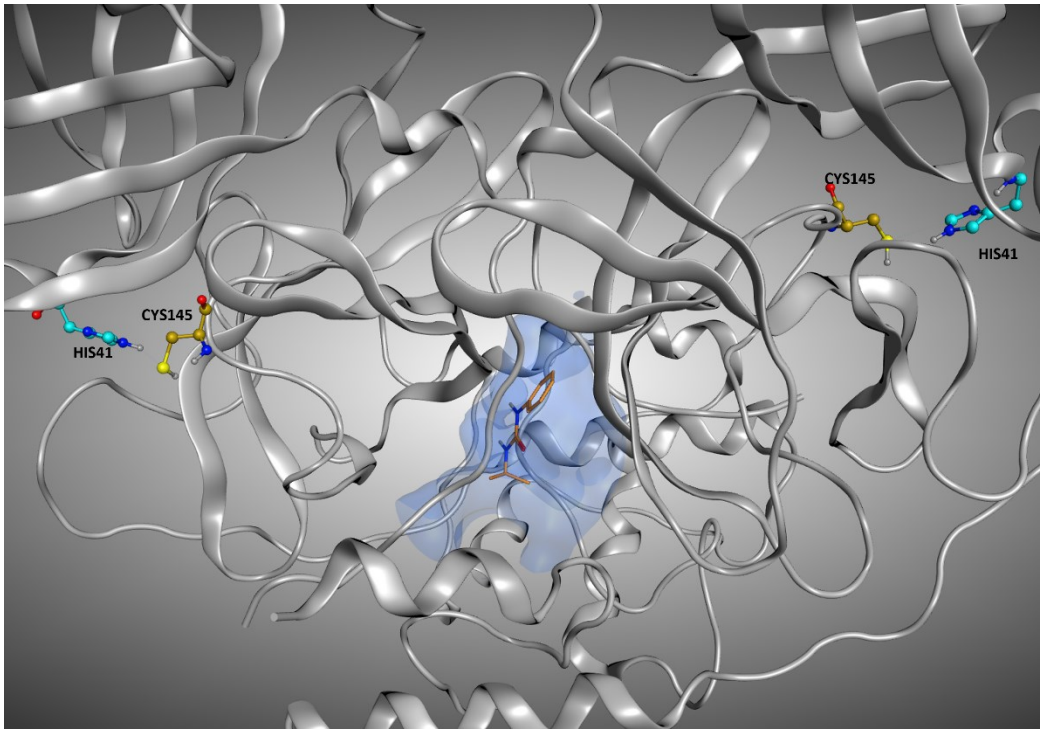


Figure 12. Representation of the crystallographic pose of 7LFP, which is one of the protein-ligand complexes in which, even if the crystallographic ligand is located outside the orthosteric binding site, the RMSD values between the original coordinates and the ones given from the docking runs are considerably low. The reason for this can be found in the very low solvent exposure of this ligand, which is located in the interface between the monomers, and so is shielded by them. In the Figure, the ligand is represented in orange and the catalytic dyad (Cys145 and His41) of both monomers is highlighted. To give a better representation, the surface of the protein in a 5 Å radius from the ligand is represented and colored in blue.

2.4. Molecular dynamics simulations

For each of the 119 crystallographic complexes, 3 different molecular dynamics simulations (MD) of 5 ns each were collected, to examine the behavior of the ligands in a dynamic environment. The trajectories were wrapped, aligned to the first frame and the root-mean-square fluctuation (RMSF) of the ligand, as well as the RMSD between its crystallographic and final coordinates ("RMSD_final"), were calculated for every single experiment. For each protein, the values coming from the average of the RMSFs and "RMSD_final" derived from the replicas were considered. Considering all the simulations collected, the average of all the ligand RMSF values was calculated to be 5.28 Å, while the average of the RMSD values between the coordinates of the crystallographic

conformation of the ligand and the ones coming from the last frame of the trajectory ("RMSD_final") was of 8.89 Å.

As already done for the docking results analysis, we firstly focused on the complexes in which the crystallographic ligand is originally located inside the orthosteric pocket. For these systems, the average of all the RMSFs coming from the simulations was 2.19 Å. The mean value of the RMSDs of the ligands in the last frame of each trajectory ("RMSD_final") was instead calculated to be 4.43 Å.

Second, we concentrated on the systems in which the crystallographic position of the ligand (and so its initial location) is outside the catalytic pocket. For these systems, the average value of all the ligand RMSFs during the trajectories was calculated to be 11.15 Å. Then, the RMSD value between the final coordinates of the ligands and their crystallographic ones ("RMSD_final") were considered. The average of these values, for all the trajectories collected for these complexes, was 17.66 Å. The output files of the molecular dynamics simulation geometric analysis are available in the supplementary material ("MD_data.csv").

As already done for the docking experiments, also for MD results the average values of RMSF and "RMSD_final" were plotted against the percentage of solvent exposure of the crystallographic conformation of the ligand, and the plots obtained have been reported in Figures 13 and 14. As expected, the complexes in which the ligand is crystallized in the orthosteric site (marked with the red dots in the scatter plot) tended to fluctuate much less than the ligands which are complexed in the external parts of the protease (represented with the blue dots in the graphs). As depicted, MD analysis confirms that the best values in terms of RMSF and "RMSD_final", which are correlated to a more energetically stable situation for the protein-ligand complex, come from the systems in which the crystallographic ligand is localized inside the catalytic pocket and are characterized by a low percentage of solvent exposure. These outcomes further support the already mentioned hypothesis about the correlation between the improvement of the docking performances in the case in which the binding site is a pocket rather than a surface.

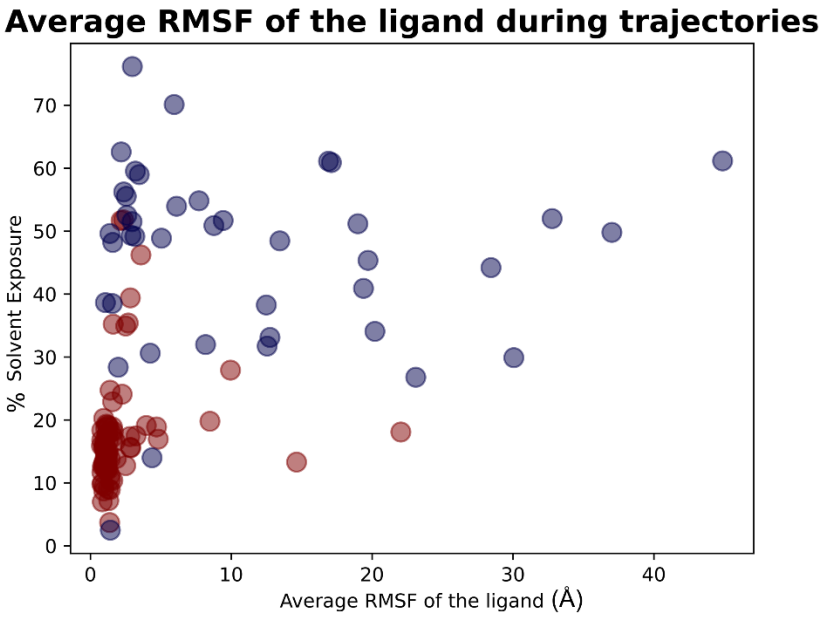


Figure 13. Scatter plots showing the different distribution of the mean RMSF values between the coordinates of the M^{pro} ligands compared to crystallographic ones after the molecular dynamics simulations in respect to the solvent exposure of the corresponding crystallographic ligands. The red dots represent the ligands that are originally crystallized inside the catalytic pocket, while the blue dots represent the ligands crystallized in the other parts of M^{pro}. As can be noticed, the molecules showing the best values of RMSF after the analysis of the trajectories are mainly located inside the catalytic pocket and characterized by a low solvent exposure of the original crystallographic pose.

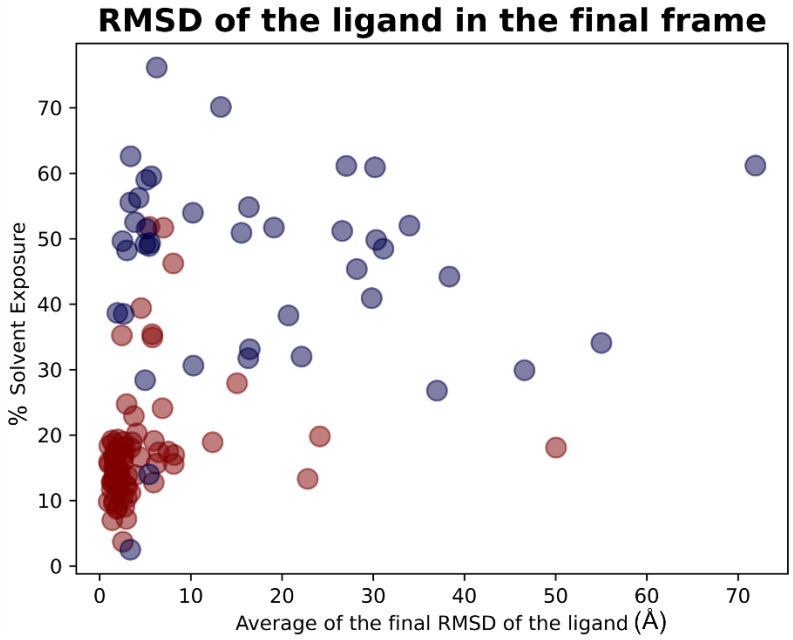


Figure 14. Scatter plots showing the different distribution of the mean RMSD values between the final coordinates of the M^{pro} ligands compared to crystallographic ones after the molecular dynamics simulations in respect to the solvent exposure of the corresponding crystallographic ligands. The red dots represent the ligands that are originally crystallized inside the catalytic pocket, while the blue dots represent the ligands crystallized in the other parts of M^{pro}. As can be noticed, the molecules showing the best values of RMSF after the analysis of the trajectories are mainly located inside the catalytic pocket and characterized by a low solvent exposure of the original crystallographic pose.

The overall results obtained with molecular dynamics simulations are summarized in Table 3. A graphical representation of the molecular dynamics simulations is reported in the Supplementary material ("Video_S1.mp4"). In this video, the ligands crystallized into the catalytic pocket were colored in magenta, while the other ligands were colored in cyan.

Results of the molecular dynamics simulations		
	RMSD_final (Å)	**RMSF_average (Å)**
All the 119 protein-ligand complexes	8.98	5.28
The 78 complexes with the ligand inside the catalytic pocket	4.43	2.19
The 41 complexes with the ligand outside the catalytic pocket	17.66	11.15

Table 3. Table representing the results of the molecular dynamics experiments.

3. Materials and Methods

3.1. Software overview

The molecular modeling operations were executed with the Molecular Operating Environment (MOE) suite (version 2019.01)[32]. The molecular docking calculations have been carried out with CCDC GOLD (version 2020), Schrodinger Glide (from the Schrödinger suite 2021.3), and PLANTS. The solvent exposure calculation was performed with a series of SVL commands (exploiting "moebatch" of the MOE suite) implemented into a python script. The systems for molecular dynamics simulations were prepared using tleap[33] and VMD[34]. The simulations were then carried out with ACEMD3[35] (version 3.3.0), a licensed program based upon OpenMM[36] (version 7.4.0). The modeling and docking calculations have been performed on a 12 CPU (Intel Xeon E5-1620 3.50 GHz) Linux Workstation, while

the MD simulations were carried out on a GPUs-cluster based composed of 20 NVIDIA GPUs.

3.2. Structure preparation for docking calculations

The different crystal structures of M^{pro} were collected from the Protein Data Bank[37]. Among these, the proteins which did not present any ligand, or which were complexed with a covalent ligand, were excluded. This way, only the non-covalent protein-ligand complexes were retained, and the complete list of all the 119 complexes used is available in the Supplementary material (Table S1). The structures were grouped into a database and were prepared with MOE "QuickPrep" tool. With this tool, each complex was properly prepared to recreate the small missing loops in the structure, assigning the proper conformation to the residues with alternate orientation (based on occupancy) and adding the hydrogens to the system (this last passage was performed with the MOE "Protonate 3D" tool). The hydrogens added this way were then minimized using the AMBER10:EHT force field implemented in MOE[38].

After these preliminary but crucial steps, each complex was manually examined and treated to eliminate every molecule except the crystallographic waters and the main ligand. Each complex was then independently saved.

3.3. Docking calculations

For each of the complexes prepared, the crystallographic ligand was separated from the protein and self-docked into its binding site. For each docking program, all the scoring functions available were used for separate runs, and each run 5 poses were collected for the ligand. GOLD supports 4 different scoring functions: goldscore, chemscore, asp, and plp; Glide supports two main functions for docking, which are Glide-SP and Glide-XP, while PLANTS implements plp and chemplp.

For each docking program-scoring function couple, the docking calculation has been carried out in two different scenarios: one in which the crystallographic water molecules were not considered (which we will refer to as "scenario 1") and one in which also the water molecules 5 Å or nearer from the ligand atoms were taken into account into the computation (which we will refer to as "scenario 2").

When all the docking calculations were executed, the ligand root-mean-square deviation (RMSD) between the coordinates of each one of the poses and the crystallographic conformations were computed. The data of major interest were: the RMSD in respect to the pose which is marked with the highest score by the program (RMSD_scor_func), the lowest RMSD of the docking run (RMSD_sorted), and the average of the RMSDs of all the poses generated (RMSD_average). The output files of the self-docking experiments executed are available in the supplementary material ("Selfdocking_scenario1.csv" and "Selfdocking_scenario2.csv").

3.4. Solvent exposure calculation

For each M^{pro} complex, the solvent exposure of the main crystallographic ligand was calculated with an SVL script based on MOE "moebatch". The output of such calculation was the percentage of the ligand surface which is exposed to the solvent in the protein-ligand crystallographic complex. A table comprising all the percentages obtained is present in the Supplementary material (Table S2).

3.5. Molecular dynamics simulations set up and execution

All the protein-ligand M^{pro} systems were independently prepared for molecular dynamics simulations. The program tleap was used for the creation of the simulation box, which was set to be cubic and characterized by a 15 Å padding. The solvation model used was the explicit TIP3P, and the neutralization of the system was performed adding Na^+ and Cl^- ions until the salt concentration inside the box reached the value of 0.154 M.

The systems then underwent a two-passage equilibration. In the first one, both protein and ligand atoms were subjected to a harmonic position restrain of 5 kcal/mol. The length of this step was set to 0.1 ns and the ensemble used was the canonical one (NVT). During the second equilibration step, the ensemble was set to NPT (isothermal-isobaric), the length was 0.5 ns and the harmonic restrains (always 5 kcal/mol) were applied both on the ligand and on the alpha-carbons of the protein backbone.

After these preliminary steps, 3 replicas of 5 ns each were collected for each system, the ensemble was again the NVT one and no restraints were applied. At the end of the simulations, the average root-mean-square fluctuation (RMSF) of the ligand during the

trajectory, as well as the RMSD between crystallographic coordinates of the ligand and the ones coming from the last frame of the trajectory were collected.

4. Conclusion

In this study, we have evaluated the performance of three orthogonal docking algorithms in reproducing the crystallographic pose of several ligands, located in different parts of the same target, which in our case was the SARS-CoV 2 main protease (M^{pro}). Our analysis revealed how, even if the docking programs used operate in different ways to give the final conformations to the user, all of them perform much better in the case in which the ligands are located in a binding pocket rather than crystallized outside of it. Specifically, we reported that their performance tends to decrease with the increment of the exposure of the crystallographic pose to the solvent. This was confirmed both from the experiments executed without considering the crystallographic water molecules in the docking calculations and from the ones taking into account the waters 5 Å or nearer to the ligand. Molecular dynamics simulations further give credit to our study, demonstrating how the less-fluctuating ligands (and so the most stable) through the trajectories were the once crystallized inside the orthosteric binding site of M^{pro}.

References

1. Kuntz, I. D., Blaney, J. M., Oatley, S. J., Langridge, R. & Ferrin, T. E. A geometric approach to macromolecule-ligand interactions. *J Mol Biol* **161**, 269–288 (1982).

2. Meng, X.-Y., Zhang, H.-X., Mezei, M. & Cui, M. Molecular Docking: A Powerful Approach for Structure-Based Drug Discovery. *Current Computer Aided-Drug Design* **7**, 146–157 (2011).

3. Lengauer, T. & Rarey, M. Computational methods for biomolecular docking. *Curr Opin Struct Biol* **6**, 402–406 (1996).

4. Kitchen, D. B., Decornez, H., Furr, J. R. & Bajorath, J. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat Rev Drug Discov* **3**, 935–949 (2004).

5. Jones, G., Willett, P., Glen, R. C., Leach, A. R. & Taylor, R. Development and validation of a genetic algorithm for flexible docking 1 1Edited by F. E. Cohen. *J Mol Biol* **267**, (1997).

6. Halgren, T. A. *et al.* Glide: A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening. *J Med Chem* (2004) doi:10.1021/jm030644s.

7. Morris, G. M. *et al.* AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J Comput Chem* **30**, (2009).

8. Trott, O. & Olson, A. J. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem* NA-NA (2009) doi:10.1002/jcc.21334.

9. Korb, O., Stützle, T. & Exner, T. E. PLANTS: Application of Ant Colony Optimization to Structure-Based Drug Design. in (2006). doi:10.1007/11839088_22.

10. Yusuf, D., Davis, A. M., Kleywegt, G. J. & Schmitt, S. An Alternative Method for the Evaluation of Docking Performance: RSR vs RMSD. *J Chem Inf Model* **48**, 1411–1422 (2008).

11. Boittier, E. D. *et al.* Assessing Molecular Docking Tools to Guide Targeted Drug Discovery of CD38 Inhibitors. *Int J Mol Sci* **21**, 5183 (2020).

12. Wang, Z. *et al.* Comprehensive evaluation of ten docking programs on a diverse set of protein–ligand complexes: the prediction accuracy of sampling power and scoring power. *Physical Chemistry Chemical Physics* **18**, 12964–12975 (2016).

13. Ramirez, U. D., Myachina, F., Stith, L. & Jaffe, E. K. Docking to Large Allosteric Binding Sites on Protein Surfaces. in 481–488 (2010). doi:10.1007/978-1-4419-5913-3_54.

14. Jacquemard, C., Drwal, M. N., Desaphy, J. & Kellenberger, E. Binding mode information improves fragment docking. *J Cheminform* **11**, 24 (2019).

15. Fan, Y., Zhao, K., Shi, Z.-L. & Zhou, P. Bat Coronaviruses in China. *Viruses* **11**, 210 (2019).

16. Lotfi, M., Hamblin, M. R. & Rezaei, N. COVID-19: Transmission, prevention, and potential therapeutic opportunities. *Clinica Chimica Acta* **508**, 254–266 (2020).

17. Latest access: 8th november 2021. https://covid19.who.int /.

18. Jin, Z. *et al.* Structure of Mpro from SARS-CoV-2 and discovery of its inhibitors. *Nature* **582**, 289–293 (2020).

19. Zhang, C.-H. *et al.* Potent Noncovalent Inhibitors of the Main Protease of SARS-CoV-2 from Molecular Sculpting of the Drug Perampanel Guided by Free Energy Perturbation Calculations. *ACS Cent Sci* **7**, 467–475 (2021).

20. Owen, D. R. *et al.* An oral SARS-CoV-2 M ^{pro} inhibitor clinical candidate for the treatment of COVID-19. *Science (1979)* (2021) doi:10.1126/science.abl4784.

21. Pavan, M., Bolcato, G., Bassani, D., Sturlese, M. & Moro, S. Supervised Molecular Dynamics (SuMD) Insights into the mechanism of action of SARS-CoV-2 main protease inhibitor PF-07321332. *J Enzyme Inhib Med Chem* **36**, 1646–1650 (2021).

22. Rudrapal, M. *et al.* Identification of bioactive molecules from Triphala (Ayurvedic herbal formulation) as potential inhibitors of SARS-CoV-2 main protease (Mpro) through computational investigations. *J King Saud Univ Sci* **34**, 101826 (2022).

23. Di Sarno, V. *et al.* Identification of a dual acting SARS-CoV-2 proteases inhibitor through in silico design and step-by-step biological characterization. *Eur J Med Chem* **226**, 113863 (2021).

24. Cuzzolin, A., Sturlese, M., Malvacio, I., Ciancetta, A. & Moro, S. DockBench: An Integrated Informatic Platform Bridging the Gap between the Robust Validation of Docking Protocols and Virtual Screening Simulations. *Molecules* **20**, 9977–9993 (2015).

25. Bolcato, G., Cuzzolin, A., Bissaro, M., Moro, S. & Sturlese, M. Can We Still Trust Docking Results? An Extension of the Applicability of DockBench on PDBbind Database. *Int J Mol Sci* **20**, (2019).

26. Zev, S. *et al.* Benchmarking the Ability of Common Docking Programs to Correctly Reproduce and Score Binding Modes in SARS-CoV-2 Protease Mpro. *J Chem Inf Model* **61**, 2957–2966 (2021).

27. Bolcato, G., Bissaro, M., Sturlese, M. & Moro, S. Comparing Fragment Binding Poses Prediction Using HSP90 as a Key Study: When Bound Water Makes the Difference. *Molecules* **25**, 4651 (2020).

28. Bolcato, G. *et al.* A Computational Workflow for the Identification of Novel Fragments Acting as Inhibitors of the Activity of Protein Kinase CK1δ. *Int J Mol Sci* **22**, 9741 (2021).

29. Yan, Y. & Huang, S.-Y. Pushing the accuracy limit of shape complementarity for protein-protein docking. *BMC Bioinformatics* **20**, 696 (2019).

30. Gabb, H. A., Jackson, R. M. & Sternberg, M. J. E. Modelling protein docking using shape complementarity, electrostatics and biochemical information 1 1Edited by J. Thornton. *J Mol Biol* **272**, 106–120 (1997).

31. Verdonk, M. L. *et al.* Docking Performance of Fragments and Druglike Compounds. *J Med Chem* **54**, 5422–5431 (2011).

32. Molecular Operating Environment (MOE), 2019.01; Chemical Computing Group ULC, 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, 2021.

33. D.A. Case et al. Amber 2021, University of California, San Francisco. Preprint at (2021).

34. Humphrey, W., Dalke, A. & Schulten, K. VMD: Visual molecular dynamics. *J Mol Graph* **14**, 33–38 (1996).

35. Harvey, M. J., Giupponi, G. & Fabritiis, G. De. ACEMD: Accelerating Biomolecular Dynamics in the Microsecond Time Scale. *J Chem Theory Comput* **5**, (2009).

36. Eastman, P. *et al.* OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Comput Biol* **13**, e1005659 (2017).

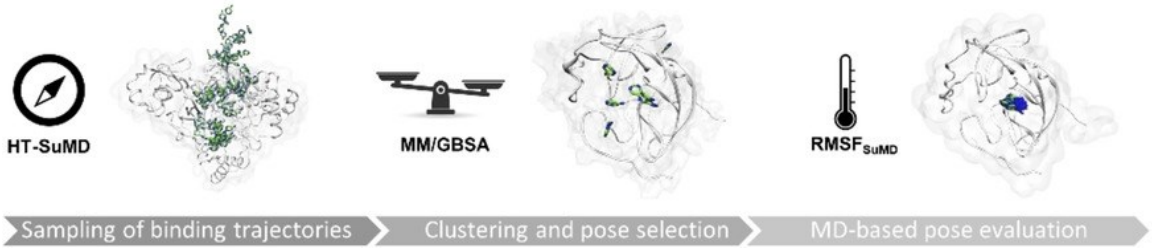
37. Berman, H. M. The Protein Data Bank. *Nucleic Acids Res* **28**, 235–242 (2000).

38. D.A. Case et al. Amber 10, University of California, San Francisco.

Inspecting the Mechanism of Fragment Hits Binding on SARS-CoV-2 M^{pro} by Using Supervised Molecular Dynamics (SuMD) Simulations

Maicol Bissaro, Giovanni Bolcato, **Matteo Pavan**, Davide Bassani, Mattia Sturlese and Stefano Moro

Bissaro, M. *et al.* Inspecting the Mechanism of Fragment Hits Binding on SARS-CoV-2 M ^{pro} by Using Supervised Molecular Dynamics (SuMD) Simulations. *ChemMedChem* **16**, 2075–2081 (2021).



Abstract

Computational approaches supporting the early characterization of fragment molecular recognition mechanism represent a valuable complement to more expansive and low-throughput experimental techniques. In this retrospective study, we have investigated the geometric accuracy with which high-throughput supervised molecular dynamics simulations (HT-SuMD) can anticipate the experimental bound state for a set of 23 fragments targeting the SARS-CoV-2 main protease. Despite the encouraging results herein reported, in line with those previously described for other MD-based posing approaches, a high number of incorrect binding modes still complicate HT-SuMD routine application. To overcome this limitation, fragment pose stability has been investigated and integrated as part of our in-silico pipeline, allowing us to prioritize only the more reliable predictions.

1. Introduction

Fragment-based drug discovery (FBDD) has progressively established as a game-changing approach to navigate the chemical space in the drug discovery pipelines, both on academic and industrial early discovery stages[1,2,3]. By definition, fragments are low molecular weight organic molecules able to recognize a target of therapeutic interest in a

mild affinity range and with a poor selectivity profile[4]. Intriguingly, the screening of small sized fragment libraries in place of conventional larger ones has proven to provide better coverage of the chemical diversity and higher hit rates[5,6]. The identification of such weak binders, however, strictly depends on the implementation of biophysical screening techniques, such as X-Ray Crystallography (XRC), Nuclear Magnetic Resonance (NMR), surface plasmon resonance (SPR), or Thermal Shift Assay (TSA)[1,7,8]. Anyway, broad differences exist among such methods and each of them suffers unique limitations in the challenging identification of reliable fragment; indeed the agreement in the hits identified is surprisingly limited[9,10,11]. Besides, only XRC and NMR offer the possibility to investigate the binding mode of weak binders. In light of this, structure-based computational strategies have increasingly gained appeal[12,13,14]. As highlighted in a recent review, during the last decade Molecular Dynamics (MD) simulations have been extensively applied also in the FBDD field, providing an atomistic insight on the fragment-receptor binding mechanisms, with a femtosecond temporal resolution[15]. From this perspective, we recently developed HT-SuMD, a computational protocol exploiting supervised MD simulations to perform the screening of a small fragments library in a competitive timescale[16]. The performance of the protocol in prioritizing the most promising fragment binders was compared with NMR-based screening, against the oncological protein target Bcl-xL. Despite the notable agreement with NMR in identifying the most promising hits, the lack of structural data prevent the assessment of HT-SuMD accuracy in fragments binding mode prediction, which would represent a valuable set of information to guide the subsequent hit to lead (H2L) optimization steps. In this methodological study, we have therefore retrospectively investigated the accuracy of HT-SuMD simulations in reproducing the experimental binding mode of several fragment-protein complexes, exploiting the 3C-like main protease (M^{pro}) of the novel SARS-CoV-2 coronavirus as a relevant case study. Following indeed the dramatic spread of the COVID-19 pandemic, a collaborative XRC fragment screening against the protein M^{pro} has timely offered to the scientific community valuable structural information to accelerate the rational design of new protease inhibitors[17,18,19]. For this validation study in detail, among the 71 fragments targeting the catalytic site of M^{pro} originally identified by the XRC screening, only the 23 presenting a reversible mechanism of recognition were taken into consideration, due to

the impossibility of modeling covalent reactivity through classical molecular mechanics (MM) force fields[20,21].

2. Results and Discussion

2.1 Characterization of fragment-receptor complexes

The high-quality M^{pro} crystallographic structures were collected from the Protein Data Bank database (PDB ID are reported in Table 1 of SI) and prepared by applying symmetric transformation to each asymmetric unit, thus recreating the original functional dimer[22]. A visual inspection of the catalytic clefts has revealed how the 23 non-covalent fragments comprehensively explore most protease binding subsites (S1, S2, S3, and S1') while providing decent coverage of chemical diversity. Besides, M^{pro} catalytic cleft is easily accessible from the bulk solvent and hence suitable to SuMD studies, as recently demonstrated for a couple of M^{pro} inhibitors[23]. The complexity, as well as the plasticity of the M^{pro} binding pocket, made this test case particularly challenging, the reason why an MD-based stability characterization of all the experimental-solved crystallographic complexes was performed, before investigating HT-SuMD accuracy in the fragment posing process.

For this purpose, the AMBER14SB force field was combined with the general amber force field (GAFF) to parameterize respectively the protein biopolymers and the small organic fragments[24,25]. To ensure results robustness, 5 trajectories each 20 ns long were collected for all M^{pro} complexes, resulting in a total of 2.3 μs of conventional MD simulations. The content of information extrapolated from a single trajectory has been hence doubled by simply repeating the analysis against the two distal and independent catalytic sites of the homodimeric SARS-CoV-2 M^{pro}. To characterize the geometric stability of the experimental-solved fragment complexes the root-mean-square fluctuation (RMSF) of ligands heavy atoms has been chosen as a metric, then summarizing the results through a heatmap representation, as reported in Figure 1.

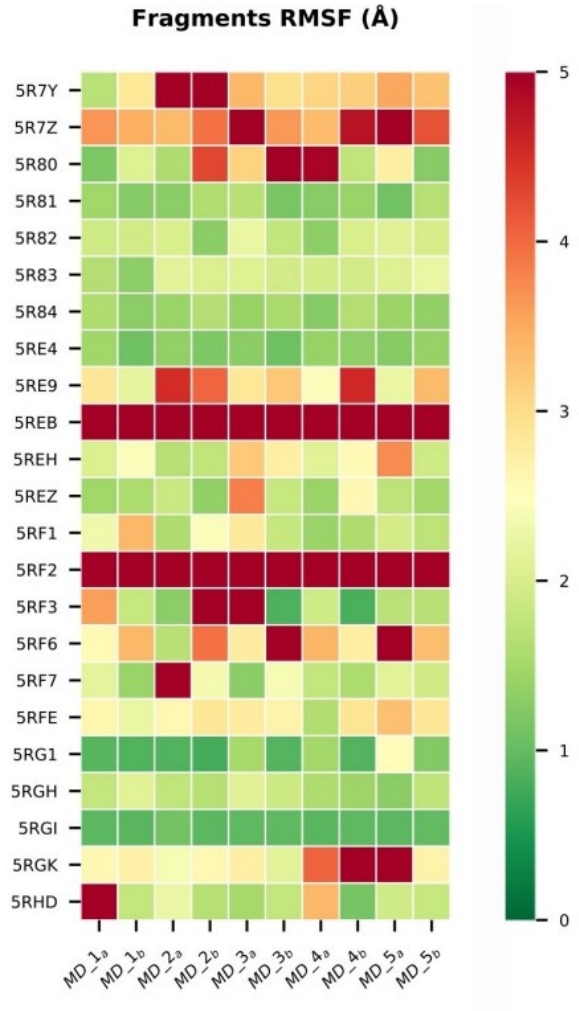


Figure 1. Fragment stability assessed by classical MD of the 23 crystallographic complexes. For each MD simulation collected (x-axis) starting from the crystallographic ligand-receptor complexes (y-axis), the pose stability value of the fragment is herein reported through a heatmap representation. The colorimetric scale, from green to red, quantitatively represents the RMSF computed for each ligand heavy atoms (0 to 5 Å scale). The MD simulation were carried out on each subunit of the Mpro functional dimer resulting in two set (labelled a and b) for each of the 5 runs.

The colorimetric scale helps in differentiating those fragments which maintained the original binding mode during all the collected replicates (green color), from others undergoing a neat perturbation of the recognition modality (yellow color) or that even experience a spontaneous unbinding event, repetitively leaving the catalytic cleft (red color). Interestingly, a strong correlation was identified between the topological localization of the fragments and their $RMSF_{avg}$, with those ligands occupying the highly flexible S2 subsite also showing the more pronounced propensity in losing the experimental solved binding mode. This information not only offers valuable insights for

the H2L optimization phase but also opens up questions about the suitability of MD-based approaches for the posing of ligands characterized by such limited structural stability.

2.2 Fragments posing through HT-SuMD

HT-SuMD protocol has been applied to investigate the binding mechanism of the 23 non-covalent fragments against the unliganded crystal structure of the SARS-CoV-2 M^{pro} (PDB ID 6YB7). As accurately described in the original paper, HT-SuMD manages the preparation, collection, and analysis of multiple SuMD simulations in an automatic modality, only requiring the binding pocket localization as initial information. SuMD, briefly, exploiting a tabu-like supervision algorithm that monitors in times variations in the ligand-protein binding site distances, could be considered an enhanced sampling approach improving the efficiency with which rare events, such as binding, are described[26,27]. For each fragment investigated, a solvated MD simulation box has been set up (a detailed description is reported on supplementary materials) and equilibrated after distancing the ligand at least 30 Å away from the protein catalytic cleft, to avoid premature intramolecular interactions. Also in this case, as an attempt to increase the robustness of the results, 10 SuMD replicates have been collected, resulting in a total of 6.3 μs of simulation time. The ensemble of 230 trajectories describing different fragment binding pathways has been later geometrically discretized through DBSCAN, a density-based clustering algorithm, which allows all the most populated ligand-protein states to emerge from the background noise[28,29]. In detail, a cluster is initialized if it contains at least 25 similar fragments conformations, which therefore differ from each other by no more than 1.5 Å. Finally, each binding mode was qualitatively evaluated using the MM/GBSA approach to approximate the ligand-protein free energy of binding, thus allowing to perform a ranking of the predicted poses[30]. The accuracy of the predictions was assessed by comparing each cluster of fragment conformations identified with the respective crystallographic reference, computing the root-mean-square deviations (RMSD) of non-hydrogen atomic coordinates. The results obtained for the 23 M^{pro} crystallographic inhibitors have been extensively reported in the supplementary information (SI_HT-SuMD.xlsx) and graphically summarized in Figure 2, exploiting a colorimetric map to differentiate the correctness of the posing protocol.

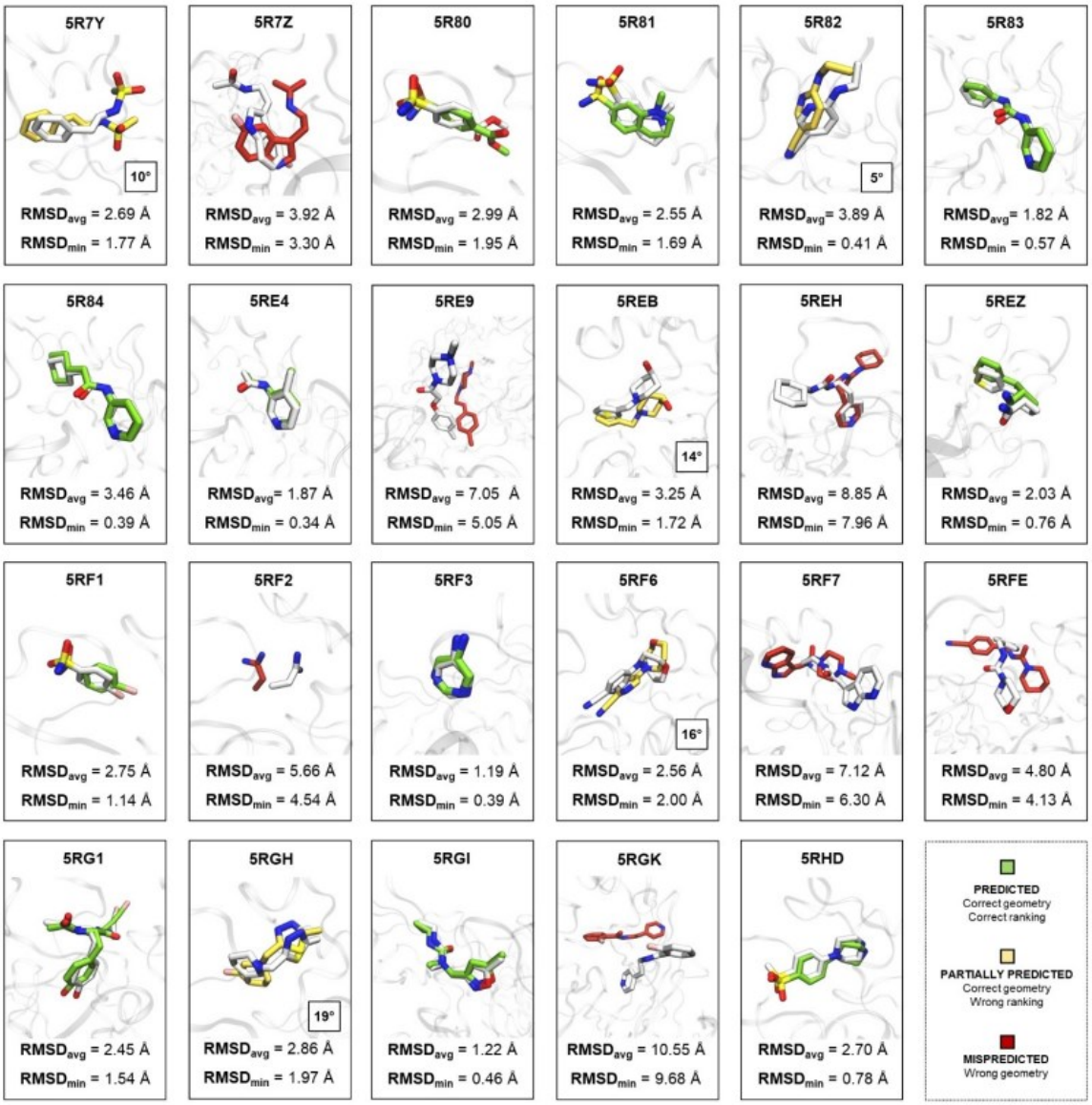


Figure 2. The results of the HT-SuMD posing protocol have been herein summarized. For each of the 23 fragments investigated the cluster of ligand conformations closes to the experimentally solved binding mode was reported, measuring the accuracy of the prediction through the RMSDavg and RMSDmin values of the selected cluster. The crystallographic reference has been rendered in white color, while the HT-SuMD predicted binding modes have been differentiated in green, yellow, and red color, following the criteria described in the legend. In the case of partially predicted fragments, in which a good binding geometry was retrieved but erroneously ranked, the magnitude of the error has been underlined reporting the incorrect ranking position.

More specifically, for each fragment, the minimum RMSD ($RMSD_{min}$) and the average RMSD ($RMSD_{avg}$) values for the best cluster, i.e. the cluster closer to the crystallographic reference, were reported then comparing the predicted binding mode with the

experimental one. The fragment posing exercise was considered correctly achieved if the $RMSD_{min}$ of the cluster selected falls below the cut-off value of 2 Å.

For 11 fragments out of 23, representing almost half of the considered cases, the protocol was able to identify and correctly rank the experimental binding mode (green-coloured molecules). Among these, the most noteworthy case is represented by the fragment with the PDB ID 5RGI, the only one targeting the S1' subsite. HT-SuMD posing approach, fully exploring the conformational flexibility of the receptor, was able to reproduce the fragment crystallographic binding mode in an extremely accurate way, with an $RMSD_{min}$ value of 0.46 Å. This result is impressive since, in the unliganded Mpro structure chosen in this study, the S1' pocket, due to a different orientation of the residues composing the catalytic dyad (H41 and C145), is initially inaccessible.

For the remaining 12 fragments, an in-depth analysis highlighted two orthogonal reasons underneath the HT-SuMD based posing failures. In 5 cases the MM/GBSA-based scoring method was unable to prioritize the experimental binding mode, even if it was exhaustively sampled by SuMD simulations (yellow-colored molecules). The incorrect ranking position was then reported in Figure 2 within a squared box, to underline the magnitude of the scoring error. This disagreement may be caused by limitations affecting the MM models, as errors in the fragments force field parameters or, more intriguingly, the crystallographic structures could capture only one of the possible accommodation states that the ligand can explore within the binding site[31]. In the other 7 cases instead, the experimental conformation was never sampled (red-colored molecules), suggesting possible MD-sampling issues that may be addressed by widening the number of SuMD replicates performed for each compound, however increasing the computational cost of our approach. The accuracy of HT-SuMD protocol, therefore, with 48% of correct binding mode predictions is greater than non-native docking-based protocols reported in the literature and in line with that of other MD based fragments posing approaches[32,33]. It appears however evident how the posing of fragments still represents a tough pharmaceutical challenge, in particular, as suggested by Verdonk, for those characterized by a low-ligand efficiency (LE). Even our computational approach, in about half of the

examined cases, fails to return a reliable result making its routine application very complex in a pharmaceutical drug discovery context.

To elucidate the applicability domain of HT-SuMD and better understand the limitations related to the implementation of MD-based protocols for the fragment binding modes prediction, we have therefore investigated if the fragment pose stability, a geometric-dynamic property, could impact the predictivity of our method. The fragment pose stability retraces the concept behind the structural stability criterion that has recently been discussed also by Barril's research group, as a complement to more traditional thermodynamic-based approaches in the identification of correct fragment-receptor binding mode[34].

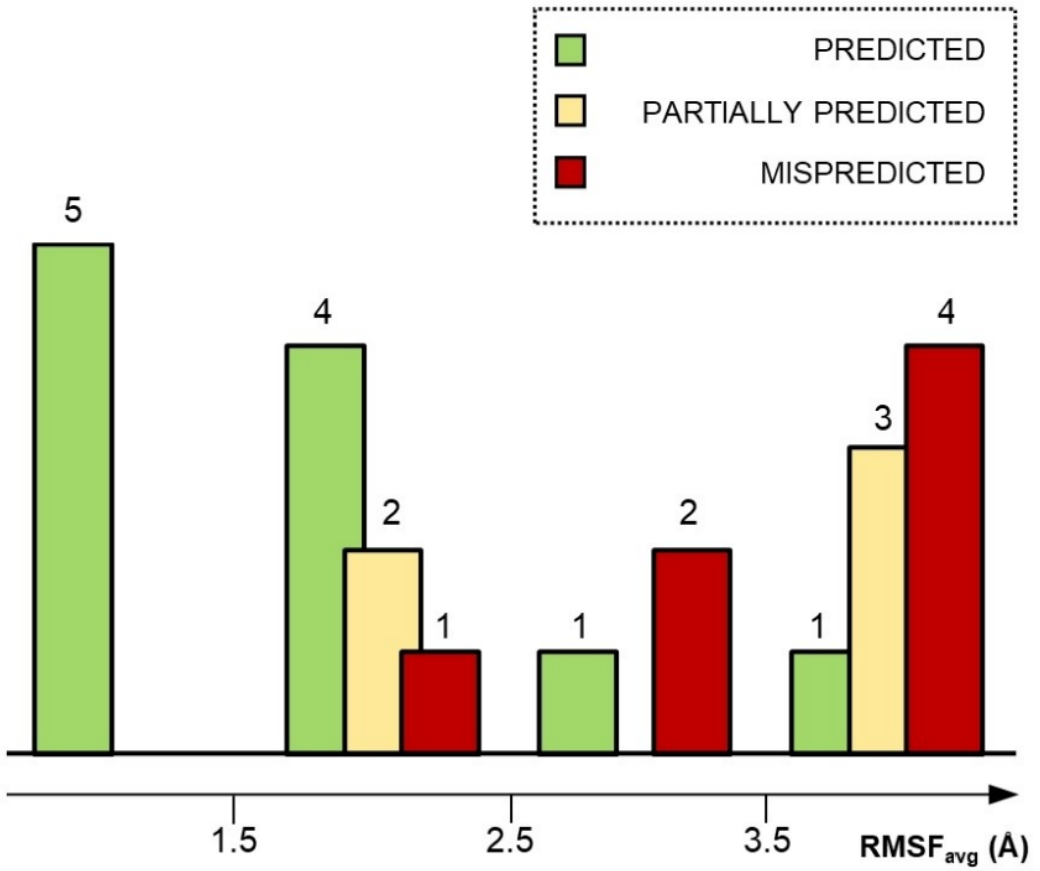


Figure 3. HT-SuMD predictions have been analyzed and related to the average fragment pose stability values ($RMSF_{avg}$) showed by each original crystallographic complex during the MD-based characterization study

HT-SuMD outcomes have therefore been compared, as reported in Figure 3, with the average values of atomic coordinates fluctuation (i.e. $RMSF_{avg}$) respectively showed by each crystallographic fragment in the classical MD study previously discussed. Intriguingly, a clear pattern is noticeable since almost the totality of the correctly predicted binding modes (9/11) has been recovered for those fragments characterized by marked structural stability, with an $RMSF_{avg}$ value lower than 2.5 Å. Above this empirical cut-off, consistently most of the incorrect predictions concentrate, thus corroborating the existence of an inverse relationship linking together the stability of a crystallographic final state and the ability to correctly anticipate it through MD-based approaches, as our protocol configure.

Fragment poses stability as a confidence metric.

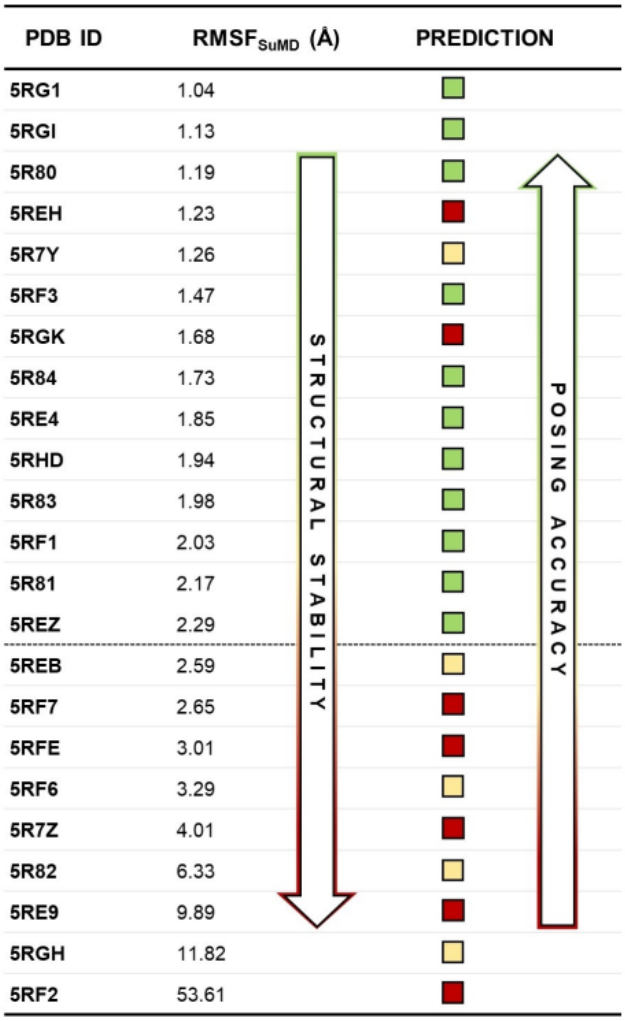


Figure 4. HT-SuMD predicted binding modes (i.e. the cluster of fragments conformations characterized by the lowest MM/GBSA value) have undergone an MD-based refinement step. The fragment poses stability of each prediction, measured as the $RMSF_{SuMD}$, has been exploited to rank HT-SuMD results, allowing in this way to efficiently prioritizing the correct binding modes at the expense of the incorrect ones. The dashed line delimits the empirical cut-off of 2.5 Å used to discriminate the reliability of HT-SuMD posing prediction

The relationship described above could therefore be exploited to drive the analysis and the interpretation of HT-SuMD results, providing an observable with which distinguish reliable binding modes predictions from decoys. To test this hypothesis, the results collected through HT-SuMD posing protocol were retrospectively evaluated simulating a real screening scenario, in which crystallographic references are not available. Hence, for each of the 23 M^{pro} fragments previously investigated through HT-SuMD, the binding mode with the lowest MM/ GBSA score was blindly selected, regardless of whether it corresponds to the original experimental pose. Then, multiple classical MD simulations 20 ns long were started from the predicted final states, to characterize their relative fragment pose stability. Results of this study have been summarized in Figure 4, sorting the data concerning the $RMSF_{SuMD}$ values, or the average fluctuations of SuMD-predicted binding poses, computed on the fragment's heavy atoms. A first interesting aspect to underline is how almost the totality of the correct binding modes anticipated by HT-SuMD (green-colored molecules) only undergoes a mild conformational perturbation during classical MD simulations, in agreement with the results described in the first part of the manuscript for the crystallographic complexes. On the contrary, incorrect binding mode (yellow and red-colored molecules) in most of the cases experience great lability when refined through MD simulations, sometimes even culminating in a spontaneous unbinding event of the fragment.

These observations corroborate the initial hypothesis, suggesting how a combination of HT-SuMD protocol for the posing of fragments with classical MD simulation for the refinement of results could represent an optimal operative pipeline, which allows overcoming some of the previously discussed methodological limitations. In this specific case indeed, the implementation of a geometric-dynamic property, namely the $RMSF_{SuMD}$, results extremely useful to qualitatively estimate the reliability of the in-silico predicted poses.

Observing the ranking reported in Figure 3, as the structural stability of the HT-SuMD predicted binding mode decreases, a worsening in posing accuracy occurs contextually. Intriguingly, also, in this case, 2.5 Å configure as a valuable empirical threshold which allows us to prioritize all the 11 correct fragment binding mode predictions. However, it is worth noting how the same cut-off is also responsible for the incorporation of three false positives, predictions characterized by remarkable structural stability, but which are nevertheless geometrically far from the crystallographic reference. For what concerns the fragment belonging to the PDB ID 5R7Y complex, HT-SuMD protocol has probably prioritized a metastable binding mode anticipating the experimental one, that has been nevertheless sampled through MD simulations but incorrectly scored by MM/GBSA. In the other two cases (PDB ID 5REH and 5RGK) the misprediction affects two fragments sharing a similar structure and interactivity. In the specific case of the 5REH complex, the HT-SuMD posing protocol has prioritized an alternative binding mode in which the pyridine portion of the fragment is correctly predicted, reproducing the key hydrogen bond interaction with H163 residue, while the remaining flexible portion is erroneously accommodated in the subsite S2 causing, as indicated in Figure 2, the high RMSD value of the cluster. This aspect is particularly interesting in the FBDD context, considering how the mild affinity profile characterizing these compounds could determine multiple recognition modes.

Conclusion

The elucidation of fragment binding modes in the early stages of FBDD campaigns still represents a tough medicinal chemistry task, which can be mitigated by the concomitant application of in-silico approaches. In this work, we have therefore investigated the geometric accuracy with which our recently developed computational protocol can reproduce experimentally solved fragment-receptor complexes. For this purpose, the XRC structures of 23 non-covalent fragments targeting SARS-CoV-2 M^{pro}, a pharmaceutical hot target in this actual COVID-19 pandemic, were exploited. HT-SuMD, as summarized in Figure 5, samples for each fragment multiple binding trajectories (Box 1), which are subsequently geometrically discretized through DBSCAN clustering and energetically evaluated using the MM/GBSA approach (Box 2). Our methodology was able to recover

and prioritize in almost half of the cases taken into consideration (48%) the original fragment bound geometry, with an accuracy comparable to that described for other MD-based posing approaches.

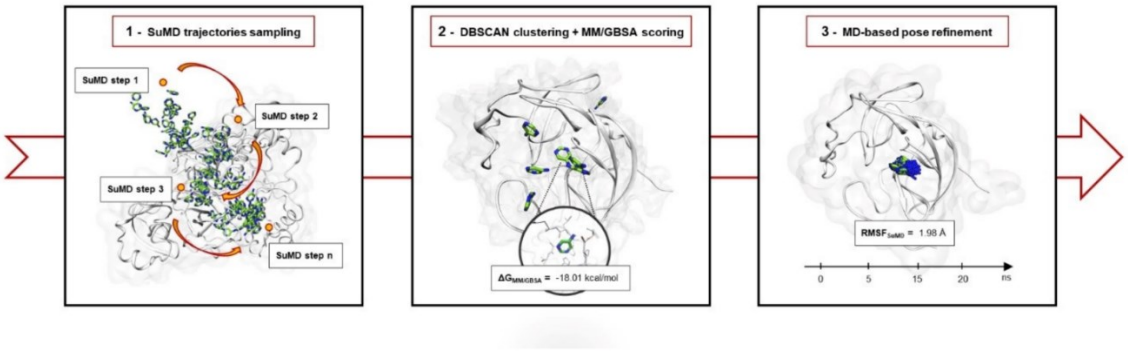


Figure 5. HT-SuMD protocol for the posing of fragments mainly consists in three operative steps, that are respectively summarized in this graphical workflow. In detail, supervised MD simulations are exploited to sample multiple binding trajectories for all the fragments analyzed (1), then DBSCAN clustering algorithm allows to identify of the most populated ligand conformation, which is energetically evaluated using MM/GBSA scoring method (2). The in-silico predicted binding modes finally undergo an MD-based refinement step, using the $RMSF_{SuMD}$ as a metric to qualitatively characterize the posing reliability.

Intriguingly, a clear correlation has been identified between HT-SuMD posing accuracy and the stability of the respective crystallographic complexes, with most of the correct binding modes predictions retrieved for those fragments characterized by a low $RMSF_{avg}$. In light of this aspect, a refinement step of HT-SuMD results through classical MD simulations has become an integrative part of our posing protocol (Figure 5 – Box 3). More specifically, the structural stability of the predicted binding mode, i.e. the $RMSF_{SuMD}$, has been exploited and validated as a metric to qualitatively estimate the reliability of each single in-silico prediction. In this way, it was possible to effectively rank and prioritize the 11 correct HT-SuMD binding poses while discharging the ones characterized by a marked instability that was mainly revealed as incorrect predictions. This concept is exemplified in Video1 (supplementary information), reporting how MM/GBSA, a thermodynamic-based approach, fails in distinguishing a correct form and incorrect fragment binding pose, while the subsequent MD refinement steps allow highlighting a marked difference between the two different predictions, in terms of $RMSF_{SuMD}$.

Despite these preliminary encouraging results, which must be certainly consolidated with further case studies, an improvement in the fragment posing accuracy is however still

desirable. From this perspective, the ever-increasing computing power that will be available in the next years coupled with the continuous optimization of the conformational sampling algorithm, as well as the force fields model used, could pave the way for the development of more accurate fragment posing protocols, that could massively impact many in-silico FBDD pipelines.

References

1. Murray, C. W. & Rees, D. C. The rise of fragment-based drug discovery. *Nat Chem* **1**, 187–192 (2009).

2. Erlanson, D. A., Fesik, S. W., Hubbard, R. E., Jahnke, W. & Jhoti, H. Twenty years on: the impact of fragments on drug discovery. *Nat. Rev. Drug Discov.* **15**, 605–619 (2016).

3. Jacquemard, C. & Kellenberger, E. A bright future for fragment-based drug discovery: what does it hold? *Expert Opin. Drug Discov.* **14**, 413–416 (2019).

4. Giordanetto, F., Jin, C., Willmore, L., Feher, M. & Shaw, D. E. Fragment Hits: What do They Look Like and How do They Bind? *J. Med. Chem.* **62**, 3381–3394 (2019).

5. Hall, R. J., Mortenson, P. N. & Murray, C. W. Efficient exploration of chemical space by fragment-based screening. *Prog. Biophys. Mol. Biol.* (2014) doi:10.1016/j.pbiomolbio.2014.09.007.

6. Hajduk, P. J. & Greer, J. A decade of fragment-based drug design: strategic advances and lessons learned. *Nat. Rev. Drug Discov.* **6**, 211–219 (2007).

7. Davis, B. J. & Roughley, S. D. Fragment-Based Lead Discovery. in 371–439 (2017). doi:10.1016/bs.armc.2017.07.002.

8. Ma, R., Wang, P., Wu, J. & Ruan, K. Process of Fragment-Based Lead Discovery—A Perspective from NMR. *Molecules* **21**, 854 (2016).

9. Davis, B. J. & Erlanson, D. A. Learning from our mistakes: The 'unknown knowns' in fragment screening. *Bioorg. Med. Chem. Lett.* **23**, 2844–2852 (2013).

10. Schiebel, J. *et al.* Six Biophysical Screening Methods Miss a Large Proportion of Crystallographically Discovered Fragment Hits: A Case Study. *ACS Chem. Biol.* **11**, 1693–1701 (2016).

11. Wielens, J. *et al.* Parallel Screening of Low Molecular Weight Fragment Libraries. *J. Biomol. Screen.* **18**, 147–159 (2013).

12. Sheng, C. & Zhang, W. Fragment Informatics and Computational Fragment-Based Drug Design: An Overview and Update. *Med. Res. Rev.* **33**, 554–598 (2013).

13. Mortier, J., Rakers, C., Frederick, R. & Wolber, G. Computational Tools for In Silico Fragment-Based Drug Design. *Curr. Top. Med. Chem.* **12**, 1935–1943 (2012).

14. de Souza Neto, L. R. *et al.* In silico Strategies to Support Fragment-to-Lead Optimization in Drug Discovery. *Frontiers in Chemistry* (2020) doi:10.3389/fchem.2020.00093.

15. Bissaro, M., Sturlese, M. & Moro, S. The rise of molecular simulations in fragment-based drug design (FBDD): an overview. *Drug Discov. Today* **25**, 1693–1701 (2020).

16. Ferrari, F. *et al.* HT-SuMD: making molecular dynamics simulations suitable for fragment-based screening. A comparative study with NMR. *J. Enzyme Inhib. Med. Chem.* (2021) doi:10.1080/14756366.2020.1838499.

17. No Title. Who. Coronavirus disease (COVID-19) Global epidemiological situation. .

18. Douangamath, A. *et al.* Crystallographic and electrophilic fragment screening of the SARS-CoV-2 main protease. *bioRxiv* 2020.05.27.118117 (2020) doi:10.1101/2020.05.27.118117.

19. T. C. M. Consortium, H. Achdout, A. Aimon, E. Bar-David, H. Barr, A. Ben-Shmuel, J. Bennett, M. L. Bobby, J. Brun, S. BVNBS, M. Calmiano, A. Carbery, E. Cattermole, J. D. Chodera, A. Clyde, J. E. Coffland, G. Cohen, J. Cole, A. Contini, L. Cox, M. Cvitkov, N. Z. COVID Moonshot: Open Science Discovery of SARS-CoV-2 Main Protease Inhibitors by Combining Crowdsourcing, High-Throughput Experiments, Computational Simulations, and Machine Learning.

20. De Vivo, M., Masetti, M., Bottegoni, G. & Cavalli, A. Role of Molecular Dynamics and Related Methods in Drug Discovery. *Journal of Medicinal Chemistry* vol. 59 4035–4061 (2016).

21. Salmaso, V. & Moro, S. Bridging molecular docking to molecular dynamics in exploring ligand-protein recognition process: An overview. *Frontiers in Pharmacology* vol. 9 (2018).

22. Berman, H. M. The Protein Data Bank http://www.rcsb.org/pdb/. *Nucleic Acids Res.* **28**, 235–242 (2000).

23. Bolcato, G., Bissaro, M., Pavan, M., Sturlese, M. & Moro, S. Targeting the coronavirus SARS-CoV-2: computational insights into the mechanism of action of the protease inhibitors lopinavir, ritonavir and nelfinavir. *Sci. Rep.* **10**, 20927 (2020).

24. Maier, J. A. *et al.* ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* (2015) doi:10.1021/acs.jctc.5b00255.

25. Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A. & Case, D. A. Development and testing of a general amber force field. *J. Comput. Chem.* **25**, 1157–1174 (2004).

26. Sabbadin, D. & Moro, S. Supervised Molecular Dynamics (SuMD) as a Helpful Tool To Depict GPCR–Ligand Recognition Pathway in a Nanosecond Time Scale. *J. Chem. Inf. Model.* **54**, 372–376 (2014).

27. Cuzzolin, A. *et al.* Deciphering the Complexity of Ligand-Protein Recognition Pathways Using Supervised Molecular Dynamics (SuMD) Simulations. *J. Chem. Inf. Model.* **56**, 687–705 (2016).

28. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* (2011).

29. Ester, M., H. P. Kriegel, J. Sander, X. X. *A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise in Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*. (AAAI press, 1996).

30. D.A. Case, H.M. Aktulga, K. Belfon, I.Y. Ben-Shalom, S.R. Brozell, D.S. Cerutti, T.E. Cheatham, III, G.A. Cisneros, V.W.D. Cruzeiro, T.A. Darden, R.E. Duke, G. Giambasu, M.K. Gilson, H. Gohlke, A.W. Goetz, R. Harris, S. Izadi, S.A. Izmailov, C. Jin, K. Ka, and P. A. K. Amber 2021. (2020).

31. Mobley, D. L. & Dill, K. A. Binding of Small-Molecule Ligands to Proteins: "What You See" Is Not Always "What You Get". *Structure* **17**, 489–498 (2009).

32. Verdonk, M. L. *et al.* Docking performance of fragments and druglike compounds. *J. Med. Chem.* (2011) doi:10.1021/jm200558u.

33. Lim, N. M., Osato, M., Warren, G. L. & Mobley, D. L. Fragment Pose Prediction Using Non-equilibrium Candidate Monte Carlo and Molecular Dynamics Simulations. *J. Chem. Theory Comput.* (2020) doi:10.1021/acs.jctc.9b01096.

34. Majewski, M. & Barril, X. Structural Stability Predicts the Binding Mode of Protein–Ligand Complexes. *J. Chem. Inf. Model.* **60**, 1644–1651 (2020).

Targeting the Coronavirus SARS-CoV-2: computational insights into the mechanism of action of the protease inhibitors Lopinavir, Ritonavir and Nelfinavir.

Giovanni Bolcato, Maicol Bissaro, **Matteo Pavan**, Mattia Sturlese and Stefano Moro

Bolcato, G., Bissaro, M., Pavan, M., Sturlese, M. & Moro, S. Targeting the coronavirus SARS-CoV-2: computational insights into the mechanism of action of the protease inhibitors lopinavir, ritonavir and nelfinavir. *Sci Rep* **10**, 20927 (2020).

Abstract

Coronavirus SARS-CoV-2 is a recently discovered single-stranded RNA (ssRNA) betacoronavirus, responsible for a severe respiratory disease known as coronavirus disease 2019 (COVID-19), which is rapidly spreading. Chinese health authorities, as a response to the lack of an effective therapeutic strategy, started to investigate the use of lopinavir and ritonavir, previously optimized for the treatment and prevention of HIV/AIDS viral infection. Despite the clinical use of these two drugs, no information regarding their possible mechanism of action at the molecular level is still known for SARS-CoV-2. Very recently, the crystallographic structure of the SARS-CoV-2 main protease (M^{pro}), also known as C30 Endopeptidase, was published. Starting from this essential structural information, in the present work we have exploited Supervised Molecular Dynamics (SuMD), an emerging computational technique that allows investigating at an atomic level the recognition process of a ligand from its unbound to the final bound state. In this research, we provided molecular insight on the whole recognition pathway of Lopinavir, Ritonavir, and Nelfinavir, three potential C30 Endopeptidase inhibitors, with the last one taken into consideration due to the promising in-vitro activity shown against the structurally related SARS-CoV protease.

1. Introduction

Coronavirus SARS-CoV-2, previously known as 2019-nCoV, is a recently discovered single-stranded RNA (ssRNA) betacoronavirus, responsible for a severe pathological condition known as coronavirus disease 2019 (COVID-19).[1] Since it was first identified in December 2019, this novel coronavirus has rapidly spread all around the world, being since now responsible for the death of nearly one milion of people, which have lost their lives due to a severe respiratory illness.[2]

The first outbreak of this new disease originally took place in the city of Wuhan (China), rapidly spreading in the southeast of Asia and, recently, in other continents like Europe, North America and Africa.[1] The astonishing rate at which COVID is expanding compared to previous coronavirus related diseases (SARS-CoV and MERS-CoV), in conjunction with the absence of approved drugs or effective therapeutic approaches for its treatment, has gathered the attention of the international community, which is promoting a cooperative effort to face this emergency.[3,4] On January 2020 indeed, the International Health Regulations Emergency Committee of the World Health Organization declared the outbreak a “public health emergency of international concern” in responding to SARS-CoV-2.

Unfortunately, the timeline characterizing a typical drug discovery process badly couples with the urgency of finding a cure for the already infected patients as rapidly as possible. In this kind of scenario, it is of paramount importance to accelerate the early stages of the drug discovery process for COVID-19 treatment, and for all possible future emergencies.[5]

The early isolation of the SARS-CoV-2 genome from ill patients represented a first crucial outcome, making it possible to highlight an important sequence identity (~80% of conserved nucleotides) with respect to the original SARS-CoV epidemic virus.[6] In light of this similarity, some therapeutic strategies could be inherited from other genetically related CoV diseases.

A possible target is for example represented by structural viral proteins, therefore interfering with the assembly and the internalization of the pathogen into the host, which was shown to occur also in this case through the Angiotensin-converting enzyme II (ACE2) receptor. From this perspective, the development of a vaccine is desirable, and it is

foreseen that the first candidates will be advanced to clinical phase I around mid-2020.[7–9]

In the meantime, however, a great effort involves the targeting of non-structural viral proteins which are instead essential for the viral replication and the maturation processes, thus representing a crucial and specific target for anti-COVID drug development.[3,10] In this regard, the crystallographic structure of the SARS-CoV-2 main protease (M^{pro}), also known as C30 Endopeptidase, was elucidated and made available to the scientific community with impressive timing, just a few weeks after the first COVID-19 outbreak (PDB ID: 6LU7). The structural characterization of the protease, which shares 96.1% of its sequence with those of SARS-CoV, has revealed a highly conserved architecture of the catalytic binding site.

As a result, Structure-Based Drug Discovery techniques (SBDD) can now be applied to efficiently speed up the rational identification of putative M^{pro} inhibitors or to drive the repurposing process of known therapy. This latter route is particularly attractive, as it allows to significantly shrink the time required to access the first phases of clinical trials, without compromising patient safety. A multitude of research groups has begun to apply computational approaches, such as molecular docking based virtual screening (VS), to evaluate already approved FDA approved drugs against the aforementioned viral protease.[11–14] Many of these studies have found convergence in suggesting compounds inhibitors of the human immunodeficiency viruses (HIV) as possible anti-COVID candidates; this is surprising considering the important structural differences exiting among these two homologous enzymes. The repositioning of HIV antiviral drugs for the treatment of coronavirus infections found, however, a foundation in the scientific literature of the past 20 years. Some of these compounds have therefore been experimentally investigated, showing promising activity, both in the case of SARS-CoV and MERS-CoV outbreak.[15,16]

Moreover, at least three randomized clinical trials are currently been held in China in order to evaluate the therapeutic efficacy of Lopinavir and Ritonavir, a combination of HIV protease inhibitors, in COVID-19 treatment.[7] In this perspective and preliminary computational research, we took advantage of the recently solved crystallographic structure of SARS-CoV-2 M^{pro} to perform a cutting edge *in-silico* investigation.

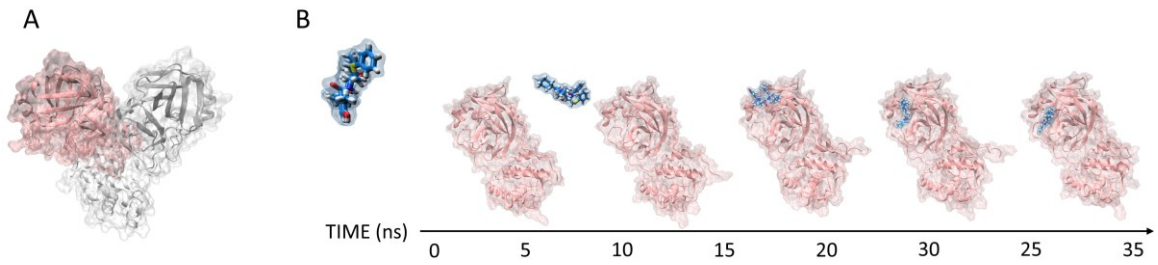


Figure 1. The crystallographic structure of SARS-CoV-2 C30 Endopeptidase exploited in our computational investigation (PDB ID : 6LU7) is reported in Panel A. The two different monomers composing the homodimeric proteases are depicted using different colors (i.e. pink and white respectively for monomer A and B). As represented on Panel B, only one chain (monomer A) was exploited in our SuMD protocol to describe the putative inhibitor binding mechanism.

Supervised Molecular Dynamics (SuMD), an emerging technique allowing to investigate at an atomic level of detail the molecular recognition process, was exploited to characterize the putative binding mechanism of three HIV protease inhibitors.[17–19] In detail, along with the aforementioned combination of Lopinavir and Ritonavir, also Nelfinavir was taken into consideration, due to the promising in-vitro activity shown by this compound against the structurally related SARS-CoV protease.[20] SuMD protocol implements a tabu-like algorithm that controls the sampling of short unbiased MD trajectories, each of which hundreds of picoseconds (ps) long. In detail, simulation steps are accepted only when describing a ligand approaching a known binding site, otherwise, the simulation is discharged and restarted from the previous coordinate set. The combination of all productive SuMD simulation steps represents, therefore, a putative molecular recognition trajectory collected, differently from brute force MD, in a very competitive computational time not exceeding the nanoseconds (ns) timescale. Contrary to molecular docking, SuMD simulations fully consider both the flexibility characterizing the protein target during the binding event and the contribution played by water molecules during the recognition. Moreover, the study is not limited to a possible final state but allows peeking dynamically at the whole process of recognition, also identifying putative metastable binding sites.

2. Results

The combination of the structurally related antiviral protease inhibitor Lopinavir and Ritonavir, commercially known with the name Kaletra, represent an effective therapeutic weapon ensuring an adequate and durable suppression of viral load in HIV positive patients. The synergistic coadministration of these two compounds exploits low-dosage concentration of Ritonavir which, inhibiting the metabolic inactivation of Lopinavir, acts as a pharmacokinetic enhancer.[21] Following a preliminary favorable clinical response in SARS-CoV related diseases, the combination of the drug is currently under investigation also against SARS-CoV-2, with at least three randomized clinical trials undergoing with Chinese infected patients.[15] In our computational study, we considered Lopinavir and Ritonavir as two independent inhibitors, performing separate SuMD binding simulations, which results are herein reported an analyzed.

As highlighted in Figure 2 (Panel B) about 20 ns proved to be sufficient to sample a putative Lopinavir recognition trajectory with SARS-CoV-2 protease. At a distance of about 15 Å from the binding site, the first molecular contacts are recorded (Figure 2 – Panel C, D and Video 1), which guide the subsequent accommodation of the ligand into the catalytic site. The predicted final state is stabilized by a double hydrogen bond interaction with residue Glu166 backbone, tightly anchoring the inhibitor (Figure2 - Panel A). This strong and persistent interaction (Figure 2 – Panel B) is known to be crucial in many SARS-CoV complexes and moreover, was also found to stabilize the covalent peptidomimetic compound crystallized in the recently published SARS-CoV-2 M^{pro} structure. In addition, the cyclic urea moiety of Lopinavir mediates a hydrogen bond interaction with the side chain of Gln189, another residue whose importance has been elucidated by means of several SARS-CoV three-dimensional complexes.

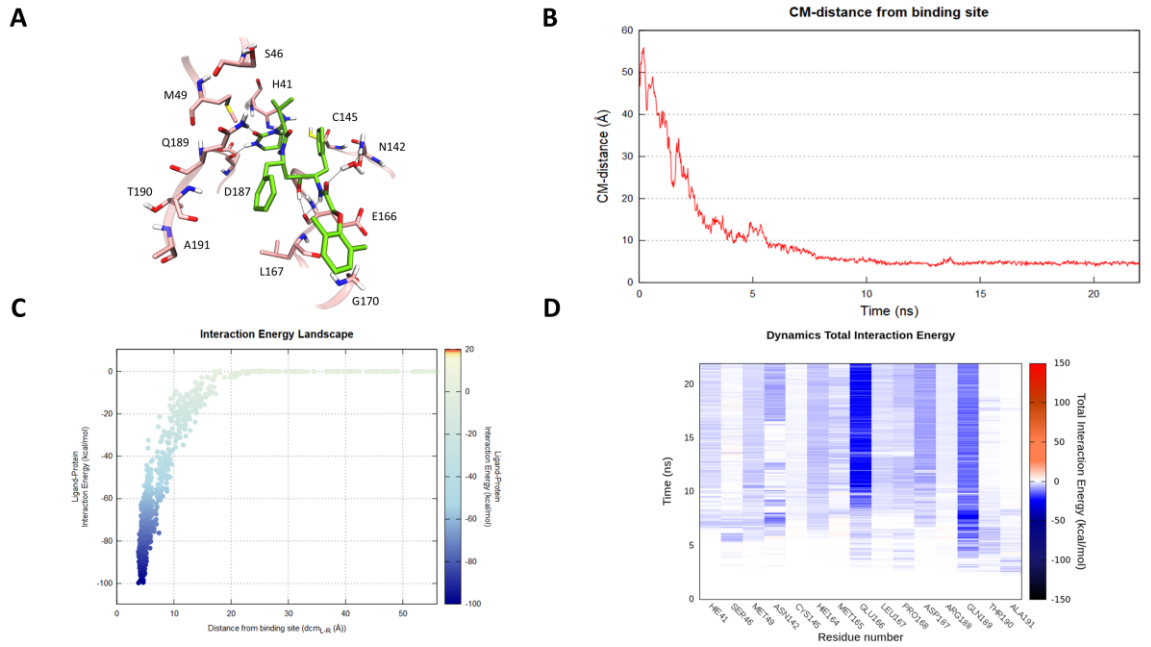


Figure 2. This panel summarizes the recognition pathway of Lopinavir against the SARS-CoV-2 main protease. (A) Lopinavir conformation sampled in the last frame of the SuMD trajectory (green-colored molecule). The residues surrounding the binding site are reported in pink color. (B) Distance between the ligand center of mass (Cm) and the catalytic binding site of the M^{pro}during the SuMD simulation. (C) Interaction Energy Landscape describing the protein-ligand recognition process; values are arranged according to the distances between ligand and protein target mass centers. (D) Dynamic total interaction energy (electrostatic + vdW) computed for most contacted M^{pro} residues.

Despite the modest pharmacodynamic contribution made by Ritonavir in the combined formulation under investigation by the Chinese scientific community, in which the drugs act as a pharmacokinetic enhancer rather than a protease inhibitor, we still tried to elucidate its putative molecular recognition pathway. Also, in this case, 20 ns of SuMD simulation time were sufficient to sample a binding trajectory (Figure 3 – Panel B). Although some key interactions – i.e. hydrogen bond network with residue Glu166 and Gln189 – are appreciable also in this final state (Figure 3 – Panel A,D and Video 2), a comparative analysis of the Interaction Energy Landscape graphs (Panel C of Figure 2 and 3) suggests lower energy stability of the SuMD predicted binding mode, when compared with that characterizing Lopinavir. A reason could be seeking on the non-optimal accommodation of Ritonavir urea moiety, which floats outside the binding site exposed to the bulk solvent during all the simulation (Video 2).

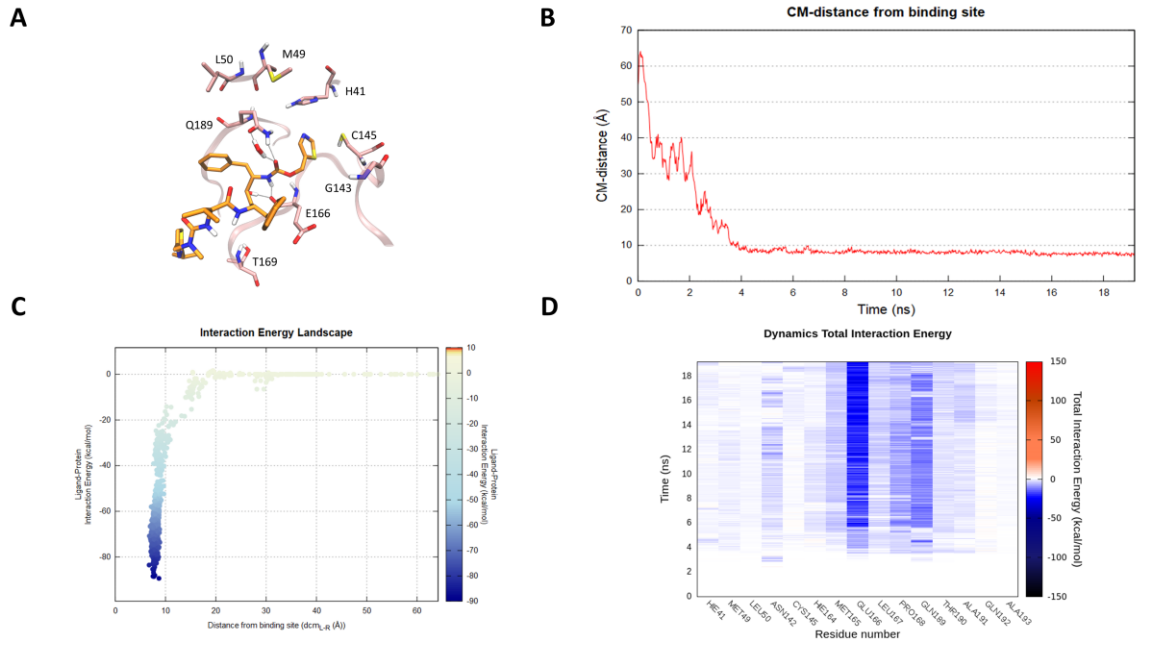


Figure 3. This panel summarizes the recognition pathway of Ritonavir against the SARS-CoV-2 main protease. (A) Ritonavir conformation sampled in the last frame of the SuMD trajectory (orange-colored molecule). The residues surrounding the binding site are reported in pink color. (B) Distance between the ligand center of mass (Cm) and the catalytic binding site of the M^{pro}during the SuMD simulation. (C) Interaction Energy Landscape describing the protein-ligand recognition process; values are arranged according to the distances between ligand and protein target mass centers. (D) Dynamic total interaction energy (electrostatic + vdW) computed for most contacted M^{pro} residues.

In light of the promising experimental results shown by Nelfinavir, which milded the cytopathic effect induced by SARS-CoV infection strongly inhibiting the virus replication, we decided to computationally evaluate its possible molecular recognition mechanism also against SARS-CoV-2 protease. As reported in Figure 4 (Panel B), a slightly longer SuMD simulation was necessary to fully describe a putative Nelfinavir binding trajectory. Once it has approached the vestibular region of the protease catalytic site, the ligand spends the first 20 ns negotiating the accommodation with a series of polar residues with which it mediates intermittent interactions, as highlighted in the interaction energy fingerprint (Figure 4 - Panel D, Video 3). The importance of this metastable site is also depicted in the Interaction Energy Landscape (IEL) graphic (Figure 4 – Panel C, Figure S3 – Panel A and B,), from which it is possible to notice a highly populated region presenting ligand-protein interaction energy comparable to the final states previously described for the other two inhibitors. The last 10 ns of the simulation were characterized by a series of

conformational rearrangements, which resulted in an optimal Nelfinavir accommodation within the protease binding cleft stabilized through a dense hydrogen bond network, tightly anchoring the inhibitor to the protease. As shown in Figure 4 (Panel A), SuMD predicted binding mode of Nelfinavir is characterized by great analogies with that of the originally crystallized covalent peptidomimetic compound. Residues His164, Glu166, Gln189, Thr190, and Gln196 mediate a series of directed or water-bridged hydrogen bonds interactions. Moreover, as highlighted in Figure 4 (Panel D), on the last ns of the simulation a stabilizing salt bridge interaction occurs between the side chain of residue Glu166 and the octahydro-1H-isoquinoline charged moiety of Nelfinavir. Intriguingly, mutagenesis studies have corroborated the crucial role played by this residue. Mutation of Glu166 correlated therefore with the block of substrate-induced dimerization of the main protease, both in SARS-CoV and in MERS-CoV.[22,23]

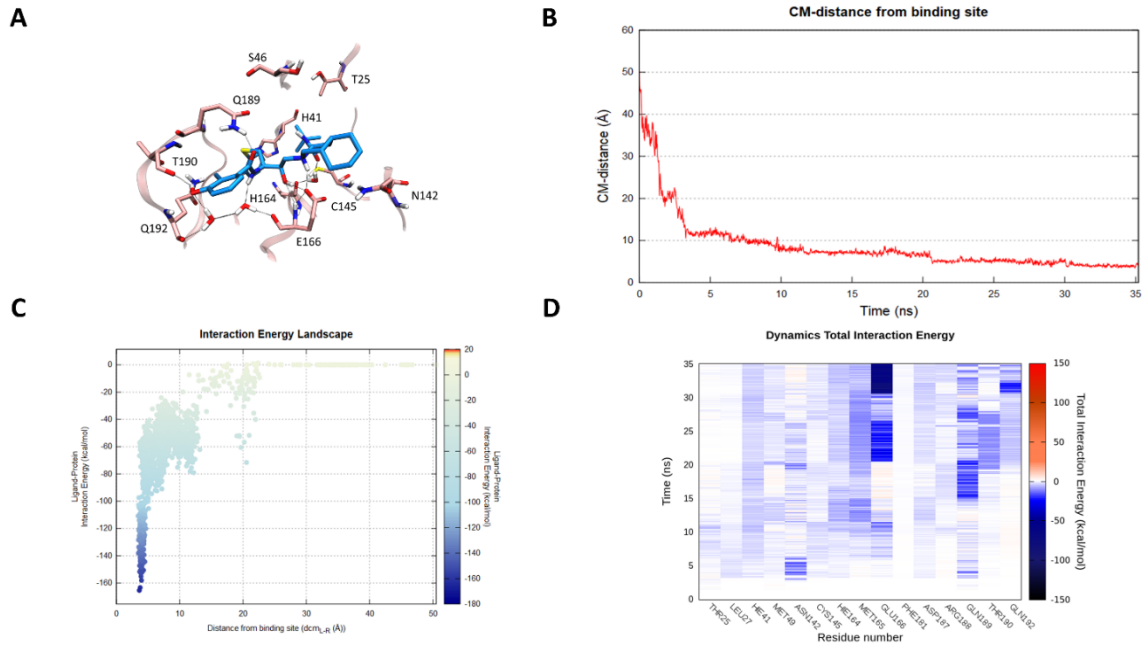


Figure 4. This panel summarizes the recognition pathway of Nelfinavir against the SARS-CoV-2 main protease. (A) Nelfinavir conformation sampled in the last frame of the SuMD trajectory (cyan-colored molecule). The residues surrounding the binding site are reported in pink color. (B) Distance between the ligand center of mass (Cm) and the catalytic binding site of the M^{pro} during the SuMD simulation. (C) Interaction Energy Landscape describing the protein-ligand recognition process; values are arranged according to the distances between ligand and protein target mass centers. (D) Dynamic total interaction energy (electrostatic + vdW) computed for most contacted M^{pro} residues.

3. Discussion

In the last two decades, three major outbreaks of coronavirus-related diseases SARS-CoV, MERS-CoV and ultimately SARS-CoV-2 have been responsible for significant public health issues, along with dramatic social-economic consequences. The process of drug discovery often undergoes timelines which are difficult to reconcile with the urgency and the need to provide an effective therapeutic response to such an emergency health situation. Drug repurposing could represent a viable possibility, and this is the case for some anti-HIV compounds targeting SARS-CoV-2 C30 Endopeptidase. The molecular basis underneath their therapeutic action remains however often obscure. In this preliminary computational investigation, we have taken advantage of the recently published crystallographic structure of SARS-CoV-2 M^{pro} to investigate the putative binding mechanism of three antiviral compounds, previously designed as selective HIV protease inhibitors and now under investigation as anti-COVID-19 emergency treatments. SuMD protocol was in detail exploited to collect, for each of the three inhibitors, MD simulation describing the possible mechanism of molecular recognition, thus providing an atomistic insight to interpret their data of therapeutic efficacy. An interesting aspect is represented by the speed of this approach: a few days of calculation in a modest GPU cluster allowed to collect a multitude of simulations, from which it was possible to hypothesize the recognition mechanism of Lopinavir, Ritonavir, and Nelfinavir. An approach of this type, therefore, becomes crucial in all emergencies, making it possible to overcome the lack of structural data to guide and understand the possible repositioning of already approved drugs. In this particular case study,the SuMD protocol not only allowed to hypothesize a possible recognition method for each antiviral but also to advance some preliminary comparative considerations. Nelfinavir, in particular, showed the best fitting for the catalytic site of SARS-CoV-2 M^{pro}, establishing an interactions network similar to those elucidated in the crystallographic complex for the covalent peptidomimetic compound N3. More specifically, the phenyl sulfanyl moiety of the protease inhibitor at the end of the simulation was completely buried within the hydrophobic sub-pocket S2, which is delimited by residues His41, Cys44, Met49 and Met165. The stabilizing vdW contribution mediated by these residues has been dynamically mapped during the entire simulation

and it is appreciable in Figure S3. Encouragingly, a recent fragment crystallographic screening has highlighted how this site, precisely renamed “aromatic wheel”, consistently accommodates aromatic fragments mediating hydrophobic interactions with the surrounding residues.[24] Furthermore, Nelfinavir hydroxyl group engages a hydrogen bond interaction with the carbonyl backbone of Glu166, a key residue found to stabilize most of the aforementioned non-covalent fragments as well as many covalent peptidomimetic inhibitors. The optimal interactive network differentiating Nelfinavir from the other two protease inhibitors is probably responsible for its total interaction energy which, as reported in Figure 4 (Panel C), is quantitatively greater than that computed for Lopinavir and Ritonavir (Figure 2 and 3 – Panel C). Intriguingly, this in-silico hypothesis has recently found two independent experimental validations, which have highlighted a mild inhibitory activity of Nelfinavir against the SARS-CoV-2 M^{pro} (estimated between 250 and 600 μM).[25,26]

4. Methods

4.1 Software overview

MOE suite (Molecular Operating Environment, version 2018.0101) was used to perform most of the general molecular modeling operations, such as proteins and ligands preparation.[27] All these operations have been performed on an 8 CPU (Intel® Xeon® CPU E5-1620 3.50 GHz) Linux workstation. Molecular dynamics (MD) simulations were performed with an ACEMD3 engine on an Nvidia GPU cluster composed of 20 NVIDIA drivers, whose models go from GTX 1080 to Titan V.[28] For all the simulations, the ff14SB force field was adopted to describe C30 Endopeptidase protein while general Amber force field (GAFF) was adopted to parameterize small organic molecules.[29–31]

4.2 Structures Preparation

The three-dimensional coordinates of C30 Endopeptidase protein in complex with a covalent peptidomimetic inhibitor (N3) were retrieved from the RCSB PDB database and prepared for SuMD simulations as herein described.[32] Considering the perfect symmetry that characterizes this homodimeric protein, and therefore its two catalytic binding sites, only one of the two monomers was used in this computational investigation. Once the

covalent ligand was removed, residue Cys145 was restored to its reduced form. Protein was then processed by means of MOE protein structure preparation tool: residues missing atoms were built according to AMBER14 force field topology. Missing hydrogen atoms were added to X-ray derived complexes and appropriate ionization states were assigned by the Protonate-3D tool.[33] The coordinates of three antiviral compounds were prepared through MOE builder tool and subsequently moved at least 30 Å away from the catalytic protease binding cleft, a distance bigger than the electrostatic cut-off term used in the simulation (9 Å with Amber force field) to avoid premature interaction at the initial phases of the SuMD simulations.

4.3 Solvated System Setup and Equilibration

Each system investigated by means of SuMD contains a C30 Endopeptidase target macromolecule and respectively one of the three HIV antiviral compounds taken into consideration in this study, moved far away from the protein binding site as previously described. The systems were explicitly solvated by a cubic water box with cell borders placed at least 15 Å away from any protein/ligand atom, using TIP3P as a water model. To neutralize the total charge of each system, Na^+/Cl^- counterions were added to a final salt concentration of 0.154 M. The systems were energy minimized by 500 steps with the conjugate-gradient method, then 500000 steps (1 ns) of NVE followed by 500000 steps (1 ns) of NPT simulations were carried out, both using 2 fs as time step and applying harmonic positional constraints on protease and ligands heavy atoms by a force constant of 1 kcal mol^{-1} $Å^{-2}$, gradually reduced with a scaling factor of 0.1. During this step, the temperature was maintained at 310 K by a Langevin thermostat with low dumping of 1 ps^{-1} and the pressure at 1 atm by a Monte Carlo barostat[34]. The M-SHAKE algorithm was applied to constrain the bond lengths involving hydrogen atoms. The particle-mesh Ewald (PME) method was exploited to calculate electrostatic interactions with a cubic spline interpolation and 1 Å grid spacing, and a 9.0 Å cutoff was applied for Lennard-Jones interactions [35].

4.4 Supervised Molecular Dynamics (SuMD) Simulations

SuMD code, in this implementation, is written in Python and exploits the ProDy python package to perform the geometrical ligand-target supervision process [36]. SuMD protocol reduces the timescale, and consequently the computational effort, required to sample a binding event in the range of nanoseconds, instead of hundreds of nanoseconds or microseconds usually necessary with unbiased MD. Sampling is improved by applying a tabu-like algorithm that monitors the distance between the ligand center of mass with respect to the protein binding site, during short unbiased MD simulations of 600 ps. Once a SuMD step has been collected, the distance points calculated at regular time intervals are fitted into a linear function. Only productive MD steps are maintained, those in which the computed slope is negative, indicating a ligand approach toward the protease catalytic binding site. Otherwise, the simulation is restarted by randomly assigning the atomic velocities. Supervision algorithm controlled the sampling of short simulations until the distance between the ligand and the protein binding site dropped below 5 Å, then was disabled, and a classical MD simulation was performed. For each case study up to a maximum of ten SuMD binding simulations were collected, of which only the best was thoroughly analyzed and discussed in the manuscript.

4.5 SuMD Trajectories Analysis

All the SuMD trajectories collected were analyzed by an in-house tool written in tcl and python languages, as described in the original publication[19]. Briefly, the dimension of each trajectory was reduced saving MD frames at a 20 ps interval, each trajectory was then superposed and aligned on the proteaseCα atoms of the first frames and wrapped into an image of the system simulated under periodic boundary condition. The molecular recognition was monitored by calculating for each simulation step the distance between the catalytic binding site and the center of mass of the ligand taken into consideration (Figure F2 to F4 – Panel A). A ligand-protein interaction energy estimation during the recognition process was calculated using an NAMD engine, plotting the ligand-receptor interaction energy values over time.[37] These values were also arranged according to the distances between ligand and protease binding site mass centers in the Interaction Energy Landscape plots (Figure F2 to F4 – Panel B). Here, the distances between mass centers are

reported on the x-axis, while the ligand-receptor interaction energy values on the y-axis, and are rendered by a colorimetric scale going from blue to red for negative to positive energetic values. These graphs allow evaluating the variation of the interaction energy profile at different ligand-protein distances, helping to individuate meta-stable binding states during the binding process. Furthermore, for each target investigated in this work, the residues within a distance of 4 Å from the respective ligand atoms were dynamically selected, to qualitatively and quantitatively evaluated the number of contacts during the entire binding process. The most contacted residues were thus selected, to compute a per-residues electrostatic and vdW interaction energy contribution with the protease target. NAMD was used for post-processing computation of electrostatic interactions, using AMBER ff14SB force field. The cumulative electrostatic interactions were computed for the same target residues by summing the energy values frame by frame along the trajectory, and the resulting graphs were reported at the lower-right of movies provided on supplementary material (Video V1 to V3). Representations of the molecular structures were prepared with VMD software [38].

4.6 SuMD videos

Each video is composed of four synchronized and animated panels that depict the molecular trajectory obtained by the SuMD simulation considering different aspects of the simulation. The time evolution is reported on an ns scale. In the first panel (upper-left), the molecular representation of the SARS-CoV-2 main protease is shown. The protein backbone is represented by the ribbon style (pink color) and the residues within 4 Å of each ligand investigated are shown in green, orange and cyan colors respectively for Lopinavir, Ritonavir, and Nelfinavir. In the second panel (upper-right), the dynamic distance of each ligand center of mass (CM) from the respective protein catalytic binding site during the trajectory is reported. In the third panel (lower-left), the ligand-protein interaction energy profile is reported. The animated red circle highlights the value of the corresponding frame. The trend is depicted by a continuous black line obtained by smoothing the raw data (grey circles) using a Bezier curve procedure. In the fourth panel (lower-right) cumulative electrostatic interactions are reported for the 15 protein residues most contacted by each ligand during the whole simulation.

References

1. Guarner, J. Three Emerging Coronaviruses in Two DecadesThe Story of SARS, MERS, and Now COVID-19. *Am. J. Clin. Pathol.* doi:10.1093/AJCP/AQAA029

2. Who. *Coronavirus disease (COVID-19) Global epidemiological situation*.

3. Zhang, L. & Liu, Y. Potential Interventions for Novel Coronavirus in China: A Systemic Review. *J. Med. Virol.* jmv.25707 (2020). doi:10.1002/jmv.25707

4. Heymann, D. L., Shindo, N. & WHO Scientific and Technical Advisory Group for Infectious Hazards. COVID-19: what is next for public health? *Lancet (London, England)* **395**, 542–545 (2020).

5. Mani, D., Wadhwani, A. & Krishnamurthy, P. T. Drug Repurposing in Antiviral Research: A Current Scenario. *J. Young Pharm.* **11**, 117–121 (2019).

6. Gralinski, L. E. & Menachery, V. D. Return of the Coronavirus: 2019-nCoV. *Viruses* **12**, 135 (2020).

7. Keener, A. B. Four ways researchers are responding to the COVID-19 outbreak. *Nat. Med.* (2020). doi:10.1038/d41591-020-00002-4

8. Letko, M. & Munster, V. Functional assessment of cell entry and receptor usage for lineage B β-coronaviruses, including 2019-nCoV. *bioRxiv* 2020.01.22.915660 (2020). doi:10.1101/2020.01.22.915660

9. Wrapp, D. *et al.* Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science* (2020). doi:10.1126/science.abb2507

10. Anand, K., Yang, H., Bartlam, M., Rao, Z. & Hilgenfeld, R. Coronavirus main proteinase: target for antiviral drug therapy. in *Coronaviruses with Special Emphasis on First Insights Concerning SARS* 173–199 (Birkhäuser-Verlag, 2005). doi:10.1007/3-7643-7339-3_9

11. Li, Y. *et al.* Therapeutic Drugs Targeting 2019-nCoV Main Protease by High-Throughput Screening. *bioRxiv* 2020.01.28.922922 (2020). doi:10.1101/2020.01.28.922922

12. Xu, Z. *et al.* Nelfinavir was predicted to be a potential inhibitor of 2019-nCov main protease by an integrative approach combining homology modelling, molecular docking and binding free energy calculation. *bioRxiv* 2020.01.27.921627 (2020). doi:10.1101/2020.01.27.921627

13. Liu, X. & Wang, X.-J. Potential inhibitors for 2019-nCoV coronavirus M protease from clinically approved medicines. *bioRxiv* 2020.01.29.924100 (2020). doi:10.1101/2020.01.29.924100

14. Contini, A. Virtual Screening of an FDA Approved Drugs Database on Two COVID-19 Coronavirus Proteins. (2020). doi:10.26434/CHEMRXIV.11847381.V1

15. Chu, C. M. *et al.* Role of lopinavir/ritonavir in the treatment of SARS: Initial virological and clinical findings. *Thorax* **59**, 252–256 (2004).

16. Sheahan, T. P. *et al.* Comparative therapeutic efficacy of remdesivir and combination lopinavir, ritonavir, and interferon beta against MERS-CoV. *Nat. Commun.* **11**, 1–14 (2020).

17. Sabbadin, D. & Moro, S. Supervised molecular dynamics (SuMD) as a helpful tool to depict GPCR–ligand recognition pathway in a nanosecond time scale. *J. Chem. Inf. Model.* **54**, 372–376 (2014).

18. Cuzzolin, A. *et al.* Deciphering the Complexity of Ligand-Protein Recognition Pathways Using Supervised Molecular Dynamics (SuMD) Simulations. *J. Chem. Inf. Model.* **56**, 687–705 (2016).

19. Salmaso, V., Sturlese, M., Cuzzolin, A. & Moro, S. Exploring Protein-Peptide Recognition Pathways Using a Supervised Molecular Dynamics Approach. *Structure* **25**, 655-662.e2 (2017).

20. Yamamoto, N. *et al.* HIV protease inhibitor nelfinavir inhibits replication of SARS-associated coronavirus. *Biochem. Biophys. Res. Commun.* **318**, 719–725 (2004).

21. Cvetkovic, R. S. & Goa, K. L. Lopinavir/ritonavir: A review of its use in the management of HIV infection. *Drugs* **63**, 769–802 (2003).

22. Cheng, S. C., Chang, G. G. & Chou, C. Y. Mutation of glu-166 blocks the substrate-induced dimerization of SARS coronavirus main protease. *Biophys. J.* **98**, 1327–1336 (2010).

23. Ho, B. L. *et al.* Critical assessment of the important residues involved in the dimerization and catalysis of MERS Coronavirus Main Protease. *PLoS One* **10**, (2015).

24. Douangamath, A. *et al.* Crystallographic and electrophilic fragment screening of the SARS-CoV-2 main protease. *bioRxiv* 2020.05.27.118117 (2020). doi:10.1101/2020.05.27.118117

25. Ghahremanpour, M. M. *et al.* Identification of 14 Known Drugs as Inhibitors of the Main Protease of SARS-CoV-2. doi:10.1101/2020.08.28.271957

26. Vatansever, E. C. *et al.* Targeting the SARS-CoV-2 Main Protease to Repurpose Drugs for COVID-19. *bioRxiv Prepr. Serv. Biol.* (2020). doi:10.1101/2020.05.23.112235

27. Chemical Computing Group (CCG) Inc. Molecular Operating Environment (MOE). (2018).

28. Harvey, M. J., Giupponi, G. & Fabritiis, G. De. ACEMD: accelerating biomolecular dynamics in the microsecond time scale. *J. Chem. Theory Comput.* **5**, 1632–1639 (2009).

29. Tan, D., Piana, S., Dirks, R. M. & Shaw, D. E. RNA force field with accuracy comparable to state-of-the-art protein force fields. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E1346–E1355 (2018).

30. Wang, J., Wang, W., Kollman, P. A. & Case, D. A. Automatic atom type and bond type perception in molecular mechanical calculations. *J. Mol. Graph. Model.* **25**, 247–260 (2006).

31. Sprenger, K. G., Jaeger, V. W. & Pfaendtner, J. The general AMBER force field (GAFF) can accurately predict thermodynamic and transport properties of many ionic liquids. *J Phys Chem B* **119**, 5882–5895 (2015).

32. Berman, H. M. *et al.* The protein data bank. *Nucleic Acids Res.* **28**, 235–242 (2000).

33. Labute, P. Protonate3D: assignment of ionization states and hydrogen coordinates to macromolecular structures. *Proteins* **75**, 187–205 (2009).

34. Loncharich, R. J., Brooks, B. R. & Pastor, R. W. Langevin dynamics of peptides: The frictional dependence of isomerization rates ofN-acetylalanyl-N?-methylamide. *Biopolymers* **32**, 523–535 (1992).

35. Essmann, U. *et al.* A smooth particle mesh Ewald method. *J. Chem. Phys.* **103**, 8577–8593 (1995).

36. Bakan, A., Meireles, L. M. & Bahar, I. ProDy: protein dynamics inferred from theory and experiments. *Bioinformatics* **27**, 1575–1577 (2011).

37. Phillips, J. C. *et al.* Scalable molecular dynamics with NAMD. *Journal of Computational*

Chemistry **26**, 1781–1802 (2005).

38. Humphrey, W., Dalke, A. & Schulten, K. VMD - Visual Molecular Dynamics. *J. Mol. Graph.* **14**, 33–38 (1996).

Supervised Molecular Dynamics (SuMD) Insights into the mechanism of action of SARS-CoV-2 main protease inhibitor PF-07321332

Matteo Pavan, Giovanni Bolcato, Davide Bassani, Mattia Sturlese, Stefano Moro.

Pavan, M., Bolcato, G., Bassani, D., Sturlese, M. & Moro, S. Supervised Molecular Dynamics (SuMD) Insights into the mechanism of action of SARS-CoV-2 main protease inhibitor PF-07321332. *J Enzyme Inhib Med Chem* **36**, 1646–1650 (2021).

Abstract

The chemical structure of PF-07321332, the first orally available Covid-19 clinical candidate, has recently been revealed by Pfizer. No information has been provided about the interaction pattern between PF-07321332 and its biomolecular counterpart, the SARS-CoV-2 main protease (M^{pro}). In the present work, we exploited Supervised Molecular Dynamics (SuMD) simulations to elucidate the key features that characterize the interaction between this drug candidate and the protease, emphasizing similarities and differences with other structurally related inhibitors such as Boceprevir and PF-07304814. The structural insights provided by SuMD will hopefully be able to inspire the rational discovery of other potent and selective protease inhibitors.

1. Introduction

The Covid-19 pandemic, caused by a single-stranded RNA betacoronavirus known as SARS-CoV-2, has caused the death of more than 3 million people around the world since its outbreak in December 2019[1,2]. Despite the impressive cooperative effort promoted by the international community and by medicinal chemists around the world[3,4], to date, there is only one drug approved by the Food and Drug Administration (FDA) for the treatment of Covid-19 patients.

Remdesivir, a polymerase inhibitor initially conceived to target Ebola Virus, proved to be efficient in shortening the recovery time in adult patients hospitalized with Covid-19[5,6] and has therefore been granted Emergency Use Authorization (EUA). Unfortunately, due to its pharmacokinetic profile, this drug has to be administered intravenously in a hospital setting, thereby limiting its use for Covid-19 treatment on a massive scale. The first

attempts to face this lack of pharmacological tools to contrast the Covid-19 pandemic involved the repurposing of antiviral drugs designed for the treatment of other virus-related illnesses against Covid-19: this approach, despite being very appealing from a timescale perspective[7], did not bring any significant results, with several clinical trials showing little to no efficacy of those active principles against SARS-CoV-2[8].

Meanwhile, the early release to the scientific community of the crystallographic structure of the SARS-CoV-2 main protease (M^{pro}) (PDB ID: 6LU7), caused a shift in the attention of researchers around the world towards the Structure-Based approach to the rational design of new potential protease inhibitors[9,10]. Among all the different chemical entities developed to target the main protease, PF-07321332 is, to date, the first and only orally available COVID-19 antiviral clinical candidate.

Designed amid the pandemic, the structure of PF-07321332 was unveiled by Pfizer on April 6th at the American Chemical Society Spring 2021 meeting[11]. This compound, which has recently entered clinical phase I, was developed to target SARS-CoV-2 main protease, thereby impairing the virus's ability to reproduce itself, and it is intended as a pharmacological tool to prevent the development of COVID-19 in people who have been exposed to the pathogen. Even though the compound structure has been revealed, no further information has been provided yet about the way PF-07321332 interacts with the main protease active site, except for the fact that it reacts reversibly with a cysteine residue located in the binding site[11].

In this perspective computational investigation, we exploited Supervised Molecular Dynamics (SuMD)[12], an emerging protocol allowing to decipher at an atomic level of detail the recognition process between two molecular entities, to sample and characterize a putative binding pathway for PF-07321332. As described in the original publication, SuMD simulations fully consider both the protein flexibility and the contribution of the solvent molecules, which are explicitly simulated, throughout the binding process. As shown by previous scientific works[13,14], this makes it possible to overcome the limitations of traditional techniques such as molecular docking when working on challenging targets such as M^{pro}, whose active site is relatively shallow, plastic and solvent exposed[15].

2. Methods

2.1 Software overview

For every general molecular modeling operation, such as protein and ligand structure preparation, MOE suite (Molecular Operating Environment, version 2019.01[16]) was used, exploiting an 8 CPU (Intel Xeon E5-1620 3.50 GHz) Linux Workstation. Molecular Dynamics simulations were carried out with ACEMD[17] (version 3.3.0), which is based upon OpenMM[18] (version 7.4.0), on a cluster composed of 20 NVIDIA GPUs.

2.2 Structure preparation

The crystallographic structure of the unliganded M^{pro} was retrieved from the Protein Data Bank (PDB ID: 7K3T). At first, the active functional dimer of the protease was restored applying the symmetric crystallographic transformation to each asymmetric unit. Residues with alternative conformation were assigned to the one with the highest occupancy. The Protonate3D tool was then used to add missing hydrogen atoms, evaluating the most probable protonation state for each titratable residue at pH 7.4. Finally, each non-protein residues (e.g.: water, co-solvents, etc.) were removed before successive steps. The ligand structure was prepared exploiting tautomers, fixpka, and molcharge tools from the QUACPAC OpenEye[19] software suite to assign the most probable tautomeric and protomeric state at pH 7 and ligand partial charges according to the MMFF94 force field. Three-dimensional coordinates were generated with Corina Classic[20].

2.3 Molecular Dynamics system setup

The simulated system contained both the protein and the ligand structure prepared as described in the previous section, with the ligand positioned at least 30 Å away from the nearest receptor atoms. For system parametrization, the combination of Amber ff14SB and General Amber Force Field (GAFF) was used to describe each component of the simulation box.

The system was explicitly solvated in a cubic TIP3P[21] water box with 15 Å padding and neutralized with the addition of Na+/Cl- ions until a 0.154 M concentration was reached. Prior to the simulation, 1000 steps of energy minimization with the conjugated-gradient

algorithm were performed. A two-step equilibration stage was carried out in the following way: the first step consisted of 0.1 ns of simulation in the canonical ensemble (NVT) with harmonic positional restraints applied both on the protease and ligand atoms using a 5 Kcal mol^{-1} $Å^{-2}$ force constant, the second step consisted of 0.5 ns of simulation in the isothermal-isobaric ensemble (NPT) with the same harmonic positional restraints applied only on protein alpha carbons and ligand atoms. For each simulation, an integration timestep of 2 fs was used. To constrain bonds involving hydrogen atoms the M-SHAKE algorithm was used. A 9.0 Å cutoff was applied for the calculation of Lennard-Jones interactions, while electrostatic interactions were computed exploiting the particle-mesh Ewald method (PME). The temperature was maintained at the constant value of 310K by the Langevin thermostat, with a friction coefficient of 0.1 ps^{-1}. During the second equilibration stage, the pressure was maintained constant at 1.0 atm utilizing a Monte Carlo barostat.

2.4 Supervised Molecular Dynamics (SuMD) simulation

SuMD code is written in Python 2.7 and exploits the ProDy[22] package to perform geometrical supervision upon the ligand-binding process. This supervision allows to reduce the timescale, hence shrinking the computational effort, that is required to sample the ligand-biomolecular target recognition process to the range of nanoseconds, instead of the usual hundreds of nanoseconds or microseconds that are required by unbiased molecular dynamics (MD) simulations. The entire SuMD derived trajectory is composed by short unbiased 600 ps MD simulation runs (NVT ensemble, T= 310 K) with the ACEMD3 software: at the end of each simulation (the so-called “SuMD-step”), the distance between the center of mass of the ligand and the binding site is computed at five different points, picked at regular time intervals, and fitted into a linear function evaluated by a tabu-like algorithm. Only those SuMD-steps whose computed slope is negative (indicating that the ligand is approaching the binding site) are retained. Every time a SuMD-step is rejected (positive slope), the simulation is restarted from the previous productive step by randomly assigning the atomic velocities. The supervision algorithm is switched off after the distance between the center of mass of the ligand and the binding site drops below 5 Å: from that point on the simulation continues as a classical MD simulation.

3. Results

In our computational study, we exploited Supervised Molecular Dynamics simulations to obtain a putative binding pathway between PF-07321332 and the SARS-CoV-2 Main Protease (M^{pro}) catalytic site. A total amount of 36 ns of SuMD simulation time proved sufficient to sample the entire recognition trajectory, from the starting unbound state to the final predicted protein-ligand complex.

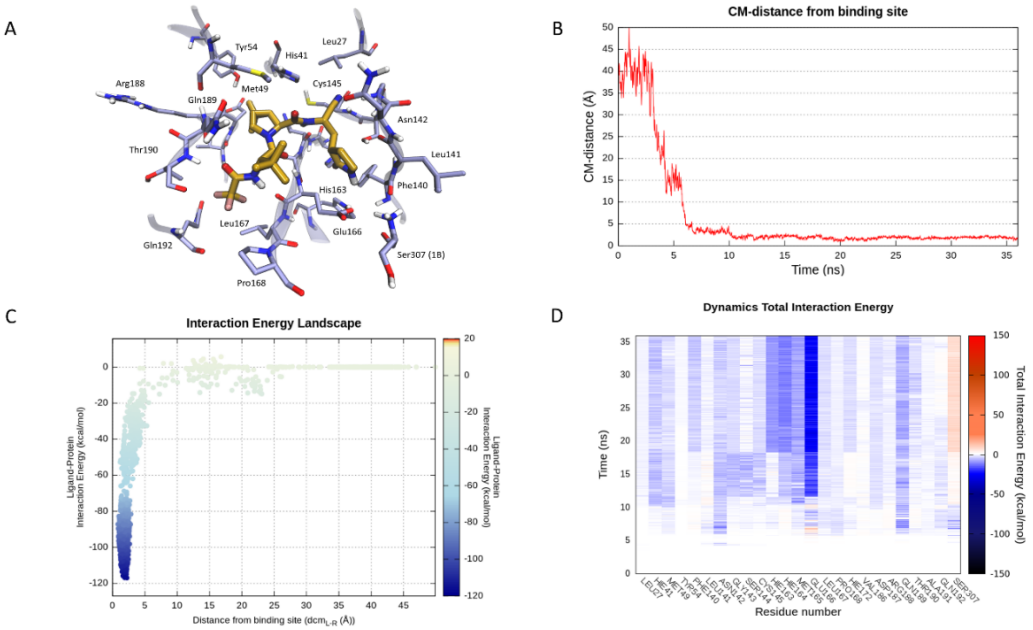


Figure 1. This panel encompasses the recognition pathway between PF-07321332 and the SARS-CoV-2 main protease predicted by SuMD. **(A)** PF-07321332 conformation within the binding site, sampled in the last SuMD trajectory frame (orange). Binding site residues within 4 Å of the ligand are depicted in ice-blue. **(B)** Profile of the distance between the center of mass of the ligand and the M^{pro} catalytic site during SuMD simulation. **(C)** Interaction Energy Landscape describing the protein-ligand binding pathway; values are arranged according to distances between the center of mass of the ligand the one of the M^{pro} catalytic site. **(D)** Dynamic total interaction energy (sum of electrostatic and van der Waals contribution) computed for the 25 most contacted residues throughout the SuMD trajectory.

As can be seen in Video1, PF-07321332 reaches M^{pro} active site after about 7 ns of simulation time, making its first contacts with Leu141, Asp 142, Gln189, and Glu166. Leu141 and Asp142 are part of the oxyanion loop (residues 138-145), which lines the binding pocket of Glutamine P1 and is assumed to stabilize the tetrahedral acyl transition state[15]. Glu166 is a key residue located in the middle of the binding site: mutagenesis studies carried out on SARS-CoV M^{pro} (which has 96% sequence identity with SARS-CoV-2

M^{pro} and is identical at the binding site level[13]) showed that this residue plays a key role in linking the dimer interface with the substrate-binding site[23]. Gln189 is located at the boundary of the S3 site and is assumed to be one of the key interactors with SARS-CoV-2 M^{pro} inhibitors, as well as Glu166[24]. Asn142 and Gln189, located on opposite sides at the boundary of the binding sites, seem to serve as electrostatic recruiters for the ligand, exploiting their polar and flexible sidechains to maneuver the entrance of the ligand into the core region of the binding site. Glu166 appears to instead serve as an electrostatic anchor that tightly hooks the middle portion of the ligand with the central region of the binding site, facilitating the formation of further interactions with residues such as His 164.

After the tri-fluoro-acetamide moiety of the compound establishes contact with the side chain of Gln189, the cyclopropyl-proline moiety occupies the central portion of the binding site, establishing a series of coordinated hydrogen bonds with the backbone of His164 and Glu166 and orientating the cyclopropyl group towards the hydrophobic S2 pocket, delimited by the side chains of His41, Met49, Tyr54, and Met165. Meanwhile, the pyrrolidone moiety is inserted in the S1 pocket, interacting with key residues of the oxyanion loop such as Asn142, Gly143, and Ser144, before undergoing a conformational rearrangement around the 18 ns simulation time mark which allows the carbonyl of the pyrrolidone to establish a hydrogen bond with His163. This interaction has been flagged as a conserved interaction across several deposited structures of non-covalent inhibitors[25]. Moreover, this interaction is conserved across all possible substrate peptide crystal structures, where the interacting group is the sidechain of the Glutamine P1 residue[26].

Subsequently, the pyrrolidone moiety rearrangement also allows the reactive nitrile group to face the catalytic Cys145, making it possible to reach the final covalent-bound state which cannot be described through molecular mechanics. Finally, in the final conformation, the tri-fluoro acetamide moiety is fully inserted in the S4 subpocket, establishing two additional hydrogen bonds with the backbone of Thr190 and Glu166.

As can be seen in Figure 1 (Supplementary Material), the ligand conformation in the final step of the SuMD simulation is superimposable to the bound state predicted by the PLANTS[27–29] docking algorithm ($RMSD_{SuMD\text{-}PLANTS}$: 0.92 Å), further corroborating the binding mode hypothesis portrayed by the SuMD protocol.

4. Discussion

Intriguingly, the binding mode proposed by the SuMD simulation for PF-07321332 is fairly superimposable to the ones of other two covalent protease inhibitor, Boceprevir (PDB ID: 6WNP) and PF-00835231 (PDB ID: 6XHM) which share common structural features with the oral candidate, validating the hypothesis that they could also share an overall similar interacting pattern (Figure 2).

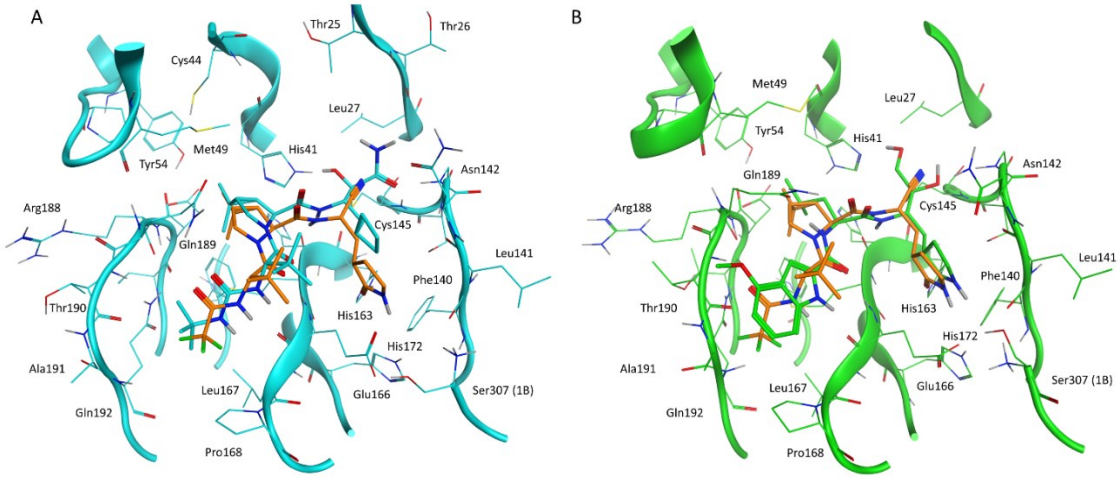


Figure 2. This panel illustrates the similarities between PF-07321332 conformation in the final SuMD trajectory frame and the crystallographic complexes of two structurally related covalent inhibitors of SARS-CoV-2 M^{pro}: Boceprevir and PF-00835231 (active metabolite of PF-07304814). **(A)** superposition between the binding mode predicted by SuMD for PF-07321332 (orange) and the crystallographic complex of Boceprevir within the catalytic site of SARS-CoV-2 M^{pro} (cyan, PDB ID: 6WNP). **(B)** superposition between the binding mode predicted by SuMD for PF-07321332 (orange) and the crystallographic complex of PF-00835231 within the catalytic site of SARS-CoV-2 M^{pro} (green, PDB ID: 6XHM)

Boceprevir is a protease inhibitor originally developed for the Hepatitis C Virus (HCV) NS3 protease[30]. It shares many common structural features with PF-07321332, such as the cyclopropyl proline residue at P2 and the alanine at the P3 position but has a different reactive group (α-ketoamide), a cyclobutyl alanine at P1, and a tert-butyl carbamate

capping moiety at P4. From a binding mode point of view, the most prominent difference between the newly developed inhibitor and Boceprevir regards the hydrogen bond with His163 (absent in Boceprevir complex with the protease) which, as previously mentioned, is a crucial interaction also for natural peptidic substrates.

PF-07304814 is a Phase I clinical candidate originally developed by Pfizer in 2002-2003 against SARS-CoV and repurposed for SARS-CoV-2 due to the aforementioned similarities between the two viruses' proteases[31]. The compound contains a hydrolyzable phosphate group which enhances its solubility and is cleaved by alkaline phosphatases in tissue releasing the active compound PF-00835231. The main limiting factor for this candidate is that, unlike its successor PF-07321332, it has to be administered intravenously, making it less appealing for massive distribution and relegating its usage to hospital settings. From a structural point of view, this latter compound is less similar to PF-07321332 compared to Boceprevir, but still retains the key features concerning its binding mode with the M^{pro} active site. The only conserved structural feature between the two inhibitors developed by Pfizer is the pyrrolidone group at the P1 position, which establishes a hydrogen bond with His163. The reactive group, in this case, is an aldehyde, the same as for Boceprevir. The hydrophobic residue at P2, in this case, is a leucine, which is the most recurrent amino acid that can be found at the P2 position in natural substrate peptides (included the N-term of M^{pro} itself)[26], while the P3 terminal residue is a 4-methoxyl indole group, which interacts through a hydrogen bond with the backbone of Glu166. Additional interaction occurs at the P1' subsite, where the two hydroxyl groups (one of which is formed upon reaction between the aldehyde group and Cys145 sidechain) form hydrogen bonds with Cys145 backbone and His41 sidechain.

Overall, PF-07321332 appears to have combined the strong points of both Boceprevir and PF-07304814 in a single molecular entity, showing that it is possible to repurpose the knowledge acquired in previous drug development campaigns on different virus proteases to rationally design SARS-CoV-2 M^{pro} inhibitors suitable for advancement to clinical phases, hence addressing the need for a quick response against a widespread disease like Covid-19. Moreover, the combination of innovative computational strategies such as SuMD with experimental data coming from X-Ray Crystallography could provide useful

structural insights to stir the rational development of antiviral drugs in a more rational and less time-consuming way.

5. Conclusions

In this computational study, we employed Supervised Molecular Dynamics (SuMD) to investigate the recognition process between PF-07321332, the first orally available Covid-19 antiviral candidate to reach clinical phase I, and its biological target, SARS-CoV-2 main protease (M^{pro}).

About 36 ns of SuMD simulations proved sufficient to sample a putative binding process, allowing to simulate the whole approaching path from the unbound state to the final protein-ligand complex. SuMD simulations suggest a possible role in the first stages of the recruitment of the ligand for residues such as Leu141, Asp 142, Gln189, and Glu166, which have already been acknowledged as crucial residues for the binding of both natural and synthetic substrates.

Finally, the binding mode predicted by SuMD for PF-07321332 is quite similar for other structurally related protease inhibitors, namely Boceprevir and PF-07304814, which could also share a similar binding pathway.

References

1. Guarner, J. Three Emerging Coronaviruses in Two Decades: The Story of SARS, MERS, and Now COVID-19. *American Journal of Clinical Pathology* vol. 153 420–421 Preprint at https://doi.org/10.1093/ajcp/aqaa029 (2020).

2. COVID Live Update: 163,750,604 Cases and 3,394,311 Deaths from the Coronavirus - Worldometer. https://www.worldometers.info/coronavirus/.

3. Zhang, L. & Liu, Y. Potential interventions for novel coronavirus in China: A systematic review. *Journal of Medical Virology* vol. 92 479–490 Preprint at https://doi.org/10.1002/jmv.25707 (2020).

4. Heymann, D. L. & Shindo, N. COVID-19: what is next for public health? *The Lancet* vol. 395 542–545 Preprint at https://doi.org/10.1016/S0140-6736(20)30374-3 (2020).

5. Kokic, G. *et al.* Mechanism of SARS-CoV-2 polymerase stalling by remdesivir. *Nat Commun* **12**, 1–7 (2021).

6. Beigel, J. H. *et al.* Remdesivir for the Treatment of Covid-19 — Final Report. *New England Journal of Medicine* **383**, 1813–1826 (2020).

7. Mani, D., Wadhwani, A. & Krishnamurthy, P. T. Drug Repurposing in Antiviral Research: A Current Scenario. *Journal of Young Pharmacists* **11**, 117–121 (2019).

8. Viveiros Rosa, S. G. & Santos, W. C. Clinical trials on drug repositioning for COVID-19 treatment. *Revista Panamericana de Salud Publica/Pan American Journal of Public Health* **44**, e40 (2020).

9. Jin, Z. *et al.* Structure of Mpro from SARS-CoV-2 and discovery of its inhibitors. *Nature* **582**, 289–293 (2020).

10. Capasso, C., Nocentini, A. & Supuran, C. T. Protease inhibitors targeting the main protease and papain-like protease of coronaviruses. *Expert Opinion on Therapeutic Patents* vol. 31 309–324 Preprint at https://doi.org/10.1080/13543776.2021.1857726 (2021).

11. Pfizer unveils its oral SARS-CoV-2 inhibitor. https://cen.acs.org/acs-news/acs-meeting-news/Pfizer-unveils-oral-SARS-CoV/99/i13.

12. Sabbadin, D. & Moro, S. Supervised molecular dynamics (SuMD) as a helpful tool to depict GPCR-ligand recognition pathway in a nanosecond time scale. *J Chem Inf Model* **54**, 372–376 (2014).

13. Bolcato, G., Bissaro, M., Pavan, M., Sturlese, M. & Moro, S. Targeting the coronavirus SARS-CoV-2: computational insights into the mechanism of action of the protease inhibitors lopinavir, ritonavir and nelfinavir. *Sci Rep* **10**, 20927 (2020).

14. Bissaro, M. *et al.* Inspecting the mechanism of fragment hit binding on SARS-CoV-2 Mpro by using supervised molecular dynamics (SuMD) simulations. *ChemMedChem* (2021) doi:10.1002/cmdc.202100156.

15. Fornasier, E. *et al.* A novel conformational state for SARS-CoV-2 main protease. *bioRxiv* 2021.03.04.433882 (2021) doi:10.1101/2021.03.04.433882.

16. Molecular Operating Environment (MOE), 2019.01; Chemical Computing Group ULC, 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, 2021. https://www.chemcomp.com/Research-Citing_MOE.htm.

17. Harvey, M. J., Giupponi, G. & de Fabritiis, G. ACEMD: Accelerating biomolecular dynamics in the microsecond time scale. *J Chem Theory Comput* **5**, 1632–1639 (2009).

18. Eastman, P. *et al.* OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Comput Biol* **13**, e1005659 (2017).

19. QUACPAC 2.0.1.2: OpenEye Scientific Software, Santa Fe, NM. https://www.eyesopen.com/.

20. Sadowski, J., Gasteiger, J. & Klebe, G. Comparison of Automatic Three-Dimensional Model Builders Using 639 X-ray Structures. *Journal of Chemical Information and Computer Sciences* **34**, 1000–1008 (1994).

21. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J Chem Phys* **79**, 926–935 (1983).

22. Bakan, A., Meireles, L. M. & Bahar, I. ProDy: Protein dynamics inferred from theory and experiments. *Bioinformatics* **27**, 1575–1577 (2011).

23. Cheng, S. C., Chang, G. G. & Chou, C. Y. Mutation of glu-166 blocks the substrate-induced dimerization of SARS coronavirus main protease. *Biophys J* **98**, 1327–1336 (2010).

24. Goyal, B. & Goyal, D. Targeting the Dimerization of the Main Protease of Coronaviruses: A Potential Broad-Spectrum Therapeutic Strategy. *ACS Combinatorial Science* vol. 22 297–305 Preprint at https://doi.org/10.1021/acscombsci.0c00058 (2020).

25. Weng, Y. L. *et al.* Molecular dynamics and in silico mutagenesis on the reversible inhibitor-bound SARS-CoV-2 main protease complexes reveal the role of lateral pocket in enhancing the ligand affinity. *Scientific Reports* **11**, 7429 (2021).

26. Rut, W. *et al.* Substrate specificity profiling of SARS-CoV-2 main protease enables design of activity-based probes for patient-sample imaging. *bioRxiv* 2020.03.07.981928 (2020) doi:10.1101/2020.03.07.981928.

27. Korb, O., Stützle, T. & Exner, T. E. PLANTS: Application of Ant Colony Optimization to Structure-Based Drug Design. 247–258 (2006) doi:10.1007/11839088_22.

28. Korb, O., Stützle, T. & Exner, T. E. An ant colony optimization approach to flexible protein–ligand docking. *Swarm Intelligence* **1**, 115–134 (2007).

29. Korb, O., Stützle, T. & Exner, T. E. Empirical scoring functions for advanced Protein-Ligand docking with PLANTS. *J Chem Inf Model* **49**, 84–96 (2009).

30. Njoroge, F. G., Chen, K. X., Shih, N.-Y. & Piwinski, J. J. Challenges in Modern Drug Discovery: A Case Study of Boceprevir, an HCV Protease Inhibitor for the Treatment of Hepatitis C Virus Infection. *Accounts of Chemical Research* **41**, 50–59 (2008).

31. Boras, B. *et al.* Discovery of a Novel Inhibitor of Coronavirus 3CL Protease as a Clinical Candidate for the Potential Treatment of COVID-19. *bioRxiv : the preprint server for biology* 2020.09.12.293498 (2020) doi:10.1101/2020.09.12.293498.

Computational strategies to identify new drug candidates against neuroinflammation.

Matteo Pavan, Davide Bassani, Giovanni Bolcato, Maicol Bissaro, Mattia Sturlese and Stefano Moro

Pavan, M. *et al.* Computational Strategies to Identify New Drug Candidates against Neuroinflammation. *Curr Med Chem* **29**, 4756–4775 (2022).

Abstract

The even more increasing application of computational approaches in these last decades has deeply modified the process of discovery and commercialization of new therapeutic entities. This is especially true in the field of neuroinflammation, in which both the peculiar anatomical localization and the presence of the blood-brain barrier make it mandatory to finely tune the candidates' physicochemical properties from the early stages of the discovery pipeline. The aim of this review is therefore to provide a general overview to the readers about the topic of neuroinflammation, together with the most common computational strategies that can be exploited to discover and design small molecules controlling neuroinflammation, especially those based on the knowledge of the three-dimensional structure of the biological targets of therapeutic interest. The techniques used to describe the molecular recognition mechanisms, such as molecular docking and molecular dynamics, will therefore be eviscerated, highlighting their advantages and their limitations. Finally, we report several case studies in which computational methods have been applied in drug discovery on neuroinflammation, focusing on the last decade's research.

Introduction

Inflammation is a vital host defense response to stimuli that undermine the homeostasis and integrity of tissues, such as traumatic injury, infection by pathogens, autoimmune responses, or hypoxic conditions. From a mechanistic point of view, it can be described as a multi-factor process that involves the invasion of circulating immune cells such as lymphocytes and monocytes and the induction of inflammatory mediators like cytokines and prostanoids[1].

Usually, inflammation consists of a cascade of events known as acute phase response aimed at resolving the initiating stimulus and restoring tissue's integrity and functionality. However, if the condition persists for a prolonged period and exceeds the physiological response it becomes detrimental to the tissue's health and function[2].

Uncontrolled inflammatory processes are associated with several neurodegenerative diseases, including both proteinopathies such as Alzheimer's disease (AD), Parkinson's disease (PD), amyotrophic lateral sclerosis (ALS), or Huntington's disease (HD), and lysosomal storage diseases (LSDs) such as Niemann-Pick type C (NPC), Gaucher Disease (GD), Mucopolysaccharidoses (MPS) and Neuronal Ceroid Lipofuscinosis (NCL), resulting in a condition defined as neuroinflammation[3,4].

Although it is still debated whether this condition has a role in the early stages of the aforementioned pathologies, contributing to their etiology, several pieces of evidence point out that sustained inflammatory states in the Central Nervous System (CNS) decisively contribute to the progression of such diseases[5,6]. Moreover, some studies even suggest that inflammation could play a protective role in the first stages of the disease[7,8], further underlying the complexity of the topic and the intricacy of the mechanisms that regulate these processes.

From an immune system point of view, the brain is compartmentalized and separated from the peripheral blood circulation by the blood-brain-barrier (BBB), therefore indicating that a critical role in the inflammatory process is portrayed by locally resident cells, namely astrocytes and microglia, rather than by leucocytes infiltrating from the periphery[9,10]. This is particularly true in the initiating phases of inflammation, while in later stages the disruption of the BBB provoked by the inflammatory cascade and by proteases such Matrix Metalloproteinases (MMPs) causes the involvement of non-resident immune cells such as monocytes, neutrophils, and CD4+/CD8+ T cells[11–14].

In healthy brain tissue, astrocytes serve as key components of the BBB. Moreover, they are involved in ionic homeostasis, removal of excess neurotransmitters, and secretion of neurotrophic factors that are vital for neuron survival[15]. Astrocytes can exist in a continuous spectrum of different phenotypes, ranging from neuroprotective to pro-

inflammatory[16]. Although their pro-inflammatory contribution is limited compared to that of microglia, under immunologic challenges or injuries, astrocytes assume an activated phenotype that leads to an increase in the production of glial fibrillary acidic proteins that eventually hinders axonal regeneration[17,18]. Furthermore, reactive phenotype astrocytes can induce the production of factors such as Interleukin-1β (IL-1β), Tumor Necrosis Factor (TNF), and nitric oxide (NO) that enhance inflammation and neuronal death[19].

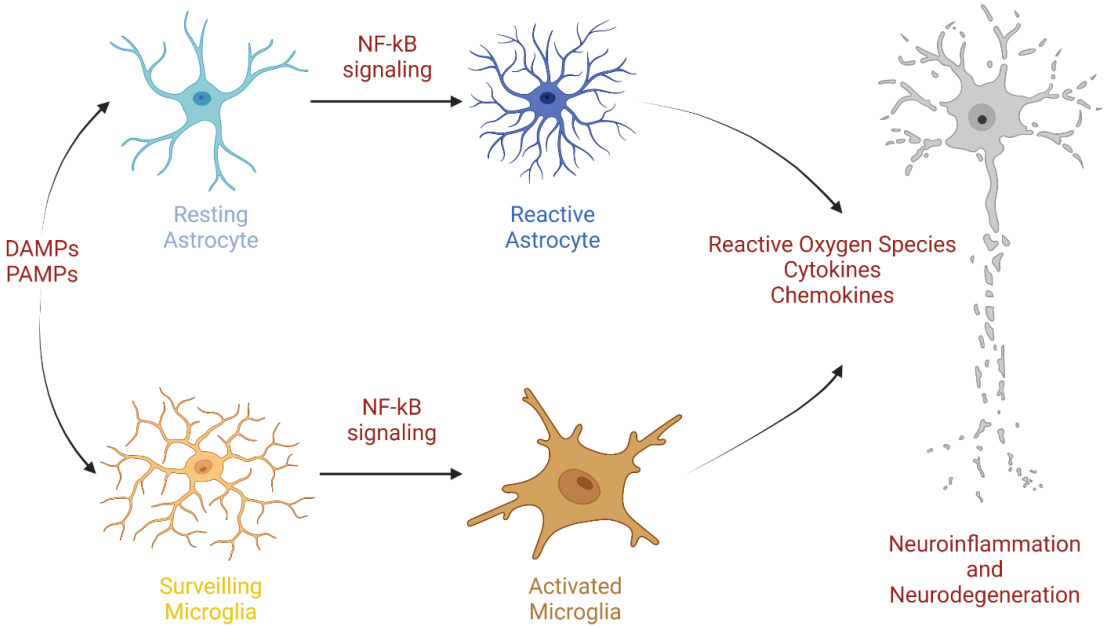


Figure 1. General scheme of the common mechanisms involved in neuroinflammation. The figure was created with 'BioRender.com'.

Microglia are the predominant immune cell species located in the brain[20]. Under physiological conditions, these macrophage-like cells exist in a deactivated ramified phenotype, and they are engaged in a process of environmental surveillance to maintain tissue homeostasis, producing neurotrophic and anti-inflammatory factors[21,22]. Under the influence of external stimuli such as pathogen invasion or tissue damage, they can switch to an activated amoeboid phenotype that promotes an inflammatory response to address the situation and promote tissue reparation[23]. If a persistent stimulus from either an environmental or an endogenous factor is perceived by the immune system as a threat,

an uncontrolled inflammatory state may cause the production of several neurotoxic factors such as superoxide, NO, and TNF that contribute to the progression of underlying disease states[24]. A general depiction of the common mechanisms underlying neuroinflammation is reported in Figure 1.

This inflammatory response is usually initiated by pattern recognition receptors such as Toll-Like Receptors (TLRs)[25], but they can also be started by purinergic[26] and scavenger receptors[27]. Particularly, TLRs such as TLR4 are overexpressed on the surface of CNS resident immune cells such as microglia and are responsible for the recognition of both pathogen-associated molecular patterns (PAMPs) and damage/danger-associated wide-array of patterns (DAMPs) coming from the host itself[25]. The inflammatory signal is then passed on to several transduction systems, such as Myeloid differentiation primary response 88 (MyD88), TIR-domain-containing adapter-inducing interferon-β (TRIF), phosphoinositide 3-kinases (PI3Ks), IkB kinases (IKK), and mitogen-activated protein kinases (MAPKs), that cooperate in a combinatorial manner to regulate hundreds of different genes based upon the activated target cells employing transcription factors such as the ubiquitous nuclear factor kappa-light-chain-enhancer of activated B cells (NF-kB)[28–33]. Among these signal transduction pathways, particular importance is portrayed by the Wnt signaling, whose pro-inflammatory activity is finely regulated by several mechanisms such as the formation of the "β-catenin destruction complex" which involves neurodegenerative disease-related kinases such as casein kinase 1 delta (CK1□) and glycogen synthase kinase 3 (GSK-3□)[34–36]. Finally, gene expression then leads to the signal amplification through prostaglandins, cytokines such as TNF or interleukin 1 beta (IL-1β), and chemokines such as monocyte chemoattractant protein-1 (MCP-1), C-C motif chemokine ligand 2 (CCL2), and C-X-C Motif Chemokine Ligand 5 (CXCL5) that allow for the recruitment of other immune cells[37,38]. Furthermore, the generation of reactive oxygen species (ROS) mainly through the nicotinamide adenine dinucleotide phosphate oxidase (NADPH oxidase) system and of nitric oxide through the inducible nitric oxide synthase (iNOS) further contribute to the defense response in an unspecific manner[39,40].

By contrast, several negative feedback mechanisms exist to attenuate or arrest this inflammatory cascade. These involve proteins with different functions, such as inhibition

of signal transduction (SOCS proteins), induction of transcriptional repressors (Nurr1), and production of soluble anti-inflammatory factors (IL-10, TGF-β)[41–43]. Furthermore, resolution of inflammation is also modulated by a family of specialized pro-resolving lipid mediators (SPMs) formed by lipoxins, resolvins, protectins, and maresins, either by inhibition of pro-inflammatory pathways or by activation of protective ones through interaction with a series of G-protein coupled receptors (GPCRs)[44,45].

Although a general convergence in the mechanism of the inflammatory processes that lead to neurotoxicity can be found, each pathology has its own set of markers that characterize the specific role of neuroinflammation in the insurgence and progression of the disease[46].

Alzheimer's disease was the first neurodegenerative disease to be linked to an inflammatory response[47]. Characterized by symptoms such as memory loss and cognitive impairment, from a pathological point of view this illness is associated with extracellular amyloid plaques formed by cleavage products of the amyloid precursor protein (APP) such as the beta-amyloid peptide (Aβ peptide) and with intracellular neurofibrillary tangles (NFTs) that are composed of hyperphosphorylated forms of the microtubule-binding protein tau[48]. The hallmarks of the inflammatory response in AD include a change in the microglia phenotype, which switches from a resting to an activated pro-inflammatory one, astrogliosis, which consists of an increase of astrocytes number, dimensions, and motility, other than high levels of cytokines and chemokines found in senile plaques[49].

Particularly, aggregates of Aβ peptide can activate microglia through TLRs such as TLR4 and TLR2, scavenger receptors CD36, CD47, and SCARA1, receptor for advanced glycoxidation end-products (RAGE) and NOD-like receptors (NLRs) like NALP3 or NALP1, leading to the phagocytosis of Aβ plaques, degradation of Aβ fibrils through the endolysosomal pathway and the subsequent activation of an inflammatory cascade that promotes neuronal death[50]. Not only pro-inflammatory cytokines can induce neuronal apoptosis directly, but they can also activate an astrocyte response that further enhances microglia-driven inflammation[17]. On the contrary, a protective role against this inflammatory cascade is played by TREM2[51] (triggering receptor expressed on myeloid cells 2), a transmembrane glycoprotein exclusively expressed on immune cells that

represses microglia-driven cytokine production and secretion[52] and regulate the phagocytic pathways involved in the removal of neuronal debris[53]. Several variants of the gene encoding for this receptor were associated with an increased risk of developing AD and other neurodegenerative diseases, such as PD and ALS, indicating that targeting TREM2 could be vital in the fight against these diseases[54,55].

Parkinson's disease is the second most common neurodegenerative disease after AD[56]. Differently from AD, it is mainly identified by motor symptoms such as bradykinesia, tremor, and rigidity, while from a neuropathological point of view it is characterized by intracellular inclusions of a misfolded protein called α-synuclein defined as Lewy bodies accompanied by the loss of dopaminergic neurons in the substantia nigra[57]. Regarding inflammation, microglial activation can be observed in PD, accompanied by astrogliosis and lymphocyte infiltration[58,59]. The presence of α-synuclein aggregates is sensed and internalized by microglia, causing the activation of NADPH oxidase and iNOS which leads to the release of ROS and NO that indeed causes the switch of microglia to the activated phenotype[60]. Pro-inflammatory microglia then release factors such as TNF, NO, and IL-1β that modulate and drive the neuroinflammatory process[61,62]. Moreover, high levels of neuromelanin derived from an increased catecholamine metabolism further contribute to neuron oxidative stress and microglial activation[63]. On the contrary, C-X3-C Motif Chemokine Receptor 1 (CX3CR1), CD200R, and Nurr1 can mediate anti-inflammatory responses that mitigate the negative effect of microglia activation in PD[64–66].

Amyotrophic Lateral Sclerosis is a neurodegenerative disease identified for the first time more than 150 years ago[67]. Clinically, it is characterized by a progressive motor neurons degeneration that eventually leads to the patient's death caused by respiratory muscles' paralysis[68]. The most prominent pathological marker of ALS is a strong inflammatory reaction caused by the immunoreactive inclusions of ubiquitin in the cytoplasm of degenerating neurons, which are mainly composed of TAR DBA-binding protein 43 (TDP-43)[69,70]. This inflammation process involves both astrocytes and microglia, whose activation causes the release of a high quantity of cytotoxic and pro-inflammatory molecules such as ROS, prostanoids, ILs, and TNF[71,72]. A prominent role in familiar forms of ALS is portrayed by mutant forms of superoxide dismutase 1 (SOD1), which cause an

incremented NADPH oxidase-dependent oxidative stress that boosts microglia activation and subsequent inflammatory response[73]. Neuron cell death, in this case, is induced by both NO and Fas ligand (FasL), while a protective role against this process is carried out by infiltrating Th2 cells by inducing microglia to produce neurotrophic factors[74–76].

Huntington's disease has been known for centuries, but only recently its causative agent has been identified[77]. HD is caused by a CAG triplet expansion in the huntingtin (HTT) gene[78]. This expansion leads to a conformational alteration of HTT protein which assumes a toxic phenotype, provoking neuronal damage and subsequent death either through the formation of intranuclear inclusion bodies, formation of toxic fragments, or impairment of metabolic pathways[79–82]. The main clinical feature of HD is the extensive striatal neuronal cell death, but also white matter is involved at various stages in the degeneration process[83]. This disease manifests itself with a set of clearly defined symptoms, such as chorea, bradykinesia, lack of coordination, and motor impairment[84]. An inflammatory response carried out by activated microglia is correlated with disease severity and progression: the up-regulation of pro-inflammatory proteins such as complement proteins and clusterin, and an increased level of cytokines are the main markers of inflammation in HD[85,86]. 8-OHdG, an oxidative stress marker circulating in the blood, is also up-regulated in patients with HD[87]. Finally, a microglial enzyme involved in excitotoxicity and ROS generation known as kynurenine 3-monooxygenase (KMO) has been found implicated in the regulation of HTT induced neurotoxicity[88].

Mucopolysaccharidoses are the most representative group of lysosomal storage diseases, accounting for 30% of all LSDs[89]. It is a heterogeneous group of pathologies caused by a deficiency in the lysosomal clearance of glycosaminoglycan (GAG)[90]. As for other LSDs, the alteration of lysosome function and integrity leads to the partial release of its content within the cellular lumen, with cathepsins and hydrolases acting as DAMPs and subsequently triggering the inflammatory cascade that leads to neuronal damage[91].

Although clinical manifestations of MPS can vary heavily depending on the illness type, peripheral organ dysfunction is generally a common trait across different MPS, with neurodegeneration and related cognitive and behavioral deficits being another quite

common clinical marker of this set of pathologies, particularly in the most aggressive forms of MPS type I (Hurler syndrome) and MPS type III (Sanfilippo syndrome)[90,92–94].

From an immune system point of view, MPS are associated with microglia-induced inflammatory response, characterized by high levels of chemokines such as CCL3 and CCL4 in the brain[95,96]. Moreover, treatments with anti-inflammatory/immunosuppressive drugs such as aspirin or prednisolone were able to lower the levels of both cytokines and chemokines associated with the inflammatory response[95,97].

Neuronal Ceroid Lipofuscinosis are a group of LSDs caused by mutation of one of 14 ceroid lipofuscinosis genes[98]. These mutations lead to alteration in the composition and structure of lysosomes[99,100]. The most recurrent symptoms of these pathologies are visual failure, neurocognitive and motor decline, seizures, and premature death[101]. The most deadly forms of these diseases are the ones associated with a more prominent neuroinflammatory response, characterized by increased microglial and astrocyte activation leading to high levels of pro-inflammatory mediators such as TNF and various interleukins[102,103]. Some studies suggested that anti-inflammatory treatment may be beneficial in hampering inflammation-caused neurodegeneration[104].

Gaucher Disease, the most common lysosomal storage disease, arises from a mutation in the lysosomal enzyme glucosylceramidase, which leads to intracellular accumulation of glucosylceramide[89,105,106]. Out of the three types of GD, only Type 2 and 3 are associated with neurological symptoms, which in the most severe cases can even lead to the premature death of the patient[92,107,108]. Neurodegeneration associated with inflammatory response is, in this case, triggered by an alteration of calcium homeostasis, which causes increased sensitivity of neurons to neurotoxic agents, thus inducing the release of DAMPs that induces the inflammatory response[109].

Niemann-Pick type C disease is caused by a mutation in either NPC1 or NPC2 gene[110]. These two genes codify for proteins involved in the regulation of intracellular trafficking, other than the processing of cholesterol and other lipids[111,112]. The alteration of their function leads to a lysosomal accumulation of cholesterol and sphingolipids, with lysosomal perturbation eventually resulting in the activation of the subsequent

inflammatory cascade, most probably due to the release in the cytoplasm of lysosome content (e.g. hydrolases and cathepsins) that acts as DAMPs[91,113]. NPC shows an age-dependent evolution of the main symptomatologic manifestation, with peripheral symptoms such as hepatosplenomegaly being predominant in early stages, while neurological manifestation such as cerebellar ataxia, dementia, cognitive decline, and seizures are more prominent at later stages of the disease, with the most severe forms leading to the patient's early death[114,115]. The involvement of a neuroinflammatory process in the progression of the pathology is highlighted by the high levels of TNF and IL-1β in brain regions associated with major neuron loss at later stages of the disease[116,117]. Furthermore, the disease progression can be altered and slowed down by exploiting anti-inflammatory compounds such as non-steroid anti-inflammatory drugs (NSAIDs)[118,119].

Considering the huge impact that these pathologies have on patient's expectancy and quality of life[120], the social cost for the treatment of such diseases[121] and the current lack of cure for this deadly neurodegenerative illnesses[122], the possibility to rationally develop pharmacological tools that modulate inflammatory response in neurodegenerative disease is highly appealing.

Computational methodologies in neurodegeneration

It is nowadays widely accepted that Computational-Aided Drug Discovery (CADD) techniques have given a huge impact on both the time and money required for bringing new therapeutical entities to the market[123].

The computational approach to drug design can be divided into two main classes: structure-based drug design (SBDD) and ligand-based drug design (LBDD)[124].

The first family of techniques strongly relies on the availability of structural data about the target investigated for the pathology considered[125]. These data mainly come from X-ray crystallography (XRC) nowadays (the public database Protein Data Bank contains 180207 structures at the present day) but can also derive from NMR spectroscopy. XRC allows to obtain, with a high resolution, objects with low limitations in terms of dimensions (roughly 150 kDa), it cannot resolve hydrogen atoms and gives just a "static" overview of the substrate. On the other hand, NMR spectroscopy can give access to data of the object

which is much more similar to reality (dynamic data in solution) but it cannot be reliable with structures above 50 kDa, other than requiring lots of runs to achieve an acceptable signal-to-noise ratio[126]. Other techniques have emerged in the latest decades, such as Cryo-Electron Microscopy, which is acquiring a lot of interest for its ability to resolve complicated (such as membrane proteins) and high-molecular-weight (such as entire mitoribosomes[127]) structures[128].

The ligand-based drug discovery comprises a variety of techniques that can be both used alone -if no structural data of the target is available[129]- and in consensus with SBDD protocols[130]. LBDD, as the name suggests, is based on the evaluation of known ligands of a certain target or simply on ligands that are active against a pathology on which no mechanism has been elucidated yet (so no target is known). The known actives are evaluated in several features, such as the 3D conformation, the electrostatic fields, the molecular descriptors, etc. After having obtained enough information, some algorithms can be developed to search molecules that are predicted to be similar to the known actives[131]. To pursue this task, in the latest years the experts in the field of machine learning and artificial intelligence have put some effort into developing reliable prediction models, both in industry and in academia[132].

In the case of neuroinflammation, considering the targets that we have summed up in the previous section, we can assess that SBDD can be implemented for drug discovery purposes, considering the considerable amount of structural data present for the targets of interest, as the PDB database witnesses[133,134,135,136].

Here we briefly report the main techniques which can be used for SBDD on neuroinflammation (which are also depicted in Figure 2), with a description of the tools and some cases in which they were adopted.

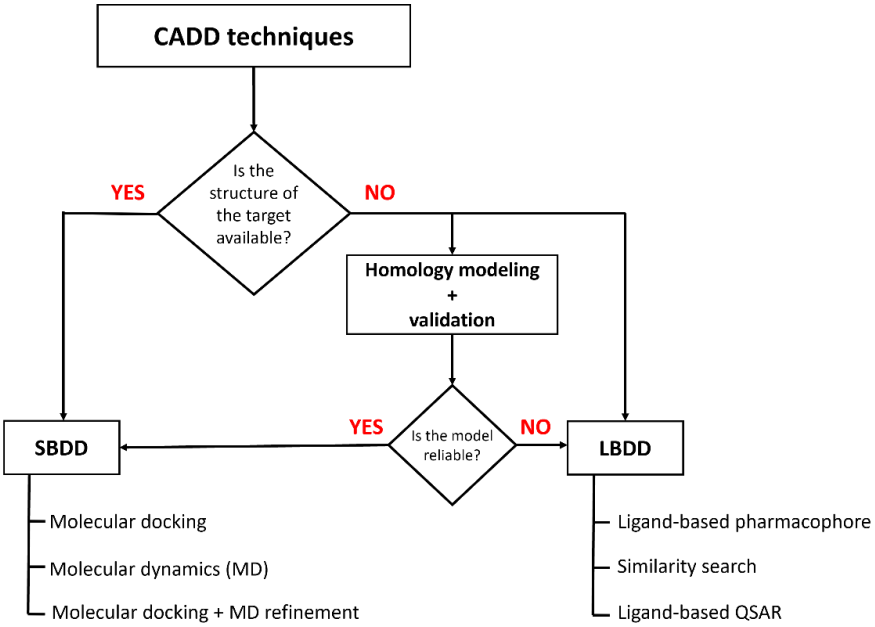


Figure 2. Representation of a general workflow scheme of the computational techniques that can be used to find new candidate molecules for neuroinflammation and related diseases.

Structure preparation

Just as an introduction, is always useful to remember that all the computational tools require both the ligands and the targets to be properly prepared for the calculation. The ligands must be in 3D shape, with a proper protonation state and, depending on the cases, with a partial charge assigned to each atom. The biological targets must also be suitable for calculation: the structure must be complete, the hydrogen atoms added (if the structure comes from XRC) and a protonation state assigned to all amino acids. It is also advisable for added hydrogens to be energetically minimized[137]. If some portions of the targets are missing, some programs may help in building them (through homology modeling, see chapter below) or, if it is the case, they can be capped (just for a small loop). This second option has reasonably been discarded if the capping site is very near to the active site, and mostly in docking calculations.

Homology modeling

If no structural data is available for a certain target, the SBDD approach can be used if a proper model is generated. Homology modeling is a technique used to predict the structure of a certain entity (mainly proteins and nucleic acids, in the case of medicinal chemistry) using available data[138]. The approach in these cases can be dual: in the case in

which some structures with reasonable sequence similarity (usually around 30% at least) are present, is possible to build the target (or even just a portion of it) using that as a template[139]. On the other hand, if no structure of the desired target is available at all, is possible to predict its 3D structure using different methodologies, such as the use of comparative models[140]. Many tools allow performing these tasks, such as MODELLER[141], Schrödinger Prime[142], MOE homology modeler[143], SWISS-MODEL[144], and others[145].

Another approach that has gained relevance recently is the prediction of protein structure based on artificial intelligence techniques. A famous example of this is represented by AlphaFold[146], a program based upon a deep learning algorithm (which has recently been updated to its second version, AlphaFold2[147]). The AlphaFold neural network has been properly trained to assign a three-dimensional shape to a sequence of amino acids based upon their similarities with existing crystallographic homologues[147]. This program has already proven to be very accurate in predicting the secondary structure of several targets[148,149,150]. Nevertheless, the application of models coming from AlphaFold to molecular docking experiments still needs some further optimization, mainly because of the higher difficulty in foretelling the conformation of a ligand-bound protein, which represents the main case of interest for drug discovery scientists[151].

It is a good practice to assess the stability and the reliability of the models produced, and this can be done through Molecular Dynamics (MD) refinement[152], evaluating various energetics and geometric parameters of the trajectory obtained.

There are some interesting situations regarding neuroinflammation in which no structural data is available for a target. When this circumstance takes place, both homology modeling and LBDD techniques can be evaluated. This is the case for the cited TSPO, on which some groups put efforts in generating valuable homology models, e.g., Bhargavi et al. used Modeller to create a virtual structure suitable for docking calculations[153]. Nguyen T. and Lai T. created and validated, through molecular dynamics simulations, a homology model for hTSPO based on the RsTSPO structure[154].

Molecular docking

Maybe the most used computational technique applied to drug discovery so far[155,156], molecular docking is a procedure consisting of generating several conformations of a molecule and evaluating the best among them for binding to a substrate (in our case, a protein or a nucleic acid). The docking programs can be roughly divided into two parts: a search algorithm and a scoring function. The first has the task to search the conformational space of the ligand, while the second has the objective to select, for each docking run, the best options among the conformations generated to fit with the target[157]. After several runs, the program returns the conformations (known as "poses") which are predicted to be the best for the scoring function applied. The docking programs can be differentiated mainly by how they search, optimize, and score the conformations of the molecules examined. A well-known family of programs use "genetic algorithms" to pursue this task (the most famous are CCDC GOLD[158] and AutoDock[159]), other use "systematic search algorithms" (such as Schrödinger Glide[160]) and other programs exploit more peculiar algorithms (such as the ACO used by PLANTS[161] and the molecular dynamics-based algorithm used by CDOCKER[162]). Several benchmarks have tried to highlight the best protocol to be used[163,164], but the most shared opinion is that each docking run must be optimized looking at the target considered[165]. A tool named "DockBench" was developed by Cuzzolin et. al to compare the efficiency of several program-scoring function pairs to reproduce the crystallographic pose of the ligand inside its substrate. This allows also us to select the best crystallographic structure to use for docking calculation (if more are present)[166].

To select among the poses obtained with molecular docking, there are several options. The first is to exclusively rely on the scoring function of the algorithm[167], another is to exclusively select poses in respect to certain parameters (such as electrostatic interaction, solvent exposure, etc.), but one of the most reliable methods remains the pharmacophore implementation[168]. A pharmacophore is an ensemble of features that are considered to be necessary for molecular recognition. If many XRC structures are present for a target, many tools offer the possibility to create a consensus "pharmacophore" (e.g., Schrödinger Phase[169] or MOE Pharmacophore editor), otherwise the features can be selected and

customed with the scientist knowledge. Another approach for docking poses scoring is the molecular dynamics refinement[170], of which the details are reported in the following section.

Molecular docking offers several advantages, such as flexibility, wide applicability, and speed[171]. With the technologies available nowadays, several billion molecules can be docked in a few days[172]. On the other hand, this technique has several limitations[173]. First of all, the scientist has to remind that the "classical" docking calculations are performed in a vacuum, which is does not take into account the presence of the water molecules solvating the biological target. Moreover, even if the solvent is considered, the water molecules should be classified for their ability to be displaced by the ligand[174], and special consideration should be given to the crystallographic waters around the ligand[175]. Another strong limitation is the number of false positives given by the algorithms. The docking programs have been developed making them very efficient in finding a good conformation for the binding, even if it is not consistent. To attenuate this feature, a particular protocol is known as "consensus docking" can be used. This approach consists of repeating the same docking calculation with different programs, which must implement different search-score algorithms[176,177]. The molecules that are selected from all the programs are more prone to be "real" positives.

In the field of neuroinflammation research, molecular docking has been applied several times. Cheng et al. implemented this technique for the development of novel MAPK inhibitors[178], while Rippin et al. used Glide docking for the design of novel GSK3 inhibitors[179].

Several efforts have been put into the discovery of CK1 inhibitors through molecular docking-based methods. Cescon et al. were successful in repurposing two molecules on CK1 through a docking-based protocol[180], while in a recent paper the same technique allowed Redenti et al. to design a dual GSK3 covalent/CK1 reversible inhibitor active in the very low micromolar range[181].

Tandon and Sinha used AutoDock and Flexidock to rationalize the binding mode of three groups of known MMP-9 inhibitors[182], while Razak et al. used the same software to

rationalize the activity of their compound of interest (2-(5-methoxy-2-methyl-1H-indol-3-yl)-N[(E)-(3-nitrophenyl) methylidene] acetohydrazide, called MMINA) against various targets of neuroinflammation, such as COX-2, STAT3, and TNF[183]. Docking-based virtual screening (VS) technique was chosen by Liu et al. to select (among over 90000 natural compounds) an inhibitor of STAT3 protein DNA-binding activity and dimerization[184], while Ray et al., through a large Glide-based VS, selected 15 molecules potentially able to stabilize SOD1 dimer, preventing its aggregation[185].

Looking at the TLRs, CDOCKER, and AutoDockVina were included by Durai et al. in the workflow for the discovery of two new TLR-2 antagonists active in the micromolar concentration range [186]. Mahita et al. applied molecular docking to evaluate the possibility of abrogating the TRAM-mediated signaling on Toll-Like Receptor 4[187]. Yilmazer et al. successfully implemented a virtual screening approach based on the Glide docking program to select the most promising candidates as novel chaperones for beta-glucocerebrosidase (GBA), one of the main pharmacological targets of Gaucher's disease[188].

El-Zohairy et al. recently implemented a GOLD-based virtual screening protocol comprising a pharmacophore filtration for the design of novel low-micromolar inhibitors of CCR5[189], and a similar workflow (using MolDock instead of GOLD) was applied by Ahmad et al. for the selection of potential inhibitors of caspase 8 for treating neurodegeneration and neuroinflammation[190].

García-Aranda et al., with AutoDock 4.2 docking, evaluated and compared the candidates of a series of 1,2-diphenylbenzimidazoles (DPBI) with different aromatic substitutions for their activity on bot COX-2 and h-iNOS active sites[191].

Molecular dynamics

If Molecular docking offers a static representation of the “binding event” between ligand and receptor, Molecular Dynamics (MD) allows exploring the behavior of the complex during the time. In MD, a system (protein-ligand, but also protein alone or ligand alone) is generated, solvated, relaxed, and simulated during the time to see the evolution in time of some desired observables. In the case of medicinal chemistry, MD is implemented in

various steps of the drug discovery process[192]. It can be used, as cited above, to assess the stability of a certain model created for a biological entity of interest. For model refinement, the trajectory obtained should show any huge change in RMSF of the backbone, as it could be a signal for model instability[193]. Even some ligand-protein structural data are present for a target, MD can help to understand the importance of ligand present in the complex (evaluating the geometrical and energetics parameters of the protein deprived of the ligand).

One of the uses of MD in SBDD is the docking poses refinement[170]. It is reasonable for docking poses to be compared for their ability to preserve their shape in a dynamic environment. The poses in which the ligands preserve the conformation obtained with the docking procedure are associated with a higher "virtual stability" and scored higher in respect to the poses which present a higher RMSD during the trajectory[194].

To perform MD simulations, several packages are available nowadays. Among the others, the most famous are Amber[195], GROMACS[196], CHARMM[197], and Schrödinger Desmond[198]. All these suites offer specific tools for system preparation, equilibration, and analysis (e.g., Amber relies on AmberTools for these purposes).

The classic MD protocols are based on molecular mechanics, so each parameter of the system (bonds, angles, partial charges, etc.) is set up at the beginning and does not change during the simulation. If the case considered by the scientist involves the presence of coordination bonds or the formation/break of bonds, different MD techniques must be implemented. This is the case of Quantum-Mechanics/Molecular-Mechanics (QM/MM) simulations, in which the site interested in the events cited above is treated with a QM level of theory, while the rest of the system is left to MM[199]. This allows obtaining reasonable results with a modest computational effort, which would become huge if all the systems were considered with QM theory.

Molecular dynamics is also applied to elucidate the mechanistic aspect of some biological events of difficult experimental evaluation, as several studies have witnessed[200]. Indeed, to investigate the inhibition mechanism of CCR5, Salmas et al. used MD, and the trajectories were processed with a principal component analysis and clustering method[201].

Banu et al., in a recent study, evaluated the interaction between NALP3 and Apoptosis-associated speck-like protein (ASC) in the inflammasome exploiting NAMD as the software for MD simulations[202]. A molecular dynamics approach was implemented by Tanwar et al. to explore the effect of different mutations on the structural behavior of N-sulphoglucosamine sulphohydrolase (SGSH), the main protein involved in mucopolysaccharidosis type IIIA[203]. Remaining in the field of LSDs, Hodošček et al. used MD simulations to investigate on the stability of the complex between the Niemann Pick type C proteins (NPC1 and NPC2) and cholesterol, considering the cases in which this molecule is bound first to one and then to the other of the pockets available in the system[204].

MD refinement of docking poses has been applied in various studies regarding neuroinflammation. Among the cited works, Ahmad et al.[190] and Tandon-Sinha[182] implemented this technique in the last part of their drug discovery workflow. Szefler and Czelen used an Amber-based MD approach for evaluating the most promising candidates in a library of oxindole derivatives as inhibitors of GSK3[205], while Kalva et al. analyzed the most promising candidates for MMP-9 inhibition derived from a cross-docking selectivity study[206]. Özkılıç and Tüzün analyzed with MD the best candidates resulting from a massive VS, which were firstly filtrated for ADME properties, highlighting two promising inhibitors of KMO[207].

In a recent methodological study, Jamal et al. implemented MD simulations for the evaluation of the structural and thermodynamic stability of the best candidates from a Glide docking VS[208].

The importance of ADMET in computational drug discovery for neuroinflammation

In drug discovery, and mainly in lead optimization, much importance must be given to the Blood-Brain-Barrier (BBB) passage of the candidate molecules. The BBB is a complex structure composed predominantly of endothelial cells, which are typically in contact with one another through tight junctions, which grant a very high filtration efficiency for the substances passing through this border.

Said this, it appears clear that a proper pharmacokinetic profile prediction is needed when the activity of a molecule versus a certain target located in the CNS is assessed. Moreover, looking at the topic of neuroinflammation, it is obvious that a molecule acting in the CNS is required to pass the BBB and reach a therapeutical concentration at the site of action. To exactly foretell if this is going to happen a priori is not a trivial task.

Moreover, the active transport mechanisms are active across the BBB in a very distinct way[209]. Among these systems, one of the most known and studied is the P-glycoprotein (P-gp) mediated efflux. The ability of this protein engine to transport potential bioactive molecules from inside to the outside of the BBB granted to this machinery the second name of Multidrug Resistance Protein 1 (MDR1)[210]. One of the most known examples in which the P-gp importance appears clear is the one related to the loperamide, an opioid-receptor agonist mainly acting on the μ-opioid receptors. The high lipophilicity of this molecule allows it to easily pass the BBB, but the P-gp efflux system can efficiently pump the drug against gradient[211], almost eliminating its cerebral side effects at the therapeutical dosage and limiting its action to the intestine.

From a computational point of view, a proper ADMET prediction can be achieved in various ways. An early evaluation can certainly be done on various molecular descriptors, coming both from the 2D structure of a molecule and from the 3D arrangement of its atoms. Among the 2D descriptors, the combination of the ones involved in Lipinski's rule of five (molecular weight, logarithm of partition coefficient, number of hydrogen bond acceptors, and donors) can certainly be considered[212]. Of course, the very basic knowledge coming from a 2D descriptor must be enriched with data coming from several other sources. Furthermore, the success rate of ADMET models can be improved if also high quality in vitro and in vivo data are included[213]. At the present day, machine learning and artificial intelligence-driven protocols are widely used, both in academics and in industry[214].

As for everything related to data, the efficiency of these model-based methods relies on the quality (and, to a lesser extent, on the quantity) of data used[215]. Different assays, different assay conditions, and differences in procedure protocols are all factors that strongly impact the final overall quality of the dataset. This makes industries the most

advantaged environment for the creation of clean datasets, mainly due to the remarkable consistency in their assay conditions and proceeding.

Some of the most known software used to predict the pharmacokinetic profile of molecules are Schrödinger QikProp[216], OptibriumStarDrop, Metasite[217], and SwissADME[218]. An exhaustive list of such tools can be found on more specialistic articles, like the one made by Peach et al[219]. Many of these tools do not simply calculate descriptors but are also able to model them for several ADMET predictions. These programs are in some cases also able to depict important features in terms of CYP interaction and molecule metabolism. In the last decades, the machine learning approaches allowed even to obtain models for P-gp interactions, as the scientific literature witnesses[220221222]. On the other hand, if the ability of prediction from data is in a constant increase, is always useful to remember the complexity of the biological systems considered in drug discovery. Even if the computational tools converge in one direction, the scientist has always to remember that all the predictive outputs coming from them are based on a limited (even if large, in some cases) dataset. It is also important to notice that the majority of the tools nowadays present can perform well in predicting molecules distribution based on passive transport in the various tissues of the organism but foretelling the effects of the active transport mechanisms remain a tough task[223]. The efficiency of the prediction relies on several key factors: the diversity of the dataset, the number of elements in the dataset which are similar to the examined molecules, the quality of the data, etc. Much importance has also to be given to the right choice of the variables to build the model upon. Indeed, it is not necessary the case that more variables considered equals a better model. The most suitable models, in fact, are generally simple, easily interpretable[224] and built on few, well-chosen variables (which could be descriptors or data coming from physicochemical, biological, or clinical data). A scheme of the features that a proper prediction model for BBB permeation should present is reported in Figure 3.

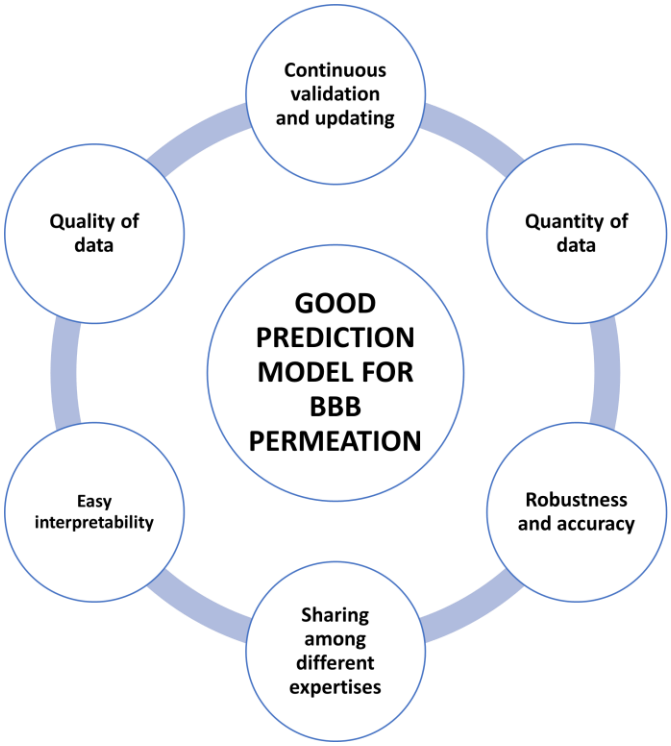


Figure 3. Scheme reporting the main characteristics that a proper model for blood-brain barrier (BBB) passage prediction should present.

A good prediction system for BBB passage based on molecular descriptors was developed by Hemmateenejad et al[225], exploiting a genetic algorithm-based artificial neural network, while Muehlbacher et al used a random forest-based method[226]. An ensemble method was recently applied in BBB passage prediction by Liu et al[227].

In the field of neuroinflammation, computational prediction of BBB passage has already been applied in various drug discovery protocols. Among these, Shahbazi et al. used Schrödinger QikProp for the development of potential NO-releasing anti-inflammatory molecules[228], while Dileep et al, with the same software, evaluated the BBB passage prediction of in silico designed potential multi-target ligands for neuroinflammation[229]. QikProp was also applied by Elrayess et al. for the ADME evaluation of potential triazole inhibitors of COX-2 against neurodegeneration[230].

Conclusion

In this review we highlighted the complexity beneath neuroinflammation, elucidating the main biological actors which play a role in its development. We then tried to give a comprehensive overview of the fundamental instruments and approaches nowadays

applied from computational medicinal chemists for the discovery and the development of innovative therapeutical entities, underlying for each its advantages and its limitations. For each technique, we reported some cases in which its application was implemented for drug discovery in neuroinflammation, looking mainly at the last decade of research. In the years to come, we hope that a deeper knowledge of the neuroinflammation biological mechanism, together with a further improvement in both the efficiency and the predictivity of the computational tools will be useful in finding innovative, valid, and safe therapies for the pathological states associated to neuroinflammation and neurodegeneration.

References

1. Lucas, S.-M., Rothwell, N. J. & Gibson, R. M. The role of inflammation in CNS injury and disease. Br. J. Pharmacol. 147, S232–S240 (2006).

2. Furman, D. et al. Chronic inflammation in the etiology of disease across the life span. Nat. Med. 2019 2512 25, 1822–1832 (2019).

3. Glass, C. K., Saijo, K., Winner, B., Marchetto, M. C. & Gage, F. H. Mechanisms Underlying Inflammation in Neurodegeneration. Cell 140, 918–934 (2010).

4. Gambino, C. M., Sasso, B. Lo, Bivona, G., Agnello, L. & Ciaccio, M. Aging and Neuroinflammatory Disorders: New Biomarkers and Therapeutic Targets. Curr. Pharm. Des. 25, 4168–4174 (2019).

5. Block, M. L. & Hong, J.-S. Microglia and inflammation-mediated neurodegeneration: Multiple triggers with a common mechanism. Prog. Neurobiol. 76, 77–98 (2005).

6. Kempuraj, D. et al. Neuroinflammation Induces Neurodegeneration. J. Neurol. Neurosurg. spine 1, (2016).

7. Woodling, N. S. & Andreasson, K. I. Untangling the Web: Toxic and Protective Effects of Neuroinflammation and PGE2 Signaling in Alzheimer's Disease. ACS Chem. Neurosci. 7, 454–463 (2016).

8. Hooten, K. G., Beers, D. R., Zhao, W. & Appel, S. H. Protective and Toxic Neuroinflammation in Amyotrophic Lateral Sclerosis. Neurotherapeutics 12, 364 (2015).

9. Obermeier, B., Daneman, R. & Ransohoff, R. M. Development, maintenance and disruption of the blood-brain barrier. Nat. Med. 19, 1584–1596 (2013).

10. Carson, M. J., Thrash, J. C. & Walter, B. The cellular response in neuroinflammation: The role of leukocytes, microglia and astrocytes in neuronal death and survival. Clin. Neurosci. Res. 6, 237 (2006).

11. Patel, J. P. & Frey, B. N. Disruption in the Blood-Brain Barrier: The Missing Link between Brain and Body Inflammation in Bipolar Disorder? Neural Plast. 2015, 1–12 (2015).

12. Wraith, D. C. & Nicholson, L. B. The adaptive immune system in diseases of the central nervous system. J. Clin. Invest. 122, 1172–1179 (2012).

13. Korn, T. & Kallies, A. T cell responses in the central nervous system. Nat. Rev. Immunol. 17, 179–194 (2017).

14. Fields, G. B. The Rebirth of Matrix Metalloproteinase Inhibitors: Moving Beyond the Dogma. Cells 8, 984 (2019).

15. Colombo, E. & Farina, C. Astrocytes: Key Regulators of Neuroinflammation. Trends Immunol. 37, 608–620 (2016).

16. Liddelow, S. A. & Barres, B. A. Reactive Astrocytes: Production, Function, and Therapeutic Potential. Immunity 46, 957–967 (2017).

17. Kwon, H. S. & Koh, S.-H. Neuroinflammation in neurodegenerative disorders: the roles of microglia and astrocytes. Transl. Neurodegener. 2020 91 9, 1–12 (2020).

18. Stephenson, J., Nutma, E., Valk, P. van der & Amor, S. Inflammation in CNS neurodegenerative diseases. Immunology 154, 204 (2018).

19. MV, S. Molecular dissection of reactive astrogliosis and glial scar formation. Trends Neurosci. 32, 638–647 (2009).

20. Block, M. L., Zecca, L. & Hong, J.-S. Microglia-mediated neurotoxicity: uncovering the molecular mechanisms. Nat. Rev. Neurosci. 8, 57–69 (2007).

21. Nimmerjahn, A. Resting Microglial Cells Are Highly Dynamic Surveillants of Brain Parenchyma in Vivo. Science (80-.). 308, 1314–1318 (2005).

22. Fetler, L. NEUROSCIENCE: Brain Under Surveillance: The Microglia Patrol. Science (80-.). 309, 392–393 (2005).

23. Rock, R. B. et al. Role of Microglia in Central Nervous System Infections. Clin. Microbiol. Rev. 17, 942–964 (2004).

24. E, P. & A, C. Reciprocal interactions between microglia and neurons: from survival to neuropathology. Rev. Neurosci. 13, 221–242 (2002).

25. V, K. Toll-like receptors in the pathogenesis of neuroinflammation. J. Neuroimmunol. 332, 16–30 (2019).

26. F, D. V., S, C., P, B. & MP, A. Purinergic signalling in inflammation of the central nervous system. Trends Neurosci. 32, 79–87 (2009).

27. J, H., JD, L., R, A., M, F. & SC, S. Scavenger receptors in neurobiology and neuropathology: their role on microglia and other cells of the nervous system. Glia 40, 195–205 (2002).

28. Kaminska, B., Mota, M. & Pizzi, M. Signal transduction and epigenetic mechanisms in the control of microglia activation during neuroinflammation. Biochim. Biophys. Acta - Mol. Basis Dis. 1862, 339–351 (2016).

29. Liu, T., Zhang, L., Joo, D. & Sun, S.-C. NF-κB signaling in inflammation. Signal Transduct. Target. Ther. 2017 21 2, 1–9 (2017).

30. Ji, R.-R., Xu, Z.-Z. & Gao, Y.-J. Emerging targets in neuroinflammation-driven chronic pain. Nat. Rev. Drug Discov. 13, 533–548 (2014).

31. Cianciulli, A. et al. Microglia Mediated Neuroinflammation: Focus on PI3K Modulation. Biomolecules 10, 137 (2020).

32. Karin, M. How NF-κB is activated: the role of the IκB kinase (IKK) complex. Oncogene 18, 6867–6874 (1999).

33. Ghosh, S. & Hayden, M. S. New regulators of NF-κB in inflammation. Nat. Rev. Immunol. 8, 837–848 (2008).

34. Jridi, I., Canté-Barrett, K., Pike-Overzet, K. & Staal, F. J. T. Inflammation and Wnt Signaling: Target for Immunomodulatory Therapy? Front. Cell Dev. Biol. 8, (2021).

35. Jia, L., Piña-Crespo, J. & Li, Y. Restoring Wnt/β-catenin signaling is a promising therapeutic strategy for Alzheimer's disease. Mol. Brain 2019 121 12, 1–11 (2019).

36. Stamos, J. L. & Weis, W. I. The β-Catenin Destruction Complex. Cold Spring Harb. Perspect. Biol. 5, (2013).

37. Becher, B., Spath, S. & Goverman, J. Cytokine networks in neuroinflammation. Nat. Rev. Immunol. 2016 171 17, 49–59 (2016).

38. Ramesh, G., Maclean, A. G. & Philipp, M. T. Cytokines and chemokines at the crossroads of neuroinflammation, neurodegeneration, and neuropathic pain. Mediators Inflamm. 2013, (2013).

39. JK, A. Oxidative stress in neurodegeneration: cause or consequence? Nat. Med. 10 Suppl, S18 (2004).

40. B, H. Oxidative stress and neurodegeneration: where are we now? J. Neurochem. 97, 1634–1658 (2006).

41. Qin, H., Niyongere, S. A., Lee, S. J., Baker, B. J. & Benveniste, E. N. Expression and Functional Significance of SOCS-1 and SOCS-3 in Astrocytes. J. Immunol. 181, 3167–3176 (2008).

42. Lallier, S. W., Graf, A. E., Waidyarante, G. R. & Rogers, L. K. Nurr1 expression is modified by inflammation in microglia. Neuroreport 27, 1120–1127 (2016).

43. Zhang, J.-M. & An, J. Cytokines, Inflammation, and Pain. Int. Anesthesiol. Clin. 45, 27–37 (2007).

44. Tiberi, M. & Chiurchiù, V. Specialized Pro-resolving Lipid Mediators and Glial Cells: Emerging Candidates for Brain Homeostasis and Repair. Front. Cell. Neurosci. 0, 136 (2021).

45. Serhan, C. N., Chiang, N. & Dyke, T. E. Van. Resolving inflammation: dual anti-inflammatory and pro-resolution lipid mediators. Nat. Rev. Immunol. 8, 349 (2008).

46. Guzman-Martinez, L. et al. Neuroinflammation as a Common Feature of Neurodegenerative Disorders. Front. Pharmacol. 10, (2019).

47. H, A. Inflammatory response in Alzheimer's disease. Tohoku J. Exp. Med. 174, 295–303 (1994).

48. J, H. & DJ, S. The amyloid hypothesis of Alzheimer's disease: progress and problems on the road to therapeutics. Science 297, 353–356 (2002).

49. Heppner, F. L., Ransohoff, R. M. & Becher, B. Immune attack: the role of inflammation in Alzheimer disease. Nat. Rev. Neurosci. 16, 358–372 (2015).

50. Heneka, M. T. et al. Neuroinflammation in Alzheimer's disease. Lancet Neurol. 14, 388–405 (2015).

51. Jiang, T., Yu, J. T., Zhu, X. C. & Tan, L. TREM2 in Alzheimer's disease. Mol. Neurobiol. 48, 180–185 (2013).

52. J. Bajramovic, J. Regulation of Innate Immune Responses in the Central Nervous System. CNS Neurol. Disord. - Drug Targets 10, 4–24 (2011).

53. Takahashi, K., Rochford, C. D. P. & Neumann, H. Clearance of apoptotic neurons without inflammation by microglial triggering receptor expressed on myeloid cells-2. J. Exp. Med. 201, 647–657 (2005).

54. Jay, T. R., Von Saucken, V. E. & Landreth, G. E. TREM2 in Neurodegenerative Diseases. Mol. Neurodegener. 2017 121 12, 1–33 (2017).

55. Zhou, S. L. et al. TREM2 Variants and Neurodegenerative Diseases: A Systematic Review and Meta-Analysis. J. Alzheimers. Dis. 68, 1171–1184 (2019).

56. Wang, Q., Liu, Y. & Zhou, J. Neuroinflammation in Parkinson's disease and its potential as therapeutic target. Transl. Neurodegener. 4, 19 (2015).

57. Braak, H. et al. Staging of brain pathology related to sporadic Parkinson's disease. Neurobiol. Aging 24, 197–211 (2003).

58. Oksanen, M. et al. Astrocyte alterations in neurodegenerative pathologies and their modeling in human induced pluripotent stem cell platforms. Cell. Mol. Life Sci. 76, 2739 (2019).

59. Rocha, N. P., Miranda, A. S. de & Teixeira, A. L. Insights into Neuroinflammation in Parkinson's Disease: From Biomarkers to Anti-Inflammatory Based Therapies. Biomed Res. Int. 2015, (2015).

60. Zhang, W. et al. Aggregated α-synuclein activates microglia: a process leading to disease progression in Parkinson's disease. FASEB J. 19, 533–542 (2005).

61. Rojanathammanee, L., Murphy, E. J. & Combs, C. K. Expression of mutant alpha-synuclein modulates microglial phenotype in vitro. J. Neuroinflammation 8, 44 (2011).

62. Hirsch, E. C. & Hunot, S. Neuroinflammation in Parkinson's disease: a target for neuroprotection? Lancet Neurol. 8, 382–397 (2009).

63. H, W. et al. Activation of microglia by human neuromelanin is NF-kappaB dependent and involves p38 mitogen-activated protein kinase: implications for Parkinson's disease. FASEB J. 17, 500–502 (2003).

64. Wang, X.-J. et al. Impaired CD200–CD200R-mediated microglia silencing enhances midbrain dopaminergic neurodegeneration: Roles of aging, superoxide, NADPH oxidase, and p38 MAPK. Free Radic. Biol. Med. 50, 1094–1106 (2011).

65. Sheridan, G. K. & Murphy, K. J. Neuron–glia crosstalk in health and disease: fractalkine and CX 3 CR1 take centre stage. Open Biol. 3, 130181 (2013).

66. M, D., N, V., A, B. & T, P. NURR1 in Parkinson disease--from pathogenesis to therapeutic potential. Nat. Rev. Neurol. 9, 629–636 (2013).

67. Rowland, L. P. & Shneider, N. A. Amyotrophic Lateral Sclerosis. N. Engl. J. Med. 344, 1688–1700 (2001).

68. Kiernan, M. C. et al. Amyotrophic lateral sclerosis. Lancet 377, 942–955 (2011).

69. M, N. et al. Ubiquitinated TDP-43 in frontotemporal lobar degeneration and amyotrophic lateral sclerosis. Science 314, 130–133 (2006).

70. VM, V. D. et al. TARDBP mutations in amyotrophic lateral sclerosis with TDP-43 neuropathology: a genetic and histopathological analysis. Lancet. Neurol. 7, 409–416 (2008).

71. J, L. & F, W. Role of Neuroinflammation in Amyotrophic Lateral Sclerosis: Cellular Mechanisms and Therapeutic Implications. Front. Immunol. 8, (2017).

72. Philips, T. & Robberecht, W. Neuroinflammation in amyotrophic lateral sclerosis: role of glial activation in motor neuron disease. Lancet Neurol. 10, 253–263 (2011).

73. Liu, Y., Hao, W., Dawson, A., Liu, S. & Fassbender, K. Expression of Amyotrophic Lateral Sclerosis-linked SOD1 Mutant Increases the Neurotoxic Potential of Microglia via TLR2. J. Biol. Chem. 284, 3691–3699 (2009).

74. Neymotin, A. et al. Lenalidomide (Revlimid®) administration at symptom onset is neuroprotective in a mouse model of amyotrophic lateral sclerosis. Exp. Neurol. 220, 191 (2009).

75. LH, B. et al. A role for astrocytes in motor neuron loss in amyotrophic lateral sclerosis. Brain Res. Brain Res. Rev. 47, 263–274 (2004).

76. Chiu, I. M. et al. T lymphocytes potentiate endogenous neuroprotective inflammation in a mouse model of ALS. Proc. Natl. Acad. Sci. 105, 17913–17918 (2008).

77. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. The Huntington's Disease Collaborative Research Group. Cell 72, 971–983 (1993).

78. CA, R. & SJ, T. Huntington's disease: from molecular pathogenesis to clinical treatment. Lancet. Neurol. 10, 83–98 (2011).

79. Browne, S. E. Mitochondria and Huntington's Disease Pathogenesis. Ann. N. Y. Acad. Sci. 1147, 358–382 (2008).

80. Schilling, G. et al. Characterization of Huntingtin Pathologic Fragments in Human Huntington Disease, Transgenic Mice, and Cell Models. J. Neuropathol. Exp. Neurol. 66, 313–320 (2007).

81. Ross, C. A. & Poirier, M. A. What is the role of protein aggregation in neurodegeneration? Nat. Rev. Mol. Cell Biol. 6, 891–898 (2005).

82. Truant, R., Atwal, R. S., Desmond, C., Munsie, L. & Tran, T. Huntington's disease: revisiting the aggregation hypothesis in polyglutamine neurodegenerative diseases. FEBS J. 275, 4252–4262 (2008).

83. Tabrizi, S. J. et al. Biological and clinical manifestations of Huntington's disease in the longitudinal TRACK-HD study: cross-sectional analysis of baseline data. Lancet Neurol. 8, 791–801 (2009).

84. Walker, F. O. Huntington's disease. Lancet 369, 218–228 (2007).

85. Hodges, A. et al. Regional and cellular gene expression changes in human Huntington's disease brain. Hum. Mol. Genet. 15, 965–977 (2006).

86. Tai, Y. F. et al. Microglial activation in presymptomatic Huntington's disease gene carriers. Brain 130, 1759–1766 (2007).

87. Hersch, S. M. et al. Creatine in Huntington disease is safe, tolerable, bioavailable in brain and reduces serum 8OH2'dG. Neurology 66, 250–252 (2006).

88. Giorgini, F., Guidetti, P., Nguyen, Q., Bennett, S. C. & Muchowski, P. J. A genomic screen in yeast implicates kynurenine 3-monooxygenase as a therapeutic target for Huntington disease. Nat. Genet. 37, 526–531 (2005).

89. Bosch, M. E. & Kielian, T. Neuroinflammatory paradigms in lysosomal storage diseases. Front. Neurosci. 9, 417 (2015).

90. Archer, L. D., Langford-Smith, K. J., Bigger, B. W. & Fildes, J. E. Mucopolysaccharide diseases: A complex interplay between neuroinflammation, microglial activation and adaptive immunity. J. Inherit. Metab. Dis. 37, 1–12 (2014).

91. Futerman, A. H. & van Meer, G. The cell biology of lysosomal storage disorders. Nat. Rev. Mol. Cell Biol. 5, 554–565 (2004).

92. Giugliani, R. et al. Mucopolysaccharidosis I, II, and VI: Brief review and guidelines for treatment. Genet. Mol. Biol. 33, 589 (2010).

93. Suarez-Guerrero, J. L., Gómez Higuera, P. J. I., Arias Flórez, J. S. & Contreras-García, G. A. Mucopolysaccharidosis: clinical features, diagnosis and management. Rev. Chil. pediatría 87, 295–304 (2016).

94. Valstar, M. J., Ruijter, G. J. G., van Diggelen, O. P., Poorthuis, B. J. & Wijburg, F. A. Sanfilippo syndrome: A mini-review. J. Inherit. Metab. Dis. 31, 240–252 (2008).

95. DiRosario, J. et al. Innate and adaptive immune activation in the brain of MPS IIIB mouse model. J. Neurosci. Res. 87, 978–990 (2009).

96. Villani, G. R. D. et al. Cytokines, neurotrophins, and oxidative stress in brain disease from mucopolysaccharidosis IIIB. J. Neurosci. Res. 85, 612–622 (2007).

97. Arfi, A. et al. Neuroinflammatory and oxidative stress phenomena in MPS IIIA mouse model: The positive effect of long-term aspirin treatment. Mol. Genet. Metab. 103, 18–25 (2011).

98. Kollmann, K. et al. Cell biology and function of neuronal ceroid lipofuscinosis-related proteins. Biochim. Biophys. Acta - Mol. Basis Dis. 1832, 1866–1881 (2013).

99. Haltia, M. The neuronal ceroid-lipofuscinoses. J. Neuropathol. Exp. Neurol. 62, 1–13 (2003).

100. Mole, S. E., Williams, R. E. & Goebel, H. H. Correlations between genotype, ultrastructural morphology and clinical phenotype in the neuronal ceroid lipofuscinoses. Neurogenetics 6, 107–126 (2005).

101. Anderson, G. W., Goebel, H. H. & Simonati, A. Human pathology in NCL. Biochim. Biophys. Acta - Mol. Basis Dis. 1832, 1807–1826 (2013).

102. Dolisca, S.-B., Mehta, M., Pearce, D. A., Mink, J. W. & Maria, B. L. Batten Disease. J. Child Neurol. 28, 1074–1100 (2013).

103. Mencarelli, C. & Martinez–Martinez, P. Ceramide function in the brain: when a slight tilt is enough. Cell. Mol. Life Sci. 70, 181–203 (2013).

104. Macauley, S. L. et al. Synergistic effects of central nervous system-directed gene therapy and bone marrow transplantation in the murine model of infantile neuronal ceroid lipofuscinosis. Ann. Neurol. 71, 797–804 (2012).

105. Butters, T. D. Gaucher disease. Curr. Opin. Chem. Biol. 11, 412–418 (2007).

106. Stirnemann, J. Ô. et al. A Review of Gaucher Disease Pathophysiology, Clinical Presentation and Treatments. Int. J. Mol. Sci. 18, 441 (2017).

107. Nagral, A. Gaucher Disease. J. Clin. Exp. Hepatol. 4, 37 (2014).

108. Nalysnyk, L., Rotella, P., Simeone, J. C., Hamed, A. & Weinreb, N. Gaucher disease epidemiology and natural history: a comprehensive review of the literature. Hematology 22, 65–73 (2017).

109. Sama, D. M. & Norris, C. M. Calcium dysregulation and neuroinflammation: Discrete and integrated mechanisms for age-related synaptic dysfunction. Ageing Res. Rev. 12, 982–995 (2013).

110. Alobaidy, H. Recent Advances in the Diagnosis and Treatment of Niemann-Pick Disease Type C in Children: A Guide to Early Diagnosis for the General Pediatrician. Int. J. Pediatr. 2015, 1–10 (2015).

111. Baudry, M., Yao, Y., Simmons, D., Liu, J. & Bi, X. Postnatal development of inflammation in a murine model of Niemann–Pick type C disease: immunohistochemical observations of microglia and astroglia. Exp. Neurol. 184, 887–903 (2003).

112. Gallala, H. D., Breiden, B. & Sandhoff, K. Regulation of the NPC2 protein-mediated cholesterol trafficking by membrane lipids. J. Neurochem. 116, 702–707 (2011).

113. Rosenbaum, A. I. & Maxfield, F. R. Niemann-Pick type C disease: molecular mechanisms and potential therapeutic approaches. J. Neurochem. 116, 789–795 (2011).

114. Patterson, M. Niemann-Pick Disease Type C. GeneReviews® (2020).

115. Vanier, M. T. Niemann-Pick disease type C. Orphanet J. Rare Dis. 5, 16 (2010).

116. Wang, M. L. et al. Identification of Surface Residues on Niemann-Pick C2 Essential for Hydrophobic Handoff of Cholesterol to NPC1 in Lysosomes. Cell Metab. 12, 166–173 (2010).

117. Peake, K. B., Campenot, R. B., Vance, D. E. & Vance, J. E. Niemann-Pick Type C1 deficiency in microglia does not cause neuron death in vitro. Biochim. Biophys. Acta - Mol. Basis Dis. 1812, 1121–1129 (2011).

118. Smith, D., Wallom, K.-L., Williams, I. M., Jeyakumar, M. & Platt, F. M. Beneficial effects of anti-inflammatory therapy in a mouse model of Niemann-Pick disease type C1. Neurobiol. Dis. 36, 242–251 (2009).

119. Williams, I. M. et al. Improved neuroprotection using miglustat, curcumin and ibuprofen as a triple combination therapy in Niemann–Pick disease type C1 mice. Neurobiol. Dis. 67, 9–17 (2014).

120. Gitler, A. D., Dhillon, P. & Shorter, J. Neurodegenerative disease: models, mechanisms, and a new hope. Dis. Model. Mech. 10, 499–502 (2017).

121. Leicht, H. et al. Predictors of Costs in Dementia in a Longitudinal Perspective. PLoS One 8, e70018 (2013).

122. Durães, F., Pinto, M. & Sousa, E. Old Drugs as New Treatments for Neurodegenerative Diseases. Pharmaceuticals 11, (2018).

123. Gurung, A. B., Ali, M. A., Lee, J., Farah, M. A. & Al-Anazi, K. M. An Updated Review of Computer-Aided Drug Design and Its Application to COVID-19. Biomed Res. Int. 2021, (2021).

124. Yu, W. & MacKerell, A. D. Computer-Aided Drug Design Methods. in (2017). doi:10.1007/978-1-4939-6634-9_5.

125. Lionta, E., Spyrou, G., Vassilatis, D. & Cournia, Z. Structure-Based Virtual Screening for Drug Discovery: Principles, Applications and Recent Advances. Curr. Top. Med. Chem. 14, (2014).

126. Krishnan, V. & Rupp, B. Macromolecular Structure Determination: Comparison of X-ray Crystallography and NMR Spectroscopy. in eLS (John Wiley & Sons, Ltd, 2012). doi:10.1002/9780470015902.a0002716.pub2.

127. Natchiar, S. K., Myasnikov, A. G., Kratzat, H., Hazemann, I. & Klaholz, B. P. Visualization of chemical modifications in the human 80S ribosome structure. Nature 551, (2017).

128. Benjin, X. & Ling, L. Developments, applications, and prospects of cryo-electron microscopy. Protein Sci. 29, (2020).

129. Aparoy, P., Kumar Reddy, K. & Reddanna, P. Structure and Ligand Based Drug Design Strategies in the Development of Novel 5- LOX Inhibitors. Curr. Med. Chem. 19, (2012).

130. Vázquez, J., López, M., Gibert, E., Herrero, E. & Luque, F. J. Merging Ligand-Based and Structure-Based Methods in Drug Discovery: An Overview of Combined Virtual Screening Approaches. Molecules 25, (2020).

131. Shim, J. & MacKerell, J. . A. D. Computational ligand-based rational design: role of conformational sampling and force fields in model development. Medchemcomm 2, (2011).

132. Lavecchia, A. Machine-learning approaches in drug discovery: methods and applications. Drug Discov. Today 20, (2015).

133. Alexander-Brett, J. M. & Fremont, D. H. Dual GPCR and GAG mimicry by the M3 chemokine decoy receptor. J. Exp. Med. 204, (2007).

134. Toledo-Sherman, L. et al. Optimization of Potent and Selective Ataxia Telangiectasia-Mutated Inhibitors Suitable for a Proof-of-Concept Study in Huntington's Disease Models. J. Med. Chem. 62, (2019).

135. Long, A., Zhao, H. & Huang, X. Structural Basis for the Interaction between Casein Kinase 1 Delta and a Potent and Selective Inhibitor. J. Med. Chem. 55, (2012).

136. Elkins, P. A. et al. Structure of the C-terminally truncated human ProMMP9, a gelatin-binding matrix metalloproteinase. Acta Crystallogr. Sect. D Biol. Crystallogr. 58, (2002).

137. Madhavi Sastry, G., Adzhigirey, M., Day, T., Annabhimoju, R. & Sherman, W. Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. J. Comput. Aided. Mol. Des. 27, (2013).

138. Cavasotto, C. N. & Phatak, S. S. Homology modeling in drug discovery: current trends and applications. Drug Discov. Today 14, (2009).

139. Xiang, Z. Advances in Homology Protein Structure Modeling. Curr. Protein Pept. Sci. 7, (2006).

140. Bender, B. J., Marlow, B. & Meiler, J. Improving homology modeling from low-sequence identity templates in Rosetta: A case study in GPCRs. PLOS Comput. Biol. 16, (2020).

141. Šali, A. & Blundell, T. L. Comparative Protein Modelling by Satisfaction of Spatial Restraints. J. Mol. Biol. 234, (1993).

142. Jacobson, M. P. et al. A hierarchical approach to all-atom protein loop prediction. Proteins Struct. Funct. Bioinforma. 55, (2004).

143. Molecular Operating Environment (MOE), 2019.01; Chemical Computing Group ULC, 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, 2021. https://www.chemcomp.com/Research-Citing_MOE.htm.

144. Waterhouse, A. et al. SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res. 46, (2018).

145. Muhammed, M. T. & Aki-Yalcin, E. Homology modeling in drug discovery: Overview, current applications, and future perspectives. Chem. Biol. Drug Des. 93, (2019).

146. Service, R. 'The game has changed.' AI triumphs at solving protein structures. Science (80-.). (2020) doi:10.1126/science.abf9367.

147. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589 (2021).

148. Callaway, E. 'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures. Nature 588, 203–204 (2020).

149. Senior, A. W. et al. Protein structure prediction using multiple deep neural networks in the 13th Critical Assessment of Protein Structure Prediction (CASP13). Proteins Struct. Funct. Bioinforma. 87, 1141–1148 (2019).

150. Kryshtafovych, A. et al. Modeling SARS-CoV-2 proteins in the CASP-commons experiment. Proteins Struct. Funct. Bioinforma. 89, 1987–1996 (2021).

151. Mullard, A. What does AlphaFold mean for drug discovery? Nat. Rev. Drug Discov. 20, 725–727 (2021).

152. Fan, H. Refinement of homology-based protein structures by molecular dynamics simulation techniques. Protein Sci. 13, (2004).

153. Bhargavi, M., Sivan, S. K. & Potlapally, S. R. Identification of novel anti cancer agents by applying insilico methods for inhibition of TSPO protein. Comput. Biol. Chem. 68, (2017).

154. Lai, H. T. T. & Nguyen, T. T. Construction of dimeric hTSPO protein model using homology modeling and molecular dynamics. J. Phys. Conf. Ser. 1932, (2021).

155. Leelananda, S. P. & Lindert, S. Computational methods in drug discovery. Beilstein J. Org. Chem. 12, (2016).

156. Wang, G. & Zhu, W. Molecular docking for drug discovery and development: a widely used approach but far from perfect. Future Med. Chem. 8, (2016).

157. Halperin, I., Ma, B., Wolfson, H. & Nussinov, R. Principles of docking: An overview of search algorithms and a guide to scoring functions. Proteins Struct. Funct. Genet. 47, (2002).

158. Jones, G., Willett, P., Glen, R. C., Leach, A. R. & Taylor, R. Development and validation of a genetic algorithm for flexible docking 1 1Edited by F. E. Cohen. J. Mol. Biol. 267, 727–748 (2002).

159. Morris, G. M. et al. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. J. Comput. Chem. 30, (2009).

160. Friesner, R. A. et al. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. J. Med. Chem. 47, (2004).

161. Korb, O., Stützle, T. & Exner, T. E. PLANTS: Application of Ant Colony Optimization to Structure-Based Drug Design. 247–258 (2006) doi:10.1007/11839088_22.

162. Gagnon, J. K., Law, S. M. & Brooks, C. L. Flexible CDOCKER: Development and application of a pseudo-explicit structure-based docking method within CHARMM. J. Comput. Chem. 37, (2016).

163. Cross, J. B. et al. Comparison of Several Molecular Docking Programs: Pose Prediction and Virtual Screening Accuracy. J. Chem. Inf. Model. 49, (2009).

164. Wang, Z. et al. Comprehensive evaluation of ten docking programs on a diverse set of protein–ligand complexes: the prediction accuracy of sampling power and scoring power. Phys. Chem. Chem. Phys. 18, (2016).

165. Lapillo, M. et al. Extensive Reliability Evaluation of Docking-Based Target-Fishing Strategies. Int. J. Mol. Sci. 20, (2019).

166. Cuzzolin, A., Sturlese, M., Malvacio, I., Ciancetta, A. & Moro, S. DockBench: An integrated informatic platform bridging the gap between the robust validation of docking protocols and virtual screening simulations. Molecules 20, 9977–9993 (2015).

167. Warren, G. L. et al. A Critical Assessment of Docking Programs and Scoring Functions. J. Med. Chem. 49, (2006).

168. Peach, M. L. & Nicklaus, M. C. Combining docking with pharmacophore filtering for improved virtual screening. J. Cheminform. 1, (2009).

169. Dixon, S. L., Smondyrev, A. M. & Rao, S. N. PHASE: A Novel Approach to Pharmacophore Modeling and 3D Database Searching. Chem. Biol. <html_ent glyph="@amp;" ascii="&"/> Drug Des. 67, (2006).

170. Rastelli, G. & Pinzi, L. Refinement and Rescoring of Virtual Screening Results. Front. Chem. 7, (2019).

171. Meng, X.-Y., Zhang, H.-X., Mezei, M. & Cui, M. Molecular Docking: A Powerful Approach for Structure-Based Drug Discovery. Curr. Comput. Aided-Drug Des. 7, 146–157 (2012).

172. Gorgulla, C. et al. An open-source drug discovery platform enables ultra-large virtual screens. Nature 580, (2020).

173. Bolcato, G., Cuzzolin, A., Bissaro, M., Moro, S. & Sturlese, M. Can We Still Trust Docking Results? An Extension of the Applicability of DockBench on PDBbind Database. Int. J. Mol. Sci. 20, (2019).

174. Cuzzolin, A., Deganutti, G., Salmaso, V., Sturlese, M. & Moro, S. AquaMMapS: An Alternative Tool to Monitor the Role of Water Molecules During Protein–Ligand Association. ChemMedChem 13, 522–531 (2018).

175. Roberts, B. C. & Mancera, R. L. Ligand−Protein Docking with Water Molecules. J. Chem. Inf. Model. 48, (2008).

176. Houston, D. R. & Walkinshaw, M. D. Consensus Docking: Improving the Reliability of Docking in a Virtual Screening Context. J. Chem. Inf. Model. 53, (2013).

177. Tuccinardi, T., Poli, G., Romboli, V., Giordano, A. & Martinelli, A. Extensive Consensus Docking Evaluation for Ligand Pose Prediction and Virtual Screening Studies. J. Chem. Inf. Model. 54, (2014).

178. Cheng, B. et al. Synthesis and Anti-neuroinflammatory Activity of Lactone Benzoyl Hydrazine and 2-nitro-1-phenyl-1 H -Indole Derivatives as p38 α MAPK Inhibitors. Chem. Biol. Drug Des. 86, (2015).

179. Rippin, I. et al. Discovery and Design of Novel Small Molecule GSK-3 Inhibitors Targeting the Substrate Binding Site. Int. J. Mol. Sci. 21, (2020).

180. Cescon, E. et al. Scaffold Repurposing of in-House Chemical Library toward the Identification of New Casein Kinase 1 δ Inhibitors. ACS Med. Chem. Lett. 11, (2020).

181. Redenti, S. et al. A Triazolotriazine-Based Dual GSK-3β/CK-1δ Ligand as a Potential Neuroprotective Agent Presenting Two Different Mechanisms of Enzymatic Inhibition. ChemMedChem 14, (2019).

182. Tandon, A. & Sinha, S. Structural insights into the binding of MMP9 inhibitors. Bioinformation 5, (2011).

183. Razak, S. et al. Molecular docking, pharmacokinetic studies, and in vivo pharmacological study of indole derivative 2-(5-methoxy-2-methyl-1H-indole-3-yl)-N'-[(E)-(3-nitrophenyl) methylidene] acetohydrazide as a promising chemoprotective agent against cisplatin induced org. Sci. Rep. 11, (2021).

184. Liu, L.-J. et al. Identification of a natural product-like STAT3 dimerization inhibitor by structure-based virtual screening. Cell Death Dis. 5, (2014).

185. Ray, S. S., Nowak, R. J., Brown, R. H. & Lansbury, P. T. Small-molecule-mediated stabilization of familial amyotrophic lateral sclerosis-linked superoxide dismutase mutants against unfolding and aggregation. Proc. Natl. Acad. Sci. 102, (2005).

186. Durai, P. et al. Toll-like receptor 2 antagonists identified through virtual screening and experimental validation. FEBS J. 284, (2017).

187. Mahita, J., Harini, K., Rao Pichika, M. & Sowdhamini, R. An in silico approach towards the identification of novel inhibitors of the TLR-4 signaling pathway. J. Biomol. Struct. Dyn. 34, (2016).

188. Yilmazer, B. et al. Investigation of novel pharmacological chaperones for Gaucher Disease. J. Mol. Graph. Model. 76, 364–378 (2017).

189. El-Zohairy, M. A., Zlotos, D. P., Berger, M. R., Adwan, H. H. & Mandour, Y. M. Discovery of Novel CCR5 Ligands as Anticolorectal Cancer Agents by Sequential Virtual Screening. ACS Omega 6, (2021).

190. Ahmad, K. et al. Targeting Caspase 8: Using Structural and Ligand-Based Approaches to Identify Potential Leads for the Treatment of Multi-Neurodegenerative Diseases. Molecules 24, (2019).

191. García-Aranda, M. I. et al. Anti-inflammatory effect and inhibition of nitric oxide production by targeting COXs and iNOS enzymes with the 1,2-diphenylbenzimidazole pharmacophore. Bioorg. Med. Chem. 28, (2020).

192. Martinez-Rosell, G., Harvey, M. J. & De Fabritiis, G. Molecular-Simulation-Driven Fragment Screening for the Discovery of New CXCL12 Inhibitors. J. Chem. Inf. Model. 58, (2018).

193. De Vivo, M., Masetti, M., Bottegoni, G. & Cavalli, A. Role of Molecular Dynamics and Related Methods in Drug Discovery. J. Med. Chem. 59, (2016).

194. Hollingsworth, S. A. & Dror, R. O. Molecular Dynamics Simulation for All. Neuron 99, (2018).

195. D.A. Case et al. Amber 2021, University of California, San Francisco. (2021).

196. Berendsen, H. J. C., van der Spoel, D. & van Drunen, R. GROMACS: A message-passing parallel molecular dynamics implementation. Comput. Phys. Commun. 91, (1995).

197. Brooks, B. R. et al. CHARMM: The biomolecular simulation program. J. Comput. Chem. 30, (2009).

198. Bowers, K. J. et al. Molecular dynamics---Scalable algorithms for molecular dynamics simulations on commodity clusters. in Proceedings of the 2006 ACM/IEEE conference on Supercomputing - SC '06 (ACM Press, 2006). doi:10.1145/1188455.1188544.

199. Lodola, A. & De Vivo, M. The Increasing Role of QM/MM in Drug Discovery. in (2012). doi:10.1016/B978-0-12-398312-1.00011-1.

200. Bunker, A. & Róg, T. Mechanistic Understanding From Molecular Dynamics Simulation in Pharmaceutical Research 1: Drug Delivery. Front. Mol. Biosci. 7, (2020).

201. Salmas, R. E., Yurtsever, M. & Durdagi, S. Investigation of Inhibition Mechanism of Chemokine Receptor CCR5 by Micro-second Molecular Dynamics Simulations. Sci. Rep. 5, (2015).

202. Banu, H., Joseph, M. C. & Nisar, M. N. In-silico approach to investigate death domains associated with nano-particle-mediated cellular responses. Comput. Biol. Chem. 75, (2018).

203. Tanwar, H., Kumar, D. T., Doss, C. G. P. & Zayed, H. Bioinformatics classification of mutations in patients with Mucopolysaccharidosis IIIA. Metab. Brain Dis. 34, 1577–1594 (2019).

204. Hodošček, M. & Elghobashi-Meinhardt, N. Simulations of NPC1(NTD):NPC2 Protein Complex Reveal Cholesterol Transfer Pathways. Int. J. Mol. Sci. 19, 2623 (2018).

205. Czeleń, P. & Szefler, B. The Oxindole Derivatives, New Promising GSK-3β Inhibitors as One of the Potential Treatments for Alzheimer's Disease—A Molecular Dynamics Approach. Biology (Basel). 10, (2021).

206. Kalva, S., Agrawal, N., Skelton, A. A. & Saleena, L. M. Identification of novel selective MMP-9 inhibitors as potential anti-metastatic lead using structure-based hierarchical virtual screening and molecular dynamics simulation. Mol. Biosyst. 12, (2016).

207. Özkılıç, Y. & Tüzün, N. Ş. In silico methods predict new blood-brain barrier permeable structure for the inhibition of kynurenine 3-monooxygenase. J. Mol. Graph. Model. 100, (2020).

208. Jamal, S., Grover, A. & Grover, S. Machine Learning From Molecular Dynamics Trajectories to Predict Caspase-8 Inhibitors Against Alzheimer's Disease. Front. Pharmacol. 10, (2019).

209. Löscher, W. & Potschka, H. Blood-brain barrier active efflux transporters: ATP-binding cassette gene family. NeuroRX 2, (2005).

210. Daneman, R. & Prat, A. The Blood–Brain Barrier. Cold Spring Harb. Perspect. Biol. 7, (2015).

211. Upton, R. N. CEREBRAL UPTAKE OF DRUGS IN HUMANS. Clin. Exp. Pharmacol. Physiol. 34, (2007).

212. Benet, L. Z., Hosey, C. M., Ursu, O. & Oprea, T. I. BDDCS, the Rule of 5 and drugability. Adv. Drug Deliv. Rev. 101, (2016).

213. Wang, Y. et al. In silico ADME/T modelling for rational drug design. Q. Rev. Biophys. 48, (2015).

214. Bhhatarai, B., Walters, W. P., Hop, C. E. C. A., Lanza, G. & Ekins, S. Opportunities and challenges using artificial intelligence in ADME/Tox. Nat. Mater. 18, (2019).

215. Göller, A. H. et al. Bayer's in silico ADMET platform: a journey of machine learning over the past two decades. Drug Discov. Today 25, (2020).

216. Jorgensen, W. L. & Duffy, E. M. Prediction of drug solubility from structure. Adv. Drug Deliv. Rev. 54, (2002).

217. Cruciani, G. et al. MetaSite: Understanding Metabolism in Human Cytochromes from the Perspective of the Chemist. J. Med. Chem. 48, (2005).

218. Daina, A., Michielin, O. & Zoete, V. SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. Sci. Rep. 7, (2017).

219. Peach, M. L. et al. Computational tools and resources for metabolism-related property predictions. 1. Overview of publicly available (free and commercial) databases and software. Future Med. Chem. 4, (2012).

220. Kadioglu, O. & Efferth, T. A Machine Learning-Based Prediction Platform for P-Glycoprotein Modulators and Its Validation by Molecular Docking. Cells 8, (2019).

221. Watanabe, R. et al. Development of an In Silico Prediction Model for P-glycoprotein Efflux Potential in Brain Capillary Endothelial Cells toward the Prediction of Brain Penetration. J. Med. Chem. 64, (2021).

222. Montanari, F. & Ecker, G. F. Prediction of drug–ABC-transporter interaction — Recent advances and future challenges. Adv. Drug Deliv. Rev. 86, (2015).

223. Yang, N. J. & Hinner, M. J. Getting Across the Cell Membrane: An Overview for Small Molecules, Peptides, and Proteins. in (2015). doi:10.1007/978-1-4939-2272-7_3.

224. Madden, J. C., Enoch, S. J., Paini, A. & Cronin, M. T. D. A Review of In Silico Tools as Alternatives to Animal Testing: Principles, Resources and Applications. Altern. to Lab. Anim. 48, (2020).

225. Hemmateenejad, B., Miri, R., Safarpour, M. A. & Mehdipour, A. R. Accurate prediction of the blood-brain partitioning of a large set of solutes using ab initio calculations and genetic neural network modeling. J. Comput. Chem. 27, (2006).

226. Muehlbacher, M., Spitzer, G. M., Liedl, K. R. & Kornhuber, J. Qualitative prediction of blood–brain barrier permeability on a large and refined dataset. J. Comput. Aided. Mol. Des. 25, (2011).

227. Liu, L. et al. Prediction of the Blood–Brain Barrier (BBB) Permeability of Chemicals Based on Machine-Learning and Ensemble Methods. Chem. Res. Toxicol. 34, (2021).

228. Shahbazi, S. et al. Impact of novel N -aryl piperamide NO donors on NF-κB translocation in neuroinflammation: rational drug-designing synthesis and biological evaluation. Innate Immun. 24, (2018).

229. Dileep, K. V., Remya, C., Tintu, I. & Sadasivan, C. Designing of multi-target-directed ligands against the enzymes associated with neuroinflammation: an in silico approach. Front. Life Sci. 7, (2013).

230. Elrayess, R. et al. Synthesis, 3D-QSAR, and Molecular Modeling Studies of Triazole Bearing Compounds as a Promising Scaffold for Cyclooxygenase-2 Inhibition. Pharmaceuticals 13, (2020).

A Computational Workflow for the Identification of Novel Fragments Acting as Inhibitors of the Activity of Protein Kinase CK1δ

Giovanni Bolcato, Eleonora Cescon, **Matteo Pavan**, Maicol Bissaro, Davide Bassani, Stephanie Federico, Giampiero Spalluto, Mattia Sturlese and Stefano Moro

Bolcato, G. *et al.* A Computational Workflow for the Identification of Novel Fragments Acting as Inhibitors of the Activity of Protein Kinase CK1δ. *Int J Mol Sci* **22**, 9741 (2021).

Abstract

Fragment-Based Drug Discovery (FBDD) has become, in recent years, a consolidated approach in the drug discovery process, leading to several drug candidates under investigation in clinical trials and some approved drugs. Among these successful applications of the FBDD approach, kinases represent a class of targets where this strategy has demonstrated its real potential with the approved kinase inhibitor Vemurafenib. In the Kinase family, protein kinase CK1 isoform δ (CK1δ) has become a promising target in the treatment of different neurodegenerative diseases such as Alzheimer's disease, Parkinson's disease, and amyotrophic lateral sclerosis. In the present work, we set up and applied a computational workflow for the identification of putative fragment binders in large virtual databases. To validate the method, the selected compounds were tested in vitro to assess the CK1δ inhibition.

1. Introduction

1.1. Protein Kinase CK1δ

Protein kinase CK1δ belongs to the family of CK1 Kinases (Casein Kinase 1), which in turn belongs to the class of Ser-Thr Kinases; seven isoforms of this family were identified in mammals: α, β, γ1, γ2, γ3, δ, ε. All the isoforms of CK1 are constitutionally active and they exhibit activity in monomeric form, They present a highly conserved catalytic domain (unlike in N and terminal C domains), they utilize ATP as a phosphate group donator and they are generally independent of the presence of a cofactor [1].

CK1δ and the other isoforms of the family of CK1 can phosphorylate Ser or Thr residues in sequences such as *(P)Ser/Thr-X_{1-2}-Ser/Thr,* where *(P)Ser/Thr* indicates a Ser or Thr pre-phosphorylated residue [2]; CK1, therefore, needs the substrate to be already

phosphorylated. Nevertheless, it has been demonstrated that a set of amino acids with acidic character in the direction of the N-terminal with respect to Ser/Thr target residue or an acidic residue in position 3 (preferably Asp) can provide for the lack of the pre-phosphorylated amino acid[3,4]. This allows CK1 to act also as a Priming Kinase activating the substrate towards a second enzyme by phosphorylation. Currently, about 140 substrates (in vitro or in vivo) recognized by CK1 have been described[1].

The activity of CK1 isoforms is regulated in different ways. Phosphorylation is the principal strategy adopted for the activity regulation of this family of kinases. CK1δ is phosphorylated by kinases such as Akt, PKA, PKCα, CLK2, and Chk1. Moreover, CK1δ can also be subjected to auto-phosphorylation [1,5,6]. Another fundamental aspect in the CK1δ activity regulation is the subcellular compartmentalization, operated through the binding to intracellular structures and other proteins [7,8]. One last mechanism reported in the literature for the CK1δ regulation is the formation of homodimers [9,10].

CK1δ, together with other CK1 isoforms, has been correlated to several neurodegenerative processes[11]; in particular, CK1 seems implied in tauopathies, among which Alzheimer's disease (AD) is the most representative one.

AD is associated with several cellular processes. The first mechanism described is correlated to Tau protein, which after phosphorylation tends to come off from the microtubules forming aggregates at a cytoplasmatic level, leading to cellular damage. A second mechanism implies instead production and accumulation, with consequent cellular death, of the β-amyloid peptide. This is produced by the cut of its precursor APP (Amyloid Precursor Protein) by β-secretase 1 and γ-secretase enzymes. The implications of CK1 isoforms in pathogenetic processes at the root of Alzheimer's disease are many. In general, CK1δ proves to be overexpressed in brain tissue, up to 30 times in patients affected by Alzheimer's disease[12,13].

Concerning Tau protein, initially, it was observed how CK1 turns out to be associated with fibrillar masses of hyperphosphorylated Tau protein (Paired Helical filaments)[14]; in particular, CK1δ seems to be accumulated within these fibrillar masses[15]. Later it was demonstrated how CK1δ can phosphorylate Tau protein causing its separation from microtubules; the residues of Tau phosphorylated by CK1δ are Ser202, Thr205, Ser396, and Thr404[11,16]. As regards β-amyloid peptide, it was described how this can stimulate the

activity of CK1 and CK2 (employing casein as a substrate)[17]. Likewise, there is evidence that CK1 activity would be proportionally correlated to β-amyloid peptide production, since in presence of constitutionally active CK1 forms the amount of this peptide increases, whereas it decreases in presence of CK1 inhibitors. CK1 interference seems to take place along with the γ-secretase enzyme[18], but it is more likely correlated to CK1ε isoform, than to CK1δ[19].

As regards Parkinson's Disease, it has been observed how CK1 isoforms phosphorylated Ser129 of α-synuclein[11,20].

Amyotrophic lateral sclerosis (ALS) is another neurodegenerative disease where CK1δ plays a role. Indeed, CK1δ phosphorylated TDP-43 (Trans Activate Response DNA Binding Protein 43) at many different residues. TDP-43 is the principal component of the protein aggregates observed in the pathogenesis of ALS[21,22].

1.2. Fragment-Based Drug Discovery (FBDD) Principles.

FBDD is a strategy used in drug discovery that has gained popularity both in the industrial and academic contexts. In a typical FBDD process a library of polar low molecular weight compounds is screened against a specific target. Usually, the screening is performed by biophysical methods including X-ray crystallography, nuclear magnetic resonance (NMR), thermal shift assay, and surface plasmon resonance (SPR). One of the key factors in the FBDD success is the smaller size of the fragment-like chemical space compared to the size of the drug-like one. The size of the drug-like chemical space has been estimated at around 10^{60} compounds, many orders of magnitude greater than that of the fragment-like compounds' chemical space[23]. This means that, through the screening of fragments, the portion of chemical space sampled is larger than the one sampled with the screening of drug-like molecules. This will promisingly also allow the attainment of innovative scaffolds for drug candidates.

Despite the hit fragments having typically a low affinity, they could be turned into a lead compound that efficiently binds the target. Fragments, having a low molecular weight, establish few interactions with the target; however, the combination of multiple fragments by linking and merging or by decorating them with adequate functional groups (fragment growing) allows the development of specific and more affine compounds.

1.3. Fragment-Based Drug Discovery and Kinase Inhibitors

Concerning the identification of kinase inhibitors through an FBDD approach, X-ray crystallography has also been largely employed because kinases represent a class of protein that provides good results with this technique.

The most outstanding example of kinase inhibitors derived from an FBDD approach is Vemurafenib (inhibitor of BRAF) which is an approved drug for the treatment of metastatic melanoma[24]. The discovery of vemurafenib started with an enzymatic assay screening of a fragment library. The hit compounds identified were analyzed through X-ray crystallography, using the structural information so obtained one fragment was chosen for optimization leading at the end to Vemurafenib[25]. Another notable example is Asciminib an allosteric inhibitor of BCR-ABL1 tyrosine kinase, now in phase III clinical trial for resistant chronic myeloid leukemia. This compound was identified from an NMR-based fragment screening; the fragment hits identified were then optimized using In Silico methodologies, X-ray crystallography, and NMR[26,27].

Many other Kinase inhibitors derived from FBDD approaches are in clinical trials; for a comprehensive review of FBDD derived drugs that have been approved or which are in clinical trials see[28].

An interesting observation is that the fragments identified often bind at the hinge region of the kinase and maintain this binding mode also in the mature compound. For this reason, the library of compounds tested in the present work has been focalized, using in silico methodologies described in the next sections, to be composed of putative hinge-binding fragments.

1.4. Computational Methods in FBDD

Since the dawn of FBDD, computational chemistry has played a major role in both fragments' hit identification and in the process of fragment optimization. The MCSS (multiple copy simultaneous search) algorithm[29] was a pioneering method for the study of fragment binding modes in a protein site. Another method for fragment posing based on grand canonical Monte Carlo (GCMC) has been reported[30].

Over the years many in silico methods have been proposed not only for fragment placement prediction but also to aid the fragment optimization process. Software like LUDI[31], HOOK[32], CAVEAT[33], RECORE[34], and many others have been developed for this

purpose. Additionally, Schrodinger[35] and CCG[36] implement in their software suites many tools to aid the fragment optimization process.

Molecular dynamics (MD)-based tools represent the most advanced in silico techniques used in FBDD. The first application of MD to FBDD was the refinement of docking poses, a method note as post-docking[37]. More advanced protocols have also been developed. Nonequilibrium candidate Monte Carlo (NCMC) is an algorithm that has been applied to enhance the sampling of fragment binding modes[38]; this method has been successfully applied to FBDD[39]. Another promising approach is the application of Markov state models to MD simulations, which has proved its potential to FBDD[40]. Recently, Supervised Molecular Dynamics (SuMD)[41] has been applied as a fragment screening tool[42].

Molecular docking has also become a routinely used tool in FBDD. While the conformational sampling performs by docking protocols is generally effective in reproducing the correct pose for a ligand, the scoring functions frequently fail in valuating this pose[43], this is especially true for Fragment-like compounds for which many doubts have been raised about the docking applicability[44]. This said, to make the docking results more reliable a consensus docking approach was used[45], and instead of the scoring function, the poses were evaluated using a pharmacophore model. A post-docking refinement of the poses was then performed. A detailed explanation of the computational workflow adopted in the present work is reported in Section 4.1, Section 4.2, and Section 4.3.

2. Results

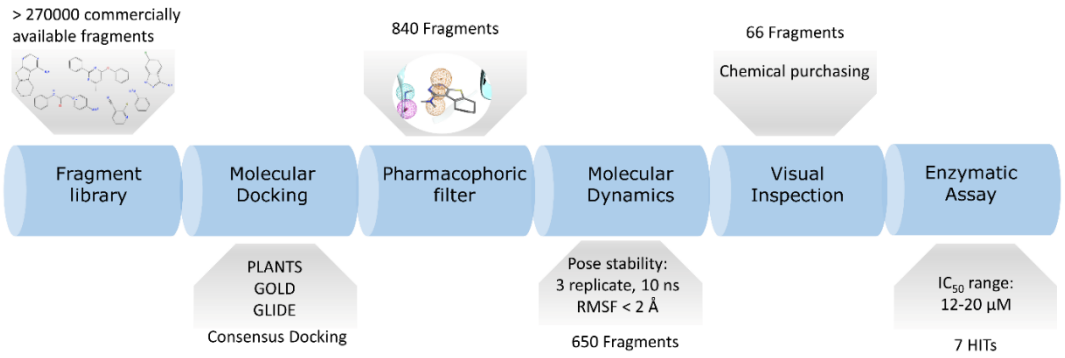


Figure 1. Schematic representation of the workflow adopted in the present work. First the fragments are retrieved from several vendors libraries. After proper preparation, the database is docked using three

different docking protocols. The resulting poses have been filtered using a pharmacophore model and only the molecule that fit the model for each protocol have been retained. The poses of these molecules were further refined using MD to assess the stability of the binding mode. The molecules that appear to be stable were finally selected through visual inspection.

2.1. Computational Results

A library of around 272,000 commercially available fragment compounds was screened in silico using an integrated structure-based approach based on different techniques such as molecular docking, molecular dynamics (MD), and pharmacophore filter. The workflow adopted is reported in Figure 1.

At first, three independent docking-based virtual screenings were performed in parallel exploiting three different protocols: PLANTS-ChemPLP, GOLD-ChemScore, and Glide-SP. PLANTS exploits an Ant-Colony Optimization (ACO) algorithm, GOLD a genetic one while Glide performs an exhaustive search. The choice of these three protocols was made to evaluate the virtual library with three orthogonal search algorithms, to minimize the false-positive rate to which traditional docking-based virtual screenings are prone. At the end of each virtual screening, a total of about 13.6 M poses (50 per ligand) was obtained for each protocol. The choice to generate such a great number of poses for each ligand was taken in order not to rely on the scoring function ability to prioritize the best binding mode for each compound, since fragments can have multiple binding modes that are similar from an energetic and qualitative point of view and are therefore difficult to distinguish for scoring functions that are trained upon mature, lead-like, compounds.

To filter this huge amount of ligand conformations and retain only the most interesting compounds, we decided to exploit the structural knowledge provided by the 23 Ck1δ protein–ligand complexes deposited in the Protein Data Bank and create a pharmacophore filter. This pharmacophore model was built to retain those features which are vital for the interaction with the hinge region of the kinase since these features are the most commonly found across the structures. The pharmacophore included three features, two of them to guarantee the interaction with Leu85 (a hydrogen bond donor and a hydrogen bond acceptor) and the presence of and a feature for an aromatic ring also in the proximity of the hinge region.

The pharmacophore filter was then applied independently on each pose database generated by the three different docking protocols. Exploiting an approach known as consensus docking, the three libraries containing those ligand conformations that fit the pharmacophore model were merged, retaining only those found within each dataset. After this consensus filtering, only 840 docking poses were left.

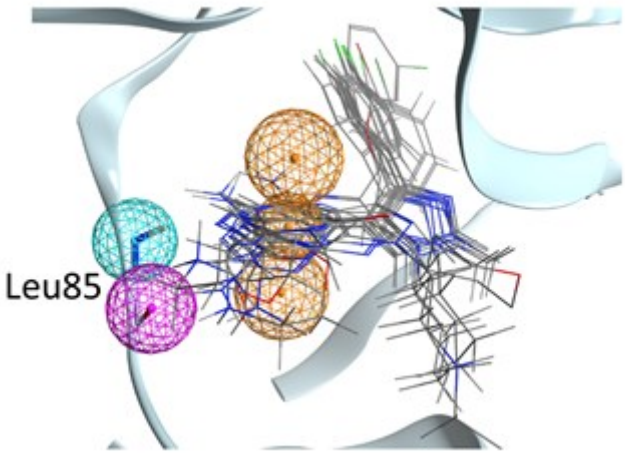


Figure 2. Representation of the pharmacophore model used in the present work. Some representative crystallographic ligands are displayed (not all for clarity). The Pharmacophore model is formed by an aromatic ring (the three orange spheres define the position and its orientation) and two hydrogen bonds with the backbone of Leu85 (an acceptor and one donor).

To further filter out those poses characterized by unstable binding modes, a post-docking molecular dynamics refinement was performed (three replicates, 10 ns each). The average Root Mean Squared Fluctuation of atomic positions (RMSF) across the three replicates was used as a cutoff to eliminate those poses characterized by conformational instability over time. After filtering out those ligand conformations with RMSF > 2 Å, 650 stable poses were maintained.

With the intent of prioritizing the most interesting compound for in vitro assays, each pose was carefully manually examined. After this visual inspection[46] step, 66 fragments were finally selected to be purchased and tested. The structure of all the 66 fragment compounds tested are reported in supplementary Table S1, while the pose of each of them resulted from the VS pipeline is reported in Video S1.

2.2. Enzymatic Assay Results

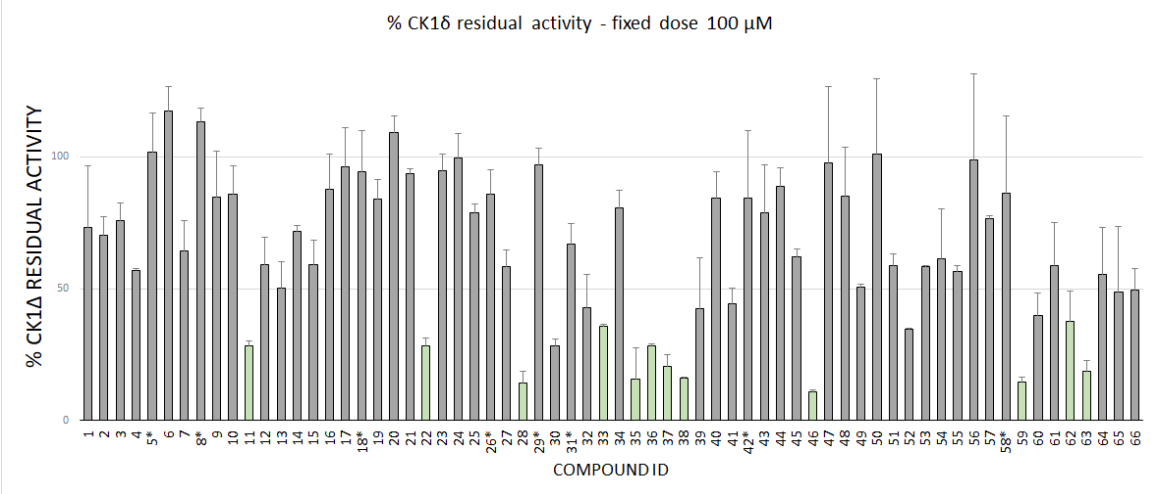


Figure 3. CK1δ residual activity at a concentration of 100 µM of the ligand under examination. Molecules marked with a star have been tested at 50 µM due to solubility issues.

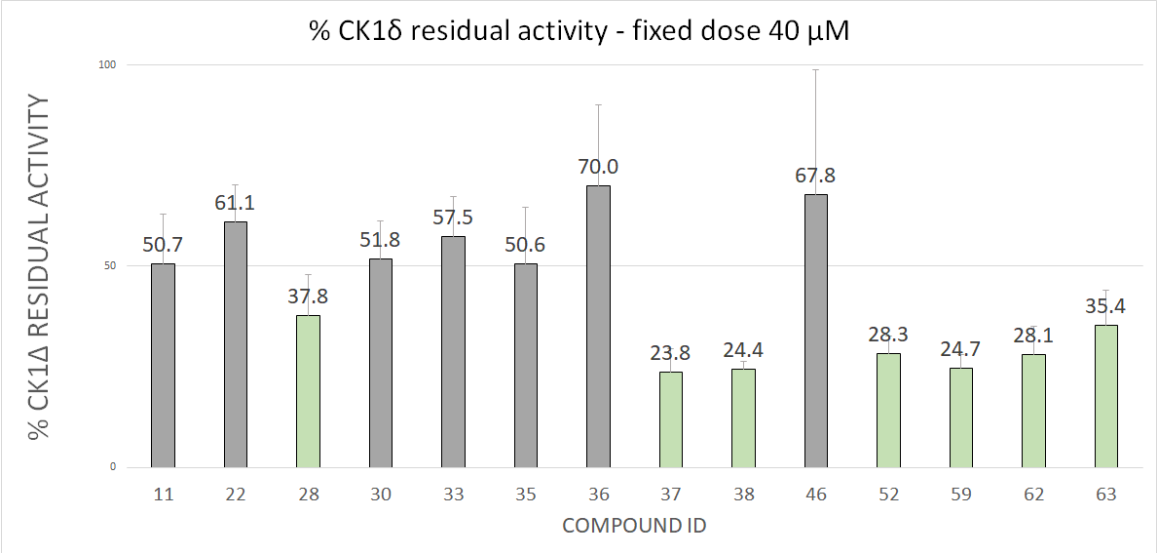


Figure 4. CK1δ Residual activity at a concentration of 40 µM of the ligands that showed a residual activity of less than 40% at 100 µM.

Fragments were tested against CK1δ using a luminescent-based assay. Compounds were evaluated at a fixed concentration of 100 µM (see Figure 3) and those that showed a kinase residual activity lower than 40% were tested also at a fixed concentration of 40 µM (see Figure 4).

IC_{50} values were calculated for compounds with a residual kinase activity lower than 40%. Compounds **37**, **38**, **52**, **59**,**62** and **63** showed IC_{50} values in the micromolar range of 12.71 µM (9.57–16.80), 20.49 µM (17.46–24.08), 13.50 µM (12.47–14.62), 13.92 µM (11.89–

16.29), 18.15 µM (16.78–19.64) and 24.86 µM (21.46–28.92), respectively. Remarkably, compound **28** shows a half-maximal inhibitory concentration of 3.31 µM (2.67–4.12). The IC_{50} curves for the seven hits are reported on SI. The value of IC50 is based on the average of three independent measurements.

2.3. Molecular Recognition Studies of the Most Promising Fragment

To shed light on the possible recognition mechanism of the most effective inhibitor, compound **28** (IC_{50} = 3.31 µM) was investigated by mean of Supervised Molecular Dynamics simulations (SuMD). The primary scope was to assess if the hypothesized bound state obtained by our computational protocol was also accessible by simulating the fragment association from the unbound state without any information about the ligand conformation. Since in our VS-pipeline the pharmacophoric filter plays a primary goal in defining the bound geometries, its validation by using a more articulated technique based on MD and in which the water molecules need to be displaced by the fragment to reach the hinge region would provide the reliability of the binding mode.
A complete recognition pathway of the length of 15 ns is reported in Video S2 (SI). Compound **28** showed three steps during the recognition, with two stable states (Figure 6A).

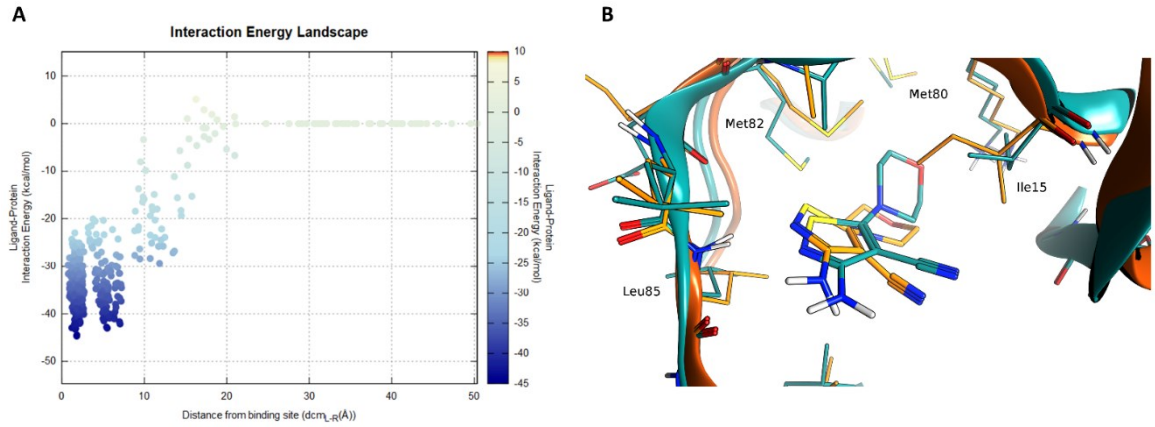


Figure 6. SuMD simulation of compound **28.** In panel (**A)** the interaction energy landscape is reported for the recognition trajectory displaying the ligand–protein interaction energy plotted against the distances between the protein–ligand center of mass. In panel (**B)**, the superposition of the VS-pose (cyan) for compound **28** against the lowest energy frame from the SuMD trajectory (orange).

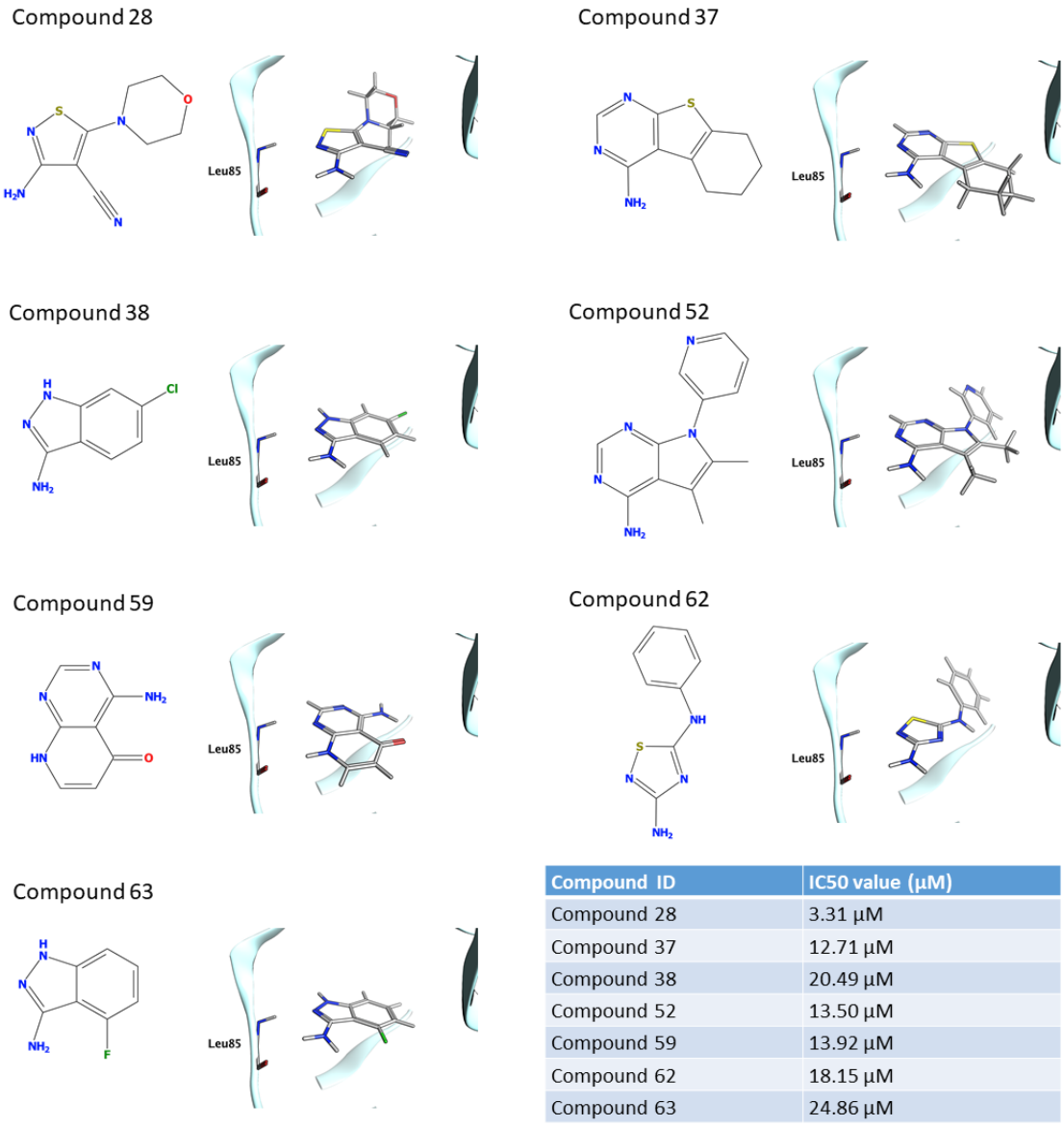
A pivotal role in the first phases (around 1 ns time mark) of the ligand recruitment within the binding site is played by Asp149, which acts as an electrostatic recruiter for the amino-

thiophene moiety of the ligand. By contrast, the vicinal residue Lys38 hampers the ligand entrance into the core portion of the binding site due to the electrostatic repulsion between the charged amino group of the amino acid side chain and the non-charged amino group of the ligand. The balance in attraction and repulsion between the flexible side chains of these two amino acids located at the boundary of the binding site is depicted also by the large energetic funnel shown in Figure 6A at around 10 Å with regard to the distance between the centers of mass of the binding site and the ligand (dcm_{L-R}). Afterwards, the binding pathway is characterized by two stable ligand conformations within the binding site. The first state (S1) occurred at a dcm_{L-R} distance of 4.5 Å, with the ligand interacting with the backbone of Leu85 through its amino-thiophene moiety and the morpholine moiety oriented towards the external part of the binding site (solvent-exposed), while the second one (S2) at a dcm_{L-R} distance of 1.5 Å is characterized by a bivalent hydrogen bond with Leu85 and the morpholine moiety of the ligand buried within the hydrophobic selectivity pocket defined by Met80, Met82, Ile23 and the alkyl portion of the Lys38 side chain. Although these two states are characterized by similar interaction energy values (according to the AMBER forcefield), their energetic funnels have different shapes: the final state (S2) shows a narrower profile than the S1 state, suggesting that the pharmacophore binding mode (S2) has a higher stability than S1. Furthermore, the final bound state nicely retraced the pose obtained with the VS pipeline, validating both the pharmacophore model used in this work and the binding mode proposed by molecular docking for this compound (Figure 6B)

3. Discussion

The seven fragments that were characterized by calculating the IC_{50} showed a noticeable chemical diversity including scaffolds spanning from one to three nitrogen-containing fused rings. The poses of the seven hits as obtained in the VS are reported in Figure 5. All the fragments logically share the common interaction pattern required by the pharmacophore filter. Interestingly, compounds **28**, **37**, **38**, **52**, **62**, and **63** showed a similar interaction scheme in which an aromatic amine moiety was able to establish a hydrogen bond with the carbonyl oxygen of the Leu85 backbone while a further hydrogen bond between the Leu85 backbone amide is guaranteed by aromatic nitrogen in ortho to

the amine group. Compounds **37**, **52**, and **59** share a conserved pyrimidine ring that is part of different fused systems. Compound **59** also has the pyrimidine ring in a different orientation: it restores the hydrogen bond donor by its fused pyridone ring. Compounds **38** and **63** present the same scaffold.



Compound ID	IC50 value (µM)
Compound 28	3.31 µM
Compound 37	12.71 µM
Compound 38	20.49 µM
Compound 52	13.50 µM
Compound 59	13.92 µM
Compound 62	18.15 µM
Compound 63	24.86 µM

Figure 5. The structure and binding mode for the seven compounds for which the IC_{50} value is reported. The value of IC_{50} is based on the average of three independent measurements

To assess the novelty of the identified fragments, a substructure search was performed against ChEMBL using the main ring recognized by the pharmacophore as a query; except for compounds **38** e **52**, which resulted in **34** and **20** already known CK1δ inhibitors, for all the remaining hits none known inhibitors were found sharing the principal ring. The 3-

amino-indazole scaffold of compound **38** was found in a multikinase inhibitor (CHEMBL1999931) with a Ki of 316.23 nM[47]. For compound **52** a couple of ligands with low uM activity were found; in particular CHEMBL2000114 with a Ki of 1 uM arose from the same kinome scan from Abbott Labs[47].

Additionally, compound GSK1838705A showed the same scaffold of **52**, in this case the Ki reported is 3.5 uM but it is a residual activity since the compound is a potent inhibitor of ALK kinase (IC_{50} = 0.5 nM)[48].

.4. Materials and Methods

4.1. Molecular Modelling and Docking

The virtual library used in this work was obtained through the merging of different libraries of commercially available compounds designed for FBDD. The vendors are Asinex, Chembridge, Enamine, Life Chemicals, Maybridge, Otava, Timtec, Vitas. The total number of fragments in the merged library is about 272,000 virtual compounds.

The merged library was prepared to be suitable for the Docking-Based Virtual Screening. This preparation consists of the following steps: the tautomeric state enumeration for each compound and determination of the most probable tautomer (for each molecule at the three most tautomeric states was retained), the most probable ionization state at pH 7.4 calculation, the atomic partial charge calculation (using MMFF94 force field), the 3D coordinates generation. All these steps were performed using QUACPAC of the Openeye suite[49] except for the 3D coordinated generation for which Corina Classic was used[50].

The protein used both for Docking and for MD simulation was prepared using MOE. The preparation consists of the removal of the crystallographic water molecules and other solvent molecules together with ions and the ligand. The correct protonation state for each residue at pH 7.4 was calculated with the Protonate3D tool of MOE.

For the Consensus Docking strategy, three different Molecular Docking protocols were used. To make the results more robust, the three docking protocols chosen rely on search algorithms of different types. The Molecular Docking Protocols are PLANTS[51] which is based on an Ant Colony Optimization algorithm, GOLD[52,53] which employs a genetic algorithm, and Glide[54,55] which use a systematic searching approach. The Scoring Functions adopted are CHEMPLP for PLANTS, ChemScore for GOLD, and Glide SP for Glide.

For each fragment 50 poses were generated using each Docking Protocol even if the termination criteria and the nature of the algorithms did not always provide 50 poses, in particular for Glide SP.

Similarity and substructure searches were performed with MOE using the ChEMBL29 database.

4.2. Pharmacophore Modeling

Each ensemble of poses (one for each docking protocol) was then filtered using a pharmacophore model. This pharmacophore model was calculated using MOE: all the holo crystal structures available on the PDB for human CK1δ were superposed and the common features of each ligand were analyzed. The list of complexes included 23 complexes with PDB ID: 3UYT, 3UZP, 4HGT, 4HNF, 4KB8, 4KBA, 4KBC, 4KBK, 4TN6, 4TW9, 4TWC, 5IH5, 5IH6, 5MQV, 5OKT, 5W4W, 6F1W, 6F26, 6GZM, 6HMP, 6HMR, 6RCG, 6RCH.

Since the ligands present in the crystal structures are drug-like molecules, it is difficult that a fragment can comply with all the common features observed in the crystal structures. For this reason (and because as stated above the first fragment identified in an FBDD process of a kinase inhibitor is a hinge binding fragment) the pharmacophore model was built using only the features involved in the interaction with the hinge region of the kinase. The model included three features: one hydrogen bond donor and one hydrogen bond acceptor to guarantee the interaction with the backbone of Leu85 (Figure 2). The last feature represents an aromatic ring also in the proximity of the hinge region. Only the molecule that has passed the Pharmacophore filtering for each protocol was retained (*consensus*).

4.3. Molecular Dynamics

The molecules retained after the consensus filtering were subjected to a post-docking refinement. The docking pose used in this step is the one obtained from Glide. All the simulations were carried out using ACEMD3[56] with ff14SB as force field[57], the system preparation was conducted with MOE concerning protein preparation and with the use of AmberTools14 for the simulation box preparation.

For each complex, a simulation box was prepared: the protein was immersed in an explicit TIP3P[58] solvent box, with an ionic strength of 0.154 M obtained using Na^+/Cl^-. The protein is 15 Å away from the border of the box.

Using the conjugate gradient method, the system energy was minimized for 500 steps; after this minimization the system was equilibrated in two stages. The first equilibration consists of 1 ns of NVT simulation with harmonic positional constraints of 1 kcal mol^{-1} $Å^{-2}$ on the protein. In the second equilibration step, which consists in this case of 1 ns of NPT simulation, the constraints of 1 kcal mol^{-1} $Å^{-2}$ were applied only on the α carbons of the protein. After the equilibration for each protein–pose complex, three NVT trajectories of 10 ns were produced. The average RMSF of the ligand during these three replicas was calculated and if this value was greater than 2 Å the molecule was discarded.

A Supervised Molecular Dynamics[59,41] simulation was performed to gain some insights into the binding process of the most potent fragment (Compound **28**). SuMD is an MD-based method developed to investigate molecular binding events without energetic biases. The algorithm is based on the supervision of the ligand–protein binding site center of mass distance during a classical short MD simulation. At the end of each small simulation (SuMD step), this distance is measured: if it has shortened during the SuMD step, the simulation continues with another SuMD step, otherwise, it is stopped, and the simulation restarts from the previous set of coordinates. The fragment was placed 30 Å away from the protein. Each SuMD step was set to 300 ps.

4.4. Enzymatic Assay

Compounds were evaluated towards CK1δ (aa 1-294, Merck Millipore) with the KinaseGlo® luminescence assay (Promega) following procedures reported in the literature[22]. In detail, luminescent assays were performed in white 96-well plates, using the following buffer: 50 mM HEPES (pH 7.5),1 mM EDTA, 1 mM EGTA, and 15 mM $MgCl_2$. Compound PF-670462 (IC50 = 14 nM) was used as a positive control for CK1δ[60] and DMSO/buffer solution was used as a negative control. In a typical assay, 10 µL of inhibitor solution (dissolved in DMSO at 10 mM concentration and diluted in assay buffer to the desired concentration) and 10 µL (16 ng) of enzyme solution were added to each well, followed by 20 µL of assay buffer containing 0.1% casein substrate and 4 µM ATP. The final DMSO concentration in the reaction mixture did not exceed 1%.

After 60 min of incubation at 30 °C, the enzymatic reactions were stopped with 40 µL of KinaseGlo® reagent (Promega). The luminescence signal (relative light unit, RLU) was recorded after 10 min at 25 °C using Tecan Infinite M100. Fixed-dose experiments were

performed at 100 µM and for more potent compounds also at 40 µM. Two independent experiments were performed in duplicate and the corresponding residual activity of CK1δ was obtained. Data were analyzed using Excel and reported as the mean of the two experiments with standard deviation. For IC_{50} determination ten different inhibitor concentrations ranging from 100 to 0.026 µM were used and each point was assessed in duplicate. IC_{50} values are the mean of three independent experiments and 95% confidence limits were also reported. Data were analyzed using GraphPad Prism software (version 8.0).

5. Conclusions

In the present work to find new potential CK1δ inhibitors, we elaborated a computational workflow for the identification of candidate hinge binding fragments. This workflow consists of the generation of a large number of poses for each compound of a virtual library of commercially available fragments using three different Docking protocols. These poses were filtered using a pharmacophore model and only the fragment for which each docking protocol was able to produce a pose that fits the model was retained (consensus docking). In the next step each protein-fragment complex that passed the previous filter was subjected to an MD-driven post-docking refinement to inspect the geometric stability of the pose. Finally, some fragments were manually selected among the group that demonstrated a good performance in the post-docking refinement; to validate the method these fragments were tested using an enzymatic assay test to assess the CK1δ residual activity, and for the most promising candidates, the IC_{50} value was determined, with a value in the low micromolar range. Five of the seven fragments showed novel scaffolds for CK1δ, confirming that the proposed pipeline could be particularly useful to identify novel structures.

References

1. Knippschild, U. *et al.* The CK1 family: Contribution to cellular stress response and its role in carcinogenesis. *Front. Oncol.* **4 MAY**, 1–33 (2014).

2. Meggio, F., Perich, J. W., Reynolds, E. C. & Pinna, L. A. A synthetic β-casein phosphopeptide and analogues as model substrates for casein kinase-1, a ubiquitous, phosphate directed protein kinase. *FEBS Lett.* (1991) doi:10.1016/0014-5793(91)80614-9.

3. Pulgar, V. *et al.* Optimal sequences for non-phosphate-directed phoshorylation by protein kinase CK1 (casein kinase-1) - A re-evaluation. *Eur. J. Biochem.* (1999) doi:10.1046/j.1432-1327.1999.00195.x.

4. MARIN, O., MEGGIO, F., SARNO, S., ANDRETTA, M. & PINNA, L. A. Phosphorylation of synthetic fragments of inhibitor-2 of protein phosphatase-1 by casein kinase-1 and -2: Evidence that phosphorylated residues are not strictly required for efficient targeting by casein kinase-1. *Eur. J. Biochem.* (1994) doi:10.1111/j.1432-1033.1994.tb19037.x.

5. Bischof, J. *et al.* CK1δ Kinase Activity Is Modulated by Chk1-Mediated Phosphorylation. *PLoS One* (2013) doi:10.1371/journal.pone.0068803.

6. Graves, P. R. & Roach, P. J. Role of COOH-terminal phosphorylation in the regulation of casein kinase Iδ. *J. Biol. Chem.* (1995) doi:10.1074/jbc.270.37.21689.

7. Milne, D. M., Looby, P. & Meek, D. W. Catalytic activity of protein kinase CK1δ (casein kinase 1 δ) is essential for its normal subcellular localization. *Exp. Cell Res.* (2001) doi:10.1006/excr.2000.5100.

8. Xu, P. *et al. Structure, regulation, and (patho-)physiological functions of the stress-induced protein kinase CK1 delta (CSNK1D)*. *Gene* vol. 715 (Elsevier B.V, 2019).

9. Longenecker, K. L., Roach, P. J. & Hurley, T. D. Crystallographic studies of casein kinasc I δ: Toward a structural understanding of auto-inhibition. *Acta Crystallogr. Sect. D Biol. Crystallogr.* (1998) doi:10.1107/s0907444997011724.

10. Hirner, H. *et al.* Impaired CK1 delta activity attenuates SV40-induced cellular transformation in vitro and mouse mammary carcinogenesis in Vivo. *PLoS One* (2012) doi:10.1371/journal.pone.0029709.

11. Perez, D. I., Gil, C. & Martinez, A. Protein kinases CK1 and CK2 as new targets for neurodegenerative diseases. *Medicinal Research Reviews* (2011) doi:10.1002/med.20207.

12. Ghoshal, N. *et al.* A new molecular link between the fibrillar and granulovacuolar lesions of Alzheimer's disease. *Am. J. Pathol.* (1999) doi:10.1016/S0002-9440(10)65219-4.

13. Yasojima, K., Kuret, J., Demaggio, A. J., McGeer, E. & McGeer, P. L. Casein kinase 1 delta mRNA is upregulated in Alzheimer disease brain. *Brain Res.* (2000) doi:10.1016/S0006-8993(00)02200-9.

14. Kuret, J. *et al.* Casein kinase 1 is tightly associated with paired-helical filaments isolated from Alzheimer's disease brain. *J. Neurochem.* (1997) doi:10.1046/j.1471-4159.1997.69062506.x.

15. Schwab, C. *et al.* Casein kinase 1 delta is associated with pathological accumulation of tau in several neurodegenerative diseases. *Neurobiol. Aging* (2000) doi:10.1016/S0197-4580(00)00110-X.

16. Li, G., Yin, H. & Kuret, J. Casein Kinase 1 Delta Phosphorylates Tau and Disrupts Its Binding to Microtubules. *J. Biol. Chem.* (2004) doi:10.1074/jbc.M314116200.

17. Chauhan, A., Chauhan, V. P. S., Murakami, N., Brockerhoff, H. & Wisniewski, H. M. Amyloid β-protein stimulates casein kinase I and casein kinase II activities. *Brain Res.* (1993) doi:10.1016/0006-8993(93)90479-7.

18. Flajolet, M. *et al.* Regulation of Alzheimer's disease amyloid-β formation by casein kinase I. *Proc. Natl. Acad. Sci. U. S. A.* (2007) doi:10.1073/pnas.0611236104.

19. Höttecke, N. *et al.* Inhibition of γ-secretase by the CK1 inhibitor IC261 does not depend on CK1δ. *Bioorganic Med. Chem. Lett.* (2010) doi:10.1016/j.bmcl.2010.02.110.

20. Kosten, J. *et al.* Efficient modification of alpha-synuclein serine 129 by protein kinase CK1 requires phosphorylation of tyrosine 125 as a priming event. *ACS Chem. Neurosci.* (2014) doi:10.1021/cn5002254.

21. Nonaka, T. *et al.* Phosphorylation of TAR DNA-binding protein of 43 kDa (TDP-43) by truncated casein kinase 1δ triggers mislocalization and accumulation of TDP-43. *J. Biol. Chem.* (2016) doi:10.1074/jbc.M115.695379.

22. Salado, I. G. *et al.* Protein kinase CK-1 inhibitors as new potential drugs for amyotrophic lateral sclerosis. *J. Med. Chem.* (2014) doi:10.1021/jm500065f.

23. Hall, R. J., Mortenson, P. N. & Murray, C. W. Efficient exploration of chemical space by fragment-based screening. *Prog. Biophys. Mol. Biol.* (2014) doi:10.1016/j.pbiomolbio.2014.09.007.

24. Flaherty, K. T., Yasothan, U. & Kirkpatrick, P. Vemurafenib. *Nat. Rev. Drug Discov.* (2011) doi:10.1038/nrd3579.

25. Bollag, G. *et al.* Vemurafenib: The first drug approved for BRAF-mutant cancer. *Nature Reviews Drug Discovery* (2012) doi:10.1038/nrd3847.

26. Romero, D. Initial results with asciminib in CML. *Nature Reviews Clinical Oncology* (2020) doi:10.1038/s41571-019-0324-z.

27. Schoepfer, J. *et al.* Discovery of Asciminib (ABL001), an Allosteric Inhibitor of the Tyrosine Kinase Activity of BCR-ABL1. *J. Med. Chem.* (2018) doi:10.1021/acs.jmedchem.8b01040.

28. Erlanson, D. A., Fesik, S. W., Hubbard, R. E., Jahnke, W. & Jhoti, H. Twenty years on: The impact of fragments on drug discovery. *Nature Reviews Drug Discovery* (2016) doi:10.1038/nrd.2016.109.

29. Miranker, A. & Karplus, M. Functionality maps of binding sites: A multiple copy simultaneous search method. *Proteins Struct. Funct. Bioinforma.* (1991) doi:10.1002/prot.340110104.

30. Clark, M., Meshkat, S., Talbot, G. T., Carnevali, P. & Wiseman, J. S. Fragment-based computation of binding free energies by systematic sampling. *J. Chem. Inf. Model.* (2009) doi:10.1021/ci900132r.

31. Böhm, H. J. The computer program LUDI: A new method for the de novo design of enzyme inhibitors. *J. Comput. Aided. Mol. Des.* (1992) doi:10.1007/BF00124387.

32. Eisen, M. B., Wiley, D. C., Karplus, M. & Hubbard, R. E. HOOK: A program for finding novel molecular architectures that satisfy the chemical and steric requirements of a macromolecule binding site. *Proteins Struct. Funct. Bioinforma.* (1994) doi:10.1002/prot.340190305.

33. Lauri, G. & Bartlett, P. A. CAVEAT: A program to facilitate the design of organic molecules. *J. Comput. Aided. Mol. Des.* (1994) doi:10.1007/BF00124349.

34. Maass, P., Schulz-Gasch, T., Stahl, M. & Rarey, M. Recore: A fast and versatile method for scaffold hopping based on small molecule crystal structure conformations. *J. Chem. Inf. Model.* (2007) doi:10.1021/ci060094h.

35. Schrödinger Release 2020-4: Maestro, Schrödinger, LLC, New York, NY, 2020.

36. Chemical Computing Group ULC, Molecular Operating Environment (MOE), 2019.01. 1010 Sherbrooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, 2019.

37. Alonso, H., Bliznyuk, A. A. & Gready, J. E. Combining docking and molecular dynamic simulations in drug design. *Medicinal Research Reviews* (2006) doi:10.1002/med.20067.

38. Gill, S. C. *et al.* Binding Modes of Ligands Using Enhanced Sampling (BLUES): Rapid Decorrelation of Ligand Binding Modes via Nonequilibrium Candidate Monte Carlo. *J. Phys. Chem. B* (2018) doi:10.1021/acs.jpcb.7b11820.

39. Lim, N. M., Osato, M., Warren, G. L. & Mobley, D. L. Fragment Pose Prediction Using Non-equilibrium Candidate Monte Carlo and Molecular Dynamics Simulations. *J. Chem. Theory Comput.* (2020) doi:10.1021/acs.jctc.9b01096.

40. Linker, S. M., Magarkar, A., Köfinger, J., Hummer, G. & Seeliger, D. Fragment Binding Pose Predictions Using Unbiased Simulations and Markov-State Models. *J. Chem. Theory Comput.* (2019) doi:10.1021/acs.jctc.9b00069.

41. Sabbadin, D. & Moro, S. Supervised molecular dynamics (SuMD) as a helpful tool to depict GPCR-ligand recognition pathway in a nanosecond time scale. *J. Chem. Inf. Model.* (2014) doi:10.1021/ci400766b.

42. Ferrari, F. *et al.* HT-SuMD: making molecular dynamics simulations suitable for fragment-based screening. A comparative study with NMR. *J. Enzyme Inhib. Med. Chem.* (2021) doi:10.1080/14756366.2020.1838499.

43. Chaput, L. & Mouawad, L. Efficient conformational sampling and weak scoring in docking programs? Strategy of the wisdom of crowds. *J. Cheminform.* (2017) doi:10.1186/s13321-017-0227-x.

44. de Souza Neto, L. R. *et al.* In silico Strategies to Support Fragment-to-Lead Optimization in Drug Discovery. *Frontiers in Chemistry* (2020) doi:10.3389/fchem.2020.00093.

45. Houston, D. R. & Walkinshaw, M. D. Consensus docking: Improving the reliability of docking in a virtual screening context. *J. Chem. Inf. Model.* (2013) doi:10.1021/ci300399w.

46. Fischer, A., Smieško, M., Sellner, M. & Lill, M. A. Decision Making in Structure-Based Drug Discovery: Visual Inspection of Docking Results. *J. Med. Chem.* **64**, 2489–2500 (2021).

47. Metz, J. T. *et al.* Navigating the kinome. *Nat. Chem. Biol.* **7**, 200–202 (2011).

48. P, S. *et al.* GSK1838705A inhibits the insulin-like growth factor-1 receptor and anaplastic lymphoma kinase and shows antitumor activity in experimental models of human cancers. *Mol. Cancer Ther.* **8**, 2811–2820 (2009).

49. QUACPAC 2.1.1.0: OpenEye Scientific Software, Santa Fe, NM. http://www.eyesopen.com.

50. Sadowski, J., Gasteiger, J. & Klebe, G. Comparison of Automatic Three-Dimensional Model Builders Using 639 X-ray Structures. *J. Chem. Inf. Comput. Sci.* (1994) doi:10.1021/ci00020a039.

51. Korb, O., Stützle, T. & Exner, T. E. PLANTS: Application of ant colony optimization to structure-based drug design. in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (2006). doi:10.1007/11839088_22.

52. Jones, G., Willett, P., Glen, R. C., Leach, A. R. & Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* (1997) doi:10.1006/jmbi.1996.0897.

53. Verdonk, M. L., Cole, J. C., Hartshorn, M. J., Murray, C. W. & Taylor, R. D. Improved protein-ligand docking using GOLD. *Proteins Struct. Funct. Genet.* (2003) doi:10.1002/prot.10465.

54. Halgren, T. A. *et al.* Glide: A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening. *J. Med. Chem.* (2004) doi:10.1021/jm030644s.

55. Sándor, M., Kiss, R. & Keseru, G. M. Virtual fragment docking by glide: A validation study on 190 protein-fragment complexes. *J. Chem. Inf. Model.* (2010) doi:10.1021/ci1000407.

56. Harvey, M. J., Giupponi, G. & De Fabritiis, G. ACEMD: Accelerating biomolecular dynamics in the microsecond time scale. *J. Chem. Theory Comput.* (2009) doi:10.1021/ct9000685.

57. Maier, J. A. *et al.* ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* (2015) doi:10.1021/acs.jctc.5b00255.

58. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* (1983) doi:10.1063/1.445869.

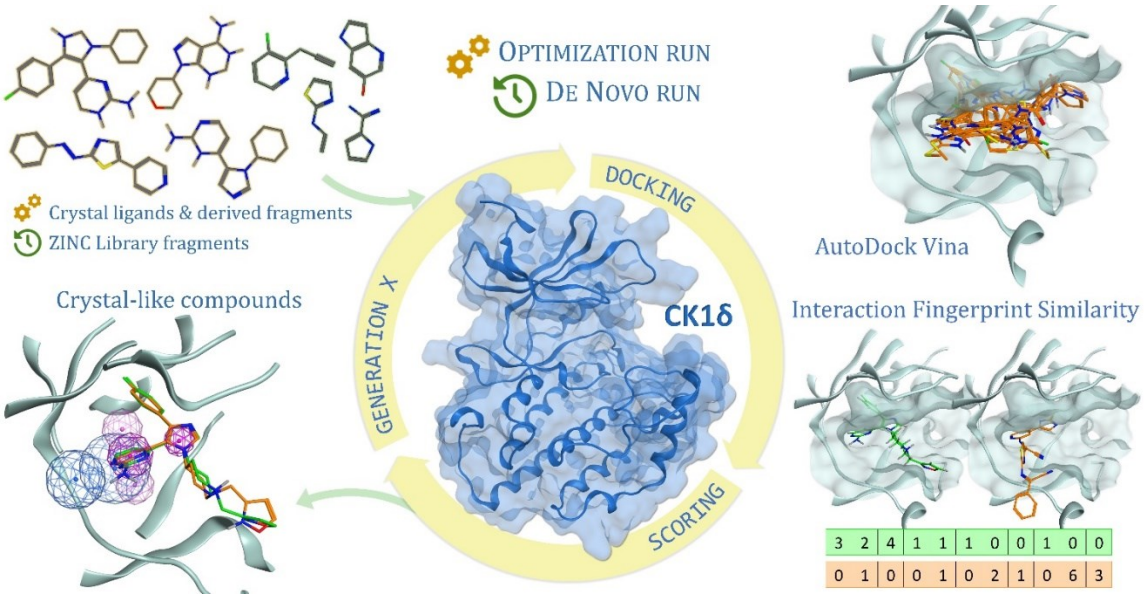
59. Cuzzolin, A. *et al.* Deciphering the Complexity of Ligand-Protein Recognition Pathways Using Supervised Molecular Dynamics (SuMD) Simulations. *J. Chem. Inf. Model.* **56**, 687–705 (2016).

60. Badura, L. *et al.* An inhibitor of casein kinase Iε induces phase delays in circadian rhythms under free-running and entrained conditions. *J. Pharmacol. Exp. Ther.* (2007) doi:10.1124/jpet.107.122846.

Implementing a scoring function based on interaction fingerprint for Autogrow4: Protein Kinase CK1δ as a case study

Matteo Pavan, Silvia Menin, Davide Bassani, Mattia Sturlese, Stefano Moro

Pavan, M., Menin, S., Bassani, D., Sturlese, M. & Moro, S. Implementing a Scoring Function Based on Interaction Fingerprint for Autogrow4: Protein Kinase CK1δ as a Case Study. *Front Mol Biosci* **0**, 629 (2022).



Abstract

In the last twenty years, Fragment-Based Drug Discovery (FBDD) has become a popular and consolidated approach within the drug-discovery pipeline, due to its ability to bring several drug candidates to clinical trials, some of them even being approved and introduced to the market.

A class of targets that have proven to be particularly suitable for this method is represented by kinases, as demonstrated by the approval of BRAF inhibitor Vemurafenib. Within this wide and diverse set of proteins, protein kinase CK1δ is a particularly interesting target for the treatment of several widespread neurodegenerative diseases such as Alzheimer's disease, Parkinson's disease, and amyotrophic lateral sclerosis.

Computational methodologies such as molecular docking are already routinely and successfully applied in fragment-based drug discovery campaigns alongside experimental techniques, both in the hit-discovery and in the hit-optimization stage. Concerning this,

the open-source software Autogrow, developed by the Durrant lab, is a semi-automated computational protocol that exploits a combination between a genetic algorithm and a molecular docking software for de-novo drug design and lead optimization.

In the present work, we present and discuss a modified version of the Autogrow code that implements a custom scoring function based on the similarity between the interaction fingerprint of investigated compounds to a crystal reference. To validate its performance, we performed both a *denovo* and a lead-optimization run (as described in the original publication), evaluating the ability of our fingerprint-based protocol to generate compounds similar to known CK1δ inhibitors based on both the predicted binding mode and electrostatic and shape similarity in comparison with the standard Autogrow protocol.

Introduction

Protein kinase CK1δ is a Ser/Thr protein kinase belonging to the Casein Kinase 1 family. In mammals, 7 distinct genes encoding for Casein kinase proteins are present, each producing a different isoform (α, β, γ1, γ2, γ3, δ, and ε)[1]. CK1 family proteins use exclusively ATP as a phosphate source for their kinase activity, which is carried out by the protein in its monomeric form. Each isoform is constitutionally active and does not require the presence of a cofactor to exert its activity[2].

From a biological function point of view, the members of this family have been historically related to different physiological mechanisms, such as cell replication[3], DNA repair[4], and circadian rhythm[5].

From a structural perspective, the members of the CK1 family are characterized by the typical bilobed structure of the globular Ser-Thr kinase proteins, with the N-term lobe consisting mainly of β-sheets and a larger C-term lobe, constituted primarily of α-helices. The two domains are connected by a protein region named the "hinge region", which forms a highly conserved pocket for ATP binding[2].

As for other members of the CK1 family, CK1δ recognizes the canonical phospho-primed structural motif *pSer/pThr-X_{1-2}-Ser/Thr*, where X stands for any amino acid and pSer/pThr represents the phospho-primed residue[6]. The CK1 kinases are also able to recognize non-phosphorylated sequences, as far as they contain strongly acidic residues (Asp or Glu) that can make up for the absence of the phosphorylated residue[3]. The structural motif that can be recognized by the CK1 proteins is widespread in many cellular proteins and, because of this, over 140 substrates have been reported both *in vitro* and *in vivo*[2], underlining the pleiotropic character of this protein family. Due to the great variability of its substrates, CK1δ is involved in many cellular pathways, among which the main ones are the Wnt-pathway, the Hippo pathway, the p53 regulation pathway, and the Hedgehog pathway[3].

The endogenous regulation of CK1δ, on the other hand, can be carried out through various mechanisms, including autophosphorylation or phosphorylation by other protein kinases[7,8], interactions with other protein and/or cellular components, and subcellular

sequestration[3,9]. In addition, homodimerization excludes ATP from the binding site, thus inhibiting kinase activity[10,11].

In recent years, several studies have highlighted the importance of CK1δ in neurodegenerative diseases, particularly tauopathies such as Alzheimer's disease (AD), Parkinson's disease (PD), and amyotrophic lateral sclerosis (ALS)[12]. In addition to having unknown etiology, these illnesses are all characterized by loss of neuronal function, with neurotransmitter deficiency, misfolding, and protein aggregation[13]. Clinical symptoms are manifested differently, depending on the neuronal area involved[14].

AD is a progressive neurodegenerative disorder that mainly involves the neurons of the hippocampus[15]. On the extracellular side, the main marker of the disease is represented by the accumulation of β-amyloid peptides, produced by β-secretase 1 and γ-secretase enzymes, which leads to neuronal death[16]. Meanwhile, on the intracellular part, the illness presents lesions related to both cytoplasmic accumulations of vacuoles with abnormal dimensions and dense granular content and the assembly of fibrils and filaments within the neuronal body. These types of lesions are both characterized by the accumulation of hyperphosphorylated Tau protein in the filaments, but also within the vacuoles[17].

The correlation between CK1δ activity and tau protein aggregates in various neurodegenerative diseases has been confirmed by co-immunoprecipitation studies, which highlight that the presence of CK1δ is associated with hyperphosphorylated tau aggregates[18,19]. CK1δ phosphorylates tau protein at the Ser202 / Thr205 and Ser369 / Ser404 residues *in vitro*[12,20]. The phosphorylation sites are the same as those involved in binding with tubulin, highlighting the key role of kinase in the pathogenesis of AD[18]. It is not clear whether the hyperactivity of CK1δ is due to an over-transcription of its gene, altered protein turnover, or both causes, but it has been observed that the concentration of the protein CK1δ in an AD-affected hippocampus is 30 times higher than normal[17].

In PD, on the other hand, the pathology is characterized by the accumulation of Lewy bodies, consisting of aggregates of α-synuclein hyperphosphorylated by CK1δ at the level of Ser129 residues[21]. This process determines a massive loss of neuronal function at the substantia nigra level[22].

CK1δ also plays a key role in Amyotrophic Lateral Sclerosis (ALS), a neurodegenerative disorder in which intracellular inclusions of TDP-43 (TAR DNA-binding Protein) are found in the frontotemporal lobe. It was established that TDP-43 can be phosphorylated by CK1δ at 29 different sites[23].

These pathologies are all characterized by the absence of effective pharmacological therapy: in fact, there are no EMA-approved drugs on the market that can solve, and therefore cure, these diseases, but there are only palliative therapies for the temporary improvement of the patient's quality of life, thus resulting in a high social cost[24]. For these reasons, CK1δ appears as an interesting therapeutical target in the field of neurodegeneration, as witnessed by the increasing interest in the research for inhibitory candidates for this protein during the last 15 years.

Concerning the identification of novel kinase inhibitors, an approach that has proven to be particularly successful is the so-called Fragment-Based Drug Discovery (FBDD), as demonstrated by the approval of the BRAF inhibitor Vemurafenib[25] (employed in the treatment of metastatic melanoma) and by several other kinase inhibitors which are at various stage of clinical trials[26,27].

This approach revolves around the exploitation of "fragments", i.e., compounds that respect the "Rule of Three" (molecular weight < 300, number of hydrogen bond donor/acceptor <= 3, logP<=3), as a starting point for the rational development of novel mature, drug-like, active molecules[28,29].The main reason for the success of FBDD is the ability to sample a larger portion of the chemical space compared to the one occupied by drug-like molecules, thus increasing the success rate in finding novel scaffolds for targets of interest[30].

This methodology heavily relies on very sensitive biophysical methods such as X-ray crystallography (XRC), nuclear magnetic resonance (NMR), or surface plasmon resonance (SPR), to perform large screening campaigns on libraries composed of molecules with low molecular weight and high solubility, to find hit compounds[31,32]. These hit fragments have usually a low affinity for the target, ranging from low mM to high µM (hence the need for very sensitive screening techniques), but usually have a higher binding efficiency

compared to traditional drug-like molecules, being able to establish high-quality interaction with the target[33]. Fragment hits can then be easily combined (either through a linking or a merging process) or chemically modified (growing) to increase their affinity for the target, allowing for the development of potent and selective active compounds[34].

Alongside the aforementioned experimental techniques, in the last decade, a prominent role in FBDD campaigns has been played by computer-aided drug discovery (CADD) techniques such as molecular docking or molecular dynamics[35]. These computational approaches have been routinely and successfully applied for performing large screening on virtual fragment libraries, for the characterization of the fragment interaction mode with the target and to aid the fragment-to-lead optimization in a less time-consuming, more rational, and more efficient way. Some examples of software developed specifically designed for FBDD are LUDI[36], HOOK[37], CAVEAT[38], RECORE[39]. Moreover, commercial drug discovery suites such as Schrödinger, MOE, and OpenEye have implemented several tools related to the fragment optimization process.

Among the plethora of software available for FBDD, the open-source software Autogrow, developed by the Durrant lab, is particularly interesting. As thoroughly described in the work of Spiegel et al.[40], the open-source software Autogrow is a Python written code that combines a genetic algorithm with docking calculation based on the Vina[41] docking software to perform a semi-automatized process for both de-novo drug design and lead optimization. The latest release of the Autogrow (version 4.0.3, the one used in this work), was developed with the idea of making the codebase modular thus allowing the third-party implementation of different conversion scripts, molecular docking programs, scoring functions, and reaction libraries, to better suit the need of different research groups.

A recent scientific work published by our laboratory led to the identification of seven novel fragment compounds that bind the hinge region of CK1δ with a low-micromolar IC_{50}[42]. Attracted by the idea of exploiting a semi-automatized computational protocol for the optimization of our newly discovered fragment compounds, we decided to investigate if this protocol would be suitable for our needs. Since it is notorious that molecular docking programs are usually very efficient and optimized with regards to the

conformational search, but are usually lacking in the scoring phase[43,44] (especially for molecules like fragments that deviates from the drug-like chemical space on which these scoring functions have been trained[45,46]), we decided to investigate if the implementation of a different scoring protocol based on protein-ligand interaction fingerprint would improve the performance of the Autogrow protocol, concerning the ability of the program to generate compounds similar to known inhibitors based on their interaction scheme and electrostatic and shape similarity.

Materials and Methods

2.1 Hardware Overview

Each general molecular modeling operation has been performed on a Linux Workstation equipped with an 8 core Intel Xeon® CPU E5-1620 CPU. For more intensive calculations, such as the Autogrow runs, a 64 cores AMD Opteron™ Processor 6376 CPU cluster was exploited. Both the workstation and the cluster run Ubuntu 16.04 as their operating system.

2.2 Structures Preparation

In the case of protein kinase CK1δ, 23 protein-ligand complexes between the protein and small drug-like molecules are available in the Protein Data Bank[47](PDB ID; 3UYT, 3UZP, 4HGT, 4HNF, 4KB8, 4KBA, 4KBC, 4KBK, 4TN6, 4TW9, 4TWC, 5IH5,5IH6, 5MQV, 5OKT, 5W4W, 6F1W, 6F26, 6GZM, 6HMP, 6HMR, 6RCG, 6RCH). In the context of this work, the crystals with codes 6RU6, 6RU7, and 6RU8 were not considered in the study because they contain the natural substrate adenosine-5'-diphosphate. One of the structures (PDB ID: 4KB8[48]) is composed of two different CK1δ-ligand complexes. For this reason, the system has been separated into two different entries (namely 4KB8-A and 4KB8-B). Because of this, the total number of complexes considered in our study is 24.

Each of the mentioned complexes has been downloaded and properly prepared for subsequent computational analysis with the “Structure Preparation” tool implemented in the Molecular Operating Environment (MOE)[49] 2019.01 suite. The missing hydrogen atoms were appropriately added with the MOE “Protonate 3D” program (setting the pH for the protonation at a value of 7.4) and were then energetically minimized according to

the AMBER10: EHT[50] force field implemented in MOE. After the preparation phase, the protein-ligand complexes have been properly aligned and superposed with the MOE dedicated tool, to make the binding site coordinates coherent among the different crystallographic structures. These complexes were saved and used at a later stage for the generation of the pharmacophore model (see section 2.4).

Afterward, each ligand has been separated from its respective protein. All the small molecules were collected in a database and prepared for docking calculations exploiting several packages from the QUACPAC OpenEye[51] suite. For each molecule the most probable tautomeric state was selected with the "tautomers" program, the three-dimensional coordinates were rebuilt using the "Omega" tool, the partial charges were attributed with the "MolCharge" program according to the MMFF94 force field, and finally, the dominant protonation state at pH 7.4 was determined by the "FixPka" tool.

2.3 Cross-docking

Each of the aforementioned 24 CK1δ crystallographic ligands, prepared as described in section 2.2, was docked inside each of the correspondent 24 CK1δ protein structures exploiting two different molecular docking pieces of software, namely GOLD[52] (based on a genetic algorithm, developed and distributed with a commercial license from CCDC) and PLANTS[53] (an Ant-Colony-Optimization docking algorithm, developed by the University of Tübingen and free-for-use for academics).

This approach was chosen to follow the principles of "consensus docking"[54], which is based on the fact that data obtained combining results coming from docking programs that operate in an orthogonal way are associated with higher robustness.

For both GOLD and PLANTS, 10 poses per molecule were collected. The default parameters were used for both protocols. Concerning the choice of the scoring function, Chemscore was selected for GOLD, while $PLANTS_{ChemPLP}$ was selected for PLANTS.

A total of 1152 (24 ligands x 24 proteins x 2 docking protocols) independent docking runs were performed, and the results were then analyzed by making use of an in-house Python script. The script collects the RMSD between each docking pose and the correspondent crystal reference pose, outputting two different plots. The first plot is a heatmap that

illustrates the RMSD values for the best docking pose generated for each ligand onto each protein. The second plot is a histogram that re-elaborates the previous results to give a visual representation of the "success rate" of each protein: a successful docking run is obtained when the RMSD between the docking pose and the crystal reference is below the arbitrary chosen 2 Å threshold value so that the "success-rate" is defined by the percentage ratio of the successful docking runs for each protein (i.e., the percentage of docking experiments where the RMSD falls below the threshold value).

2.4 Pharmacophore Modeling

Based on previously published works on the same target, we took advantage of the structural information about known inhibitors of CK1δ in the form of crystal structures of their complex with the kinase deposited in the Protein Data Bank. The same 24 protein-ligand complexes mentioned in section 2.2 were subjected to the MOE Pharmacophore model tool: shared interaction features (with a 50% threshold value for feature retention) were then used in the generation of the pharmacophore model.

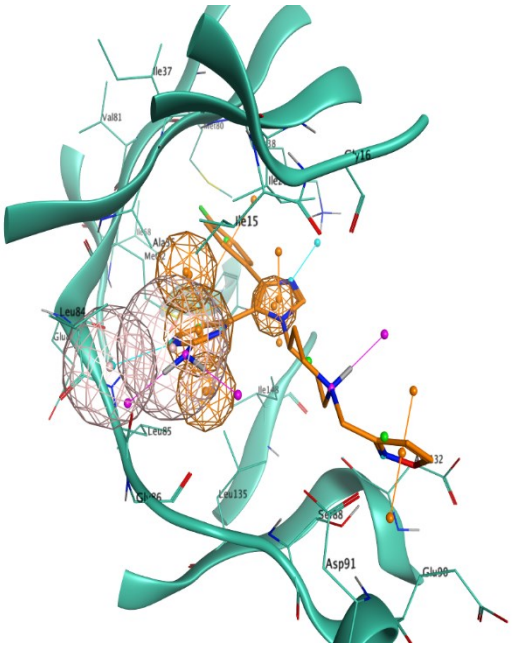


Figure 5. Visual representation of the pharmacophore model used in this scientific work. Features are represented as spheres. Orange spheres indicate an aromatic ring, with an orientation determined by the small orange pin, while the pink spheres indicate a hydrogen bond donor/acceptor. For visual reference, the 4TN6 complex is also reported in this figure, with the protein represented in teal ribbons and the PFO ligand represented as orange sticks.

As can be seen in Figure 1, the final model consisted of 4 features (represented as spheres in the image), namely a hydrogen bond donor and a hydrogen bond acceptor interacting with Leu85, an aromatic ring in the proximity of the hinge region, and another aromatic ring adjacent to the first one in the inner part of the binding pocket.

2.5 Autogrow

Autogrow4[55] is a fully open-source code written in Python and developed by the Durrant lab that combines a genetic algorithm with docking calculation based on the Vina[56] docking software (version 1.2.0) to perform a semi-automatized process for both de-novo drug design and lead optimization.

Molecules are submitted to the program in the form of SMILES strings. The genetic algorithm part of the code uses a series of synthetically feasible reactions to perform a defined number of mutation and crossover operations (i.e., growing and merging) on submitted chemical entities, creating a full population (called generation) of molecules to feed to the molecular docking program.This generation is then docked using the Vina docking software. After the docking stage, the genetic algorithm retrieves the score for each docking pose, which it uses to rank molecules and pick the most fitted members of the generation to promote them to the next generation. This iterative process is repeated for a user-defined number of generations, or until an earlier termination criterion is met.

The code is released under the Apache2 license, is freely available at https://durrantlab.pitt.edu/autogrow4/, and works both in a Python 2.7 and >=3.6 environment. A detailed description of how the latest Autogrow release works is provided in the work of Spiegel et al.[40].

Two different versions of the Autogrow code were used in this scientific work. The first one was downloaded from the official repository and used as is, without any modifications to the source code. The second one was the result of an in-house modification of the source code performed to customize the scoring stage of the docking process. The traditional Autogrow protocol uses the Vina standard scoring function (from now on, defined as VINA), which encompasses some elements of knowledge-based potentials and others of a typical empiric scoring function[56]. Instead, our modified version of the

Autogrow code implements an alternative scoring function (from now on, defined as IFP_{CS}) based on the similarity between protein-ligand interaction fingerprints.

The crystal complex of a known inhibitor is chosen as reference (in our case, the ligand PFO from complex 4TN6 was chosen) and its binding mode is codified into a bit vector exploiting the InteractionFingerprint function from the fingerprint module of the Open Drug Discovery Toolkit[57] Python Library. This function converts the protein-ligand interaction into a bit array according to the residue of choice and the type of interaction. Each protein residue is represented by eight bits, one for each type of interaction considered (hydrophobic contacts, aromatic face to face, aromatic edge to face, hydrogen bond with protein acting as donor, hydrogen bond with protein acting as acceptor, salt bridge with protein acting as the positively charged member, salt bridge with protein acting as the positively negative member and ionic bond with a metal ion), so that the final vector will have a size of r*8, where r stands for the number of protein residues.

During the scoring phase of our custom Autogrow run, each docking pose is also codified into an Interaction Fingerprint vector, the same way as for the crystal reference. Then, the two vectors are transformed from sparse to dense making use of the appropriate functions from the Numpy Python library, before the comparison between the reference and the query fingerprint is executed using the cosine similarity metrics, exploiting the appropriate function of the Scikit-learn Python library. The resulting score, which ranges from 1 (indicating a complete agreement and coherence between the two binding modes) to 0 (indicating that the two binding modes are not coherent), is then multiplied by -1 to comply with the selection mechanism of Autogrow genetic algorithm, which favors the most negative scores, as is usually the case for most classic scoring functions, like the one used by Vina.

$$IFP_{CS} = \frac{A \cdot B}{\|A\|\|B\|} * (-1)$$

Equation 1: mathematic formulation of the IFP_{CS} scoring function. The IFP_{CS} scoring function is the inverse of the cosine similarity between two vectors, A and B, representing the Interaction Fingerprint for the reference and the query ligand respectively. Values range from -1 (indicating maximum coherence between the two binding modes) to 0 (indicating the lowest possible correspondence between the two binding modes).

Results

3.1 Cross-docking

Since 24 different protein-ligand complexes were available for CK1δ (the target for our computational study), but only one at a given time can be used for docking calculations, we had to carefully evaluate which one was the most suitable for our needs. The choice of the protein structure to use for docking calculation is not trivial, for several reasons. When a ligand gets in contact with a protein, the binding event may cause a change in the structure of the protein itself[58]. These modifications are mainly depictable in the binding site and may also be extended to other regions. In a crystallographic complex, this effect is highlightable by differences in the shape of the binding site among the different crystal structures available for a single protein[59].

One of the possible approaches to accomplish this task, which is the one that we used in our workflow, is known as "cross-docking"[60]. This technique consists in taking all the protein-ligand complexes available for a target, separating all the ligands from their respective co-crystallized structure, and docking all the different ligands in the binding site of each different protein structure. By analyzing the docking results, it is possible to define which is the crystallographic protein structure that has the highest tendency to correctly reproduce ligands' crystallographic conformation.

For these reasons, we performed a cross-docking experiment on our 24 CK1δ complexes to decide which one to pick for subsequent calculation. Each ligand was docked into each protein structure using two different docking protocols, GOLD-Chemscore and PLANTS-$PLANTS_{ChemPLP}$, for a total of 1152 independent docking runs. For each ligand, the root-mean-square deviation (RMSD) between each docking pose and the crystallographic conformation was calculated. The poses with the lowest RMSD in each docking run were selected and their RMSDs were plotted, obtaining the graphs represented in Figure 2. A detailed description of the methodology used for the cross-docking experiment is provided in the Materials and Methods, section 2.3.

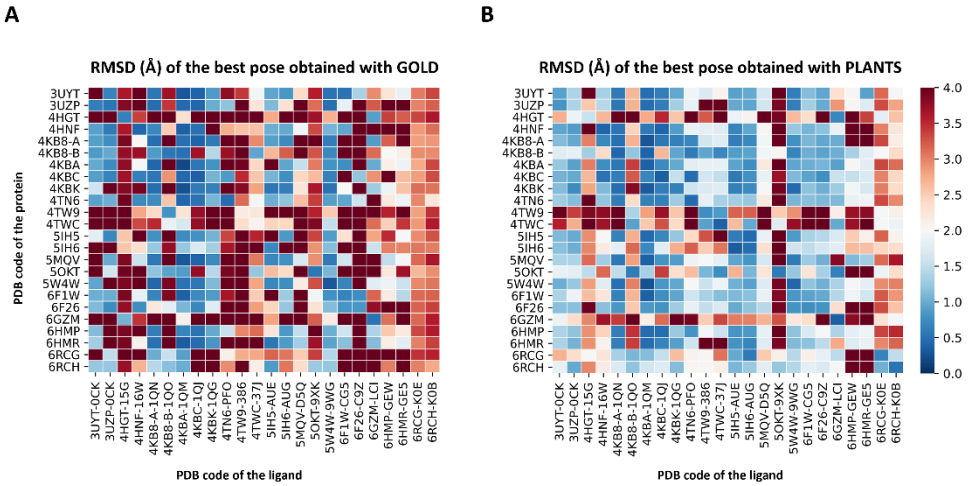


Figure 6. This figure contains two heatmaps that summarize the results of the cross-docking experiment performed before the Autogrow runs to select the protein structure to use for subsequent calculations. Panel A reports the results for the GOLD-Chemscore protocol, while Panel B encompasses the results of the PLANTS-PLANTS$_{ChemPLP}$ one. On the vertical axis, the PDB code of the protein is reported, while on the horizontal axis the PDB code of the ligand is indicated. The colored squares report the RMSD values for the best docking pose generated by the two docking protocols according to the color bar located on the right side of the image: color ranges from Blue (indicating a low RMSD (minimum value is 0 Å, indicating a perfect superposition between the docking pose and the crystal reference) to Red (maximum value is 4 Å, indicating a high deviation between the docking pose and the crystal reference).

To visualize the results more clearly, the data from the plots reported in Figure 2 were re-elaborated to obtain a single indicator of the performance of each protein in reproducing the correct binding mode for docked ligands. We opted for calculating the “success rate” for each protein structure: a 2 Å threshold value was chosen to discriminate between successful and unsuccessful docking runs. For each protein, the percentage of successful docking runs (the “success rate”) was calculated accordingly and plotted in a histogram.

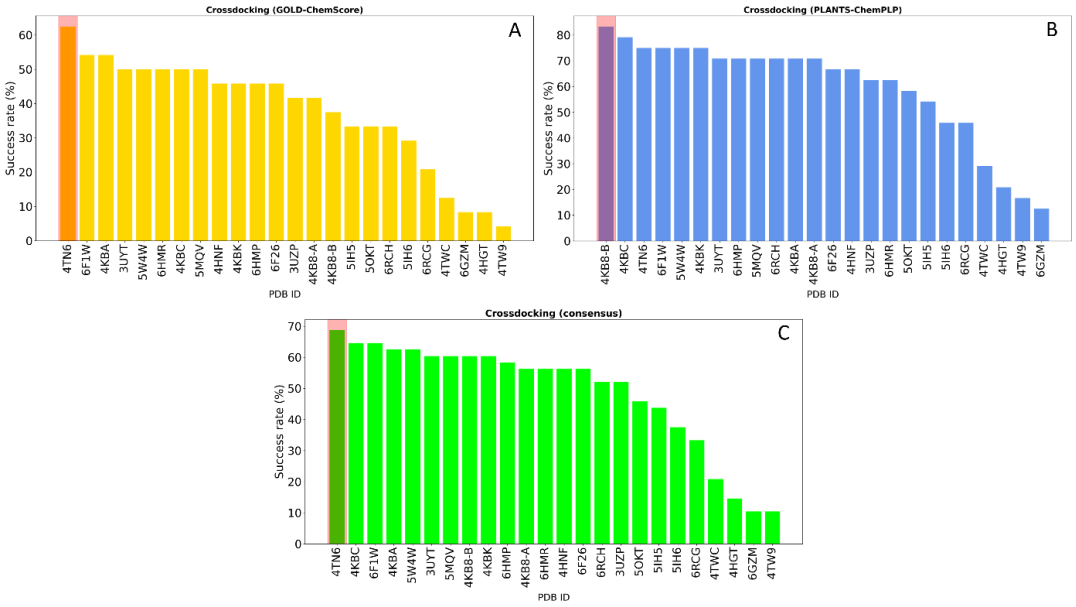


Figure 7. This figure shows the overall "success rate" in reproducing the correct crystallographic binding mode for each of the 24 CK1δ complexes considered in the study. The "success rate" is defined as the percentage of successful docking runs for each protein in the cross-docking experiment, where a successful docking run is defined as a docking calculation where the RMSD between the best docking pose and the crystal reference falls below an arbitrarily chosen threshold value of 2 Å. Panel A reports the results for the GOLD-Chemscore protocol. Panel B reports the results for the PLANTS-PLANTS$_{ChemPLP}$ protocol. Panel C encompasses the combined "success rate" for each protein, defined as the average between the success rate for each protocol. Protein from the complex 4TN6 was chosen as the most representative CK1δ structure for successive calculations.

Figure 3 encompasses the results of this second analysis, reporting the success rate for both the GOLD-Chemscore and PLANTS-PLANTS$_{ChemPLP}$ protocols. Moreover, since we adopted the principle of "consensus docking", as mentioned in section 2.3, we decided to calculate the average success rate between the two docking protocols. As can be seen in Figure 3, the overall "success rate" obtained by the combination of data from the two docking protocols indicates the protein from the complex 4TN6 as the protein that is, on average, more able than the other ones to correctly reproduce the crystallographic binding mode of docked ligands. Although the difference in the success rate between the first and the second protein is low, in the context of several consequential docking runs where thousands of compounds are considered at a given time even small differences in the percentage success rate could have a big impact on the quality of the run, considering

that the prioritization of compounds from one generation to another is based upon their docking-predicted ability to retain the interaction features that characterize the binding mode of known inhibitors. For this reason, we used the protein 4TN6 as a representative CK1δ structure for our subsequent calculations with Autogrow.

3.2 Benchmark DeNovo Run

To assess the performance of our alternative, fingerprint-based, Autogrow protocol (defined as IFP_{CS}, while the traditional one is defined from now on as VINA), we first performed a benchmark *denovo* run, using the same conditions as the ones described in the work of Spiegel et al.[40]

A 30-generation run was performed for each protocol, using the "Fragment_MW_100_to_150.smi" library provided in the Autogrow repository and described in the original publication. Configuration files for both *denovo* runs in the JSON format are available in the Supplementary Materials, while a detailed description of both Autogrow and our alternative scoring approach is described in Materials and Methods, section 2.5.

In order to validate the performance of both protocols, we opted for evaluating the quality of the generated compounds by filtering each generation of poses using a pharmacophore model. This filter, which has already proved to identify true binders in previous related works[42,61], was used to retain only those poses which complied with known requirements for binding to the CK1δ pocket. This metric was used to determine if there is any advantage in incorporating a knowledge-based element in the generation of novel potential inhibitors of CK1δ, steering the compound selection process towards the ones that assume a pharmacophore-like binding mode. These pharmacophore-like compounds were then characterized by calculating their molecular weight and the similarity of their shape and electrostatic properties to crystal CK1δ inhibitors taken as reference. For this purpose, the EON[56] package from the OpenEye suite was used. Each compound passing the pharmacophore filter was compared with each crystallographic ligand, calculating the electrostatic and shape similarity (ET_{combo}). The best value for each

ligand was extracted and used for the elaboration of the a posteriori analysis, whose results are reported in Figure 4 and Figure 5.

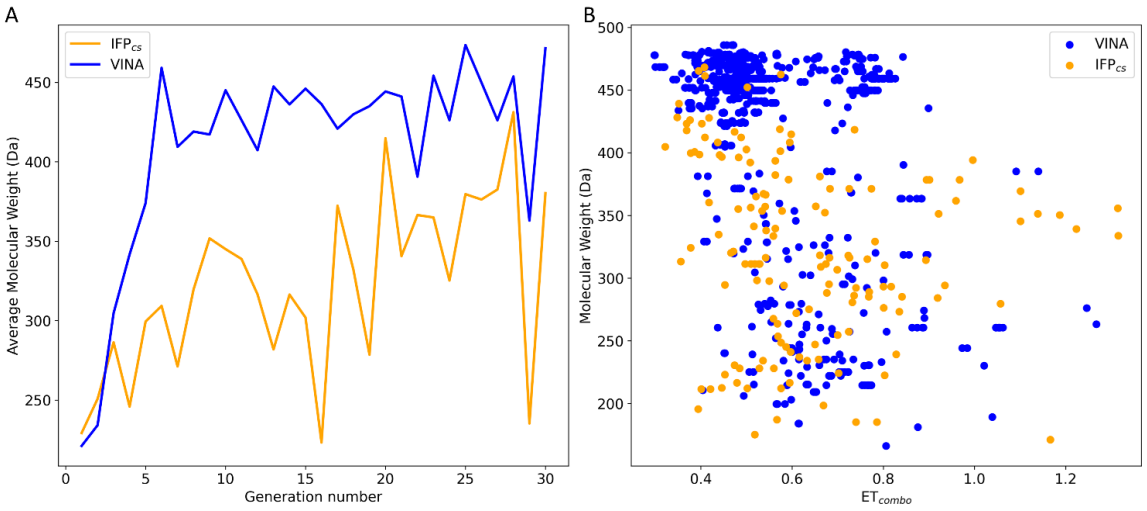


Figure 8. This figure compares the performance of the two Autogrow protocols in the benchmark *denovo* runs regarding their ability to generate compounds that pass the pharmacophore filter. The VINA protocol is reported as a blue line, while the IFP$_{CS}$ one is reported as an orange line. Panel A depicts, for each protocol, the average molecular weight of compounds within the population that pass the pharmacophore filter on a per-generation basis. The vertical axis reports the molecular weight, while the horizontal axis reports the generation number. Panel B depicts, for each protocol, the distribution of generated compounds that pass the pharmacophore filter regarding their molecular weight and the similarity of shape and electrostatic properties to crystal inhibitors taken as reference. The vertical axis reports the average molecular weight in Da, while the horizontal axis reports the ET_{combo} value. Blue dots represent compounds generated by the VINA protocol, while orange dots represent compounds generated by the IFP$_{CS}$ one.

As can be seen in Figure 4 (panel A), which shows the average molecular weight of compounds that pass the pharmacophore filter for each generation, the VINA protocol rapidly reaches the peak of the average molecular weight (around generation 6), while our IFP$_{CS}$ protocol has a slower but regular growth that reaches values comparable to the VINA protocol from around generation 27 onwards. This difference is probably related to the fact that the VINA scoring function is biased towards the selection of larger compounds, which can make a good number of non-specific interactions with the target,

while our IFP_{CS} one is biased towards the selection of compounds that have a similar interaction pattern compared to a reference compound, regardless of their dimensions.

As depicted in Figure 4 (panel B), which illustrates the distribution of generated compounds across all generations concerning their molecular weight and their electrostatic and shape similarity with crystal CK1δ inhibitors, this different selection process results in the production of compounds with different properties: the blue dots, which represent the compounds generated by the VINA protocol, are mostly located in the left-upper portion of the graph, indicating that most of the compounds generated by the traditional protocol have a high molecular weight but a low level of similarity with known inhibitors. On the contrary, the upper-right part of the graph (high molecular weight, high electrostatic, and shape similarity) is mostly populated with orange dots, which represent the compounds generated by our IFP_{CS} protocol.

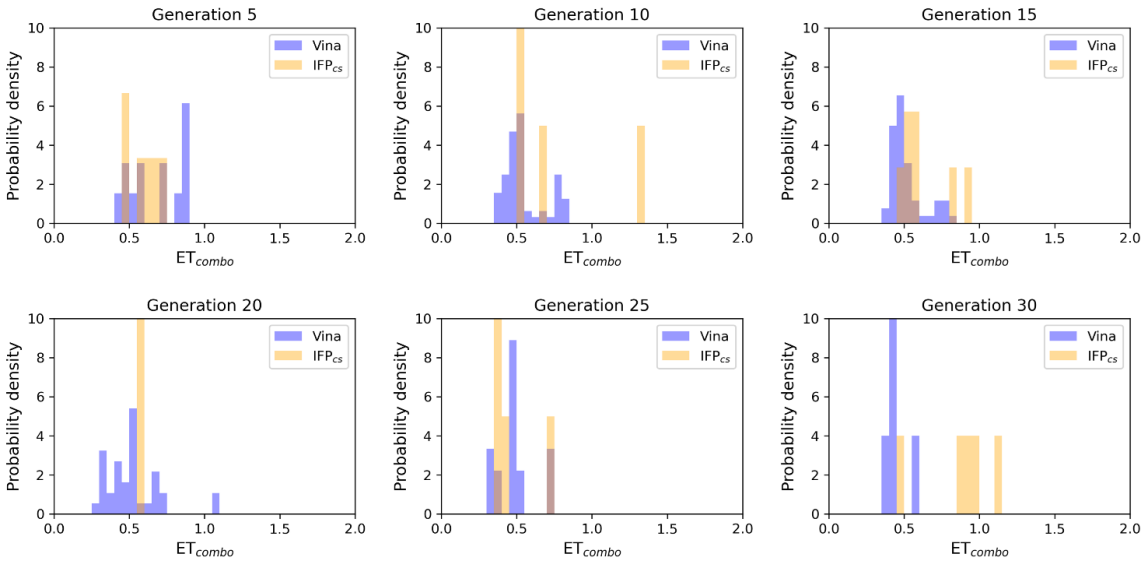


Figure 9. This figure encompasses the ability of the two Autogrow protocols in the benchmark *denovo* run to produce compounds that have a high degree of similarity concerning shape and electrostatic properties to the crystallographic ligands, chosen as reference. The probability distribution of the ET_{combo} score for compounds populating each generation is reported as a histogram, where the vertical axis reports the probability density while the horizontal axis reports the ET_{combo} value. Two distributions are reported within each plot: the blue bars refer to compounds generated with the VINA protocol, while the orange bars refer to compounds generated with the IFP_{CS} one.

The difference in the selection process is also highlighted in Figure 5, which illustrates the distribution of compounds across a representative subset of generations concerning their electrostatic and shape similarity: the graph clearly shows how the VINA protocol does not improve the similarity of generated compounds while increasing the number of generations. On the contrary, the orange population (which represents the compounds generated by the IFP_{CS} protocol) gradually shifts towards the right part of the plot passing from earlier to later stage generations, indicating that the compounds passing the pharmacophore filter increase their electrostatic and shape similarity passing from one generation to another. Another comparison of the performances of the two protocols is given in Figure 6, which reports the progressive enrichment in compounds with a high degree of similarity to reference inhibitors within the total population. An example of a high-scoring compound generated by our IFP_{CS} protocol is reported in Figure 7, where its chemical structure and the comparison between its docking-predicted binding mode and crystal pose of the PFO ligand from reference crystal complex 4TN6 is shown.

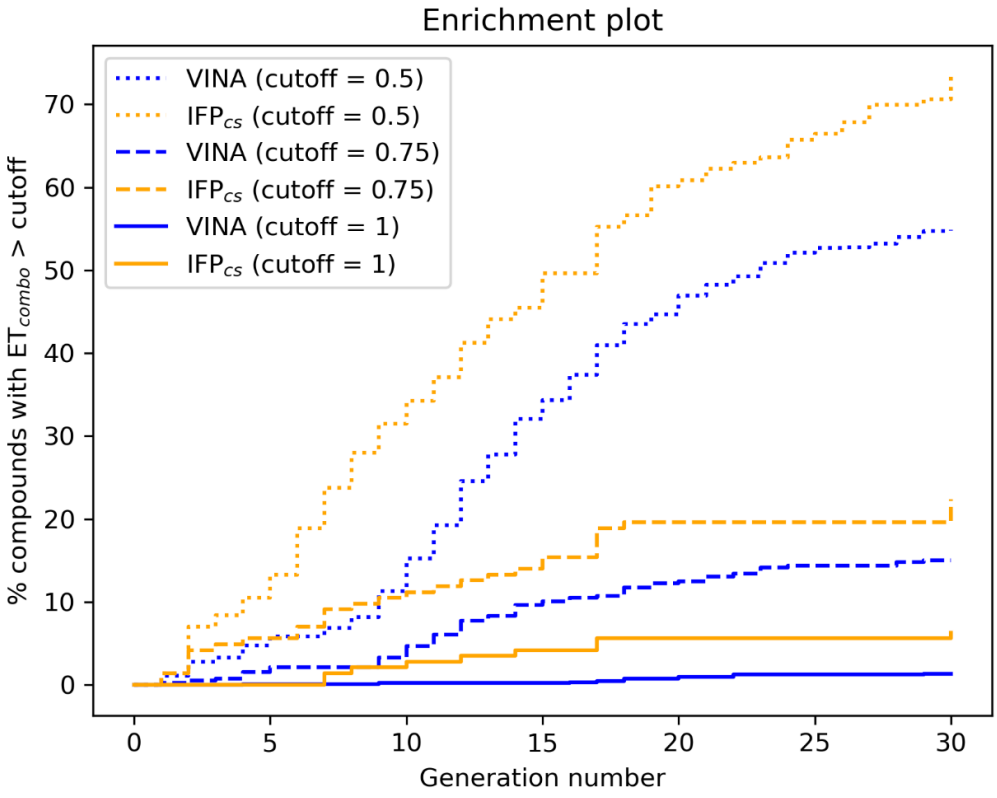


Figure 6. This figure illustrates the capability of the two different Autogrow protocols in the benchmark *denovo* run to produce compounds that have a high degree of similarity concerning shape and electrostatic properties to the crystallographic ligands, chosen as reference. For each generation, the percentage of compounds within the total population whose ET_{combo} exceeds a defined threshold value is reported. Three different cutoff values are reported, 0.50, 0.75, and 1.00 respectively.

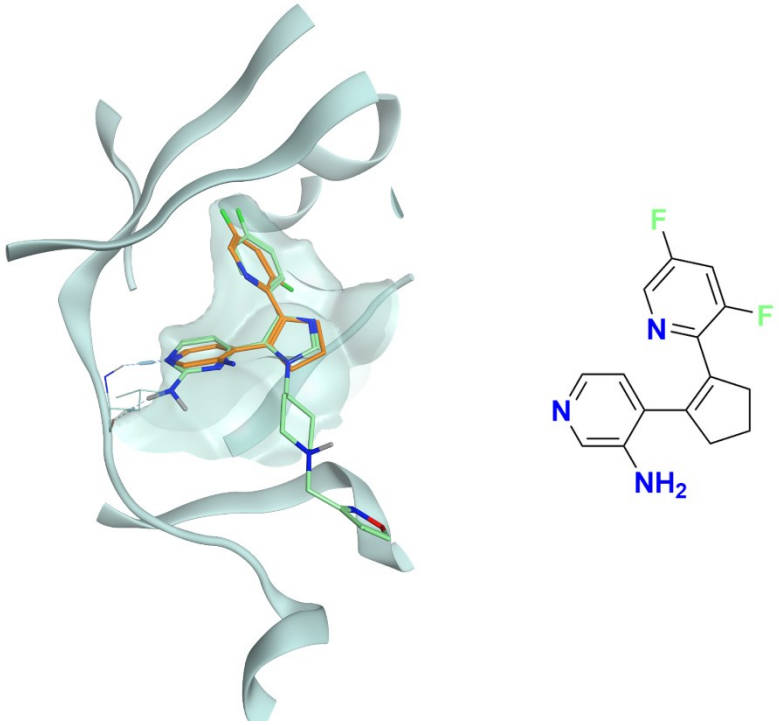


Figure 7. This figure reports the superposition between the docking-predicted binding mode of a high scoring compound (MMS1) from the benchmark *denovo* run performed with the IFP_{CS} scoring protocol and the reference crystal binding pose of compound PFO from the structure deposited in the Protein Data Bank with accession code 4TN6. On the left part of the image, the protein kinase CK1δ ATP binding site is reported in teal ribbon, the pose of compound MMS1 is shown in orange sticks while the pose of compound PFO is shown in green sticks. On the right part of the image, the chemical structure of compound MMS1 is reported.

3.3 Benchmark Lead Optimization Run

To further evaluate the validity of our custom scoring protocol, we performed also a benchmark lead optimization run, using once again the same conditions as the ones reported in the work of Spiegel et al.[40] A 5-generation run was performed for each protocol, using a library composed of the 24 crystallographic ligands mentioned in the previous sections and other 316 fragments obtained from the fragmentation of crystallographic ligands exploiting the “fragmenter_of_smi_mol.py” Python script

provided by the Autogrow developers, using the BRICS fragmentation rule, for a total of 340 compounds fed to the algorithm. In this case, also, configuration files for both benchmark runs in the JSON format are available in the Supplementary Materials.

To assess the performance of both protocols, we applied the same criteria described previously for the *denovo* runs, focusing once again on compounds passing the pharmacophore filter described in section 2.4 and characterizing them about their molecular weight and electrostatic and shape similarity compared to crystal CK1δ inhibitors.

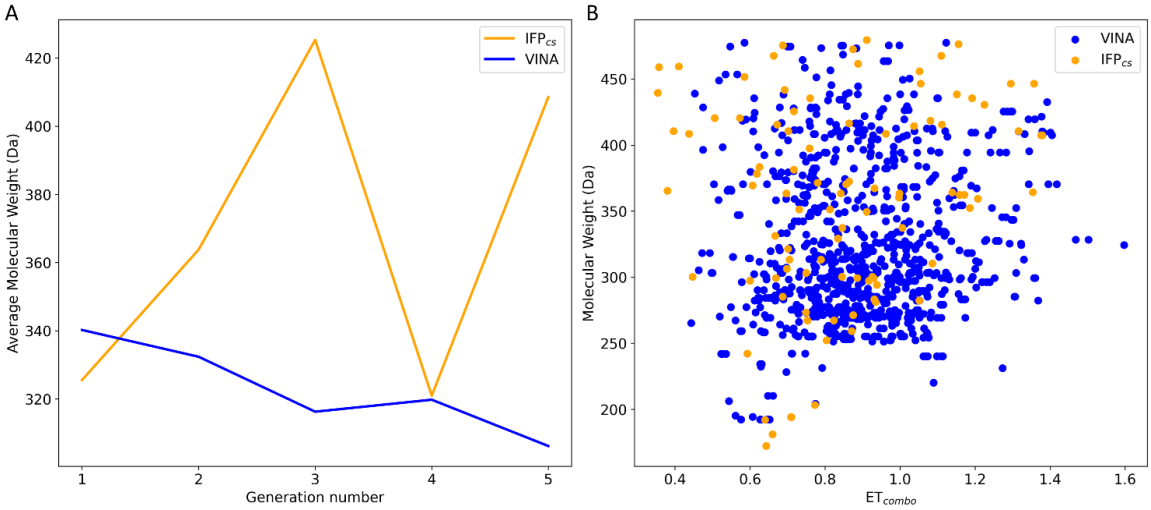


Figure 8. This figure compares the performance of the two Autogrow protocols in the benchmark lead optimization runs regarding their ability to generate compounds that pass the pharmacophore filter. The VINA protocol is reported as a blue line, while the IFP$_{CS}$ one is reported as an orange line. Panel A depicts, for each protocol, the average molecular weight of compounds within the population that pass the pharmacophore filter on a per-generation basis. The vertical axis reports the molecular weight, while the horizontal axis reports the generation number. Panel B depicts, for each protocol, the distribution of generated compounds that pass the pharmacophore filter regarding their molecular weight and the similarity of shape and electrostatic properties to crystal inhibitors taken as reference. The vertical axis reports the average molecular weight in Da, while the horizontal axis reports the ET_{combo} value. Blue dots represent compounds generated by the VINA protocol, while orange dots represent compounds generated by the IFP$_{CS}$ one.

Figure 8 (panel B) illustrates the distribution of compounds across all five generations regarding their ET_{combo} and their molecular weight: as can be seen, there is little to no difference between the two protocols, with the two populations being practically superimposable. However, contrary to what might be suggested by this plot, there is a significant difference in the performances of the two protocols, which is highlighted in both Figure 8 (panel A), Figure 9, and Figure 10.

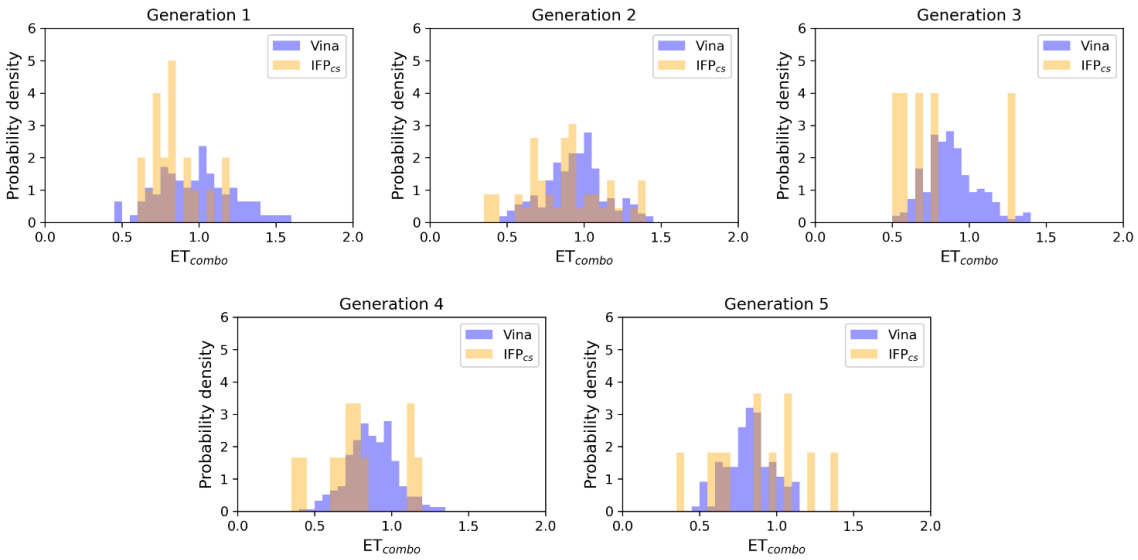


Figure 9. This figure encompasses the ability of the two Autogrow protocols in the benchmark lead optimization run to produce compounds that have a high degree of similarity with regards to shape and electrostatic properties to the crystallographic ligands, chosen as reference. The probability distribution of the *ET$_{combo}$* score for compounds populating each generation is reported as a histogram, where the vertical axis reports the probability density while the horizontal axis reports the *ET$_{combo}$* value. Two distributions are reported within each plot: the blue bars refer to compounds generated with the VINA protocol, while the orange bars refer to compounds generated with the IFP_{CS} one.

As can be noticed in Figure 8 (panel A), the average molecular weight of compounds passing the pharmacophore filter grows by about 90 Da passing from the first to the last generation in the case of our IFP_{CS} protocol. On the contrary, not only the average molecular weight of pharmacophore-like compounds generated by the traditional VINA protocol does not increase with the number of generations, but slightly decreases over time, falling even below the average molecular weight of the first generation derived from

the IFP_{CS} protocol. Furthermore, Figure 9 illustrates how, as previously seen in the benchmark *denovo* run, the similarity of compounds passing the pharmacophore filter increases over time when the IFP_{CS} scoring protocol is adopted, while it slightly decreases and does not improve over time in the case of the traditional VINA scoring protocol.

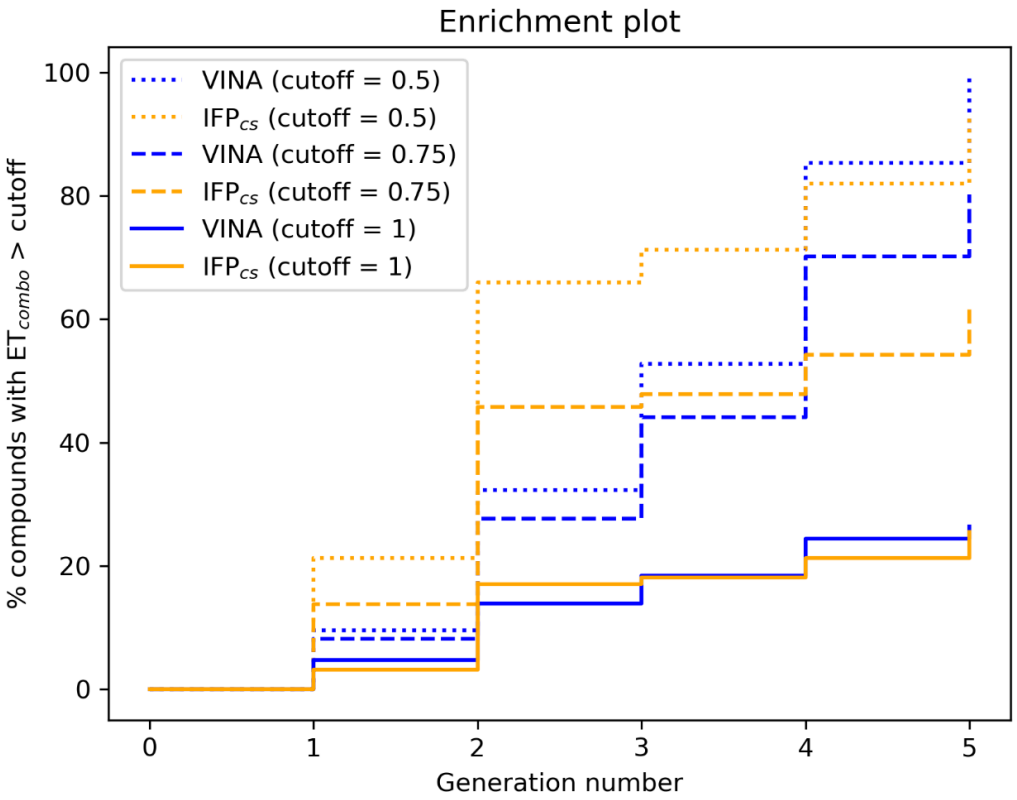


Figure 10. This figure illustrates the capability of the two different Autogrow protocols in the benchmark lead optimization run to produce compounds that have a high degree of similarity concerning shape and electrostatic properties to the crystallographic ligands, chosen as reference. For each generation, the percentage of compounds within the total population whose ET_{combo} exceeds a defined threshold value is reported. Three different cutoff values are reported, 0.50, 0.75, and 1.00 respectively.

Particularly, this trend is also confirmed by Figure 10, which shows how the IFP_{CS} protocol can produce a quicker enrichment of the population in high similarity compounds compared to the traditional VINA one. As for the previous case, an example of a high-scoring compound generated in the last and final generation of the IFP_{CS} run is reported in Figure 11.

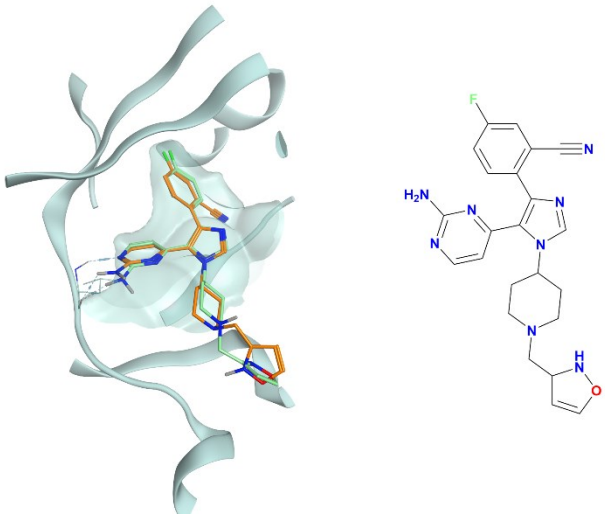


Figure 11. This figure reports the superposition between the docking-predicted binding mode of a high scoring compound (MMS2) from the benchmark lead optimization run performed with the IFP_{CS} scoring protocol and the reference crystal binding pose of compound PFO from the structure deposited in the Protein Data Bank with accession code 4TN6. On the left part of the image, the protein kinase CK1δ ATP binding site is reported in teal ribbon, the pose of compound MMS2 is shown in orange sticks while the pose of compound PFO is shown in green sticks. On the right part of the image, the chemical structure of compound MMS2 is reported.

3.4 Prospective DeNovo Run

Encouraged by the results of our benchmark runs, we decided to perform a prospective run with the IFP_{CS} protocol, applying the same operating conditions as before. This time, the starting library was modified to add to the compounds used for the benchmark runs also 7 fragment ATP-competitive CK1δ inhibitors identified during a previous virtual screening campaign from our laboratory[42]. The idea behind this run was to evaluate the ability of our IFP_{CS} scoring protocol to generate interesting novel potential CK1δ inhibitors derived from in-house, readily available compounds.

The chemical structure of the seven fragments used in this run is reported in Figure 12.

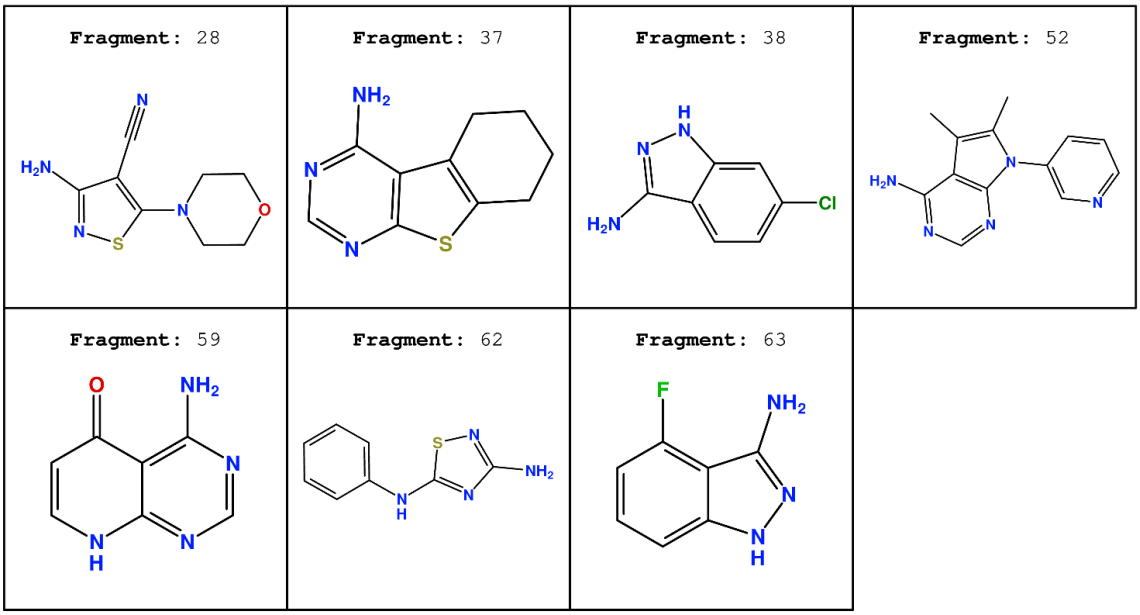


Figure 12. Chemical structure of the seven fragment CK1δ inhibitors derived from the work of Bolcato et al.[42]

To verify the quality of this run, we performed the same analysis as for the benchmark runs. The results of this analysis are summarized in Figures 13, 14, and 15 respectively.

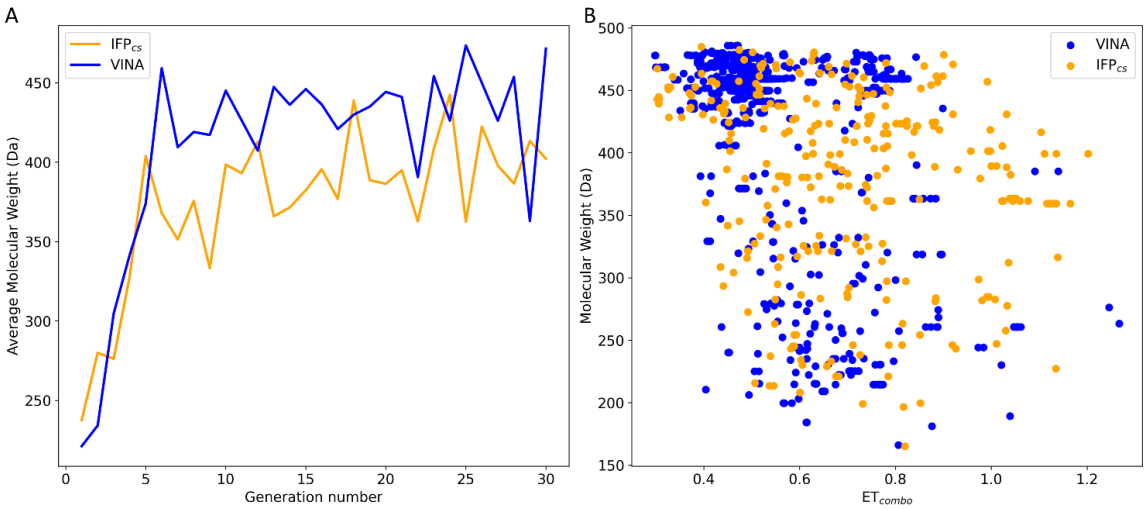


Figure 13. This figure compares the performance of the two Autogrow protocols in the prospective *denovo* runs regarding their ability to generate compounds that pass the pharmacophore filter. The VINA protocol is reported as a blue line, while the IFP$_{CS}$ one is reported as an orange line. Panel A depicts, for each protocol, the average molecular weight of compounds within the population that pass the pharmacophore filter on

a per-generation basis. The vertical axis reports the molecular weight, while the horizontal axis reports the generation number. Panel B depicts, for each protocol, the distribution of generated compounds that pass the pharmacophore filter regarding their molecular weight and the similarity of shape and electrostatic properties to crystal inhibitors taken as reference. The vertical axis reports the average molecular weight in Da, while the horizontal axis reports the ET_{combo} value. Blue dots represent compounds generated by the VINA protocol, while orange dots represent compounds generated by the IFP_{CS} one.

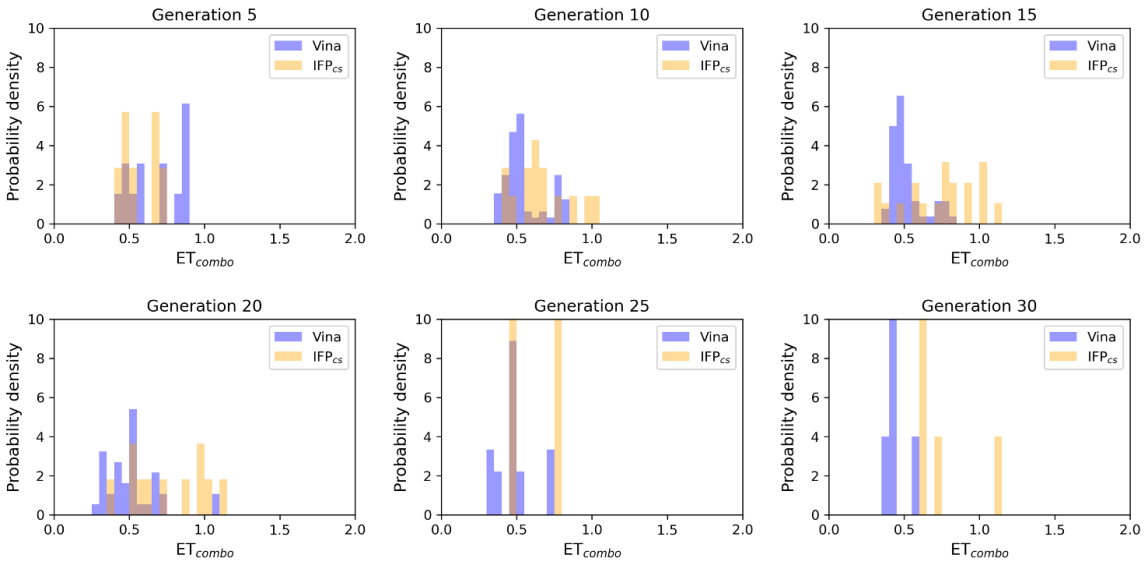


Figure 14. This figure encompasses the ability of the two Autogrow protocols in the prospective *denovo* run to produce compounds that have a high degree of similarity with regards to shape and electrostatic properties to the crystallographic ligands, chosen as reference. The probability distribution of the ET_{combo} score for compounds populating each generation is reported as a histogram, where the vertical axis reports the probability density while the horizontal axis reports the ET_{combo} value. Two distributions are reported within each plot: the blue bars refer to compounds generated with the VINA protocol, while the orange bars refer to compounds generated with the IFP_{CS} one.

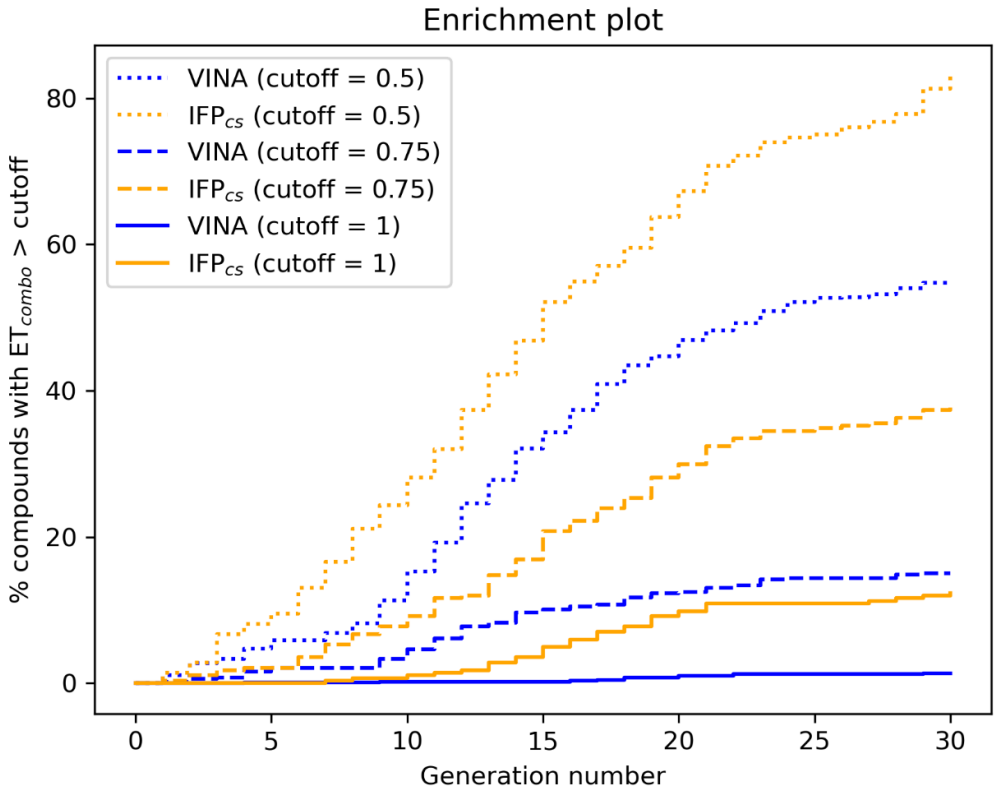


Figure 15. This figure illustrates the capability of the two different Autogrow protocols in the prospective *denovo* run to produce compounds that have a high degree of similarity concerning shape and electrostatic properties to the crystallographic ligands, chosen as reference. For each generation, the percentage of compounds within the total population whose ET_{combo} exceeds a defined threshold value is reported. Three different cutoff values are reported, 0.50, 0.75, and 1.00 respectively.

As remarked in Figure 14, the same trend seen in the benchmark *denovo* run can be observed also in the case of this prospective run: while the VINA protocol is not able to increase the shape and electrostatic similarity to known inhibitors over time, the IFP_{CS} protocol can produce a shift of the orange population towards higher ET_{combo} values. As illustrated by Figure 13, which reports a comparison between the benchmark *denovo* run performed with the VINA protocol and the prospective *denovo* run carried out with the IFP_{CS} protocol, the trend in both the distribution of compounds regarding their molecular weight and ET_{combo} and the growth of molecular weight over time are similar to the benchmark *denovo* run. Figure 13 (panel B) clearly shows how the upper-right portion of the graph, which hosts compounds with both high molecular weight and ET_{combo} values, is populated exclusively by orange dots, which represent compounds generated with the

IFP_{CS} scoring protocol. Interestingly, Figure 14 (panel A) highlights how there is much less difference in the growth rate of molecular weight between the IFP_{CS} run (which is contaminated by the presence of our 7 CK1δ-inhibiting fragments) and the benchmark VINA run, suggesting that performances of the IFP_{CS} could improve if some high-quality pharmacophore-like fragments are included in the starting library. However, despite the quicker growth of molecular weight, the quality of generated compounds follows the same trend seen in the benchmark *denovo* run, as reported in Figure 15. As for the previous cases, an example of a high-scoring compound is reported in Figure 16.

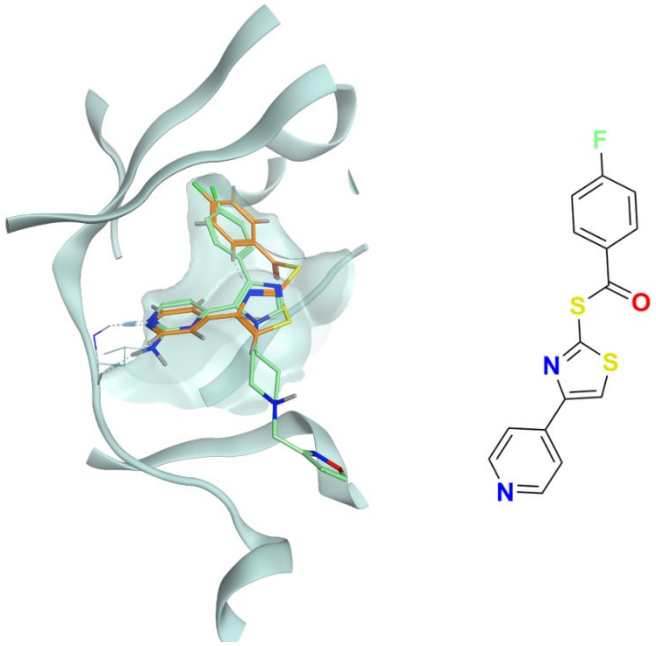


Figure 16. This figure reports the superposition between the docking-predicted binding mode of a high scoring compound (MMS3) from the benchmark *denovo* run performed with the IFP_{CS} scoring protocol and the reference crystal binding pose of compound PFO from the structure deposited in the Protein Data Bank with accession code 4TN6. On the left part of the image, the protein kinase CK1δ ATP binding site is reported in teal ribbon, the pose of compound MMS3 is shown in orange sticks while the pose of compound PFO is shown in green sticks. On the right part of the image, the chemical structure of compound MMS3 is reported.

Discussion

The open-source software Autogrow4[55] is an interesting piece of code that utilizes a combination between a genetic algorithm and the Vina[41] molecular docking software to semi-automatize the processes of fragment growing and lead optimization. Thanks to the modular nature of the codebase, we implemented an alternative scoring protocol (IFP_{CS}) based on the similarity of protein-ligand interaction fingerprint between a crystal reference and query compounds, exploiting the appropriate function from the open-source library Open Drug Discovery Toolkit[57], and we compared its performances with the traditional Autogrow scoring protocol (VINA), which is based on the Autodock Vina scoring function.

The VINA protocol uses a scoring function that incorporates some elements of knowledge-based potentials and some others of empirical scoring functions. As is the case for many scoring functions, the score is biased towards higher molecular weight compounds, which can establish a higher number of non-specific interactions with the target[44]. For this reason, usually, molecular docking programs are efficient in sampling the conformational space available for the ligand within the binding site but are weaker in prioritizing the right binding mode among a set of reasonable hypotheses generated by the search algorithm[43]. This is especially true in the case of fragments, which deviate from the drug-like properties of compounds upon which traditional scoring functions have been trained[46].

As thoroughly discussed in the work of Bolcato et al.[42], one possible solution to the scoring problem is to apply a pharmacophore filter to poses generated by the molecular docking program. When several structural pieces of information are available in the form of protein-ligand crystal complexes for a certain target (as is the case for protein kinase CK1δ, the case study for this work), a good solution to reduce the false positive rate of molecular docking programs is to build a pharmacophore model that encompasses the most prominent interaction features that are required to ligands to bind to the target active site[62]. In the case of a program like Autogrow, where the selection mechanism that determines which compounds to promote to the next generation is based on the docking score, we thought it would be interesting to incorporate a knowledge-based element in the pose selection mechanism in the form of a comparison between the interaction

fingerprint of query compounds and known inhibitors, to bias the selection mechanism towards molecules that respect the required features to bind to the target.

To validate our IFP_{CS} scoring protocol, we performed both a *denovo* and a lead-optimization benchmark run, using the same operative conditions described in the original work of Spiegel et al.[40] but on a different target. The protein target of choice was the protein kinase CK1δ, a pharmaceutically relevant target in the field of neurodegenerative diseases for which several crystal complexes with inhibitors are available in the Protein Data Bank. The benchmark *denovo* run was performed on a library composed of 6103 fragment compounds whose molecular weight falls between 100 and 150 Da, while the benchmark lead-optimization run was carried out on a library composed of 24 crystallographic ligands of the protein kinase CK1δ and 316 fragments derived from the fragmentation of crystallographic ligands using the BRICS rule. To compare the capabilities of the two protocols, we filtered each generation of compounds with the same pharmacophore filter already utilized in previous scientific works on the target[42,61]. We then proceeded to evaluate the quality of compounds that pass the pharmacophore filter, considering both the size and the similarity of shape and electrostatic properties of query compounds compared to the crystallographic ligands taken as reference.

As illustrated by the results of our analysis (Section 3.3 and 3.4, respectively), there is a substantial difference in the performances of the two protocols: while both protocols can generate a certain amount of compounds that pass the pharmacophore filter (therefore possessing the right structural features that are required for the interaction with the target), in both scenarios the IFP_{CS} scoring function outperforms the traditional VINA one regarding the ability to select and prioritize pharmacophore-like compounds that have a similar shape and electrostatic properties compared to known inhibitors of the protein kinase CK1δ. This is particularly evident in the lead-optimization scenario, where within each generation passage, the average molecular weight of compounds that pass the pharmacophore filter steadily increases, passing from the typical MW of a fragment-like compound to the MW of a grown, mature, lead candidate, while the contrary happens in the case of the VINA protocol, with the average MW of the compounds that pass the pharmacophore filter steadily decreasing, falling even below the value of the first

generation from the IFP_{CS} protocol. Moreover, when poses from each generation are compared with the ones of crystallographic ligands concerning the shape and electrostatic similarity, a similar trend can be noticed. While the VINA protocol can select high-quality compounds in the first generation, compared to the IFP_{CS} one, at later stages during the run a progressive reduction in the similarity between the query and reference compounds can be noticed, contrary to what happens when the IFP_{CS} scoring protocol is utilized. This can be explained considering the different nature of the two scoring functions: the VINA protocol is biased towards bigger, therefore higher scoring, compounds, while the IFP_{CS} protocol favors compounds that respect the interaction pattern of the reference crystallographic ligand, regardless of their size. For this reason, the IFP_{CS} protocol tends to favor smaller compounds in the first generations, as long as they are complying with the constraint imposed by the reference interaction fingerprint, increasing the possibility to maintain in the population high-quality fragment to be optimized via the mutation and crossover operation of the genetic algorithm. On the contrary, the VINA protocol selects high MW compounds in the first generation that have little to no space for meaningful chemical modifications, giving low priority to those smaller compounds that will have a lower number of interactions with the target, thus resulting in lower docking score. Overall, our IFP_{CS} protocol seems preferable in those cases where structural data is available in the form of protein-ligand complex structure, as is the case for a good number of targets nowadays, while the traditional protocol seems a valid choice in those cases where such structural information is missing.

A recent virtual screening campaign performed in our laboratory led to the identification of 7 novel fragment compounds that are ATP competitive CK1δ inhibitors[42]. Curious to see if our protocol would have been able to produce novel potential CK1δ that incorporate structural features of our 7 fragments, we performed a second *denovo* run, using the same conditions as for the benchmark one, except for the introduction in the starting library of those 7 fragment compounds. The same analysis performed on the benchmark runs showed that the performance of our IFP_{CS} scoring protocol is even better when the Autogrow protocol is seeded with high-quality fragments that have the right structural feature to interact with the target. Usually, in a typical fragment-based drug discovery

campaign, the identification of fragment binders either through virtual or experimental screening leads to the discovery of several potential starting points for the hit-to-lead fragment optimization phase. Our preliminary study showed that it is possible to obtain meaningful results even in those cases where the starting library is populated by fragments that are randomly selected and not specifically tuned for the target of choice, but it certainly benefits from the contamination of the starting library with fragments that are known binders, indicating that the application of the IFP_{CS} protocol could lead to some interesting results in those cases where the known binders constitute a bigger fraction of the starting library. Concerning this, this approach could be utilized to evaluate the competitiveness of newly found scaffolds with already existing ones, based on the simplicity to derive those scaffolds with common and feasible chemical reactions, therefore, producing a good number of derivatives with increased affinity for the target.

Conclusions

In the present work, we presented and benchmarked a custom version of the open-source Autogrow4 which implements an alternative scoring protocol based on the similarity between protein-ligand interaction fingerprint of query compounds compared to a crystal reference. To demonstrate the applicability of our protocol in a pharmaceutically relevant scenario, we tested its capability to generate compounds that have similar binding and structural features to known inhibitors of the protein kinase CK1δ, a protein that is involved in several neurodegenerative diseases such as Alzheimer's Disease, Parkinson's Disease, and Amyotrophic Lateral Sclerosis.

A benchmark *denovo* run and a lead-optimization one were both carried out to compare the performance of our IFP_{CS} scoring protocol against the traditional one implemented in the original version of the Autogrow code, using the same conditions as the one reported in the original publication by Spiegel et al. Compared to the traditional Autogrow protocol, which uses the default scoring function of the Vina docking software, our IFP_{CS} protocol was able to generate, on average, compounds that were bigger and more similar from a shape and electrostatic properties point of view to crystallographic ligands, while retaining the key protein-ligand interaction features required for the inhibition of CK1δ.

The custom Autogrow version used in this work, which implements our alternative IFP_{CS} scoring protocol, along with the JSON configuration files used for each run and a YAML file to reconstitute the Python environment to run the custom version of the code, are available in the Supplementary Material.

References

1. Knippschild, U. *et al.* The casein kinase 1 family: participation in multiple cellular processes in eukaryotes. *Cell. Signal.***17**, 675–689 (2005).

2. Knippschild, U. *et al.* The CK1 Family: Contribution to Cellular Stress Response and Its Role in Carcinogenesis. *Front. Oncol.***4**, (2014).

3. Xu, P. *et al.* Structure, regulation, and (patho-)physiological functions of the stress-induced protein kinase CK1 delta (CSNK1D). *Gene***715**, (2019).

4. Behrend, L. *et al.* Interaction of casein kinase 1 delta (CK1δ) with post-Golgi structures, microtubules and the spindle apparatus. *Eur. J. Cell Biol.***79**, 240–251 (2000).

5. Lee, H., Chen, R., Lee, Y., Yoo, S. & Lee, C. Essential roles of CKIδ and CKIε in the mammalian circadian clock. *Proc. Natl. Acad. Sci.***106**, 21359–21364 (2009).

6. Meggio, F., Perich, J. W., Reynolds, E. C. & Pinna, L. A. A synthetic beta-casein phosphopeptide and analogues as model substrates for casein kinase-1, a ubiquitous, phosphate directed protein kinase. *FEBS Lett.***283**, 303–306 (1991).

7. Bischof, J. *et al.* CK1δ Kinase Activity Is Modulated by Chk1-Mediated Phosphorylation. *PLoS One***8**, e68803 (2013).

8. Graves, P. R. & Roach, P. J. Role of COOH-terminal phosphorylation in the regulation of casein kinase I delta. *J. Biol. Chem.***270**, 21689–21694 (1995).

9. Milne, D. M., Looby, P. & Meek, D. W. Catalytic activity of protein kinase CK1 delta (casein kinase 1delta) is essential for its normal subcellular localization. *Exp. Cell Res.***263**, 43–54 (2001).

10. Longenecker, K. L., Roach, P. J. & Hurley, T. D. Crystallographic studies of casein kinase I delta toward a structural understanding of auto-inhibition. *Acta Crystallogr. D. Biol. Crystallogr.***54**, 473–475 (1998).

11. Hirner, H. *et al.* Impaired CK1 Delta Activity Attenuates SV40-Induced Cellular Transformation In Vitro and Mouse Mammary Carcinogenesis In Vivo. *PLoS One***7**, 29709 (2012).

12. Perez, D. I., Gil, C. & Martinez, A. Protein kinases CK1 and CK2 as new targets for neurodegenerative diseases. *Med. Res. Rev.***31**, 924–954 (2011).

13. Breijyeh, Z. & Karaman, R. Comprehensive Review on Alzheimer's Disease: Causes and Treatment. *Molecules***25**, 5789 (2020).

14. Lee, V. M. Y., Goedert, M. & Trojanowski, J. Q. Neurodegenerative tauopathies. *Annu. Rev. Neurosci.***24**, 1121–1159 (2001).

15. Selkoe, D. J. Alzheimer's disease: Genes, proteins, and therapy. *Physiol. Rev.***81**, 741–766 (2001).

16. J, H. & DJ, S. The amyloid hypothesis of Alzheimer's disease: progress and problems on the road to therapeutics. *Science***297**, 353–356 (2002).

17. Ghoshal, N. *et al.* A New Molecular Link between the Fibrillar and Granulovacuolar Lesions of Alzheimer's Disease. *Am. J. Pathol.***155**, 1163–1172 (1999).

18. Li, G., Yin, H. & Kuret, J. Casein Kinase 1 Delta Phosphorylates Tau and Disrupts Its Binding to Microtubules. *J. Biol. Chem.***279**, 15938–15945 (2004).

19. Schwab, C. *et al.* Casein kinase 1 delta is associated with pathological accumulation of tau in several neurodegenerative diseases. *Neurobiol. Aging***21**, 503–510 (2000).

20. Li, G., Yin, H. & Kuret, J. Casein kinase 1 delta phosphorylates tau and disrupts its binding to microtubules. *J. Biol. Chem.***279**, 15938–15945 (2004).

21. Okochi, M. *et al.* Constitutive Phosphorylation of the Parkinson's Disease Associated α-Synuclein. *J. Biol. Chem.***275**, 390–397 (2000).

22. Surmeier, D. J. Determinants of dopaminergic neuron loss in Parkinson's disease. *FEBS J.***285**, 3657–3668 (2018).

23. Kametani, F. *et al.* Identification of casein kinase-1 phosphorylation sites on TDP-43. *Biochem. Biophys. Res. Commun.***382**, 405–409 (2009).

24. 'Dementia', WHO, https://www.who.int/news-room/fact-sheets/detail/dementia.

25. Bollag, G. *et al.* Vemurafenib: the first drug approved for BRAF-mutant cancer. *Nat. Rev. Drug Discov.***11**, 873–886 (2012).

26. Schoepfer, J. *et al.* Discovery of Asciminib (ABL001), an Allosteric Inhibitor of the Tyrosine Kinase Activity of BCR-ABL1. *J. Med. Chem.***61**, 8120–8135 (2018).

27. Erlanson, D. A., Fesik, S. W., Hubbard, R. E., Jahnke, W. & Jhoti, H. Twenty years on: The impact of fragments on drug discovery. *Nat. Rev. Drug Discov.***15**, 605–619 (2016).

28. Jhoti, H., Williams, G., Rees, D. C. & Murray, C. W. The 'rule of three' for fragment-based drug discovery: Where are we now? *Nat. Rev. Drug Discov.***12**, 644 (2013).

29. Hajduk, P. J. Fragment-based drug design: How big is too big? *J. Med. Chem.***49**, 6972–6976 (2006).

30. Hall, R. J., Mortenson, P. N. & Murray, C. W. Efficient exploration of chemical space by fragment-based screening. *Prog. Biophys. Mol. Biol.***116**, 82–91 (2014).

31. Erlanson, D. A., McDowell, R. S. & O'Brien, T. Fragment-Based Drug Discovery. *J. Med. Chem.***47**, 3463–3482 (2004).

32. Murray, C. W. & Rees, D. C. The rise of fragment-based drug discovery. *Nat. Chem. 2009 13***1**, 187–192 (2009).

33. Schultes, S. *et al.* Ligand efficiency as a guide in fragment hit selection and optimization. *Drug Discov. Today Technol.***7**, e157–e162 (2010).

34. Rees, D. C., Congreve, M., Murray, C. W. & Carr, R. Fragment-based lead discovery. *Nat. Rev. Drug Discov.***3**, 660–72 (2004).

35. Bissaro, M., Sturlese, M. & Moro, S. The rise of molecular simulations in fragment-based drug design (FBDD): an overview. *Drug Discov. Today***25**, 1693–1701 (2020).

36. Böhm, H. J. The computer program LUDI: a new method for the de novo design of enzyme inhibitors. *J. Comput. Aided. Mol. Des.***6**, 61–78 (1992).

37. Eisen, M. B., Wiley, D. C., Karplus, M. & Hubbard, R. E. HOOK: A program for finding novel molecular architectures that satisfy the chemical and steric requirements of a macromolecule binding site. *Proteins Struct. Funct. Bioinforma.***19**, 199–221 (1994).

38. Lauri, G. & Bartlett, P. A. CAVEAT: A program to facilitate the design of organic molecules. *J. Comput. Mol. Des. 1994 81***8**, 51–66 (1994).

39. Maass, P., Schulz-Gasch, T., Stahl, M. & Rarey, M. Recore: a fast and versatile method for scaffold hopping based on small molecule crystal structure conformations. *J. Chem. Inf. Model.***47**, 390–399 (2007).

40. Spiegel, J. O. & Durrant, J. D. AutoGrow4: an open-source genetic algorithm for de novo drug design and lead optimization. *J. Cheminformatics 2020 121***12**, 1–16 (2020).

41. Trott, O. & Olson, A. J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading. *J. Comput. Chem.***31**, 455 (2010).

42. Bolcato, G. *et al.* A Computational Workflow for the Identification of Novel Fragments Acting as Inhibitors of the Activity of Protein Kinase CK1δ. *Int. J. Mol. Sci.***22**, 9741 (2021).

43. Chaput, L. & Mouawad, L. Efficient conformational sampling and weak scoring in docking programs? Strategy of the wisdom of crowds. *J. Cheminform.***9**, 1–18 (2017).

44. Chen, Y. C. Beware of docking! *Trends in Pharmacological Sciences* vol. 36 78–95 (2015).

45. de Souza Neto, L. R. *et al.*In silico Strategies to Support Fragment-to-Lead Optimization in Drug Discovery. *Front. Chem.***8**, 93 (2020).

46. Verdonk, M. L. *et al.* Docking performance of fragments and druglike compounds. *J. Med. Chem.***54**, 5422–5431 (2011).

47. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res.***28**, 235 (2000).

48. Mente, S. *et al.* Ligand-protein interactions of selective casein kinase 1δ inhibitors. *J. Med. Chem.***56**, 6819–6828 (2013).

49. Molecular Operating Environment (MOE), 2019.01; Chemical Computing Group ULC, 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, 2021. https://www.chemcomp.com/Research-Citing_MOE.htm.

50. D.A. Case et al. Amber 10, University of California, San Francisco.

51. QUACPAC 2.1.3.0: OpenEye Scientific Software, Santa Fe, NM. http://www.eyesopen.com.

52. Jones, G., Willett, P., Glen, R. C., Leach, A. R. & Taylor, R. Development and validation of a genetic algorithm for flexible docking 1 1Edited by F. E. Cohen. *J. Mol. Biol.***267**, 727–748 (2002).

53. Korb, O., Stützle, T. & Exner, T. E. PLANTS: Application of Ant Colony Optimization to Structure-Based Drug Design. 247–258 (2006) doi:10.1007/11839088_22.

54. Houston, D. R. & Walkinshaw, M. D. Consensus Docking: Improving the Reliability of Docking in a Virtual Screening Context. *J. Chem. Inf. Model.***53**, (2013).

55. AutoGrow4 – Durrant Lab. https://durrantlab.pitt.edu/autogrow4/.

56. Eberhardt, J., Santos-Martins, D., Tillack, A. F. & Forli, S. AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings. *J. Chem. Inf. Model.***61**, 3891–3898 (2021).

57. Wójcikowski, M., Zielenkiewicz, P. & Siedlecki, P. Open Drug Discovery Toolkit (ODDT): a new open-source player in the drug discovery field. *J. Cheminform.***7**, 26 (2015).

58. Li, D. & Ji, B. Protein conformational transitions coupling with ligand interactions: Simulations from molecules to medicine. *Med. Nov. Technol. Devices***3**, 100026 (2019).

59. Csermely, P., Palotai, R. & Nussinov, R. Induced fit, conformational selection and independent dynamic segments: an extended view of binding events. *Trends Biochem. Sci.***35**, 539–546 (2010).

60. Wierbowski, S. D., Wingert, B. M., Zheng, J. & Camacho, C. J. Cross-docking benchmark for automated pose and ranking prediction of ligand binding. *Protein Sci.***29**, 298–305 (2020).

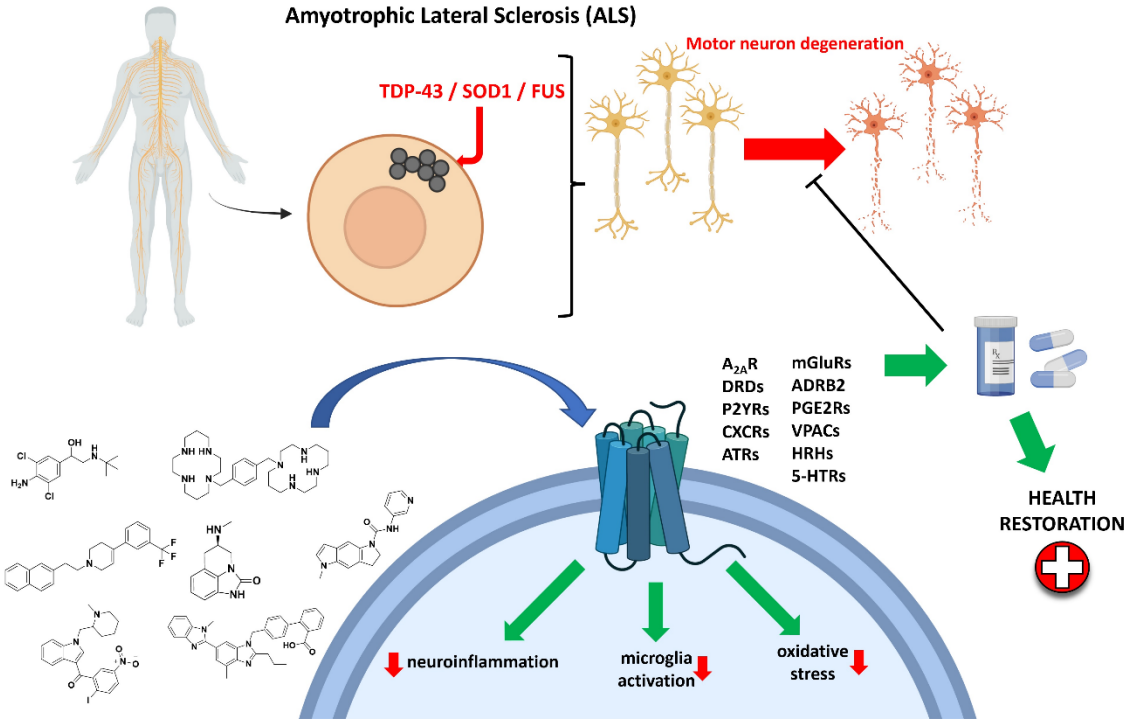
61. Cescon, E. *et al.* Scaffold Repurposing of in-House Chemical Library toward the Identification of New Casein Kinase 1 δ Inhibitors. *ACS Med. Chem. Lett.***11**, (2020).

62. Peach, M. L. & Nicklaus, M. C. Combining docking with pharmacophore filtering for improved virtual screening. *J. Cheminform.***1**, 6 (2009).

The Multifaceted Role of GPCRs in Amyotrophic Lateral Sclerosis: A New Therapeutic Perspective?

Davide Bassani, **Matteo Pavan**, Stephanie Federico, Giampiero Spalluto, Mattia Sturlese and Stefano Moro

Bassani, D. *et al.* The Multifaceted Role of GPCRs in Amyotrophic Lateral Sclerosis: A New Therapeutic Perspective? *International Journal of Molecular Sciences 2022, Vol. 23, Page 4504* **23**, 4504 (2022).



Abstract

Amyotrophic lateral sclerosis (ALS) is a degenerating disease involving the motor neurons, which causes a progressive loss of movement ability, usually leading to death within 2 to 5 years from the diagnosis. Much effort has been put into research for an effective therapy for its eradication, but still, no cure is available. The only two drugs approved for this pathology, Riluzole and Edaravone, are only able to slow down the inevitable disease progression. As assessed in the literature, drug targets such as protein kinases have already been extensively examined as potential drug targets for ALS, with some molecules already in clinical trials. Here, we focus on the involvement of another very important and studied class of biological entities, G protein-coupled receptors (GPCRs), in the onset and progression of ALS. This work aims to give an overview of what has been already

discovered on the topic, providing useful information and insights that can be used by scientists all around the world who are putting efforts into the fight against this very important neurodegenerating disease.

1. Introduction

Amyotrophic Lateral Sclerosis (ALS, also referred to as “motor neuron disease”) indicates a clinical situation in which the motor neurons of patients undergo a progressive loss in their function and number[1]. This type of neuronal cell, whose cell body is localized in the motor cortex, the brainstem, and the spinal cord, is responsible for the innervation and the control of muscle fibers, essential for voluntary muscle contraction[2]. Their loss has very important consequences on the patient’s life, firstly impairing the ability to chew and walk, then to speak and to move, until even the ability to breath is affected, leading, after 2–5 years, to death due to respiratory failure[3]. ALS can be classified into two main types, “sporadic ALS” (the great majority of all cases), which has no known cause and typically has its onset between the ages of 58 and 63 years, and “familial ALS” (about 5–10% of cases), which is linked to genetic factors, and has its onset between the ages of 47 and 52 years[4].

In both scenarios, the pathology starts with the manifestation of muscle weakness and atrophy, with methods and timing very variable based on the patient and on the parts of the motor neurons that are affected first[5]. Indeed, a classification of the onset of the pathology can be made with reference to the site of its onset. For two-thirds of patients, the limb muscles are affected first (“spinal ALS”), with manifestations mainly in the distal muscles of the dominant hand for the upper limb and in the hamstrings for the lower limb. For the greater part of the remaining patients, the bulbar muscles represent the onset site (“bulbar ALS”), and in this case, dysphagia and chewing problems represent the first manifestations of the pathology[6].

The next steps of the disease involve the progressive spreading of the neurodegeneration process to the unaffected motor neurons, causing an increasing worsening in the patient’s daily life, making activities such as eating and walking continuously more difficult and leading to their complete loss. The final and worst clinical scenario has its onset when the respiratory function is significantly affected, progressively increasing the risk of respiratory failure, which is the main cause of death due to ALS[7].

Even if much effort has been made among both academic and industrial scientific groups, no cure has yet emerged for ALS. Riluzole[8] and the recently FDA-approved drug Edaravone[9] (both represented in Figure 1) constitute the only two small molecules used for the ALS treatment, and only succeed in slowing down the disease's progression[10].

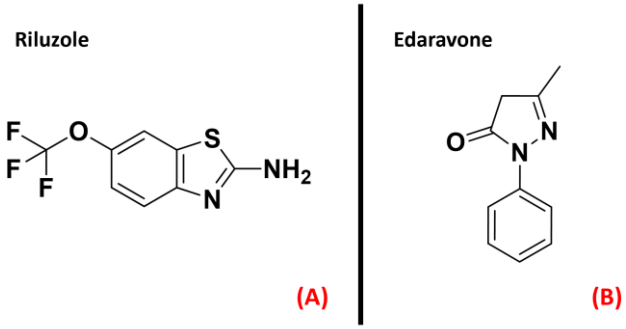


Figure 1. The chemical structures of Riluzole (**A**) and Edaravone (**B**), which are the only two small-molecule drugs approved currently for ALS treatment.

Such a neurodegenerating process, affecting 1.75–3 people per 100,000[11], has, in the great majority of cases, no known cause[12], making it even harder to design a therapy for this disease. From a biochemical point of view, the hallmark of ALS is considered to be the presence of inclusion bodies in the cytoplasm of motor neurons. These aggregates are formed by the TAR DNA-binding protein 43 (TDP-43)[13], a protein involved in several important physiological functions such as DNA repair, splicing, and transcriptional regulation. Even if its main localization site is the nucleus, processes such as its hyperphosphorylation or the mutation of its gene (TARDBP) lead to its aggregation in the cytoplasm[14]. This mislocalization directly causes the dysregulation of several cellular events related to RNA metabolism, DNA replication, and oxidative stress management, leading to the loss of the motor neurons affected[15]. Other molecular targets that have been demonstrated to be important for ALS onset and progression are superoxide dismutase (SOD1)[16] and DNA/RNA-binding protein FUS/TLS (FUused in Sarcoma/Translocated in LipoSarcoma, also called "FUS")[17], which appear to be mutated in the patients.

The knowledge that hyperphosphorylation of TDP-43 is one of the main processes leading to its aggregation has led the scientific community to devote some effort to identifying

the protein kinases responsible for such processes, in order to find proper inhibitors for such species[18]. Recent work by Guo et al. goes deep in the examination of the involvement of kinases in ALS progression, enucleating species such as CK1, ERK, GSK3 β, and JAK3 as promising targets for the treatment of this neurodegenerating disease[19]. Another article by Palomo et al. gives an exhaustive panoramic view of the protein kinase inhibitors currently in clinical trials for ALS treatment[20]. Riluzole has proven to increase life expectancy by about 2–3 months[21], and even if its main target still remains the NMDA receptor, recent work by Bissaro et al. suggested that this mechanism could be due to its action on the delta isoform of CK1[22].

Many molecular candidates (both new chemical entities and compounds coming from repurposing strategies) are nowadays in clinical trials for ALS[23], acting on different biological pathways, with the common aim being to restore the neuronal health status in the affected patients, possibly trying to go in the direction to find a proper cure for this pathology[24]. A comprehensive list of the potential small-molecule drugs now being evaluated by the FDA in clinical phases is reported in Table 1.

Molecule	**Target/Mechanism**	**Developer**	**Clinical Phase**
Ibudilast	Macrophage migration inhibitory factor inhibitor	MediciNova	Phase II/III
Prosetin	Mitogen-activated protein kinase inhibitor	ProJenX	Phase I
Sotuletinib	Macrophage colony-stimulating factor receptor antagonist	Novartis	Phase II
EPI 589	NAD(P)H dehydrogenase modulator	PTC Therapeutics	Phase II
DNL 343	Eukaryotic initiation factor 2b stimulant	Denali Therapeutics Inc	Phase I
Celecoxib/ciprofloxacin	Cyclo-oxygenase 2 inhibitors/DNA gyrase inhibitors	NeuroSense Therapeutics	Phase I
Fingolimod	Apoptosis stimulant and immunosuppressant	ALS Therapy Development Institute	Phase II
Trehalose	Autophagy stimulant and protein aggregation inhibitor	Massachusetts General Hospital	Phase II/III
Sodium cromoglicate	Glial cell modulator and mast cell stabilizer	AZTherapies	Phase II
Dexpramipexole	Antioxidant and apoptosis inhibitor	Knopp Biosciences	Phase II
Masitinib	Tyrosine kinase inhibitor	AB Science	Phase III
NP 001	Macrophage modulator	Neuvivo	Phase II

Fasudil	Rho-associated kinase inhibitor and vasodilatator	Woolsey Pharmaceuticals	Phase II
Levosimendan	Calcium-sensitising phosphodiesterase inhibitor and potassium channel agonist	Orion	Phase III
Apilimoddimesylate	Interleukin 12 inhibitor and interleukin 23 inhibitor	AI Therapeutics	Phase II
Verdiperstat	Peroxidase inhibitor	Biohaven Pharmaceuticals	Phase II/III
Pridopidine	Sigma-1 receptor agonist	Massachusetts General Hospital, Prilenia Therapeutics	Phase II/III
Triheptanoin	Triglyceride replacement agent	Ultragenyx Pharmaceutical	Phase I/II
Reldesemtiv	Troponin stimulant	Cytokinetics	Phase III
BIIB 100	Exportin-1 protein inhibitor	Biogen	Phase I
AGX 201	Histamine receptor modulator	AgoneX Biopharmaceuticals	Phase I
Ranolazine extended release	Sodium channel antagonist	Gilead Sciences	Phase II
GDC 0134	Mitogen-activated protein kinase 12 inhibitor	Genentech	Phase I
NPT520 34	Phosphatidylinositol 3 kinase modulator	Neuropore Therapies	Phase I

Table 1. Table reporting the different small molecules currently in FDA clinical trials for ALS treatment (updated 13 April 2022).

Even if some effort has been directed toward trying to highlight the role of protein kinases in ALS progression, this has not been recently or extensively done with respect to G-protein coupled receptors, biological actors which have been demonstrated to be detrimental to neuronal and physiological conditions. It is important to remember that ALS is a non-cell-autonomous disease, which means that the neuronal damage characterizing the pathology is caused by aberrant processes also happening outside the neurons themselves. Indeed, ALS progression has been demonstrated to be strongly related to glial cell dysregulation (mainly microglia and astrocytes)[25]. GPCRs are very widely expressed proteins in the human organism[26], and so a beneficial effect could also be obtained by targeting extraneuronal receptors, which could trigger biological processes that, in the overall scenario, could mitigate if not reverse the disease progression.

GPCRs are membrane receptors and constitute one of the main protein families encoded by human genes, with more than 800 members already identified[27], divided into six different classes (identified alphabetically with letters from “A” to “F”) based on their similarities in sequence and function. They all share a common architecture formed of a seven-α-helix transmembrane domain (usually referred to as “7-TM”), an extracellular N-terminal domain, and an intracellular C-terminal domain. These proteins exert their roles by coupling with an intracellular messenger called “heterotrimeric G protein”, which is formed by α,β, and γ subunits, and interacts with different intracellular partners based on its type. The α subunit is displaced from the βγ-complex upon GPCR–ligand binding, and its fate depends on its Gα family belonging. Indeed, activated $G_{i/0}$ proteins inhibit adenylyl cyclase (AC), reducing the production of the second messenger cyclic adenosine monophosphate (cAMP); G_s, conversely, activates adenylyl cyclase, and $G_q\alpha$ subunits activate phospholipase C (PLC), leading to an increase in Ca^{2+} influx in the cytoplasm[28]. The physiological roles of GPCRs include homeostasis modulation, mood balancing, immune system regulation, neuronal plasticity, and many more[29]. The goal of the present work is to give a panoramic view of the GPCRs which have been linked to ALS onset and progression, presenting what has already been done to modulate their action, and highlighting new potential therapeutical scenarios. Table 2 summarizes the outcomes of our study, listing the GPCR targets that will be discussed and highlighting the new possible paths that can be taken in order to exploit their therapeutic potential for ALS. Our work will be beneficial for all of the scientists who are dedicating their knowledge and efforts to the eradication of ALS.

Receptor/Receptor Family	**Cellular Expression**	**Potential for ALS Treatment**	**References**
Adenosine receptors	Circulatory, immune, respiratory, and nervous systems	Ambiguous	[32–39]
Purinergic receptors P2Y	Almost all human tissues	Antagonism	[44–49]
Chemokine receptors	Predominantly on leukocytes surface	CXCR3, CXCR4, and CCR2 Antagonism	[53–56]
Angiotensin II receptors	Adrenal cortex, kidneys, vascular and cardiac muscles, nervous system	AT_1 Antagonism	[64–66]
Dopamine receptors	Arteries, heart, kidneys, CNS	D2R Agonism	[71–78]
Serotonin receptors	Almost all human tissues	Ambiguous	[84–88]
GPR17 receptor	CNS, kidneys, heart	Antagonism	[95–102]

Adrenergic receptor β_2	GI tract, respiratory system, blood vessels, pancreas, nervous system	Agonism	[106,107]
Histamine receptors	GI tract, circulatory, immune, and nervous systems.	Ambiguous	[111–114]
Cannabinoid receptors	CNS and immune system	CB_2 agonism	[119–122]
Prostaglandin E_2 receptor	GI tract, kidneys, reproductive, skeletal, immune, and nervous systems.	Ambiguous	[136–138]
Vasoactive intestinal peptide receptors	Almost all human tissues	Agonism	[145–149]
Metabotropic glutamate receptors	Nervous system	mGluR I antagonism/ mGluR II and mGluR III agonism	[152–156]

Table 2. Table summarizing the evidence about the therapeutic potential of the GPCRs examined in this article for ALS treatment.

2. GPCRs Involved in ALS

2.1. Puringergic Receptors P2Y and Adenosine Receptor A_{2A}AR

Purinergic receptors are a peculiar class of membrane receptors, sensitive to a wide series of purinergic ligands such as ATP, ADP, UTP, UDP, UDP-glucose, and adenosine. The kind of molecules interacting with them defines their classification into one of the three subfamilies forming this class. The first group, called “P1 receptors”, is formed of GPCRs activated upon adenosine binding (and for these reasons are also known as “adenosine receptors”), while “P2Y receptors” are GPCRs that can bind to ATP, UDP, and their diphosphate analogs ADP and UDP (with the addition of UDP-glucose). The last subfamily, named “P2X receptors”, are ligand-gated ion channels exclusively sensitive to ATP[30]. Being the P2X family not formed by GPCRs, our evaluations will focus on the first two families of receptors.

The “P1” subfamily, more commonly referred to as adenosine receptors (ARs), is a group of purinergic class A GPCRs divided into four subtypes, A_1AR, A_{2A}AR, A_{2B}AR, and A_3AR, each involved in many different physiological processes. While the functions of A_{2B}AR and A_{3A}R are mainly related to the circulatory, immune and respiratory systems, the A_1AR and A_{2A}AR proteins are importantly present in the central nervous system (CNS)[31]. Moreover, the A_1AR and the A_{2A}AR receptors have been demonstrated to play a crucial role in

neuroprotection, neuronal survival, and neuroinflammation[32]. A study from Vincenzi et al. reported an upregulation of A_{2A}AR receptors in the lymphocytes of people affected by ALS[33], while Yoshida et al. measured adenosine levels in the cerebrospinal fluid of ALS patients, finding out that these were significantly higher with respect to the control subjects[34]. Unexpectedly, treatment with the A_{2A}AR antagonist caffeine (Figure 2, panel A), which is usually referred to as a protective agent against Alzheimer's Disease (AD) and Parkinson's Disease (PD), was demonstrated to shorten the survival of ALS-affected $SOD1^{G93A}$ mice, a well-known experimental model for ALS (Potenza et al.)[35], even if some explanation for this phenomenon can be provided by the non-selectivity of caffeine[36]. Indeed, Ng et al. showed that suppression of A_{2A}AR signaling delays the progression of ALS in the same $SOD1^{G93A}$ mouse model[37]. This is in accordance with the study of Mojsilovic-Petrovic et al., which demonstrated that A_{2A}AR inhibitors (such as the non-selective enprofylline (Figure 2, panel B), and the A_{2A}AR-selective KW-6002 (Figure 2, panel C), also called Istradefylline) protect motor neurons from toxic insult, highlighting the beneficial effects of such activity for ALS patients[38].

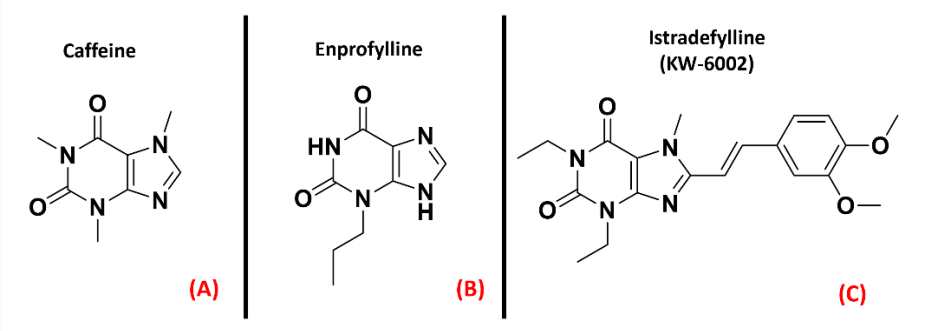


Figure 2. The chemical structures of the non-selective adenosine receptors antagonists caffeine (**A**) and Enprofylline (**B**), and the selective A_{2A}AR antagonist Istradefylline (**C**).

Despite all this evidence, a study from Liu et al. showed that A_{2A}AR activation, suppressing AMPK activation, suppressed TDP-43 mislocalization[39]. The multifactorial nature of ALS makes it very difficult to define sharply whether agonism or antagonism of the A_{2A}AR receptor has the best risk/benefit ratio, but the literature clearly defines this GPCR (represented in Figure 3) as one of the promising targets for ALS treatment.

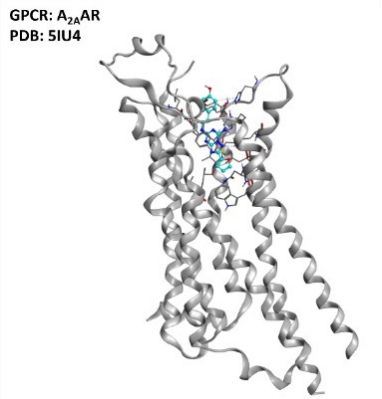


Figure 3. Representation of the structure of the $A_{2A}AR$ receptor (sourced from the Protein Data Bank[40], PDB code: 5IU4[41], method: X-ray diffraction, resolution: 1.72 Å). The image was created and rendered with the Molecular Operating Environment (MOE) suite[42].

The second subfamily of purinergic receptors, "P2Y", is present in a great variety of human tissues, but their main biological roles are identifiable in blood clotting, vasodilatation, and immune response[43]. The P2Y family comprises eight different isoforms, among which $P2Y_{12}$ has gained the interest of the scientific community for its role in neuroinflammation, as addressed by Morillas et al.[44] and Amadio et al.[45].

Jacobson et al. recently highlighted the proinflammatory effect of P2Y agonists, reporting that antagonizing this class of GPCRs could be considered a way of treating inflammatory conditions[46]. Even if inhibitors of the $P2Y_{12}$ isoform are already marketed as antiplatelet drugs (e.g., Clopidogrel, Prasugrel, Ticagrelor, all represented in Figure 4), Jacobson et al. pointed out that there is a lack of selective and versatile P2Y ligands for each subtype, meaning that the drug discovery process is still very active in this specific field.

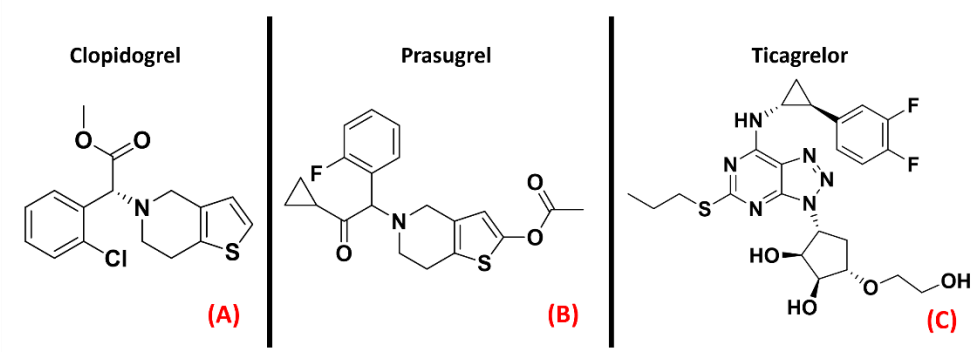


Figure 4. The chemical structures of the selective and irreversible $P2Y_{12}$ receptor antagonists Clopidogrel (**A**) and Prasugrel (**B**). Ticagrelor, a selective, reversible, allosteric $P2Y_{12}$ receptor antagonist, is also reported (**C**).

Specifically, D'Ambrosi et al. reported an upregulation of $P2Y_6$ receptors in the microglia SOD1 mutant models of ALS, also remembering that this phenomenon is associated with brain damage[47]. Moreover, $P2Y_{12}$ (represented in Figure 5) is upregulated in spinal cord microglia upon nerve injury, as pointed out by Kobayashi et al.[48]. Converging information is provided by a study from Moore et al.[49], further confirming $P2Y_{12}$ as a potential target for modulating neuroinflammation and neuronal damage. The data currently available help in suggesting the practical possibility of ALS regulation through purinergic receptor modulation, and this will be realizable as soon as proper inhibitors can be designed, potentially avoiding the non-desired antiplatelet effect of these molecules.

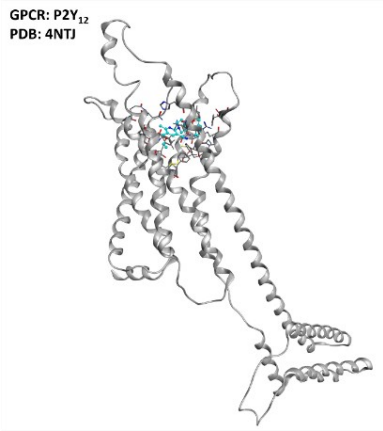


Figure 5. Representation of the structure of the $P2Y_{12}$ receptor (sourced from the Protein Data Bank, PDB code: 4NTJ[50], method: X-ray diffraction, resolution: 2.62 Å). The image was created and rendered with MOE.

2.2. Chemokine Receptors

Chemokine receptors constitute a group of about 20 classes of GPCRs found mainly on the surface of leukocytes, which respond to specific ligands to control chemotaxis. The ligands for these proteins are called chemokines and are a peculiar kind of cytokine used for inducing a directional movement of certain types of cells, such as epithelial and immune ones. The chemokines, as well as the receptors they act on, can be divided into four families, namely CC (e.g., chemokine CCL4), CXC (e.g., chemokine CXCL8, also known

as IL-8), XC (e.g., XCL1), and CX3C (of which the only member today is CX3CL1, also called neurotactin)[51]. The activation of chemokine receptors leads to Ca^{2+} influx and cell mobilization[52]. Several of these receptors are important in the progression of motor neuron damage. La Cognata et al. highlighted an upregulation of CXCR2 in both sporadic ALS patients and $SOD1^{G93A}$ mice, showing that treating the mouse models with the CXCR2 allosteric inhibitor Reparixin (Figure 6, panel B), the neuromuscular function of the subjects was improved[53]. Another interesting paper published by Rabinovich-Nikitin et al. highlighted the benefits in terms of lifespan and motor function obtained on $SOD1^{G93A}$ mouse models through the administration of the CXCR4 antagonist AMD3100 (also known as "Plerixafor", Figure 6, panel A)[54].

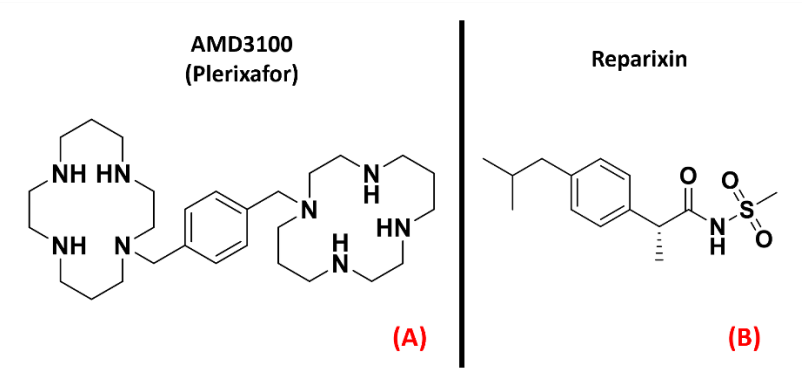


Figure 6. The chemical structures of the CXCR4 antagonist AMD3100 (also known as "Plerixafor", (**A**) and the inhibitor of CXCR1 and CXCR2 known as Reparixin (**B**).

Several scientific works have reported an increase in circulating chemokines and cytokines in ALS patients, as recently detailed by Liu et al.[55], and the upregulation of chemokine receptors CXCR3, CXCR4, and CCR2 was also highlighted in the pathology of interest by Perner et al.[56], who also proposed CXCR3 and its ligands as possible therapeutic targets for ALS. These scientific works converge in addressing chemokine receptor modulation as a possibility for ALS treatment, focusing on the antagonism of certain isoforms. The three-dimensional structures of CXCR2, CXCR3, and CXCR4 are represented in Figure 7.

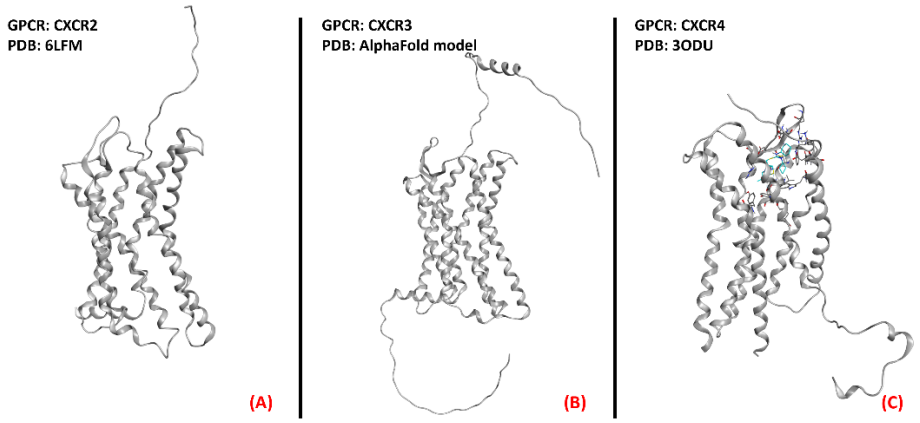


Figure 7. Representation of the three-dimensional structures of the chemokine receptors that could be considered for ALS treatment.CXCR2 **(A)** receptor (sourced from the Protein Data Bank, PDB code: 6LFM[57], method: cryo-EM, resolution: 2.50 Å), CXCR3 (**B**) (with no experimentally resolved structure available, the model from the AlphaFold[58] database is presented), and CXCR4 **(C)** receptor (sourced from the Protein Data Bank, PDB code: 3ODU[59], method: X-ray diffraction, resolution: 2.50 Å). The images were created and rendered with MOE.

2.3. Angiotensin II Receptors (ATRs)

Angiotensin II receptors (ATRs) are a group of GPCRs that have gained fame for their importance in the therapy of hypertension. Indeed, their main physiological role is related to the renin–angiotensin–aldosterone system, one of the main physiological pathways for blood pressure regulation and fluid and electrolyte balance[60]. Briefly, the peptide hormone angiotensinogen is secreted by the liver and cleaved by renin to form angiotensin I, which is then converted to angiotensin II by the angiotensin-converting enzyme (ACE), produced by the lungs. Angiotensin II acts on its receptors and modulates several processes, such as aldosterone secretion (in the adrenal glands), water and sodium retention (in the kidneys), sanguine pressure, and vasopressin production (in the CNS)[61]. To treat hypertension, many efforts have been directed towards the creation of drugs acting as ATRs inhibitors. The most famous drug family designed for this purpose is represented by the “Sartans”, which selectively bind to the first isoform of angiotensin receptors[62]. Indeed, ATRs can be divided into four isoforms, AT_1, AT_2, AT_3, and AT_4. While the latter two are still in the early stages of research, the first two isoforms have been more deeply characterized. AT_1 is mainly found in blood vessels, heart, kidney, brain, and adrenal cortex, mediating vasoconstrictive effects[60]. AT_2 receptors are more concentrated

in the fetus and neonate, and their functions are more strongly related to neuronal development and excitability[63]. A study by Kawajiri et al. highlighted a reduction in angiotensin II levels in the CSF coming from ALS patients, reporting two opposite consequences: the reduction of protection and repair mediated by AT_2, on the one hand, and the reduction of oxidative stress due to AT_1 on the other. Indeed, Kawajiri et al. hypothesized that angiotensin II could be downregulated in CSF of ALS patients as a protective reaction, avoiding excessive activation of AT_1[64]. The benefits of AT_1 antagonism have also been underlined by Iwasaki et al., who reported evidence of the neurotrophic effects on spinal motor neurons of the drug Olmesartan (Figure 8, panel A), specifically referring to its potential application in ALS[65]. Furthermore, an article by Mammana et al. highlighted the AT_1 antagonism-mediated neuroprotective effects of Telmisartan (Figure 8, panel B), also outlining the decrease in neuronal injury and microglial activation caused by it[66].

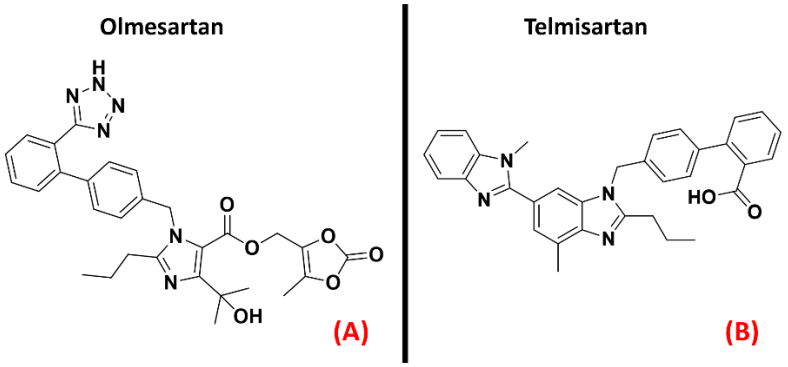


Figure 8. The chemical structures of the AT_1 antagonists Olmesartan (**A**) and Telmisartan (**B**).

Summing up the information obtainable from the literature, AT_1 inhibition could be examined as a potential new therapeutic method of fighting ALS conditions. Both AT_1 and AT_2 receptors are represented in Figure 9 below.

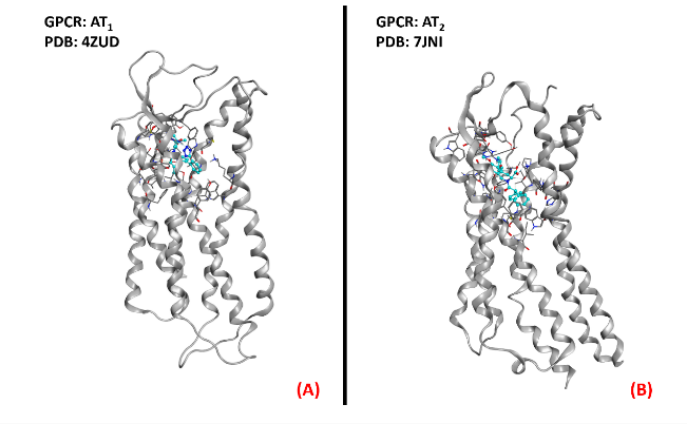


Figure 9. Representation of the three-dimensional structures of the angiotensin II receptors that could be considered for ALS treatment. AT_1 receptor **(A)** (sourced from the Protein Data Bank, PDB code: 4ZUD[67], method: X-ray diffraction, resolution: 2.80 Å) and AT_2 receptor **(B)** (sourced from the Protein Data Bank, PDB code: 7JNI[68], method: X-ray diffraction, resolution: 3.00 Å). The images were created and rendered with MOE.

2.4. Dopamine Receptors

Dopamine receptors are among the most important and widely studied G-protein coupled receptors, mainly for their important physiological roles in neurotransmission. This class of GPCRs is divided into five isoforms, which are separated into two classes. The first, also called the "D1-like family", comprises the D1R and D5R receptors, which are coupled to a G_s protein responsible for adenylyl cyclase activation upon binding. The second family, also known as the "D2-like" family, comprises the D2R, D3R, and D4R proteins, all coupled with a G_i protein with inhibitory activity on adenylyl cyclase. Dopamine receptors are localized in different peripheral parts of the organism, such as arteries, heart, and kidneys, but their activities much more determinant within the CNS. Indeed, dopamine is the main neurotransmitter involved in the reward system, and its signaling is of crucial importance for processes such as cognition, memory, and motor control[69]. Dysregulation of the dopaminergic system in the brain represents the main cause of very important diseases such as schizophrenia, attention deficit hyperactivity disorder (ADHD), and Parkinson's disease[70]. Several articles highlight a correlation between dopamine signaling and ALS development[71]. We have previously reported what was assessed by Liu et al. regarding the protective effect of the A_{2A} receptor on TDP-43 mislocalization[39]. A recent study by Lai et al. details how this beneficial activity can be blocked by D2R activation[72]. Despite this,

D2R was also identified as important for the modulation of motor neuron excitability by Huang et al.[73]

Fujimori et al. showed how the treatment with Ropinirole (Figure 10, panel A), an agonist for receptors D2R, D3R, and D4R mainly used for Parkinson's disease, has neuroprotective effects in ALS models[74]. Additionally, D2R agonists such as Bromocriptine and Sumanirole (both represented in Figure 10, panels B and C, respectively) were tested by Huang et al., who reported that the final effect of such activity on ALS models was an increase in motor neuron survival[75]. Another agonist for the "D2-like family" of dopamine receptors is the R(+) enantiomer of the Parkinson's disease drug Pramipexole, known as Dexpramipexole (Figure 10, panel D).

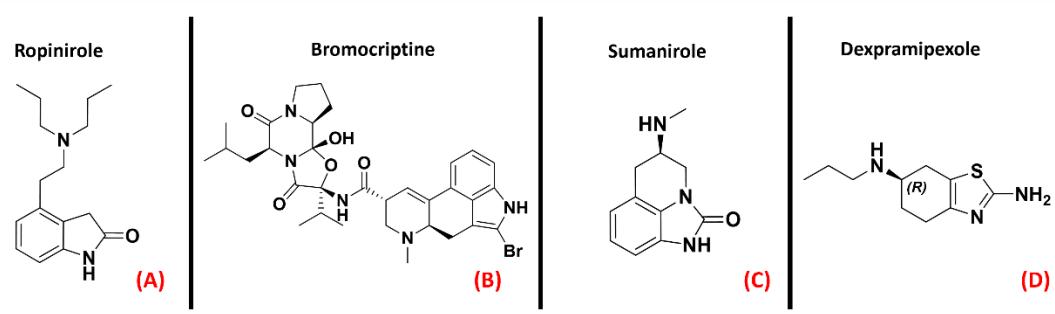


Figure 10. The chemical structures of the dopamine receptor agonists Ropinirole (**A**), which has an affinity for D2R, D3R, and D4R, Bromocriptine (**B**), also non-selective with an affinity for D2R, D3R, and D4R, and Sumanirole (**C**), selective for D2R. (**D**) Chemical structure of Dexpramipexole (its neuroprotective effects are attributed to dopaminergic-independent activities).

Even if Pramipexole is a powerful agonist of D2R, D3R, and D4R (all depicted in Figure 11)[76], its R(+) enantiomer has a very low affinity for dopamine receptors, so its neuroprotective effects have to be due to a non-dopaminergic action[77]. This molecule has been considered a promising candidate for ALS conditions[78]. After the phase III clinical trial, however, its development in Europe was discontinued[79].

D2R is still a very relevant target for Parkinson's disease treatment, but the presented literature concords in considering it also a protein of high therapeutic potential for treating ALS.

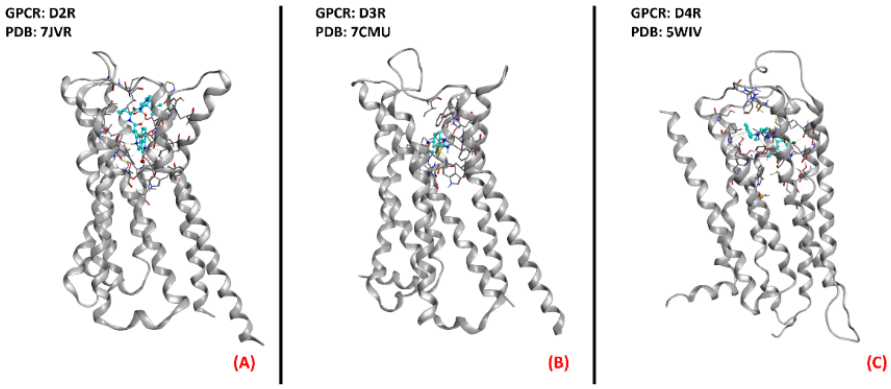


Figure 11. Representation of the three-dimensional structures of the dopamine receptors that could be considered for ALS treatment. (**A**) D2R (sourced the Protein Data Bank, PDB code: 7JVR[80], method: cryo-EM, resolution: 2.80 Å), (**B**) D3R (sourced from the Protein Data Bank, PDB code: 7CMU[81], method: cryo-EM, resolution: 3.00 Å), and (**C**) D4R (sourced from the Protein Data Bank, PDB code: 5WIV[82], method: X-ray diffraction, resolution: 2.14 Å). All of the images were created and rendered with MOE.

2.5. Serotonin (5-HT) Receptors

Serotonin (also called 5-hydroxytryptamine, or 5-HT) receptors represent one of the most populated subfamilies of class A GPCRs, consisting of 13 G-protein coupled receptor isoforms (5-HT_{1A}, 5-HT_{1B}, 5-HT_{1D}, 5-HT_{1E}, 5-HT_{1F}, 5-HT_{2A}, 5-HT_{2B}, 5-HT_{2C}, 5-HT_{4}, 5-HT_{5A}, 5-HT_{5B}, 5-HT_{6}, and 5-HT_{7}) distributed throughout the entire human organism, and a cation channel (5-HT_{3}), mainly involved in gastrointestinal motility[83]. It is also interesting to note that almost all of these isoforms are present in the CNS[84]. Concerning ALS, the serotonin receptor which has gained the greatest popularity is 5-HT_{2B}. An article from Oussini et al. reported that the activity of this biological entity could limit the degeneration of spinal cord mononuclear phagocytes, which is a process typical of neurodegenerative diseases. This article highlighted that the ablation of the 5-HT_{2B} gene resulted in an acceleration of ALS progression in mutant SOD1 mouse models. Indeed, they showed that the administration of a 5-HT_{2B} selective antagonist (SB204741, Figure 12, panel A) caused an important reduction in microglia viability, while treatment with the agonist BW723C86 (Figure 12, panel B) induced an increase in viability[85]. Another work by Dentel et al. reported that the spasticity associated with ALS progression could be strongly alleviated by the administration of inverse agonists of 5-$HT_{2B/C}$ such as SB206553 and

Cyproheptadine (both depicted in Figure 12, panels C and D, respectively)[86]. A recent article by Arnoux et al., on the other hand, highlighted the lack of beneficial effects when ALS-affected SOD1^{G86R} mutants were treated with the 5-HT$_{2B}$ agonist BW723C86[87]. The main factor that has always limited the development of 5-HT$_{2B}$ agonists is their inherent cardiotoxicity, which can determine valvular heart disease[88]. Two randomized, double-blind, placebo-controlled multicenter studies (phase III) were conducted in 2004 by Meininger et al. to evaluate the potential benefits of Xaliproden (Figure 12, panel E), a 5-HT$_{1A}$ receptor agonist with neuroprotective effects, in ALS patients.

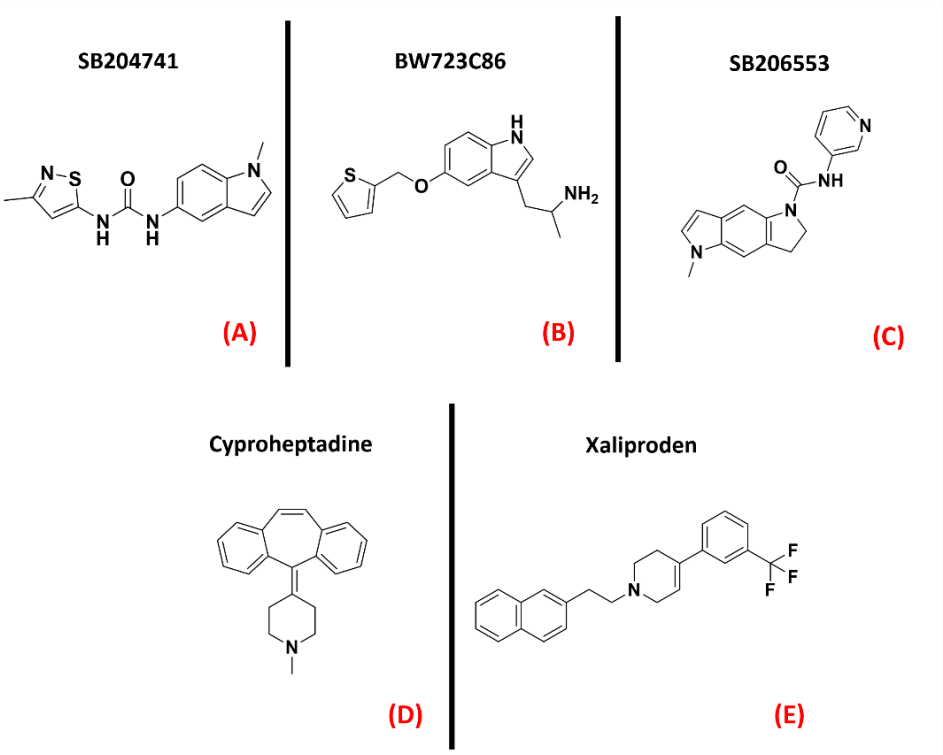


Figure 12. The chemical structures of the serotonin receptor modulators are treated in this review. The selective 5-HT$_{2B}$ antagonist SB204741 (**A**), the 5-HT$_{2B}$ agonist BW723C86 (**B**). (**C**) Structure of the 5-HT$_{2B/C}$ mixed inverse agonist SB206553; (**D**) chemical structure of Cyproheptadine, a non-selective mixed serotonin receptor antagonist (which is an inverse agonist of 5-HT$_{2B}$). (**E**) Structure of the 5-HT$_{1A}$ receptor agonist Xaliproden.

Despite the promising outcomes of the prior experiments[89], no effective slowing down in the progression of the pathology was evidenced. Even if some limitations have been encountered in serotoninergic modulation for motor neuron disease, the 5-HT receptors

have been demonstrated to be targets of relevance in the ALS scenario. The three-dimensional structures of 5-HT_{1A}, 5-HT_{2B}, and 5-HT_{2C} are represented in Figure 13.

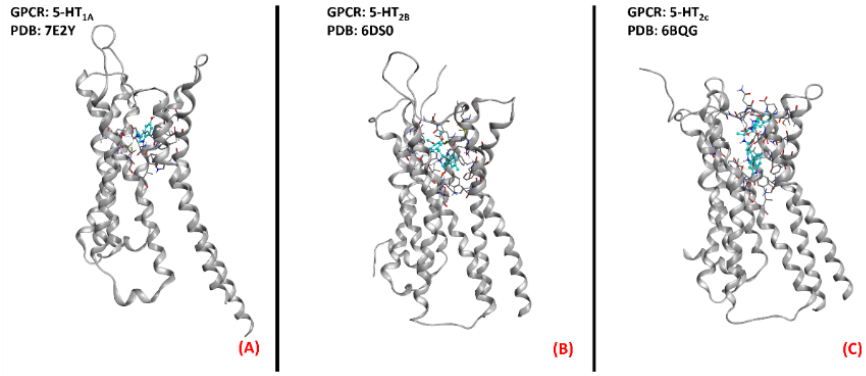


Figure 13. Three-dimensional structures of the serotonin receptors that could be considered for ALS treatment. (**A**) 5-HT_{1A} (sourced from the Protein Data Bank, PDB code: 7E2Y[90], method: cryo-EM,resolution: 3.00 Å), (**B**) 5-HT_{2B} (sourced from the Protein Data Bank, PDB code: 6DS0[91], method: X-ray diffraction, resolution: 3.19 Å), and (**C**) 5-HT_{2C} (sourced from the Protein Data Bank, PDB code: 6BQG[92], method: X-ray diffraction, resolution: 3.00 Å). All of the images were created and rendered with MOE.

2.6. GPR17 Receptor

GPR17 (also known as "uracil nucleotide/cysteinyl leukotriene receptor") is a protein belonging to the 15th subfamily of class A GPCRs. One of its peculiarities is that its structure is phylogenetically related to both cysteinyl leukotriene (CysLT) receptors and to purinergic P2Y receptors[93]. This receptor is activated by uracil nucleotides such as UDP, UDP-glucose, and UDP-galactose, but is also sensitive to CysLTs, like Leukotriene D4 and C4[94]. GPR17 is mainly expressed in the CNS (but also in kidneys, heart, and generally in organs that can experience ischemic damage), and a more pronounced presence of this protein has been highlighted in oligodendrocyte precursor cells (OPCs). Upregulation of GPR17 can be observed in neuronal cells surrounding an ischemic-injured area, making this protein a marker for cellular stress and death. It has been reported in the literature that in the case of a demyelinating event, GPR17 is involved in the remyelination process, but the mechanism of its involvement is still debated[95]. What is known is that GPR17 is deputed to accompany the OPCs in the early stages of their differentiation process, and so its downregulation is necessary for these cells to complete their maturation. As a result

of this, overexpression of this protein leads to incomplete OPC development, impairing myelination and promoting inflammatory responses[96]. Moreover, GPR17 upregulation in neurons was linked to increased cell damage by Zhao et al., who also observed that its knockdown attenuated neuronal injury and microgliosis[97]. In the field of ALS, GPR17 was demonstrated by Bonfanti et al. to be upregulated in the spinal cord of SOD1^{G93A} mouse models[98]. As reported in a recent study by Raffaele et al., while the application of non-selective GPR17 antagonists such as HAMI3379 (Figure 14, panel A) or Montelukast (a marketed CysLT receptor inhibitor, which is represented in Figure 14, panel B) has been shown to improve remyelination processes (as also demonstrated by Merten et al. for the first of these two molecules[99]), the implementation of agonists has also been demonstrated to be beneficial in pushing OPCs to start differentiating[100]. Jin et al. asserted that the inhibition of GPR17 by Cangrelor (Figure 14, panel C) results in the amelioration of cognitive deficits through the inhibition of oxidative stress and neuroinflammation in Alzheimer's Disease mouse models[101].

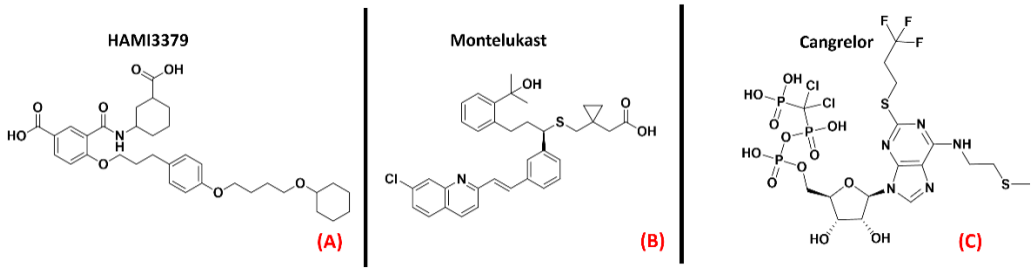


Figure 14. The chemical structures of the GPR17 inhibitors. The non-selective inhibitors HAMI3397, Montelukast (sold as a CysLT receptors inhibitor for asthma), and Cangrelor (an antiplatelet drug, reversible inhibitor of $P2Y_{12}$ receptor), respectively (**A**–**C**).

Marschallinger et al., in a recent paper, highlighted a restoration in cognitive function and a reduction in neuroinflammation in rats treated with Montelukast[102]. Another study by Burnstock et al. indicated that the in vivo knockdown of GPR17 markedly reduced brain damage[103].

The information available nowadays converges in indicating GPR17 (which three-dimensiona structure is provided in Figure 15) as a promising target for neuroinflammation and neurodegeneration diseases. Even if, at the present moment,

these efforts are more focused on multiple sclerosis treatment, GPR17 regulation for ALS is also attracting increasing interest from the scientific community.

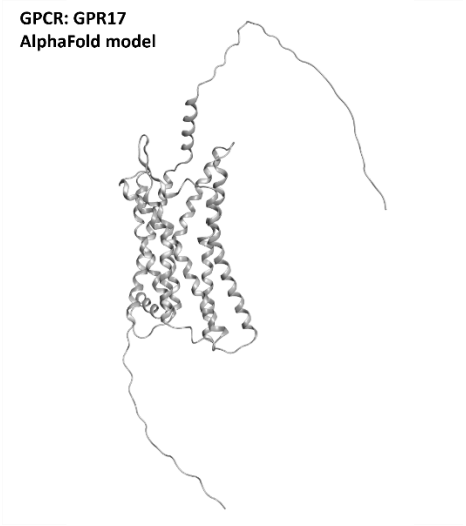


Figure 15. Structure of the GPR17 receptors (with no experimentally resolved structure available, the model sourced from the AlphaFold database is presented). The image was created and rendered with MOE.

2.7. Adrenergic Receptor β_2

Adrenergic receptors are part of the 17th subfamily of class A GPCRs and are divided into nine different isoforms (α_{1A}, α_{1B}, α_{1D}, α_{2A}, α_{2B}, α_{2C}, β_1, β_2, β_3), which are involved in very important physiological functions such as smooth muscle contraction and relaxation, heart muscle contraction (mainly receptors β_1 and β_2)[104], and glycogenolysis[105]. The research to find a link between adrenergic transmission and ALS has been focused on the β_2 isoforms of this GPCR. Historically, β_2 agonists represent one of the main solutions for asthma therapy[106], and drugs such as Salbutamol, Clenbuterol, and Formoterol (all depicted in Figure 16) are an example of this. A recent work by Bartus et al. highlighted the potentialities of β_2 agonists for ALS, reporting that the downstream effects of these molecules can be useful for protecting spinal cord neurons, both preserving and/or restoring their function[107].

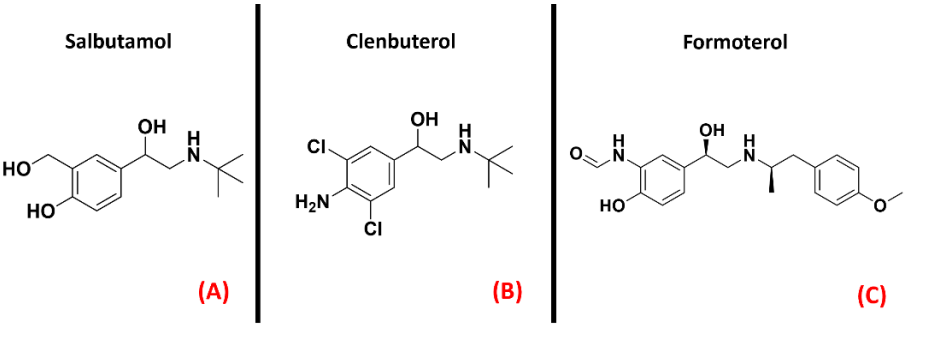


Figure 16. The chemical structures of the adrenergic β_2 receptor agonists Salbutamol (**A**), Clenbuterol (**B**), and Formoterol (**C**).

The pathways responsible for such outcomes presented by Bartus et al. are very biologically important for cell homeostasis, such as the cAMP/PKA/CREB pathway, the PI3K-Akt-mTOR pathway, and the PKA/SIRT1 pathway. Other than neuroprotection, other effects attributed by the authors to this class of ligands are the increase in muscle strength and the amelioration of mitochondrial function. Another study by Teng et al. reported a favorable effect of the β_2 agonist Clenbuterol on $SOD1^{G93A}$ mice[108]. The outcomes of these studies open new possibilities in drug discovery for ALS, focusing special attention on adrenergic β_2 receptor modulation. A three-dimensional representation of the adrenergic β_2 receptor is provided in Figure 17.

Figure 17. Representation of the structure of the adrenergic β_2 receptor (sourced from the Protein Data Bank, PDB code: 7DHI[109], method: cryo-EM, resolution: 3.26 Å). The image was created and rendered with MOE.

2.8. Histamine Receptors

Histamine receptors represent a group of class A GPCRs that has attracted a lot of interest in the pharmaceutical world in recent decades. This family is composed of four different members (H_1, H_2, H_3, and H_4), each with a specific localization in the organism. Their functions, vary from one isoform to another, ranging from vasoconstriction (H_1) to gastric acid secretion (H_2), to neurotransmitter release (H_3), to immunoregulation (mainly H_2 and H_4)[110]. For each histamine receptor, the research has mainly focused on the mechanism of antagonism, of which several marketed drugs are still now relevant examples (e.g., Cetirizine, Figure 18, panel A, for H_1 antagonism; or Famotidine, represented in Figure 18, panel B, for H_2 blockage)[111].

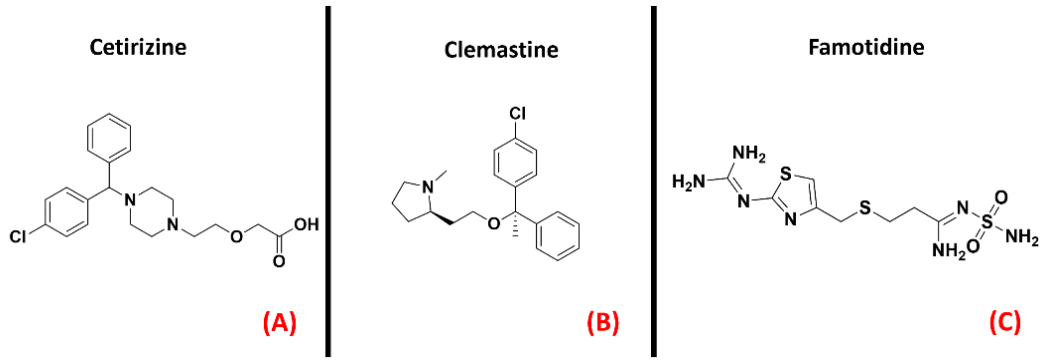


Figure 18. The chemical structures of the H_1 receptor antagonists Cetirizine (**A**) and Clemastine (**B**). The H_2 receptor antagonist Famotidine is also reported (**C**).

An article from Apolloni et al. reported the involvement of histaminergic signals in ALS progression, highlighting that histamine receptors are dysregulated in the cortex, spinal cord, and hypothalamus of $SOD1^{G93A}$ ALS-affected mice[112]. This study reported that histamine could counteract the pro-inflammatory phenotype of microglia, mainly through its H_1 and H_4 receptor isoforms. This would be mediated by both the reduction of NOX-2 and NF-kB expression and the increase in production of other species, such as IL-6 and IL-10. Another work by Volontè et al. highlighted the neuroprotective effects of histamine signaling in ALS, again giving higher relevance to H1R (represented in Figure 19, panel A) and H4R[113]. On the other hand, Zhang et al. described H_1 and H_4 receptors as being responsible for pro-inflammatory cytokine release in microglia, while H_2 and H_3 were considered to be the main actors of anti-inflammation in that environment (the H_2

receptor is depicted in its three-dimensional structure in Figure 19, panel B)[114]. Another article by Apolloni et al. reported an amelioration in ALS progression of SOD1^{G93A} mice treated with the anti-histaminergic drug Clemastine[115]. Even if much more remains to be understood about the specific role of each histamine receptor isoform in ALS progression, what is certain is that this GPCR family has already proven to be a promising target for drug development against neuroinflammation and neurodegeneration.

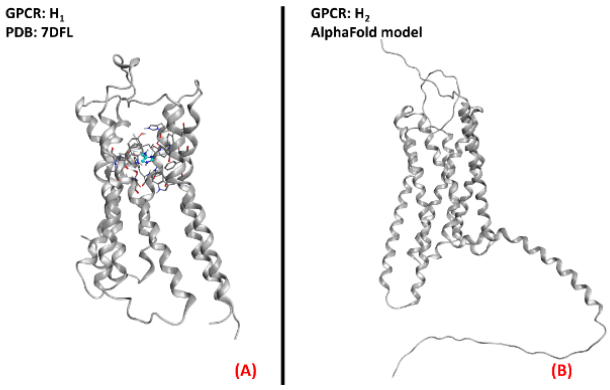


Figure 19. Representation of the three-dimensional structures of the histamine receptors that could be considered for ALS treatment. (**A**) The H_1 receptor (sourced from the Protein Data Bank, PDB code: 7DFL [116], method: cryo-EM, resolution: 3.30 Å) and (**B**) the H_2 receptor (with no experimentally resolved structure available, the model sourced from the AlphaFold[117] database is presented). The images were created and rendered with MOE. The H3 and H4 receptors both lack experimentally resolved structures, but their AlphaFold models are available.

2.9. Cannabinoid Receptors

A group of class A GPCRs that are of high interest at the present date is certainly the cannabinoid receptors. These biological entities are the main actors in the endocannabinoid system, and play relevant roles in several physiological processes. Indeed, the first of its two main isoforms, called CB_1, is mainly located in both the central and peripheral nervous system, acting as a neurotransmitter release modulator in response to the binding of its agonists (mainly anandamide, but also 2-arachidonoylglycerol, both represented in Figure 20, panels A and B, respectively). In the majority of cases, CB_1 is coupled with $G_{i/o}$ protein, leading to adenylyl cyclase inhibition and consequent decrease in cAMP upon activation. The final effect of such an action is the reduction of neurotransmitter release in the synapse. On the other hand, the CB_2

receptor is mainly localized on the surface of the immune system cells. Its main agonist is 2-arachidonoylgycerol, binding of which leads to the inhibition of adenylyl cyclase through $G_{i/o}$ subunit action[118]. The final main effect is immunosuppression[119]. As reported by an article from Giacoppo and Mazzon, several studies have shown how the application of cannabinoid receptor agonists in $SOD1^{G93A}$ mouse models of ALS could be beneficial for the neuroprotective effects mediated by them[120]. Similarly, in 2019, Urbi et al. performed a meta-analysis on the studies regarding the application of cannabinoids in ALS murine models, highlighting the effective concordance in assessing that their application leads to a delay in disease progression[121]. A study by Shoemaker et al. highlighted that the increase in survival could be more addressable to the CB_2 isoform, showing that the administration of the CB_2 selective agonist AM-1241 (Figure 20, panel C) increased survival by 56%[122]. This molecule was also examined for cannabinoid-mediated ALS treatment by Kim et al., with similar results[123].

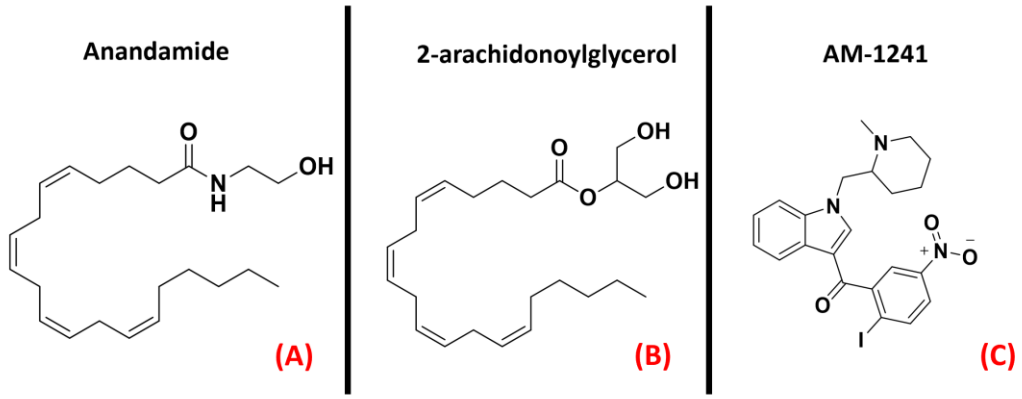


Figure 20. The chemical structures of the endogenous CB_1 and CB_2 receptors agonists anandamide (**A**) and 2-arachidonoylglycerol (**B**). The selective CB_2 receptor agonist AM-1241 is also reported (**C**).

In addition to this, Bilsland et al. reported that the knock-out of CB_1 receptors in $SOD1^{G39A}$ ALS-affected mice had no appreciable effect on disease onset[124], and regarding this, Shoemaker et al. reported that the activation of CB_1 could exacerbate disease progression[122]. The literature available today regarding the application of molecules acting on the endocannabinoid system for ALS treatment converges in the possible evaluation

of a therapy based on CB_2 selective agonists. A three-dimensional representation of both CB_1 and CB_2 receptors is provided in Figure 21.

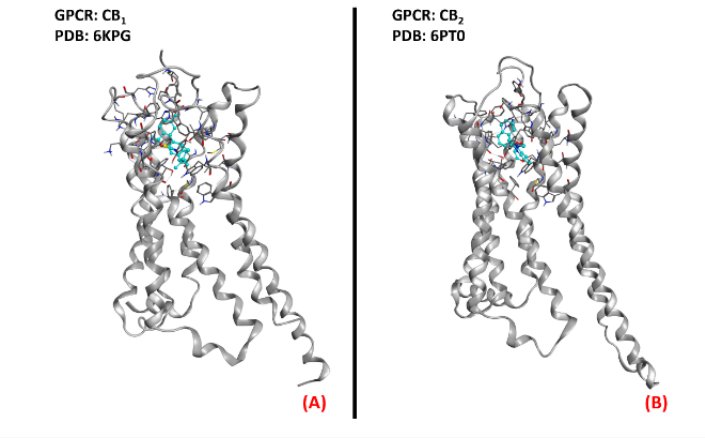


Figure 21. The structures of (**A**) the CB_1 receptor (sourced from the Protein Data Bank, PDB code: 6KPG[125], method: cryo-EM, resolution: 3.00 Å) and (**B**) the CB_2 receptor (sourced from the Protein Data Bank, PDB code: 6PT0[126], method: cryo-EM, resolution: 3.20 Å). The images were created and rendered with MOE.

2.10. Prostaglandin E_2 Receptor (PGE2R)

Prostaglandin E_2 receptors (PGE_2) are a series of class A GPCRs that selectively bind to prostaglandin E_2 (also known as dinoprostone), an endogenous arachidonic acid derivative of high importance for several physiological functions. PGE_2 can be divided into four isoforms, named E_1, E_2, E_3, and E_4 (all represented in their three-dimensional structure in Figure 22). With the exception of the first, which stimulates phospholipase C if agonized, the other isoforms act on adenylyl cyclase and, specifically, the EP_2 and EP_4 isoforms (coupled with a G_s subunit) stimulate its function when agonized, while EP_3 inhibits AC through its action (being coupled to a $G_{i/o}$ subunit)[127]. EP_1 function has been correlated with hyperalgesia[128], immunoregulation[129], and colon cancer progression[130]. EP_2, which is active in the reproductive, visual, cardiovascular, skeletal, and nervous systems, has also been strictly related to tumor promotion, as highlighted in a 2018 study by Sun and Li[131]. Minor correlations with cancer have been reported for EP_3, which is also important for a large variety of functions, ranging from digestion[132] to blood pressure[133] and clotting[134], in addition to pain management[135]. The spectrum of systems in which the fourth isoform of PGE_2 receptors, EP_4, is involved is also very wide. Additionally, in this

case, EP_4 has been reported to be hyper-expressed in various types of cancer, mainly prostate cancer[136]. Talking about ALS onset and progression, Iłzecka found increased levels of PGE_2 in the cerebrospinal fluid of ALS patients, and therefore concluded that this mediator could play a role in disease progression, suggesting that its inhibition could be beneficial[137]. Additionally, Kosuge et al. highlighted the role of PGE_2 in the ROS generation pathway, focusing on its impact on ALS conditions. The same conclusion was reported in 2008 by Liang et al., who suggested EP_2 inhibition as a novel way to treat the neuroinflammation typical in ALS[138]. On the other hand, PGE_2 receptors were proposed to have an unexpected neuroprotective effect on motor neurons by Bilak et al., who reported that the neuroinflammatory process typical of ALS was mainly due to COX-2-mediated, prostaglandin-independent processes[139]. Taken together, all of these studies converge in evaluating PGE_2 receptors as interesting pharmacological targets for ALS, being strongly correlated with the significant neuroinflammation characterizing the pathology.

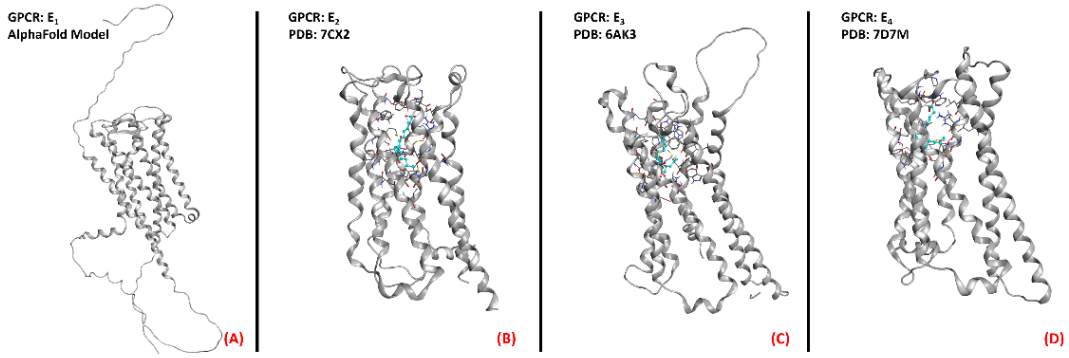


Figure 22. Structure of the PGE_2 receptors (**A**) EP_1 (with no experimentally resolved structure available, the model sourced from the AlphaFold database is presented), (**B**) EP_2 (sourced from the Protein Data Bank, PDB code: 7CX2[140], method: cryo-EM, resolution: 2.80 Å), (**C**) EP_3 (sourced from the Protein Data Bank, PDB code: 6AK3[141], method: X-ray diffraction, resolution: 2.90 Å), and (**D**) EP_4 (sourced from the Protein Data Bank, PDB code: 7D7M[142], method: cryo-EM, resolution: 3.30 Å). All of the images were created and rendered with MOE.

2.11. Vasoactive Intestinal Peptide Receptors

The receptors for the vasoactive intestinal peptide are part of the first subfamily of class B GPCRs. As part of this group of proteins, these receptors are responsive to signals

mediated by the peptide hormone VIP (vasoactive intestinal polypeptide), formed by 28 amino acids and belonging to the glucagon/secretin superfamily[143]. After being produced by organs such as the gut, pancreas, and brain, VIP act in different physiological functions depending on the target tissue and the receptor isoform interacting with it. Indeed, two vasoactive intestinal polypeptide receptor isoforms are known, namely $VPAC_1$ and $VPAC_2$ (both represented in Figure 23). Both of these proteins are highly expressed throughout the human body, from the smooth muscle of the GI tract and blood vessels to the reproductive system, lungs, spleen, and brain[144]. Their activity is mediated by a G_s protein, and involves the activation of adenylyl cyclase upon binding to VIP, consequently activating the protein kinase A (PKA)[145].

The implication of VIP in the CNS has been noticed when studying both the circadian rhythm and schizophrenia[144], but its importance as a potential target for ALS therapy is of more recent discovery. In 2008, Staines considered the possibility of studying vasoactive neuropeptides for degenerating pathologies such as MS and ALS[146], and a previous article by Iwasaki et al. specifically referred to the neurotrophic properties exerted by VIP in the degenerating diseases of motor neurons[147]. Solés-Tarrés et al. analyzed the neuroprotective effects of both VIP and PACAP (pituitary adenylate cyclase-activating polypeptide, a peptide hormone binding to both VPACs and PACAP receptors), also evaluating the synthetic derivatives available to mimic their action, with a special focus on the intrinsic pharmacokinetic problems of these species[148]. Waschek already identified both VIP and PACAP as promising targets for neuroinflammation in the CNS[149], as pointed out again in recent work by Martinez et al.[150].

The interaction with receptors of vasoactive peptides has been demonstrated to be a promising way to counteract the neuroinflammatory and degenerative effects of ALS, mainly through biological and/or chemical species mimicking the functions of the endogenous peptides. Further research on this topic will define the best way to accomplish this task.

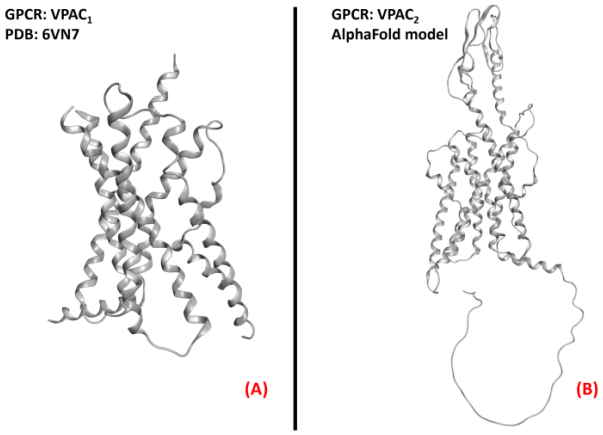


Figure 23. Structure of the receptors (**A**) $VPAC_1$ (sourced from the Protein Data Bank, PDB code: 6VN7[151], method: cryo-EM, resolution: 3.20 Å), and (**B**) $VPAC_2$ (with no experimentally resolved structure available, the model sourced from the AlphaFold database is presented). The images were created and rendered with MOE.

2.12. Metabotropic Glutamate Receptors (mGluRs)

Metabotropic glutamate receptors (mGluRs) are GPCRs belonging to class C of this family of proteins. As the name suggests, these entities bind to the neurotransmitter glutamate, exerting different functions in both the central and peripheral nervous systems. They can be divided into three groups, with the first being composed of mGluR1 and mGluR5, predominantly postsynaptic, which, once activated, cause the stimulation of phospholipase C (PLC) through G_q-mediated signaling. Group II (formed by mGluRs 2 and 3) and III (of which mGluRs 4,6,7, and 8 are a part) receptors are mainly presynaptic and are all coupled with a $G_{i/0}$ subunit, which inhibits the activation of adenylyl cyclase, causing presynaptic inhibition. The functions of this family of proteins are majorly related to the nervous system, from the modulation of neurotransmission (e.g., gabaergic and dopaminergic) and of other proteins' signaling (e.g., NMDA receptors), to synaptic plasticity regulation[152]. This being said, it appears clear that the possibility of their involvement in ALS onset and progression is more than possible. Anneser et al. found a strong upregulation of mGluRs in the spinal cord with ALS, leading to the propagation of glial proliferation[153]. Hyperactivity of group I mGluRs has been correlated with neuroinflammation. Indeed, as demonstrated by Milanese et al., $SOD1^{G93A}$ ALS-affected mice with mGluR1 knockdown experience a reduction in microglia and astrocyte

activation, decreasing mitochondrial damage and improving survival[154], and this phenomenon was also highlighted by Rossi et al.[155]

Anneser et al. showed the beneficial and protective effects of both agonism and antagonism of group I mGluRs for motor neuron disease, while less promising effects were derived from modulation of other mGluRs[156]. Crupi et al. recently pointed out that the beneficial therapeutic modulation of mGluRs is usually achieved through the reduction of the excitotoxicity drive via mGluR I inhibition or mGluR II and III agonism[157]. In conclusion, the literature supports the possibility of investing resources in the treatment of motor neuron diseases via mGluR modulation. A three-dimensional representation of mGluR1, mGluR2, and mGluR4 receptors is provided in Figure 24.

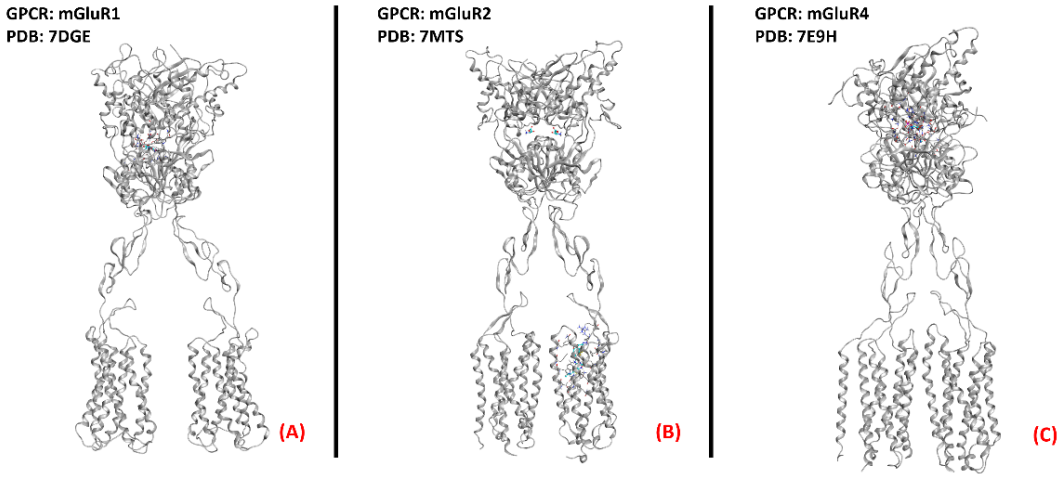


Figure 24. One example of each group of the metabotropic glutamate receptors. (**A**) mGluR1, a member of the first group of mGluRs (sourced from the Protein Data Bank, PDB code: 7DGE[158], method: cryo-EM, resolution: 3.65 Å), (**B**) mGluR2 (owing to mGluRs group II, sourced from the Protein Data Bank, PDB code: 7MTS[159], method: cryo-EM, resolution: 3.20 Å), and (**C**) mGluR4, part of group III of the mGluRs (sourced from the Protein Data Bank, PDB code: 7E9H[160], method: cryo-EM, resolution: 4.00 Å). All of the images were created and rendered with MOE.

3. Conclusions

In this review, we provided a panoramic view of the involvement of different G-protein-coupled receptors in the onset and progression of ALS, evaluating what has already been discovered on these biological entities, and highlighting what the next steps in research could be, always on the basis of the present literature on the topic. Our analysis shows

that a GPCR-based therapy for ALS could be considered a practical possibility for the eradication of the ALS condition, and we encourage scientific groups all around the world in directing efforts towards this field.

References

1. Hardiman, O. *et al.* Amyotrophic lateral sclerosis. *Nature Reviews Disease Primers* **3**, 17071 (2017).

2. Heckman, C. J. & Enoka, R. M. Physiology of the motor neuron and the motor unit. in 119–147 (2004). doi:10.1016/S1567-4231(04)04006-7.

3. Wijesekera, L. C. & Nigel Leigh, P. Amyotrophic lateral sclerosis. *Orphanet Journal of Rare Diseases* **4**, 3 (2009).

4. Kiernan, M. C. *et al.* Amyotrophic lateral sclerosis. *The Lancet* **377**, 942–955 (2011).

5. de la Rubia Ortí, J. E. *et al.* Muscle Function Differences between Patients with Bulbar and Spinal Onset Amyotrophic Lateral Sclerosis. Does It Depend on Peripheral Glucose? *Journal of Clinical Medicine* **10**, 1582 (2021).

6. van Es, M. A. *et al.* Amyotrophic lateral sclerosis. *The Lancet* **390**, 2084–2098 (2017).

7. www.ninds.nih.gov, published J. 2013, updated 26th M. 2021. Amyotrophic Lateral Sclerosis (ALS) Fact Sheet | National Institute of Neurological Disorders and Stroke".

8. Saitoh, Y. & Takahashi, Y. Riluzole for the treatment of amyotrophic lateral sclerosis. *Neurodegenerative Disease Management* **10**, 343–355 (2020).

9. Breiner, A., Zinman, L. & Bourque, P. R. Edaravone for amyotrophic lateral sclerosis: barriers to access and lifeboat ethics. *Canadian Medical Association Journal* **192**, E319–E320 (2020).

10. Jaiswal, M. K. Riluzole and edaravone: A tale of two amyotrophic lateral sclerosis drugs. *Medicinal Research Reviews* **39**, 733–748 (2019).

11. Masrori, P. & Van Damme, P. Amyotrophic lateral sclerosis: a clinical review. *European Journal of Neurology* **27**, 1918–1929 (2020).

12. Boddy, S. L. *et al.* The gut microbiome: a key player in the complexity of amyotrophic lateral sclerosis (ALS). *BMC Medicine* **19**, 13 (2021).

13. Neumann, M. *et al.* Ubiquitinated TDP-43 in Frontotemporal Lobar Degeneration and Amyotrophic Lateral Sclerosis. *Science* **314**, 130–133 (2006).

14. Jo, M. *et al.* The role of TDP-43 propagation in neurodegenerative diseases: integrating insights from clinical and experimental studies. *Experimental & Molecular Medicine* **52**, 1652–1662 (2020).

15. Suk, T. R. & Rousseaux, M. W. C. The role of TDP-43 mislocalization in amyotrophic lateral sclerosis. *Molecular Neurodegeneration* **15**, 45 (2020).

16. Berdyński, M. *et al.* SOD1 mutations associated with amyotrophic lateral sclerosis analysis of variant severity. *Scientific Reports* **12**, 103 (2022).

17. Ishigaki, S. & Sobue, G. Importance of Functional Loss of FUS in FTLD/ALS. *Frontiers in Molecular Biosciences* **5**, (2018).

18. Eck, R. J., Kraemer, B. C. & Liachko, N. F. Regulation of TDP-43 phosphorylation in aging and disease. *GeroScience* **43**, 1605–1614 (2021).

19. Guo, W., Vandoorne, T., Steyaert, J., Staats, K. A. & Van Den Bosch, L. The multifaceted role of kinases in amyotrophic lateral sclerosis: genetic, pathological and therapeutic implications. *Brain* **143**, 1651–1673 (2020).

20. Palomo, V., Nozal, V., Rojas-Prats, E., Gil, C. & Martinez, A. Protein kinase inhibitors for amyotrophic lateral sclerosis therapy. *British Journal of Pharmacology* **178**, 1316–1335 (2021).

21. Miller, R. G., Mitchell, J. D. & Moore, D. H. Riluzole for amyotrophic lateral sclerosis (ALS)/motor neuron disease (MND). *Cochrane Database of Systematic Reviews* (2012) doi:10.1002/14651858.CD001447.pub3.

22. Bissaro, M. & Moro, S. Rethinking to riluzole mechanism of action: the molecular link among protein kinase CK1δ activity, TDP-43 phosphorylation, and amyotrophic lateral sclerosis pharmacological treatment. *Neural Regeneration Research* **14**, 2083 (2019).

23. Corcia, P., Beltran, S., Bakkouche, S. E. & Couratier, P. Therapeutic news in ALS. *Revue Neurologique* **177**, 544–549 (2021).

24. Springer Nature. AdisInsight - small-molecules in clinical trials for ALS. https://adisinsight.springer.com/search.

25. Chen, H., Kankel, M. W., Su, S. C., Han, S. W. S. & Ofengeim, D. Exploring the genetics and non-cell autonomous mechanisms underlying ALS/FTLD. *Cell Death & Differentiation* **25**, 648–662 (2018).

26. Insel, P. A. *et al.* G Protein–Coupled Receptor (GPCR) Expression in Native Cells: "Novel" endoGPCRs as Physiologic Regulators and Therapeutic Targets. *Molecular Pharmacology* **88**, 181–187 (2015).

27. B. Gacasan, S., L. Baker, D. & L. Parrill, A. G protein-coupled receptors: the evolution of structural insight. *AIMS Biophysics* **4**, 491–527 (2017).

28. Lee, Y., Basith, S. & Choi, S. Recent Advances in Structure-Based Drug Design Targeting Class A G Protein-Coupled Receptors Utilizing Crystal Structures and Computational Simulations. *Journal of Medicinal Chemistry* **61**, 1–46 (2018).

29. Heng, B. C., Aubel, D. & Fussenegger, M. An overview of the diverse roles of G-protein coupled receptors (GPCRs) in the pathophysiology of various human diseases. *Biotechnology Advances* **31**, 1676–1694 (2013).

30. Zarrinmayeh, H. & Territo, P. R. Purinergic Receptors of the Central Nervous System: Biology, PET Ligands, and Their Applications. *Molecular Imaging* **19**, 153601212092760 (2020).

31. Borea, P. A., Gessi, S., Merighi, S., Vincenzi, F. & Varani, K. Pharmacology of Adenosine Receptors: The State of the Art. *Physiological Reviews* **98**, 1591–1625 (2018).

32. Gomes, C. V., Kaster, M. P., Tomé, A. R., Agostinho, P. M. & Cunha, R. A. Adenosine receptors and brain diseases: Neuroprotection and neurodegeneration. *Biochimica et Biophysica Acta (BBA) - Biomembranes* **1808**, 1380–1399 (2011).

33. Vincenzi, F. *et al.* A $_{2A}$ adenosine receptors are up-regulated in lymphocytes from amyotrophic lateral sclerosis patients. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration* **14**, 406–413 (2013).

34. YOSHIDA, Y. *et al.* Adenosine and Neopterin Levels in Cerebrospinal Fluid of Patients with Neurological Disorders. *Internal Medicine* **38**, 133–139 (1999).

35. Potenza, R. L. *et al.* Effects of chronic caffeine intake in a mouse model of amyotrophic lateral sclerosis. *Journal of Neuroscience Research* **91**, 585–592 (2013).

36. Mori, A., Cross, B., Uchida, S., Kerrick Walker, J. & Ristuccia, R. How Are Adenosine and Adenosine A2A Receptors Involved in the Pathophysiology of Amyotrophic Lateral Sclerosis? *Biomedicines* **9**, 1027 (2021).

37. Ng, S. kah, Higashimori, H., Tolman, M. & Yang, Y. Suppression of adenosine 2a receptor (A2aR)-mediated adenosine signaling improves disease phenotypes in a mouse model of amyotrophic lateral sclerosis. *Experimental Neurology* **267**, 115–122 (2015).

38. Mojsilovic-Petrovic, J. Protecting Motor Neurons from Toxic Insult by Antagonism of Adenosine A2a and Trk Receptors. *Journal of Neuroscience* **26**, 9250–9263 (2006).

39. Liu, Y.-J. *et al.* Activation of AMP-activated protein kinase α1 mediates mislocalization of TDP-43 in amyotrophic lateral sclerosis. *Human Molecular Genetics* **24**, 787–801 (2015).

40. Berman, H. M. The Protein Data Bank. *Nucleic Acids Res* **28**, 235–242 (2000).

41. Segala, E. *et al.* Controlling the Dissociation of Ligands from the Adenosine A $_{2A}$ Receptor through Modulation of Salt Bridge Strength. *Journal of Medicinal Chemistry* **59**, 6470–6479 (2016).

42. Molecular Operating Environment (MOE), 2019.01; Chemical Computing Group ULC, 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, 2021.

43. Burnstock, G. Purinergic Signalling: Therapeutic Developments. *Frontiers in Pharmacology* **8**, (2017).

44. Gómez Morillas, A., Besson, V. C. & Lerouet, D. Microglia and Neuroinflammation: What Place for P2RY12? *International Journal of Molecular Sciences* **22**, 1636 (2021).

45. Amadio, S. *et al.* P2Y $_{12}$ Receptor on the Verge of a Neuroinflammatory Breakdown. *Mediators of Inflammation* **2014**, 1–15 (2014).

46. Jacobson, K. A. *et al.* Update of P2Y receptor pharmacology: IUPHAR Review 27. *British Journal of Pharmacology* **177**, 2413–2433 (2020).

47. D'Ambrosi, N. *et al.* The Proinflammatory Action of Microglial P2 Receptors Is Enhanced in SOD1 Models for Amyotrophic Lateral Sclerosis. *The Journal of Immunology* **183**, 4648–4656 (2009).

48. Kobayashi, K. *et al.* P2Y12 Receptor Upregulation in Activated Microglia Is a Gateway of p38 Signaling and Neuropathic Pain. *Journal of Neuroscience* **28**, 2892–2902 (2008).

49. Moore, C. S. *et al.* P2Y12 expression and function in alternatively activated human microglia. *Neurology - Neuroimmunology Neuroinflammation* **2**, e80 (2015).

50. Zhang, K. *et al.* Structure of the human P2Y12 receptor in complex with an antithrombotic drug. *Nature* **509**, 115–118 (2014).

51. Hughes, C. E. & Nibbs, R. J. B. A guide to chemokines and their receptors. *The FEBS Journal* **285**, 2944–2971 (2018).

52. Agle, K. A., Vongsa, R. A. & Dwinell, M. B. Calcium Mobilization Triggered by the Chemokine CXCL12 Regulates Migration in Wounded Intestinal Epithelial Monolayers. *Journal of Biological Chemistry* **285**, 16066–16075 (2010).

53. La Cognata, V. *et al.* CXCR2 increases in ALS cortical neurons and its inhibition prevents motor neuron degeneration in vitro and improves neuromuscular function in SOD1G93A mice. *Neurobiology of Disease* **160**, 105538 (2021).

54. Rabinovich-Nikitin, I., Ezra, A., Barbiro, B., Rabinovich-Toidman, P. & Solomon, B. Chronic administration of AMD3100 increases survival and alleviates pathology in SOD1G93A mice model of ALS. *Journal of Neuroinflammation* **13**, 123 (2016).

55. Liu, E., Karpf, L. & Bohl, D. Neuroinflammation in Amyotrophic Lateral Sclerosis and Frontotemporal Dementia and the Interest of Induced Pluripotent Stem Cells to Study Immune Cells Interactions With Neurons. *Frontiers in Molecular Neuroscience* **14**, (2021).

56. Perner, C. *et al.* Dysregulation of chemokine receptor expression and function in leukocytes from ALS patients. *Journal of Neuroinflammation* **15**, 99 (2018).

57. Liu, K. *et al.* Structural basis of CXC chemokine receptor 2 activation and signalling. *Nature* **585**, 135–140 (2020).

58. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

59. Wu, B. *et al.* Structures of the CXCR4 Chemokine GPCR with Small-Molecule and Cyclic Peptide Antagonists. *Science* **330**, 1066–1071 (2010).

60. Singh, K. D. & Karnik, S. S. Angiotensin Receptors: Structure, Function, Signaling and Clinical Applications. *Journal of cell signaling* **1**, (2016).

61. Ames, M. K., Atkins, C. E. & Pitt, B. The renin-angiotensin-aldosterone system and its suppression. *Journal of Veterinary Internal Medicine* **33**, 363–382 (2019).

62. Miura, S., Karnik, S. S. & Saku, K. Review: Angiotensin II type 1 receptor blockers: class effects versus molecular effects. *Journal of the Renin-Angiotensin-Aldosterone System* **12**, 1–7 (2011).

63. Akishita, M. *et al.* Expression of the AT2 receptor developmentally programs extracellular signal-regulated kinase activity and influences fetal vascular growth. *Journal of Clinical Investigation* **103**, 63–71 (1999).

64. Kawajiri, M. *et al.* Reduced angiotensin II levels in the cerebrospinal fluid of patients with amyotrophic lateral sclerosis. *Acta Neurologica Scandinavica* **119**, 341–344 (2009).

65. Iwasaki, Y., Ichikawa, Y., Igarashi, O., Kinoshita, M. & Ikeda, K. Trophic effect of Olmesartan, a novel AT $_1$ R antagonist, on spinal motor neurons *in vitro* and *in vivo*. *Neurological Research* **24**, 468–472 (2002).

66. Mammana, S. *et al.* The Role of Macrophages in Neuroinflammatory and Neurodegenerative Pathways of Alzheimer's Disease, Amyotrophic Lateral Sclerosis, and Multiple Sclerosis: Pathogenetic Cellular Effectors and Potential Therapeutic Targets. *International Journal of Molecular Sciences* **19**, 831 (2018).

67. Zhang, H. *et al.* Structural Basis for Ligand Recognition and Functional Selectivity at Angiotensin Receptor. *Journal of Biological Chemistry* **290**, 29127–29139 (2015).

68. Perryman, R. et al. Inhibition of the angiotensin II type 2 receptor AT2R is a novel therapeutic strategy for glioblastoma. *To be published (PDB doi: 10.2210/pdb7JNI/pdb)*.

69. Beaulieu, J.-M. & Gainetdinov, R. R. The Physiology, Signaling, and Pharmacology of Dopamine Receptors. *Pharmacological Reviews* **63**, 182–217 (2011).

70. Vallone, D., Picetti, R. & Borrelli, E. Structure and function of dopamine receptors. *Neuroscience & Biobehavioral Reviews* **24**, 125–132 (2000).

71. Borasio, G. D. *et al.* Dopaminergic deficit in amyotrophic lateral sclerosis assessed with [I-123] IPT single photon emission computed tomography. *Journal of Neurology, Neurosurgery & Psychiatry* **65**, 263–265 (1998).

72. Lai, C.-Y. *et al.* The D2 Dopamine Receptor Interferes With the Protective Effect of the A2A Adenosine Receptor on TDP-43 Mislocalization in Experimental Models of Motor Neuron Degeneration. *Frontiers in Neuroscience* **12**, (2018).

73. Huang, X. *et al.* Human amyotrophic lateral sclerosis excitability phenotype screen: Target discovery and validation. *Cell Reports* **35**, 109224 (2021).

74. Fujimori, K. *et al.* Modeling sporadic ALS in iPSC-derived motor neurons identifies a potential therapeutic agent. *Nature Medicine* **24**, 1579–1589 (2018).

75. Huang, X. *et al.* Human amyotrophic lateral sclerosis excitability phenotype screen: Target discovery and validation. *Cell Reports* **35**, 109224 (2021).

76. Dooley, M. & Markham, A. Pramipexole. *Drugs & Aging* **12**, 495–514 (1998).

77. Gribkoff, V. K. & Bozik, M. E. KNS-760704 [(6R)-4,5,6,7-tetrahydro-N6-propyl-2, 6-benzothiazole-diamine dihydrochloride monohydrate] for the Treatment of Amyotrophic Lateral Sclerosis. *CNS Neuroscience & Therapeutics* **14**, 215–226 (2008).

78. Kingwell, K. Dexpramipexole shows promise for ALS in phase II trial. *Nature Reviews Neurology* **8**, 4–4 (2012).

79. Cudkowicz, M. E. *et al.* Dexpramipexole versus placebo for patients with amyotrophic lateral sclerosis (EMPOWER): a randomised, double-blind, phase 3 trial. *The Lancet Neurology* **12**, 1059–1067 (2013).

80. Zhuang, Y. *et al.* Structural insights into the human D1 and D2 dopamine receptor signaling complexes. *Cell* **184**, 931-942.e18 (2021).

81. Xu, P. *et al.* Structures of the human dopamine D3 receptor-Gi complexes. *Molecular Cell* **81**, 1147-1159.e4 (2021).

82. Wang, S. *et al.* D $_4$ dopamine receptor high-resolution structures enable the discovery of selective agonists. *Science* **358**, 381–386 (2017).

83. Nichols, D. E. & Nichols, C. D. Serotonin Receptors. *Chemical Reviews* **108**, 1614–1641 (2008).

84. *Frazer A, Hensler JG. Serotonin Receptors. In: Siegel GJ, Agranoff BW, Albers RW, et al., editors. Basic Neurochemistry: Molecular, Cellular and Medical Aspects. 6th edition. Philadelphia: Lippincott-Raven; 1999. Available from: https://www.ncbi.nlm.nih.gov/books/NBK28234/*.

85. El Oussini, H. *et al.* Serotonin 2B receptor slows disease progression and prevents degeneration of spinal cord mononuclear phagocytes in amyotrophic lateral sclerosis. *Acta Neuropathologica* **131**, 465–480 (2016).

86. Dentel, C. *et al.* Degeneration of serotonergic neurons in amyotrophic lateral sclerosis: a link to spasticity. *Brain* **136**, 483–493 (2013).

87. Arnoux, A. *et al.* Evaluation of a 5-HT2B receptor agonist in a murine model of amyotrophic lateral sclerosis. *Scientific Reports* **11**, 23582 (2021).

88. Elangbam, C. S. *et al.* 5-Hydroxytryptamine (5HT)-induced valvulopathy: Compositional valvular alterations are associated with 5HT2B receptor and 5HT transporter transcript changes in Sprague-Dawley rats. *Experimental and Toxicologic Pathology* **60**, 253–262 (2008).

89. Lacomblez, L. *et al.* Xaliproden in amyotrophic lateral sclerosis: early clinical trials. *Amyotrophic Lateral Sclerosis and Other Motor Neuron Disorders* **5**, 99–106 (2004).

90. Xu, P. *et al.* Structural insights into the lipid and ligand regulation of serotonin receptors. *Nature* **592**, 469–473 (2021).

91. McCorvy, J. D. *et al.* Structural determinants of 5-HT2B receptor activation and biased agonism. *Nature Structural & Molecular Biology* **25**, 787–796 (2018).

92. Peng, Y. *et al.* 5-HT2C Receptor Structures Reveal the Structural Basis of GPCR Polypharmacology. *Cell* **172**, 719-730.e14 (2018).

93. Marucci, G. *et al.* GPR17 receptor modulators and their therapeutic implications: review of recent patents. *Expert Opinion on Therapeutic Patents* **29**, 85–95 (2019).

94. Dziedzic, A., Miller, E., Saluk-Bijak, J. & Bijak, M. The GPR17 Receptor—A Promising Goal for Therapy and a Potential Marker of the Neurodegenerative Process in Multiple Sclerosis. *International Journal of Molecular Sciences* **21**, 1852 (2020).

95. Marucci, G. *et al.* The G Protein-Coupled Receptor GPR17: Overview and Update. *ChemMedChem* **11**, 2567–2574 (2016).

96. Lecca, D., Raffaele, S., Abbracchio, M. P. & Fumagalli, M. Regulation and signaling of the GPR17 receptor in oligodendroglial cells. *Glia* **68**, 1957–1967 (2020).

97. Zhao, B. *et al.* The new P2Y-like receptor G protein-coupled receptor 17 mediates acute neuronal injury and late microgliosis after focal cerebral ischemia in rats. *Neuroscience* **202**, 42–57 (2012).

98. Bonfanti, E. *et al.* Abnormal Upregulation of GPR17 Receptor Contributes to Oligodendrocyte Dysfunction in SOD1 G93A Mice. *International Journal of Molecular Sciences* **21**, 2395 (2020).

99. Merten, N. *et al.* Repurposing HAMI3379 to Block GPR17 and Promote Rodent and Human Oligodendrocyte Differentiation. *Cell Chemical Biology* **25**, 775-786.e5 (2018).

100. Raffaele, S., Boccazzi, M. & Fumagalli, M. Oligodendrocyte Dysfunction in Amyotrophic Lateral Sclerosis: Mechanisms and Therapeutic Perspectives. *Cells* **10**, 565 (2021).

101. Jin, S. *et al.* Inhibition of GPR17 with cangrelor improves cognitive impairment and synaptic deficits induced by Aβ1–42 through Nrf2/HO-1 and NF-κB signaling pathway in mice. *International Immunopharmacology* **101**, 108335 (2021).

102. Marschallinger, J. *et al.* Structural and functional rejuvenation of the aged brain by an approved anti-asthmatic drug. *Nature Communications* **6**, 8466 (2015).

103. Burnstock, G. An introduction to the roles of purinergic signalling in neurodegeneration, neuroprotection and neuroregeneration. *Neuropharmacology* **104**, 4–17 (2016).

104. Insel, P. A. Adrenergic Receptors — Evolving Concepts and Clinical Implications. *New England Journal of Medicine* **334**, 580–585 (1996).

105. Subbarao, K. V. & Hertz, L. Effect of adrenergic agonists on glycogenolysis in primary cultures of astrocytes. *Brain Research* **536**, 220–226 (1990).

106. Patel, M. & Shaw, D. A review of standard pharmacological therapy for adult asthma – Steps 1 to 5. *Chronic Respiratory Disease* **12**, 165–176 (2015).

107. Bartus, R. T. *et al.* β 2 -Adrenoceptor agonists as novel, safe and potentially effective therapies for Amyotrophic lateral sclerosis (ALS). *Neurobiology of Disease* **85**, 11–24 (2016).

108. Teng, Y. D. *et al.* Therapeutic effects of clenbuterol in a murine model of amyotrophic lateral sclerosis. *Neuroscience Letters* **397**, 155–158 (2006).

109. Yang, F. *et al.* Different conformational responses of the β2-adrenergic receptor-Gs complex upon binding of the partial agonist salbutamol or the full agonist isoprenaline. *National Science Review* **8**, (2021).

110. Parsons, M. E. & Ganellin, C. R. Histamine and its receptors. *British Journal of Pharmacology* **147**, S127–S135 (2006).

111. Tiligada, E. & Ennis, M. Histamine pharmacology: from Sir Henry Dale to the 21st century. *British Journal of Pharmacology* **177**, 469–489 (2020).

112. Apolloni, S. *et al.* Histamine Regulates the Inflammatory Profile of SOD1-G93A Microglia and the Histaminergic System Is Dysregulated in Amyotrophic Lateral Sclerosis. *Frontiers in Immunology* **8**, (2017).

113. Volonté, C., Apolloni, S. & Sabatelli, M. Histamine beyond its effects on allergy: Potential therapeutic benefits for the treatment of Amyotrophic Lateral Sclerosis (ALS). *Pharmacology & Therapeutics* **202**, 120–131 (2019).

114. Zhang, W. *et al.* Histamine Induces Microglia Activation and the Release of Proinflammatory Mediators in Rat Brain Via H1R or H4R. *Journal of Neuroimmune Pharmacology* **15**, 280–291 (2020).

115. Apolloni, S., Fabbrizio, P., Parisi, C., Amadio, S. & Volonté, C. Clemastine Confers Neuroprotection and Induces an Anti-Inflammatory Phenotype in SOD1G93A Mouse Model of Amyotrophic Lateral Sclerosis. *Molecular Neurobiology* **53**, 518–531 (2016).

116. Xia, R. *et al.* Cryo-EM structure of the human histamine H1 receptor/Gq complex. *Nature Communications* **12**, 2086 (2021).

117. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature 2021 596:7873* **596**, 583–589 (2021).

118. Zou, S. & Kumar, U. Cannabinoid Receptors and the Endocannabinoid System: Signaling and Function in the Central Nervous System. *International Journal of Molecular Sciences* **19**, 833 (2018).

119. Kaminski, N. E. Immune regulation by cannabinoid compounds through the inhibition of the cyclic AMP signaling cascade and altered gene expression. *Biochemical Pharmacology* **52**, 1133–1140 (1996).

120. Giacoppo, S. & Mazzon, E. Can cannabinoids be a potential therapeutic tool in amyotrophic lateral sclerosis? *Neural Regeneration Research* **11**, 1896 (2016).

121. Urbi, B. *et al.* Effects of cannabinoids in Amyotrophic Lateral Sclerosis (<scp>ALS</scp>) murine models: a systematic review and meta-analysis. *Journal of Neurochemistry* **149**, 284–297 (2019).

122. Shoemaker, J. L., Seely, K. A., Reed, R. L., Crow, J. P. & Prather, P. L. The CB2 cannabinoid agonist AM-1241 prolongs survival in a transgenic mouse model of amyotrophic lateral sclerosis when initiated at symptom onset. *Journal of Neurochemistry* **101**, 87–98 (2006).

123. Kim, K., Moore, D. H., Makriyannis, A. & Abood, M. E. AM1241, a cannabinoid CB2 receptor selective compound, delays disease progression in a mouse model of amyotrophic lateral sclerosis. *European Journal of Pharmacology* **542**, 100–105 (2006).

124. Bilsland, L. G. *et al.* Increasing cannabinoid levels by pharmacological and genetic manipulation delays disease progression in SOD1 mice. *The FASEB Journal* **20**, 1003–1005 (2006).

125. Hua, T. *et al.* Activation and Signaling Mechanism Revealed by Cannabinoid Receptor-Gi Complex Structures. *Cell* **180**, 655-665.e18 (2020).

126. Xing, C. *et al.* Cryo-EM Structure of the Human Cannabinoid Receptor CB2-Gi Signaling Complex. *Cell* **180**, 645-654.e13 (2020).

127. Reader, J., Holt, D. & Fulton, A. Prostaglandin E2 EP receptors as therapeutic targets in breast cancer. *Cancer and Metastasis Reviews* **30**, 449–463 (2011).

128. Johansson, T., Narumiya, S. & Zeilhofer, H. U. Contribution of peripheral versus central EP1 prostaglandin receptors to inflammatory pain. *Neuroscience Letters* **495**, 98–101 (2011).

129. Fedyk, E. R. & Phipps, R. P. Prostaglandin E2 receptors of the EP2 and EP4 subtypes regulate activation and differentiation of mouse B lymphocytes to IgE-secreting cells. *Proceedings of the National Academy of Sciences* **93**, 10978–10983 (1996).

130. Watanabe, K. *et al.* Role of the prostaglandin E receptor subtype EP1 in colon carcinogenesis. *Cancer research* **59**, 5093–6 (1999).

131. Sun, X. & Li, Q. Prostaglandin EP2 receptor: Novel therapeutic target for human cancers (Review). *International Journal of Molecular Medicine* (2018) doi:10.3892/ijmm.2018.3744.

132. Takeuchi, K., Kato, S. & Amagase, K. Prostaglandin EP Receptors Involved in Modulating Gastrointestinal Mucosal Integrity. *Journal of Pharmacological Sciences* **114**, 248–261 (2010).

133. Yang, T. & Du, Y. Distinct Roles of Central and Peripheral Prostaglandin E2 and EP Subtypes in Blood Pressure Regulation. *American Journal of Hypertension* **25**, 1042–1049 (2012).

134. Friedman, E. A., Ogletree, M. L., Haddad, E. V. & Boutaud, O. Understanding the role of prostaglandin E2 in regulating human platelet activity in health and disease. *Thrombosis Research* **136**, 493–503 (2015).

135. TAKASAKI, I. *et al.* Involvement of cyclooxygenase-2 and EP prostaglandin receptor in acute herpetic but not postherpetic pain in mice. *Neuropharmacology* **49**, 283–292 (2005).

136. Xu, S., Zhou, W., Ge, J. & Zhang, Z. Prostaglandin E2 receptor EP4 is involved in the cell growth and invasion of prostate cancer via the cAMP-PKA/PI3K-Akt signaling pathway. *Molecular Medicine Reports* (2018) doi:10.3892/mmr.2018.8415.

137. Iłżecka, J. Prostaglandin E2 is increased in amyotrophic lateral sclerosis patients. *Acta Neurologica Scandinavica* **108**, 125–129 (2003).

138. Kosuge, Y. *et al.* Generation of Cellular Reactive Oxygen Species by Activation of the EP2 Receptor Contributes to Prostaglandin E2-Induced Cytotoxicity in Motor Neuron-Like NSC-34 Cells. *Oxidative Medicine and Cellular Longevity* **2020**, 1–14 (2020).

139. Bilak, M. *et al.* PGE2 receptors rescue motor neurons in a model of amyotrophic lateral sclerosis. *Annals of Neurology* **56**, 240–248 (2004).

140. Qu, C. *et al.* Ligand recognition, unconventional activation, and G protein coupling of the prostaglandin E_2 receptor EP2 subtype. *Science Advances* **7**, (2021).

141. Morimoto, K. *et al.* Crystal structure of the endogenous agonist-bound prostanoid receptor EP3. *Nature Chemical Biology* **15**, 8–10 (2019).

142. Nojima, S. *et al.* Cryo-EM Structure of the Prostaglandin E Receptor EP4 Coupled to G Protein. *Structure* **29**, 252-260.e6 (2021).

143. Umetsu, Y. *et al.* Structural difference of vasoactive intestinal peptide in two distinct membrane-mimicking environments. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* **1814**, 724–730 (2011).

144. Harmar, A. J. *et al.* Pharmacology and functions of receptors for vasoactive intestinal peptide and pituitary adenylate cyclase-activating polypeptide: IUPHAR Review 1. *British Journal of Pharmacology* **166**, 4–17 (2012).

145. Couvineau, A. & Laburthe, M. VPAC receptors: structure, molecular pharmacology and interaction with accessory proteins. *British Journal of Pharmacology* **166**, 42–50 (2012).

146. Staines, D. R. Are multiple sclerosis and amyotrophic lateral sclerosis autoimmune disorders of endogenous vasoactive neuropeptides? *Medical Hypotheses* **70**, 413–418 (2008).

147. Iwasaki, Y., Ikeda, K., Ichikawa, Y. & Igarashi, O. Vasoactive intestinal peptide influences neurite outgrowth in cultured rat spinal cord neurons. *Neurological Research* **23**, 851–854 (2001).

148. Solés-Tarrés, I., Cabezas-Llobet, N., Vaudry, D. & Xifró, X. Protective Effects of Pituitary Adenylate Cyclase-Activating Polypeptide and Vasoactive Intestinal Peptide Against Cognitive Decline in Neurodegenerative Diseases. *Frontiers in Cellular Neuroscience* **14**, (2020).

149. Waschek, J. VIP and PACAP: neuropeptide modulators of CNS inflammation, injury, and repair. *British Journal of Pharmacology* **169**, 512–523 (2013).

150. Martínez, C. *et al.* A Clinical Approach for the Use of VIP Axis in Inflammatory and Autoimmune Diseases. *International Journal of Molecular Sciences* **21**, 65 (2019).

151. Duan, J. *et al.* Cryo-EM structure of an activated VIP1 receptor-G protein complex revealed by a NanoBiT tethering strategy. *Nature Communications* **11**, 4121 (2020).

152. Niswender, C. M. & Conn, P. J. Metabotropic Glutamate Receptors: Physiology, Pharmacology, and Disease. *Annual Review of Pharmacology and Toxicology* **50**, 295–322 (2010).

153. Anneser, J. M. H., Chahli, C., Ince, P. G., Borasio, G. D. & Shaw, P. J. Glial Proliferation and Metabotropic Glutamate Receptor Expression in Amyotrophic Lateral Sclerosis. *Journal of Neuropathology & Experimental Neurology* **63**, 831–840 (2004).

154. Milanese, M. *et al.* Knocking down metabotropic glutamate receptor 1 improves survival and disease progression in the SOD1G93A mouse model of amyotrophic lateral sclerosis. *Neurobiology of Disease* **64**, 48–59 (2014).

155. Rossi, D. *et al.* Focal degeneration of astrocytes in amyotrophic lateral sclerosis. *Cell Death & Differentiation* **15**, 1691–1700 (2008).

156. Anneser, J. M. H., Chahli, C. & Borasio, G. D. Protective effect of metabotropic glutamate receptor inhibition on amyotrophic lateral sclerosis–cerebrospinal fluid toxicity in vitro. *Neuroscience* **141**, 1879–1886 (2006).

157. Crupi, R., Impellizzeri, D. & Cuzzocrea, S. Role of Metabotropic Glutamate Receptors in Neurological Disorders. *Frontiers in Molecular Neuroscience* **12**, (2019).

158. Zhang, J. *et al.* Structural insights into the activation initiation of full-length mGlu1. *Protein & Cell* **12**, 662–667 (2021).

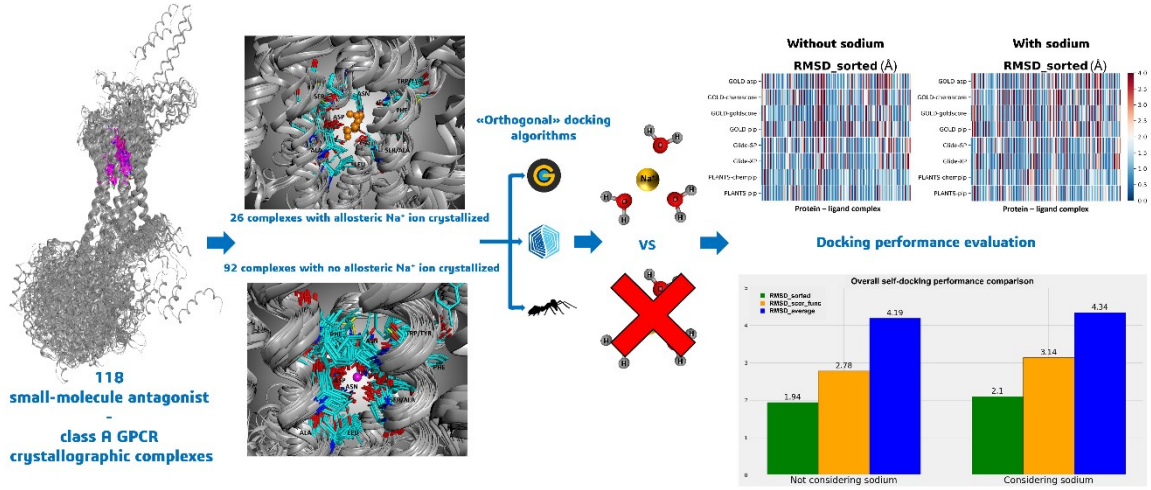
159. Seven, A. B. *et al.* G-protein activation by a metabotropic glutamate receptor. *Nature* **595**, 450–454 (2021).

160. Lin, S. *et al.* Structures of Gi-bound metabotropic glutamate receptors mGlu2 and mGlu4. *Nature* **594**, 583–588 (2021).

Sodium or Not Sodium: Should Its Presence Affect the Accuracy of Pose Prediction in Docking GPCR Antagonists?

Davide Bassani, **Matteo Pavan**, Mattia Sturlese and Stefano Moro

Bassani, D., Pavan, M., Sturlese, M. & Moro, S. Sodium or Not Sodium: Should Its Presence Affect the Accuracy of Pose Prediction in Docking GPCR Antagonists? *Pharmaceuticals* **15**, 346 (2022).



Abstract

The function of the allosteric sodium ion in stabilizing the inactive form of GPCRs has been extensively described in the past decades. Its presence has been reported to be essential for the binding of antagonist molecules in the orthosteric site of these very important therapeutical targets. Among the GPCR–antagonist crystal structures available, in most cases, the sodium ion could not be experimentally resolved, obliging computational scientists using GPCRs as targets for virtual screening to ask: “Should the sodium ion affect the accuracy of pose prediction in docking GPCR antagonists?” In the present study, we examined the performance of three orthogonal docking programs in the self-docking of GPCR antagonists to try to answer this question. The results of the present work highlight that if the sodium ion is resolved in the crystal structure used as the target, it should also be taken into account during the docking calculations. If the crystallographic studies were not able to resolve the sodium ion then no advantage would be obtained if this is manually inserted in the virtual target. The outcomes of the present analysis are useful for

researchers exploiting molecular docking-based virtual screening to efficiently identify novel GPCR antagonists.

1. Introduction

G protein-coupled receptors (GPCRs) represent one of the most important protein superfamilies encoded by the human genome. The members of this protein superfamily (more than 800 entities[1]) have been proven to perform a great variety of biological functions in the organism. Among these, very remarkable are the regulation of senses (e.g., smell, taste, gustatory), the regulation of the nervous and immune systems, homeostasis modulation, pain control, and mood balancing[2]. Indeed, it becomes clear why GPCRs are one of the most interesting protein superfamilies for drug discovery, with more than 160 validated drug targets among them[3]. The fact that encourages the scientific community in putting efforts into the research about GPCRs is their huge therapeutic potential. At the present date, about 35% of the FDA-approved drugs are directed towards a GPCR[4,5], and more than 300 molecules are currently in clinical trials, with near one-fifth targeting a novel GPCR protein[4]. These data make clear that the drug discovery research in this field is very active, and much about this superfamily of proteins has yet to be understood.

GPCRs are cataloged into six classes based on their sequence and function similarities: class A (rhodopsin-like receptors), class B (known as “secretin family”), class C (metabotropic glutamate receptors), class D (fungal mating pheromone receptors), class E (cyclic adenosine monophosphate (cAMP) receptors), and class F (frizzled and smoothened receptors)[6]. All GPCRs share a similar organization in their three-dimensional structure; they are membrane protein receptors constituted of a transmembrane domain formed by seven α-helices (7TM domain), which are linked by three extracellular and three intracellular loops (three ECLs and three ICLs, respectively). The N-terminal (N-ter) domain is located in the extracellular side, while the C-terminal (C-ter) is found intracellularly. The functions of GPCRs are strongly dependent on their conformation and on the changes of this confirmation in time. They exist in an equilibrium between an active and an inactive state[7], and this balance can be shifted upon ligand binding. Indeed, three main families of GPCR ligands have been reported: agonists, antagonists, and inverse

agonists. The first group of binders shifts the equilibrium towards the active arrangement of the receptor, while inverse agonists exert the opposite effect, increasing the conformational inactive population and decreasing the GPCR basal activity. Antagonists simply bind to the receptor and prevent the binding of other ligands, without affecting the conformational balance of the GPCR[8].

In drug discovery campaigns aimed to find new molecular entities for GPCR binding, several techniques are used to select, prioritize, and optimize the most promising compounds. Computational tools have acquired a very important role in the latest decades for drug design and discovery, strongly reducing both the time and money required to obtain new drug candidates and elucidating the most important features required to achieve a desired therapeutic effect. The approach chosen from computational medicinal chemistry to reach these ambitious goals depends on the data available about the target of interest. The presence of a three-dimensional structure of the drug target implies the possibility to exploit a structure-based drug design (SBDD) procedure, while its absence prompts the prioritization of the ligand-based drug design (LBDD) techniques. SBDD has proven to be very successful through the pharmaceutical history[9,10], with several campaigns leading to approved drugs or the repositioning of existing drugs on different targets. The most applied technique belonging to the SBDD family is surely “molecular docking”[11].

Molecular docking is a computational approach that aims to find the conformation in which a molecule binds to its recognition site, forming a stable complex[12]. Specifically, in the case of drug discovery, the main goal is to elucidate how a ligand (which could be a small molecule, a peptide, or a macromolecule) binds to a biological target of interest (usually a protein or a nucleic acid). Docking algorithms are composed of two main parts: a conformational search algorithm and a scoring function. The first aims to search through the conformational space of the ligand, while the second has the goal of ranking the conformations obtained based on their eligibility for target binding. This fitness evaluation is based on several factors, taking into account different geometrical and energetical parameters. Molecular docking has been successfully applied multiple times for virtual screening (VS) aimed at GPCR drug discovery, both in academic and industrial

environments[13]. In these specific cases, attention must be paid to obtain reasonable results from the VS, tuning the docking experiment with respect to both the specific target and the family of ligands considered. A recent study demonstrated that the results of molecular docking on adenosine receptor A_{2A} change if the sodium ion stabilized in the transmembrane domain is considered or not during the calculations[14]. Specifically, that work highlighted a concordance between the computational data and the literature regarding A_{2A} receptor modulation, showing that docking algorithms tend to more efficiently reproduce antagonists' crystallographic binding modes when the sodium ion also is taken into consideration during the calculations. Indeed, the sodium ion has been reported to be present in the middle of the 7TM region of the receptor in several structures of class A GPCRs, helping stabilize the inactive conformation. The sodium ion, together with its solvation sphere, has been demonstrated to negatively modulate the binding of agonists, without influencing the binding of antagonists[15].

To date, the GPCR group which has prevailed for importance for drug discovery is class A (known as rhodopsin-like receptors), mainly for their centrality in the diseases in which they are involved, as well as for the abundance of resolved structures[16]. These proteins are divided into 19 subfamilies (A1–A19) based on phylogenetic analysis[17], including some receptors which have already become very famous in the pharmaceutical world, such as opioid, adrenergic, histaminergic, cannabinoid, and adenosine receptors.

Our evaluation starts from the already cited work of Margiotta et al.[14] to explore the influence of the allosteric sodium ion when molecular docking experiments for the diverse class A GPCR antagonists are performed. Indeed, we evaluated the performance of three different and orthogonal docking algorithms (GOLD, Glide, and PLANTS) in reproducing the ligand crystallographic pose of protein–ligand complexes involving an antagonist bound to a class A GPCR. We extended the study to the class A subfamilies of which some antagonist–protein experimental structure is available, also taking into consideration the eventually present complexes involving a reverse agonist bound to the orthosteric binding site.

2. Results and Discussion

The complete results of the docking runs are reported in the Supplementary Materials (files "Selfdocking_without_sodium.csv" and "Selfdocking_with_sodium.csv"), while a brief per-protocol report is here described by Tables 1 and 2. A graphical representation of the outcomes of the docking runs is also reported using a colormap representation in Figures 1 and 2. In these plots, the colorimetric scale delineating the RMSD values starts from 0 Å, corresponding to a docking pose perfectly superposable to the crystallographic one (maximum docking performance, represented by the dark blue color), and reaches values of 4 Å or higher (minimum docking performance, all represented by the dark red color), which stands for a very suboptimal overlay between the coordinates of the pose produced and the ones of the crystallographic conformation. The results have been reported using three different metrics: "RMSD_average", which represents the mean RMSD of all the poses obtained; "RMSD_scor_func", which is the average value of the RMSDs obtained by the poses which were top-ranked by the scoring functions in each docking run; and "RMSD_sorted", which represents the mean value of the RMSDs obtained from the poses with the lowest RMSD value in each docking calculation. As mentioned, the analysis of the results has also been executed on each docking program–scoring function pair exploited in the study (Tables 1 and 2). Moreover, to better inspect the effect of the sodium ion in the docking simulations, the analysis has also been applied to separate the group of proteins in which the sodium ion is present in the crystallographic structure considered (26 systems) from the other entries (92 complexes). The per-protocol inspections of these last results are reported in the Supplementary Materials (Tables S2–S5).

Self-Docking Results—Na^+ and H_2O Not Considered			
	RMSD_average (Å)	**RMSD_scor_func (Å)**	**RMSD_sorted (Å)**
GOLD-goldscore	3.60	2.83	1.86
GOLD-chemscore	4.45	3.25	2.46
GOLD-asp	3.87	2.91	2.14

GOLD-plp	4.60	3.48	2.56
Glide-SP	4.16	2.57	1.73
Glide-XP	2.67	2.46	1.89
PLANTS$_{CHEMPLP}$	4.96	2.12	1.35
PLANTS$_{PLP}$	5.18	2.58	1.54
All the molecular docking experiments	**4.19**	**2.78**	**1.94**

Table 1. Table showing the results of the self-docking calculations executed without considering the sodium ion.

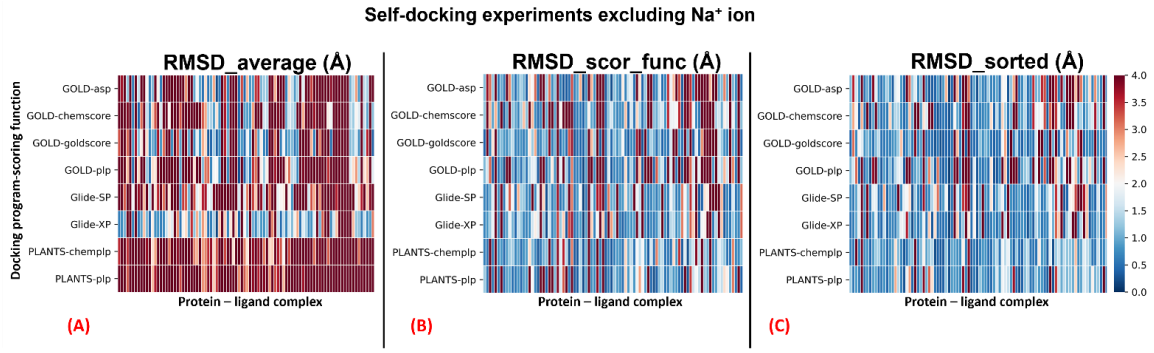


Figure 1. Colormaps show the results of the self-docking calculations not considering the sodium ion within the GPCR receptor 7TM region of the 118 complexes examined in this study. The three plots depict respectively: (**A**) the outcomes derived from the average of the RMSDs of all the poses for each docking run ("RMSD_average"); (**B**) the results obtained just from the RMSD between the crystallographic ligand coordinates and the best-ranked pose from the scoring function ("RMSD_scor_func"); (**C**) the results of the self-docking experiments if just the pose showing the best RMSD value between its coordinates and the crystallographic ones are considered ("RMSD_sorted"). The x-axis enumerates all the different GPCR–antagonist complexes, which are plotted against the different docking program–scoring function pairs used for our study, reported on the y-axis.

Self-Docking Results—Na^+ and H_2O Placed at 4 Å or Nearer to It Both Considered			
	RMSD_average (Å)	**RMSD_scor_func (Å)**	**RMSD_sorted (Å)**
GOLD-goldscore	4.07	3.93	2.33
GOLD-chemscore	4.53	3.90	2.82
GOLD-asp	4.13	3.25	2.40

GOLD-plp	4.52	3.51	2.56
Glide-SP	4.40	2.55	1.67
Glide-XP	2.81	2.62	1.90
$PLANTS_{CHEMPLP}$	5.16	2.60	1.49
$PLANTS_{PLP}$	5.14	2.76	1.62
All the molecular docking experiments	**4.34**	**3.14**	**2.10**

Table 2. Table showing the results of the self-docking calculations executed considering the sodium ion and the water molecules surrounding it.

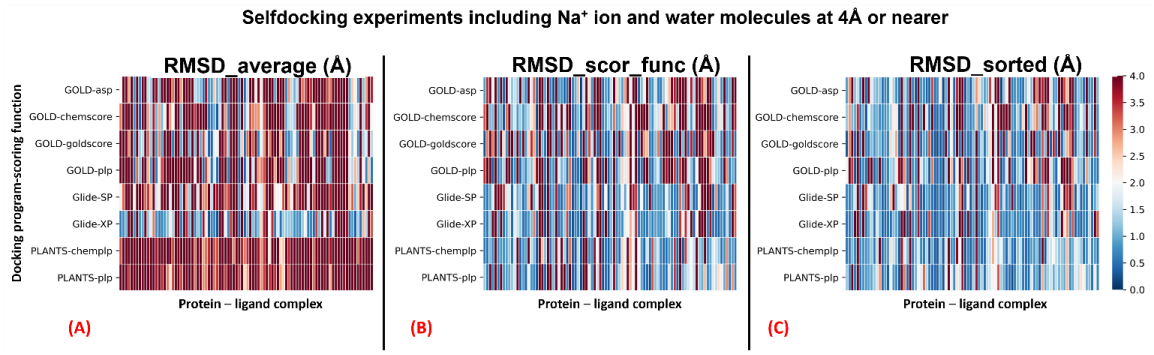


Figure 2. Colormaps show the results of the self-docking calculations executed considering the sodium ion and the water molecules at 4 Å or nearer to it within the GPCR receptor 7TM region of the 118 complexes examined in this study. The three plots depict respectively: (**A**) the outcomes derived from the average of the RMSDs of all the poses for each docking run ("RMSD_average"); (**B**) the results obtained just from the RMSD between the crystallographic ligand coordinates and the best-ranked pose from the scoring function ("RMSD_scor_func"); (**C**) the results of the self-docking experiments if just the pose showing the best RMSD value between its coordinates and the crystallographic ones are considered ("RMSD_sorted"). The x-axis enumerates all the different GPCR–antagonist complexes, which are plotted against the different docking program–scoring function pairs used for our study, reported on the y-axis.

The outcomes of our experiment highlight how all the algorithms used show an overall good performance in GPCR–antagonist self-docking. Among the others, the pairs "Glide-SP", "$PLANTS_{CHEMPLP}$", and "$PLANTS_{PLP}$" were always able to produce an "RMSD_sorted" value of less than 2 Å with respect to the crystallographic coordinates. Even if the scoring functions allowed obtaining reasonable RMSD values (as observable from the "RMSD_scor_func" columns in Tables 1 and 2), the solutions are given by them rarely

corresponded to the ones with the lowest RMSD. As expected, an ant colony optimization algorithm such as PLANTS tends to produce poses with less three-dimensional conservation compared to a genetic algorithm such as GOLD or a systematic method such as Glide, and this is evidenced by the higher values of "RMSD_average" given by both docking program–scoring function pairs involving PLANTS. On the other hand, the higher variability in the poses produced could be the reason for the fact that PLANTS can obtain solutions with very low RMSD, as demonstrated by the "RMSD_sorted" results, which are far below 2 Å in all the cases reported in this study (also when the complexes are separated based on the presence of the sodium ion in the original PDB structure, as depicted in Supplementary Materials, Tables S2–S5). GOLD and Glide both performed remarkably, with "goldscore" giving the best results among the scoring functions implemented for GOLD in all the metrics used for the analysis (exception made for the "RMSD_scor_func" value when considering the sodium ion and the water molecules at 4 Å or nearer to it in the calculations). Comparing "Glide-SP" and "Glide-XP" outcomes, even if the first can obtain lower "RMSD_sorted" values, is important to notice that the XP protocol is the overall best performing when considering the "RMSD_average", always giving a value below 3 Å for this parameter. The choice between the two for GPCR antagonist virtual screening (VS) should so be based on the specific case examined. Indeed, "Glide-SP" would be more beneficial in the VS of a GPCR antagonist with already known scaffold and properties (eventually coming from "focused libraries"), while "Glide-XP" would be more effective when a library with molecules characterized by higher diversity is taken into account. When considering the use of "Glide-XP" instead of "Glide-SP" for large VS of high-diversity entities for GPCR antagonism, the medicinal chemists should always consider the higher computational times required for the XP function (passing from the 10 s/compound of "Glide-SP" to about 2 min/compound of "Glide-XP", as reported on the developer's page[18]).

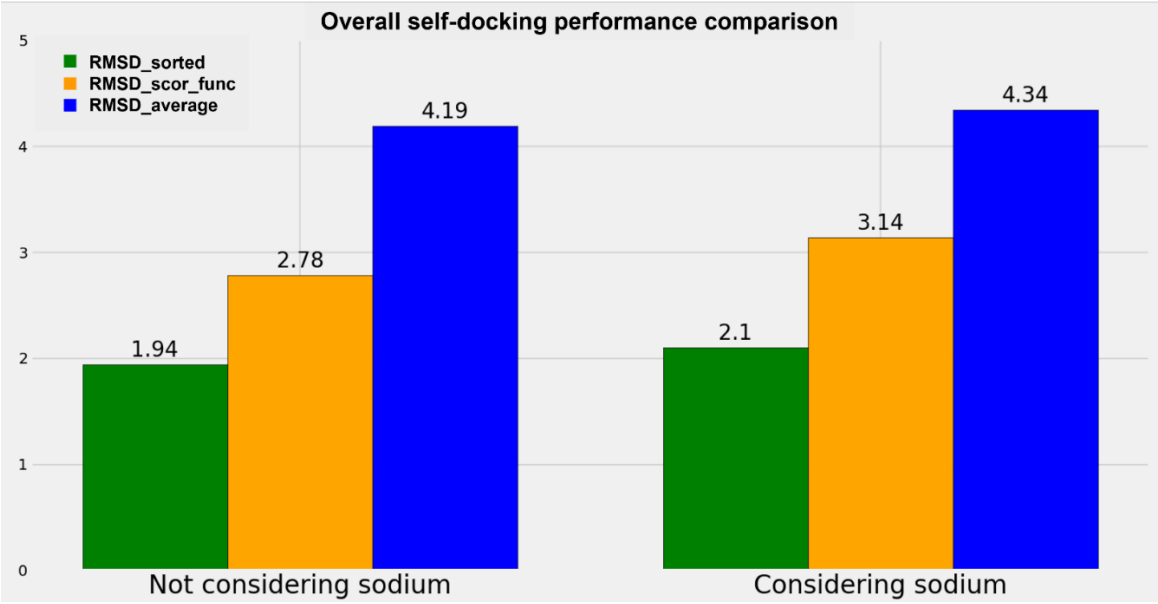


Figure 3. Plot representing the comparison of the overall performance of the different docking algorithms implemented in this study when the sodium ion is not considered during the calculation (on the **left**) and when both the sodium ion and the crystal water molecules at 4 Å or nearer to it are included (on the **right**). The metrics used for the comparison are the "RMSD_average", the "RMSD_scor_func", and the "RMSD_sorted" values already described in the present study.

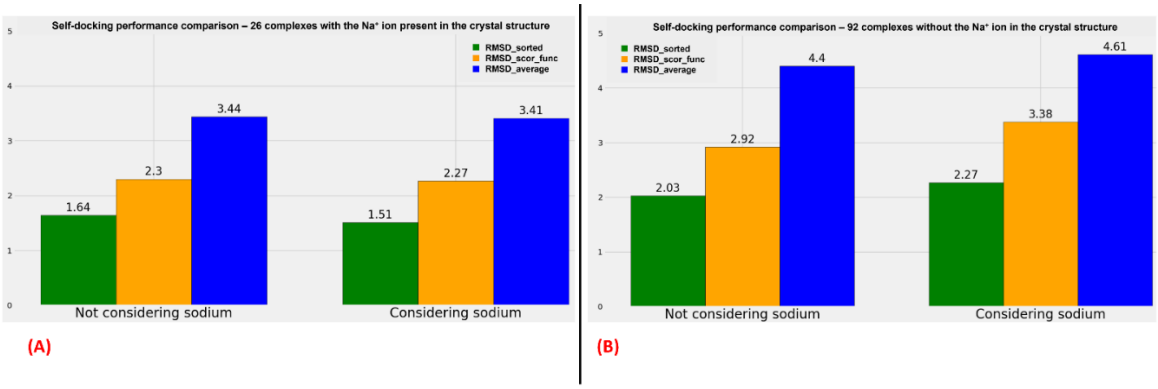


Figure 4. Panel (**A**) Plot representing the comparison of the overall performance of the different docking algorithms implemented in this study when the sodium ion is not considered during the calculation (on the **left**) and when both the sodium ion and the crystal water molecules at 4 Å or nearer to it are included (on the **right**), focusing just on the 26 GPCR–antagonist complexes in which the sodium ion is already present in the crystal structure. Panel (**B**) Graphical representation of the comparison of the overall performance of the different docking algorithms used in this study when the sodium ion is not considered during the calculation (on the **left**) and when both the sodium ion and the crystal water molecules at 4 Å or nearer to it are included (on the **right**), focusing just on the 92 GPCR–antagonist complexes in which the sodium ion is not present on the original crystal structure.

A graphical representation of the comparison between the performance of the algorithms when the sodium ions are considered or not is reported in Figure 3, while two analog diagrams are reported in Figure 4 (based on Tables S2–S5, which can be examined in the Supplementary Materials) to give a more immediate visualization of the outcomes divided based on whether the sodium ion is present in the original crystallographic complexes.

The results obtained show a small decrease in all the metrics used when the sodium ion is not considered in the docking runs. Specifically, the decrement in “RMSD_average” is 3.46%, the reduction in “RMSD_scor_func” is 11.47% (this higher value has to be attributed to the scoring functions), and the diminution in “RMSD_sorted” is 7.62%. Considering the decreases in the order of magnitude of the RMSD of the docking results (which is around the very promising value of 2 Å for the best pose obtained and around 3 Å for the best solution given by the scoring functions), we can conclude that no big difference in the docking performance regarding a GPCR–antagonist system is achieved if the sodium ion is taken into account during the calculation.

The metrics used for the comparisons are the “RMSD_average”, the “RMSD_scor_func”, and the “RMSD_sorted” values already described in the present study. The overall figure is useful to compare the performance of the docking algorithms when the sodium is present in the original crystal structure and when it is not.

With an examination of the data coming from Tables S2–S5, of which the comparison of the overall results is plotted in Figure 4, we can see an analog trend of the outcomes when sodium is considered or not during the docking runs. It is interesting to notice that when the sodium is already present in the crystal structure, the RMSD values obtained from the docking poses tend to be more promising, but this has to be weighed on the fact that, in that case, the exact position of the sodium is known, and so the possible error coming from the manual placing of this alkaline ion in the 7TM region is removed. Moreover, a comparison should be made very carefully when data coming from only 26 complexes (the ones having the sodium crystallographically resolved) are juxtaposed to the ones derived from a larger set of 92 structures (the complexes in which the sodium ion is missing in the crystal structure).

On the contrary, important information is obtainable if the comparison is limited between the two groups of proteins. Indeed, as shown in Figure 4 (as well as Tables S2–S5), if the sodium ions are already present in the GPCR–antagonist crystallographic complex, no relevant difference can be noticed between the results coming from the docking runs in which the alkaline ion is considered and the ones derived from the calculation in which also sodium and the water molecules surrounding it are taken into account. Furthermore, the analysis shows a very slight decrease in the RMSD values when the crystallographic sodium and the water molecules at 4 Å or nearer to it are retained during the calculations.

On the other side, more important changes in the RMSD metrics used are highlightable comparing the outcomes of the docking runs for the complexes in which the sodium ion is not present in the crystal structure. In this case, all "RMSD_average", "RMSD_scor_func", and "RMSD_sorted" values show an increase if the alkaline ion with its surrounding water molecules is inserted in the complex and considered during the calculation.

The results of our study show that when performing molecular docking experiments on GPCR antagonists, the sodium ion present in the allosteric 7TM pocket should be considered during the calculation only if it is already present in the crystal structure used as the protein target. If the GPCR on which the research is based does not have antagonist-bound crystallographic structures in which the sodium ion is present, any advantage will be obtained if it is manually placed in its allosteric pocket, and so the execution of the docking calculations without this alkaline ion should be considered. A possible reason for this behavior could be related to the fact that the small benefit coming from taking into account the sodium ion when performing the virtual screening would be demolished by the inevitable error coming from the manual placing of this ion in its allosteric pocket. We also assert that this type of uncertainty would not be canceled even if advanced computational approaches would be used for sodium placement, because of the errors that these techniques inexorably bring with them.

The importance of the allosteric sodium ion for the binding of antagonists to class A GPCRs has been extensively described in the literature[15,19]. Moreover, as observable from Figure 5, the side chains of the amino acids located in the allosteric sodium binding site are

conservatively orientated towards the alkaline ion location even if all the structures represented do not have the sodium ion present in the crystallographic complex, showing that this alkaline ion has to be present in its site to guarantee the antagonist activity.

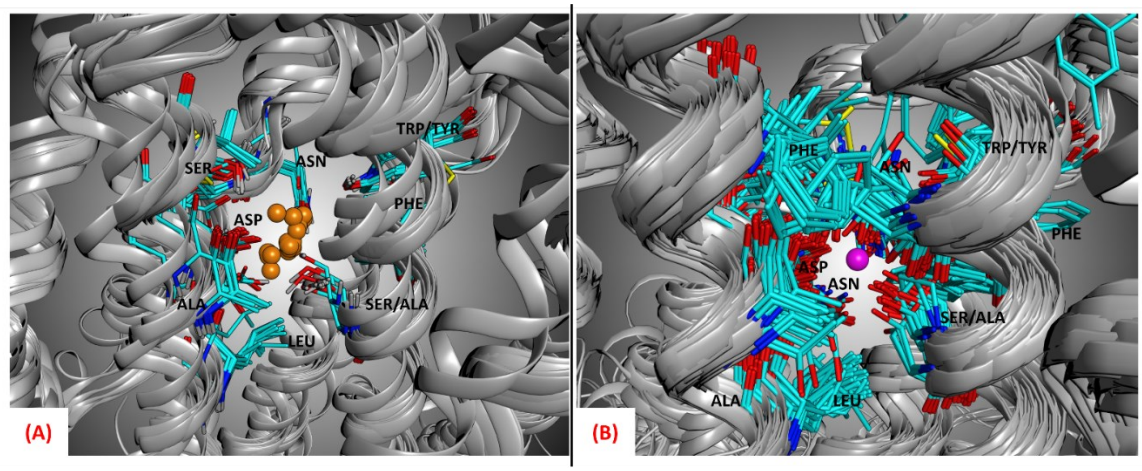


Figure 5. (**A**) Representation of the allosteric sodium binding site of all the 26 GPCR–antagonist complexes considered in this study which had the sodium ion present in their crystallographic structures. As depicted, the position of the sodium (the orange atoms in the image) is well conserved, as are the type and orientation of the side chains of the amino acids surrounding it, which help stabilize the alkaline ion. (**B**) Representation of the allosteric sodium binding site of all the 92 GPCR–antagonist complexes considered in this study in which the sodium ion is not present in the crystallographic structures. As can be seen, the potential position of the sodium (the purple atom in the image, which comes from the crystal upon which all proteins have been superposed, PDB code: 5IU4) is well conserved, as are the type and orientation of the side chains of the amino acids surrounding it, which help to stabilize the alkaline ion in its 7TM allosteric pocket.

Molecular docking techniques are known for not being able to distinguish agonism from antagonism. Indeed, this family of computational approaches has the goal of highlighting the potential binders for a target, but their results cannot be related to a specific type of outcomes that this binding will have on the target itself (which has to be evaluated by the medicinal chemistry experts, based on their expertise and the communication with other professionals of the pharmaceutical world). These limitations of the technique of molecular docking may be the reason for the very low difference between the results coming from the cases in which sodium is considered or not in the calculations.

3. Materials and Methods

For each of the 19 GPCR class A subfamilies, the crystal structures available in the Protein Data Bank[20] (PDB, latest access 15 January 2022) were inspected. Each entry with a human GPCR protein complexed with a small molecule orthosteric antagonist crystallized together was selected to build the starting database of our study. If multiple crystals of a protein bound to the same ligand existed, only the highest resolution crystal with the sodium ion present was selected. To give a more comprehensive panoramic of the role of allosteric sodium in GPCR binding, the structures with an inverse agonist were also considered for this study (e.g., 6K1Q, 7F83, 7B6W, 7BVQ). In the end, 118 protein–ligand complexes involving a GPCR and a small molecule antagonist were obtained (a comprehensive list is reported in the Supplementary Materials, Table S1).

The 118 complexes were downloaded from the PDB and imported into Molecular Operating Environment (MOE) suite[21], the main molecular modeling program that we used in this work. Each system was then prepared with a protocol involving the tools included in the MOE package. First, the "Structure Preparation" program was used to rebuild the small missing loops in the structures and to adequately select the orientation of alternate crystallographic residues based on occupancy. Then, the most proper protonation state for each amino acid was determined with the "Protonate 3D" tool, setting 7.4 as the pH value for the environment. Subsequently, the added hydrogen atoms were minimized with the AMBER10:EHT[22] force field implemented in MOE. Finally, each non-protein, non-ligand, and non-sodium molecule was deleted from the systems, with an exception made for the water molecules solvating the sodium ion (we used 4 Å as the cut-off radius), when present.

The systems were then separated based on whether they had the sodium ion crystallized in their original PDB structure. Among all the complexes downloaded, 26 already had the sodium present in the crystal, while 92 did not (the distinction is highlighted in the Supplementary Materials, Table S1). All the systems in which the sodium was not present were properly treated, inserting the sodium ion with its solvation water molecules. The position of the sodium and the water molecules was chosen according to the PDB crystal 5IU4, the complex with the best resolution, R-value, and R-free value balance among all

the entries considered. This choice was also supported by the fact that when superposing all the 7TM regions of the protein–ligand systems with the sodium crystallized, the position of this alkaline ion is very conservative, as observable from Figure 5 (the average RMSD between the coordinates of the sodium ions and the sodium ion of the reference structure 5IU4 was calculated to be 0.75 Å). It is important to mention that the crystal 5IU4 was not considered for the docking calculations because, even if its resolution is optimal, it is significantly mutated in the 7TM region. Another ADORA2 crystal structure bearing the same ligand (ZMA), 6LPJ, shows a very similar resolution (1.80 Å versus the 1.72 Å of 5IU4) and does not show mutations in the 7TM domain.

Our self-docking approach consisted of the separation of each ligand from its crystallographic GPCR structure, its preparation, and its molecular docking inside the orthosteric binding site with three different orthogonal programs, namely GOLD[23] (based on a genetic algorithm, developed and licensed by CCDC), Glide[24](a systematic docking program developed and distributed by Schrödinger), and PLANTS[25] (an ant colony optimization algorithm developed by the University of Tübingen). For each of the programs, all the scoring functions supported were used. Specifically, GOLD was used in four different parallel runs, applying the scoring functions “goldscore”, “chemscore”, “asp”, and “plp”. The two Glide calculations for each ligand were executed first with “Glide-SP” and then with “Glide-XP”, while the docking runs with PLANTS exploited the scoring functions “$PLANTS_{CHEMPLP}$” and “$PLANTS_{PLP}$”. For each program–scoring function pair, five poses were produced for each ligand, and each of those was compared with the crystallographic pose to calculate the root-mean-square deviation (RMSD) between the coordinates of the two conformations.

This whole procedure was executed twice, first setting the docking programs to not consider the sodium ion and the water molecules solvating it and then setting the programs to take into account both sodium and the water molecules placed at 4 Å or nearer to the alkaline ion.

4. Conclusions

In the present study, we examined the effect of considering the allosteric sodium ion when molecular docking approaches are applied to GPCR antagonists. To accomplish our task, we collected 118 GPCR–antagonist complexes, both with and without the sodium ion present in the crystallographic structure. For the systems in which this alkaline ion was not present, a manual insertion of the sodium and its surrounding water molecules was executed based on superposition with a very high resolution structure (PDB: 5IU4), after having established that the position of the ion is very conservative in the GPCR–antagonist crystals. Then, we executed self-docking experiments of the orthosteric GPCR ligands with three orthogonal docking programs (GOLD, Glide, and PLANTS) both considering and not considering the sodium ion and its surrounding water molecules. What emerged from the present work is the finding that the performance of the docking programs (enucleated in three different metrics, “RMSD_average”, “RMSD_sorted”, and “RMSD_scor_func”) does not significantly change between the two cited scenarios. Going deeper into the analysis of the results, we highlighted that a small increment in the docking programs’ performance is observable if the sodium ion is kept during the docking runs just for those crystal structures in which the alkaline ion was resolved, while for the other complexes the trend is the opposite, favoring the solution of not considering sodium during the docking calculations. The outcomes of the present work are helpful to increase the knowledge about the performance of docking programs when applied to research about GPCR antagonists, and we are confident that the pharmaceutical experts that are putting effort into this fascinating field will benefit from our work.

References

1. Gacasan, S. B., Baker, D. L. & Parrill, A. L. G protein-coupled receptors: the evolution of structural insight. *AIMS Biophys* **4**, 491–527 (2017).

2. Heng, B. C., Aubel, D. & Fussenegger, M. An overview of the diverse roles of G-protein coupled receptors (GPCRs) in the pathophysiology of various human diseases. *Biotechnol Adv* **31**, 1676–1694 (2013).

3. Yang, D. *et al.* G protein-coupled receptors: structure- and function-based drug discovery. *Signal Transduct Target Ther* **6**, (2021).

4. Hauser, A. S., Attwood, M. M., Rask-Andersen, M., Schiöth, H. B. & Gloriam, D. E. Trends in GPCR drug discovery: new agents, targets and indications. *Nat Rev Drug Discov* **16**, 829–842 (2017).

5. Insel, P. A. *et al.* GPCRomics: An Approach to Discover GPCR Drug Targets. *Trends Pharmacol Sci* **40**, 378–387 (2019).

6. Lee, Y., Basith, S. & Choi, S. Recent Advances in Structure-Based Drug Design Targeting Class A G Protein-Coupled Receptors Utilizing Crystal Structures and Computational Simulations. *J Med Chem* **61**, 1–46 (2018).

7. Weis, W. I. & Kobilka, B. K. The Molecular Basis of G Protein-Coupled Receptor Activation. *Annu Rev Biochem* **87**, 897–919 (2018).

8. Syrovatkina, V., Alegre, K. O., Dey, R. & Huang, X. Y. Regulation, Signaling, and Physiological Functions of G-Proteins. *J Mol Biol* **428**, 3850–3868 (2016).

9. Meng, X.-Y., Zhang, H.-X., Mezei, M. & Cui, M. Molecular Docking: A Powerful Approach for Structure-Based Drug Discovery. *Current Computer Aided-Drug Design* **7**, 146–157 (2012).

10. Anderson, A. C. The Process of Structure-Based Drug Design. *Chem Biol* **10**, 787–797 (2003).

11. Wang, G. & Zhu, W. Molecular docking for drug discovery and development: a widely used approach but far from perfect. *http://dx.doi.org/10.4155/fmc-2016-0143* **8**, (2016).

12. Lengauer, T. & Rarey, M. Computational methods for biomolecular docking. *Curr Opin Struct Biol* **6**, 402–406 (1996).

13. Ballante, F., Kooistra, A. J., Kampen, S., de Graaf, C. & Carlsson, J. Structure-Based Virtual Screening for Ligands of G Protein-Coupled Receptors: What Can Molecular Docking Do for You? *Pharmacol Rev* **73**, 527–656 (2021).

14. Margiotta, E., Deganutti, G. & Moro, S. Could the presence of sodium ion influence the accuracy and precision of the ligand-posing in the human A2A adenosine receptor orthosteric binding site using a molecular docking approach? Insights from Dockbench. *J Comput Aided Mol Des* **32**, 1337–1346 (2018).

15. Katritch, V. *et al.* Allosteric sodium in class A GPCR signaling. *Trends Biochem Sci* **39**, 233–244 (2014).

16. Basith, S. *et al.* Exploring G protein-coupled receptors (GPCRs) ligand space via cheminformatics approaches: Impact on rational drug design. *Front Pharmacol* **9**, 128 (2018).

17. Joost, P. & Methner, A. Phylogenetic analysis of 277 human G-protein-coupled receptors as a tool for the prediction of orphan receptor ligands. *Genome Biol* **3**, 1–16 (2002).

18. Friesner, R. A. *et al.* Extra precision glide: Docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. *J Med Chem* **49**, 6177–6196 (2006).

19. Gentry, P. R., Sexton, P. M. & Christopoulos, A. Novel Allosteric Modulators of G Protein-coupled Receptors *. *Journal of Biological Chemistry* **290**, 19478–19488 (2015).

20. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res* **28**, 235–242 (2000).

21. Molecular Operating Environment (MOE), 2019.01; Chemical Computing Group ULC, 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, 2021. https://www.chemcomp.com/Research-Citing_MOE.htm.

22. Case, D. A. *et al.* Amber 10. (2008).

23. Jones, G., Willett, P., Glen, R. C., Leach, A. R. & Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J Mol Biol* **267**, 727–748 (1997).

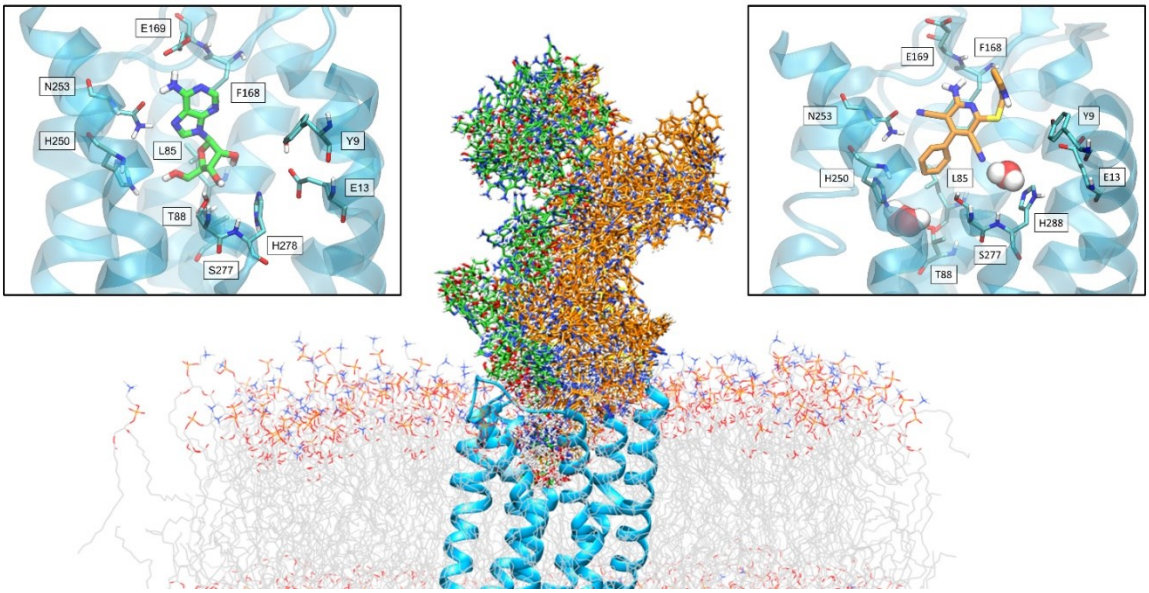
24. Friesner, R. A. *et al.* Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J Med Chem* **47**, 1739–1749 (2004).

25. Korb, O., Stützle, T. & Exner, T. E. PLANTS: Application of Ant Colony Optimization to Structure-Based Drug Design. 247–258 (2006) doi:10.1007/11839088_22.

Ribose and Non-Ribose A2A Adenosine Receptor Agonists: Do They Share the Same Receptor Recognition Mechanism?

Giovanni Bolcato, **Matteo Pavan**, Davide Bassani, Mattia Sturlese and Stefano Moro

Bolcato, G., Pavan, M., Bassani, D., Sturlese, M. & Moro, S. Ribose and Non-Ribose A2A Adenosine Receptor Agonists: Do They Share the Same Receptor Recognition Mechanism? *Biomedicines 2022, Vol. 10, Page 515* **10**, 515 (2022).



Abstract

Adenosine receptors have been a promising class of targets for the development of new therapies for several diseases. In recent years a renewed interest in this field has risen, thanks to the implementation of a novel class of agonists that lack the ribose moiety, once considered essential for the agonistic profile. Recently an X-ray crystal structure of A_{2A} Adenosine Receptor has been solved providing insights about the receptor activation from this novel class of agonists. Starting from this structural information, we have performed supervised molecular dynamics (SuMD) simulations to investigate the binding pathway of a non-nucleoside adenosine receptor agonist, as well as one of three classic agonists. Furthermore, we analyzed the possible role of water molecules in receptor activation.

1. Introduction

Adenosine is the endogenous agonist of a group of class A G protein-coupled receptors (GPCRs) named adenosine receptors (AR); four receptors belong to this family: A_1, A_{2A}, A_{2B}, and A_3. A_1 and A_3 are generally coupled to $G_{\alpha i}$ protein (so they inhibit adenylate cyclase enzyme upon activation), A_{2A} and A_{2B} are coupled to $G_{\alpha s}$ protein (and therefore stimulate the production of cAMP upon activation)[1].

AR are targets of interest for the treatment of several diseases[2]: Parkinson's disease[3,4], asthma[5], pain treatment[6], several cancer types[7], and cardiovascular diseases[8]. Despite this broad range of potential therapeutic applications, only two AR ligands have been approved: the A_{2A} antagonist Istradefylline, approved for the treatment of Parkinson's disease, and the A_{2A} agonist Regadenoson, used as a coronary vasodilator. One of the main problems in the translation of AR ligands into therapeutic agents is the presence of unacceptable side effects due to the lack of selectivity of the drug candidates among different AR subtypes as well as off-target effects[9].

Progress was made in the field of AR agonists as therapeutic agents[10] with the publication by Bayer of some patents regarding non-nucleoside AR agonists[11]. While this novel class of AR agonists presents several advantages over classic adenosine-derived ligands (easier synthesis, improved pharmacokinetics, and oral bioavailability), the AR activation for this class of compounds has been difficult to understand since they lack the ribose moiety which was considered essential for the agonistic profile of AR ligands[12]. Some modifications on this moiety are tolerated and can improve both metabolic stability and potency, but often these alterations on the ribose unit lead to inactive compounds or switch the ligand activity toward an antagonistic profile[13].

To gain some insights on the structural basis of AR activation several site-directed mutagenesis data have been collected over the years[14]. Interestingly it was proven that the non-nucleoside A_{2A} agonist LUF5834 is sensitive to mutagenesis experiments in a different way compared to classic adenosine-like AR agonists like CGS21680[15]. In particular, the agonistic profile of LUF5834 is not affected when Ser-277 and Thr-88 are mutated in alanine (note that the enumeration, as well as in all this work, refers to A_{2A}).

These two residues are essential for the agonistic activity of classic AR agonists along with His-278.

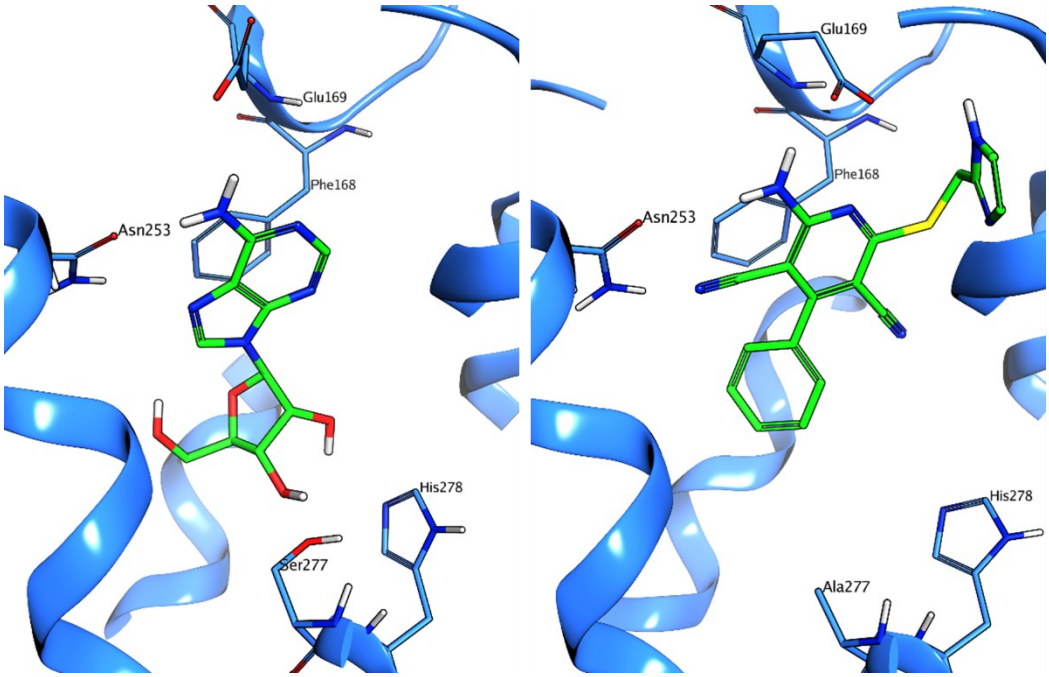


Figure 10. Binding modes of Adenosine (left) and of LUF5833 (right) as observed in X-ray crystal structures (PDB code 2YDO and 7ARO respectively). Please note that the AR used to obtain the crystal structure with LUF5833 presents some thermostabilizing mutations including Ser-277. Other AR agonists, like CGS21680 and NECA, also interact with His-250 and Thr-88. The binding mode of these two ligands can be found in Figure 2.

Fundamental progress in the comprehension of the structural basis of the agonistic action of non-nucleoside agonists has been made recently with the obtainment of the X-ray crystal structure of A_{2A} AR in complex with a close analog of LUF5834, LUF5833[16]. Interestingly the ligand does not form any hydrogen bonds with the abovementioned residues that are considered essential for the activation of A_{2A} AR (Thr-88, Ser-277, His-278).

In Figure 1, a comparison of the binding mode (as observed in X-ray crystal structures) that the endogenous agonist adenosine and of LUF5833 adopt in the orthosteric binding site of A_{2A} AR is reported.

Starting from this structural information, in the present work we have investigated the recognition process of LUF5833 and different classic adenosine-like agonists: CGS21680, NECA, and adenosine itself. The study has been carried out using supervised molecular dynamics simulations (SuMD) in order to gain structural information beyond the observed final bound state. SuMD is indeed a molecular dynamics-based approach that allows the sampling of events involving infrequent particle collision such as protein-ligand binding, without applying any energetic bias to the system.

The comparison between the binding trajectories collected for the two different classes of AR agonists reveals a different recognition pathway. Moreover, a detailed analysis of the behavior of water molecules during the binding event provides some insights into the possible role of the solvent molecules in the activation of the A_{2A} adenosine receptor.

2. Materials and Methods

2.1. System Setup

The three-dimensional structure of the protein-ligand complexes examined in this work (PDB codes: 2YDO, 2YDV, 4UG2, and 7ARO) was retrieved from the Protein Data Bank (PDB) and prepared for subsequent calculations using various tools provided by the Molecular Operating Environment (MOE) suite, version 2019.01[17]. Residues with alternate conformation were assigned to the one with the highest occupancy. Missing hydrogen atoms were added to the system with the Protonate3D tool, assigning each titratable residue to the most probable protonation state at pH = 7.4. Crystallographic water molecules, ions, and other molecules present in the crystallization buffer were then removed, and the ligand was moved away from the binding site into the bulk, at a distance of at least 30 Å from the nearest receptor atom (higher than the cutoff chosen for electrostatic interaction computation).

The system preparation for Supervised Molecular Dynamics (SuMD) simulations was carried out with Visualize Molecular Dynamics (VMD[18]) version 1.9.3. At first, the protein-ligand system was explicitly solvated in a cubic TIP3P[19] water box, ensuring a distance of 15 Å between the box borders and any protein atom. Then, the system charge was neutralized by the addition of sodium and chlorine ions until a physiological concentration

(0.154 M) was reached. Finally, the receptor was embedded in a lipid bilayer consisting of phosphatidylcholine (POPC) units.

From a methodological point of view, one main limitation of the SuMD technique, as is the case for traditional molecular dynamics (MD), is the fact that simulations are carried out assuming fixed protonation states. The prediction of the protonation state of titratable residues relies on a static structure (the crystal complex, which is the starting point for the simulations) and can sometimes be imprecise in those cases where the protein is flexible[20] or the residues are highly buried[21]. Furthermore, the co-existence of protonated and deprotonated states and dynamical processes coupled to a change in protonation states cannot be directly studied if the protonation states are fixed.

A second limitation is represented by the fact that the lipid constitution of the phospholipid membrane does not include the presence of cholesterol, which could exert some form of allosteric modulation on AR[22].

2.2. Molecular Dynamics

All simulations were carried out using the ACEMD[23] molecular dynamics engine. The system was described using parameters from the CHARMM36[24] force field (protein, lipids, ions, and water molecules), while ligand parameters were retrieved from Paramchem[25], a web front-end for the CgenFF[26] force field. If the parameters associated with specific dihedral angles of ligands presented high penalties, these have been parametrized using FFParam[27]. A QM scansion of the dihedral angle has been performed using MP2 level of theory with 6-31G** basis set, then the QM profile has then been fitted to retrieve the new force field parameters.

The simulation protocol consisted of a four-stage equilibration phase, followed by a productive SuMD simulation phase. For both equilibration and productive simulations, the integration timestep was set to 2 fs, the temperature was set to 310 K through a Langevin thermostat (friction coefficient = 0.1 ps^{-1}), the M-SHAKE algorithm was employed to constrain the length of bonds involving hydrogen atoms and the particle-mesh Ewald (PME[28]) was exploited to compute electrostatic interactions (grid length = 1 Å). Finally, a 9.0 Å cutoff was applied to long-term interactions.

2.3. Equilibration phase

Before equilibration MD simulations, 1500 steps of energy minimization using the conjugate-gradient method were performed, to remove clashes and bad contacts within the system. The first three equilibration MD simulations were carried out in the isothermal-isobaric ensemble (NPT), maintaining the system pressure fixed at 1 atm through the Berendsen barostat[29], while the fourth and final one was performed in the isothermal ensemble (NVT).

The first equilibration stage consisted of a 5 ns simulation with 1 Kcal mol^{-1} $Å^{-2}$ harmonic positional constraints applied on each receptor, ligand, and membrane atom. The second equilibration stage consisted of a 10 ns simulation with the same constraints applied only on each protein, ligand, and phosphorus atom. The third equilibration stage consisted of a 5 ns simulation with the same constraints applied only on the protein alpha carbons and on ligand atoms. Finally, a 10 ns equilibration MD simulation was performed without any constraints applied to the system.

2.4. Supervised Molecular Dynamics simulations

SuMD[30] is an enhanced sampling MD method that allows investigating molecular recognition processes at an atomistic level of detail in the nanosecond timescale without any energetic bias. The SuMD code is written in Python 2.7 and mainly exploits the Numpy and ProDy[31] packages to perform geometric supervision over a series of short classic MD trajectories (defined as “SuMD steps”) carried out with the ACEMD engine. As reported in the original publication, each suMD step lasts 600 ps.

During each SuMD step, the distance between the center of mass (i.e., the hypothetical point where the entire mass of an object is assumed to be concentrated) of both the ligand and the binding site is monitored and collected at 5 evenly spaced time intervals. At the end of each step, these data are fitted in a straight line, which is then processed by a tabù-like algorithm: if the line slope is negative (indicating that the ligand is approaching the binding site), the step is considered productive and retained for the generation of the final MD trajectory, while the final state of this simulation is set as the initial state for the successive step. On the contrary, in the case where the slope is positive (indicating that

the ligand is not approaching the binding site), the step is considered not productive and is discarded: in this case, the step is repeated reassigning the velocities through the Langevin thermostat. This process continues until the distance between the two centers of mass gets below 5 Å: from that point on, the supervision is turned off and the simulation proceeds as a classic MD simulation for other 30 SuMD steps.

2.5. Trajectory analysis

A per-residue energetic analysis was performed using an in-house developed Python script.

At first, the MDAnalysis[32,33] Python package is exploited to parse each MD trajectory and compute the number of contacts between the ligand and each protein residue, using a cutoff distance of 4.5 Å.

Afterward, the interaction energy (defined as the sum of the electrostatic and van der Waals contribution) is computed between the ligand and each one of the top 25 most contacted residues alongside each MD trajectory using the NAMD Energy Plugin[34] for VMD (version 1.4).

Finally, a heatmap is generated exploiting the Seaborn Python package: on the horizontal axis the simulation time in nanoseconds is reported, while on the vertical axis the residue name and index are reported for each residue considered for this analysis. The interaction energy is then plotted onto the heatmap using a colormap which ranges from red (indicating positive energy values, *i.e.,* a repulsive interaction) to blue (indicating negative energy values, *i.e.,* an attractive interaction). The first and the third quartile with regards to the distribution of interaction energy values are used as mask values for the heatmap generation.

To inspect the peculiar hydrodynamic profile of ligand LUF5833, the trajectory was analyzed with AquaMMapS[35], an in-house developed tool that allows investigating the behavior of water molecules within a receptor, based on their persistency across an MD trajectory. For this purpose, the simulation box is discretized in a voxel grid and the occupancy value for each cell is calculated as the ratio between the number of frames in which a water molecule occupies that cell and the total number of frames.

3. Results and Discussion

For all the four ligands a SuMD trajectory where the crystallographic observed binding mode is well reproduced has been collected. In Figure 2 the crystallographic binding mode and the final pose obtained with SuMD for the four ligands under examination are reported: as it can be seen the experimentally observed binding mode is well reproduced.

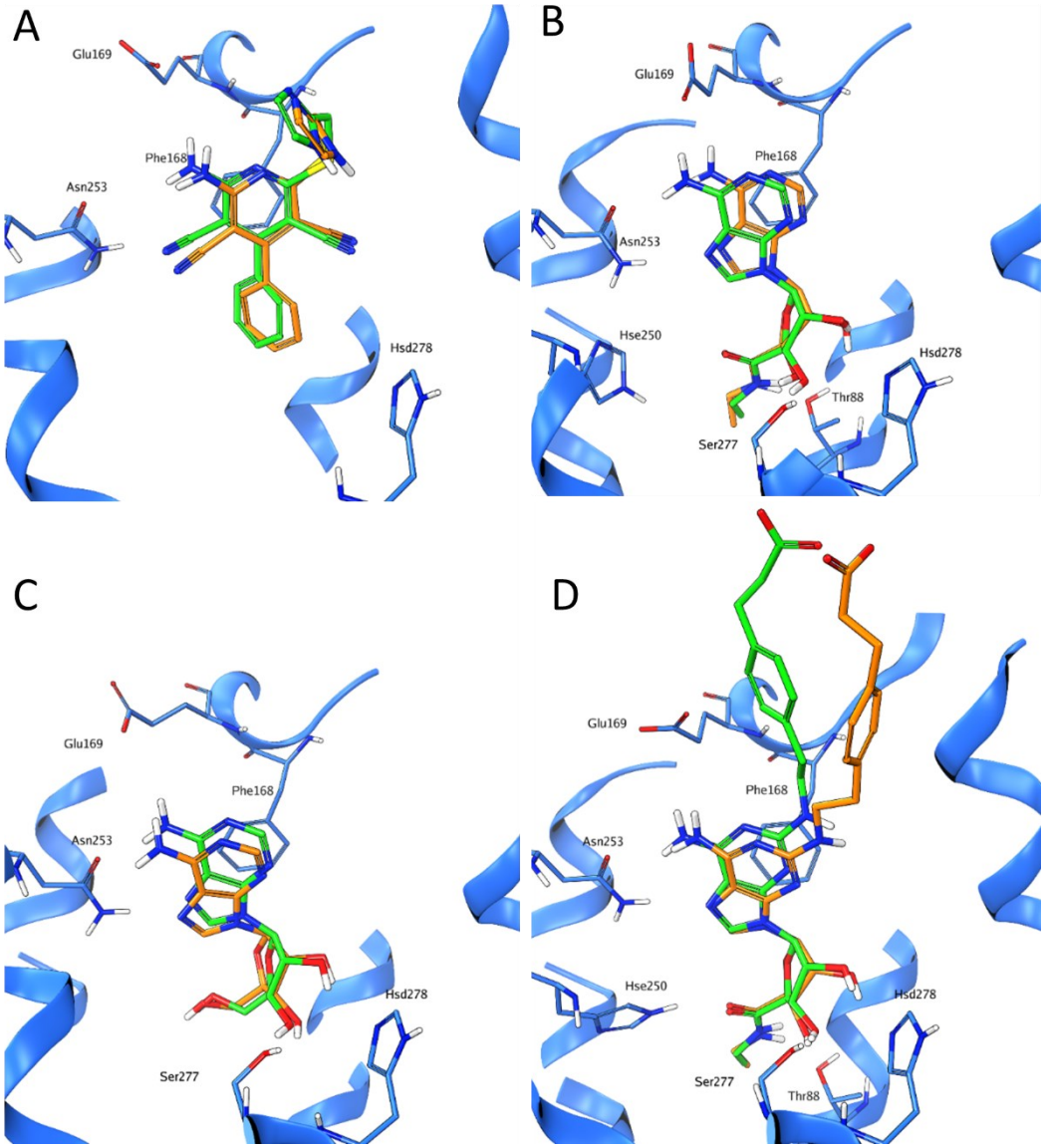


Figure 2. The crystallographic binding mode (green) and the final pose obtained using SuMD (orange) for the four agonists used in this study (A: LUF5833; B: NECA, C: Adenosine; D: CGS21680).

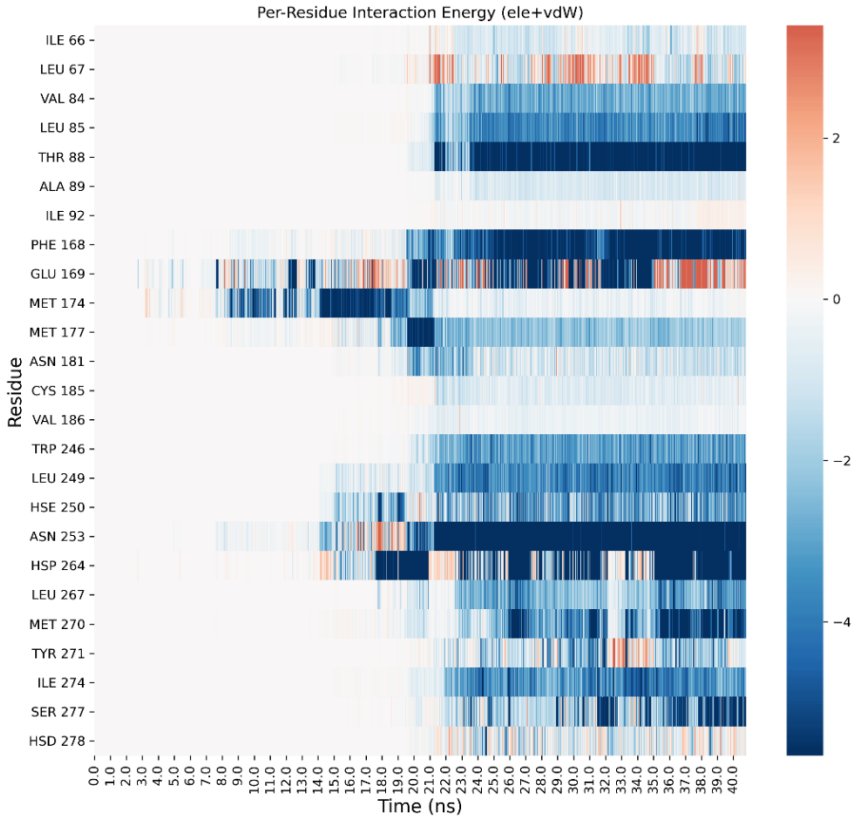


Figure 3. Per-residue energetic analysis of the SuMD trajectory of CGS21680.

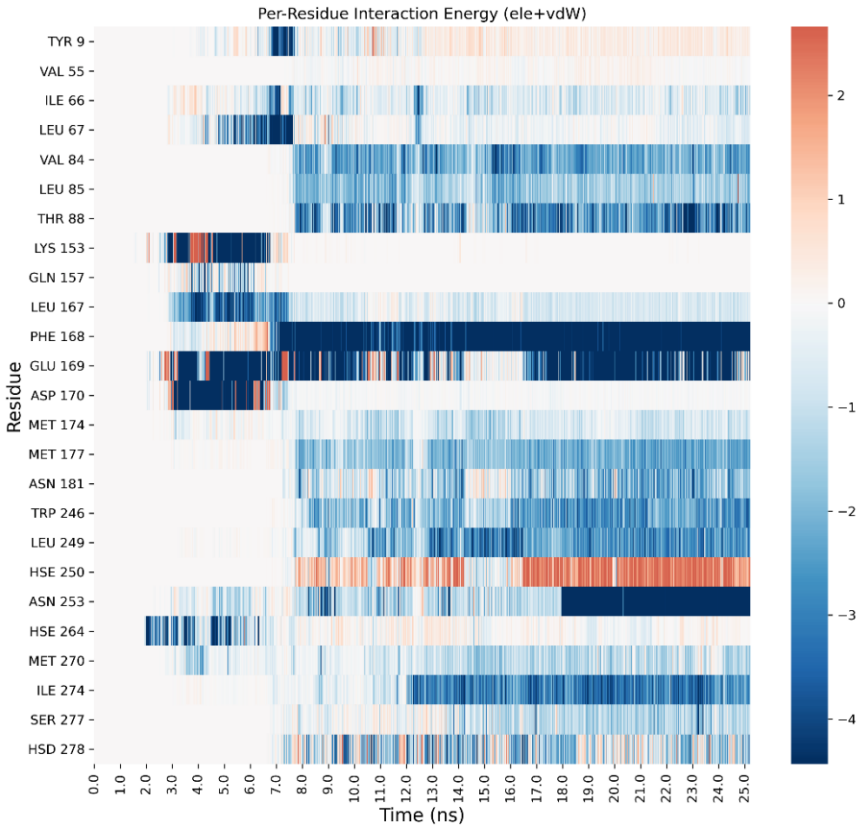


Figure 4. Per-residue energetic analysis of the SuMD trajectory of Adenosine.

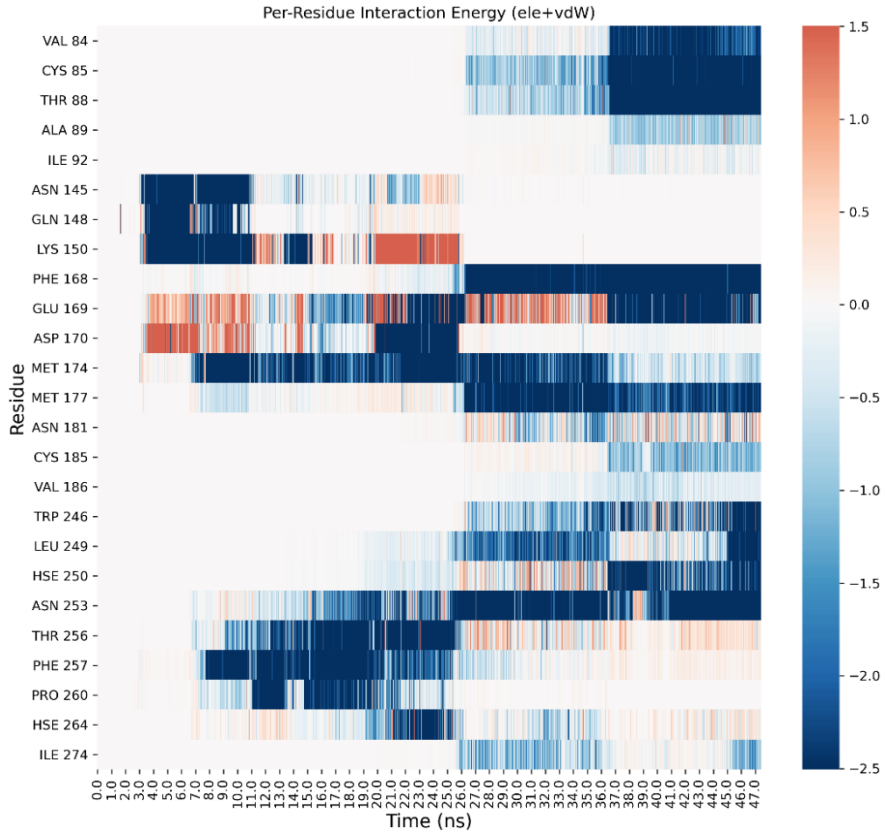


Figure 5. Per-residue energetic analysis of the SuMD trajectory of NECA.

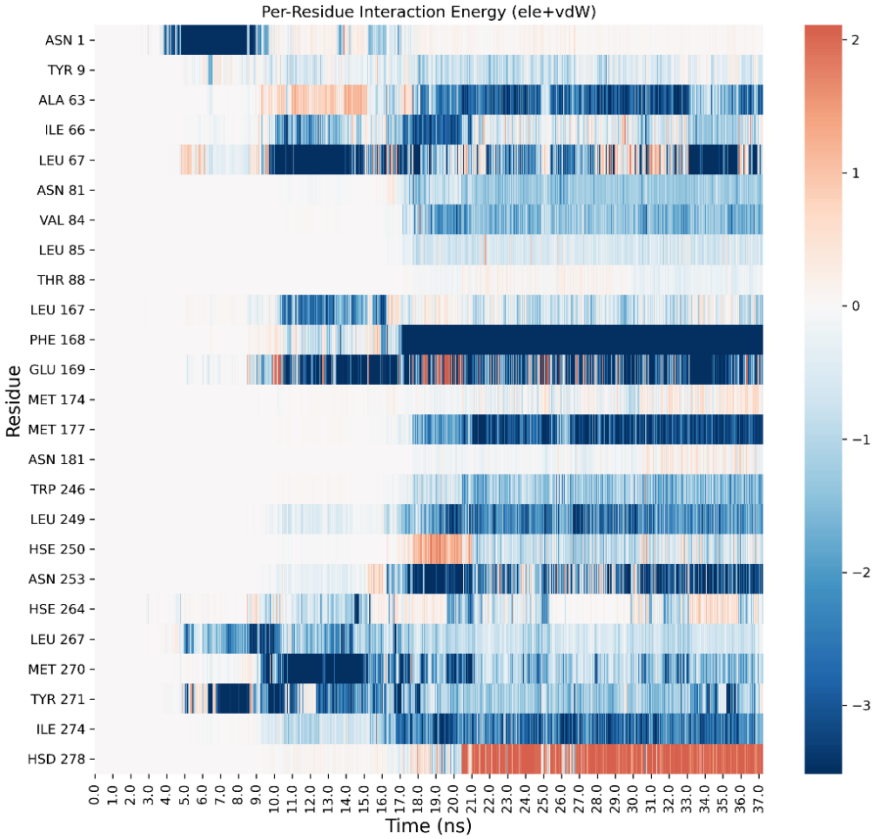


Figure 6. Per-residue energetic analysis of the SuMD trajectory of LUF5833.

Detailed analysis on each trajectory has been performed to understand the recognition process for the four agonists (Figures 3-6). This analysis consists of a per-residue decomposition of the interaction energy between the ligand and the protein, during the binding event.

The binding pathway for the four ligands as well as the most contacted regions of the protein can be visualized in Figure 7 and the videos collected in the Supplementary Information.

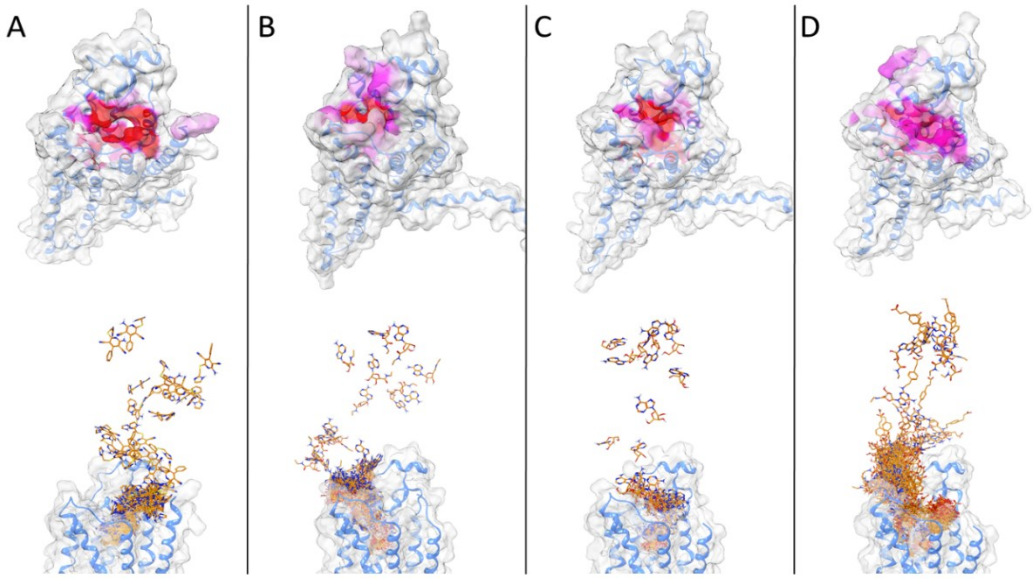


Figure 7. In the upper part of the Figure the protein surface is colored according to the number of contacts with the ligand during the trajectory (scale white to red, from less contacted to more contacted residues). In the lower part the SuMD trajectory is displayed superposing each frame. A: LUF5833; B: NECA; C: Adenosine; D: CGS21680.

The trajectory of LUF5833 has been prolonged for 25 ns at the end of the SuMD simulation. This prolonged trajectory has been analyzed using AquaMMapS (see Materials and Methods) to gain information on the possible role of the solvent in the activation mechanism of AR by non-nucleoside agonists.

For comparison also the first part (before the ligand reaches the orthosteric site) of the SuMD trajectory of LUF5833 has been analyzed using AquaMMapS. This analysis can

provide some additional information on the solvent behavior in the apo form of the receptor. The results of these analyses are reported in Figure 8.

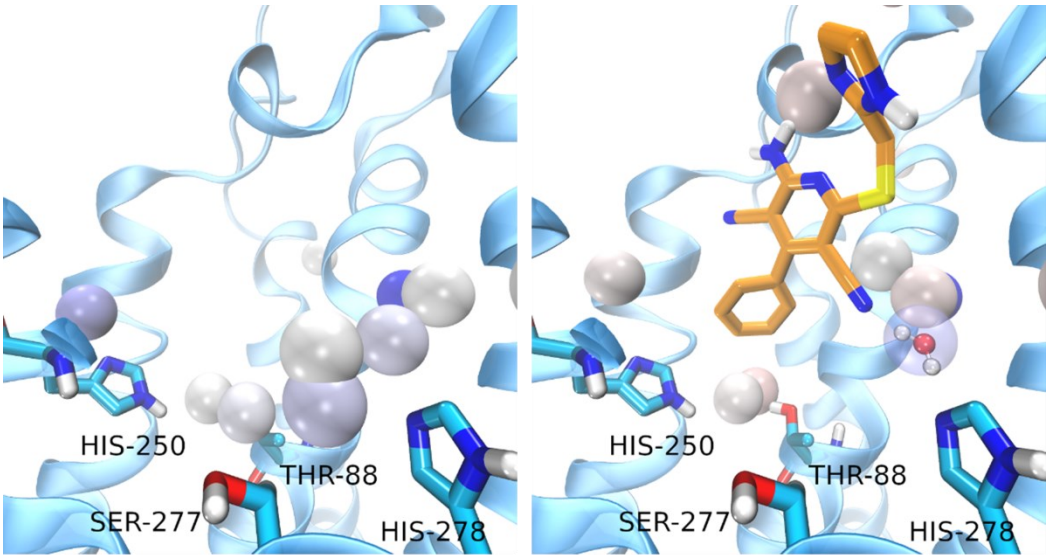


Figure 8. Results of the AquaMMapS analysis for the "APO" trajectory on the left and for the prolonged SuMD trajectory of LUF5833. The cells where the water molecules have an occupancy value higher than 25% are displayed as spheres coloured according to the occupancy value (from white to blue).

As it can be observed, the three classic ribose-containing agonists approach the protein taking contacts in a region that includes residues of the extracellular loop (ECL) 2 and 3 and transmembrane helical segments (TMs) 4, 5, and 6. This meta-binding site has already been described in our previous works[36,37]. LUF5833 instead approaches the receptor from the other side, making contacts in the region between the extracellular portion of TMs 1, 2, and 7. This suggests a different binding pathway for the two classes of AR agonists. Trajectory analysis correctly highlights the pivotal role that is played by Phe-168 (which is involved in a π-stacking with both types of agonists), according to mutagenesis studies which flagged this residue as fundamental for ligand binding[15]. Moreover, Asn-253, which establishes a double hydrogen bond with the adenine moiety of ribose agonists and a single hydrogen bond through one of the two nitrile groups of LUF5833, is also marked as an important residue for the recognition of both classes of agonists, according to mutagenesis data that illustrate how an N253A mutation would be detrimental for the activity of both ribose and non-ribose agonists[15]. Aside from these common interaction

features that regard the adenine or “pseudo-adenine” portion of the molecule, the main difference in the recognition pattern of these two classes of agonists is related to the role of Ser-277 and Thr-88: as highlighted by our trajectory analysis, neither of these two residues establishes a direct interaction with LUF5833, in agreement with mutagenesis data which shows that mutation of these two residues negatively impact ribose agonists but have no effect on the affinity of non-ribose one[15].

Regarding the solvent behavior in the orthosteric site, it is interesting to note that the water molecules in the apo form of the receptor seem to adopt an interactive pattern that mimics the one observed for agonists ligands. Indeed, key residues for the activation of the receptor, like Thr-88, His-250, Ser-277, and His-278 are well solvated and stabilize water molecules through hydrogen bonds.

It is tempting to argue that this observation (the organization of solvent molecules in a way that mimics agonists interactions) can provide a possible explanation for the concept of receptor basal activity, defined as the activation of the receptor in the absence of the ligand.

In detail, it seems that the stable water molecule interacting with Ser-277 is displaced upon LUF5833 binding while the water molecule interacting with His-278 is further stabilized by the cyano group in position 3. This water molecule is displayed in Figure 8. So, while it is true that LUF5833 does not interact directly with any key residues for the receptor activation, at least one of these interactions (the one with His-278) is still present and is mediated by a stable water molecule. Interestingly also the interaction between Adenosine and His-250 is mediated by a bridging water molecule, while NECA and CGS21680 interact with this residue using their amide tail.

Concerning water-bridged interactions, AquaMMapS analysis illustrates how LUF5833 seems to stabilize two water molecules that form a hydrogen bond bridge between His-250 and Thr-88 (Figure 8), playing a similar role to the amide tail of both NECA and CGS. Mutation of both residues has a detrimental role on ribose agonists' affinity, coherently with their direct interaction with the ribose moiety[38,39]. This could indicate that, while not interacting with His-250 and Thr-88, non-ribose agonists such as LUF5833 could stabilize

a water molecule network that mimics the same interaction pattern of ribose agonists. The hydrophobic pocket which houses these stable water molecules is completed by Leu-85: this residue was determined to have a big impact on the affinity of ribose agonists such as CGS but has a smaller effect on the affinity of non-ribose agonists such as LUF5833[15]. This could be explained by the fact that this residue interacts directly with the ligand in the case of CGS, while in the case of LUF5833 its main involvement seems to be in the definition of a "hydrophobic" trap for these two water molecules that mimic the interaction pattern of ribose agonists. Notably, this bound water network extends also towards Asn-253: based on mutagenesis studies which show that, in the case of non-ribose agonists, the reduction of potency is mainly related to efficiency rather than on binding affinity, it is also tempting to speculate that this water network stabilized by LUF5833 is somehow involved in receptor activation, thereby validating the role of LUF5833 as a partial agonist.

Altogether, our SuMD simulations provide an overview of the mechanistic details regarding the recognition process between AR and their agonists, shedding light upon differences in the binding event between nucleoside and non-nucleoside ones. Despite the useful information that can be gathered from our simulations, some AR-specific features cannot be captured by the SuMD technique, thereby impairing a clear and complete depiction of the agonist mechanism. Firstly, our simulations consider the interaction between one single ligand molecule and an individual receptor in a defined lipidic and ionic environment: despite being a sufficiently accurate approximation of reality for the evaluation of geometric properties related to the binding event, these boundary conditions cannot take into account the complex network of interactions of AR within a cellular environment, including the ones with themselves, other GPCRs and a plethora of ancillary factors[40], which leads to surprising pharmacological properties[41,42]. Secondly, a key aspect of AR agonist signaling is portrayed by the ligand residence time, which has been flagged as a more efficient predictor of "in vivo" functional efficacy than binding affinity[43]. Although the evaluation of this aspect of agonist signaling was beyond the scope of this scientific work, it is important to underline that the association process is only the first part of a more complex and intricate story.

References

1. Müller, C. E. & Jacobson, K. A. Recent developments in adenosine receptor ligands and their potential as novel drugs. *Biochimica et Biophysica Acta (BBA) - Biomembranes* **1808**, 1290–1308 (2011).

2. Chen, J. F., Eltzschig, H. K. & Fredholm, B. B. Adenosine receptors as drug targets-what are the challenges? *Nature Reviews Drug Discovery* (2013) doi:10.1038/nrd3955.

3. Navarro, G., Borroto-Escuela, D. O., Fuxe, K. & Franco, R. Purinergic signaling in Parkinson's disease. Relevance for treatment. *Neuropharmacology* **104**, 161–168 (2016).

4. Shah, U. & Hodgson, R. Recent progress in the discovery of adenosine A2A receptor antagonists for the treatment of Parkinson's disease. *Current Opinion in Drug Discovery and Development* Preprint at (2010).

5. Gao, Z.-G. G. & Jacobson, K. A. Purinergic Signaling in Mast Cell Degranulation and Asthma. *Frontiers in Pharmacology* **8**, (2017).

6. Vincenzi, F., Pasquini, S., Borea, P. A. & Varani, K. Targeting Adenosine Receptors: A Potential Pharmacological Avenue for Acute and Chronic Pain. *International Journal of Molecular Sciences* **21**, 8710 (2020).

7. Fishman, P. *et al.* Adenosine Receptors and Cancer. in 399–441 (2009). doi:10.1007/978-3-540-89615-9_14.

8. Mustafa, S. J., Morrison, R. R., Teng, B. & Pelleg, A. Adenosine Receptors and the Heart: Role in Regulation of Coronary Blood Flow and Cardiac Electrophysiology. in 161–188 (2009). doi:10.1007/978-3-540-89615-9_6.

9. Schmidt, J. & Ferk, P. Safety issues of compounds acting on adenosinergic signalling. *Journal of Pharmacy and Pharmacology* **69**, 790–806 (2017).

10. Jacobson, K. A., Tosh, D. K., Jain, S. & Gao, Z.-G. Historical and Current Adenosine Receptor Agonists in Preclinical and Clinical Development. *Frontiers in Cellular Neuroscience* **13**, (2019).

11. Albrecht-Küpper, B. E., Leineweber, K. & Nell, P. G. Partial adenosine A1 receptor agonists for cardiovascular therapies. *Purinergic Signalling* **8**, 91–99 (2012).

12. Soudijn, W., Wijngaarden, I. & Ijzerman, A. Medicinal Chemistry of Adenosine A1 Receptor Ligands. *Current Topics in Medicinal Chemistry* **3**, 355–367 (2003).

13. Cristalli, G., Lambertucci, C., Marucci, G., Volpini, R. & Ben, D. A2A Adenosine Receptor and its Modulators: Overview on a Druggable GPCR and on Structure-Activity Relationship Analysis and Binding Requirements of Agonists and Antagonists. *Current Pharmaceutical Design* **14**, 1525–1552 (2008).

14. Jespers, W. *et al.* Structural Mapping of Adenosine Receptor Mutations: Ligand Binding and Signaling Mechanisms. *Trends in Pharmacological Sciences* **39**, 75–89 (2018).

15. Lane, J. R. *et al.* A Novel Nonribose Agonist, LUF5834, Engages Residues That Are Distinct from Those of Adenosine-Like Ligands to Activate the Adenosine A 2a Receptor. *Molecular Pharmacology* **81**, 475–487 (2012).

16. Amelia, T. *et al.* Crystal Structure and Subsequent Ligand Design of a Nonriboside Partial Agonist Bound to the Adenosine A 2A Receptor. *Journal of Medicinal Chemistry* **64**, 3827–3842 (2021).

17. Molecular Operating Environment (MOE), 2019.01; Chemical Computing Group ULC, 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, 2021. https://www.chemcomp.com/Research-Citing_MOE.htm.

18. Humphrey, W., Dalke, A. & Schulten, K. Sartorius products. *Journal of molecular graphics* **14**, 33–38 (1996).

19. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J Chem Phys* **79**, 926–935 (1983).

20. Wallace, J. A. & Shen, J. K. Predicting pKa values with continuous constant pH molecular dynamics. *Methods in enzymology* **466**, 455–475 (2009).

21. Wallace, J. A. *et al.* Toward accurate prediction of pKa values for internal protein residues: The importance of conformational relaxation and desolvation energy. *Proteins: Structure, Function, and Bioinformatics* **79**, 3364–3373 (2011).

22. Huang, S. K. *et al.* Allosteric modulation of the adenosine A2A receptor by cholesterol. *eLife* **11**, (2022).

23. Harvey, M. J., Giupponi, G. & de Fabritiis, G. ACEMD: Accelerating biomolecular dynamics in the microsecond time scale. *J Chem Theory Comput* **5**, 1632–1639 (2009).

24. Brooks, B. R. *et al.* No Title. **30**, 1545–1614 (2009).

25. CGenFF Home. https://cgenff.umaryland.edu/.

26. Vanommeslaeghe, K. *et al.* CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J Comput Chem* **31**, 671–690 (2010).

27. Kumar, A., Yoluk, O. & MacKerell, A. D. FFParam: Standalone package for CHARMM additive and Drude polarizable force field parametrization of small molecules. *Journal of Computational Chemistry* **41**, 958–970 (2020).

28. Essmann, U. *et al.* A smooth particle mesh Ewald method. *J Chem Phys* **103**, 8577 (1998).

29. Berendsen, H. J. C., van der Spoel, D. & van Drunen, R. GROMACS: A message-passing parallel molecular dynamics implementation. *Computer Physics Communications* **91**, (1995).

30. Sabbadin, D. & Moro, S. Supervised molecular dynamics (SuMD) as a helpful tool to depict GPCR-ligand recognition pathway in a nanosecond time scale. *J Chem Inf Model* **54**, 372–376 (2014).

31. Bakan, A., Meireles, L. M. & Bahar, I. ProDy: Protein dynamics inferred from theory and experiments. *Bioinformatics* **27**, 1575–1577 (2011).

32. Michaud-Agrawal, N., Denning, E. J., Woolf, T. B. & Beckstein, O. MDAnalysis: A toolkit for the analysis of molecular dynamics simulations. *J Comput Chem* **32**, 2319–2327 (2011).

33. Gowers, R. J. *et al.* MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations. *Proceedings of the 15th Python in Science Conference* 98–105 (2016) doi:10.25080/MAJORA-629E541A-00E.

34. Phillips, J. C. *et al.* Scalable molecular dynamics on CPU and GPU architectures with NAMD. *J Chem Phys* **153**, 044130 (2020).

35. Cuzzolin, A., Deganutti, G., Salmaso, V., Sturlese, M. & Moro, S. AquaMMapS: An Alternative Tool to Monitor the Role of Water Molecules During Protein–Ligand Association. *ChemMedChem* **13**, 522–531 (2018).

36. Sabbadin, D., Ciancetta, A., Deganutti, G., Cuzzolin, A. & Moro, S. Exploring the recognition pathway at the human A 2A adenosine receptor of the endogenous agonist adenosine using supervised molecular dynamics simulations. *MedChemComm* **6**, 1081–1085 (2015).

37. Deganutti, G., Welihinda, A. & Moro, S. Comparison of the Human A 2A Adenosine Receptor Recognition by Adenosine and Inosine: New Insight from Supervised Molecular Dynamics Simulations. *ChemMedChem* **12**, 1319–1326 (2017).

38. Jiang, Q. *et al.* Hydrophilic side chains in the third and seventh transmembrane helical domains of human A2A adenosine receptors are required for ligand recognition. *Molecular Pharmacology* **50**, (1996).

39. Kim, J., Wess, J., van Rhee, A. M., Schöneberg, T. & Jacobson, K. A. Site-directed Mutagenesis Identifies Residues Involved in Ligand Recognition in the Human A2a Adenosine Receptor. *Journal of Biological Chemistry* **270**, 13987–13997 (1995).

40. Fredholm, B. B., Ijzerman, A. P., Jacobson, K. A., Klotz, K. N. & Linden, J. International Union of Pharmacology. XXV. Nomenclature and classification of adenosine receptors. *Pharmacological Reviews* **53**, 527–552 (2001).

41. Lopes, L. V., Cunha, R. A. & Ribeiro, J. A. Cross talk between A(1) and A(2A) adenosine receptors in the hippocampus and cortex of young adult and old rats. *Journal of neurophysiology* **82**, 3196–3203 (1999).

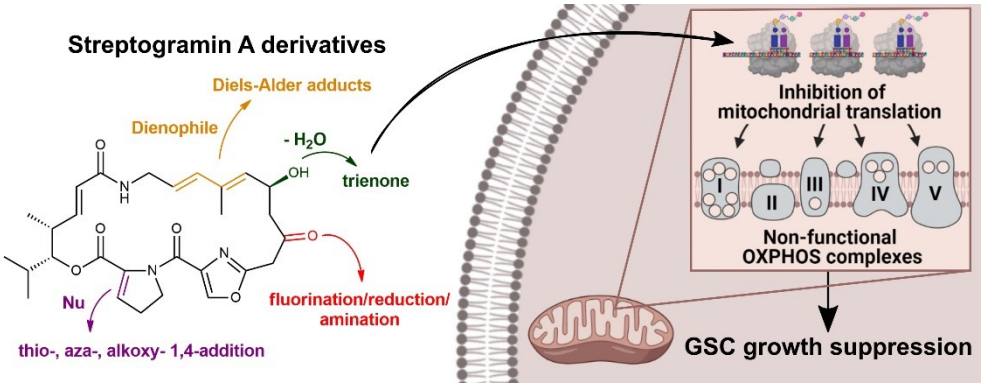
42. Cunha, R. A., Correia-de-Sá, P., Sebastião, A. M. & Ribeiro, J. A. Preferential activation of excitatory adenosine receptors at rat hippocampal and neuromuscular synapses by adenosine formed from released adenine nucleotides. *British journal of pharmacology* **119**, 253–260 (1996).

43. Guo, D., Mulder-Krieger, T., IJzerman, A. P. & Heitman, L. H. Functional efficacy of adenosine A2A receptor agonists is positively correlated to their receptor residence time. *British Journal of Pharmacology* **166**, 1846–1859 (2012).

Streptogramin A derivatives as mitochondrial translation inhibitors to suppress glioblastoma stem cell growth

Denise Sighel, Giulia Battistini, Emanuele Filiberto Rosatti, Jacopo Vigna, **Matteo Pavan**, Romina Belli, Daniele Peroni, Federica Alessandrini, Sara Longhi, Michael Pancher, Joanna Rorbach, Stefano Moro, Alessandro Quattrone, Ines Mancini

Sighel, D. *et al.* Streptogramin A derivatives as mitochondrial translation inhibitors to suppress glioblastoma stem cell growth. *Eur J Med Chem* 114979 (2022) doi:10.1016/J.EJMECH.2022.114979.



Abstract

New therapeutic strategies for glioblastoma treatment, especially tackling the tumour's glioblastoma stem cell (GSC) component, are an urgent medical need. Recently, mitochondrial translation inhibition has been shown to affect GSC growth, clonogenicity, and self-renewal capability, therefore becoming an attractive therapeutic target. The combination of streptogramins B and A antibiotics quinupristin/dalfopristin (Q/D), which inhibits mitochondrial ribosome function, affects GSCs more effectively *in vitro* than the standard of care temozolomide. Here, docking calculations based on the cryo-EM structure of the Q/D-bound mitochondrial ribosome have been used to develop a series of streptogramin A derivatives. We obtained twenty-two new and known molecules starting from the dalfopristin and virginiamycin M1 scaffolds. A structure-activity relationship refinement was performed to evaluate the capability of these compounds to suppress GSC growth and inhibit mitochondrial translation, either alone or in combination with quinupristin. Finally, quantitative ultra HPLC-mass spectrometry allowed us to assess the cell penetration of some of these derivatives. Among all, the fluorine derivatives of dalfopristin and virginiamycin M1, **(16*R*)-1e** and **(16*R*)-2e,** respectively, and flopristin

resulted in being more potent than the corresponding lead compounds and penetrating to a greater extent into the cells. We, therefore, propose these three compounds for further evaluation *in vivo* as antineoplastic agents.

1. Introduction

IDH-wild type glioblastoma (GBM) is the most common primary brain tumour in adults, classified as grade IV by the World Health Organisation[1]. GBM is highly aggressive and inevitably lethal. Indeed, despite the radical standard of care treatment, based on extensive surgical resection, cycles of radiotherapy and temozolomide-based chemotherapy, the median survival rate is around 12-15 months after diagnosis, and the patients' 5-year survival remains very low, around 5.5%[2,3]. Therefore, new therapeutic approaches for GBM treatment are desperately needed.

We recently showed that inhibition of mitochondrial translation suppresses glioblastoma stem cell (GSC) growth *in vitro*[4]. Among the different cell types comprising the tumour, GSCs are responsible for tumour initiation and development, treatment resistance, and hence disease recurrence[5,6] and are a promising therapeutic target. GSCs have been shown to rely on oxidative phosphorylation (OXPHOS) for their metabolic and energetic demands[7,8], along with several types of other cancer stem cells[9–11]. OXPHOS complexes are composed of proteins most of which are translated by the cytosolic ribosomes while a minor part, thirteen proteins, are translated by the mitoribosomes. Thus, the assembly of functional OXPHOS complexes depends both on the cytosolic and mitochondrial translation machinery. Human mitoribosomes are descendants of bacterial ribosomes and, despite the evolutionary changes that occurred, still share with their bacterial counterparts structural and functional similarities, particularly considering the functional ribosomal core[12,13]. Based on these similarities, we hypothesised that antibiotics targeting bacterial ribosomes could bind to mitoribosomes and inhibit mitochondrial translation[4].

Among the different compounds screened, the combinations of streptogramins B and A quinupristin/dalfopristin (Q/D) and quinupristin/virginiamycin M1 (Q/VM1) were selected as the most potent (structures depicted in **Fig. 1**)[4]. Dalfopristin (D, **1**) is a semisynthetic derivative of virginiamycin M1 (VM1, **2**), and it rapidly hydrolyses to VM1 (**2**) at pH 7.4 and

37 °C[14]. Given that Q/D (30:70 w/w, natural ratio preserved in the therapeutic formulation) is an FDA-approved antibiotic for the treatment of skin infections[15], it was selected for further investigation. *Via* Cryo-EM, we showed that Q/D acts by binding to the large subunit of the mitoribosome, with Q (**3**) at the entrance to the exit tunnel for nascent polypeptides and VM1 (**2**), the hydrolysis product of D (**1**), at the peptidyl-transferase centre (PDB ID: 6I9R). Upon binding, Q/D inhibits mitochondrial protein synthesis, affecting functional OXPHOS complexes assembly, finally leading to detrimental effects on GSCs viability[4]. Interestingly, we found that Q/D is almost 15 times more effective than temozolomide (TMZ) in terms of growth inhibition of GSCs, regardless of MGMT promoter methylation status. Indeed, the growth inhibition 50 (GI_{50}) value for Q/D was 6.5 ± 1.1 µM and 20.2 ± 1.4 µM, for COMI and VIPI GSCs, respectively, whereas the GI_{50} value for TMZ was 96.5 ± 15.2 µM and 337.7 ± 35.5 µM, for COMI and VIPI cells, respectively[4].

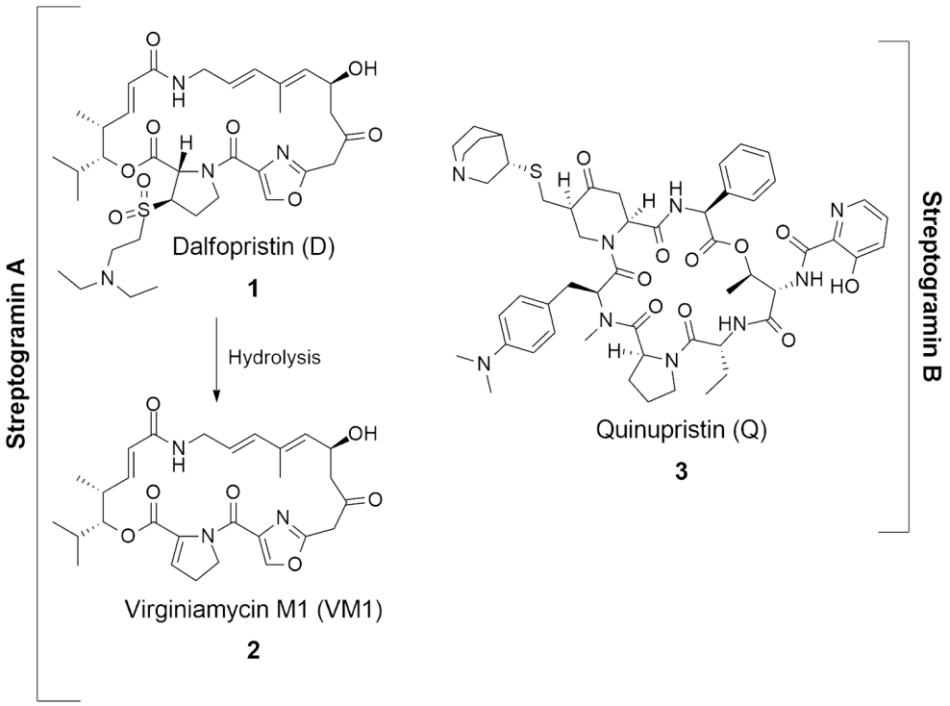


Figure 1. Molecular structures of dalfopristin (**1**), virginiamycin M1 (**2**), and quinupristin (**3**).

Although Q/D is an already FDA-approved bacterial antibiotic traded with the name of Synercid®, we wondered if Q and D specificity and potency could be improved by

increasing their affinity toward the mitoribosome by selectively introducing suitable modifications.

Herein, we report the assessment of the biological activity of Q (**3**) and D (**1**) tested as separate molecules, and the docking-supported design, synthesis, and biological evaluation of D (**1**) and VM1 (**2**) derivatives as anti-IDH-wt glioblastoma agents. All derivatives were obtained by introducing chemical modifications on different positions of the D scaffold by a semisynthetic approach. All compounds were tested for their ability to suppress GSC growth, either alone or in combination with Q (**3**), and the deriving growth inhibition 50 (GI_{50}) values were calculated. In addition, all compounds were evaluated for their ability to inhibit mitochondrial translation. At last, we developed a quantitative ultrahigh-performance liquid chromatography-mass spectroscopy (UHPLC-MS) method to evaluate the capability of some derivatives to enter the cell and determine the extent of such penetration compared to D (**1**) and VM1 (**2**).

2. Results and Discussion

2.1 Choosing between quinupristin and dalfopristin

The cryo-EM structure of Q/D bound to the human mitoribosome showed the specific binding sites of the two molecules and the consequent rearrangements they induce[4]. This structural information sets the basis for the rational design of new chemical derivatives. To determine whether to focus on one or both the molecular scaffolds, D (**1**) and Q (**3**) were evaluated as separate molecules, taking into account their effect on GSC growth and their ability to inhibit mitochondrial translation.

Two GSC lines, namely COMI and VIPI, were treated with a range of drug concentrations for 48 h, and cell viability was determined using Hoechst 33342 and propidium iodide (PI). In detail, PI-positive cells were subtracted from Hoechst 33342 positive cells, in order to calculate the number of live cells. Dose-response curves were constructed, and the deriving GI_{50} values were calculated. Both D (**1**) and Q (**3**) inhibited GSC growth more or less to the same extent as the Q/D combination (**Fig. 2a**).

Next, the effect of the single molecules on mitochondrial translation was determined using two different techniques: ^{35}S metabolic labelling, which allows the detection of nascent proteins, and immunoblotting. Regarding immunoblotting, the expression levels

of NADH dehydrogenase 3 (ND3), which belongs to complex I of the electron transport chain, and of cytochrome c oxidase subunit 1 (COX1), cytochrome c oxidase subunit 2 (COX2) and cytochrome c oxidase subunit 4 (COX4), three proteins belonging to complex IV of the electron transport chain, were taken into account as representative of mitochondrial and cytosolic translation activity. Indeed, ND3, COX1, and COX2 are mitochondrially encoded and translated by the mitochondrial translation system, whereas COX4 is nuclearly encoded and translated by the cytosolic translation system.

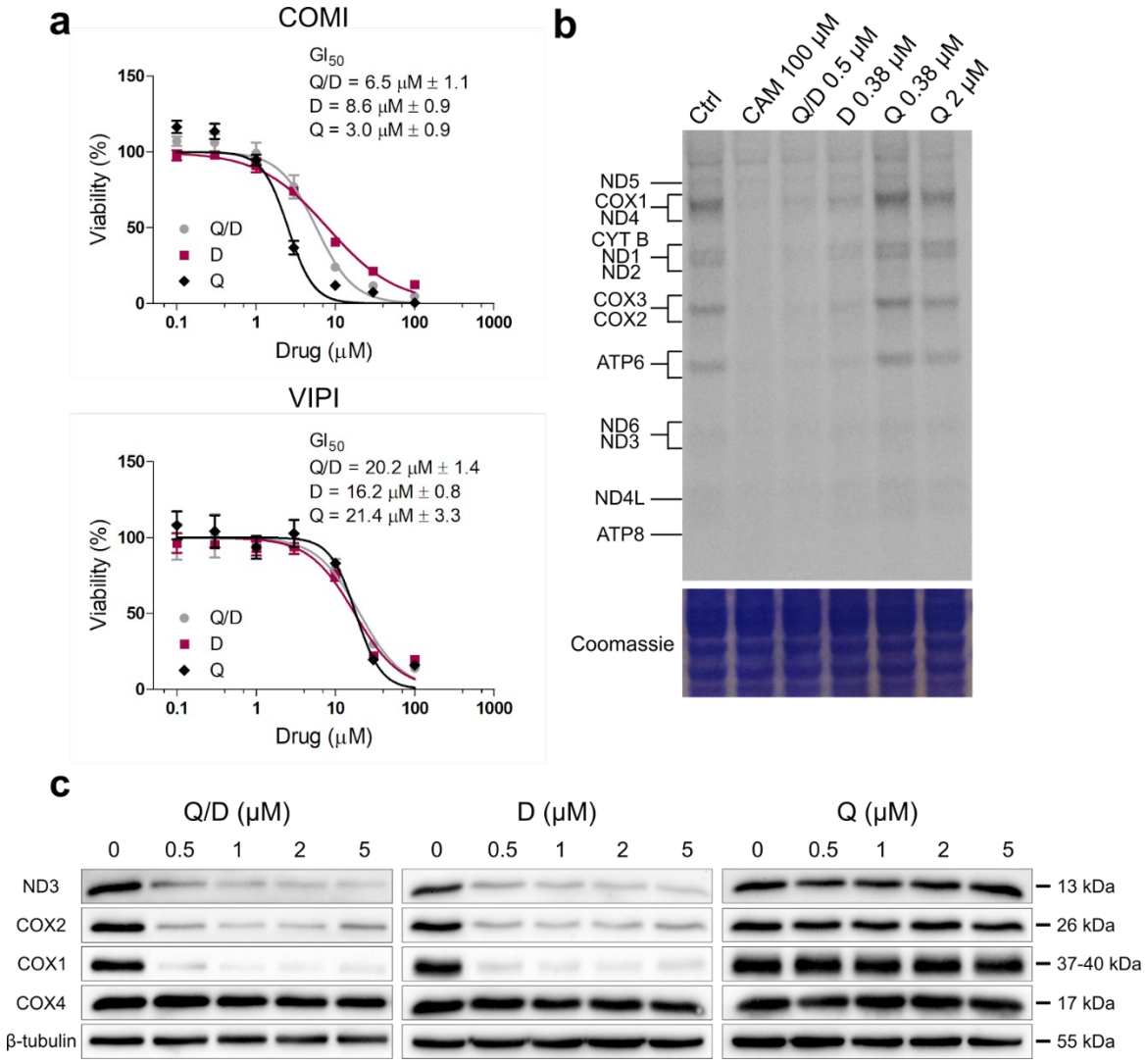


Figure 2. Comparison of quinupristin (Q, **3**) and dalfopristin (D, **1**).

a. Representative dose-response curves for Q/D, Q (**3**), and D (**1**) on COMI and VIPI cells. The GI_{50} values were calculated from n=3 biological replicates, n=4 technical replicates each, mean ± SD. **b.** Effects of Q/D, Q, and D 24 h treatment on mitochondrial translation on COMI cells as assayed by ^{35}S metabolic labelling assay. Chloramphenicol (CAM) was used as a positive control. One representative result is shown, n=3 biological replicates. **c.** Immunoblotting on COMI cells after 48 h treatment with Q/D, D, or Q. Effects on

ND3, COX1, COX2, and COX4 proteins are shown. Beta tubulin was used as a loading control. One representative result, n=3 biological replicates.

If the mitochondrial translation is inhibited, a decrease or absence of newly synthesised mitochondrial proteins and a selective decrease in ND3, COX1, and COX2 expression levels, but not in those of COX4, should be observed.

While D (**1**) inhibited mitochondrial translation as efficiently as Q/D, Q (**3**) did not (**Fig. 2b-c**). The incapability of Q (**3**) alone to inhibit mitochondrial translation could be related to the observation in bacteria that streptogramins B alone have a lower affinity for the ribosome compared to the combination of streptogramins A and B[16,17]. In fact, when streptogramins A bind to the bacterial ribosome, they induce a conformational change of the A2062 residue (corresponding to A2725 in the mitoribosome), which favours the subsequent binding of streptogramins B[14]. On the contrary, streptogramins A do not require streptogramins B to bind to the bacterial ribosome, even though the presence of streptogramins B potentiates their activity in bacteria. Based on these results, we decided to focus on D (**1**) to design and synthesise derivatives with increased affinity for the mitoribosome.

2.2 Chemistry

Dalfopristin (D, **1**) belongs to group A streptogramins, which are 23-membered macrocyclic polyketide/non-ribosomal peptide hybrids. It was developed by Rhône-Poulenc as a water-soluble salt derivative starting from virginiamycin M1 (VM1, **2**) upon the introduction of an *N*,*N*-diethylaminoethylsulfonyl chain on C26.

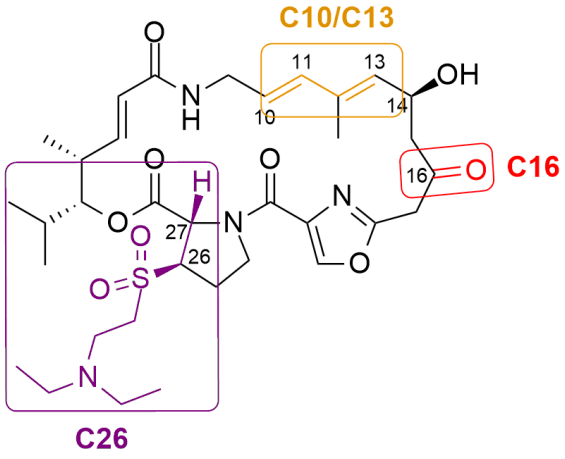


Figure 3. Sites of structural modifications on dalfopristin (D, **1**) scaffold.

We identified three accessible positions on the D (**1**) scaffold to introduce structural modifications by semisynthesis: C10/C13, C16, and C26, as depicted in **Fig. 3**. Some of the derivatives produced are already known; others instead are new. D (**1**) was obtained by chromatographic isolation from the commercial Synercid®; VM1, of which a total synthesis consisting of 6-8 linear steps from simple chemical building blocks has been recently reported[18,19], was obtained by D (**1**) hydrolysis (**Fig. 1**) in suitable amount for the production of derivatives.

2.2.1 C10/C13 modifications

A visual inspection of the VM1 (**2**) binding mode within the mitoribosome streptogramin A binding site (PDB ID: 6I9R) showed a relatively unoccupied region near the C10/C13 diene. This observation led us to investigate the possibility of obtaining new Diels-Alder products in this position.

The reaction of D (**1**) or VM1 (**2**) with maleic anhydride or maleimide as dienophiles produced the Diels-Alder adducts **1a**, **1b**, and **2a**, **2b**, respectively, in moderate to good yields under $AlCl_3$ catalysis. Our choice for the dienophiles was led by the possibility of introducing groups with different properties for interacting with the mitoribosome space. Docking calculations were performed to evaluate the putative binding mode of these four Diels-Alder derivatives within the mitoribosome streptogramin A binding site. As well known, Diels-Alder cycloaddition generates up to four stereocentres in a single concerted reaction step, increasing significantly the molecular complexity. Each of the four possible diastereomers generated (structures depicted in **Fig. S1**) was evaluated independently. The results of the docking calculations for these compounds and the comparison between the predicted binding mode for these derivatives and the reference VM1 (**2**) are reported in **Fig. S2** and **Video V1**. As can be seen by the per-residue decomposition of electrostatic and hydrophobic interactions, the diastereomer presenting all *syn* substituents on the stereocentres in the cyclohexane ring (diastereomer 4, **Fig. S1**) has the most favourable interaction profile out of all four possible combinations. In detail, the added moiety stretches below the plane defined by the macrocycle ring, while the defining features of VM1 (**2**) binding mode, such as interaction with residues G2724, A2725, A2938, C2939, G2992, U2993, and U3072, are generally conserved or strengthened (**Fig. 4**). On the

contrary, the other diastereomers likely have a worse predicted interaction with the mitoribosome streptogramin A binding site. This can be because the orientation of the added moiety causes steric hindrance or conformational changes to the macrocycle ring preventing the conservation of the key VM1 interaction features, as highlighted by the per-residue interaction energy analysis (**Fig. S2**).

Since interactions with the target are predicted to involve mainly the carbonyl groups, which are shared between the maleimide and the maleic anhydride derivatives, there is no notable difference in the predicted interaction pattern between them.

Aside from the predicted interaction pattern, the cyclization with the maleic anhydride or maleimide dienophiles leads to the reduction of the conformational freedom of the macrocycle ring increasing, in principle, the fraction of time that the compounds spend in the conformation competent for the binding to the mitoribosome.

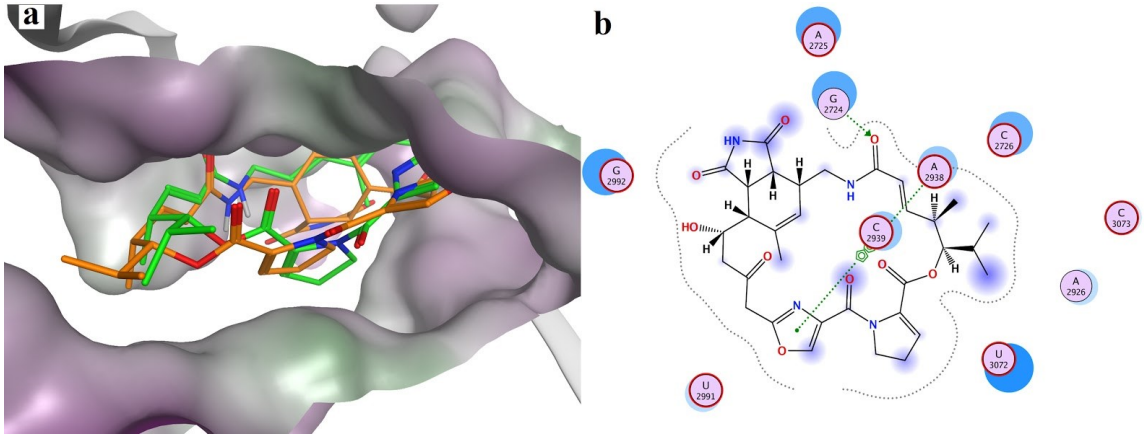
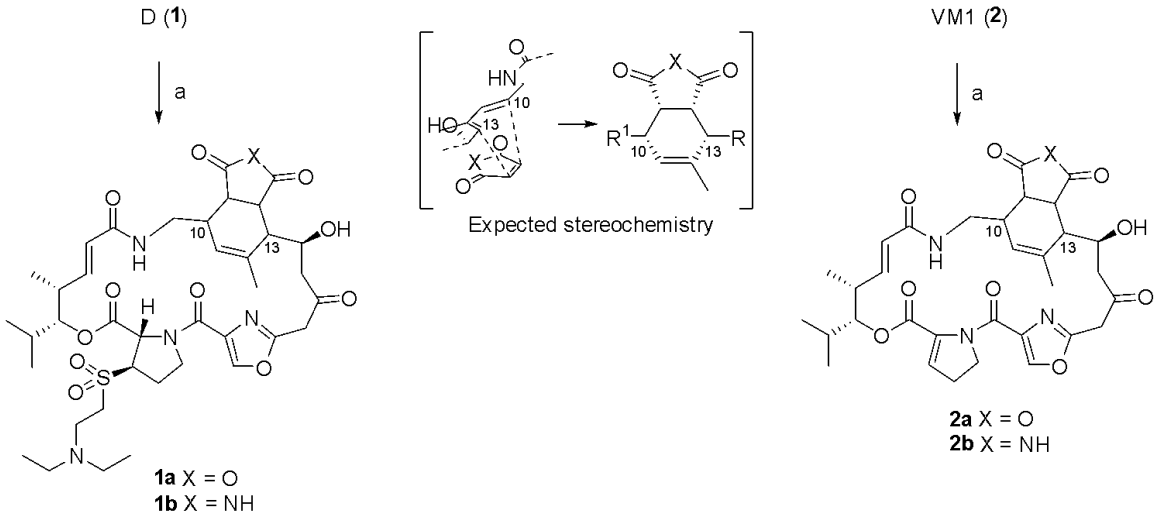


Figure 4. a. Superposition between the Cryo-EM structure of VM1 (**2**) within the mitoribosome streptogramin A binding site (PDB ID: 6I9R, green) and the docking pose of compound **2b-4** (VM1 maleimide adduct, orange). The surface of the binding site is coloured according to residue lipophilicity, ranging from green (lipophilic) to violet (hydrophilic). **b.** 2D interaction diagram of the same pose reported in panel A for compound **2b-4**.

When we performed the Diels-Alder reaction, only one of the possible stereoisomers was produced in a substantial yield. Because of NMR spectra complexity[20,21], which is also increased by the presence of conformational equilibria, and the difficult crystallisation of these macrocycles, the stereochemistry of the major isomers produced was not determined. However, according to several examples in the literature[22,23], the major

product obtained may be the one resulting from the less hindered *endo* approach of the dienophile to the diene in its *cis* conformations as shown in **Scheme 1** (diastereomer 1, **Fig. S1**). This is further supported by the use of $AlCl_3$ as a catalyst, which is known to enhance the stereoselectivity of Diels-Alder reactions through stabilisation of the *endo* transition state[24].



Scheme 1. Diels-Alder derivatives obtained by reaction of C10/C13 diene of **1** or **2** with either maleic anhydride or maleimide as dienophiles. Reagents and conditions: (a) maleic anhydride or maleimide, dichloromethane, N_2, rt, 1 h; $AlCl_3$, -35 °C, 3 h; reflux, 16-36 h.

2.2.2 C16 modifications

Due to the chemical versatility of the carbonyl group present in C16, this position has been already extensively investigated in the past[25–27]. **Fig. 5** shows the comparison between the VM1 (**2**) binding mode and the docking-predicted binding mode of a possible C16 derivative (specifically, the 16*R*-hydroxyl derivative). As can be seen, the presence of a small, slightly hydrophilic and unoccupied subpocket near the C16 carbonyl group of **2** suggests that the introduction of small substituents in the C16 position could theoretically improve the interaction with the mitoribosome.

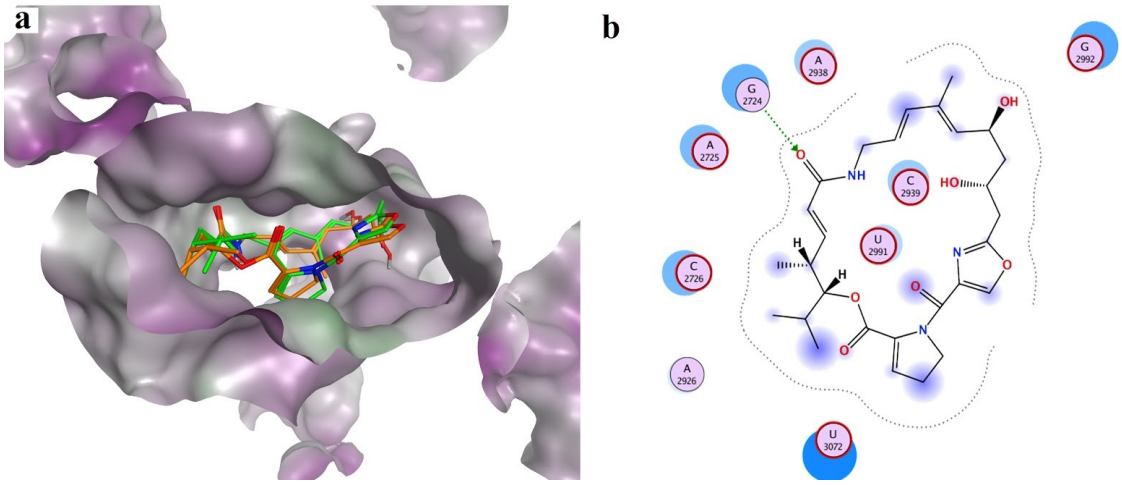
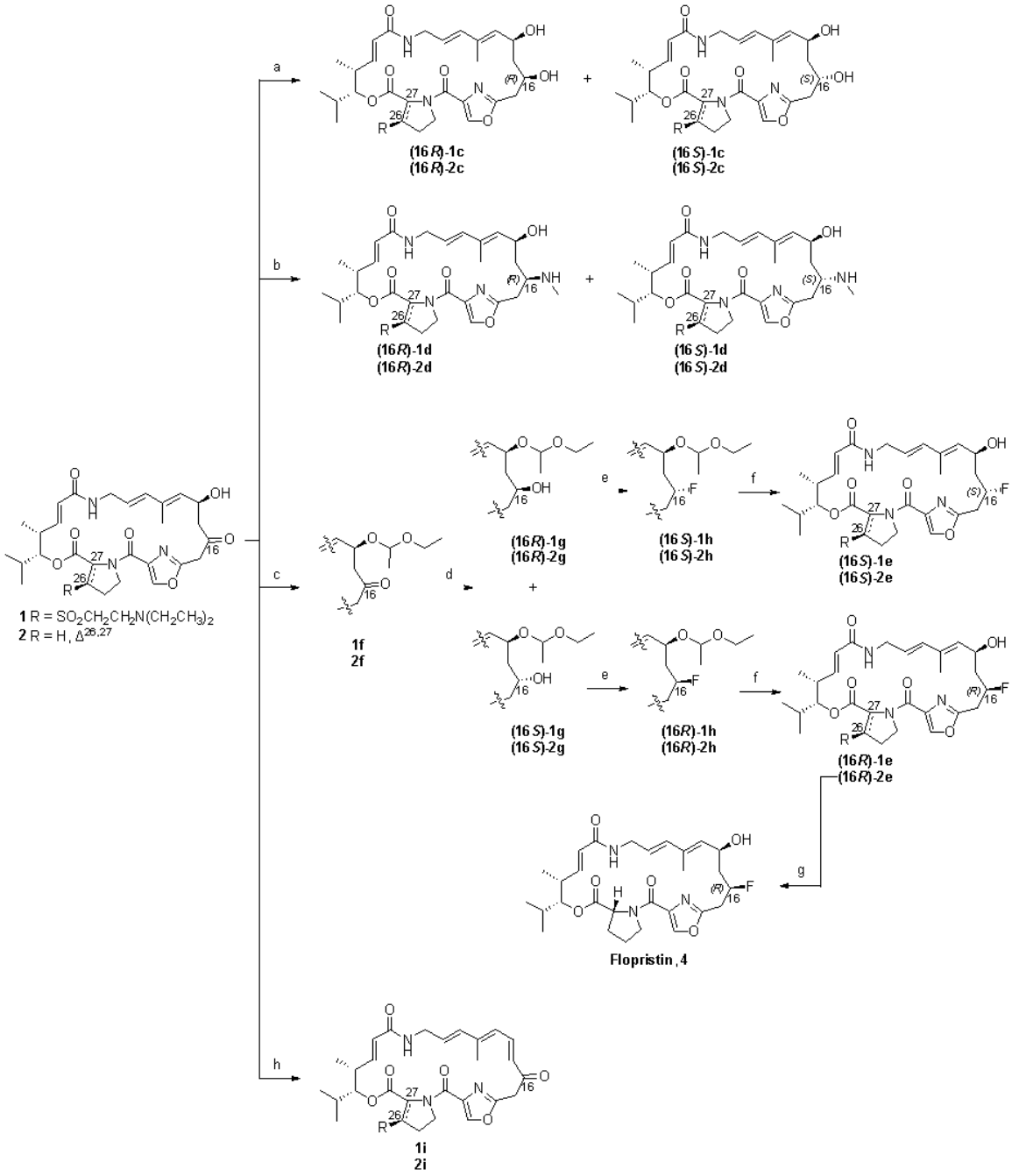


Figure 5. **a.** Superposition between the Cryo-EM structure of VM1 (**2**) within the mitoribosome streptogramin A binding site (PDB ID: 6I9R, green) and the docking pose of compound **(16*R*)-2c** (VM1 hydroxyl derivative, orange). The surface of the binding site is coloured according to residue lipophilicity, ranging from green (lipophilic) to violet (hydrophilic). **b.** 2D interaction diagram of the same pose reported in panel A for compound **(16*R*)-2c**.

Several substituents were considered for the C16 positions, and their putative binding mode within the mitoribosome streptogramin A binding site was evaluated through docking calculations. These results and the comparison between the predicted binding mode for these derivatives and the reference VM1 (**2**) are reported in **Fig. S3** and **Video V2**. Altogether, all examined substitutions are predicted not to alter the binding mode of VM1 (**2**), maintaining the binding features present in the 6I9R structure. In detail, the hydroxyl and *N*-methyl derivatives are predicted to improve the affinity for the receptor, respectively establishing hydrogen bonds or charge-π interactions with residues G2992, U2991 and C2939. In contrast, the fluorine derivative is predicted to ameliorate the affinity for the receptor through improved hydrophobic interaction with C2939 and U2991.

Based on the computational analysis, the C16 derivatives reported in **Scheme 2** were synthesised.



Scheme 2. Synthetic pathways to introduce C16 modifications.

Series 1 R = $SO_2CH_2CH_2N(CH_2CH_3)_2$; series 2 R = H, $\Delta^{26,27}$. Reagents and conditions: (**a**) $NaBH_4$, EtOH, 0 °C, 1 h, C16 epimers separated by column chromatography on silica gel; (**b**) CH_3NH_2, CHCOOH (0.1 eq.), acetonitrile, N_2, 0 °C to rt, 48 h; $NaCNBH_3$, 0 °C, 3 h, C16 epimers separated by column chromatography on silica gel; (**c**) ethyl vinyl ether, pyridinium *p*-toluenesulfonate, dichloromethane, rt, 16 h; (**d**) $NaBH_4$, ethanol, 0 °C, 5 h, C16 epimers separated by column chromatography on silica gel; (**e**) diethylaminosulfur trifluoride, dichloromethane, N_2, 0 °C to rt, 24 h; (**f**) pyridinium *p*-toluenesulfonate, methanol, rt, 24 h; (**g**) $NaBH_4$, ethanol, 0 °C, 1 h; (**h**) Propyl isocyanate, 4-dimethylaminopyridine, pyridine, dichloromethane, 16 h.

The selective reduction of the C16 keto group led to the formation of an additional stereocentre, resulting in a diastereoisomeric mixture of 16*R* and 16*S* diols, which were successfully separated through column chromatography[28]. Compounds **(16*R*)-1c**, **(16*S*)-1c** and **(16*R*)-2c**, **(16*S*)-2c** were obtained starting from **1** and **2**, respectively.

The C16 reductive amination by reacting **1** or **2** with methylamine followed by $NaBH_3CN$ addition provided the corresponding secondary amines (**(16*R*)-1d**, **(16*S*)-1d**, and **(16*R*)-2d**, **(16*S*)-2d**) (**Scheme 2**) as a mixture of the two possible epimers, which were separated through column chromatography[29].

Fluorinated compounds are frequently considered in medicinal chemistry to increase metabolic stability, modulate physicochemical properties and favour the binding affinity for the biological target, leading to the development of a wide series of highly effective drugs[30]. During the optimisation of the antibacterial activity of **2** in the 1990s, fluorine introduction was the most interesting and promising modification performed on the C16 position[26,27]. Therefore, we have considered the production of 16*R* and 16*S* fluorine derivatives starting from either **1** or **2**. Compounds **(16*R*)-1e**, **(16*S*)-1e** and **(16*R*)-2e**, **(16*S*)-2e** were obtained as depicted in **Scheme 2**. The hydroxyl group in C14 of **1** or **2** was protected via acetal formation using ethyl vinyl ether (EVE) in solvent-free conditions and pyridinium *p*-toluenesulfonate (PPTS) as a weakly acidic catalyst. The use of EVE led to the introduction of a further stereocentre on the inserted protecting group. The two epimeric acetals, obtained in comparable amounts, were not separated for the subsequent steps since the stereocentre introduced is removed in the deprotection step. Subsequently, the C16 carbonyl group was reduced, obtaining a diastereomeric mixture of 16*R* and 16*S* compounds, which were successfully separated through column chromatography on silica gel. Fluorination of the C16 hydroxyl group occurred rapidly using diethylaminosulfur trifluoride (DAST), as previously reported[26,27,31,32]. The DAST-mediated S_N2 reaction occurred with the inversion of the configuration of the incoming substituent and proceeded till the complete conversion of the starting material. Removal of the protecting group was obtained in high yields under mildly acidic conditions.

After FDA approval of Q/D, Rhône-Poulenc (then Aventis Pharma, now Sanofi) actively worked on the synthesis of second-generation orally available streptogramin analogues. This campaign led to the selection of the combination of the streptogramin A flopristin (**4)**

and the streptogramin B linopristin, also known as NXL103. NXL103 is now in phase II clinical trials for acute bacterial skin and skin structure infections and community-acquired pneumonia[33]. Since flopristin (**4**) was not available on the market, we produced it starting from **(16*R*)-2e** by C26-C27 double bond reduction using $NaBH_4$ in ethanol. As expected, the reaction produced a highly complex mixture of products, but we managed to isolate flopristin (**4**) in pure form (**Scheme 2**).

As previously mentioned, it has been reported that heating or treating **2** with strong acids leads to dehydration of the C14 hydroxyl group to give a highly stable trienone[25]. Effectively, we obtained the formation of **1i** and **2i**, the trienone derivatives of **1** and **2**, respectively, even in the presence of mild acidic conditions (**Scheme 2**).

2.2.3 C26 modifications

A visual comparison of the docking-predicted binding mode for D (**1**) and VM1 (**2**) in the 6I9R complex showed that **1** retains the key features of the binding mode of **2**, and that is capable of forming two additional hydrogen bonds with U3072 through the *N*,*N*-diethylaminoethylsulfonyl chain on C26 (**Fig. 6**). As previously reported (**Fig. 1**), the *N*,*N*-diethylaminoethylsulfonyl chain on C26 is known to rapidly hydrolyse at pH = 7.4 and 37 °C[14], leading to the formation of **2**. Given the increased number of interactions of **1**, introducing a non-hydrolysable chain in C26 would potentially increase the affinity for the mitoribosome and hence the compound's activity in suppressing GSC growth.

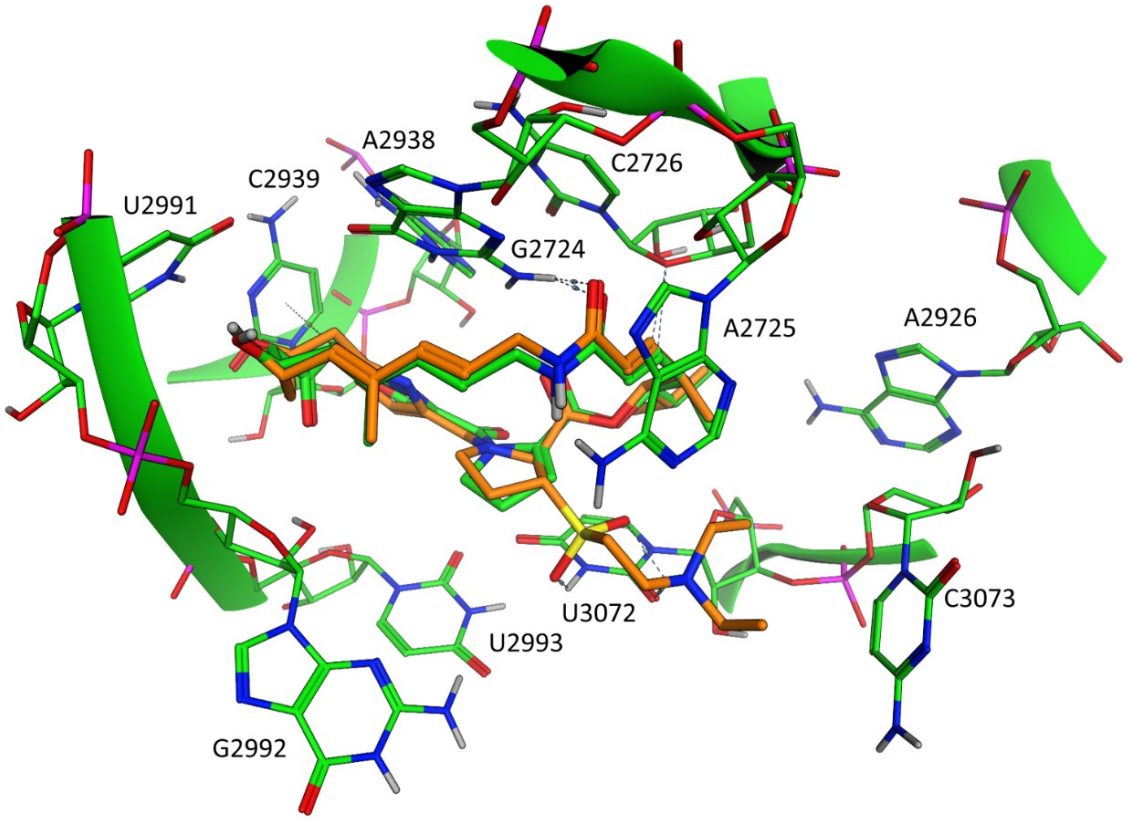
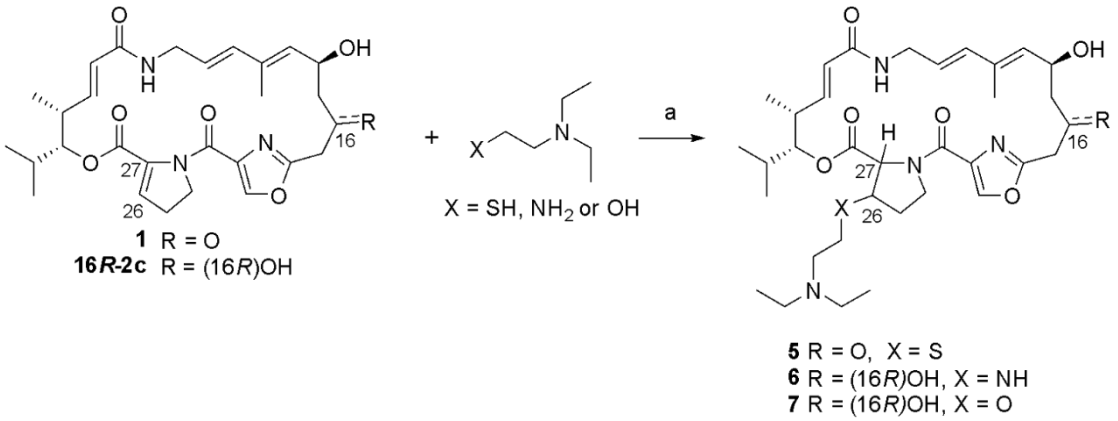


Figure 6. Superposition between the Cryo-EM structure of VM1 (**2**) within the mitochondrial ribosome streptogramin A binding site (PDB ID: 6I9R, green) and the docking pose of D (1, orange). Ribosome residues within 4.5 Å from the VM1 (**2**) solved structure are reported and depicted as sticks in the figure. Hydrogen bonds between the ligand and the receptor are shown as light blue dashed lines.

Structural modifications at the C26 position have already been investigated and exploited to introduce the *N*,*N*-diethylaminoethylsulfonyl chain which led to dalfopristin (D, **1**) development. The introduction of this chemical moiety in C26 conferring greater water-solubility was achieved by the 1,4-addition of a thiol-based nucleophile, which was successively oxidised to the sulfone derivative and converted into a mesylate salt[25,34]. Following the conjugate addition of a nucleophile in C26, four different diastereomers are expected to be produced. Kinetic factors, and in particular the temperature parameter, have been reported to play a fundamental role in stereochemical control[25,34]. The addition of 2-diethylaminoethanethiol to the α,β-unsaturated ester of **2** gave the (26*R*,27*S*) derivative **5** in good yield, in agreement with literature data[25,34,35] (**Scheme 3**). By oxidising **5** with sodium tungstate in combination with hydrogen peroxide, **1** was obtained, further

confirming **5** stereochemistry. Indeed, the ^{1}H and ^{13}C NMR signals and the retention time by HPLC analysis of **5** after oxidation were completely superimposable with data of **1**.



Scheme 3. 1,4-addition of a nucleophile in C26. Reagents and conditions: (a) *S*-, *N*- or *O*-nucleophile, methanol, 10-72 h; *O*-nucleophile generated *in situ*: NaH, tetrahydrofuran, 0 °C, 1 h.

Aiming at introducing non-hydrolysable chains, other types of nucleophiles, such as amines and alkoxides, were considered. Since the C16 carbonyl group of compound **2** could also undergo a nucleophilic attack, we used the *R* hydroxyl derivative **(16*R*)-2c** as starting material.

The conjugate addition of the *N*-nucleophile on **(16*R*)-2c** proceeded in a few days at room temperature with the almost full conversion of the starting material, giving **6** in good yield as a single isomer (**Scheme 3**). Even the reaction with the *O*-based nucleophile, generated *in situ* by deprotonation of the corresponding alcohol, produced the more abundant diastereomer **7**, which was isolated through HPLC purification (**Scheme 3**). Working at a lower temperature (-78 °C) significantly increased the stereoselectivity of the conjugate addition. Because of NMR complexity and the difficult crystallization of these compounds, the stereochemistry of the C26 derivatives produced was not determined, except for compound **5**.

Docking calculations were performed to evaluate the putative binding mode of C26 derivatives, considering each of the four possible diastereomers formed by each reaction independently. Docking results and the comparison of predicted binding modes for these

compounds with respect to the reference VM1 (**2**) are summarized in **Fig. S4** and **Video V3**.

Regarding the thiol derivative **5**, the favorite diastereomer retains the same (26*R*,27*S*) configuration as D (**1**). Also for compounds **6** and **7**, the favorite diastereomer seems to be the one presenting the added moiety oriented relatively to the macrocycle ring as the *N*,*N*-diethylaminoethylsulfonyl of D (**1**). Of note, in this case, the preferred configuration is (26*R*,27*R*), based on the lower priority of N or O atoms compared to S. Docking analysis suggests that these compounds could theoretically retain all the characterizing interaction features of VM1 (**2**) while improving the affinity for the mitoribosome through additional/strengthened interaction with U3072 through the C26 chain.

2.3 Evaluation of streptogramin A derivatives as mitochondrial translation inhibitors to suppress glioblastoma stem cells growth

The effects of the synthesised derivatives on cell growth were evaluated by treating COMI cells with a range of drug concentrations for 48 hours, and by determining cell viability using Hoechst 33342 and propidium iodide. We constructed dose-response curves and calculated the deriving GI_{50} values. **Table 1** summarises the GI_{50} values obtained for the different derivatives tested alone or in combination with Q (**3**) in a 70:30 w/w ratio since streptogramins are known to act synergistically in bacteria.

	alone	**+ Q (3)**		**alone**	**+ Q (3)**
D (1)	8.65 ± 0.9	6.50 ± 1.10	**VM1 (2)**	17.06 ± 1.10	7.25 ± 1.39
1a	> 100	14.75 ± 2.00	**2a**	> 100	19.59 ± 2.51
1b	> 100	14.46 ± 1.76	**2b**	> 100	16.39 ± 1.06
(16*R*)-1c	18.79 ± 1.52	5.17 ± 0.78	**(16*R*)-2c**	21.65 ± 3.02	5.28 ± 1.34
(16*S*)-1c	37.53 ± 1.54	7.05 ± 1.02	**(16*S*)-2c**	52.45 ± 2.55	8.74 ± 0.51
(16*R*)-1d	14.25 ± 1.17	4.85 ± 0.39	**(16*R*)-2d**	19.45 ± 4.95	6.57 ± 0.6
(16*S*)-1d	1.56 ± 0.15	1.32 ± 0.15	**(16*S*)-2d**	1.16 ± 0.55	1.5 ± 0.81
(16*R*)-1e	6.73 ± 1.24	5.71 ± 1.2	**(16*R*)-2e**	4.82 ± 1.23	5.42 ± 1.67
(16*S*)-1e	20.53 ± 2.31	8.37 ± 1.08	**(16*S*)-2e**	36.14 ± 0.84	12.26 ± 2.56

flopristin (4)	7.37 ± 1.44	5.83 ± 1.04			
1i	10.91 ± 2.39	9.9 ± 2.27	**2i**	20.52 ± 1.93	15.62 ± 0.71
5	7.02 ± 0.75	8.00 ± 1.46			
6	> 100	16.88 ± 1.93			
7	85.27 ± 5.01	13.99 ± 1.52			

Table 1. GI_{50} values in µM for dalfopristin and virginiamycin M1 derivatives[1]: Diels-Alder (a, b), C16-hydroxyl (c), C16-NHMethyl (d), C16-fluoro (e), trienone (i), and C26 (**5**-**7**) products.

[1]GI_{50} values were calculated upon treating COMI cells with different concentrations of the derivatives for 48 h, alone or in combination with Q (**3**), in a 70:30 w/w ratio. The GI_{50} values were calculated from n=3 biological replicates, n=3 technical replicates each, mean ± SD.

In parallel, the capability of the derivatives to inhibit mitochondrial translation was also evaluated by assessing the expression of ND3, COX1, COX2, and COX4 proteins *via* immunoblotting. As previously mentioned, ND3, COX1, and COX2 are translated by the mitochondrial translation system, whereas COX4 is translated by the cytosolic translation system. Therefore ND3, COX1, and COX 2 were taken into account as representatives of the mitochondrial translation activity, while COX4 was taken into account as a representative of the cytosolic translation activity.

Concerning the C10/C13 derivatives, none of the four derivatives **1a**, **1b**, **2a**, and **2b** suppressed the growth of the GSC line tested, as evident by the GI_{50} values resulting above the maximum concentration tested (100 µM). Accordingly, none of the Diels-Alder derivatives was able to inhibit mitochondrial translation (**Fig. S5**). This result is not surprising. Indeed, based on docking calculations, the stereoisomer produced (diastereomer 1) showed a worse binding profile compared to the more favourable diastereomer 4 or D (**1**) or VM1 (**2**). Of note, the growth inhibition effect observed for the Diels-Alder derivatives in combination with Q (**3**) is only due to the Q contribution. Indeed, dose-response curves of these derivatives in combination with Q plotted as a function of Q contribution are superimposable with that of Q used as a single agent.

As previously observed for the antibacterial properties, by modifying the C16 position it is possible to modulate activity. The stereochemistry at this position has also been reported to play an important role in the activity[26]. Indeed, compounds with (16*R*) configuration

have been reported to be more potent than the *S* epimers[26]. In our hands, the *R* epimers displayed lower GI_{50} values than the *S* epimers when considering the diols (**(16*R*)-1c**, **(16*S*)-1c** and **(16*R*)-2c**, **(16*S*)-2c**) and the fluorine (**(16*R*)-1e**, **(16*S*)-1e** and **(16*R*)-2e**, **(16*S*)-2e**) derivatives. In detail, the *R* diols (**(16*R*)-1c** and **(16*R*)-2c**) were slightly less potent than the corresponding lead compound, while the *S* diols (**(16*S*)-1c** and **(16*S*)-2c**) displayed much higher GI_{50} values. The *R* fluorine derivatives (**(16*R*)-1e**, **(16*R*)-2e** and **4**) were, instead, slightly more active than the corresponding lead compounds, while the *S* epimers (**(16*S*)-1e** and **(16*S*)-2e**) were less potent. Accordingly, all the tested compounds were able to inhibit mitochondrial translation, with the *R* epimers presenting the highest activity (**Fig. 7**, **Fig. S6**, **Fig. S7**). All these compounds presented GI_{50} values comparable to those of the lead compounds when tested in combination with Q (**3**).

The bioisosteric substitution of the C16 keto group with a C-F bond increases lipophilicity which is expected to favour cell penetration, explaining the improved GI_{50} values compared to the more hydrophilic diols (**(16*R*)-1c**, **(16S)-1c** and **(16*R*)-2c**, **(16S)-2c**) and the lead compounds (**1** and **2**). Unfortunately, the higher lipophilicity has been reported to result in a reduced aqueous solubility and increased metabolization rate *in vivo*[27], which is in turn compensated by the overcome of the chemical instability of streptogramins A at acidic/basic pH due mainly to the sensitive β-hydroxy ketone system[27]. Our *in vitro* data show that compounds **(16*R*)-1e**, **(16*R*)-2e**, and **4**, the last of them developed as an orally available second-generation streptogramin A, are more potent than the corresponding lead compounds, and are, therefore, good candidates for further studies aimed at evaluating their pharmacokinetics properties and their *in vivo* activity in xenografted mouse models.

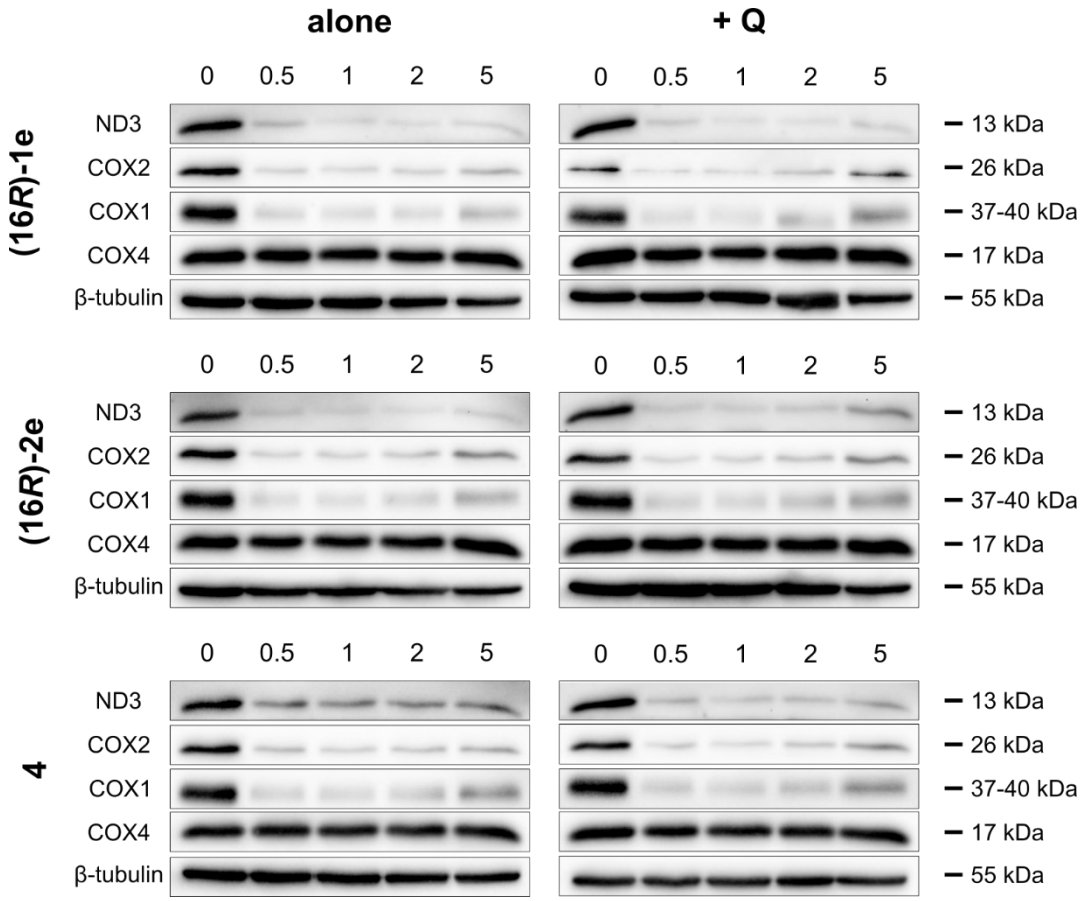


Figure 7. (16*R*) fluoro derivatives inhibit mitochondrial translation. Immunoblotting on COMI cells after 48 h treatment with different concentrations (µM) of **(16*R*)-1e**, **(16*R*)-2e**, and **4**, alone (*left column*) or in combination with Q (**3**) (*right column*). Effects on ND3, COX1, COX2, and COX4 proteins are shown. Beta tubulin was used as a loading control. One representative result, n=3 biological replicates.

Among all the C16 derivatives, compounds **(16*S*)-1d** and **(16*S*)-2d** had the lowest GI_{50} values. Based on the *in vitro* results, the compounds having (16*S*) configuration were more active in comparison to the *R* epimers (**(16*R*)-1d** and **(16*R*)-2d**). This observation contrasts with what has been previously observed for the bacterial activity or for our 16*R* and 16*S* diols and fluorine derivatives.

Surprisingly, when evaluating the capability of these compounds to inhibit mitochondrial translation, we observed no effect. Indeed, in COMI cells treated with increasing concentrations of **(16*S*)-1d**, **(16*S*)-2d**, **(16*R*)-1d** and **(16*R*)-2d**, the expression levels of the mitochondrially encoded proteins ND3, COX1, and COX2 remained unaltered, as well as those of the nuclearly encoded protein COX4 (**Fig. 8** and **Fig. S8**). Based on these results,

we speculated that the capability of the C16-NHMe derivatives to suppress GSC growth derives from an unknown mechanism of action different from mitochondrial translation inhibition, and therefore we did not further investigate these compounds.

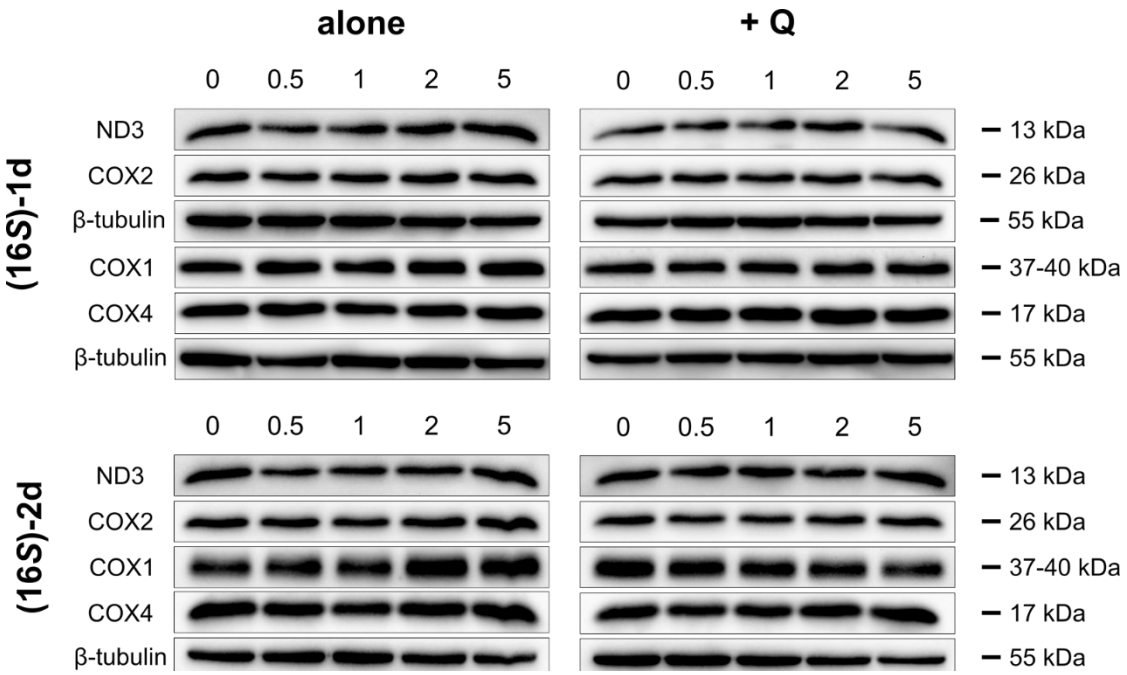


Figure 8. (16*S*) NHMe derivatives do not inhibit mitochondrial translation. Immunoblotting on COMI cells after 48 h treatment with different concentrations (μM) of **(16*S*)-1d,** and **(16*S*)-2d**, alone (*left column*) or in combination with Q (**3**) (*right column*). Effects on ND3, COX1, COX2, and COX4 proteins are shown. Beta tubulin was used as a loading control. One representative result, n=3 biological replicates.

Compounds **1i** and **2i** inhibited GSC proliferation slightly less efficiently than the respective lead compounds. Of note, the loss of the C14 hydroxyl group to produce the trienone unit has been reported to cause a major reduction of the antibacterial activity[25], which instead was not observed for GSC growth inhibition.

Compound **5** inhibited GSC growth to the same extent as **1** (**Table 1**). Unexpectedly, the introduction of an *N*- or *O*-non-hydrolysable moiety in the C26 position resulted in a major reduction of GSC growth inhibition capability (**Table 1**). Indeed, **6** displayed a GI_{50} value above the maximum concentration tested (100 μM), while **7** displayed a GI_{50} value of 85.27 ± 5.01 μM. Accordingly, **6** and **7** inhibit mitochondrial translation only at very high concentrations (> 25 μM) (**Fig. S9**). The immunoblotting results suggest that **6** and **7** can

permeate the cellular membranes, enter the cells and bind to the mitoribosome at least to a certain extent. It has to be clarified whether the low potency observed is due to low membrane permeability or low affinity for the target.

2.4 Evaluation of streptogramin A derivative capability to permeate the cellular membrane and enter the cell via UHPLC-MS

To evaluate whether the lack of the effect observed for the Diels-Alder derivatives (**1a**, **1b**, **2a**, and **2b**) and especially for the derivatives presenting *N*- or *O*-non-hydrolysable chain in C26 (**6**, **7**) was due to their low membrane permeability, we developed a sensitive ultra-high-performance liquid chromatography coupled with mass spectrometry (UHPLC-MS) method for the determination of the amount of these compounds inside the cells. Compounds **1** and **2** were used as a reference, while **1b**, **2b**, and **6** were used as representatives for Diels-Alder and *N*- or *O*-non-hydrolysable chain C26 derivatives, respectively. Moreover, we assessed compounds **(16*R*)-1c** and **(16*R*)-2c** since they present a (16*R*)-hydroxyl group as compound **6**. Finally, we also evaluated the fluorinated derivatives **(16*R*)-1e** and **(16*R*)-2e**, which are slightly more potent than **1** and **2** in GSC growth inhibition. After incubating COMI cells with **1** or **2** derivatives at 100 µM for 3 h, the metabolites were extracted with methanol and the chromatographic separation was carried out using a C18 column, with a linear gradient of acetonitrile-water-formic acid as the mobile phase.

For each compound, a calibration curve was constructed by spiking pure compound in non-treated-cell metabolite extracts resuspended in acetonitrile-water-formic acid (95:5:0.1%, v/v/v) and performing serial dilutions to obtain five different concentrations, ranging from 0.005 ng/µL to 50 ng/µL. The calibration curves were linear over the five orders of magnitude tested (**Fig. S10-12, Table S1**). The lower limit of quantification of this method was 0.005 ng/µL for all compounds. To be noted that for **1** and its derivatives, which are readily hydrolysed in water, each point of the calibration curve was loaded on the instrument just after preparing the spike solution and the contribution of both the hydrolysed and non-hydrolysed peaks was taken into account. By doing so, more than 95% of the compounds were detected in the non-hydrolysed form.

Subsequently, the metabolite extracts derived from cells treated with the compounds were loaded on the UHPLC-MS and the concentration was inferred using the respective calibration curve. In **1**-derivative-treated samples, more than 99% of the compound resulted in the hydrolysed form, coherent with a treatment of 3 hours in an aqueous environment. Still, the hydrolysed and non-hydrolysed peaks were considered and their concentration was derived by comparison to their respective calibration curve. The final concentration of these samples was determined by summing the concentrations of both forms.

Table 2 shows the concentrations of the evaluated compounds.

	C (ng/µL)		C(ng/µL)
D (1)	12.5 ± 0.9	**VM1 (2)**	11.5 ± 1.6
1b	9.6 ± 1.7	**2b**	13.3 ± 1.2
(16*R*)-1c	2.3 ± 1.3	**(16*R*)-2c**	4.6 ± 0.9
(16*R*)-1e	25.7 ± 4.9	**(16*R*)-2e**	38.4 ± 4.4
5	10.8 ± 2.8		
6	6.2 ± 1.6		

Table 2. Concentrations inside the cells calculated for dalfopristin and virginiamycin M1 and their derivatives[1]: Diels-Alder (b), C16-hydroxyl (c), C16-fluoro (e), and C26 (**5**-**6**) products.

[1]COMI cells were treated with 100 µM of a given compound for 3 h. Successively, the metabolite extracts were loaded on the HPLC-MS/MS. The concentration values (expressed in ng/µL) were inferred using the compound calibration curve. The concentration values were calculated from n=5 replicates, mean ± SD.

In detail, compounds **1** and **2** showed similar membrane permeability, with final calculated concentrations of around 12.5 ± 0.9 ng/µL and 11.5 ± 1.6 ng/µL, respectively.

Compounds **1b** and **2b** presented similar permeability compared to their precursors **1** and **2**, respectively, suggesting that these derivatives can enter the cells to the same extent as their precursors. Instead, the lack of effect in mitochondrial translation inhibition could be due to an excessive steric hindrance around the C10/C13 diene moiety that could prevent the molecule from entering and/or interacting with the mitoribosome.

The diol derivatives **(16*R*)-1c** and **(16*R*)-2c** displayed a significantly reduced membrane permeability compared to **1** and **2**, with concentration values of 2.3 ± 1.3 ng/µL and 4.6 ± 0.9 ng/µL, respectively, in line with what has been previously observed by Bacque et al.[27]. In contrast, the fluorine derivatives **(16*R*)-1e** and **(16*R*)-2e** showed a greatly increased permeability compared to their precursors, with concentrations around two to three times those of **1** and **2**. Interestingly, **(16*R*)-2e** had a higher permeability than **(16*R*)-1e**.

The reduced or increased permeability of the diol or fluorinated compounds can be explained, at least in part, by the hydrophilic/lipophilic properties that the inserted group bears, which contribute to modulating the passage through the biological membranes. Indeed, the more hydrophilic diol derivatives present reduced permeability that could explain, at least in part, the decreased capability of these compounds to inhibit GSC growth. On the contrary, the increased permeability of the fluorine derivatives could justify their increased biological activity due to a higher amount of the compound reaching the mitoribosome and being able to interact.

Compound **5**, which presents an *N*,*N*-diethylaminoethylthioether chain in the C26 position, showed a similar membrane permeability to **1** and **2**, which are its *N*,*N*-diethylaminoethylsulfonyl analogue and its precursor, respectively. Indeed, compound **5** presents similar GSC growth inhibition capability to **1** and **2**.

Finally, compound **6**, which bears a (16*R*)-hydroxyl group, had reduced membrane permeability compared to **1** and **2**, in line with what was observed for **(16*R*)-2c**, which is its precursor. As previously described, compounds **6** and **7** displayed a strong reduction in biological activity compared to **1**. There are two factors to be considered to explain this reduction: the entrance into the mitoribosome and the stereochemistry of the C26 moiety. Indeed, based on the immunoblotting results, **6** and **7** can enter the cells and bind to the ribosome to a certain extent. Moreover, the docking calculations show that the addition of a tertiary amine-bearing non-hydrolysable moiety in C26 increases the

interactions of the derivatives inside the mitoribosome binding site. However, the presence of this moiety could hamper the entrance of the compound inside the mitoribosome itself, a parameter that is not taken into account by the docking calculations. Finally, the stereochemistry of the C26 for compounds **6** and **7** has not been clarified yet. It could also be possible that the moiety added in C26 clashes against the mitoribosome while entering or even inside the binding site.

3. Conclusions

In conclusion, a series of streptogramin A derivatives possibly endowed with a higher affinity for the mitochondrial ribosome was designed with the support of docking calculations and was produced by semisynthesis. Three accessible positions (C10/C13, C16, and C26) were identified to introduce structural modifications, and twenty-two derivatives were synthesised, including novel and already reported compounds.

For all the synthesised derivatives, the capability of suppressing GSC growth and of inhibiting mitochondrial translation was evaluated, either alone or in combination with quinupristin (**3**). Moreover, a quantitative UHPLC-MS method was developed to assess the cell penetration of some of these derivatives. The Diels-Alder derivatives in C10/C13 position and the derivatives presenting an *N*- or *O*-non-hydrolysable moiety in C26, although being able to enter the cells as well as the reference compounds, did not suppress GSC growth or inhibit mitochondrial translation at low concentrations. Unfortunately, this was probably because the diastereomers produced did not preserve the interaction pattern inside the mitoribosome of the reference compounds (**1** and **2**), and the introduced moiety clashed against the mitoribosome cavity. On the other hand, most of the C16 derivatives did preserve the binding mode of the reference compounds inside the mitoribosome and their biological activity. In general, the *R* epimers were more potent than the *S* epimers. Among all the derivatives tested, the *N*-methyl derivatives **(16*S*)-1d** and **(16*S*)-2d** were the most potent in suppressing GSC growth, but surprisingly, they could not inhibit mitochondrial translation and thus were not further investigated. The fluorine derivatives **(16*R*)-1e**, **(16*R*)-2e**, and flopristin (**4**) displayed GI_{50} values lower than the corresponding lead compounds and penetrated to a greater extent into the cells.

Further *in vivo* experiments using xenograft models of glioblastoma are needed to evaluate the activity of the fluorine derivatives, alone or in combination with quinupristin, in comparison to Q/D, to assess which molecule or molecule combination is worthier to be further developed as an antineoplastic drug.

4. Experimental Section

4.1. Chemistry

All chemicals and reagents were purchased from Sigma Aldrich, Alfa Aesar and Apollo Scientific. Thin-layer chromatography (TLC) was performed using Merck silica gel F_{254} or reversed-phase Merck RP-18 F_{254}, using short-wave UV light as the visualising agent and $KMnO_4$ or $Ce(SO_4)_2$ as developing agents upon heating. Preparative thin-layer chromatography (PLC) was performed using 20x20 cm Merck Kieselgel 60 F_{254} 0.5/2-mm plates. Column chromatography was performed using Merck Si 45-60 µm as the stationary phase. NMR spectra were recorded on a Bruker-Avance 400 spectrometer using a 5-mm BBI probe ^{1}H at 400 MHz and ^{13}C at 100 MHz and calibrated using residual non-deuterated solvent for $CDCl_3$ (relative to δ_H 7.27 and δ_C 77.0 ppm, respectively) or CD_3OD (δ_H 3.31 and δ_C 49.0 ppm) with chemical shift values in ppm and *J* values in Hz. The following abbreviations were used to describe multiplicities: s = singlet, d = doublet, t = triplet, q = quartet, m = multiplet, br = broad. NMR data were analysed using BrukerTopspin software 3.6.1 version. Electrospray ionisation ESI(+)MS mass spectra were recorded using a Bruker Esquire-LC spectrometer by direct infusion of a methanol solution (source temperature 300 °C, drying gas N_2, 4 L/min, scan range *m/z* 100:1000). High-resolution ESI(+)MS spectra were obtained by direct infusion of a methanol solution using an Orbitrap Fusion Tribrid® mass spectrometer. The exact mass of compounds has been derived from an average of 30 spectra sets. Infrared spectra were recorded by using an FT-IR Tensor 27 Bruker spectrometer equipped with an Attenuated Transmitter Reflection (ATR) device at 1 cm^{-1} resolution in the absorption region Δν 4000–1000 cm^{-1}. Spectral processing was made using the Opus software package. Polarimetric data were obtained with a Jasco DIP-181 apparatus, using Na source. The purity of the compounds was determined by High-Performance Liquid Chromatography (HPLC) using an Agilent 1200 HPLC system equipped

with an autosampler, a binary pump, a diode array detector (Agilent Technologies Waldbronn, Germany) and Phenomenex® Gemini 5u C18 110A column, in gradient conditions with eluent water:acetonitrile (acetonitrile t_0 30%, $t_{8\ min}$ min 80%, $t_{22\ min}$ 80%) flow 1 mL/min (HPLC method A) or Luna 5u C18 100A column, in isocratic conditions with eluent water/acetonitrile 70:30 + 0.1% TFA, flow 1 mL/min (HPLC method B), unless otherwise specified. All tested compounds are >95% pure by HPLC analysis. ^{1}H-NMR, ^{13}C-NMR and High-resolution ESI(+) MS spectra and high-performance liquid chromatograms of tested compounds are reported in the Appendix A Supplementary Data section (**Fig. S13-S77)**.

4.1.01. Isolation of Dalfopristin (1)

The compound was isolated from the commercial Synercid® (500 mg of Q/D combination in a ratio of 30:70 w/w) by liquid chromatography on silica gel or using 0.5 mm PLC using DCM/EtOH 80:20 as the eluting solution (D: R_f = 0.86; Q: R_f = 0.12).

*4.1.01.1. (1^2Z,3^2S,3^3R,6R,7R,8E,13E,15E,17S)-3^3-((2-(diethylamino)ethyl)sulfonyl)-17-hydroxy-6-isopropyl-7,15-dimethyl-5-oxa-11-aza-1(4,2)-oxazola-3(1,2)-pyrrolidinacycloicosaphane-8,13,15-triene-2,4,10,19-tetraone (**1**)*

White powder; 347 mg (99%); R_f = 0.44 (DCM/EtOH = 90:10); ^{1}H-NMR (400 MHz, $CDCl_3$) δ: 8.06 (s, 1H), 6.74 (brt, 1H), 6.58 (dd, *J* = 16.5, 4.5 Hz, 1H), 6.13 (d, *J* = 15.9 Hz, 1H), 5.79 (d, *J* = 16.5 Hz, 1H), 5.66 (ddd, *J* = 15.9, 7.7, 4.1 Hz, 1H), 5.40 (d, *J* = 7.7 Hz, 1H), 5.33 (s, 1H), 4.90 (dq, *J* = 9.7, 5.6 Hz, 1H), 4.83 (d, *J* = 10.3 Hz, 1H), 4.26 (m, 1H), 4.19 (m, 1H), 3.85 (m, 1H), 3.79 (d, *J* = 8.8 Hz, 2H), 3.72 (m, 1H), 3.56 (ddd, *J* = 14.9, 9.8, 5.6 Hz, 1H), 3.24 (m, 2H), 3.03 (m, 3H), 2.87 (dd, *J* = 16.5, 5.2 Hz, 1H), 2.76 (m, 1H), 2.56 (m, 5H), 2.14 (m, 1H), 1.93 (m, 1H), 1.71 (s, 3H), 1.01 (t, *J* = 7.4 Hz, 9H), 0.94 (d, *J* = 6.7 Hz, 6H); ^{13}C-NMR (100 MHz, $CDCl_3$) δ: 201.9, 168.6, 165.7, 160.3, 157.2, 145.1, 144.3, 136.8, 136.2, 134.4, 132.9, 125.1, 124.1, 82.8, 64.9, 62.6, 59.7, 49.2, 49.1, 47.0, 46.8 (2C), 45.4, 43.3, 40.7, 29.4, 25.6, 19.7, 18.6, 12.8, 11.5 (2C), 11.0; HPLC: 15.095 min (acetonitrile/water = 30:70 + 1% TFA); ESI(+)MS: *m/z* 691 $[M+H]^+$, 713 $[M+Na]^+$, 673 $[M+H-H_2O]^+$.

4.1.02. Hydrolysis of Dalfopristin to Virginiamycin M1 (2)

Compound **1** (230 mg, 0.33 mmol) was dissolved in THF (12 mL) and phosphate-buffered saline (60 mL). The resulting solution was sonicated for 5 min and stirred at rt for 5 h. THF was evaporated under vacuum and the remaining aqueous phase was extracted with DCM (3x40 mL). The combined organic phases were washed with brine (40 mL), dried over anhydrous Na_2SO_4 and concentrated under low pressure to give compound **2**. The crude product was used in the next steps without further purification.

*4.1.02.1. (1^2Z,6R,7R,8E,13E,15E,17S)-17-hydroxy-6-isopropyl-7,15-dimethyl-3^2,3^3-dihydro-3^1H-5-oxa-11-aza-1(4,2)-oxazola-3(1,5)-pyrrolacycloicosaphane-8,13,15-triene-2,4,10,19-tetraone (**2**)*

Light-yellow powder; 172 mg (99%); R_f = 0.70 (DCM/EtOH = 90:10); ^{1}H-NMR (400 MHz, $CDCl_3$) δ: 7.89 (s, 1H), 7.53 (brt, 1H), 6.61 (dd, *J* = 7.2, 15.7 Hz, 1H), 6.15 (t, *J* = 2.9 Hz, 1H), 6.03 (dd, *J* = 16.3, 1.0 Hz, 1H), 5.84 (brd, 1H), 5.60 (ddd, *J* = 16.3, 3.7 Hz, 1H), 4.97 (dd, *J* = 10.0, 1.9 Hz, 1H), 4.92 (d, *J* = 9.2 Hz, 1H), 4.8 (m, 1H), 4.35 (m, 2H), 4.23 (m, 1H), 3.93 (d, *J* = 13.0 Hz, 2H), 3.72 (m, 2H), 3.17 (dd, *J* = 14.2, 9.5 Hz, 1H), 2.77 (dd, *J* = 14.0, 3.4 Hz, 1H), 2.03 (m, 2H), 1.59 (s, 3H), 1.13 (d, *J* = 6.7 Hz, 3H), 0.99 (t, *J* = 6.7 Hz, 6H); ^{13}C-NMR (100 MHz, $CDCl_3$) δ: 200.9, 167.5, 160.8, 160.0, 156.0, 145.4, 143.0, 137.0, 136.0, 134.6, 133.6, 126.1, 125.2, 122.8, 81.3, 66.0, 50.4, 47.5, 45.7, 37.6, 30.11, 29.8, 19.5, 18.8, 12.7, 12.2; HPLC: 21.020 min (HPLC method B); ESI(+)MS: *m/z* 548 $[M+Na]^+$, 508 $[M+H-H_2O]^+$.

4.1.03. General procedure for the synthesis of Diels-Alder adducts 1a, 1b, 2a and 2b

The diene (1.5 eq.) was added to a solution of **1** or **2** (0.04 mmol scale, 1 eq.) in anhydrous DCM (0.015 M) and stirred for 1 h at room temperature. The solution was cooled to -35 °C and $AlCl_3$ (1.5 eq.) was added. After 3 h at -35 °C, the reaction was left to rise back to room temperature overnight and finally heated under reflux conditions for 16 h. The reaction mixture was treated with saturated $NaHCO_3$ (1 mL), extracted with DCM (3x5 mL) and dried over anhydrous Na_2SO_4. The crude product was then purified via PLC.

4.1.03.1. (1^2S,1^3R,3^2Z,7S,12E,14R,15R)-1^3-((2-(diethylamino)ethyl)sulfonyl)-7-hydroxy-15-isopropyl-8^5,14-dimethyl-8^1,8^3,8^{3a},8^4,8^7,8^{7a}-hexahydro-16-oxa-10-aza-3(4,2)-oxazola-

*1(1,2)-pyrrolidina-8(4,7)-isobenzofuranacycloheptadecaphan-12-ene-8^{1},8^{3},2,5,11,17-hexaone (**1a**)*

The title compound was obtained from **1** (40 mg, 0.076 mmol) and maleic anhydride as diene (1.5 eq.) using the general procedure.

White powder; 25 mg (56%); R_f= 0.2 (DCM/EtOH = 90:10); $[\alpha]_D$=+70° (*c* 0.21, DCM); ^{1}H-NMR (400 MHz, $CDCl_3$) δ: 8.21 (s, 1H), 6.83 (d, *J* = 12.6 Hz, 1H), 5.92 (d, *J* = 15.2 Hz, 1H), 5.38 (s, 1H), 5.13 (s, 1H), 4.80 (d, *J* = 9.9 Hz, 1H), 4.24 (brs, 1H), 4.03-3.44 (m, 10H), 3.39-3.25 (m, 2H), 3.24-3.18 (m, 3H), 2.99-2.53 (m, 10H), 2.29 (brs, 1H), 1.95 (brs, 1H), 1.62 (s, 3H), 1.28-1.15 (m, 6H), 1.08 (brs, 3H), 0.95 (brs, 6H). ^{13}C-NMR (100 MHz, $CDCl_3$) δ: 200.8, 178.1, 174.8, 169.7, 167.7, 167.4, 159.7, 157.7, 147.8, 144.8, 136.6, 135.2, 123.1, 83.0, 65.5, 62.4, 59.4, 49.1, 48.9, 47.2, 46.9 (2C), 45.3, 44.4, 43.5, 43.1, 42.7, 40.2, 37.1, 35.7, 29.2, 26.1, 19.7, 19.6, 18.7, 11.7 (2C), 9.9; FT-IR: 2980, 2940, 1746, 1624, 1618, 1495, 1190, 1135, 880, 731 cm^{-1}; HPLC: 6.632 min (acetonitrile/water = 35:65 + 1% TFA); ESI(+)MS: *m/z* 789 $[M+H]^+$, 811 $[M+Na]^+$; HR-ESI(+)MS: calcd for $C_{38}H_{53}N_4O_{12}S$ *m/z* 789.3380; found *m/z* 789.3381.

*4.1.03.2. (1^{2}S,1^{3}R,3^{2}Z,7S,12E,14R,15R)-1^{3}-((2-(diethylamino)ethyl)sulfonyl)-7-hydroxy-15-isopropyl-8^{5},14-dimethyl-8^{2},8^{3},8^{3a},8^{4},8^{7},8^{7a}-hexahydro-8^{1}H-16-oxa-10-aza-3(4,2)-oxazola-8(4,7)-isoindola-1(1,2)-pyrrolidinacycloheptadecaphan-12-ene-8^{1},8^{3},2,5,11,17-hexaone (**1b**)*

The title compound was obtained from **1** (20 mg, 0.038 mmol) and maleimide as diene (1.5 eq.) using the general procedure.

Pale yellow powder; 9.8 mg (43%); R_f= 0.34 (DCM/EtOH = 9:1); $[\alpha]_D$=+14° (*c* 0.22, DCM); ^{1}H-NMR (400 MHz, $CDCl_3$) δ: 8.18 (s, 1H), 6.55 (m, 1H), 6.54 (dd, *J* = 16.4, 5.2 Hz, 1H), 5.79 (d, *J* = 16.4 Hz, 1H), 5.41 (brs, 1H), 5.36 (brs, 1H), 4.77 (dd, *J* = 9.8, 1.8 Hz, 1H), 4.74-4.69 (m, 1H), 4.40 (dt, *J* = 10.3, 7.6 Hz, 1H), 4.14-3.75 (m, 5H), 3.58-3.53 (m, 1H), 3.43 (t, *J* = 8.0 Hz, 1H), 3.37-3.01 (8H), 2.98-2.87 (m, 1H), 2.84-2.73 (m, 1H), 2.62-2.54 (m, 6H), 2.39-2.23 (2H), 2.05-1.93 (m, 1H), 2.88 (s, 3H), 2.13-2.03 (m, 9H), 0.98 (t, *J* = 7.4 Hz, 6H); ^{13}C-NMR (100 MHz, $CDCl_3$) δ: 201.2, 180.4, 178.1, 168.7, 166.5, 159.9, 157.9, 145.0, 144.2, 138.9,

136.4, 125.4 124.3, 82.9, 65.4, 62.3, 59.4, 49.1, 48.9, 47.3, 47.0 (2C), 45.3, 44.4, 43.5, 43.1, 42.7, 40.2, 37.1, 35.7, 29.2, 26.1, 19.7, 19.6, 18.6, 11.7 (2C), 9.9; FT-IR: br 3367, 2967, 2927, 1713, 1669, 1620, 1578, 1423, 1312, 1191, 1136, 1046, 988, 891, 813, 748, 635 cm^{-1}; HPLC: 5.343 min (acetonitrile/water = 65:35 + 1% TFA); ESI(+)MS: *m/z* 788 $[M+H]^+$, 810 $[M+Na]^+$; HR-ESI(+)MS: calcd for $C_{38}H_{54}N_5O_{11}S$ *m/z* 788.3540; found *m/z* 788.3533.

*4.1.03.3. (3²Z,7S,12E,14R,15R)-7-hydroxy-15-isopropyl-8⁵,14-dimethyl-1²,1³,8¹,8³,8ᵃ,8⁴,8⁷,8⁷ᵃ-octahydro-1¹H-16-oxa-10-aza-3(4,2)-oxazola-1(1,5)-pyrrola-8(4,7)-isobenzofuranacycloheptadecaphan-12-ene-8¹,8³,2,5,11,17-hexaone (**2a**)*

The title compound was obtained from **2** (15 mg) and maleic anhydride as diene (3 eq.) using the general procedure A without $AlCl_3$ catalysis.

White powder; 12 mg (76%); R_f= 0.60 (DCM/EtOH = 9:1); $[\alpha]_D$=+16° (*c* 0.43, DCM); ^{1}H-NMR (400 MHz, CD_3OD) δ: 8.35 (s, 1H), 6.78 (dd, *J* = 14.1, 4.4 Hz, 1H), 6.39 (s, 1H), 6.36 (d, *J* = 12.6 Hz, 1H), 6.26 (s, 1H), 6.15 (d, *J* = 11.2 Hz, 1H), 5.95 (d, *J* = 15.7 Hz, 1H), 5.37 (brs, 1H), 4.22 (q, *J* = 7.8 Hz, 2H), 4.10-3.95 (m, 3H), 3.63 (dd, *J* = 12.7, 6.3 Hz, 1H), 3.46 (dd, *J* = 15.6, 4.8 Hz, 1H), 3.28-3.12 (m, 2H), 3.05 (d, *J* = 16.6 Hz, 1H), 2.89-2.71 (m, 2H), 2.65-2.53 (m, 1H), 2.51-2.41 (m, 1H), 2.12-1.95 (m, 1H), 1.77 (s, 3H), 1.31 (t, *J* = 6.7 Hz, 3H), 1.18 (d, *J* = 6.2 Hz, 3H), 1.01 (dd, *J* = 9.5, 7.2 Hz, 6H); ^{13}C-NMR (100 MHz, CD_3OD) δ: 201.8, 178.4, 173.4, 169.3, 167.5, 166.2, 160.7, 159.6, 158.9, 146.7, 144.1, 136.4, 136.2, 135.6, 135.3, 132.9, 127.4, 126.5, 123.2, 122.5, 81.5, 78.4, 60.5, 51.4, 42.6, 41.7, 40.9, 38.8, 36.7, 29.5, 29.2, 19.2, 18.7, 17.3, 12.8, 9.2; FT-IR: 3386 (br), 2970, 2930, 1767, 1720, 1663, 1619, 1577, 1416, 1369, 1214, 1172, 1039, 989, 868, 743, 729 cm^{-1}; HPLC: 13.394 min (HPLC method B); ESI(+)MS *m/z* 646 $[M+Na]^+$; HR-ESI(+)MS calcd for $C_{32}H_{38}N_3O_{10}$ 624.2552; found *m/z* 624.2553.

*4.1.03.4. (3²Z,7S,12E,14R,15R)-7-hydroxy-15-isopropyl-8⁵,14-dimethyl-1²,1³,8²,8³,8ᵃ,8⁴,8⁷,8⁷ᵃ-octahydro-1¹H,8¹H-16-oxa-10-aza-3(4,2)-oxazola-8(4,7)-isoindola-1(1,5)-pyrrolacycloheptadecaphan-12-ene-8¹,8³,2,5,11,17-hexaone (**2b**)*

The title compound was obtained from **2** (10 mg, 0.015 mmol) and maleimide as diene (1.5 eq.) using the general procedure.

Pale yellow solid; 5 mg (42%); R_f = 0.48 (DCM/EtOH = 9:1); $[\alpha]_D$ = −80° (*c* 0.22, DCM); ^{1}H-NMR (400 MHz, $CDCl_3$) δ: 8.15 (s, 1H), 6.91-6.86 (m, 1H), 6.54 (dd, *J* = 16.7, 5.6 Hz, 1H), 6.23 (t, *J* = 2.8 Hz, 1H), 5.85 (d, *J* = 16.5 Hz, 1H), 5.36 (brs, 1H), 4.87 (brdd, *J* = 10.3, 1.7 Hz, 1H), 4.71-4.64 (m, 1H), 4.28-4.02 (m, 4H), 3.76 (d, *J* = 14.8 Hz, 1H), 3.48-3.42 (m, 2H), 3.21-3.11 (m, 2H), 2.96 (dd, *J* = 15.7, 4.9 Hz, 1H), 2.79-2.72 (m, 3H), 2.49-2.40 (m, 1H), 2.30-2.24 (m, 1H), 2.06-1.95 (m, 2H), 1.88 (s, 3H), 1.12 (d, *J* = 6.9 Hz, 3H), 0.98 (d, *J* = 6.8 Hz, 6H); ^{13}C-NMR (101 MHz, DMSO-d6) (mixture of isomers in a 1:1 ratio) δ: 202.2, 178.2, 174.2, 168.10, 161.2, 159.6, 147.1, 144.31, 136.5, 136.1, 126.9, 125.8, 123.4, 122.8, 120.9, 81.0, 55.3, 54.7, 53.4, 51.3, 43.0, 41.5, 41.2, 37.0, 34.7, 31.7, 29.6, 28.2, 22.3, 22.0, 20.8, 20.5, 19.2, 18.8, 17.6, 17.5, 13.0, 8.9, 7.7; FT-IR: 3278 (br), 2958, 2916, 2853, 1725, 1706, 1621, 1406, 1263, 1263, 1404, 1177, 1017, 802 cm^{-1}; HPLC: 14.847 min (HPLC method B); ESI(+)MS *m/z* 645 $[M+Na]^+$; HR-ESI(+)MS calcd for $C_{32}H_{38}N_4O_9Na$ *m/z* 645.2531; found *m/z* 645.2563.

4.1.04. General procedure for the synthesis of C16 alcohol compounds (16R)-1c, (16S)-1c, (16R)-2c and (16S)-2c.

$NaBH_4$ (0.5 eq.) was added at 0 °C to a solution of **1** or **2** (0.05 mmol scale) in EtOH (3 mL). The solution was stirred for 1 h and was subsequently quenched with 100 µL of acetone. The solvent was removed under reduced pressure, the residue was resuspended in DCM (5 mL), and extracted with DCM (3x). The combined organic phases were washed with water (10 mL) and brine (10 mL), dried over anhydrous Na_2SO_4 and evaporated under low pressure to give a white solid powder. The raw reaction was purified *via* PLC or column chromatography (silica gel, DCM/EtOH = 80:20).

*4.1.04.1. (1²Z,3²S,3³R,6R,7R,8E,13E,15E,17S,19R)-3³-((2-(diethylamino)ethyl)sulfonyl)-17,19-dihydroxy-6-isopropyl-7,15-dimethyl-5-oxa-11-aza-1(4,2)-oxazola-3(1,2)-pyrrolidinacycloicosaphane-8,13,15-triene-2,4,10-trione (**(16R)-1c**)*

The title compound was obtained from **1** (32 mg, 0.05 mmol) and purified *via* PLC (DCM/EtOH = 80:20).

White powder: 17.3 mg (54%); R_f= 0.43 (DCM/EtOH = 90:10); $[\alpha]_D$= −71° (*c* 0.16, DCM); ^{1}H-NMR (400 MHz, $CDCl_3$) δ: 8.15 (s, 1H), 6.66 (dd, *J* = 16.0, 4.3 Hz, 1H), 6.22 (d, *J* = 15.7 Hz,

1H), 6.15 (dd, *J* = 6.8, 4.9 Hz, 1H), 5.80 (dd, *J* = 16.0, 1.8 Hz, 1H), 5.66 (ddd, *J* = 15.5, 13.1, 4.2 Hz, 1H), 5.45 (m, 2H), 4.89 (dd, *J* = 10.3, 1.9 Hz, 1H), 4.82 (dt, *J* = 9.1, 6.5 Hz, 1H), 4.24 (m, 3H), 3.99 (dt, *J* = 9.5, 3.8 Hz, 1H), 3.65 (m, 2H), 3.24 (m, 2H), 3.75-3.55 (m, 2H), 3.00 (dd, *J* = 16.1, 5.4 Hz, 1H), 2.82 (dd, *J* = 16.4, 6.7 Hz, 1H), 2.90-2.73 (m, 1H), 2.59 (dq, *J* = 7.2, 2.8 Hz, 4H), 2.60-2.42 (m, 1H), 2.19-1.90 (m, 1H), 2.00-1.95 (m, 1H), 1.92-1.83 (m, 2H), 1.80 (s, 3H), 1.12 (d, *J* = 7.0 Hz, 3H), 1.07 (t, *J* = 7.1 Hz, 6H), 0.97 (t, *J* = 6.0 Hz, 6H); ^{13}C-NMR (100 MHz, $CDCl_3$) δ: 168.0, 165.3, 161.6, 160.3, 146.0, 143.9, 136.5, 135.8, 134.6, 134.0, 124.6, 123.7, 82.8, 68.0, 67.3, 62.4, 59.2, 49.1, 47.0, 46.9 (2C), 45.2, 43.3, 40.9, 36.4, 35.7, 29.3, 25.5, 19.7, 18.4, 13.1, 11.6 (2C), 10.6; FT-IR: br 3409, 2362, 2339, 1747, 1669, 1617, 1428, 1188, 1048, 968, 892, 747, 542 cm^{-1}; HPLC: 15.125 min (isocratic ACN/water 70:30 + 1% TFA); ESI(+)MS: *m/z* 693 $[M+H]^+$, 715 $[M+Na]^+$, 675 $[M+H-H_2O]^+$; HR-ESI(+)MS calcd for $C_{34}H_{53}N_4O_9S$ *m/z* 693.3528; found *m/z* 693.3546.

*4.1.04.2. (1²Z,3²S,3³R,6R,7R,8E,13E,15E,17S,19S)-3³-((2-(diethylamino)ethyl)sulfonyl)-17,19-dihydroxy-6-isopropyl-7,15-dimethyl-5-oxa-11-aza-1(4,2)-oxazola-3(1,2)-pyrrolidinacycloicosaphane-8,13,15-triene-2,4,10-trione (**(16S)-1c**)*

The title compound was obtained from **1** (32 mg, 0.05 mmol) and purified *via* PLC (DCM/EtOH = 80:20).

White powder; 11.8 mg (37%); R_f= 0.56 (DCM/EtOH = 90:10); $[\alpha]_D$=−83° (*c* 0.58, DCM); ^{1}H-NMR (400 MHz, $CDCl_3$) δ: 8.15 (s, 1H), 6.61 (dd, *J* = 16.2, 4.1 Hz, 1H), 6.27 (d, *J* = 15.6 Hz, 1H), 5.98 (brs, 1H), 5.82 (d, *J* = 9.4 Hz, 1H), 5.77 (dd, *J* = 16.2, 2.0 Hz, 1H), 5.86-5.65 (m, 1H), 5.38 (s, 1H), 5.02 (dt, *J*=9.4, 4.0 Hz, 1H), 4.84 (dd, *J* = 10.4, 1.9 Hz, 1H), 4.52 (brt, 1H), 4.30-4.15 (m, 1H), 4.11 (dt, *J* = 11.2, 8.2 Hz, 1H), 3.94 (dt, *J* = 10.3, 3.9 Hz, 1H), 3.57 (m, 1H), 3.49 (brdt, *J* = 8.0, 1.9 Hz, 1H), 3.36-3.25 (m, 1H), 3.26-3.13 (m, 1H), 3.12-3.00 (m, 3H), 2.92 (dd, *J* = 16.7, 9.9 Hz, 1H), 2.81 (m, 1H), 2.59 (dq, *J* = 7.2, 1.9 Hz, 2H), 2.47-2.41 (m, 1H), 2.17 (brd, *J* = 14.3 Hz, 1H), 2.18-2.00 (m, 1H), 2.00-1.92 (m, 2H), 1.76 (s, 3H), 1.11 (d, *J* = 7.0 Hz, 3H), 1.08 (t, *J* = 7.2 Hz, 6H), 1.00 (dd. *J* = 15.2, 6.6 Hz, 6H); ^{13}C-NMR (100 MHz, $CDCl_3$) δ: 168.8, 165.7, 161.3, 160.3, 145.7, 143.9, 137.4, 136.0, 133.6, 133.6, 125.0, 123.5, 82.9, 67.6, 66.7, 62.3, 59.3, 49.2, 47.0 (2C), 46.7, 45.2, 41.4, 41.2, 36.6, 35.2, 29.3, 25.8, 19.8, 18.5, 12.7, 11.7 (2C), 9.9; FT-IR: 3303 (br), 2968, 2929, 2358, 1748, 1669, 1620, 1582, 1427, 1189, 1134, 1063, 967, 891, 749, 541 cm^{-1}; HPLC: 15.732 min (isocratic

acetonitrile/Water 70:30 + 1% TFA); ESI(+)MS: *m/z* 693 $[M+H]^+$, 715 $[M+Na]^+$, 675 $[M+H-H_2O]^+$; HR-ESI(+)MS: calcd for $C_{34}H_{53}N_4O_9S$ *m/z* 693.3528; found *m/z* 693.3546.

*4.1.04.3. (1²Z,6R,7R,8E,13E,15E,17S,19R)-17,19-dihydroxy-6-isopropyl-7,15-dimethyl-3²,3³-dihydro-3¹H-5-oxa-11-aza-1(4,2)-oxazola-3(1,5)-pyrrolacycloicosaphane-8,13,15-triene-2,4,10-trione (**(16R)-2c**)*

The title compound was obtained from **2** (40 mg, 0.076 mmol) and purified *via* PLC (DCM/EtOH = 90:10).

White powder; 16 mg (40%); R_f= 0.55 (DCM/EtOH = 90:10); $[\alpha]_D$=−102° (*c* 0.87, DCM); ^{1}H-NMR (400 MHz, $CDCl_3$) δ: 7.97 (s, 1H), 6.88 (brt, 1H), 6.60 (dd, *J* = 16.1, 6.4 Hz, 1H), 6.16 (s, 1H), 6.00-5.90 (m, 2H), 5.64 (dt, *J* = 15.7, 5.4 Hz, 1H), 5.09 (d, *J* = 8.9 Hz, 1H), 4.94 (d, *J* = 9.7 Hz, 1H), 4.73-6.64 (m, 1H), 4.36-4.14 (m, 2H), 4.14-4.05 (m, 1H), 4.04-4.90 (m, 2H), 3.03 (dd, *J* = 15.0, 4.6 Hz, 1H), 2.93 (dd, *J* = 15.0, 8.2 Hz, 1H), 2.77 (m, 3H), 2.08-1.80 (m, 3H), 1.76 (s, 3H), 1.13 (d, *J* = 6.9 Hz, 3H), 0.99 (d, *J* = 6.1 Hz, 6H); ^{13}C-NMR (100 MHz, $CDCl_3$) δ: 167.3, 161.5, 160.8, 160.5, 144.0, 143.7, 137.1, 135.6, 134.6, 134.6, 132.8, 125.4, 124.8, 123.8, 81.2, 67.7, 67.2, 50.9, 43.9, 40.7, 37.5, 36.6, 30.1, 29.8, 19.6, 18.9, 13.4, 12.0; FT-IR: 3309 (br), 2919, 1728, 1617, 1586, 1416, 1224, 1179, 1161, 1109, 1063, 987, 941, 886, 746 cm^{-1}; HPLC: 6.973 min (HPLC method A); ESI(+)MS: *m/z* 550 $[M+Na]^+$, 532 $[M+Na-H_2O]^+$; HR-ESI(+)MS: calcd for $C_{28}H_{37}N_3O_7Na$ *m/z* 550.2524; found *m/z* 550.2542.

*4.1.04.4. (1²Z,6R,7R,8E,13E,15E,17S,19S)-17,19-dihydroxy-6-isopropyl-7,15-dimethyl-3²,3³-dihydro-3¹H-5-oxa-11-aza-1(4,2)-oxazola-3(1,5)-pyrrolacycloicosaphane-8,13,15-triene-2,4,10-trione (**(16S)-2c**)*

The title compound was obtained from **2** (40 mg, 0.076 mmol) and purified *via* PLC (DCM/EtOH = 90:10).

White powder; 13 mg (32%); R_f= 0.76 (DCM/EtOH = 90:10); $[\alpha]_D$=−66° (*c* 0.18, DCM); ^{1}H-NMR (400 MHz, $CDCl_3$) δ: 8.08 (s, 1H), 6.52 (dd, *J* = 16.2, 5.3 Hz, 1H), 6.43 (brd, *J* = 7.4 Hz, 1H), 6.09 (d, *J* = 15.4 Hz, 1H), 6.07 (d, *J* = 3.0 Hz, 1H), 5.91 (d, *J* = 16.2 Hz, 1H), 5.61 (d, *J* = 8.4 Hz, 1H), 5.58 (dd, *J* = 10.0, 4.0 Hz, 1H), 4.98-4.91 (m, 2H), 4.55-4.42 (m, 2H), 4.31-4.13 (m, 2H), 3.45-3.33 (m, 1H), 3.11 (dd, *J* = 15.9, 3.3 Hz, 1H), 2.89 (dd, *J* = 15.9, 10.3 Hz, 1H),

2.81-2.62 (m, 3H), 2.10-1.94 (m, 3H), 1.69 (s, 3H), 1.13 (d, *J* = 7.2 Hz, 3H), 1.00 (t, *J* = 6.2 Hz, 6H); ^{13}C-NMR (100 MHz, $CDCl_3$) δ: 167.4, 161.4, 161.2, 159.8, 143.9, 143.3, 137.3, 136.2, 135.9, 133.6, 133.1, 125.6, 124.8, 123.5, 80.9, 67.6, 66.7, 50.4, 42.2, 41.5, 37.2, 35.5, 30.0, 29.7, 19.6, 18.8, 13.0, 11.2; FT-IR: 3330 (br), 2919, 1728, 1617, 1416, 1224, 1179, 1161, 1109, 1063, 987, 941, 886, 746 cm^{-1}; HPLC: 7.206 min (HPLC method A); ESI(+)MS: *m/z* 550 $[M+Na]^+$, 532 $[M+Na-H_2O]^+$; HR-ESI(+)MS: calcd for $C_{28}H_{37}N_3O_7Na$ *m/z* 550.2524; found *m/z* 550.2542.

4.1.05. General procedure for the synthesis of NHMe-C16 compounds (16R)-1d, (16S)-1d, (16R)-2d and (16S)-2d

Methylamine (5 eq.) and acetic acid (0.1 eq.) were added to a stirred solution of **1** or **2** (0.04 mmol scale, 1 eq.) in dry ACN (3 mL) at 0 °C under N_2 atmosphere. After 48 h, $NaCNBH_3$ (1.5 eq.) was added to the reaction mixture and stirred for 3 h at 0 °C. The reaction was quenched with acetone (1 mL), evaporated under reduced pressure and resuspended in DCM (10 mL). The organic phase was then washed with saturated $NaHCO_3$ (2x5 mL).

*4.1.05.1. (1²Z,3²S,3³R,6R,7R,8E,13E,15E,17S,19R)-3³-((2-(diethylamino)ethyl)sulfonyl)-17-hydroxy-6-isopropyl-7,15-dimethyl-19-(methylamino)-5-oxa-11-aza-1(4,2)-oxazola-3(1,2)-pyrrolidinacycloicosaphane-8,13,15-triene-2,4,10-trione (**(16R)-1d**)*

The title compound was obtained from **1** (30 mg, 0.043 mmol) and purified *via* PLC (DCM/MeOH = 90:10 + 1% TEA).

White powder; 2.4 mg (7.8%); R_f= 0.41 (DCM/EtOH = 90:10 + 1% of TEA); $[\alpha]_D$= −35° (*c* 0.19, DCM); ^{1}H-NMR (400 MHz, $CDCl_3$) δ: 8.12 (s, 1H), 6.70 (dd, *J* = 16.0, 4.1 Hz, 1H), 6.24 (d, *J* = 15.8 Hz, 1H), 6.12 (br, 1H), 5.80 (d, *J* = 16.0 Hz, 1H), 5.66 (ddd, *J* = 15.3, 8.6, 4.5 Hz, 1H), 5.46 (d, *J* = 8.4 Hz, 1H), 5.39 (s, 1H), 4.90 (d, *J* = 10.1 Hz, 1H), 4.75-4.68 (m, 1H), 4.28-3.76 (m, 3H), 3.60 (d, *J* = 8.1 Hz, 1H), 3.36-2.68 (m, 10H), 2.67-2.57 (m, 4H), 2.53 (s, 3H), 2.50-2.24 (m, 2H), 2.09-2.03 (m, 3H), 1.80 (s, 3H), 1.13-1.06 (m, 9H), 0.99 (t, *J* = 6.5 Hz, 6H); ^{13}C-NMR (100 MHz, $CDCl_3$ + CD_3OD) δ: 168.1, 165.3, 160.9, 160.6, 146.3, 143.8, 137.4, 135.9, 134.4, 133.9, 124.4, 123.6, 82.8, 68.1, 62.7, 59.3, 57.3, 47.1, 46.9 (2C), 45.0, 40.7, 40.3, 36.6, 32.6, 31.9, 29.3, 25.5, 22.6, 19.6, 14.1, 12.9, 11.5 (2C), 10.3; FT-IR: 3332 (br),

2922, 2840, 1743, 1621, 1430, 1187, 1122, 1039, 969, 891, 810, 747 cm^{-1}; HPLC: 14.293 min (HPLC method B); ESI(+)MS: *m/z* 705 $[M+H]^+$, 728 $[M+Na]^+$; HR-ESI(+)MS: calcd for $C_{35}H_{56}N_5O_8S$ *m/z* 706.3844; found *m/z* 706.3841.

*4.1.05.2. (1²Z,3²S,3³R,6R,7R,8E,13E,15E,17S,19S)-3³-((2-(diethylamino)ethyl)sulfonyl)-17-hydroxy-6-isopropyl-7,15-dimethyl-19-(methylamino)-5-oxa-11-aza-1(4,2)-oxazola-3(1,2)-pyrrolidinacycloicosaphane-8,13,15-triene-2,4,10-trione (**(16S)-1d**)*

The title compound was obtained from **1** (30 mg, 0.043 mmol) and purified *via* PLC (DCM/MeOH = 90:10 + 1% TEA).

White powder; 3.2 mg (10%); R_f= 0.41 (DCM/EtOH = 90:10 + 1% of TEA); $[\alpha]_D$=−50° (*c* 0.11, DCM); ^{1}H-NMR (400 MHz, $CDCl_3$) δ: 8.14 (s, 1H), 6.62 (dd, *J* = 16.6, 4.3 Hz, 1H), 6.29 (d, *J* = 15.6 Hz, 1H), 5.83-5.74 (m, 2H), 5.73-5.61 (m, 1H), 5.38 (brs, 1H), 4.89 (brd, *J* = 9.6 Hz, 1H), 4.37-3.41 (m, 8H), 3.38-2.87 (m, 7H), 2.87-2.56 (m, 6H), 2.53 (s, 3H), 2.13-1.82 (m, 4H), 1.76 (s, 3H), 1.08-1.05 (m, 3H), 1.02-0.95 (m, 6H), 0.88 (t, *J* = 6.5 Hz, 6H); ^{13}C-NMR (100 MHz, $CDCl_3$ + 5% CD_3OD) δ: 168.1, 165.7, 160.5, 159.9, 146.0, 144.0, 138.0, 136.1, 134.5, 133.3, 124.2, 123.4, 82.8, 67.0, 62.4, 59.3, 54.8, 47.0 (2C), 46.7, 45.0, 36.6, 33.5, 32.4, 31.9, 29.3, 25.9, 22.7, 19.7, 14.0, 12.5, 11.6 (2C), 9.9; FT-IR: 3274 (br), 2970, 1741, 1623, 1431, 1191, 1123, 970, 809 cm^{-1}; HPLC: 13.958 min (HPLC method B); ESI(+)MS: *m/z* 705 $[M+H]^+$, 728 $[M+Na]^+$; HR-ESI(+)MS: calcd for $C_{35}H_{56}N_5O_8S$ *m/z* 706.3844; found *m/z* 706.3841.

*4.1.05.3. (1²Z,6R,7R,8E,13E,15E,17S,19R)-17-hydroxy-6-isopropyl-7,15-dimethyl-19-(methylamino)-3²,3³-dihydro-3¹H-5-oxa-11-aza-1(4,2)-oxazola-3(1,5)-pyrrolacycloicosaphane-8,13,15-triene-2,4,10-trione (**(16R)-2d**)*

The title compound was obtained from **2** (30 mg, 0.06 mmol) and purified *via* PLC (DCM/EtOH = 90:10 + 1% TEA).

White powder; 3.6 mg (11.5%); R_f= 0.35 (DCM/EtOH = 90:10 + 0.1% TEA); $[\alpha]_D$=−60° (*c* 0.16, DCM); ^{1}H-NMR (400 MHz, $CDCl_3$) δ: 7.90 (s, 1H), 7.33 (m, 1H), 6.60 (dd, *J* = 16.5, 7.3 Hz, 1H), 6.13 (t, *J* = 2.8 Hz, 1H), 6.01 (d, *J* = 16.2 Hz, 1H), 5.90 (d, *J* = 15.7 Hz, 1H), 5.62 (dt, *J* = 15.9, 4.9 Hz, 1H), 4.95 (dd, *J* = 10.3, 1.5 Hz, 1H), 4.90 (d, *J* = 8.8 Hz, 1H), 4.52 (ddd, *J* =

14.6, 9.2, 3.8 Hz, 1H), 4.39-4.20 (m, 3H), 3.88 (brd, *J* = 17.8 Hz, 1H), 3.10 (d, *J* = 12.0 Hz, 1H), 2.90-2.78 (m, 1H), 2.78-2.63 (m, 4H), 2.57 (s, 3H), 2.10-1.90 (m, 4H), 1.85-1.77 (m, 1H), 1.72 (s, 3H), 1.12 (d, *J* = 6.7 Hz, 3H), 1.00-0.96 (m, 6H); ^{13}C-NMR (100 MHz, $CDCl_3$) δ: 167.9, 161.2, 161.1, 160.7, 144.7, 144.1, 136.8, 135.2, 133.8, 133.6, 132.2, 124.6, 124.5, 123.6, 81.5, 57.7, 55.4, 50.7, 41.3, 40.0, 37.6, 31.8, 30.1, 19.4, 18.8, 17.8, 13.9, 13.0, 12.0; FT-IR: 3343 (br), 2969, 2921, 2850, 1727, 1667, 1610, 1587, 1420, 1372, 1164, 1117, 978, 746 cm^{-1}; HPLC: 1.945 min (HPLC method A); ESI(+)MS: *m/z* 541 $[M+H]^+$, 563 $[M+Na]^+$, 523 $[M-H_2O]^+$; HR-ESI(+)MS: calcd for $C_{29}H_{41}N_4O_6$ *m/z* 541.3021; found *m/z* 541.3015.

*4.1.05.4. (1^{2}Z,6R,7R,8E,13E,15E,17S,19S)-17-hydroxy-6-isopropyl-7,15-dimethyl-19-(methylamino)-3^2,3^3-dihydro-3^{1}H-5-oxa-11-aza-1(4,2)-oxazola-3(1,5)-pyrrolacycloicosaphane-8,13,15-triene-2,4,10-trione (**(16S)-2d**)*

The title compound was obtained from **2** (30 mg, 0.06 mmol) and purified *via* PLC (DCM/EtOH = 90:10 + 1% TEA).

White powder; 4.1 mg (13%); R_f= 0.35 (DCM/EtOH = 90:10 + 0.1% TEA); $[\alpha]_D$=−77° (*c* 0.26, DCM); ^{1}H-NMR (400 MHz, $CDCl_3$) δ: 8.07 (s, 1H), 6.79 (m, 1H), 6.55 (dd, *J* = 15.8, 5.5 Hz, 1H), 6.09 (brs, 1H), 6.01 (d, *J* = 15.5 Hz, 1H), 5.94 (d, *J* = 15.8 Hz, 1H), 5.56-5.42 (m, 2H), 4.94 (d, *J* = 9.4 Hz, 1H), 4.89-4.83 (m, 1H), 4.55-4.38 (m, 2H), 4.20-4.11 (m, 1H), 3.62-3.53 (m, 1H), 3.45-3.38 (m, 1H), 3.35 (d, *J* = 15.3 Hz, 1H), 3.15-3.08 (m, 1H), 3.07-2.97 (m, 1H), 2.77-2.73 (m, 2H), 2.69 (s, 3H), 2.06-1.90 (m, 4H), 1.78-1.72 (m, 1H), 1.61 (s, 3H), 1.14 (d, *J* = 7.2 Hz, 3H), 0.99 (d, *J* = 6.6 Hz, 6H); ^{13}C-NMR (100 MHz, CD_3OD) δ: 165.3, 160.5, 160.1, 159.9, 145.4, 144.3, 137.0, 136.6, 135.2, 134.2, 132.3, 124.6, 123.7 (2C), 80.1, 65.9, 54.0, 42.1, 40.5, 36.9, 35.6, 34.6, 31.7, 22.3, 17.4, 13.0, 11.4, 10.2, 9.1; FT-IR: 3329 (br), 2955, 2919, 2850, 1730, 1667, 1617, 1587, 1417, 1371, 1227, 1163, 1109, 1056, 997, 746 cm^{-1}; HPLC: 2.118 min (HPLC method A); ESI(+)MS: *m/z* 541 $[M+H]^+$, 563 $[M+Na]^+$, 523 $[M-H_2O]^+$; HR-ESI(+)MS: calcd for $C_{29}H_{41}N_4O_6$ *m/z* 541.3021; found *m/z* 541.30162.

4.1.06. General procedure for the synthesis of C14-EVE protected compounds 1f and 2f

Compound **1** or **2** (0.1-0.01 mmol scale) was dissolved in DCM/EVE 1:1, and successively PPTS (1 eq.) was added. The reaction was stirred overnight at room temperature. The solvents were evaporated under vacuum, the residue was resuspended in DCM (10 mL)

and washed with water (3x15 mL) and brine (15 mL). The organic phase was dried over anhydrous Na_2SO_4 and evaporated under low pressure. The crude oil was used in the next step without further purification.

*4.1.06.1. (1²Z,3²S,3³R,6R,7R,8E,13E,15E,17S)-3³-((2-(diethylamino)ethyl)sulfonyl)-17-(1-ethoxyethoxy)-6-isopropyl-7,15-dimethyl-5-oxa-11-aza-1(4,2)-oxazola-3(1,2)-pyrrolidinacycloicosaphane-8,13,15-triene-2,4,10,19-tetraone (**1f**)*

The title compound was obtained from **1** (100 mg, 0.15 mmol) in 0.3 mL DCM and 0.3 mL EVE as a mixture of isomers in a 1:1 ratio.

Yellow oil; 77 mg (70.6%); R_f = 0.37 (DCM/EtOH = 95:5); ^{1}H-NMR (400 MHz, $CDCl_3$) δ (mixture of isomers in a 1:1 ratio): 8.03 [8.02] (s, 1H), 6.91 [6.91] (bs, 1H), 6.59 [6.55] (dd, *J* = 16.0, 1.6 Hz, 1H), 6.09 [6.09] (d, *J* = 15.6 Hz, 1H), 5.84 [5.84] (d, *J* = 15.6 Hz, 1H), 5.71-5.58 [5.71-5.58] (m, 1H), 5.33 [5.33] (s, 1H), 5.28 [5.13] (d, *J* = 9.2 Hz, 1H), 4.87-4.53 [4.87-4.53] (m, 4H), 4.31-4.13 [4.31-4.13] (m, 3H), 3.88 [3.88] (d, *J* = 15.6 Hz, 1H), 3.83-3.77 [3.83-3.77] (m, 1H), 3.69 [3.69] (d, *J* = 16.0 Hz, 1H), 3.65-3.20 [3.65-3.20] (m, 6H), 2.97 [2.93] (d, *J* = 7.6 Hz, 1H), 2.88-2.78 [2.88-2.78] (m, 1H), 2.77-2.47 [2.77-2.47] (m, 5H), 2.29-2.14 [2.29-2.14] (m, 1H), 1.90-1.49 [1.90-1.49] (m, 2H), 1.72 [1.68] (s, 3H), 1.28-1.12 [1.28-1.12] (m, 6H), 1.09 [1.09] (t, *J* = 6.8 Hz, 6H), 1.08 [1.02] (d, *J* = 6.8 Hz, 3H), 0.93 [0.91] (d, *J* = 6.4 Hz, 6H); ^{13}C-NMR (100 MHz, $CDCl_3$) δ (mixture of isomers in a 1:1 ratio): 199.9 [199.8], 168.5 [168.4], 165.8 [165.7], 160.3 [160.3], 157.1 [157.0], 144.6 [144.6], 144.3 [144.2], 136.8 [136.5], 136.2 [134.2], 132.3 [131.1], 129.1 [125.6], 125.3 [125.1], 124.5 [124.5], 98.9 [97.2], 82.6 [82.5], 67.9 [67.4], 62.9 [62.9], 61.0 [61.0], 59.5 [59.5], 48.8 [48.8], 47.8 [47.7], 47.0 [47.0], 49.9 [46.9] (2C), 45.6 [45.6], 43.7 [43.7], 40.6 [40.5], 36.5 [36.5], 29.4 [29.4], 25.9 [25.9], 20.3 [20.2], 19.6 [19.6] (2C), 18.7 [18.7], 12.7 [12.7], 11.2 [11.1], 10.9 [10.9] (2C); ESI(+)MS: *m/z* 763 $[M+H]^+$, 785 $[M+Na]^+$.

*4.1.06.2. (1²Z,6R,7R,8E,13E,15E,17S)-17-(1-ethoxyethoxy)-6-isopropyl-7,15-dimethyl-3²,3³-dihydro-3¹H-5-oxa-11-aza-1(4,2)-oxazola-3(1,5)-pyrrolacycloicosaphane-8,13,15-triene-2,4,10,19-tetraone (**2f**)*

The title compound was obtained from **2** (30 mg, 0.0571 mmol) in 0.2 mL DCM and 0.2 mL EVE as a mixture of isomers in a 1:1 ratio.

Yellow oil; 34 mg (98%); R_f = 0.77 (DCM/EtOH = 95:5); ^{1}H-NMR (400 MHz, $CDCl_3$) δ (mixture of isomers in a 1:1 ratio): 7.83 [7.81] (s, 1H), 7.71 [7.65] (brdd, *J* = 8.0, 3.0 Hz, 1H), 6.60 [6.59] (dd, *J* = 16.3, 7.6 Hz, 1H), 6.13 [6.13] (m, 1H), 6.01 [6.01] (brd, *J* = 16.4, 1H), 5.83 [5.82] (dd, *J* = 15.9, 4.8 Hz, 1H), 4.95 [4.95] (d, *J* = 10.3 Hz, 1H), 4.89 [4.89] (d, *J* = 9.6 Hz, 1H), 4.84 [4.84] (m, 1H), 4.7-4.57 (m, 2H), 4.47 [4.47] (m, 1H), 4.42-4.20 (m, 3H), 3.95 [3.94] (d, *J* = 13.3 Hz, 1H), 3.90-3.73 [3.90-3.73] (m, 1H), 3.65 [3.64] (d, *J* = 13.5 Hz, 1H), 3.45-3.35 [3.45-3.35] (m, 1H), 3.28-3.14 [3.28-3.14] (m, 1H), 2.96-2.82 [2.96-2.82] (m, 1H), 2.76-2.58 [2.76-2.58] (m, 3H), 2-08-2.00 [2.00] (m, 1H), 1.56 [1.56] (s, 3H), 1.55 [1.55] (s, 3H), 1.23- 1.07 (m, 6H), 0.95 [0.95] (m, 6H); ^{13}C-NMR (100 MHz, $CDCl_3$) δ (the values of the other isomer are reported in brackets): 200.0 [199.8], 167.5 [167.5], 160.8 [160.7], 160.3 [160.3], 155.9 [155.9], 145.4 [145.3], 142.9 [142.9], 137.0 [136.9], 136.6 [135.8], 134.6 [133.8], 133.3 [132.9], 129.8 [128.9], 125.9 [125.8], 125.4 [125.2], 122.9 [122.7], 98.3 [97.1], 81.4 [81.2], 68.4 [68.2], 61.2 [60.4], 50.6 [50.5], 46.5 [46.2], 45.8 [45.7], 40.3 [40.0], 37.8 [37.6], 30.1 [30.1], 29.9 [29.8], 20.4 [20.3], 19.5 [19.5], 18.8 [18.8], 15.4 [15.0], 12.7 [12.6], 12.3 [12.23]; ESI(+)MS: *m/z* 620 $[M+Na]^+$.

4.1.07. Synthesis of OH-C16 C14-protected compounds (16R)-1g, (16S)-1g, (16R)-1g and (16S)-1g

Starting from **1f** or **2f** the general procedure reported in section 4.1.04 was followed.

*4.1.07.1. (1^{2}Z,3^{2}S,3^{3}R,6R,7R,8E,13E,15E,17S,19R)-3^3-((2-(diethylamino)ethyl)sulfonyl)-17-(1-ethoxyethoxy)-19-hydroxy-6-isopropyl-7,15-dimethyl-5-oxa-11-aza-1(4,2)-oxazola-3(1,2)-pyrrolidinacycloicosaphane-8,13,15-triene-2,4,10-trione (**(16R)-1g**)*

The title compound was obtained from **1f** (97 mg, 0.13 mmol) and separated from 16*S* enantiomer *via* PLC (silica gel, DCM/EtOH = 95:5).

Yellow solid; 25 mg (25.7%); R_f = 0.22 (DCM/EtOH = 95:5); ^{1}H-NMR (400 MHz, $CDCl_3$) δ (mixture of isomers in a 1:1 ratio): 8.16 [8.15] (s, 1H), 6.66 [6.62] (dd, *J* = 9.5, 4.0 Hz, 1H), 6.23 [6.22] (d, *J* = 15.7 Hz, 1H), 6.08-6.00 [6.04] (m, 1H), 5.78 [5.78] (d, *J* = 15.7 Hz, 1H), 5.63 [5.63] (m, 1H), 5.46 [5.46] (d, *J* = 7.8 Hz, 1H), 5.43 [5.43] (d, *J* = 9.4 Hz, 1H), 4.88 [4.87] (d, *J* = 10.3 Hz, 1H), 4.85-4.69 [4.85-4.69] (m, 1H), 4.72-4.63 [4.72-4.63] (m, 1H), 4.30-4.17 [4.30-4.17] (m, 3H), 4.00 [4.00] (dq, *J* = 10.6, 4.1 Hz, 1H), 3.74-3.45 [3.74-3.45] (m, 4H),

3.33-3.17 [3.33-3.17] (m, 2H), 3.16-2.90 [3.16-2.90] (m, 3H), 2.88-2.71 [2.88-2.71] (m, 2H), 2.65-2.45 [2.65-2.45] (m, 5H), 2.15-1.69 [2.15-1.69] (m, 4H), 1.82 [1.80] (s, 3H), 1.30 [1.27] (d, *J* = 5.3 Hz, 3H), 1.22 [1.15] (t, *J* = 7.2 Hz, 3H), 1.11 [1.11] (d, *J* = 7.0 Hz, 3H), 1.07 [1.07] (t, *J* = 7.2 Hz, 6H), 0.96 [0.96] (t, *J* = 7.2 Hz, 6H); ^{13}C-NMR (100 MHz, $CDCl_3$) δ (mixture of isomers in a 1:1 ratio): 168.4 [168.3], 165.3 [165.2], 162.1 [161.9], 160.2 [160.1], 145.9 [145.7], 144.3 [144.1], 136.9 [136.8], 135.9 [135.6], 134.1 [133.7], 132.7 [132.7], 124.7 [124.6], 123.6 [123.6], 98.9 [97.5], 82.6 [82.6], 70.9 [70.9], 68.1 [65.5], 62.5 [62.4], 61.4 [59.9], 59.3 [59.3], 49.2 [49.1], 47.0 [47.0] (2C), 45.3 [45.2], 41.3 [41.3], 41.3 [41.2], 36.5 [36.4], 35.7 [35.4], 29.3 [29.3], 28.5 [28.5], 25.5 [25.5], 20.6 [20.4], 19.8 [19.8], 18.5 [18.5], 15.5 [15.1], 13.2 [13.1], 11.6 [11.6] (2C), 10.6 [10.5]; ESI(+)MS: *m/z* 787 $[M+Na]^+$.

*4.1.07.2. (1²Z,3²S,3³R,6R,7R,8E,13E,15E,17S,19S)-3³-((2-(diethylamino)ethyl)sulfonyl)-17-(1-ethoxyethoxy)-19-hydroxy-6-isopropyl-7,15-dimethyl-5-oxa-11-aza-1(4,2)-oxazola-3(1,2)-pyrrolidinacycloicosaphane-8,13,15-triene-2,4,10-trione (**(16S)-1g**)*

The title compound was obtained from **1f** (97 mg, 0.13 mmol) and separated from 16*R* enantiomer *via* PLC (silica gel, DCM/EtOH = 95:5).Yellow solid; 60 mg (67.6%); R_f = 0.36 (DCM/EtOH = 95:5); ^{1}H-NMR (400 MHz, $CDCl_3$) δ (the values of the other isomer are reported in brackets): 8.14 [8.14] (s, 1H), 6.58 [6.57] (dd, *J* = 16.4, 4.0 Hz, 1H), 6.27 [6.27] (dd, *J* = 15.6, 4.4 Hz, 1H), 6.05-5.94 [6.05-5.94] (m, 1H), 5.83-5.64 [5.83-5.64] (m, 3H), 5.36 [5.36] (s, 1H), 5.01-4.80 [5.01-4.80] (m, 1H), 4.81 [4.81] (d, *J* = 10.4 Hz, 1H), 4.73 [4.64] (q, *J* = 5.2 Hz, 1H), 4.48-4.25 [4.48-4.25] (m, 2H), 4.15-3.85 [4.15-3.85] (m, 3H), 3.73-3.60 [3.73-3.60] (m, 1H), 3.56-3.37 [3.56-3.37] (m, 2H), 3.31-3.13 [3.31-3.13] (m, 2H), 3.12-2.97 [3.12-2.97] (m, 3H), 2.84 [2.82] (d, *J* = 10.0 Hz, 1H), 2.64-2.53 [2.64-2.53] (m, 4H), 2.51-2.40 [2.51-2.40] (m, 1H), 2.30-2.18 [2.30-2.18] (m, 1H), 2.16-1.89 [2.16-1.89] (m, 4H), 1.78 [1.75] (s, 3H), 1.29 [1.29] (d, *J* = 6.0 Hz, 3H), 1.21 [1.16] (t, *J* = 6.8 Hz, 3H), 1.26-1.05 [1.26-1.05] (m, 9H), 1.03 [0.98] (d, *J* = 6.4 Hz, 6H); ^{13}C-NMR (100 MHz, $CDCl_3$) δ (mixture of isomers in a 1:1 ratio): 169.1 [169.1], 165.8 [165.8], 161.4 [161.3], 160.2 [160.2], 145.3 [145.2], 143.8 [143.8], 137.6 [137.5], 136.0 [135.1], 133.1 [132.8], 131.6 [131.6], 125.0 [124.7], 123.6 [123.6], 99.3 [97.4], 82.9 [82.9], 70.7 [70.5], 66.2 [66.2], 62.3 [62.2], 61.7 [59.8], 59.5 [59.5], 49.3 [49.3], 47.0 [47.0] (2C), 46.7 [46.6], 45.3 [45.3], 41.4 [41.1], 41.3 [41.3],

36.7 [36.7], 34.9 [34.8], 29.2 [29.2], 25.8 [25.8], 20.4 [20.3], 19.8 [19.8], 18.5 [18.5], 15.4 [15.1], 12.7 [12.6], 11.7 [11.7] (2C), 9.7 [9.7]; ESI(+)MS: *m/z* 787 $[M+Na]^+$.

*4.1.07.3. (1²Z,6R,7R,8E,13E,15E,17S,19R)-17-(1-ethoxyethoxy)-19-hydroxy-6-isopropyl-7,15-dimethyl-3²,3³-dihydro-3¹H-5-oxa-11-aza-1(4,2)-oxazola-3(1,5)-pyrrolacycloicosaphane-8,13,15-triene-2,4,10-trione (**(16(R)-2g**)*

The title compound was obtained from **2f** (34 mg, 0.057 mmol) using the general procedure reported in 4.1.04 section and separated from 16*S* enantiomer *via* PLC (silica gel, DCM/EtOH = 95:5).

Colourless oil; 18.8 mg (56.9%); R_f= 0.78 (DCM/EtOH = 95:5); ^{1}H-NMR (400 MHz, $CDCl_3$) δ (mixture of isomers in a 1:1 ratio): 7.97 [7.96] (s, 1H), 6.92 [6.81] (brt, 1H), 6.59 [6.59] (dd, *J* = 16.4, 6.4 Hz, 1H), 6.20-6.13[6.20-6.13] (m, 1H), 6.03-5.89 [6.03-5.89] (m, *2*H), 5.64-5-54 [5.64-5.54] (m, 1H), 5.16 [4.98] (d, *J* = 9.2 Hz, 1H), 4.94 [4.94] (d, *J* = 10.8 Hz, 1H), 4.76-4-49 [4.76-4.49] (m, 2H), 4.85-4.30[4.85-4.30] (m, 4H), 3.78-3.34 (m, 3H), 3.07-2.99 [3.07-2.99] (m, 1H), 2.09-2.58 (m, 4H), 2.10-1.81 [2.10-1.81] (m, 3H), 1.76 [1.76] (s, 3H), 1.28-1.10 [1.28-1.10] (m, 9H), 0.98 [0.98] (d, *J* = 6.4 Hz, 6H); ^{13}C-NMR (100 MHz, $CDCl_3$) δ (mixture of isomers in a 1:1 ratio): 167.2 [167.2], 161.7 [161.5], 160.7 [160.7], 160.5 [160.4], 143.9 [143.7], 143.6 [143.5], 137.1 [135.9], 135.6 [135.5], 134.9 [134.5], 134.0 [132.1], 131.1 [131.0], 125.1 [124.9], 124.8 [124.7], 123.6 [123.5], 98.5 [97.4], 81.2 [81.1], 70.2 [70.1], 67.8 [67.5], 60.2 [60.2], 50.8 [50.8], 42.3 [42.1], 40.9 [40.6], 37.5 [37.4], 36.6 [36.6], 30.0 [30.0], 29.7 [29.7], 20.4 [20.4], 19.4 [19.4], 18.9 [18.9], 15.4 [15.0], 13.4 [13.2], 11.9 [11.9]; ESI(+)MS: *m/z* 622 $[M+Na]^+$.

*4.1.07.4. (1²Z,6R,7R,8E,13E,15E,17S,19S)-17-(1-ethoxyethoxy)-19-hydroxy-6-isopropyl-7,15-dimethyl-3²,3³-dihydro-3¹H-5-oxa-11-aza-1(4,2)-oxazola-3(1,5)-pyrrolacycloicosaphane-8,13,15-triene-2,4,10-trione (**(16S)-2g**)*

The title compound was obtained from **2f** (34 mg, 0.057 mmol) using the general procedure and separated from 16*R* enantiomer *via* PLC (silica gel, DCM/EtOH = 95:5).

Colourless oil; 14.2 mg (41.6%); R_f= 0.61 (DCM/EtOH = 95:5); ^{1}H-NMR (400 MHz, $CDCl_3$) δ (mixture of isomers in a 1:1 ratio): 8.01 [8.00] (s, 1H), 6.53 [6.53] (bs, 1H), 6.51 [6.51] (dd, *J* = 5.5, 3 Hz, 1H), 6.09 [6.05] (d, *J* = 3.5 Hz, 1H), 6.05 [6.05] (d, *J* = 2.9 Hz, 1H), 5.88 [5.88] (dd, *J* = 16.4, 1.0 Hz, 1H), 5.51 [5.51] (m, 2H), 4.90 [4.90] (dd, *J* = 10.0, 1.9 Hz, 1H), 4.83 [4.83] (m, 1H), 4.77-4.68 [4.77-4.68] (m, 2H), 4.49-4.39 [4.49-4.39] (m, 1H), 4.39-4.28 [4.39-4.28] (m, 1H), 4.28-4.16 [4.28-4.16] (m, 1H), 4.16-4.03 [4.16-4.03] (m, 1H), 3.73-3.62 (m, 1H), 3.57-3.37(m, 6H), 3.31-3.14(m, 2H), 3.12-2.98 (m, 2H), 3.09 [3.04] (t, *J* = 3.07 Hz, 1H), 2.83 [2.79] (dd, *J* = 10.4, 1.9 Hz, 1H), 2.70-2.50 [2.70-2.50] (m, 4H), 2.00-1.90 [2.00-1.90] (m, 3H), 1.66 [1.63] (s, 3H), 1.27-1.12 [1.27-1.12] (m, 6H), 1.14-1.00 [1.14-1.00] (m, 3H), 0.97 [0.97] (t, *J* = 6.4 Hz, 6H); ^{13}C-NMR (100 MHz, $CDCl_3$) δ (mixture of isomers in a 1:1 ratio): 167.4 [167.4], 161.4 [161.4], 161.2 [161.3], 159 [159.6], 144.0 [143.9], 143.1 [143.1], 137.3 [137.3], 136.3 [136.3], 135.9 [135.9], 134.9 [133.0], 132.7 [131.3], 125.6 [125.4], 124.9 [124.9], 123.3 [123.3], 99.1 [99.1], 98.1 [98.1], 80.8 [80.8], 71.0 [70.3], 66.5 [66.4], 61.8 [60.4], 50.3 [50.3], 41.9 [41.7], 41.6 [41.6], 37.2 [37.2], 35.3 [35.2], 30.0 [30.0], 29.8 [29.8], 20.5 [20.4], 19.6 [19.6], 18.7 [18.7], 12.8 [12.8], 11.1 [11.0]; ESI(+)MS: *m/z* 622 $[M+Na]^+$.

4.1.08. General procedure for the synthesis of F-C16 C14-protected compounds (16R)-1h, (16S)-1h, (16R)-2h and (16S)-2h

To a solution of C14-protected alcohol (1 eq., 0.06 mmol scale) in anhydrous DCM (0.2 mL), DAST (2 eq.) was added at 0 °C under nitrogen atmosphere. After stirring for 24 h, the reaction mixture was quenched with saturated $NaHCO_3$ (10 mL) and the aqueous phase was extracted with DCM (3x10 mL). The combined organic phases were washed with saturated $NaHCO_3$, dried over anhydrous Na_2SO_4 and evaporated under vacuum to dryness.

*4.1.08.1. (1²Z,3²S,3³R,6R,7R,8E,13E,15E,17S,19R)-3³-((2-(diethylamino)ethyl)sulfonyl)-17-(1-ethoxyethoxy)-19-fluoro-6-isopropyl-7,15-dimethyl-5-oxa-11-aza-1(4,2)-oxazola-3(1,2)-pyrrolidinacycloicosaphane-8,13,15-triene-2,4,10-trione (**(16R)-1h**)*

The title compound was obtained from **(16*S*)-1g**. The crude product was checked by NMR and used in the next step without further purification.

Brown oil; 15 mg (80.4%); R_f = 0.53 (DCM/EtOH = 95:5); ^{1}H-NMR (400 MHz, $CDCl_3$) δ (mixture of isomers in a 1:1 ratio): 8.12 [8.12] (s, 1H), 6.60 [6.59] (dd, *J* = 16.4, 4.8 Hz,1H), 6.24 [6.24] (brd, *J* = 15.2 Hz, 1H), 6.07-6.01 (m, 1H), 5.82 [5.81] (dd, *J* = 16.0, 2.0 Hz, 1H), 5.75-5.62 (m, 1H), 5.50-5.45 (m, 1H), 5.36 [5.25] (d, *J* = 9.2 Hz, 1H), 5.16-5.09 [5.04-4.99] (m, 1H), 4.87 [4.87] (dd, *J* = 10.4, 2.4 Hz, 1H), 4.80-4.54 (m, 2H), 4.50-4.39 (m, 1H), 4.27-4.18 (m, 1H), 4.08-4.00 (m, 1H), 3.69-3.62 (m, 1H), 3.60-3.40 (m, 4H), 3.30-2.80 (m, 7H), 2.83-2.73 (m, 1H), 2.64-2.46 (m, 1H), 2.58 [2.57] (q, *J* = 7.2 Hz, 4H), 2.01-1.92 (m, 1H), 1.82 [1.80] (s, 3H), 1.70-1.54 (m, 1H), 1.28 [1.27] (d, *J* = 5.2 Hz, 3H), 1.22-1.00 (m, 6H), 1.06 [1.06] (t, *J* = 7.2 Hz, 6H), 0.96 [0.96] (t, *J* = 6.8 Hz, 6H); ^{13}C-NMR (100 MHz, $CDCl_3$) δ (mixture of isomers in a 1:1 ratio): 168.7 [168.7], 165.5 [165.5], 160.2 [160.1], 159.9 [159.6], 145.6 [145.5], 143.7 [143.7], 136.9 [136.1], 136.0 [136.0], 135.6 [135.0], 132.9 [131.9], 125.1 [124.8], 123.8 [123.7], 98.8 [97.2], 88.7 [88.7] ($^1J_{CF}$ = 170.6 Hz), 82.7 [82.6], 68.8 [68.5], 62.4 [62.4], 61.1 [59.5], 58.8 [58.8], 49.1 [49.1], 46.9 [46.9], 46.9 [46.9] (2C), 45.2 [45.2], 41.1 [40.9], 40.5 [40.4] ($^2J_{CF}$ = 25.1 Hz), 36.4 [36.4], 33.7 [33.6] ($^2J_{CF}$ = 25.1 Hz), 29.6 [25.6], 29.2 [29.2], 20.4 [20.3], 19.8 [19.8], 18.4 [18.4], 15.4 [15.2], 13.0 [13.0], 11.6 [11.6] (2C), 10.7 [10.7]; ESI(+)MS: *m/z* 789 $[M+Na]^+$.

*4.1.08.2. (1²Z,3²S,3³R,6R,7R,8E,13E,15E,17S,19S)-3³-((2-(diethylamino)ethyl)sulfonyl)-17-(1-ethoxyethoxy)-19-fluoro-6-isopropyl-7,15-dimethyl-5-oxa-11-aza-1(4,2)-oxazola-3(1,2)-pyrrolidinacycloicosaphane-8,13,15-triene-2,4,10-trione (**(16S)-1h**)*

The title compound was obtained from **(16*R*)-1g**. The crude product was purified *via* PLC (DCM/EtOH = 95:5).

Brown oil; 15 mg (65.1%); R_f = 0.57 (DCM/EtOH = 95:5); ^{1}H-NMR (400 MHz, $CDCl_3$) δ (mixture of isomers in a 1:1 ratio): 8.13 [8.12] (s, 1H), 6.67 [6.67] (dd, *J* = 16.0, 4.4 Hz,1H), 6.26 [6.26] (dd, *J* = 16.0, 2.0 Hz, 1H), 6.14-6.06 (m, 1H), 5.78 [5.78] (dd, *J* = 16.0, 1.2 Hz, 1H), 5.70-5.50 (m, 1H), 5.58 [5.43] (d, *J* = 9.2 Hz, 1H), 5.48 [5.48] (d, *J* = 4.8 Hz, 1H), 5.14-4.95 (brd, 1H), 4.90 [4.90] (d, *J* = 10.4 Hz, 1H), 4.82-4.53 (m, 2H), 4.30-3.95 (m, 3H), 3.79-3.36 (m, 3H), 3.34-3.15 (m, 3H), 3.14-2.90 (m, 3H), 2.84-2.75 (m, 1H), 2.67-2.55 (m, 4H), 2.57-2.44 (m, 2H), 2.12-1.86 (m, 4H), 1.81 [1.77] (s, 3H), 1.32 [1.30] (d, *J* = 5.2 Hz, 3H), 1.26-1.15 (m, 6H), 1.08 [1.08] (t, *J* = 6.8 Hz, 6H), 0.99 [0.97] (d, *J* = 6.2 Hz, 6H); ^{13}C-NMR (100 MHz, $CDCl_3$) δ (mixture of isomers in a 1:1 ratio): 168.3 [168.3], 165.2 [165.2], 160.5

[160.5], 159.6 [159.5] ($^{2}J_{CF}$)=18.8 Hz), 145.9 [145.8], 143.7 [143.7], 137.7 [137.6], 136.2 [136.2]. 134.8 [134.2], 133.1 [132.9], 124.5 [124.3], 123.7 [123.7], 99.1 [97.3], 88.5 [88.3] ($^{1}J_{CF}$=173.6 Hz), 82.7 [82.7], 68.2 [67.4] ($^{3}J_{CF}$= 3.9 Hz), 62.6 [62.6], 61.2 [59.5], 59.3 [59.2], 49.1 [49.1], 46.9 [46.9], 46.9 [46.9] (2C), 45.2 [45.2], 41.3 [41.3], 40.0 [39.8], 36.6 [36.6], 33.2 [33.1] ($^{2}J_{CF}$= 24.2 Hz), 29.4 [29.4], 25.4 [25.4], 20.4 [20.3], 19.8 [19.8], 18.5 [18.5], 15.4 [15.1], 12.7 [12.6], 11.7 [11.7] (2C), 10.3 [10.3]; ESI(+)MS: *m/z* 789 $[M+Na]^{+}$.

*4.1.08.3. (1^{2}Z,6R,7R,8E,13E,15E,17S,19R)-17-(1-ethoxyethoxy)-19-fluoro-6-isopropyl-7,15-dimethyl-3^{2},3^{3}-dihydro-3^{1}H-5-oxa-11-aza-1(4,2)-oxazola-3(1,5)-pyrrolacycloicosaphane-8,13,15-triene-2,4,10-trione (**(16R)-2h**)*

The title compound was obtained from **(16*S*)-1g**. The crude product was checked by NMR and used in the next step without further purification. Brown oil; 77 mg (95%); R_f= 0.32 (DCM/EtOH = 93:7); ^{1}H-NMR (400 MHz, $CDCl_3$) δ (mixture of isomers in a 1:1 ratio): 7.94 [7.92] (s, 1H), 7.18 [7.04] (br t, 1H), 6.63 [6.59] (dd, *J* = 6.6, 1.6 Hz, 1H), 6.18 [6.17] (d, *J* = 2.8 Hz, 1H), 5.98 [5.98] (m, 2H), 5.67 [5.55] (ddd, *J* = 16.0, 6.2, 4.17 Hz, 1H), 5.02 [4.82] (d, *J* = 9.1 Hz, 1H), 4.95 [4.95] (dd, *J* = 9.9, 2.0 Hz, 1H), 4.61-4.44 [4.61-4.44] (m, 3H), 4.30-4.19 [4.30-4.19] (m, 2H), 3.60-3.41 [3.60-3.41] (m, 2H), 3.12-2.96 [3.12-2.96] (m, 2H), 2.86-2.69 [2.86-2.69] (m, 3H), 2.24 [2.24] (m, 1H), 2.00-1.90 [2.00-1.90] (m, 2H), 1.78 [1.78] (s, 3H), 1.80-1.59 [1.80-1.59] (m, 2H), 1.47 [1.47] (t, *J* = 7.3 Hz, 3H), 1.22-1.14 [1.22-1.14] (m, 6H), 0.98 [0.98] (dd, *J* = 6.9, 2.6 Hz, 6H); ^{13}C-NMR (100 MHz, $CDCl_3$) δ (mixture of isomers in a 1:1 ratio): 167.3 [167.3], 160.7 [160.7], 159.6 [159.5], 144.3 [144.0], 143.5 [143.4], 137.0 [137.0], 136.7 [135.6], 135.6 [134.7], 134.2 [133.7], 131.2 [130.1], 125.5 [125.3], 124.9 [124.7], 123.9 [123.8], 98.3 [96.9], 88.7 [88.6] ($^{1}J_{CF}$= 174.0 Hz), 81.2 [81.1], 68.8 [68.7] ($^{3}J_{CF}$= 3.0 Hz), 61.1 [59.8], 51.0 [50.9], 41.1 [40.9] ($^{2}J_{CF}$=23 Hz), 40.5 [40.3], 37.6 [37.5], 34.7 [34.6] ($^{2}J_{CF}$= 29 Hz), 30.1 [30.0], 29.7 [29.7], 20.4 [20.2], 19.5 [19.5], 18.9 [18.9], 15.4 [15.1], 13.4 [13.3], 12.1 [12.1]; ESI(+)MS: 624 $[M+Na]^{+}$.

*4.1.08.4. (1^{2}Z,6R,7R,8E,13E,15E,17S,19S)-17-(1-ethoxyethoxy)-19-fluoro-6-isopropyl-7,15-dimethyl-3^{2},3^{3}-dihydro-3^{1}H-5-oxa-11-aza-1(4,2)-oxazola-3(1,5)-pyrrolacycloicosaphane-8,13,15-triene-2,4,10-trione (**(16S)-2h**)*

The title compound was obtained from (**(16*R*)-1g**. The crude product was checked by NMR and used in the next step without further purification.

Brown oil; 14.2 mg (71%); R_f = 0.46 (DCM/EtOH = 93:7); ^{1}H-NMR (400 MHz, $CDCl_3$) δ (mixture of isomers in a 1:1 ratio): 8.02 [8.00] (s, 1H), 7.12 [7.06] (bt, 1H), 6.60 [6.57] (t, *J* = 6.0 Hz, 1H), 6.13-6.10 [6.13-6.10] (m, 1H), 5.99-5.95 [5.99-5.95] (m, *2*H), 5.56 [5.50] (ddd, *J* = 15.2, 8.4, 3.6 Hz, 1H), 5.23 [5.08] (d, *J* = 8.0 Hz, 1H), 5.20-5.00 [5.20-5.00] (m, 1H), 4.80-4.30 [4.80-4.30] (m, 5H), 3.95-3.75 (m, 2H), 3.69-2.90 (m, 5H), 2.41-2.22 (m, 1H), 2.09-1.54 (m, 4H), 1.69 [1.69] (s, 3H), 1.27 [1.27] (d, *J* = 8.0 Hz, 3H), 1.15-1.06 (m, 6H), 0.98 [0.98] (dd, *J* = 6.9, 3.0 Hz, 6H); ^{13}C-NMR (100 MHz, $CDCl_3$) δ (mixture of isomers in a 1:1 ratio): 167.3 [167.2], 161.0 [161.0], 159.9 [159.8], 159.6 [159.1], 144.6 [144.2], 143.3 [143.1], 137.3 [137.2], 135.8 [135.7], 135.5 [135.3], 133.4 [133.4], 131.7 [130.7], 125.6 [125.5], 125.1 [125.0], 124.9 [124.7], 97.8 [97.1], 87.8 [87.8] ($^{1}J_{CF}$ = 178 Hz), 80.8 [80.8], 67.6 [67.3] ($^{3}J_{CF}$ = 4.2 Hz), 60.3 [59.7], 59.6 [59.5], 42.1 [42.1], 42.0 [41.4] ($^{2}J_{CF}$ = 23 Hz), 37.1 [36.5], 34.4 [34.0] ($^{2}J_{CF}$ = 29 Hz), 29.9 [29.9], 29.6 [29.6], 20.2 [20.1], 19.5 [18.9], 15.4 [15.2], 15.6 [13.8], 12.9 [12.1], 10.9 [10.9]; ESI(+)MS: *m/z* 624[M+Na]$^{+}$.

4.1.09. General procedure for the synthesis of F-C16 compounds (16R)-1e, (16S)-1e, (16R)-2e and (16S)-2e

C-14 protected alcohols (1 eq., 0.1 mmol scale) were dissolved in 10 mL of MeOH or in 10 mL 10% acetic acid/THF 9:1 aqueous solution and PPTS (1 eq.) was added. The reaction was stirred at room temperature for 24 hours. The solvent was then evaporated under vacuum and the residue was resuspended in DCM (5 mL). The organic solution was washed with water (5 mL) and brine (5 mL), dried over anhydrous Na_2SO_4 and concentrated under low pressure.

*4.1.09.1. (1^{2}Z,3^{2}S,3^{3}R,6R,7R,8E,13E,15E,17S,19R)-3^{3}-((2-(diethylamino)ethyl)sulfonyl)-19-fluoro-17-hydroxy-6-isopropyl-7,15-dimethyl-5-oxa-11-aza-1(4,2)-oxazola-3(1,2)-pyrrolidinacycloicosaphane-8,13,15-triene-2,4,10-trione (**(16R)-1e**)*

The title compound was obtained from **(16*R*)-1h** solubilized in MeOH. The crude product was purified *via* PLC (DCM/EtOH = 95:5).

White solid; 4 mg (63,1%); R_f= 0.65 (DCM/EtOH = 95:5); $[\alpha]_D$=−68° (*c* 0.10, DCM); ^{1}H-NMR (400 MHz, $CDCl_3$) δ: 8.14 (s, 1H), 6.66 (dd, *J* = 16.0, 4.8 Hz, 1H), 6.23 (d, *J* = 15.9 Hz, 1H), 6.04-5.98 (m, 1H), 5.85 (dd, *J* = 16.1, 1.8 Hz, 1H), 5.73 (ddd, *J* = 15.8, 8.10, 4.1 Hz, 1H), 5.47 (d, *J* = 1.5 Hz, 1H), 5.42 (d, *J* = 8.9 Hz, 1H), 5.17-4.98 (m, 1H), 4.90 (dd, *J* = 10.4, 2.1 Hz, 1H), 4.82 (ddd, *J* = 13.8, 9.2, 4.6 Hz, 1H), 4.46-4.38 (m, 1H), 4.27-4.19 (m, 1H), 4.08-4.00 (m, 1H), 3.67-3.57 (m, 2H), 3.34-2.92 (m, 7H), 2.83-2.76 (m, 1H), 2.64-2.57 (m, 5H), 2.23-2.12 (m, 1H), 2.06-1.93 (m, 2H), 1.84 (s, 3H), 1.13 (d, *J* = 6.9 Hz, 3H), 1.08 (t, *J* = 7.2 Hz, 6H), 0.98 (t, *J* = 7.3 Hz, 6H); ^{13}C-NMR (100 MHz, $CDCl_3$) δ: 168.7, 160.5, 160.2, 159.8 ($^3J_{CF}$ = 11.5 Hz), 145.8, 143.8, 136.2, 135.8, 135.6, 133.3, 125.4, 123.8, 89.1 ($^1J_{CF}$= 173 Hz), 82.7, 65.8, 62.5, 58.8, 53.4, 48.8, 46.9 (2C), 45.2, 42.2 ($^2J_{CF}$= 21 Hz), 40.8, 36.5, 33.5 ($^2J_{CF}$= 26.2 Hz), 29.7, 29.3, 25.9, 19.7, 13.0, 11.4 (2C), 10.6; FT-IR: br 2964, 2919, 2874, 1747, 1670, 1621, 1425, 1259, 1189, 1134, 1053, 968, 812, 745 cm^{-1}; HPLC: 11.173 min (HPLC method B); ESI(+)MS: *m/z* 695 $[M+H]^+$, 717 $[M+Na]^+$; HR-ESI(+)MS: calcd for $C_{34}H_{51}FN_4O_8S$ *m/z* 695.3484; found *m/z* 695.3478.

*4.1.09.2. (1²Z,3²S,3³R,6R,7R,8E,13E,15E,17S,19S)-3³-((2-(diethylamino)ethyl)sulfonyl)-19-fluoro-17-hydroxy-6-isopropyl-7,15-dimethyl-5-oxa-11-aza-1(4,2)-oxazola-3(1,2)-pyrrolidinacycloicosaphane-8,13,15-triene-2,4,10-trione (**(16S)-1e**)*

The title compound was obtained from **(16*S*)-1h** solubilized in MeOH. The crude product was purified *via* PLC (DCM/EtOH = 95:5).

White powder, 7 mg (77.2%); R_f= 0.67 (DCM/EtOH = 95:5); $[\alpha]_D$=−126° (*c* 0.10, DCM); ^{1}H-NMR (400 MHz, $CDCl_3$) δ: 8.15 (s, 1H), 6.67 (dd, *J* = 16.0, 4.3 Hz, 1H), 6.26 (d, *J* = 15.6 Hz, 1H), 6.06 (brs, 1H), 5.80 (dd, *J* = 15.9, 1.8 Hz, 1H), 5.72-5.62 (m, 2H), 5.30-5.12 (m, 1H), 5.47 (s, 1H), 4.88 (dd, *J* = 10.5, 2.1 Hz, 1H), 4.86-4.82 (m, 1H), 4.27-4.18 (m, 1H), 4.16-4.08 (m, 1H), 4.02-3.94 (m, 1H), 3.70-3.62 (m, 1H), 3.60 (d, *J* = 7.6 Hz, 1H), 3.36-3.04 (m, 7H), 2.85-2.77 (m, 1H), 2.64-2.45 (m, 5H), 2.42-2.29 (m, 1H), 2.16-1.92 (m, 3H), 1.79 (s, 3H), 1.10 (t, *J* = 7.4 Hz, 9H), 0.98 (t, *J* = 6.2 Hz, 6H); ^{13}C-NMR (100 MHz, $CDCl_3$) δ: 167.6, 165.1, 160.2, 159.4 ($^3J_{CF}$= 10 Hz), 146.0, 143.9, 137.2, 134.0, 133.4, 129.2, 125.6, 123.7, 88.5 ($^1J_{CF}$= 163 Hz), 82.5, 65.3, 62.9, 58.2, 49.7, 47.2 (2C), 46.8, 45.6, 41.3 ($^2J_{CF}$= 17 Hz), 41.0, 36.7, 33.3 ($^2J_{CF}$= 25 Hz), 29.7, 29.3, 26.4, 19.8, 12.7, 10.1, 9.8 (2C); FT-IR: br 3328, 2967, 2923, 1748, 1670, 1622, 1582, 1544, 1425, 1294, 1227, 1191, 1125, 1049, 987, 813 cm^{-1}; HPLC: 13.501

min (HPLC method B); ESI(+)MS: *m/z* 695 $[M+H]^+$, 717 $[M+Na]^+$; HR-ESI(+)MS: calcd for $C_{34}H_{51}FN_4O_8S$ *m/z* 695.3484; found *m/z* 695.3476.

*4.1.09.3. (1²Z,6R,7R,8E,13E,15E,17S,19R)-19-fluoro-17-hydroxy-6-isopropyl-7,15-dimethyl-3²,3³-dihydro-3¹H-5-oxa-11-aza-1(4,2)-oxazola-3(1,5)-pyrrolacycloicosaphane-8,13,15-triene-2,4,10-trione (**(16R)-2e**)*

The title compound was obtained from **(16*R*)-2h** solubilized in 10% acetic acid/THF 9:1. The crude product was purified *via* preparative RP-HPLC with an eluent solution of acetonitrile/water = 1:1.

Ochre powder; 60 mg (88.2%); R_f= 0.40 (DCM/EtOH = 95:5); $[\alpha]_D$=−78° (*c* 0.06, DCM); ^{1}H-NMR (400 MHz, $CDCl_3$) δ: 7.94 (s, 1H), 7.09 (brt, 1H), 6.61 (dd, *J* = 16.3, 7.0 Hz, 1H), 6.18 (t, *J* = 3.0 Hz, 1H), 5.99 (dd, *J* = 16.3, 1.1 Hz, 1H), 5.97 (brd, *J* = 15.8 Hz, 1H), 5.69 (ddd, *J* = 15.9, 5.9, 4.2 Hz, 1H), 4.95 (brdd, *J* = 9.9, 2.0 Hz, 2H), 4.81-4.59 (m, 2H), 4.36-4.25 (m, 1H), 4.23-4.09 (m, 2H), 3.98 (brd, *J* = 16.4 Hz, 1H), 3.25 (dt, *J* = 14.5, 3.9 Hz, 1H), 3.12-3.01 (m, 1H), 2.90-2.79 (m, 1H), 2.78-2.65 (m, 2H), 2.38-2.20 (m, 1H), 2.00-1.86 (m, 2H), 1.78 (s, 3H), 1.13 (d, *J* = 7.0 Hz, 3H), 0.99 (dd, *J* = 6.7, 2.8 Hz, 6H); ^{13}C-NMR (100 MHz, $CDCl_3$) δ: 167.3, 160.8, 160.5, 159.5 ($^3J_{CF}$= 15.6 Hz), 144.3, 143.5, 137.2, 135.7, 135.4, 133.9, 131.6, 125.9, 124.9, 123.6, 88.7 ($^1J_{CF}$= 173 Hz), 81.3, 66.2 ($^3J_{CF}$= 3 Hz), 50.9, 42.5 ($^2J_{CF}$= 22.5 Hz), 40.4, 37.5, 34.6 ($^2J_{CF}$= 28 Hz), 30.1, 29.8, 19.6, 18.9, 13.3, 12.0; HPLC: 8.241 min (HPLC method A); ESI(+)MS: *m/z* 552 $[M+Na]^+$, 534 $[M+Na-H_2O]^+$; HR-ESI(+)MS: calcd for $C_{28}H_{36}FN_3O_6Na$ *m/z* 552.2481; found *m/z* 552.2487.

*4.1.09.4. (1²Z,6R,7R,8E,13E,15E,17S,19S)-19-fluoro-17-hydroxy-6-isopropyl-7,15-dimethyl-3²,3³-dihydro-3¹H-5-oxa-11-aza-1(4,2)-oxazola-3(1,5)-pyrrolacycloicosaphane-8,13,15-triene-2,4,10-trione (**(16S)-2e**)*

The title compound was obtained from **(16*S*)-2h** solubilized in 10% acetic acid/THF 9:1. The crude product was purified *via* preparative RP-HPLC with an eluent solution of acetonitrile/water = 1:1.

Ochre powder; 15 mg (70%); R_f= 0.85 (DCM/EtOH = 95:5, $[\alpha]_D$=−58° (*c* 0.10, DCM); ^{1}H-NMR (400 MHz, $CDCl_3$) δ: 8.02 (s, 1H), 6.93 (brt, 1H), 6.57 (dd, *J* = 16.4, 6.4 Hz, 1H), 6.10 (s, 1H), 5.97 (s, 1H), 5.93 (d, *J* = 4.0 Hz, 1H), 5.55 (ddd, *J* = 13.6, 9.2, 4.0 Hz, 1H), 5.25 (d, *J* = 8.0 Hz,

1H), 5.16-5.08 (m, 1H), 4.93 (d, *J* = 9.6 Hz, 1H), 4.80-4.71 (m, 1H), 4.48-4.36 (m, *2*H), 3.58-3.47 (m, 1H), 3.28 (dt, *J* = 14.8, 3.2 Hz, 1H), 3.16-3.02 (m, 1H), 2.87-2.64 (m, 3H), 2.35-2.19 (m, 1H), 2.08-1.05 (m, 2H), 1.77-1.69 (m, 1H), 1.65 (s, 3H), 1.12 (d, *J* = 6.8 Hz, 3H), 0.97 (t, *J* = 6.4 Hz, 6H); ^{13}C-NMR (100 MHz, $CDCl_3$) δ: 167.3, 161.2, 159.4($^3J_{CF}$ = 5.6 Hz), 159.4, 144.5, 143.3, 137.2, 135.9, 135.5, 134.4, 131.9, 126.0, 124.8, 123.1, 87.8 ($^1J_{CF}$ = 172.0 Hz), 81.0, 65.3 ($^3J_{CF}$ = 3 Hz), 50.3, 41.4, 41.2($^2J_{CF}$ = 20.4 Hz), 37.2, 34.3 ($^2J_{CF}$ = 29.3 Hz), 30.0, 29.9, 19.6, 18.9, 13.0, 11.9; FT-IR: 2959 (br), 2918, 2850, 1732, 1669, 1616, 1417, 1179, 1163, 1054, 968, 745 cm^{-1}; HPLC: 8.029 min (HPLC method A); ESI(+)MS: *m/z* 552 $[M+Na]^+$; HR-ESI(+)MS: calcd for $C_{28}H_{37}FN_3O_6$ *m/z* 530.2661; found *m/z* 530.2657.

4.1.10. General procedure for the synthesis of trienones 1i and 2i

Propylisocyanate (3 eq.) was added at room temperature to a stirring solution of **1** or **2** (0.02 mmol scale) in anhydrous DCM (500 µl). Successively, pyridine (3 eq.), and dimethylaminopyridine (DMAP, 3 eq.) were added and the solution was stirred overnight. 0.5 mL of MeOH were added and the mixture was treated with 5 mL of saturated NaCl solution containing 0.1 mL of conc. HCl and extracted with DCM (3x5 mL). The organic phases were washed with brine, dried over anhydrous Na_2SO_4 and evaporated under reduced pressure.

*4.1.10.1. (1²Z,3²S,3³R,6R,7R,8E,13E,15E,17E)-3³-((2-(diethylamino)ethyl)sulfonyl)-6-isopropyl-7,15-dimethyl-5-oxa-11-aza-1(4,2)-oxazola-3(1,2)-pyrrolidinacycloicosaphane-8,13,15,17-tetraene-2,4,10,19-tetraone (**1i**)*

The title compound was obtained from **1** using the general procedure G. The crude was purified *via* PLC (DCM/EtOH = 92:8).

Yellow powder; 8.5 mg (58.1%); R_f = 0.75 (DCM/EtOH = 90:10; $[\alpha]_D$ = −38° (*c* 0.31, DCM); ^{1}H-NMR (400 MHz, $CDCl_3$) δ: 8.18 (s, 1H), 7.59-7.50 (m, 1H), 6.78 (dd, *J* = 15.9, 5.6 Hz, 1H), 6.41 (d, *J* = 14.9 Hz, 1H), 6.36 (d, *J* = 14.9 Hz, 1H), 6.22 (d, *J* = 11.2 Hz, 1H), 6.01-5.89 (m, 2H), 5.41 (s, 1H), 4.88 (dd, *J* = 10.6, 1.8 Hz, 1H), 4.33-4.28 (m, 1H), 4.25-4.17 (m, 1H), 4.10-4.02 (m, 1H), 3.93 (d, *J* = 8.2 Hz, 2H), 3.89-3.79 (m, 2H), 3.67 (d, *J* = 8.0 Hz, 1H), 3.47-3.08 (m, 5H), 2.78-2.65 (m, 5H), 2.32-2.23 (m, 2H), 1.99 (s, 3H), 1.24-1.20 (m, 3H), 1.15-1.10 (d, *J* = 6.3 Hz, 3H), 1.00-0.85 (m, 9H); ^{13}C-NMR (100 MHz, $CDCl_3$) δ: 192.7, 168.1, 165.1, 160.1,

158.1, 147.2, 144.4, 144.1, 138.8, 136.6, 129.1, 128.9, 128.4, 123.4, 82.9, 63.1, 47.5, 47.0 (2C), 45.0, 42.4, 41.8, 40.3, 36.7, 31.9, 29.3, 23.4, 22.6, 19.6, 14.1, 13.7, 11.3 (2C), 10.2; FT-IR: 2959, 2920, 2890, 1742, 1625, 1575, 1423, 1189, 1136, 971, 892, 811, 749 cm^{-1}; HPLC: 7.078 min (HPLC method B); ESI(+)MS: 673 $[M+H]^+$; HR-ESI: calcd for $C_{34}H_{49}N_4O_8S$ *m/z* 673.3266; found *m/z* 673.3279.

*4.1.10.2. (1²Z,6R,7R,8E,13E,15E,17E)-6-isopropyl-7,15-dimethyl-3²,3³-dihydro-3¹H-5-oxa-11-aza-1(4,2)-oxazola-3(1,5)-pyrrolacycloicosaphane-8,13,15,17-tetraene-2,4,10,19-tetraone (**2i**)*

The title compound was obtained from **2**. The crude product was purified *via* PLC (DCM/EtOH = 92:8).

Yellow powder; 6.5 mg (44.8%); R_f = 0.55 (DCM/EtOH = 95:5); $[\alpha]_D$ = −46° (*c* 0.44, DCM); ^{1}H-NMR (400 MHz, $CDCl_3$) δ: 8.03 (s, 1H), 7.83-7.74 (m, 1H), 7.39-7.33 (m, 1H), 6.76 (d, *J* = 15.3 Hz, 1H), 6.49 (dd, *J* = 16.7, 7.2 Hz, 1H), 6.30 (brs, 1H), 6.10-5.89 (m, 4H), 4.90 (d, *J* = 9.7 Hz, 1H), 4.40-4.05 (m, 4H), 4.00 (d, *J* = 14.8 Hz, 1H), 3.82-3.74 (m, 1H), 2.94-2.81 (m, 2H), 2.71-2.59 (m, 2H), 1.95 (s, 3H), 1.10 (d, *J* = 6.7 Hz, 3H), 0.98-0.86 (m, 6H); ^{13}C-NMR (100 MHz, $CDCl_3$) δ: 192.8, 167.6, 160.3, 160.1, 159.9, 146.3, 143.0, 142.9, 142.4, 136.9, 136.3, 131.6, 129.2, 126.8, 126.3, 125.4, 125.2, 81.2, 51.7, 41.8, 47.3, 37.4, 29.9, 29.8, 21.3, 19.5, 18.7, 11.9; FT-IR: 3317 (br), 2966, 2924, 2852, 1729, 1665, 1616, 1530, 1417, 1315, 1223, 1074, 980 cm^{-1}; HPLC: 9.359 min (HPLC method A); ESI(+)MS: *m/z* 508 $[M+H]^+$; HR-ESI(+)MS: calcd for $C_{28}H_{34}N_3O_6$ *m/z* 508.2442; found *m/z* 508.2433.

4.1.11 General procedure for the synthesis of C26-derivatives 5-7

To a solution of **2** or **(16*R*)-2c** (0.1 mmol scale) in MeOH (2 mL), the suitable nucleophile (2-6 eq.) was added at 0 °C and the reaction was stirred at RT for 48-72 h. The solvent was evaporated under reduced pressure and the crude reaction mixture was purified *via* PLC.

*4.1.11.1. (1²Z,3²S,3³R,6R,7R,8E,13E,15E,17S)-3³-((2-(diethylamino)ethyl)thio)-17-hydroxy-6-isopropyl-7,15-dimethyl-5-oxa-11-aza-1(4,2)-oxazola-3(1,2)-pyrrolidinacycloicosaphane-8,13,15-triene-2,4,10,19-tetraone (**5**)*

The title compound was obtained from **2** with 2 eq. of 2-diethylaminoethanethiol hydrochloride as nucleophile and 2 eq. of TEA. The crude product was purified *via* PLC (DCM/EtOH = 80:20).

White powder; 2.3 mg (32%); R_f= 0.30 (DCM/MeOH = 90:10 + 0.1% TEA); $[\alpha]_D$=−44° (*c* 0.18, DCM); ^{1}H-NMR (400 MHz, $CDCl_3$) δ: 8.14 (s, 1H), 6.65 (dd, *J* = 16.7, 4.4 Hz, 1H), 6.18 (d, *J* = 15.8 Hz, 1H), 5.88 (d, *J* = 16.4 Hz, 1H), 5.67 (ddd, *J* = 15.4, 9.1, 4.3 Hz, 1H), 5.50 (d, *J* = 8.5 Hz, 1H), 4.97-4.92 (m, 1H), 4.80 (dd, *J* = 10.1, 1.5 Hz, 2H), 4.71 (s, 1H), 4.42-4.29 (m,1H), 4.16-4.03 (m, 3H), 3.91-3.85 (m, 1H), 3.82 (s, 2H), 3.56-3.44 (m,1H), 3.40 (d, *J* = 6.0 Hz, 1H), 3.15 (dd, *J* = 18.0, 4.4 Hz, 1H), 2.90 (dd, *J* = 18.0, 4.4 Hz, 1H), 2.80-2.60 (m, 9H), 2.18-1.97 (m, 5H), 1.72 (s, 3H), 1.09 (d, *J* = 6.6 Hz, 6H), 0.97 (t, *J* = 6.6 Hz, 6H); ^{13}C-NMR (100 MHz, $CDCl_3$) δ: 202.5, 170.1, 166.1, 160.4, 157.0, 144.8, 144.2, 136.9, 136.3, 134.2, 132.8, 125.3, 124.0, 81.9, 66.1, 65.4, 65.0, 52.3, 49.0, 47.0, 46.9 (2C), 45.2, 42.9, 40.9, 40.5, 36.6, 31.8, 29.4, 19.7, 18.6, 12.7 (2C), 10.2; FT-IR: 3330 (br), 2969, 2925, 2825, 1740, 1670, 1621, 1579, 1424, 1367, 1354, 1217, 1176, 1107, 961 cm^{-1}; HPLC: 2.251 min (HPLC method A); ESI(+)MS: *m/z* 659 $[M+H]^+$, 681$[M+Na]^+$, 641 $[M-H_2O]^+$; HR-ESI(+)MS: calcd for $C_{34}H_{51}N_4O_7S$ *m/z* 659.3473; found *m/z* 659.3467.

*4.1.11.2. (1²Z,6R,7R,8E,13E,15E,17S,19R)-3³-((2-(diethylamino)ethyl)amino)-17,19-dihydroxy-6-isopropyl-7,15-dimethyl-5-oxa-11-aza-1(4,2)-oxazola-3(1,2)-pyrrolidinacycloicosaphane-8,13,15-triene-2,4,10-trione (**6**)*

The title compound was obtained from **(16*R*)-2c** with 6 eq. of *N*,*N*-diethylethylenediamine as the nucleophile. The crude product was purified *via* PLC (DCM/EtOH = 80:20 + 1% TEA).

White solid; 9.3 mg (54%); R_f= 0.10 (DCM/EtOH = 90:10 + 1% TEA); $[\alpha]_D$= −32° (*c* 0.10, DCM); ^{1}H-NMR (400 MHz, $CDCl_3$) δ: 8.13 (s, 1H), 6.61 (dd, *J* = 15.9, 4.3 Hz, 1H), 6,51 (m, 1H), 6.20 (d, *J* = 15.6 Hz, 1H), 5.87 (dd, *J* = 16.0, 1.7 Hz, 1H), 5.67 (ddd, *J* = 15.6, 9.4, 4.3 Hz, 1H), 5.46 (d, *J* = 9.4 Hz, 1H), 4.87-4.79 (m, 2H), 4.75 (brs, 1H), 4.94-4.24 (m, 1H), 4.22-4.12 (m, 1H), 4.05-3.99 (m, 2H), 3.35 (s, 1H), 3.03 (dd, *J* = 16.8, 5.3 Hz, 1H), 2.88-2.72 (m, 13H), 1.99-1.88 (m, 2H), 1.87-1.81 (m, 2H), 1.79 (s, 3H), 1.18 (t, *J* = 7.1 Hz, 6H), 1.10 (d, *J* = 6.8 Hz, 3H), 0.99-0.94 (m, 6H); ^{13}C-NMR (100 MHz, $CDCl_3$) δ: 170.4, 167.0, 161.7, 160.7, 145.2,

143.6, 136.8, 136.4, 135.0, 133.6, 124.7, 123.9, 81.4, 68.1, 67.2, 64.5, 60.2, 52.2, 46.9 (2C), 45.2, 43.2, 41.2, 39.6, 36.5, 35.8, 31.4, 29.4, 19.8, 18.6, 13.0, 11.5 (2C), 10.3; FT-IR: 3291 (br), 2918, 1741, 1669, 1614, 1581, 1432, 1177, 1090, 970, 892, 753, 603 cm^{-1}; HPLC: 2.844 min (HPLC method A); ESI(+)MS: *m/z* 644 $[M+H]^+$, 666 $[M+Na]^+$, 626 $[M+H-H_2O]^+$; HR-ESI(+)MS: calcd for $C_{34}H_{54}N_5O_7$ *m/z* 644.4018; found *m/z* 644.4029.

*4.1.11.3. (1²Z,6R,7R,8E,13E,15E,17S,19R)-3³-(2-(diethylamino)ethoxy)-17,19-dihydroxy-6-isopropyl-7,15-dimethyl-5-oxa-11-aza-1(4,2)-oxazola-3(1,2)-pyrrolidinacycloicosaphane-8,13,15-triene-2,4,10-trione (**7**)*

The title compound was obtained from **(16*R*)-2c** with 3 eq. of nucleophile prepared *in-situ* from 2-(diethylamino)ethanol. Alkoxide was produced by adding 1 eq. of NaH in dry THF at 0 °C for 1 h. The crude product was purified *via* PLC (DCM/MeOH = 90:10 + 1% TEA).

White powder; major isomer: 4 mg (19%); R_f = 0.39 (DCM/EtOH = 90:10 + 1% TEA); $[\alpha]_D$ = −36° (*c* 0.13, DCM); ^{1}H-NMR (400 MHz, $CDCl_3$) δ: 8.14 (s, 1H), 6.68 (ddd, *J* = 16.3, 3.7 Hz, 1H), 6.24 (d, *J* = 15.6 Hz, 1H), 5.97 (d, *J* = 16.3 Hz, 1H), 5.70 (ddd, *J* = 14.9, 9.7, 6.3 Hz, 1H), 5.50 (d, *J* = 8.6 Hz, 1H), 4.91 (s, 1H), 4.87 (d, *J* = 9.8 Hz, 1H), 4.38-4.30 (m, 1H), 4.19-3.83 (m, 7H), 3.10-3.72 (m, 11H), 2.09-1.92 (m, 5H), 1.80 (s, 3H), 1.32-1.27 (m, 6H), 1.13 (d, *J* = 6.8 Hz, 3H), 0.97 (t, *J* = 6.5 Hz, 6H); ^{13}C-NMR (100 MHz, $CDCl_3$) δ: 169.9, 165.8, 161.7, 160.8, 145.3, 143.6, 136.7, 136.4, 134.3, 134.3, 134.2, 125.2, 123.9, 81.4, 68.3, 67.7, 64.3, 60.0, 52.0, 46.9, 46.8, 43.0, 41.1, 36.6, 35.6, 31.4, 29.7, 29.4, 19.8, 18.6, 13.1, 10.3, 10.1; FT-IR: 3291 (br), 2918, 1741, 1614, 1432, 1177, 1109, 970, 892, 753, 674, 603 cm^{-1}; HPLC: 1.935 min (HPLC method A); ESI(+)MS: *m/z* 645 $[M+H]^+$, 667 $[M+Na]^+$, 627 $[M-H_2O]^+$; HR-ESI(+)MS: calcd for $C_{34}H_{53}N_4O_8$ *m/z* 645.3858; found *m/z* 645.3865.

4.2. Molecular Modelling

The evaluation of the putative binding mode of **1** and **2** derivatives within the mitochondrial ribosome streptogramin A binding site was conducted using molecular docking calculations. Each molecular modelling operation was performed on an 8 core Intel Xeon® CPU E5-1620 CPU, running Ubuntu 16.04 as its operative system.

4.2.1. Structure preparation

The structure of the mitochondrial ribosome in complex with VM1 (**2**) (PDB ID: 6I9R) was retrieved from the Protein Data Bank and prepared for further calculations exploiting various tools implemented in the Molecular Operating Environment (MOE) 2019.01 suite[36]. At first, minor structural issues were addressed with the Structure Preparation tool: ACE and NME capping residues were added to protein chains with missing residues at the N and C terminal, while structural gaps in the central part of protein chains were addressed using the Loop Modeller tool. Secondly, missing hydrogen atoms were added to the system according to the most probable protonation state at pH 7.4 with the "Protonate 3D" and their coordinates were then energetically minimized according to the AMBER10: EHT force field. The co-crystallized VM1 (**2**) and Q (**3**), as well as magnesium ions, were removed from the final structure used for docking calculations. The centre of mass of VM1 (**2**) was used to define the location of the streptogramin A site in subsequent calculations.

4.2.2. Ligand preparation

Each ligand evaluated in this scientific work was prepared for docking calculations as follows, making use of several tools from the OpenEye suite[37]. At first, the most probable tautomeric state at pH 7 was determined using the "tautomers" program. Then, the 3D coordinates were generated by exploiting the macrocycle protocol of the OMEGA package. Finally, the most probable protonation state at pH 7 was determined using the "fixpka" tool, while partial charges were attributed using the "molcharge" tool according to the MMFF94 force field.

4.2.3. Docking calculations

Docking calculations were performed using the PLANTS[38–40] software, which is free for academic use. The docking was centred on the streptogramin A binding site of the mitochondrial ribosome, defined as a sphere with a 12 Å radius wrapped around the centre of mass of the crystal VM1 (see section 4.2.1). The best ten poses for each compound were retained. Afterwards, the van der Waals and electrostatic interaction energy values were calculated for each pose using MOE, filtering out those poses with

repulsive interaction energy values. The best scoring pose for each compound was then stored for further analysis.

4.2.4. Results analysis

Docking poses, obtained as described in section 4.2.3, were characterised and compared to the binding mode of the reference compound (VM1) making use of an in-house Python script. First, a per-residue decomposition of the electrostatic and hydrophobic contribution to the ligand-receptor was calculated using a SVL script within MOE. Afterwards, values coming from this analysis were used for the generation of an interaction energy map, which depicts the strength of interaction between the most important receptor residues and the ligand. Values are reported on the map according to a divergent colourimetric scale: for electrostatic interaction, values are reported in Kcal/mol with a colour palette ranging from black (indicating an attractive interaction) to red (indicating a repulsive interaction), while for hydrophobic interaction values are reported in arbitrary units with a colour palette ranging from white (indicating the absence of interaction) to dark teal (indicating a strong interaction). The same Python script was also used for the generation of a video, which reports each docking pose associated with its interaction energy profile in comparison with the ones of structurally related compounds and with the crystal reference.

4.3. Biology

4.3.1. Cell culture

Human glioblastoma stem cell lines COMI and VIPI were a kind gift from Antonio Daga (Azienda Ospedaliera Universitaria San Martino di Genova, Italy). The cell lines were cultured in DMEM/F-12 and Neurobasal media (Thermo Fisher Scientific, 1:1 ratio), supplemented with GlutaMAX (2 mM; Thermo Fisher Scientific), B27 supplement (1%; Thermo Fisher Scientific), Penicillin G (100 U/mL; Sigma Aldrich), recombinant human fibroblast growth factor-2 (bFGF) (10 ng/mL; R&D Systems), recombinant human epidermal growth factor (EGF) (20 ng/mL; R&D Systems) and heparin (2 µg/mL; Sigma Aldrich) at 37 °C, 5% CO_2 on laminin-coated flasks where they maintain intact self-renewal capacity[41].

4.3.2. Viability assays

For the evaluation of the effect of D (**1**) and VM1 (**2**) derivatives on GSCs viability, COMI and VIPI cells grown in culture flasks were detached using Accutase. Cells were then counted using a trypan blue staining and an automated cell counter and seeded in 96-well laminin-coated microtiter plates at a density of 3,500-4,000 cells/well in a final media volume of 100 µL. The plates were incubated for 24 hours before drug treatment. Serial drug dilutions were prepared in phosphate-buffered saline (PBS) to provide a total of seven drug concentrations plus control. 5 µL of these dilutions were added to each well, and the plates were incubated for 48 hours. Each treatment was performed in technical triplicate. After the drug treatment, the cells were stained with Hoechst 33342 (1 µg/mL; Thermo Fisher Scientific, cat. H1399) and propidium iodide (1 µg/mL; Sigma Aldrich, cat. P4170) and incubated for 1 hour at 37 °C, 5% CO_2. The fluorescent signal was then measured using the Operetta-High Content Imaging System (Perkin Elmer) and analysed using the Harmony Software. The number of viable cells was calculated by subtracting PI-positive cells from the total number of cells estimated by Hoechst 33342 staining and normalised to the control.

4.3.3. Mitochondrial and cytosolic protein synthesis assay

COMI cells grown in culture flasks were detached using Accutase. Cells were then counted using trypan blue staining and an automated cell counter and seeded in a 6-well plate at a density of 180,000 cells/well. The plate was incubated for 24 hours at 37 °C, 5% CO_2 and then treated with Q/D for further 24 h before ^{35}S labelling. To assay for mitochondrial protein synthesis, the growth medium was removed and cells were washed twice with methionine/cysteine-free DMEM medium, followed by incubation in methionine/cysteine-free DMEM medium containing 96 µg/mL Cysteine, 1% B27 supplement, 1% GlutaMax; 1% Sodium Pyruvate, 10 ng/mL bFGF, 20 ng/mL EGF, 2 µg/mL heparin and 80 µg/mL emetine (Sigma Aldrich, cat. E2375) for 15 min at 37 °C. Emetine was used to inhibit cytosolic translation. Subsequently, [^{35}S]-methionine (Perkin Elmer, cat. NEG709A005MC) was added to a final concentration of 166.6 µCi/mL and the labelling was performed for 20 min at 37 °C. The cells were then detached and centrifuged at 4,000 rpm for 5 min. The pellet obtained was washed three times with 1 mL of PBS. Cell pellets

were resuspended in protein lysis buffer containing protease inhibitors and 1.25 U/µL benzonase. Protein concentrations were measured with Pierce™ BCA Protein Assay Kit (Thermo Fisher Scientific, cat. 23227), as per the manufacturer's instructions, and equal amounts of protein samples were separated on SDS-PAGE gels (NuPAGE™ 12% Bis-Tris Protein Gels, Thermo Fisher Scientific, cat. NP0343BOX). The labelled proteins were visualised and quantified using a PhosphorImager system and ImageQuant software (Molecular Dynamics, GE Healthcare).

4.3.4. Immunoblotting

COMI cells grown in culture flasks were detached using Accutase. Cells were then counted using a trypan blue staining and an automated cell counter and seeded in a 6-well plate at a density of 180,000 cells per well in a final media volume of 2 mL. The plates were incubated for 24 hours before drug treatment. A 1 mM solution of the compound was prepared in PBS starting from the 10 mM stock. From this solution, cells were treated with the right volume to obtain 4 different concentrations of compound plus control. The plates were incubated for 48 hours. After the treatment, cells were washed once with cold PBS, detached using a cell scraper and harvested in 1.5 mL tubes. Cells were centrifuged at 5,000 rpm for 5 minutes to obtain a pellet and the supernatant was discarded. Protein extraction was performed using RIPA lysis buffer (Merck Millipore, cat. 20-188) supplemented with protease inhibitors. The protein content of the lysates was quantified using the Pierce BCA assay (Thermo Fisher Scientific), as per the manufacturer's instructions. Equal amounts of protein were separated on SDS-PAGE and transferred to a PVDF membrane. Membranes were blocked with 5% nonfat dry milk prepared in TBS-Tween and probed with anti-MTCO1 (COX1, Abcam, cat. ab14705), anti-COX4 (Cell Signaling, cat. 4850), anti-MTCO2 (COX2, Abcam, cat. ab110258), anti-ND3 (ND3, Abcam, cat. ab192306), anti-β-tubulin (Santa Cruz, cat. sc-53140) overnight at 4 °C. After three washes with TBS-Tween, they were incubated with secondary HRP-conjugated antibodies (Santa Cruz Biotechnology) for 1 hour at room temperature. Detection was performed using Amersham ECL Prime or Select Western Blotting Detection Reagent (GE Healthcare Life Sciences) and ChemiDoc Imaging System (Bio-Rad). Data were analysed using ImageLab software.

4.4. UHPLC-MS ANALYSIS

4.4.1. Sample extraction and preparation

COMI cells were seeded in a 6-well plate at a density of 180,000 cells/well in a final volume of 2 mL. The plates were incubated for 72 hours at 37 °C, 5% CO_2 before drug treatment. Cells were treated with different compounds at a concentration of 100 µM and incubated for 3 hours. For the sample extraction, culture media was removed and cells were washed with 1 mL of cold 150 mM ammonium acetate. 750 µL of 80% v/v methanol (LCMS grade), previously cooled to -80 °C, were added to each well. Plates were incubated for 20 minutes at -80 °C, then cells were detached using a cell scraper and harvested in 1.5 mL tubes. The wells were washed with another 250 µL of 80% v/v methanol. The samples were vortexed at 2,000 rpm for 5 minutes at 4 °C and subsequently centrifuged at 14,000 g for 10 minutes at 4 °C. The supernatant was transferred to a new tube. The samples were centrifuged again at 14,000 g for 10 minutes at 4 °C and the supernatant was transferred to a new tube. The samples were dried using a SpeedVac vacuum chamber and then stored at -80 °C.

Just before the analysis, samples were equilibrated to room temperature, resuspended in 30 µL of 5% acetonitrile/0.1% formic acid aqueous solution, and thoroughly mixed. 20 µL of each sample were transferred to a v-bottom 96-well plate and the plate was loaded in the LC-MS autosampler.

4.4.2. Preparation of calibration standards

To prepare the standards for the calibration curves of each compound, five serial drug dilutions, ranging from 0.005 to 50 ng/µL, were prepared in non-treated metabolite extracts resuspended in 5% acetonitrile/0.1% formic acid aqueous solution starting from a 10 mM stock of compound. 20 µL of each dilution plus a blank sample were transferred to a v-bottom 96-well plate and the plate was loaded in the LC autosampler. For D (**1**) derivatives, that are readily hydrolysed in water, each solution was prepared just a few minutes before loading into the column. The lower limit of detection (LOD) was determined by comparing the signals from a set of samples with known low concentrations of analyte with those of blank samples. The minimum signal to noise ratio for the LOD was set to the ratio of 10:1.

4.4.3. Data acquisition

The analytes of each prepared sample subjected to UHPLC-MS analysis were injected into an Ultimate 3000RS (Thermo Scientific) chromatography apparatus and analysed by Fusion Tribrid (Thermo Scientific) mass spectrometer. A Hypersil Gold C18 column (Thermofisher, 100 x 2.1 mm, particle size: 1.9 µm) was used for separation. The LC method consisted of a linear gradient from 5 to 100% B (B: acetonitrile 0.1% of formic acid; A: water + 0.1% formic acid) over 15 min, followed by 4 min at 100% at the flow of 0.2 ml/min. The MS spray voltage was set at 3500V with an ion transfer tube temperature of 300 °C (sheath and auxiliary gasses were set at 20 and 5 Arb, respectively). The MS data were acquired in full scan at 120.000 FWHM (200 m/z) in the scan range of *m/z* 100-1000.

4.4.4. Data analysis

The area under the peak of each precursor ion and the total ion current (TIC) were extracted using Skyline (MacCoss Lab Software) for standards and samples[42]. The software derived precise m/z values as well as isotope distributions for each precursor. Each analyte was also investigated for common adducts, $[M+H]^+$, $[M+K]^+$, $[M+NH_4]^+$, $[M+Na]^+$, and the $[M-H_2O+H]^+$ ions, for each considering the three most abundant isotopes. The normalisation of the area under the peak of each ion by the TIC of the sample was performed to account for variations in sampling volumes across conditions. After normalisation, the precursor level of each ion was determined by taking the median of the different isotopes. Finally, the total area of precursor ions was calculated by summing all precursor levels.

The log-transformed value of the total area of precursor ions of each standard sample was plotted in function of the concentration to construct the standard curves of each compound. Standard linear regression was performed to obtain the linear predictors that were used to compute a concentration value of the compound expressed in ng/µL. In the case of hydrolysable compounds, the contribution of both full molecule and hydrolysed form was taken into consideration. Each molecule form was analysed using its calibration curve and the two concentrations were then summed to obtain the total concentration.

References

1. Louis, D. N. *et al.* The 2021 WHO Classification of Tumors of the Central Nervous System: a summary. *Neuro Oncol* **23**, 1231–1251 (2021).

2. Ostrom, Q. T. *et al.* CBTRUS Statistical Report: Primary brain and other central nervous system tumors diagnosed in the United States in 2010–2014. *Neuro Oncol* **19**, v1–v88 (2017).

3. Stupp, R. *et al.* Radiotherapy plus Concomitant and Adjuvant Temozolomide for Glioblastoma. *New England Journal of Medicine* **352**, 987–996 (2005).

4. Sighel, D. *et al.* Inhibition of mitochondrial translation suppresses glioblastoma stem cell growth. *Cell Rep* **35**, 109024 (2021).

5. Auffinger, B. *et al.* Conversion of differentiated cancer cells into cancer stem-like cells in a glioblastoma model after primary chemotherapy. *Cell Death & Differentiation 2014 21:7* **21**, 1119–1131 (2014).

6. Chen, J. *et al.* A restricted cell population propagates glioblastoma growth after chemotherapy. *Nature 2012 488:7412* **488**, 522–526 (2012).

7. Janiszewska, M. *et al.* Imp2 controls oxidative phosphorylation and is crucial for preserving glioblastoma cancer stem cells. *Genes Dev* **26**, 1926–1944 (2012).

8. Vlashi, E. *et al.* Metabolic state of glioma stem cells and nontumorigenic cells. *Proc Natl Acad Sci U S A* **108**, 16062–16067 (2011).

9. Škrtić, M. *et al.* Inhibition of Mitochondrial Translation as a Therapeutic Strategy for Human Acute Myeloid Leukemia. *Cancer Cell* **20**, 674–688 (2011).

10. Viale, A. *et al.* Oncogene ablation-resistant pancreatic cancer cells depend on mitochondrial function. *Nature 2014 514:7524* **514**, 628–632 (2014).

11. Vlashi, E. *et al.* Metabolic differences in breast cancer stem cells and differentiated progeny. *Breast Cancer Res Treat* **146**, 525–534 (2014).

12. Amunts, A., Brown, A., Toots, J., Scheres, S. H. W. & Ramakrishnan, V. The structure of the human mitochondrial ribosome. *Science (1979)* **348**, 95–98 (2015).

13. Petrov, A. S. *et al.* Structural Patching Fosters Divergence of Mitochondrial Ribosomes. *Mol Biol Evol* **36**, 207–219 (2019).

14. Noeske, J. *et al.* Synergy of streptogramin antibiotics occurs independently of their effects on translation. *Antimicrob Agents Chemother* **58**, 5269–5279 (2014).

15. Delgado, G., Neuhauser, M. M., Bearden, D. T. & Danziger, L. H. Quinupristin-Dalfopristin: An Overview. *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy* **20**, 1469–1485 (2000).

16. CONTRERAS, A. & VÁZQUEZ, D. Synergistic interaction of the streptogramins with the ribosome. *Eur J Biochem* **74**, 549–551 (1977).

17. Parfait, R., de Béthune, M. P. & Cocito, C. A spectrofluorimetric study of the interaction between virginiamycin S and bacterial ribosomes. *Mol Gen Genet* **166**, 45–51 (1978).

18. Li, Q. & Seiple, I. B. Modular, Scalable Synthesis of Group A Streptogramin Antibiotics. *J Am Chem Soc* **139**, 13304–13307 (2017).

19. Li, Q. & Seiple, I. B. Modular Synthesis of Streptogramin Antibiotics. *Synlett* **32**, 647–654 (2021).

20. Dang, J., Bergdahl, M., Separovic, F., Brownlee, R. T. C. & Metzger, R. P. Solvent affects the conformation of virginiamycin M1 (pristinamycin IIA, streptogramin A). *Org Biomol Chem* **2**, 2919–2924 (2004).

21. Dang, J. *et al.* The conformational flexibility of the antibiotic virginiamycin M 1. *European Biophysics Journal* **34**, 383–388 (2005).

22. Martin, J. G. & Hill, R. K. Stereochemistry of the diels-alder reaction. *Chem Rev* **61**, 537–562 (1961).

23. K. C. Nicolau, Scott A. Snyder, Tamsyn Montagnon & Georgios Vassilikogiannis. The Diels–Alder Reaction in Total Synthesis. *Angewandte Chemie International Edition* **41**, 1668–1698 (2002).

24. Li, D., Shi, A., Zhang, W. X., Liu, G. & Xi, Z. Selective synthesis of multiply substituted 7-norbornenone derivatives or Diels–Alder cycloadducts from 1,2,3,4-tetrasubstituted 1,3-butadienes and maleic anhydride with or without Lewis acids. *Tetrahedron* **64**, 9895–9900 (2008).

25. Paris, J. M., Barrière, J. C., Smith, C. & Bost, P. E. The Chemistry of Pristinamycins. *Recent Progress in the Chemical Synthesis of Antibiotics* 183–248 (1990) doi:10.1007/978-3-642-75617-7_6.

26. Bacqué, E., Barrière, J. C. & Berthaud, N. Recent progress in the field of antibacterial pristinamycins. *Current Medicinal Chemistry: Anti-Infective Agents* **4**, 185–217 (2005).

27. Bacqué, E. Influence of Fluorination at Position 16 of Antibacterial Pristinamycins II#. *Chimia (Aarau)* **58**, 128 (2004).

28. Lee, C. K., Minami, M., Sakuda, S., Nihira, T. & Yamada, Y. Stereospecific reduction of virginiamycin M1 as the virginiamycin resistance pathway in Streptomyces virginiae. *Antimicrob Agents Chemother* **40**, 595–601 (1996).

29. Commerçon, A. *et al.* US6815437B1 - Streptogramines, their preparation and compositions containing them. (2004).

30. Böhm, H. J. *et al.* Fluorine in medicinal chemistry. *Chembiochem* **5**, 637–643 (2004).

31. Bacque, E., Barriere, J.-C., Doerflinger, G., Dutruc-Rosset, G. & Pantel, G. US6541451B1 - Streptogramin derivatives, their preparation and compositions containing them. (2003).

32. Achard, D. *et al.* US6569854B1 - Streptogramin derivatives, their preparation and compositions containing them. (2003).

33. Politano, A. D. & Sawyer, R. G. NXL-103, a combination of flopristin and linopristin, for the potential treatment of bacterial infections including community-acquired pneumonia and MRSA. *Curr Opin Investig Drugs* **11**, 225 (2010).

34. BARRIERE, J. C. *et al.* Recent developments in streptogramin research. *undefined* **29**, no-no (1998).

35. Barriere, J.-C., Cotrel, C. & Paris, J.-M. US4798827A - Synergistin derivatives and pharmaceutical compositions which contain them. (1989).

36. Molecular Operating Environment (MOE), 2019.01; Chemical Computing Group ULC, 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, 2021. https://www.chemcomp.com/Research-Citing_MOE.htm.

37. QUACPAC 2.1.3.0: OpenEye Scientific Software, Santa Fe, NM. http://www.eyesopen.com.

38. Korb, O., Stützle, T. & Exner, T. E. PLANTS: Application of Ant Colony Optimization to Structure-Based Drug Design. 247–258 (2006) doi:10.1007/11839088_22.

39. Korb, O., Stützle, T. & Exner, T. E. An ant colony optimization approach to flexible protein–ligand docking. *Swarm Intelligence* **1**, 115–134 (2007).

40. Korb, O., Stützle, T. & Exner, T. E. Empirical scoring functions for advanced Protein-Ligand docking with PLANTS. *J Chem Inf Model* **49**, 84–96 (2009).

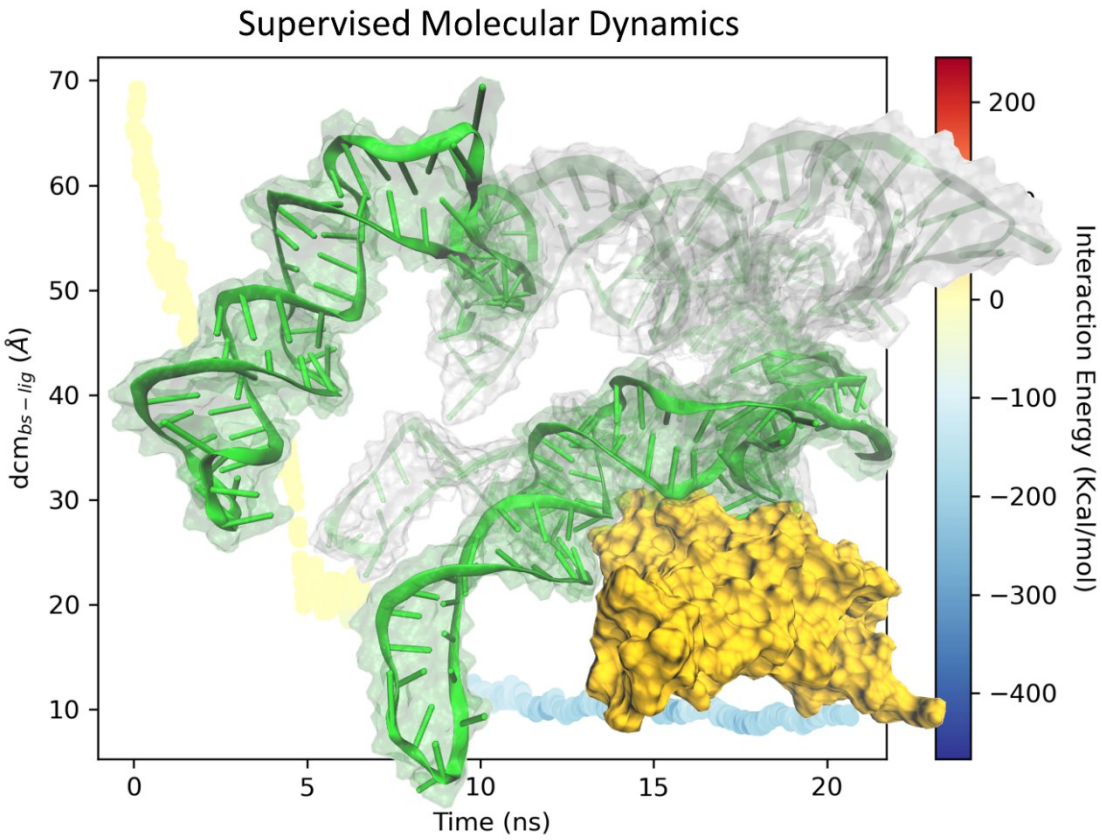
41. Pollard, S. M. *et al.* Glioma stem cell lines expanded in adherent culture have tumor-specific phenotypes and are suitable for chemical and genetic screens. *Cell Stem Cell* **4**, 568–580 (2009).

42. Adams, K. J. *et al.* Skyline for Small Molecules: A Unifying Software Package for Quantitative Metabolomics. *J Proteome Res* **19**, 1447 (2020).

Investigating RNA-Protein Recognition Mechanisms through Supervised Molecular Dynamics (SuMD) Simulations

Matteo Pavan, Davide Bassani, Mattia Sturlese and Stefano Moro

Pavan, M., Bassani, D., Sturlese, M. & Moro, S. Investigating RNA–protein recognition mechanisms through supervised molecular dynamics (SuMD) simulations. *NAR Genom Bioinform* **4**, (2022).



Abstract

Ribonucleic acid (RNA) plays a key regulatory role within the cell, cooperating with proteins to control the genome expression and several biological processes. Due to its characteristic structural features, this polymer can mold itself into different three-dimensional structures able to recognize target biomolecules with high affinity and specificity, thereby attracting the interest of drug developers and medicinal chemists. One successful example of the exploitation of RNA's structural and functional peculiarities is represented by aptamers, a class of therapeutic and diagnostic tools that can recognize and tightly bind several pharmaceutically relevant targets, ranging from small molecules to proteins, making use of the available structural and conformational freedom to

maximize the complementarity with their interacting counterparts. In this scientific work, we present the first application of Supervised Molecular Dynamics (SuMD), an enhanced sampling Molecular Dynamics-based method for the study of receptor-ligand association processes in the nanoseconds timescale, to the study of recognition pathways between RNA aptamers and proteins, elucidating the main advantages and limitations of the technique while discussing its possible role in the rational design of RNA-based therapeutics.

Introduction

According to the central dogma of molecular biology, ribonucleic acid (RNA) is considered the functional link between deoxyribonucleic acid (DNA), which is involved in the storage of genetic information, and proteins, which are the effectors of most pivotal cell functions[1]. Complementing this ancestral and simplistic description of the biological role of RNA, in recent times this polymer has been linked with a variety of regulatory activities within the cell, cooperating with proteins to finely tune the genome expression and other biological processes[2]. The consideration that the vast majority of human RNA is not translated into proteins[3] in conjunction with the fact that a large number of newly discovered non-coding RNAs are associated with various pathologies and illnesses[4,5], caused an increase in the popularity of RNA among the scientific community, both from a biological and a therapeutic perspective[6,7].

From a structural point of view, RNA exists mainly as a double-stranded molecule with a chain length that can span from a few tens of nucleobases, as in small hairpins[8], to few thousand nucleotides, as in long non-coding sequences[9]. Compared to DNA, the higher conformational freedom of ribonucleic acids implies that they can assume a wide variety of three-dimensional structures in solution[10,11], organizing themselves in functional domains specifically designed to recognize other nucleic acids[12], proteins[13], glycated derivatives[14] or small organic molecules[15].

From a functional perspective, successful exploitation of RNA's ability to mold itself into different three-dimensional structures able to recognize target biomolecules with high affinity and specificity is represented by aptamers[16]. This class of single-stranded oligonucleotides that fold into defined and complex architectures including stems, loops, bulges, hairpins, pseudoknots, triplexes, or quadruplexes, can bind several molecular targets, such as proteins, small organic molecules, and ions, thus classifying as a useful tool both for a therapeutic and diagnostic purpose[17,18].

As is the case for other nucleic acids, interactions between aptamers and proteins are characterized by a complex network of van der Waals, hydrogen bond, stacking, and general non-polar interactions that define the complementarity of shape and electrostatic properties at the surface between the two interactors and determine the specificity of binding to a certain target[19,20]. To the present date, the rising amount of experimentally

solved three-dimensional RNA structures, especially concerning their complexes with both macro and small molecules, has led to an increased interest by the scientific community in the investigation of RNA structures at an atomistic level of detail, to apply Structure-Based Drug Design (SBDD) strategies for the rational design of novel therapeutic entities[21,22].

Among the tools that are routinely used, both in the academic and industrial environment, for investigating the structural determinants of biological entities' recognition, molecular docking is by far the most widely and successfully adopted[23]. Originally developed for predicting the interaction between small organic molecules and proteins[24], throughout the years this computational technique has been also applied to the investigation of protein-protein[25], protein-peptides[26], and antigen-antibodies[27] complexes with various degrees of success. Contrary to proteins, the application of molecular docking to the study of complexes involving nucleic acids has been so far very limited, and mainly applied to the prediction of the binding mode of small molecules[28]. Compared to the aforementioned established docking protocols, a smaller number of methods are available for the investigation of nucleic acids-protein complexes, due to some intrinsic structural peculiarities of nucleic acids, particularly in the case of RNA[29]. The first limitation is represented by the distinctive charge distribution that characterizes the RNA surface compared to that of proteins[30], the second one is portrayed by the neglected treatment of the role of solvent[31] and the third one is the lack or limited consideration of the structural flexibility and dynamicity of ribonucleic acids[32].

One possible approach to overcome the limitations of molecular docking is represented by molecular dynamics (MD) simulations. Despite the enhanced description of the binding event derived from the explicit treatment of solvent molecules and the consideration of both receptor and ligand flexibility, MD simulations are rarely carried out to investigate the whole binding event due to the long simulation times and computational effort required to sample these infrequent events and are therefore mostly exploited for the refinement of docking results[33].

To mitigate the time constraints of classic molecular dynamics, one possible strategy is the exploitation of enhanced sampling techniques that allow increasing the frequency of observing desired events, such as the receptor-ligand association[34]. Among the plethora

of enhanced sampling protocols that have been developed throughout the years, Supervised Molecular Dynamics (SuMD) has proven to be particularly successful in investigating ligand-receptor recognition pathways at an atomistic level of detail without applying any energetic bias to the system while contemporarily reducing by orders of magnitude the simulation time compared to classic MD simulations[35]. Particularly, from a ligand perspective, SuMD simulations have proven useful to work with a variety of molecular entities, ranging from small organic molecules (both fragment[36–41] and mature, lead-like, compounds[42–44]), to more complex chemical species such as macrocycles[45] and peptide ligands[46]. From a receptor point of view, instead, SuMD was successfully applied to the study of both soluble[47,48] and membrane[49–51] systems, including both protein[52,53] and nucleic[54] targets.

In this scientific work, SuMD simulations were applied for the first time to the study of the recognition process between RNA macromolecules and proteins, to extend the applicability domain of the methodology. Particularly, we decided to focus our attention on RNA aptamers, due to their relevance as diagnostic and therapeutic tools and due to the variety of their structural landscape.

Briefly, we present four different applications of the SuMD methodology to RNA aptamer-protein complexes. Three test cases, involving systems in which the three-dimensional structure of the complex is known and deposited in the Protein Data Bank[55], are used to validate the ability of the SuMD protocol to correctly reproduce the experimental data while giving additional useful information that goes beyond the final state of the recognition process. The fourth and final case, concerning instead a complex whose structure has not yet been experimentally determined, is presented to show a possible prospective application of the SuMD protocol, discussing at the same time the advantages and limitations of the technique other than its possible role in a typical drug discovery pipeline.

MATERIAL AND METHODS

Hardware Overview

General molecular modeling operations, such as the preparation of RNA aptamer-protein complex structures, the system setup for molecular dynamics simulations, and trajectory analysis were performed on an 8 CPU Linux workstation equipped with an Intel Xeon E5-1620 3.50 GHz processor. All molecular dynamics simulations were performed on a GPU cluster composed of 20 NVIDIA devices ranging from GTX980 to Titan V.

Structures Preparation

The three-dimensional coordinates of the three RNA aptamer-protein complex used as control cases in this study were retrieved from the Protein Data Bank[56] (PDB ID: 3DD2, 4PDB, 5VOE) and prepared for subsequent simulations exploiting several modules from the Molecular Operating Environment (MOE) 2019.01 suite[57]. At first, structures were pre-processed through the “Structure Preparation” tool, to assign each residue with alternate conformations to the one with the highest occupancy, build missing loops through homology modeling, and correct inconsistencies between the primary sequence and the tertiary structure. For compatibility with each piece of software used in this work and for consistency with previous studies involving the SuMD approach, structures were manually edited to mutate each non-natural nucleic residue (e.g.: fluorinated nucleotides) to the corresponding natural alternative. Afterward, titratable residues were assigned to the most probable protonation state at pH = 7.40 exploiting the “Protonate3D” tool. Finally, each non-protein or non-nucleic residue was removed, and the nucleic ligand was moved away from the binding site at a distance of at least 30 Å from the nearest receptor atom, in order to explore the conformational degree of freedom of the ligand throughout the recognition process.

For the investigation of the recognition process between the SARS-CoV-2 Spike RNA aptamer, the structure of the Spike Receptor Binding Domain (RBD) was retrieved from the Protein Data Bank (accession code: 6M0J[58]). Concerning the RNA aptamer, the experimental structure was not available, therefore the primary sequence was obtained from the Supplementary Material of the original work from Valero et al.[59] and sequentially

submitted to the NUPACK[60] and 3DRNA[61] webservers to retrieve the predicted secondary and tertiary structure respectively. In both cases, default parameters were chosen, and the lowest energy structure coming from the 3DRNA webserver was used for both docking calculations and MD simulations. Every other preparation passage is identical to the ones executed for the three control cases.

System Setup and Equilibration Protocol

Nucleic-protein systems coming from the preliminary preparation stage were then further processed making use of both Visual Molecular Dynamics (VMD) 1.9.2[62] and several tools from the Ambertools14 suite[63]. Each protein or nucleic atom was parameterized according to the ff14SB force field with χ modification tuned for RNA (χ_{OL3})[64–66]. At first, each system was solvated in a cubic box of TIP3P[67] water molecules with a padding of 35 Å. Afterward, each solvated system was neutralized through the addition of an appropriate number of Na^+ and Cl^- counterions until a salt concentration of 0.154 M was reached. Before Molecular Dynamics (MD) simulations, each system was subjected to a 1500-step energy minimization phase with the conjugate-gradient algorithm.

Each minimized system underwent then a two stages equilibration protocol. The first stage consisted of 1 ns of simulation in the canonical ensemble (NVT), applying a 5 kcal mol^{-1} $Å^{-2}$ harmonic positional restraint on each protein and nucleic atom. The second stage consisted instead of a 2 ns simulation in the isothermal-isobaric ensemble (NPT), with the same restraints applied only to the backbone atoms of both the protein and the nucleic acid. For each MD simulation performed in this scientific work, an integration step of 2 fs was used, the temperature was kept at a constant value of 310 K through a Langevin thermostat[68], and the M-SHAKE algorithm[69] was used to constrain the length of bonds involving hydrogen atoms, the particle-mesh Ewald (PME)[70] method was exploited to compute electrostatic interactions using cubic spline interpolation and a 1 Å grid spacing, while a 9.0 Å cutoff was set for calculation of Lennard-Jones interactions. For simulations in the NPT ensemble, the pressure was kept at a constant value of 1 atm through a Monte Carlo barostat[71]. Finally, all MD simulations were run through the ACEMD 3[72] engine, which is based upon OpenMM 7[73].

Supervised Molecular Dynamics (SuMD) Simulations

Supervised Molecular Dynamics (SuMD) is a well-established enhanced-sampling molecular dynamics approach that has been successfully applied to the study of the recognition process between various molecular entities at an atomic level of details on the nanosecond timescale[35].

The main advantage of the SuMD approach compared to traditional molecular dynamics simulations is the improved ability to sample infrequent events such as molecular association processes, thus reducing the timescale of the simulation that is required to spontaneously observe a binding event from the microseconds range to a few nanoseconds.

In detail, this task is accomplished by performing a sequence of short, unbiased, MD simulations followed by an evaluation of the simulation progress by a tabu-like algorithm. In this case, each of these MD simulations, defined as the “SuMD-step”, is run in the canonical ensemble at a constant temperature of 310 K for 300 ps, but the length of the “SuMD-step” can vary and is chosen according to the system that is studied. At the end of each “SuMD-step”, the distance between the center of mass of the ligand and one of the user-defined binding sites is computed at each step of the simulation, and this data is then fitted into a linear function: if the slope of the resulting straight-line is negative, indicating that the ligand is approaching the binding site, the “SuMD-step” is considered productive and retained for the generation of the final trajectory, while the final state of the simulation is used as the starting point for the successive step. On the contrary, if the slope is positive, thereby indicating that the ligand is not approaching the binding site, the “SuMD-step” is considered not productive and therefore discarded: in this case, the step is repeated by randomly reassigning the particle velocities through the Langevin thermostat and retaining the final coordinates from the end of the previous “SuMD-step”. The supervision algorithm is switched off when the distance between the two centers of mass falls below a threshold value (10 Å, in this case): from that point on, the simulation proceeds for the other 10 ns of classic molecular dynamics, allowing the system to relax and reach the final state of the simulation without any external geometric biased imposed by the supervision. For each control case study, 10 SuMD simulations were collected:

while every single one was visually inspected, only the best one according to the geometrical agreement with the reference (based on the RMSD between the ligand coordinates in the final step of the simulation and the ligand coordinate in the reference experimental structure after optimal superposition of the protein backbone) was thoroughly analyzed and discussed in the manuscript. In the case where the structure of the complex was not available, the best replica was chosen based on the MMGBSA interaction energy instead. In the current implementation, the SuMD code is written in Python and exploits the Numpy and ProDy[74] modules to perform the aforementioned geometrical supervision throughout the simulation. A list of residues utilized to define the ligand and protein binding site for each case is provided in Table S4 (Supplementary Materials).

Trajectory Analysis

The SuMD trajectories were analyzed by making use of an-in house tool written in Python 3 which represents an evolution and customization of the original one which is described in the work of Salmaso et al.[46].

Initially, trajectories representing each single "SuMD-step" were merged into a single collective trajectory. Then, obtained trajectories were pre-processed by applying a stride and retaining one frame every 20 ps, superposing and aligning each frame on the protein backbone atoms of the first frame, and wrapping it into an image of the system simulated under periodic boundary conditions (PBC). Both geometric and energetic analyses were performed on the so-obtained SuMD trajectories.

Concerning the geometric properties of the system, regarding both the nucleic ligand and the protein receptor, the time-dependent evolution of both backbone RMSD and radius of gyration, a global and a time-dependent per-residue decomposition of the backbone RMSF were collected and reported in an aggregated panel. Furthermore, the geometric performance of the SuMD protocol in reproducing the experimental bound conformation of the ligand was evaluated by computing the ligand backbone RMSD compared to the experimental reference throughout the entire simulation. All these geometric analyses were performed making use of the appropriate functions of the MDAnalysis[75,76] Python library and plotted through the Matplotlib[77] module.

Regarding the energetic analysis, an estimation of the ligand-receptor interaction energy alongside the SuMD trajectory was obtained both through the MMGBSA protocol, as implemented in AMBER 14, and through the "NAMD Energy" plugin for VMD, which exploits the NAMD[78] package to retrieve an estimate of the interaction energy defined as the sum of the van der Waals and electrostatic contribution calculated according to the user-defined force field (AMBER 14, in this case). The energy values were then plotted both as a function of the simulation time and of the RMSD to the reference pose, giving both a time-dependent and a geometry-dependent energetic profile of the trajectory.

Finally, a per-residue interaction energy decomposition analysis was carried out exploiting once again the "NAMD Energy" plugin for VMD: plots report a time-dependent per-residue decomposition of the interaction energy for both the receptor and the ligand and a bidimensional interaction energy matrix in which interacting residues on the ligand side are correlated with the corresponding interacting residues on the receptor side. For all these per-residue analyses, the 25 most frequently contacted residues throughout the trajectory are considered (25 for the receptor, as well as for the ligand), defining contacting residues as the ones that are at a maximum distance of 4.5 Å from the nearest atom of the counterpart, either the ligand or the receptor.

A movie representation of the trajectory alongside the dynamic evaluation of its geometric and energetic features is also provided by the same analysis tool, which exploits VMD for the visual rendering of the simulated system. For uniformity reasons, in each plot and video, residue numbering is related to the fasta sequence for the wild-type receptor, as retrieved from the UniProt database.

Docking

To evaluate SuMD's ability to reproduce the native conformation of the RNA aptamer-protein complex, we decided to compare its performance with the one of molecular docking. The program chosen to accomplish this task was HADDOCK[25] ("High Ambiguity Driven protein-protein DOCKing", version 2.4) since it has already been extensively used for dealing with protein-nucleic acid complexes[79] and it uses *a priori* information to steer the docking calculation in a similar way to how SuMD works.

Each one of the crystal structures used as control (the ones coming from PDB codes 3DD2, 4PDB, and 5VOE) was subjected to a docking run. For all these cases, the nucleic acid was treated as the ligand, while the protein was considered as the receptor. The binding site was defined based on residues at the contact surface in the crystallographic structures, both on the protein and nucleic acid sides.

Concerning the SARS-CoV-2 Spike RBD RNA aptamer, the selected protein residues were chosen instead based on the contact surface with human ACE2 in the structure 6M0J, while for the aptamer the residues were selected based on the information coming from the original paper by Valero et. al. The list of protein and aptamer residues used as input for each docking calculation is reported in Table S3 (Supplementary Material). All user-definable parameters for molecular docking were kept as default.

HADDOCK starts with a randomization stage, in which the docking partners are placed far in space from one another (about 150 Å) and randomly rotated around their centers of mass. The following step consists of a rigid body energy minimization, which is followed by the rigid-body docking of the ligand and the receptor, allowing to obtain 1000 complexes. The 200 best solutions in terms of intermolecular energies obtained at this stage are subjected to simulated annealing refinements. Both the intra- and inter-molecular energies are evaluated by HADDOCK using full electrostatic and van der Waals energy terms with an 8.5 Å distance cutoff using the OPLS[80] nonbonded parameters. The final complexes are then clustered based on the Fraction of Common Contacts[81] (FCCs) with a 0.6 cutoff similarity for clustering, and the clusters are then ranked for energetics.

RNA Aptamer-Protein Complexes Molecular Dynamics Simulations

To evaluate the dynamic behavior of aptamer-bound protein complexes, we performed several classic molecular dynamics simulations. At first, each system was subjected to a preparation step exploiting both AmberTools14 and VMD 1.9.2, as previously mentioned in the preparation stage for SuMD simulations. Specifically, each protein-nucleic acid complex was singularly solvated in an explicit TIP3P water box with a 40 Å padding. Each of these simulation boxes was then neutralized using Na^+/Cl^- ions until reaching a physiological salt concentration of 0.154 M. The preparation phase was followed by a two-step equilibration protocol. The first equilibration was carried out in the canonical

ensemble (NVT) and was composed of 1500 steps of energy minimization with a conjugate-gradient algorithm followed by a 1 ns MD simulation. In this first passage, harmonic positional restraints of 5 kcal mol^{-1} $Å^{-2}$ were applied on both the protein and the nucleic acid, while the temperature was kept at the constant value of 310 K exploiting a Langevin thermostat (with friction coefficient set to 0.1 ps^{-1}). The second equilibration was performed in the isothermal-isobaric ensemble (NPT), also this time for 1 ns of MD simulation in which the harmonic positional restraints of 5 kcal mol^{-1} $Å^{-2}$ were applied just on the protein and nucleic acid backbones. The pressure was kept constant at the value of 1.0 atm using a Monte Carlo barostat. In each of the equilibration steps, a 2 ps integration step was adopted, the bonds involving the hydrogen atoms were constrained through the M-SHAKE algorithm, and a 9.0 Å cutoff was used for the calculation of the Lennard-Jones interaction. For the electrostatic interaction, a particle-mesh Ewald method (PME) was used. After this preparation phase, three different 50 ns replicates of classic MD simulation in the NVT ensemble at 310 K were executed.

Free RNA-Aptamer Molecular Dynamics Simulations

To complement the investigation of the structural dynamicity of investigated RNA aptamer-protein complexes, we also performed a classic MD simulation of the free RNA aptamer. To accomplish this task, we retrieved the coordinates for each nucleic acid molecule from the previously mentioned complexes. Each of these aptamers was then subjected to the same protocol described before for RNA aptamer-protein complexes, except for the parts related to the protein which, in this case, was not part of the system.

Results

To assess the applicability domain and accuracy of Supervised Molecular Dynamics simulation in the context of the nucleic acids-protein recognition processes, we opted for a retrospective validation approach, evaluating the ability of the protocol to correctly reproduce the binding mode of nucleic ligands found in experimentally solved complex structures, focusing both on the sampling and ranking capabilities of the protocol. Particularly, we decided to focus our attention on the class of RNA aptamers, both for their therapeutical relevance and for their challenging nature due to their peculiar structural features such as intrinsically higher flexibility and density of negative charge compared to ligands considered in the past applications of the SuMD protocol. In the following paragraphs, we present the application of SuMD to three different case studies for which the experimental structure of the RNA aptamer-protein complex is available on the Protein Data Bank, focusing on the geometrical accuracy of the technique in reproducing the experimentally determined binding mode and monitoring both the geometric and energetic features of the recognition process, stretching beyond the final state of the simulation. The three test cases are reported in chronological order, starting from the oldest structure to the most recent one. Furthermore, we also present a prospective application of the SuMD protocol to the investigation of a complex whose structure has not yet been experimentally determined, to present and discuss the role, the advantages, and the limitations of implementing the SuMD protocol in a pipeline for the rational design of RNA-based therapeutics. Information about each SuMD simulation reported in the manuscript are encompassed in Table S2 (Supplementary Material).

RNA aptamer bound to human thrombin (PDB ID: 3DD2)

Due to its ability to process several proteins that are part of the coagulation cascade, including the cleavage of soluble fibrinogen into fibrin, which is responsible for the formation of clots, human thrombin is a serine protease that exerts a pivotal role in blood coagulation and is, therefore, a target of interest for anticoagulation therapy[82–84]. Two surface regions of thrombin (exosite-1 and exosite-2), which are located on opposite sides of the molecule and away from the catalytic site, are responsible for its ability to interact with various macromolecular substrates. Particularly, exosite-2 is responsible for the

binding of thrombin to heparin, a clinically used oligosaccharide that mediates its anticoagulant effect by facilitating the interaction of thrombin with its endogenous inhibitor antithrombin[85].

In 2001, White et al. reported the discovery of Toggle-25, an RNA aptamer that was developed to bind with a high affinity to both human and porcine thrombin, leading to the inhibition of both plasma clot formation and platelet activation[86]. In 2008, Long et. al. were able to solve the crystal structure of Toggle-25t, a 25 nucleotide truncated version of Toggle-25, bound to the exosite-2 of human thrombin at a resolution of 1.90 Å (PDB ID: 3DD2[87]), thereby allowing a structural characterization of the complex which nicely complements previous biochemical analysis[87]. As reported in the original publication, the aptamer recognizes the thrombin exosite-2 in its native state, since no conformational changes can be observed between the unbound and bound state. Despite a relatively simple secondary structure, defined by a stem-loop with an internal bulge, Toggle-25t can achieve a selective and high-affinity binding to human thrombin (K_d = 0.54±0.1 nM) thanks to the good complementarity of shape and electrostatic properties between the negatively charged aptamer and the basic protein region responsible for its recognition. The absence of significant structural alteration of the protein upon binding, the therapeutic relevance of the target, and the relatively modest size of the aptamer (25 residues), make this complex an ideal target for the application of the SuMD protocol to the study of RNA aptamer-protein interactions. As previously introduced, two simplifications were introduced in the system investigated through SuMD: firstly, each 2' fluoro substituted pyrimidine residue was retro-mutated to the correspondent naturally occurring nucleotide, secondly, the divalent Mg^{2+} ion was not included in the system. The choice of retro mutating the fluorine-containing nucleotides was done for compatibility reasons since some of the software used in this scientific work (e.g., HADDOCK) could not work with non-natural nucleic residues. According to the original publication by Long et al.[87], the introduction of fluorine mainly impacted the aptamer resistance to ribonuclease rather than the binding affinity, as also underlined by the fact that only one single 2' fluoro group is in direct contact with the protein, specifically at the level of the U17-Arg126 interaction. Concerning the presence of Mg^{2+} ions, despite their undisputable importance

in the field of RNA folding, we opted not to include them at all in the simulations due to some intrinsic molecular mechanics limitations that hamper the possibility of fully and accurately describing their interaction with RNA[88], other than the difficulties in accurately predicting their locations without hints from experimental data[89].

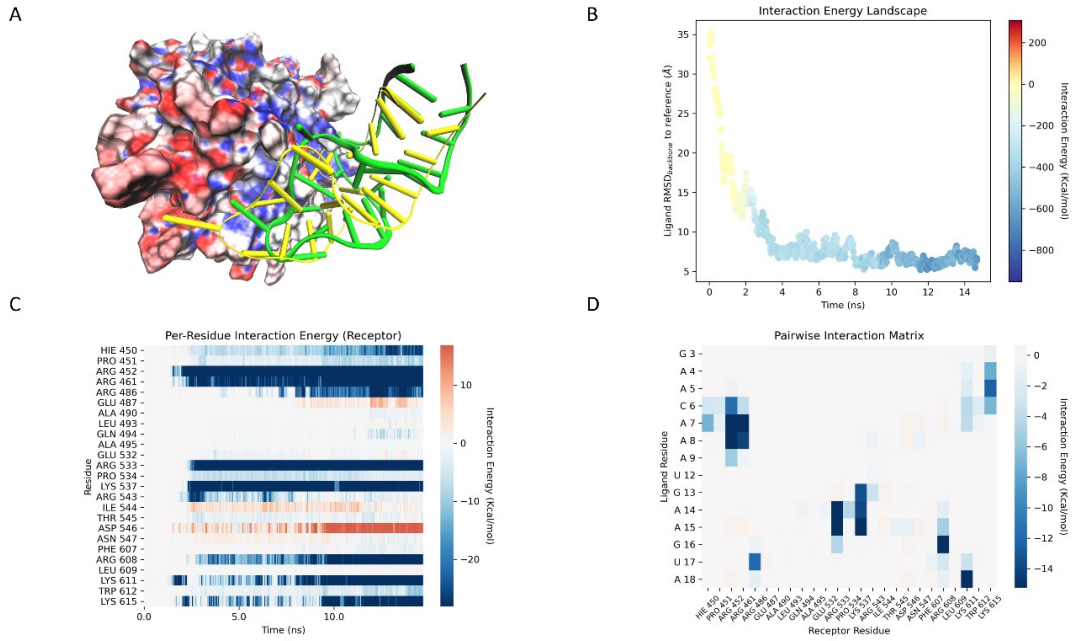


Figure 1. This panel encompasses the putative recognition pathway between the RNA aptamer Toggle-25t and the exosite-2 of human thrombin described by the best trajectory obtained through the SuMD protocol (the one with the lowest RMSD to the crystal reference). (A) Visual representation of the Toggle-25t conformation sampled in the last frame of the SuMD trajectory (green) superposed with the native Toggle-25t conformation (yellow), as found in the crystal structure deposited in the Protein Data Bank with accession code 3DD2 ($RMSD_{SuMD\text{-}Crystal}$: 6.41 Å). The aptamer is represented as a ribbon, while the protein is represented as a Connolly surface colored according to the electrostatic potential as calculated with the APBS software[90], where red indicates a negatively charged area while blue indicates a positively charged one. (B) Profile of the ligand-receptor interaction energy (defined as the sum of the electrostatic and van der Waals contribution) throughout the recognition process as a function of both the simulation time and the RMSD between the ligand position during the trajectory and the ligand position in the crystal. (C) Receptor per-residue decomposition of the receptor-ligand interaction energy throughout the SuMD trajectory as a function of the simulation time: the 25 most-contacted residues are reported in the plot. (D) Per-residue interaction energy matrix: the 25 most-contacted residues for both the receptor and the ligand are considered, while each square composing the heatmap represents the average value of the interaction energy between the two paired residues alongside the trajectory.

As can be seen in Video 1 (Supplementary Material), about 15 ns of simulation time was sufficient to sample a putative molecular recognition event between the Toggle-25t RNA aptamer and the human thrombin exosite-2. This is a quite remarkable result, considering the usual hundreds of nanoseconds that are necessary to spontaneously sample a binding event with classic MD simulations. As illustrated by Video 1 and summarized in Figure 1, the final state of the SuMD simulation converged quite well both from a geometrical and an interactive point of view toward the crystal reference.

Concerning the geometric accuracy, the RMSD between the ligand backbone conformation in the final state of the simulation and the native ligand backbone conformation observed in the crystal structure is 6.41 Å, which is quite impressive considering the intrinsic structural flexibility of these objects, which is also confirmed by the structural deviation than can be observed in classic MD simulations of both the crystal complex (average RMSD of the nucleic acid backbone in the final step of the simulation across three MD replicates: 2.87 Å) and the free aptamer (average RMSD of the nucleic acid backbone in the final step of the simulation across three MD replicates: 3.85 Å). It is not surprising, therefore, that SuMD performs worse compared to molecular docking from a geometrical point of view (RMSD between the best docking pose and the crystal binding mode of the aptamer: 2.51 Å).

Despite a lower geometrical accuracy of the method compared to the one of docking, that derives mainly from the high intrinsic flexibility of the nucleic ligand during the simulation (see also Figure S4, Supplementary Material, which reports the pairwise RMSD matrix of the free RNA aptamer during classic MD simulations), SuMD can correctly pose the negatively charged RNA-aptamer in a native-like conformation that maximizes the complementarity of shape and electrostatic features with the electropositive concave surface of the thrombin exosite-2, as highlighted in Figure 1 (panel A).

Concerning the capability of the protocol to correctly recapture the pivotal binding features despite a suboptimal geometric accuracy, SuMD can accurately describe the main interaction determinants, as illustrated by a comparison between the per-residue energy decomposition of the first 300 ps of the crystal complex classic MD simulation and the last 300 ps of the SuMD simulation (Figure S5, Supplementary Material). Particularly,

as can be seen in Figure 1 (panels C and D), SuMD correctly captures the pivotal role of both Arg 101 (461) and Arg 233 (608) in driving the recognition mechanism, serving as electrostatic recruiters in the initial phases of the process and acting as anchors to stabilize the bound state in the final part of the simulation, in agreement with mutagenesis data that assess how mutation of each of this two residue completely abrogates any aptamer's effect[91]. Particularly, Arg 233 (608) along with Arg 165 (533) are responsible for the formation of a stacked interaction domain motif defined as an "A-Arg zipper" which involves five unpaired adenine residues on the ligand side, A4, A5, A7, A15, and A18 respectively. As can be seen in Figure 1 panel D, SuMD individuates all these five adenine residues as key interaction determinants, in agreement with the experimental data. Finally, SuMD also discriminates the non-relevance of Gln 239 (614), which is not reported in the analysis as it is not one of the 25 most contacted residues during the trajectory, coherently with mutagenesis data that shows how a mutation of this residue does not affect the aptamer binding[91].

On the ligand side, SuMD once again correctly encompasses the different roles portrayed by different nucleotides. As can be seen in Figure 1 (panel D), the interaction between U17 and Arg 126 (486) is retrieved by the SuMD simulation: mutagenesis data shows how the mutation of U17, one of the flipped-out nucleotides, with adenine has essentially no effect on the aptamer's affinity for the target. Looking at our ligand-based interaction map it can be noticed how this interaction is the only one in which this nucleotide is involved, other than being less intense compared to more prominent interaction such as the aforementioned "A-Arg zipper", which suggest a non-pivotal contribution to the binding affinity. This can be related to the fact the U17 interacts with Arg126 through the backbone and not through its sidechain so that, as pointed out by the work of Jeter et al., this interaction would be maintained even when substituting the base[91]. Substitution of U12 with adenine results in a nearly three-orders-of-magnitude diminished binding affinity: this is due to the stabilizing role that U12 plays towards A15 through a non-Watson-Crick base pairing, one of the bases involved in the formation of the "A-Arg zipper"[91]. Once again, as depicted in Figure 1 panel D, SuMD correctly recognizes the pivotal role portrayed by A15, while contemporarily elucidating the indirect role of U12,

which is not involved in any major interactions with protein residues. All other geometric and energetic analyses performed on the trajectory are summarized in Figures S1-S3 (Supplementary Materials).

To assess the predictive power of the method, we retrospectively analyzed all trajectories using two different metrics, i.e. the electrostatic interaction energy and the MMGBSA interaction energy of the final state of the simulation. The idea to use these two metrics stems from the consideration that RNA binding to proteins usually requires a good level of complementarity of steric and electrostatic properties at the binding interface. As reported in Figure S19, both metrics can successfully distinguish the native and native-like poses, i.e. the ones with a superimposable interaction pattern with the reference (measured through the Mean Signed Error and the Root-Mean Squared Error of the per-residue interaction energy decomposition, panels C and D respectively), from the decoys. This observation suggests that both metrics could be utilized in a prospective application of SuMD to rank poses coming from different simulations prioritizing the ones that are most similar to the native binding mode.

RNA aptamer bound to Bacillus anthracis ribosomal protein S8 (PDB ID: 4PDB)

Thanks to its central function in the vital cycle of bacteria, the bacterial ribosome, complex machinery responsible for protein synthesis in prokaryotic organisms, has been extensively studied, both from a structural and a functional perspective, and has been validated as a target for multiple antibiotic drugs[92]. Protein-RNA interactions play a key role in the assembly, maturation, and function of the bacterial ribosome[93]. Among these, association processes involving ribosomal protein S8 are particularly relevant since it not only participates in the 30s subunit assembly by binding to 16S rRNA but, additionally, it also serves as a translational repressor of the *spc* operon mRNA, which encodes for 11 ribosomal proteins including S8 itself[94].

Due to its relevance, the complex formed between bacterial ribosomal protein S8 and 16S rRNA has been thoroughly characterized through different techniques, allowing us to establish that most of the protein–RNA contacts involve helices 21 and 25 and that a small RNA portion located in helix 21 is sufficient to confer specificity and high affinity to the S8-RNA interaction[95]. Furthermore, the interaction determinants between S8 and its RNA targets are largely conserved, and the same degree of conservation applies also to the overall fold of various S8 proteins[96–98]. Finally, the complementarity of shape and electrostatic properties that is required for the binding entails a high level of nucleotide sequence and secondary structure conservation, to impose an RNA shape that optimizes interaction properties with the protein surface[99].

To fetch RNA secondary structures that deviate from the conserved bacterial motif while retaining the ability to bind the S8 protein, in 2014 Davlieva et al. performed a SELEX experiment that led to the discovery of a 38-mer RNA aptamer that can bind the Bacillus anthracis S8 protein with high affinity (K_d = 110 ± 30 nM), determining at the same time the structure of the bound complex between the RNA aptamer and its protein target[99]. The selection process was based on an RNA stem-loop scaffold containing symmetric and asymmetric internal loops of 16 randomized nucleotides, with the resulting aptamer sequence forming a secondary structure with a symmetric internal loop[99]. The absence of major structural rearrangements on the protein side of the interaction (0.65 Å $RMSD_{backbone}$ between the free and the bound form) and the relevance of this interaction

from a mechanistic perspective led us to consider this complex as a suitable casestudy to validate the SuMD protocol.

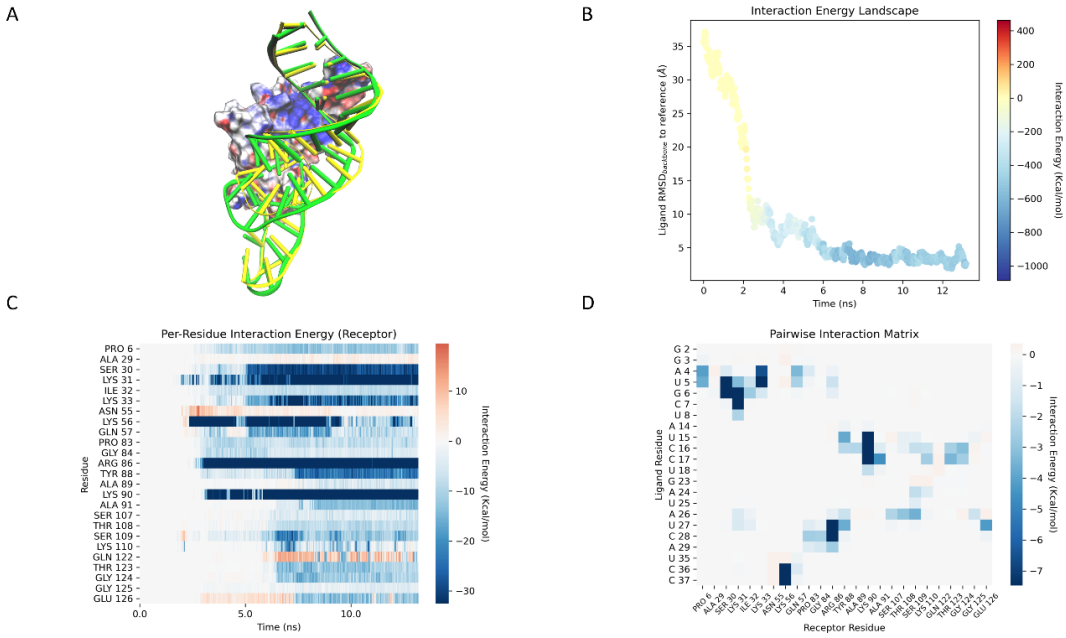


Figure 2. This panel encompasses the putative recognition pathway between the RNA aptamer and the S8 ribosomal protein of Bacillus anthracis described by the best trajectory obtained through the SuMD protocol (the one with the lowest RMSD to the crystal reference). (A) Visual representation of the RNA aptamer conformation sampled in the last frame of the SuMD trajectory (green) superposed with the native RNA aptamer conformation (yellow), as found in the crystal structure deposited in the Protein Data Bank with accession code 4PDB ($RMSD_{SuMD\text{-}Crystal}$: 2.61 Å). The aptamer is represented as a ribbon, while the protein is represented as a Connolly surface colored according to the electrostatic potential as calculated with the APBS software[90], where red indicates a negatively charged area while blue indicates a positively charged one. (B) Profile of the ligand-receptor interaction energy (defined as the sum of the electrostatic and van der Waals contribution) throughout the recognition process as a function of both the simulation time and the RMSD between the ligand position during the trajectory and the ligand position in the crystal. (C) Receptor per-residue decomposition of the receptor-ligand interaction energy throughout the SuMD trajectory as a function of the simulation time: the 25 most-contacted residues are reported in the plot. (D) Per-residue interaction energy matrix: the 25 most-contacted residues for both the receptor and the ligand are considered, while each square composing the heatmap represents the average value of the interaction energy between the two paired residues alongside the trajectory.

As can be deducted from Video 2 (Supplementary Materials), less than 15 ns of simulation time were needed to sample a putative recognition mechanism between the RNA aptamer and the ribosomal S8 protein. As underlined by both Video 2 and Figure 2, the final state of the SuMD simulation converged impressively well with the experimental data, both from a geometric and an interactive point of view.

Specifically, regarding the geometric accuracy of the method, the RMSD between the ligand backbone conformation in the final state of the simulation and the native ligand backbone conformation observed in the crystal structure is only 2.61 Å, a result comparable with the performance of molecular docking (RMSD between the best docking pose and the crystal binding mode of the aptamer: 2.20 Å). The higher geometric accuracy of the SuMD protocol, compared to 3DD2 case, can be partially explained by the lower degree of conformational freedom available to the aptamer, as can be seen in Figure S9 (Supplementary Materials, average RMSD of the nucleic acid backbone in the final step of the simulation across three MD replicates: 2.27 Å), despite similar stability of the bound state (average RMSD of the nucleic acid backbone in the final step of the simulation across three MD replicates: 2.80 Å).

Intriguingly, despite the impressive geometric convergence of the trajectory with the experimental data, the comparison between the per-residue energy decomposition from the last 300 ps of the SuMD simulation and the first 300 ps of the classic MD simulation of the crystal complex reveals a slightly lower congruence of the binding mode compared to the 3DD2 case (Figure S10, Supplementary Materials). It is important to notice that, in this case, the difference is not due to the interaction pattern, which for the most part is correctly depicted by the SuMD simulation analysis, but instead to the relative strength of the interactions. Indeed, as can be noticed in Figure 2 panel A, the final state of the SuMD simulation is slightly shifted compared to the crystal reference, considering that the predominant electrostatic component to the total interaction energy is proportional to the squared distance between the two interactors, even small differences in the relative position of interacting residues can alter the quantitative estimation of the interaction energy. Coherently with this interpretation of the data coming from the simulation

analysis, our SuMD protocol can qualitatively describe the vast majority of the key interaction determinants.

In the crystal structure, the interaction between the S8 protein and the RNA aptamer involves a strip of electropositive charge (as is also visible in Video 2 and Figure 2, panel A) along which the phosphate backbone of the aptamer traverses from residues A4-C7 and U27-A29: this behavior is correctly captured by our simulation, as depicted in Figure 2 panel D. Moreover, as can be noticed in Figure 2 panels C and D, SuMD correctly intercepts the polar interactions that form between the backbone of aptamer residues C16, C17, A24, U25, U27, and A26 and their counterparts on the protein side such as Glu 126, Ser 107, Gly 124, Lys 110, Ser 109, Ala 91 and Thr 123. Additionally, as depicted in Figure 2 panel D, the SuMD protocol is also able to spot some water-mediated interactions such as the contact between U27 to Glu 126. Finally, as can be depicted in Figure 2 panel D, SuMD can retrieve the stacked interaction between the peptide bond of highly conserved residues Ser 107 – Thr 108 – Ser 109 and the purine ring of A26: an analogous stacking interaction, involving A642, is present also in the complexes between the S8 protein and its natural RNA interactors, where it represents the only base-specific contacts of the complex[99]. All other geometric and energetic analyses performed on the trajectory are summarized in Figures S6-S8 (Supplementary Materials).

As for the previous case, we once again retrospectively analyzed all trajectories using the same metrics utilized before (electrostatic interaction energy and MMGBSA interaction energy), to assess the predictive power of the method. As reported in Figure S20, also in this case both metrics can successfully distinguish the native and native-like poses from the decoys. This observation further supports the idea that either of the two metrics could be utilized in a prospective application of SuMD to rank poses coming from different simulations to prioritize the most similar to the native binding mode.

RNA aptamer bound to human factor Xa (PDB ID: 5VOE)

One of the key events of the coagulation cascade is represented by the formation of the prothrombinase complex, macromolecular machinery formed by the serine protease factor Xa (FXa) and its cofactor factor Va (FVa)[100,101]: this membrane-mediated interaction enhances the catalytic activity of the FXa, leading to increased conversion of prothrombin into thrombin by a factor of approximately 10^5 and is, thereby, an interesting target for anticoagulation therapy[102,103]. Despite a great effort devoted to the development of both small-molecule inhibitors of the FXa catalytic site and peptide inhibitors directed at epitopes on the binding interface with FVa, both these approaches have led to disappointing therapeutic results since the interaction surface is large and involves multiple hotspots[104,105] and the inhibition at the catalytic site-level interferes with natural regulatory agents such as antithrombin III[106].

To avoid the limitations of traditional small molecule ligands, in 2010 Buddai et. al. developed RNA_{11f7t}, an RNA aptamer that exerts a potent anticoagulant effect by binding to FXa with high affinity (Kd 1.1 ± 0.2 nM) and selectivity (~3000 fold over other coagulation proteases) and inhibiting its interaction with FVa[107].

In 2018, Gunaratne et al. were able to solve the crystal structure of 11F7t bound to FXa, allowing us to better comprehend the key structural features that characterize this interaction[108]. Specifically, in agreement with previous biochemical data, the analysis of the structure confirmed that the interaction occurs at a protein site which is implicated in the binding of both anticoagulant drug heparin and coagulation factor Factor Va (FVa)[109], with the interaction surface involving a central aptamer loop formed by residues C8, A10, A21, and C28-C30 and a protease exosite formed by Leu 59, Arg 64, Val 88, Ile 89, Asn 92, Arg 93, Lys 236, and Arg 240[108]. The absence of notable structural alterations of the protease upon aptamer binding, the therapeutical relevance of the target, and the relatively contained size of the aptamer (36 residues) induced us to consider it to validate the SuMD protocol. As in the case of structure 3DD2, 2’ fluoro-modified nucleotides were retro-mutated to the corresponding natural alternatives, and the presence of the two Mg^{2+} ions was not considered in the simulations.

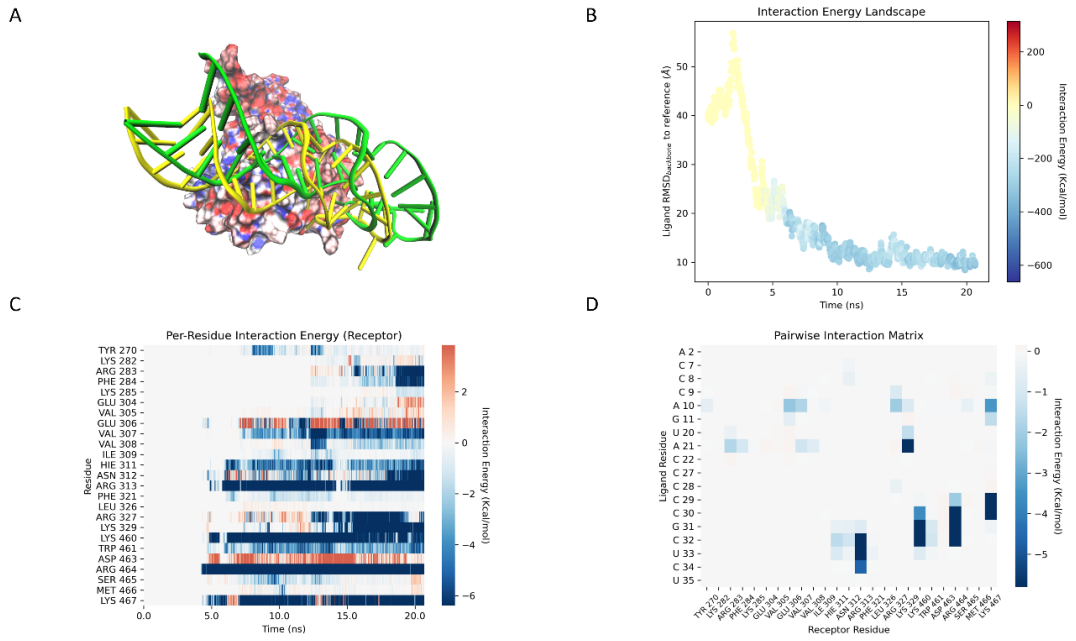


Figure 3. This panel encompasses the putative recognition pathway between the 11F7t RNA aptamer and human factor Xa described by the best trajectory obtained through the SuMD protocol (the one with the lowest RMSD to the crystal reference). (A) Visual representation of the RNA aptamer conformation sampled in the last frame of the SuMD trajectory (green) superposed with the native RNA aptamer conformation (yellow), as found in the crystal structure deposited in the Protein Data Bank with accession code 5VOE ($RMSD_{SuMD\text{-}Crystal}$: 9.12 Å). The aptamer is represented as a ribbon, while the protein is represented as a Connolly surface colored according to the electrostatic potential as calculated with the APBS software[90], where red indicates a negatively charged area while blue indicates a positively charged one. (B) Profile of the ligand-receptor interaction energy (defined as the sum of the electrostatic and van der Waals contribution) throughout the recognition process as a function of both the simulation time and the RMSD between the ligand position during the trajectory and the ligand position in the crystal. (C) Receptor per-residue decomposition of the receptor-ligand interaction energy throughout the SuMD trajectory as a function of the simulation time: the 25 most-contacted residues are reported in the plot. (D) Per-residue interaction energy matrix: the 25 most-contacted residues for both the receptor and the ligand are considered, while each square composing the heatmap represents the average value of the interaction energy between the two paired residues alongside the trajectory.

As can be deemed by Video 3 (Supplementary Materials), in this case about 20 ns of simulation time were sufficient to sample a presumptive association pathway between the RNA aptamer 11F7t and human coagulation factor Xa. As can be noticed in Figure 3 (panel A), in this case, the geometric accuracy of the SuMD protocol in reproducing the crystal complex was worse compared to the first two cases, as also denoted by the 9.12 Å

RMSD value between the final state of the simulation and the crystal reference. A first explanation of the lower geometric accuracy of the method can be found in the analysis of the classic MD simulation performed on both the crystal complex and the free aptamer: as highlighted by Figure S14 (Supplementary Materials) and by the RMSD of the nucleic backbone in the crystal complex MD (5.98 Å) and the free aptamer MD (4.93 Å), this aptamer has a significantly higher degree of conformational freedom compared to the previous two cases, which increase the difficulty of sampling the binding event due to the reduced amount of time that the aptamer spends in the binding competent conformation. Intriguingly, in this case, also, molecular docking compared worse than in the previous one, with the most correct pose having an RMSD of 3.17 Å to the crystal reference, indicating that the structural flexibility of the object (which is not considered by molecular docking) is not the only determinant of the performance of the protocol.

Contrary to the expectations, SuMD converged quite well from an interactive point of view with the experimental data: as can be seen in Figure S15 (Supplementary Materials), SuMD can qualitatively retrieve most of the native crystal interactions even if, as in the case of complex 4PDB, the estimation of the relative interaction strength is not always congruent.

Specifically, as can be noticed in Figure 3 panel D, the SuMD protocol can correctly retrieve the pivotal part played by residues A10, A21, and C29-C30, while slightly missing out on the importance of contacts with residues C8 and C28. On the protein side, as illustrated by Figure 3 panel C, SuMD precisely describes the central role portrayed by residues Arg 64 (283), Val 88 (308), Ile 89 (309), Asn 92 (312), Arg 93 (313), Lys 236 (460) and Arg 240 (464), while only passing up on the interaction with Leu 59 (278). Interestingly, Arg 240 (464) and Lys 236 (460) are key residues for the binding of heparin, according to mutagenesis studies[110]. Finally, SuMD analysis highlights how Arg 165 (387) and Lys 169 (391), two critical residues in the recognition of factor FXa by either factor Va and/or prothrombin, are not contacted during the trajectory, in agreement with both the crystal structure and previous observations which pointed out to the possibility that the abrogation of factor Va binding happened through an indirect effect rather than through occlusion of the interaction surface[107,108]. All other geometric and energetic analyses

performed on the trajectory are summarized in Figures S11-S13 (Supplementary Materials).

Finally, encouraged by the promising insights provided by the first two cases, we retrospectively analyzed all trajectories using the same scoring metrics defined before, to establish if they were once again able to distinguish the native and native-like poses from the decoys. As can be noticed in Figure S21, disappointingly both metrics fail to prioritize the most geometrically accurate solution, preferring instead the second-best one. Curiously, the reference crystal is also scored poorly by both metrics, indicating even if the SuMD protocol would have been able to sample it we would not have been able to prioritize it. As can be noticed in Table S1, this case also HADDOCK fails to rank the most geometrically accurate pose as the top solution, attributing to it a lower rank than a completely incorrect pose (RMSD: 19.79 Å), which ranks as the second best one. The observation that HADDOCK, despite incorporating information about native contacts in its scoring protocol, has trouble in correctly ranking poses for this case, combined with the intrinsic instability of the crystal complex, as indicated by the high RMSD value (5.98 Å) in the classic MD simulations and the low interaction energy values attributed by both metrics, indicate how this case might be an outlier, thus justifying the hypothesis of using the previously proposed metrics in prospective applications of the SuMD protocol.

RNA aptamer bound to SARS-CoV-2 Spike glycoprotein Receptor Binding Domain (RBD)

The outbreak of the COVID-19 pandemic in December 2019 caused an unprecedented worldwide public health crisis, leading to the death of more than six million people all over the world[111,112]. This illness is caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), a betacoronavirus able to infect human cells by expressing a surface glycoprotein known as spike (S) glycoprotein, which interacts through its receptor-binding domain (RBD) with the human angiotensin-converting enzyme 2 (hACE2) that mediates the viral uptake process in conjunction with the associated transmembrane serine protease 2 (TMPRSS2)[113]. Due to the central role that this interaction plays in SARS-CoV-2 infectivity, the vast majority of therapeutic and prophylactic efforts in contrasting the COVID-19 pandemic have therefore been directed towards the inhibition of the spike-

hACE2 interaction, either through vaccination or administration of monoclonal antibodies[114,115].

To overcome the main disadvantages of monoclonal antibodies, such as their high production costs, poor room temperature stability, and immunogenicity, in 2021 Valero et al. performed a SELEX experiment for the identification of a serum-stable RNA aptamer that could tightly bind the RBD of SARS-CoV-2 spike protein preventing the interaction with hACE2 thereby neutralizing viral infectivity[59]. This experiment led to the identification of RBD-PB6, an elongated stem-loop RNA aptamer that can selectively interact with the RBD with low nanomolar affinity ($K_D \approx 18$ nM), inhibiting the binding of RBD to hACE2 in a concentration-dependent manner[59].

Considering the encouraging results shown by the three applications of the SuMD protocol to the study of recognition processes between RNA aptamers and proteins and to illustrate a possible application of the protocol in a prospective scenario, we used SuMD simulations to shed light on the possible association pathway between the RBD-PB6 aptamer and the SARS-CoV-2 spike RBD. A model of the RNA aptamer structure was obtained through the 3dRNA webserver, based on the input primary sequence retrieved from the original publication and on the secondary structure prediction by the NUPACK webserver, while the structure of the SARS-CoV-2 spike RBD was retrieved from the crystal structure of the complex between the RBD and hACE2, deposited in the PDB with accession code 6M0J. The RBD structure that is present in this crystal comprises protein residues ranging from Thr 333 to Gly 526, slightly shorter than the construct used in the SELEX experiment which included residues from Arg 319 to Asn 532. However, residues that are not experimentally solved in the crystal structure are on the opposite side relative to the hACE2 interaction interface, so their absence should not impact the validity of the simulation. In agreement with experimental data that indicate how the RBD-PB6 RNA aptamer and hACE2 compete for the same binding site on the RBD surface, SuMD simulations were carried out to sample a putative recognition mechanism between RBD-PB6 and RBD surface that is responsible for interaction with hACE2.

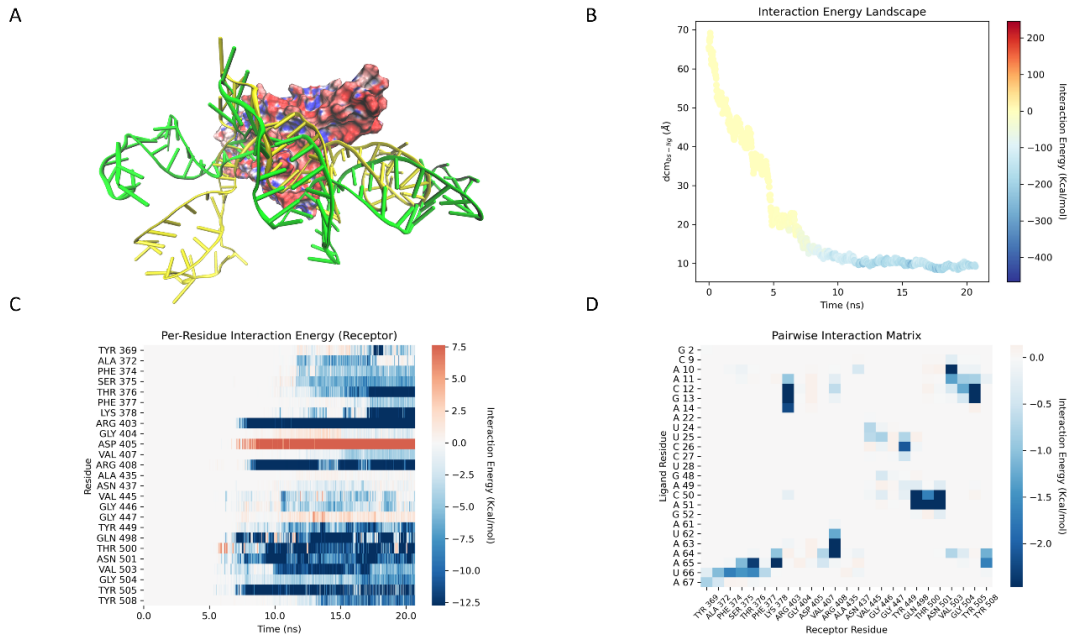


Figure 4. This panel encompasses the putative recognition pathway between the RBD-PB6 RNA aptamer and the SARS-CoV-2 Spike RBD described by the best trajectory obtained through the SuMD protocol according to the MMGBSA interaction energy. (A) Visual representation of the RNA aptamer conformation sampled in the last frame of the SuMD trajectory (green) superposed with the best docking-predicted RNA aptamer conformation (yellow). The aptamer is represented as a ribbon, while the protein is represented as a Connolly surface colored according to the electrostatic potential as calculated with the APBS software[90], where red indicates a negatively charged area while blue indicates a positively charged one. (B) Profile of the ligand-receptor interaction energy (defined as the sum of the electrostatic and van der Waals contribution) throughout the recognition process as a function of both the simulation time and the RMSD between the ligand position during the trajectory and the ligand position in the crystal. (C) Receptor per-residue decomposition of the receptor-ligand interaction energy throughout the SuMD trajectory as a function of the simulation time: the 25 most-contacted residues are reported in the plot. (D) Per-residue interaction energy matrix: the 25 most-contacted residues for both the receptor and the ligand are considered, while each square composing the heatmap represents the average value of the interaction energy between the two paired residues alongside the trajectory.

As can be observed in Video 4 (Supplementary Materials), about 20 ns of simulation time was enough to sample a putative recognition mechanism between the RBD-PB6 aptamer and the SARS-CoV-2 spike RBD. Interestingly, as can be noticed in Figure 4 panel A, there is a discrete level of convergence between the final state of the simulation and the docking-predicted binding mode, especially in the region ranging from A10 to A61, which is the minimal portion of the full-length aptamer which fully retains its binding capabilities

to the RBD ($RMSD_{backbone\ SuMD\text{-}docking}$: 11.69 Å). Regarding the recognition mechanism proposed by the SuMD protocol, an analysis of the interaction pattern reveals that the first contacts on the aptamer side involve residues A10-A14 and G48-G52, all of which fall into the conserved aptamer moiety that is required for binding to RBD. After these initial contacts take place and steer the binding event, other ancillary stabilizing interactions occur, such as the one with residues U24-U26 and the one involving residues U62-U66. On the protein side, instead, the main contacts involve polar and charged residues, such as Thr 376, Lys 378, Arg 403, Arg 408, Tyr 449, Gln 498, Thr 500, Asn 501, and Tyr 505. Interestingly, the most contacted residues during the trajectory do not include either Lys 417 or Glu 484, which are involved in the mutations K417N and E484K that characterize most viral variants with augmented infectivity compared to the wild-type virus. This evidence is in agreement with experimental data showing the RBD-PB6 affinity for the spike protein is practically unaffected by these mutations[59]. On the contrary, SuMD simulations indicate Asn 501 (involved in the N501Y mutation) as one of the most important residues in the recognition process: contrary to other previously mentioned residues, Asn 501 is surrounded by other interacting residues that can be found both in the complex between hACE2 and RBD and in the final state of the SuMD simulation, such as Gln 498, Thr 500 and Tyr 505, which could justify the contained impact of this mutation on the binding affinity of the aptamer compared to the other two. Interestingly, the most recent viral variants of concern include mutations such as Q498R, Y505H, and D405N that increase the positive charge on the spike surface: based on the analysis of the interaction pattern predicted by SuMD (Figure 4, panel C), all these mutations should increase the affinity with the negatively charged RBD-PB6 RNA aptamer, not only justifying its affinity towards the alpha and beta variant of SARS-CoV-2 but also towards the one from omicron on[116,117]. All other geometric and energetic analyses performed on the trajectory are summarized in Figures S16-S18 (Supplementary Materials).

Discussion

In this scientific work, we presented the first-ever application of Supervised Molecular Dynamics (SuMD) to the study of recognition processes between RNA macromolecules and proteins. Specifically, we concentrated our efforts on the aptamer class, due to their relevance as both therapeutic and diagnostic tools.

Three different retrospective case studies, where the structure of the RNA aptamer-protein complex was available, were presented to validate the ability of the SuMD protocol to reproduce the experimental data. In all three cases, despite the intrinsic limitations derived from the relatively high conformational freedom of the ribonucleic ligands, SuMD was able to converge quite well, both from a geometric and interactive point of view, with the experimentally solved complex structures. The lower geometric accuracy did not impair the ability to retrieve most if not all the binding determinants of the native complex, with the increased RMSD compared to docking being related to the portion of the ligand not directly involved in the binding interface. Despite the increased complexity of the considered system compared to the usually investigated protein-small molecule complexes, the simulation times required to sample a putative recognition mechanism between the RNA aptamers and their protein target were comparable: in all presented cases, indeed, 10-20 ns of simulation time were sufficient to capture the entire association pathway, from the unbound state to the final complex. The reduced computational effort that the SuMD platform provides compared to classic, unsupervised, molecular dynamics simulation, makes it more suitable for its implementation in a drug discovery pipeline, flanking and complementing the role of already established approaches such as molecular docking. Due to the limited sampling capability of molecular dynamics-based methods compared to molecular docking, now the optimal strategy would be to combine the two techniques rather than using them in a mutually exclusive fashion: the rapidity of molecular docking could be useful to generate a series of a reasonable binding hypothesis that could be then more thoroughly investigated through SuMD simulations. The rapid increase in computational power available to scientists will hopefully make it possible to solve the sampling issue of MD-based

methods, allowing them to fully replace more physically approximate methods such as molecular docking.

Concerning the applicability of the SuMD protocol in a prospective scenario, we also presented a case study where the experimental structure of the RNA aptamer-protein complex was not experimentally solved. Particularly, due to its therapeutic relevance, we decided to investigate the recognition process between the RBD-PB6 RNA aptamer, developed by Valero et al. in 2021[59], and the SARS-CoV-2 spike protein receptor-binding domain RBD. We showed how, even in the absence of an experimentally solved structure of the aptamer, the SuMD protocol can be coupled with structure prediction tools to give a structural prediction congruent with experimental evidence.

Regarding the ability of the SuMD protocol to be applied in a prospective scenario, one crucial point regards the capability to discriminate between the native-like binding mode and decoys. In the three presented cases, the simulation that was carefully analyzed and discussed in the manuscript was the one that presented the best geometric agreement with the crystal reference but such a metric could not be applied to a prospective investigation of complexes whose structures have not already been solved. By retroactively analyzing the geometric and energetic profile of the generated trajectories, we noticed that the electrostatic component of the interaction energy plays a fundamental role in steering the association process. Particularly, as can be noticed in Figures S19-S21 (Supplementary Materials), both the MMGBSA interaction energy and the electrostatic component on its own can discriminate and prioritize the native complex and native-like poses from decoys in two out of three case studies. The only exception to this rule is represented by the complex between the RNA aptamer 11F7t and human factor Xa (PDB ID: 5VOE), for which both the electrostatic and MMGBSA scoring metrics indicate the second most geometrically accurate pose as the one with the most favorable energetic profile. Interestingly, in this case, molecular docking also fails to prioritize the most geometrically accurate pose, ranking it as the third-best one (Table S1, Supplementary Materials). The indication that docking and SuMD, despite using different scoring metrics, failed to prioritize the native-like conformation suggests that there is still room for improvement regarding the scoring of complexes involving nucleic acids. However, despite this, the use of MMGBSA and or electrostatic interaction energy as

scoring metrics can still be relatively accurate in suggesting reasonable binding mode hypotheses that are congruent with experimental evidence, as also previously pointed out by a benchmark study by Chen et al. from 2018[118]. Due to these considerations, we opted for using MMGBSA as a scoring metric for our prospective study of the interaction between the RNA aptamer RBD-PB6 and the SARS-CoV-2 spike RBD.

The last aspect that is worth addressing is related to the choice of the residues to consider for the supervision of the association process throughout the SuMD trajectory. As is the case for molecular docking, where the binding site is defined by the user through a sphere or a box wrapped around the area of interest, SuMD also requires the user to specify a residue selection on both the receptor and ligand sides that are used to compute the distance between the center of mass of the binding site and of the ligand that is fed to the supervision algorithm. The choice of residues is usually based on prior knowledge of the interaction site derived from experimental evidence, but there could be some cases where this choice is not obvious. A first possible solution to this problem is represented by the analysis of the electrostatic potential of the receptor surface: as can be seen in Video 1-4, the recognition between the RNA aptamers and their protein targets usually involves a high level of complementarity of electrostatic properties, with the negatively charged ribonucleic surface being nicely harbored by positively charged patches on the protein side. A second possible solution is to perform a docking calculation to have a first indication of the preferable binding mode of the object, followed by a more extensive characterization of the binding mode through SuMD simulations. This solution was used in the context of this article for the study of the interaction between the RBD-PB6 RNA aptamer and the SARS-CoV-2 spike RBD.

The possibility to investigate different binding sites and binding mode hypotheses can also be viewed as a strong point of the SuMD technique: for example, in the case of complex ribosomal protein S8, two different RNA recognition sites are available for the aptamer, specifically the site involved in the interaction with helix 21 and the site that mediates interaction with helix 25[99]. SuMD simulations would allow investigating both possibilities at the same time, elucidating the mechanistic details that determine the preferential recognition of the primary binding site thereby helping the rational development of selective binders.

Furthermore, concerning the exploration of different binding hypotheses, SuMD would allow deciphering the possibility of alternative stoichiometries. For example, in the case of aptamer 11F7t, Buddai et al. noticed a peculiar and difficult to rationalize binding stoichiometry, other than a strong Ca^{2+} dependence of the interaction[107]. The authors discussed various possibilities, including a possible effect on the protein and/or aptamer structure, but also a possible calcium-induced aptamer dimerization[107]. In this case, the exploitation of the SuMD technique would have allowed the exploration of all these different hypotheses, which could not be investigated through static, time-independent techniques such as molecular docking.

References

1. Gilbert, W. Origin of life: The RNA world. *Nature 1986 319:6055* **319**, 618–618 (1986).

2. Breaker, R. R. & Joyce, G. F. The Expanding View of RNA and DNA Function. *Chem Biol* **21**, 1059–1065 (2014).

3. Hangauer, M. J., Vaughn, I. W. & McManus, M. T. Pervasive Transcription of the Human Genome Produces Thousands of Previously Unidentified Long Intergenic Noncoding RNAs. *PLoS Genet* **9**, (2013).

4. Cheetham, S. W., Gruhl, F., Mattick, J. S. & Dinger, M. E. Long noncoding RNAs and the genetics of cancer. *Br J Cancer* **108**, 2419–2425 (2013).

5. Esteller, M. Non-coding RNAs in human disease. *Nat Rev Genet* **12**, 861–874 (2011).

6. Morris, K. v. & Mattick, J. S. The rise of regulatory RNA. *Nature Reviews Genetics 2014 15:6* **15**, 423–437 (2014).

7. Connelly, C. M., Moon, M. H. & Schneekloth, J. S. The Emerging Role of RNA as a Therapeutic Target for Small Molecules. *Cell Chem Biol* **23**, 1077–1090 (2016).

8. Bartel, D. P. MicroRNAs: Genomics, Biogenesis, Mechanism, and Function. *Cell* **116**, 281–297 (2004).

9. Ponting, C. P., Oliver, P. L. & Reik, W. Evolution and Functions of Long Noncoding RNAs. *Cell* **136**, 629–641 (2009).

10. Salmon, L., Yang, S. & Al-Hashimi, H. M. Advances in the Determination of Nucleic Acid Conformational Ensembles. *http://dx.doi.org/10.1146/annurev-physchem-040412-110059* **65**, 293–316 (2014).

11. Cruz, J. A. & Westhof, E. The dynamic landscapes of RNA architecture. *Cell* **136**, 604–609 (2009).

12. Bartel, D. P. MicroRNAs: Target Recognition and Regulatory Functions. *Cell* **136**, 215–233 (2009).

13. Draper, D. E. Protein-RNA recognition. *Annu Rev Biochem* **64**, 593–620 (1995).

14. Lorger, M., Engstler, M., Homann, M. & Göringer, H. U. Targeting the variable surface of African trypanosomes with variant surface glycoprotein-specific, serum-stable RNA aptamers. *Eukaryot Cell* **2**, 84–94 (2003).

15. Disney, M. D. Targeting RNA with Small Molecules To Capture Opportunities at the Intersection of Chemistry, Biology, and Medicine. *J Am Chem Soc* **141**, 6776–6790 (2019).

16. Stoltenburg, R., Reinemann, C. & Strehlitz, B. SELEX—A (r)evolutionary method to generate high-affinity nucleic acid ligands. *Biomol Eng* **24**, 381–403 (2007).

17. Keefe, A. D., Pai, S. & Ellington, A. Aptamers as therapeutics. *Nature Reviews Drug Discovery 2010 9:7* **9**, 537–550 (2010).

18. Jayasena, S. D. Aptamers: An Emerging Class of Molecules That Rival Antibodies in Diagnostics. *Clin Chem* **45**, 1628–1650 (1999).

19. Jones, S., Daley, D. T. A., Luscombe, N. M., Berman, H. M. & Thornton, J. M. Protein–RNA interactions: a structural analysis. *Nucleic Acids Res* **29**, 943–954 (2001).

20. Morozova, N., Myers, J. & Shamoo, Y. Protein–RNA interactions: exploring binding patterns with a three-dimensional superposition analysis of high resolution structures. *Bioinformatics* **22**, 2746–2752 (2006).

21. Reynolds, A. *et al.* Rational siRNA design for RNA interference. *Nat Biotechnol* **22**, 326–330 (2004).

22. Boese, Q. *et al.* Mechanistic insights aid computational short interfering RNA design. *Methods Enzymol* **392**, 73–96 (2005).

23. Meng, X.-Y., Zhang, H.-X., Mezei, M. & Cui, M. Molecular Docking: A Powerful Approach for Structure-Based Drug Discovery. *Current Computer Aided-Drug Design* **7**, 146–157 (2012).

24. Kuntz, I. D., Blaney, J. M., Oatley, S. J., Langridge, R. & Ferrin, T. E. A geometric approach to macromolecule-ligand interactions. *J Mol Biol* **161**, 269–288 (1982).

25. Dominguez, C., Boelens, R. & Bonvin, A. M. J. J. HADDOCK: A protein-protein docking approach based on biochemical or biophysical information. *J Am Chem Soc* **125**, 1731–1737 (2003).

26. Ciemny, M. *et al.* Protein–peptide docking: opportunities and challenges. *Drug Discov Today* **23**, 1530–1537 (2018).

27. Pedotti, M., Simonelli, L., Livoti, E. & Varani, L. Computational Docking of Antibody-Antigen Complexes, Opportunities and Pitfalls Illustrated by Influenza Hemagglutinin. *Int J Mol Sci* **12**, 226 (2011).

28. David Morley, S. & Afshar, M. Validation of an empirical RNA-ligand scoring function for fast flexible docking using RiboDock. *J Comput Aided Mol Des* **18**, 189–208 (2004).

29. Nithin, C., Ghosh, P. & Bujnicki, J. M. Bioinformatics Tools and Benchmarks for Computational Docking and 3D Structure Prediction of RNA-Protein Complexes. *Genes 2018, Vol. 9, Page 432* **9**, 432 (2018).

30. Disney, M. D. Targeting RNA with Small Molecules To Capture Opportunities at the Intersection of Chemistry, Biology, and Medicine. *J Am Chem Soc* **141**, 6776 (2019).

31. Fulle, S. & Gohlke, H. Molecular recognition of RNA: challenges for modelling interactions and plasticity. *J Mol Recognit* **23**, 220–231 (2010).

32. Hermann, T. Rational ligand design for RNA: the role of static structure and conformational flexibility in target recognition. *Biochimie* **84**, 869–875 (2002).

33. De Vivo, M., Masetti, M., Bottegoni, G. & Cavalli, A. Role of Molecular Dynamics and Related Methods in Drug Discovery. *J Med Chem* **59**, 4035–4061 (2016).

34. Bernardi, R. C., Melo, M. C. R. & Schulten, K. Enhanced sampling techniques in molecular dynamics simulations of biological systems. *Biochimica et Biophysica Acta (BBA) - General Subjects* **1850**, 872–877 (2015).

35. Sabbadin, D. & Moro, S. Supervised molecular dynamics (SuMD) as a helpful tool to depict GPCR-ligand recognition pathway in a nanosecond time scale. *J Chem Inf Model* **54**, 372–376 (2014).

36. Deganutti, G., Moro, S., Ciruela, F. & Sotelo, E. Supporting the Identification of Novel Fragment-Based Positive Allosteric Modulators Using a Supervised Molecular Dynamics Approach: A Retrospective Analysis Considering the Human A2A Adenosine Receptor as a Key Example. *Molecules 2017, Vol. 22, Page 818* **22**, 818 (2017).

37. Bolcato, G., Bissaro, M., Sturlese, M. & Moro, S. Comparing Fragment Binding Poses Prediction Using HSP90 as a Key Study: When Bound Water Makes the Difference. *Molecules 2020, Vol. 25, Page 4651* **25**, 4651 (2020).

38. Ferrari, F. *et al.* HT-SuMD: Making Molecular Dynamics Simulations Suitable for Fragment-Based Screening. a Comparative Study with NMR. (2020) doi:10.26434/CHEMRXIV.12582662.V1.

39. Bissaro, M. *et al.* Inspecting the Mechanism of Fragment Hits Binding on SARS-CoV-2 M^{pro} by Using Supervised Molecular Dynamics (SuMD) Simulations. *ChemMedChem* **16**, 2075–2081 (2021).

40. Bolcato, G. *et al.* A Computational Workflow for the Identification of Novel Fragments Acting as Inhibitors of the Activity of Protein Kinase CK1δ. *Int J Mol Sci* **22**, 9741 (2021).

41. Grosjean, H. *et al.* SAMPL7 protein-ligand challenge: A community-wide evaluation of computational methods against fragment screening and pose-prediction. *Journal of Computer-Aided Molecular Design 2022* 1–21 (2022) doi:10.1007/S10822-022-00452-7.

42. Bolcato, G., Bissaro, M., Pavan, M., Sturlese, M. & Moro, S. Targeting the coronavirus SARS-CoV-2: computational insights into the mechanism of action of the protease inhibitors lopinavir, ritonavir and nelfinavir. *Sci Rep* **10**, 20927 (2020).

43. Pavan, M., Bolcato, G., Bassani, D., Sturlese, M. & Moro, S. Supervised Molecular Dynamics (SuMD) Insights into the mechanism of action of SARS-CoV-2 main protease inhibitor PF-07321332. *J Enzyme Inhib Med Chem* **36**, 1646–1650 (2021).

44. Bolcato, G., Pavan, M., Bassani, D., Sturlese, M. & Moro, S. Ribose and Non-Ribose A2A Adenosine Receptor Agonists: Do They Share the Same Receptor Recognition Mechanism? *Biomedicines 2022, Vol. 10, Page 515* **10**, 515 (2022).

45. Hassankalhori, M., Bolcato, G., Bissaro, M., Sturlese, M. & Moro, S. Shedding Light on the Molecular Recognition of Sub-Kilodalton Macrocyclic Peptides on Thrombin by Supervised Molecular Dynamics. *Front Mol Biosci* **8**, 730 (2021).

46. Salmaso, V., Sturlese, M., Cuzzolin, A. & Moro, S. Exploring Protein-Peptide Recognition Pathways Using a Supervised Molecular Dynamics Approach. *Structure* **25**, 655-662.e2 (2017).

47. Bissaro, M. *et al.* Targeting Protein Kinase CK1δ with Riluzole: Could It Be One of the Possible Missing Bricks to Interpret Its Effect in the Treatment of ALS from a Molecular Point of View? *ChemMedChem* **13**, 2601–2605 (2018).

48. Panday, S. K., Sturlese, M., Salmaso, V., Ghosh, I. & Moro, S. Coupling Supervised Molecular Dynamics (SuMD) with Entropy Estimations to Shine Light on the Stability of Multiple Binding Sites. *ACS Med Chem Lett* **10**, 444–449 (2019).

49. Deganutti, G., Cuzzolin, A., Ciancetta, A. & Moro, S. Understanding allosteric interactions in G protein-coupled receptors using Supervised Molecular Dynamics: A prototype study analysing the human A3 adenosine receptor positive allosteric modulator LUF6000. *Bioorg Med Chem* **23**, 4065–4071 (2015).

50. Paoletta, S. *et al.* Modeling ligand recognition at the P2Y12 receptor in light of X-ray structural information. *J Comput Aided Mol Des* **29**, 737–756 (2015).

51. Palazzotti, D. *et al.* Deciphering the Molecular Recognition Mechanism of Multidrug Resistance Staphylococcus aureus NorA Efflux Pump Using a Supervised Molecular Dynamics Approach. *International Journal of Molecular Sciences 2019, Vol. 20, Page 4041* **20**, 4041 (2019).

52. Cuzzolin, A. *et al.* Deciphering the Complexity of Ligand-Protein Recognition Pathways Using Supervised Molecular Dynamics (SuMD) Simulations. *J Chem Inf Model* **56**, 687–705 (2016).

53. Deganutti, G., Moro, S. & Reynolds, C. A. A Supervised Molecular Dynamics Approach to Unbiased Ligand-Protein Unbinding. *ACS Appl Mater Interfaces* **2020**, 1804–1817 (2020).

54. Bissaro, M., Sturlese, M. & Moro, S. Exploring the RNA-Recognition Mechanism Using Supervised Molecular Dynamics (SuMD) Simulations: Toward a Rational Design for Ribonucleic-Targeting Molecules? *Front Chem* **8**, 107 (2020).

55. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res* **28**, 235 (2000).

56. Berman, H. M. The Protein Data Bank. *Nucleic Acids Res* **28**, 235–242 (2000).

57. Molecular Operating Environment (MOE), 2019.01; Chemical Computing Group ULC, 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, 2021. https://www.chemcomp.com/Research-Citing_MOE.htm.

58. Lan, J. *et al.* Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature 2020 581:7807* **581**, 215–220 (2020).

59. Valero, J. *et al.* A serum-stable RNA aptamer specific for SARS-CoV-2 neutralizes viral entry. *Proc Natl Acad Sci U S A* **118**, (2021).

60. Zadeh, J. N. *et al.* NUPACK: Analysis and design of nucleic acid systems. *J Comput Chem* **32**, 170–173 (2011).

61. Wang, J., Wang, J., Huang, Y. & Xiao, Y. 3dRNA v2.0: An Updated Web Server for RNA 3D Structure Prediction. *Int J Mol Sci* **20**, (2019).

62. Humphrey, W., Dalke, A. & Schulten, K. VMD: Visual molecular dynamics. *J Mol Graph* **14**, 33–38 (1996).

63. Case, D. A. *et al.* The Amber biomolecular simulation programs. *J Comput Chem* **26**, 1668–1688 (2005).

64. Maier, J. A. *et al.* ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J Chem Theory Comput* **11**, 3696–3713 (2015).

65. Pérez, A. *et al.* Refinement of the AMBER Force Field for Nucleic Acids: Improving the Description of α/γ Conformers. *Biophys J* **92**, 3817–3829 (2007).

66. Zgarbová, M. *et al.* Refinement of the Cornell et al. Nucleic acids force field based on reference quantum chemical calculations of glycosidic torsion profiles. *J Chem Theory Comput* **7**, 2886–2902 (2011).

67. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J Chem Phys* **79**, 926–935 (1983).

68. Davidchack, R. L., Handel, R. & Tretyakov, M. v. Langevin thermostat for rigid body dynamics. *J Chem Phys* **130**, 234101 (2009).

69. Kräutler, V., van Gunsteren, W. F. & Hünenberger, P. H. A Fast SHAKE Algorithm to Solve Distance Constraint Equations for Small Molecules in Molecular Dynamics Simulations. *J Comput Chem* **22**, 501–508 (2001).

70. Essmann, U. *et al.* A smooth particle mesh Ewald method. *J Chem Phys* **103**, 8577 (1998).

71. Faller, R. & de Pablo, J. J. Constant pressure hybrid Molecular Dynamics–Monte Carlo simulations. *J Chem Phys* **116**, 55 (2001).

72. Harvey, M. J., Giupponi, G. & de Fabritiis, G. ACEMD: Accelerating biomolecular dynamics in the microsecond time scale. *J Chem Theory Comput* **5**, 1632–1639 (2009).

73. Eastman, P. *et al.* OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Comput Biol* **13**, e1005659 (2017).

74. Bakan, A., Meireles, L. M. & Bahar, I. ProDy: Protein dynamics inferred from theory and experiments. *Bioinformatics* **27**, 1575–1577 (2011).

75. Michaud-Agrawal, N., Denning, E. J., Woolf, T. B. & Beckstein, O. MDAnalysis: A toolkit for the analysis of molecular dynamics simulations. *J Comput Chem* **32**, 2319–2327 (2011).

76. Gowers, R. J. *et al.* MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations. *Proceedings of the 15th Python in Science Conference* 98–105 (2016) doi:10.25080/MAJORA-629E541A-00E.

77. Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Comput Sci Eng* **9**, 90–95 (2007).

78. Phillips, J. C. *et al.* Scalable molecular dynamics on CPU and GPU architectures with NAMD. *J Chem Phys* **153**, 044130 (2020).

79. van Zundert, G. C. P. *et al.* The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. *J Mol Biol* **428**, 720–725 (2016).

80. Jorgensen, W. L. & Tirado-Rives, J. The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *J Am Chem Soc* **110**, 1657–1666 (1988).

81. Rodrigues, J. P. G. L. M. *et al.* Clustering biomolecular complexes by residue contacts similarity. *Proteins: Structure, Function, and Bioinformatics* n/a-n/a (2012) doi:10.1002/prot.24078.

82. Stubbs, M. T. & Bode, W. The clot thickens: clues provided by thrombin structure. *Trends Biochem Sci* **20**, 23–28 (1995).

83. Hoffman, M. & Monroe, D. M. A cell-based model of hemostasis. *Thromb Haemost* **85**, 958–965 (2001).

84. di Cera, E. Thrombin Interactions. *Chest* **124**, 11S-17S (2003).

85. Rau, J. C., Beaulieu, L. M., Huntington, J. A. & Church, F. C. Serpins in thrombosis, hemostasis and fibrinolysis. *Journal of Thrombosis and Haemostasis* **5**, 102–115 (2007).

86. White, R. *et al.* Generation of Species Cross-reactive Aptamers Using "Toggle" SELEX. *Molecular Therapy* **4**, 567–573 (2001).

87. Long, S. B., Long, M. B., White, R. R. & Sullenger, B. A. Crystal structure of an RNA aptamer bound to thrombin. *RNA* **14**, 2504–2512 (2008).

88. Vangaveti, S., Ranganathan, S. v. & Chen, A. A. Advances in RNA molecular dynamics: a simulator's guide to RNA force fields. *Wiley Interdiscip Rev RNA* **8**, e1396 (2017).

89. Giambaşu, G. M., Case, D. A. & York, D. M. Predicting Site-Binding Modes of Ions and Water to Nucleic Acids Using Molecular Solvation Theory. *J Am Chem Soc* (2019) doi:10.1021/JACS.8B11474/ASSET/IMAGES/MEDIUM/JA-2018-11474K_M001.GIF.

90. Jurrus, E. *et al.* Improvements to the <scp>APBS</scp> biomolecular solvation software suite. *Protein Science* **27**, 112–128 (2018).

91. Jeter, M. L. *et al.* RNA aptamer to thrombin binds anion-binding exosite-2 and alters protease inhibition by heparin-binding serpins. *FEBS Lett* **568**, 10–14 (2004).

92. Poehlsgaard, J. & Douthwaite, S. The bacterial ribosome as a target for antibiotics. *Nature Reviews Microbiology 2005 3:11* **3**, 870–881 (2005).

93. Nierhaus, K. H. The assembly of the prokaryotic ribosome. *Biosystems* **12**, 273–282 (1980).

94. Nomura, M., Yates, J. L., Dean, D. & Post, L. E. Feedback regulation of ribosomal protein gene expression in Escherichia coli: Structural homology of ribosomal RNA and ribosomal protein mRNA. *Proc Natl Acad Sci U S A* **77**, 7084–7088 (1980).

95. Wu, H., Jiang, L. & Zimmermann, R. A. The binding site for ribosomal protein S8 in 16S rRNA and spc mRNA from Escherichia coli: minimum structural requirements and the effects of single bulged bases on S8-RNA interaction. *Nucleic Acids Res* **22**, 1687 (1994).

96. Merianos, H. J., Wang, J. & Moore, P. B. The structure of a ribosomal protein S8/spc operon mRNA complex. *RNA* **10**, 954–964 (2004).

97. Tishchenko, S. *et al.* Detailed analysis of RNA-protein interactions within the ribosomal protein S8-rRNA complex from the archaeon Methanococcus jannaschii. *J Mol Biol* **311**, 311–324 (2001).

98. Brodersen, D. E., Clemons, W. M., Carter, A. P., Wimberly, B. T. & Ramakrishnan, V. Crystal structure of the 30 S ribosomal subunit from Thermus thermophilus: structure of the proteins and their interactions with 16 S RNA. *J Mol Biol* **316**, 725–768 (2002).

99. Davlieva, M., Donarski, J., Wang, J., Shamoo, Y. & Nikonowicz, E. P. Structure analysis of free and bound states of an RNA aptamer against ribosomal protein S8 from Bacillus anthracis. *Nucleic Acids Res* **42**, 10795–10808 (2014).

100. Mann, K. G., Nesheim, M. E., Church, W. R., Haley, P. & Krishnaswamy, S. Surface-dependent reactions of the vitamin K-dependent enzyme complexes. *Blood* **76**, 1–16 (1990).

101. Mann, K. G., Jenny, R. J. & Krishnaswamy, S. COFACTOR PROTEINS IN THE ASSEMBLY AND EXPRESSION OF BLOOD CLOTTING ENZYME COMPLEXES. *https://doi.org/10.1146/annurev.bi.57.070188.004411* **57**, 915–956 (2003).

102. Gross, P. L. & Weitz, J. I. New anticoagulants for treatment of venous thromboembolism. *Arterioscler Thromb Vasc Biol* **28**, 380–386 (2008).

103. Bauer, K. A. New anticoagulants. *Curr Opin Hematol* **15**, 509–515 (2008).

104. Krishnaswamy, S. Prothrombinase complex assembly. Contributions of protein-protein and protein-membrane interactions toward complex formation. *Journal of Biological Chemistry* **265**, 3708–3718 (1990).

105. Krishnaswamy, S. Exosite-driven substrate specificity and function in coagulation. *Journal of Thrombosis and Haemostasis* **3**, 54–67 (2005).

106. Björk, I. & Olson, S. T. Antithrombin. *Adv Exp Med Biol* **425**, 17–33 (1997).

107. Buddai, S. K. *et al.* An Anticoagulant RNA Aptamer That Inhibits Proteinase-Cofactor Interactions within Prothrombinase *. *Journal of Biological Chemistry* **285**, 5212–5223 (2010).

108. Gunaratne, R. *et al.* Combination of aptamer and drug for reversible anticoagulation in cardiopulmonary bypass. *Nature Biotechnology 2018 36:7* **36**, 606–613 (2018).

109. Rezaie, A. R. Identification of Basic Residues in the Heparin-binding Exosite of Factor Xa Critical for Heparin and Factor Va Binding *. *Journal of Biological Chemistry* **275**, 3320–3327 (2000).

110. Rezaie, A. R. Identification of Basic Residues in the Heparin-binding Exosite of Factor Xa Critical for Heparin and Factor Va Binding. *Journal of Biological Chemistry* **275**, 3320–3327 (2000).

111. Guarner, J. Three Emerging Coronaviruses in Two Decades: The Story of SARS, MERS, and Now COVID-19. *American Journal of Clinical Pathology* vol. 153 420–421 Preprint at https://doi.org/10.1093/ajcp/aqaa029 (2020).

112. COVID Live - Coronavirus Statistics - Worldometer. https://www.worldometers.info/coronavirus/.

113. Hoffmann, M. *et al.* SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell* **181**, 271-280.e8 (2020).

114. Thanh Le, T. *et al.* The COVID-19 vaccine development landscape. *Nat Rev Drug Discov* **19**, 305–306 (2020).

115. Taylor, P. C. *et al.* Neutralizing monoclonal antibodies for treatment of COVID-19. *Nature Reviews Immunology 2021 21:6* **21**, 382–393 (2021).

116. Sartore, G. *et al.* In silico evaluation of the interaction between ACE2 and SARS-CoV-2 Spike protein in a hyperglycemic environment. *Sci Rep* **11**, (2021).

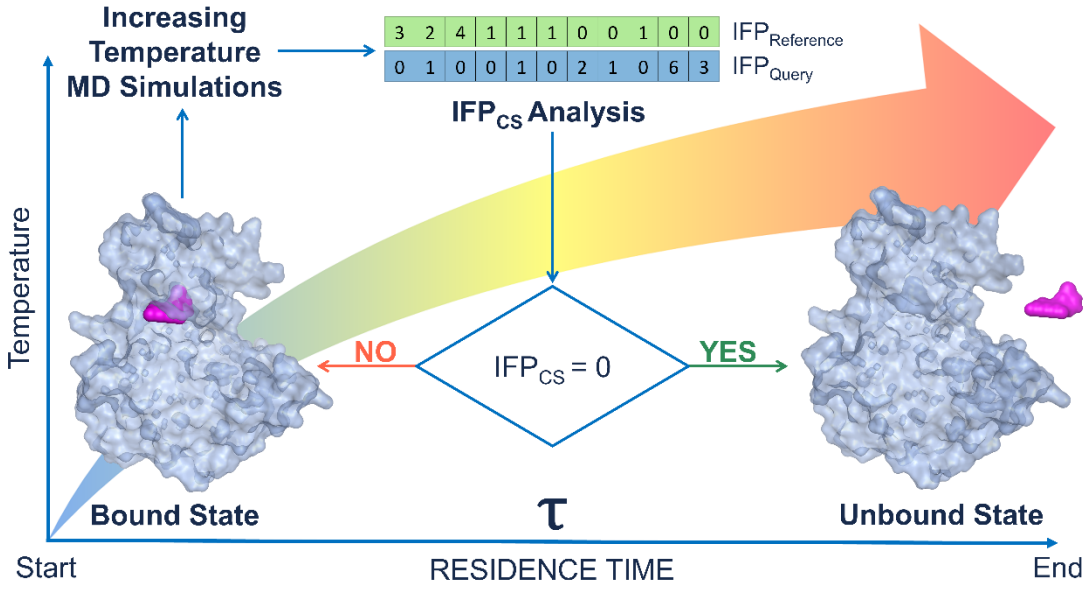
117. Bassani, D., Ragazzi, E., Lapolla, A., Sartore, G. & Moro, S. Omicron Variant of SARS-CoV-2 Virus: In Silico Evaluation of the Possible Impact on People Affected by Diabetes Mellitus. *Front Endocrinol (Lausanne)* **13**, (2022).

118. Chen, F. *et al.* Assessing the performance of MM/PBSA and MM/GBSA methods. 8. Predicting binding free energies and poses of protein-RNA complexes. *RNA* **24**, 1183–1194 (2018).

Qualitative Estimation of Protein-Ligand Complex Stability through Thermal Titration Molecular Dynamics (TTMD) Simulations

Matteo Pavan, Silvia Menin, Davide Bassani, Mattia Sturlese and Stefano Moro

Pavan, M., Menin, S., Bassani, D., Sturlese, M. & Moro, S. Qualitative Estimation of Protein–Ligand Complex Stability through Thermal Titration Molecular Dynamics Simulations. *J Chem Inf Model* (2022) doi:10.1021/ACS.JCIM.2C00995



Abstract

The prediction of ligand efficacy has long been linked to thermodynamic properties such as the equilibrium dissociation constant, which considers both the association and the dissociation rate of a defined protein-ligand complex. In the last fifteen years, there has been a paradigm shift, with an increased interest in the determination of kinetic properties such as the drug-target residence time, since they better correlate with ligand efficacy compared to other parameters. In this article, we present Thermal Titration Molecular Dynamics (TTMD), an alternative computational method that combines a series of molecular dynamics simulations performed at progressively increasing temperatures with a scoring function based on protein-ligand interaction fingerprints for the qualitative estimation of protein-ligand binding stability. The protocol has been applied to four different pharmaceutically relevant test cases, including protein kinase CK1δ, protein kinase CK2, pyruvate dehydrogenase kinase 2, and SARS-CoV-2 main protease, on a

variety of ligands with different sizes, structures, and experimentally determined affinity values. In all four cases, TTMD was successfully able to distinguish between high-affinity compounds (low nanomolar range) and low-affinity ones (micromolar), proving to be a useful screening tool for the prioritization of compounds in a drug discovery campaign.

Introduction

At the beginning of the 20th century, Paul Ehrlich's famous quote "*Corpora non agunt nisi fixata*" marked a pivotal moment in the history of modern pharmacology, molecular medicine, and drug development[1]. His statement, combined with independent observations about "receptive substances" by John Newport Langley[2], defined the birth of the receptor theory of drug action, which postulates that a drug can only work as long as it is bound to its target receptor[3,4].

Although the basic ideas of this cornerstone principle were formulated more than one hundred years ago[5], it was only in the 1970s that molecular receptors could be successfully isolated and purified[6,7]. This allowed for the development of different biochemical and cellular assays for the direct determination of the extent to which a drug is bound to its receptor under thermodynamic equilibrium conditions, i.e., the binding affinity[8,9]. Traditionally, this parameter is quantified either through the equilibrium dissociation constant (K_d) or through other proxy metrics such as the drug concentration responsible for the half-maximal inhibition/effect (IC_{50}/EC_{50}), and the inhibition constant (K_i)[10].

Although in principle, these measurements are all adequate predictors for *in vivo* efficacy, i.e., the capability of the drug to induce the desired response, they are all related to *in vitro* assays portrayed under closed system conditions[11]. Since in an open, *in vivo*, system the drug concentration is not a fixed variable and indeed varies over time because of various physiological processes, several authors thus suggested that the observables related to drug-receptor binding kinetics, such as the association (k_{on}) and dissociation (k_{off}) constant, could be better descriptors for drug efficacy[12–15]. Accordingly, while the binding affinity only depends on the free energy difference between the bound and

unbound state, which can be directly correlated to the K_d, association and dissociation rates depend on the energy barriers that separate those states[10].

Thermodynamics and kinetics of bindings are interlinked by the equation $K_d = \frac{k_{off}}{k_{on}}$[10]. While, in theory, both kinetic constants should equally contribute to the determination of the K_d, physicochemical and pharmacological limitations on the k_{on} value[16] render the in vivo duration of a receptor-ligand complex almost entirely dependent on the k_{off} value[11]. Based on this observation, Copeland et al. first suggested that the key determinant of in vivo pharmacological activity and duration is not the binding affinity but, instead, the lifetime of the receptor-ligand complex, defined as the residence time[12]. Furthermore, Copeland et al. proposed a mathematical formulation for the quantification of this parameter, defining it as the reciprocal of the k_{off} ($\tau=1/k_{off}$)[12].

From an experimental perspective, a plethora of methods for the determination of binding kinetics is available[17–20]. Each of them relies on monitoring the time-dependent evolution of a signal in response to the binding event[21]. The first strategy revolves around the radio-[22] and spectroscopic[20,23] labeling of ligands and includes techniques such as fluorescent resonance energy transfer (FRET)[24] and bioluminescence resonance energy transfer (BRET)[25]. An alternative approach revolves around the exploitation of label-free approaches such as surface plasmon resonance (SPR)[26,27], nuclear magnetic resonance (NMR)[28], surface acoustic wave method[29], and various declinations of isothermal titration calorimetry (ITC)[30,31]. Finally, another possible method is based on following enzymatic reactions, usually through the monitoring of spectroscopic parameters[32].

Alongside the aforementioned experimental protocols, various computational approaches exist than can flank and expand on the information that they provide by showcasing mechanistic information about the underlying process at an atomic level of detail[21,33,34]. Particularly, Molecular Dynamics (MD) simulations have been exploited to estimate thermodynamic properties such as the binding affinity for protein-ligand complexes, and due to the growing interest in the study of kinetics for drug discovery, they have recently been applied also to the estimation of kinetic properties[35]. Although it would theoretically be possible to exploit unbiased MD simulations for the determination

of kinetics observables, biologically relevant events such as drug-target unbinding occur at much longer timescales than that of typical MD simulations, heavily restricting its limitations in terms of computational resources' availability[36] and neglecting any real-world application of the technique[37]. For this reason, several different methods have been developed throughout the years that implement smart sampling strategies to reduce the required computational effort, such as various instances of metadynamics[38–41], which are based on repeatedly 'filling' the potential energy of the system by a sum of Gaussians centered along the trajectory followed by an appropriately chosen ensemble of collective variables (CVs)[42], scaled Molecular Dynamics (sMD), which relies on smoothing the potential energy surface (PES) by applying an appropriate scaling factor[43–45], and τ-Random Acceleration Molecular Dynamics (τ-RAMD), in which a small randomly oriented force vector is applied to the ligand[46–48].

In the present study, we present the first application of Thermal Titration Molecular Dynamics (TTMD), an alternative MD-based approach for the qualitative estimation of protein-ligand complex stability. The method relies on evaluating the conservation of the native binding mode for a ligand of interest throughout a series of molecular dynamics trajectories performed at progressively increasing temperature values. For validation purposes, the protocol has been applied to four different biomolecular targets of pharmaceutical interest: casein kinase 1δ (CK1δ), casein kinase 2 (CK2), pyruvate dehydrogenase kinase 2 (PDK2), and SARS-CoV-2 main protease (M^{pro}).

MATERIALS AND METHODS

Hardware Overview

Each general molecular modeling operation, such as the preparation of protein-ligand complex structures, the setup for molecular dynamics simulations, and trajectory analyses were conducted on a 20 CPU Linux workstation equipped with an Intel Core i9-9820X 3.3 GHz processor. All molecular dynamics simulations were carried out on a GPU cluster composed of 20 NVIDIA drivers ranging from GTX980 to RTX2080Ti.

Structures Preparation

The three-dimensional coordinates of the protein-ligand complexes used in this study were retrieved from the Protein Data Bank (PDB)[49] and processed before molecular dynamics simulations through several tools provided by the Molecular Operating Environment (MOE) 2019.01 suite[50]. Four different macromolecular targets were considered in this work: casein kinase 1δ (CK1δ), casein kinase 2 (CK2), pyruvate dehydrogenase kinase 2 (PDK2), and SARS-CoV-2 main protease (M^{pro}). For each macromolecular target, the considered structures are reported in Table 1.

CK1δ	3UZP[51]	4TN6[52]	5IH5[53]	5IH6[53]	5MQV[54]
CK2	2ZJW[55]	3H30[56]	3PE1[57]	3PE2[57]	6HOU[58]
PDK2	4MP2[59]	4V25[60]	5J71[61]	5M4M[62]	7EA0[63]
M^{pro}	6M2N[64]	7LTJ[65]	7M8P[66]	7M91[66]	7N44[67]

Table 1: List of the protein-ligand complex structures used in this work. Complexes are grouped by macromolecule target.

Each protein-ligand system was simulated in the monomeric form, except for SARS-CoV-2 M^{pro}, which was simulated in the dimeric form by applying a symmetric crystallographic transformation to each asymmetric unit. Firstly, all structures were pre-processed using the "Structure Preparation" tool, assigning alternates to the highest occupancy conformation, rebuilding missing loops through homology modeling, and correcting inconsistencies between the primary sequence and the tertiary structure. Secondly, the "Protonate3D" tool was exploited to add missing hydrogens to the system and to determine the most probable protonation state of titratable residues at pH = 7.4. Finally, every non-protein and non-ligand atom of the system was removed before saving the structure for further calculations, except for water molecules within 4.5 Å of the ligand that were not removed and were indeed considered in the simulations. Concerning the protonation state of the ligand, the most abundant protomer at pH 7.4 according to the "Protomers" tool was considered in the calculations, besides CK2 complex 2ZJW where two different protonation states were considered. Particularly, in the case of 2ZJW, the

predominant form at pH 7.4 should be the neutral, non-charged one. However, in the context of the binding pocket, the interaction network of the hydroxyl in position 3 (the one facing Lys68 and the conserved water molecule W1) suggests the prevalence of a monocharged, ionized form. Since experimental data published in the literature does not clarify the correct protonation state for ellagic acid in the context of CK2 recognition[55,68,69], we opted to consider both hypotheses equally relevant (50/50).

System Setup for Molecular Dynamics Simulations and Equilibration Protocol

Each protein-ligand complex prepared as described before was further processed through various tools from Visual Molecular Dynamics (VMD) 1.9.2[70] and the Ambertools14[71] suite. Protein atoms were parametrized through the ff14SB[72] force field, while the General Amber Force Field (GAFF)[73] was utilized to parametrize the ligands. Partial charges were attributed to the ligand through the AM1-BCC method[74]. Each investigated system was solvated in a cubic box with a padding of 15 Å utilizing the TIP3P[75] model for water molecules. The proper number of sodium and chloride ions were added to neutralize the system and reach a salt concentration of 0.154 M. Before undergoing Molecular Dynamics (MD) simulations, each system was energy minimized for a total of 500 steps with the conjugate-gradient algorithm to remove clashes and bad contacts.

Afterward, each minimized system was subjected to a two-step equilibration protocol. During the first stage, a 0.1 ns simulation in the canonical ensemble (NVT) was performed, with harmonic positional restraints (5 Kcal $mol^{-1}Å^{-2}$ force constant) applied on both protein and ligand atoms. The second stage, instead, consisted of a 0.5 ns simulation carried out in the isothermal-isobaric ensemble (NPT), applying the same restraints only to the ligand position and the protein backbone.

Each MD simulation presented in this work, both in the equilibration and the production stage, was performed using an integration timestep of 2 fs, keeping the temperature at a constant value of 310 K through a Langevin thermostat[76], constraining the length of bonds involving hydrogen bonds through M-SHAKE algorithm[77], exploiting the particle-mesh Ewald (PME)[78] method to compute electrostatic interactions using cubic spline interpolation and a 1 Å grid spacing, and setting a 9.0 Å cutoff for the calculation of

Lennard-Jones interactions. Simulations in the NPT ensemble were carried out keeping the pressure at a constant 1 atm value making use of a Monte Carlo barostat[79]. All MD simulations were run through the ACEMD 3[80] engine, which is based upon the open-source library for molecular simulations OpenMM 7[81].

Thermal Titration Molecular Dynamics (TTMD) simulations

Thermal Titration Molecular Dynamics (TTMD) is an alternative enhanced sampling molecular dynamics approach for the qualitative estimation of protein-ligand complex stability. The method relies on evaluating the conservation of the native binding mode for a ligand of interest throughout a series of molecular dynamics trajectories performed at progressively increasing temperature values. The protocol described herein is implemented as a Python 3.10 code, which relies on the Numpy, MDAnalysis[82,83], Open Drug Discovery Toolkit[84], and Scikit-learn libraries.

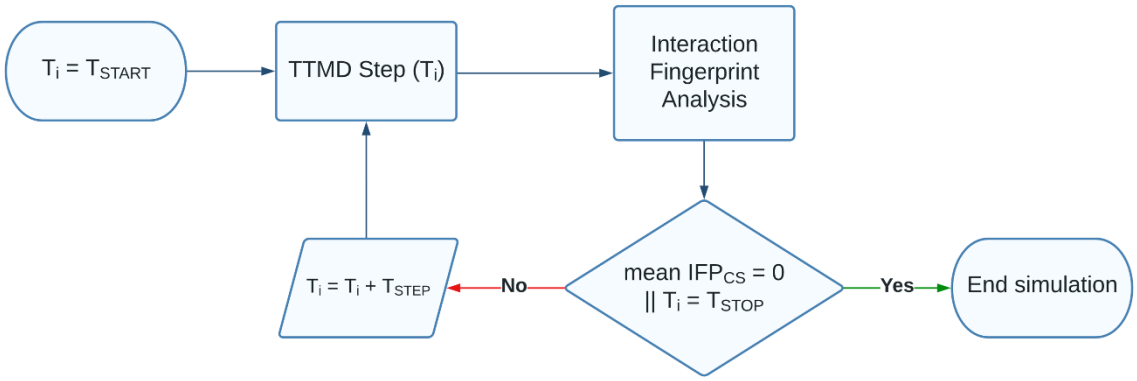


Figure 11: Computational workflow for a Thermal Titration Molecular Dynamics (TTMD) simulation.

The workflow for a TTMD simulation is reported in Figure 1 and detailed hereafter. In detail, the task is accomplished through a series of short, unbiased, MD simulations performed at different, progressively increasing, temperatures in the NVT ensemble with the ACEMD3 engine. For each TTMD run, the duration of each simulation window (defined as “TTMD-step”) is fixed and user-defined (10ns, in this case). The starting and final

temperature values as well as the temperature increase between each “TTMD-step” are also defined by the user based on prior knowledge of the target, particularly regarding the conservation of the protein fold at higher simulation temperatures (which, in the context of this article, is carried out by monitoring the protein backbone RMSD throughout the simulation). In this work, the starting temperature was set to 300 K, the ending temperature was set to 450 K, while the temperature increase between each “TTMD-step” was set to 10 K.

The progress of the simulation is monitored through a scoring function based on protein-ligand interaction fingerprints. The scoring function, defined as IFP_{CS} and originally described in previous scientific work from our laboratory[85], exploits the Open Drug Discovery Toolkit Python library to calculate protein-ligand interaction fingerprints for each frame of the TTMD trajectory and compare it through the cosine similarity metric as implemented in the Scikit-learn Python module to a reference fingerprint, based on the last trajectory frame extracted from the second and last equilibration stage. Specifically, each protein-ligand interaction fingerprint is an integer vector composed of rx8 elements, where r is the number of protein residues. Each protein residue is encoded into eight bits of information, one for each type of intermolecular interaction considered (hydrophobic contacts, aromatic face to face, aromatic edge to face, hydrogen bond with protein acting as donor, hydrogen bond with protein acting as acceptor, salt bridge with protein acting as the positively charged member, salt bridge with protein acting as the positively negative member, and ionic bond with a metal ion respectively). The mathematic formulation of the IFP_{CS} scoring function is reported in Equation 1:

$$IFP_{CS} = \frac{A \cdot B}{\|A\|\|B\|} * -1 \quad (1)$$

The IFP_{CS} value ranges from -1, indicating a total superposition between the reference and the query fingerprint, to 0, which indicates that every interaction determinant of the reference fingerprint is lost in the query.

At the end of each “TTMD-step”, the average IFP_{CS} score for the step is calculated: if the value is null, indicating that for the whole duration of the step the original binding mode

was not sampled, the TTMD trajectory is terminated, while if the value is not null the simulation proceeds to the next“TTMD-step”.

TTMD Trajectory Analyses

Each TTMD trajectory is analyzed by making use of an in-house Python 3.10 script. The root-mean-squared deviation of atomic coordinates (RMSD) for both the ligand and the protein backbone is calculated for each frame through the MDAnalysis package. The per-residue decomposition of the protein-ligand interaction energy is computed for each frame by exploiting the NAMD Energy plugin (version 1.4)[86] for VMD. Three different plots are then generated, making use of the Matplotlib and seaborn Python packages. The first plot (“Titration Profile”) reports the average IFP_{CS} value for each TTMD step as a function of the step temperature. A straight line joining the start and final state of the simulation is also drawn in the graph, and its slope is reported in the legend and stored for further analysis. The second graph illustrates the time-dependent per-residue decomposition of the interaction energy, with the 25 most contacted residues alongside the TTMD trajectory being considered. The third and final plot reports the time-dependent evolution of the ligand and protein backbone RMSD and the IFP_{CS} value.

MS coefficient determination

For each TTMD simulation, a proxy value for the protein-ligand complex stability based on the conservation of the binding mode throughout the trajectory is calculated as reported in Equation 2:

$$MS = \frac{mean\ IFP_{CS}{}^{T^{end}} - (-1)}{T^{end} - T^{start}} \quad (2)$$

The MS coefficient is the slope of the straight line that interpolates the first and the last point of the “Titration Profile” plot described in the previous paragraph. In Equation 2, $mean\ IFP_{CS}{}^{T^{end}}$ is the average IFP_{CS} value for the last temperature explored in the TTMD trajectory, -1 is the IFP_{CS} value for the initial state of the simulation, T^{end} and T^{start} are the final and starting temperatures of the simulation. Values are positive and can vary between 0 (indicative of a strong binding) and 1 (related to a weak binding).

For each ligand, five independent TTMD simulations are performed, and the average MS coefficient is then calculated upon three of them, discarding the highest and the lowest value.

Results

To test and validate the applicability of the Thermal Titration Molecular Dynamics (TTMD) protocol, we performed four different case studies on three different pharmaceutically relevant targets of interest for our laboratory, specifically casein kinase 1δ (CK1δ), casein kinase 2 (CK2), pyruvate dehydrogenase kinase 2 (PDK2), and SARS-CoV-2 main protease (M^{pro}). For each protein target, five different protein-ligand complexes were chosen, based on the availability of binding affinity data. A list of all protein-ligand complexes used in the present work can be found in Table 1 at the beginning of the Materials and Methods section, while detailed information about the ligands utilized in each test case can be found in Tables S1-S4 (Supplementary Materials). For each protein-ligand complex investigated in the article, five independent TTMD simulations were carried out. The results for each test case are reported hereafter in separate paragraphs and discussed aggregately in the Discussion section of the manuscript. For each target, the conservation of the protein fold throughout the simulation is carried out by monitoring the time-dependent evolution of the protein backbone RMSD, as reported in the detailed analysis for each representative replicate illustrated in Figures S1-S21 (Supplementary Materials).

Protein Kinase CK1δ

Protein Kinase CK1δ is a serine-threonine kinase that belongs to the family of CK1 Kinases (Casein Kinase 1)[87]. Due to its pleiotropic nature (about 140 substrates have been reported so far), this kinase is involved in the regulation of several different cellular pathways[87,88]. Particularly relevant from a medicinal chemistry perspective is its involvement in several neurodegenerative diseases such as Alzheimer's Disease (AD), Parkinson's Disease (PD), and amyotrophic lateral sclerosis (ALS) by phosphorylating protein targets such as the Tau protein, α-synuclein, and TDP-43 (TransActivate Response DNA Binding Protein 43)[89]. 34 crystal structures of CK1δ, among which several protein-ligand complexes can be found, are deposited in the Protein Data Bank, with the affinity

of co-crystallized inhibitors ranging over three orders of magnitude, making it a suitable target for the application of our computational protocol. The results of TTMD simulations performed on CK1δ crystal complexes are summarized in Table 2 and Figure 2, while a detailed analysis for a representative trajectory for each protein-ligand complex (the one highlighted in green in Table 2) is reported in Figures S1-S5 in the Supplementary Materials.

PDB ID	LIG ID	IC_{50} (nM)	MS_{MD1}	MS_{MD2}	MS_{MD3}	MS_{MD4}	MS_{MD5}	$MS_{average}$
3UZP	0CK	13	0.00203	0.00333	0.00228	0.00212	0.00275	0.0024
4TN6	PFO	3.9	0.00300	0.00345	0.00307	0.00214	0.00309	0.0031
5IH5	AUE	500	0.00539	0.00446	0.00357	0.00909	0.00347	0.0045
5IH6	AUG	2500	0.00380	0.00769	0.00420	0.00557	0.00594	0.0052
5MQV	D5Q	9	0.00283	0.00295	0.00317	0.00256	0.00340	0.0030

Table 2. Results for the Thermal Titration Molecular Dynamics (TTMD) simulations performed on the five investigated CK1δ complexes. For each protein-ligand complex, the PDB accession code, the ligand three-letter code, the experimentally determined affinity value, the MS coefficient for each simulation, and the average MS coefficient are reported. In each row, the lowest MS value is highlighted in red, while the highest value is highlighted in blue: both values were discarded for the calculation of the average MS coefficient reported in the last column. The most representative replicates, the one with the nearest MS coefficient to the average MS, is highlighted in green.

As can be deduced by the analysis of the data extracted from the various TTMD simulations, the ligands respond differently to the protocol based on the experimental affinity value. As can be noticed in Figure 2, complexes 5IH5 and 5IH6, which are characterized by the lowest affinity values (500 and 2500 µM, respectively), are the ones with the highest MS coefficient value, indicating a loss of the native binding mode throughout the simulations. On the contrary, ligands with a good experimental affinity towards the target (in the low nanomolar range), are associated with good conservation of the native binding mode, as highlighted by the lower MS coefficient value. Based on this observation, a cutoff MS value of 0.004, able to distinguish between the tight and

weak binders for CK1δ, can be determined. The detailed trajectory analyses provided in Figures S1-S5 illustrate how the loss of the native binding mode is primarily driven by the loss of crucial hydrogen bond interactions with the hinge region, particularly with Leu85 and Glu83. This evidence is in agreement with previously published works from our laboratory, which indicates how using an appropriate pharmacophore filter that takes into account the crucial hydrogen bond with the backbone of Leu85 leads to good results in virtual screening[85,90,91]. For visual reference, a comparison between the representative replicate for the 3UZP and 5IH6 is reported in Video S1.

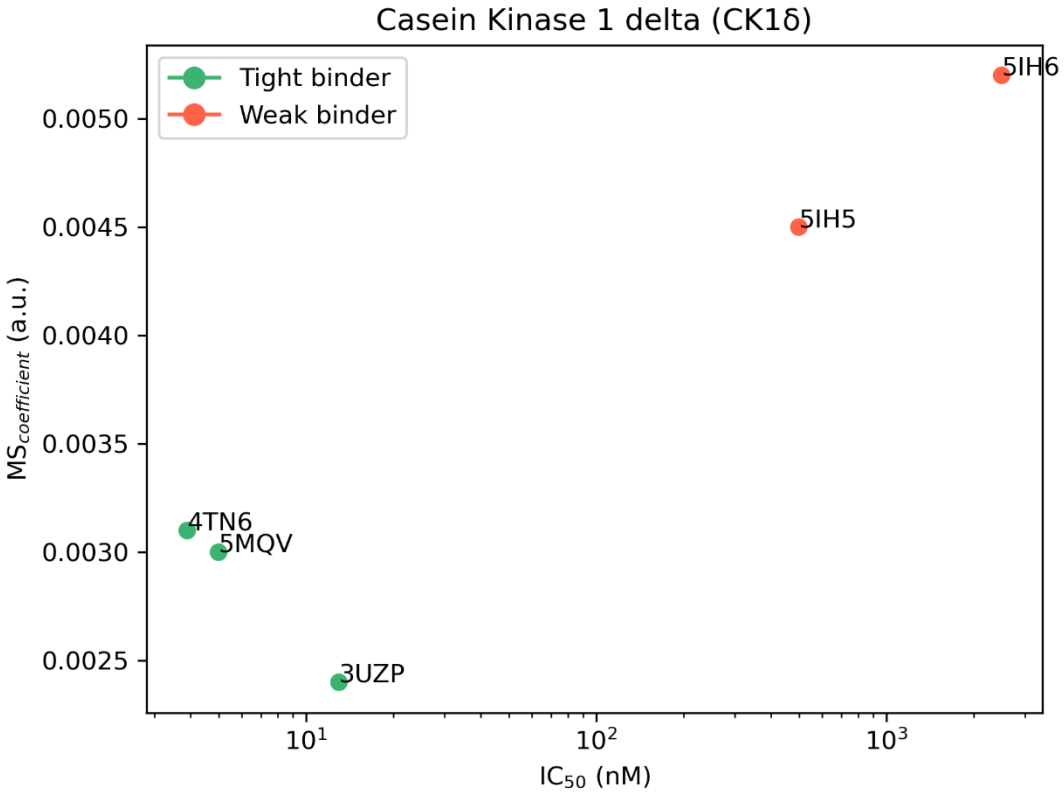


Figure 2. Aggregate results of the Thermal Titration Molecular Dynamics (TTMD) simulations performed on the five investigated CK1δ complexes. On the horizontal axis, the experimentally determined affinity value (expressed as IC_{50}) is reported, while on the vertical axis the average MS coefficient is indicated. Each dot is color-coded as green or red and classified as a tight or weak binder based on the MS cutoff value of 0.004.

Protein Kinase CK2

Protein Kinase CK2 is a serine-threonine kinase and represents one of the first identified protein kinases[92]. Similar to CK1δ, CK2 can phosphorylate a plethora of different

substrates, being therefore involved in the regulation of several biological pathways[93]. The variety of biologically relevant scenarios in which CK2 is involved makes it a hot target from a pharmaceutical perspective, being related to several types of cancer, different neurodegenerative diseases (similarly to CK1δ), and also viral infections[94]. As of today, 214 crystal structures of CK2, among which several protein-ligand complexes can be found, are deposited in the Protein Data Bank, with the affinity of co-crystallized inhibitors ranging over five orders of magnitude, making it also a suitable target for the application of our computational protocol.

PDB ID	LIG ID	K_d (nM)	MS_{MD1}	MS_{MD2}	MS_{MD3}	MS_{MD4}	MS_{MD5}	$MS_{average}$
2ZJW (0)	REF	40	0.00318	0.00282	0.00269	0.00391	0.00205	0.0029
2ZJW (-)	REF	40	0.00192	0.00189	0.00227	0.00205	0.00133	0.0020
3H30	RFZ	13000	0.00504	0.00484	0.00395	0.00450	0.00453	0.0046
3PE1	3NG	1.5	0.00178	0.00231	0.00198	0.00219	0.00373	0.0022
3PE2	E1B	2.3	0.00179	0.00179	0.00171	0.00169	0.00170	0.0017
6HOU	V55	53400	0.0035	0.00667	0.00769	0.00364	0.00667	0.0057

Table 3. Results for the Thermal Titration Molecular Dynamics (TTMD) simulations performed on the five investigated CK2 complexes. For each protein-ligand complex, the PDB accession code, the ligand three-letter code, the experimentally determined affinity value, the MS coefficient for each simulation, and the average MS coefficient are reported. In each row, the lowest MS value is highlighted in red, while the highest value is highlighted in blue: both values were discarded for the calculation of the average MS coefficient reported in the last column. The most representative replicates, the one with the nearest MS coefficient to the average MS, is highlighted in green. For the complex 2ZJW, two different protonation states were independently considered in the simulations and are reported separately.

The results of TTMD simulations carried out on CK2 crystal complexes are summarized in Table 3 and Figure 3, while a detailed analysis for a representative trajectory for each protein-ligand complex (the one highlighted in green in Table 3) is reported in Figures S6-S11 in the Supplementary Materials. As explained in the Materials and Methods section, two different ligand protonation states are considered for the 2ZJW complex, the neutral

and the negatively charged one. Although they are reported separately in both Table 3 and Figures S6-S7, they are considered as a single entity in Figure 3, where the average MS value between the two different protonation states is reported.

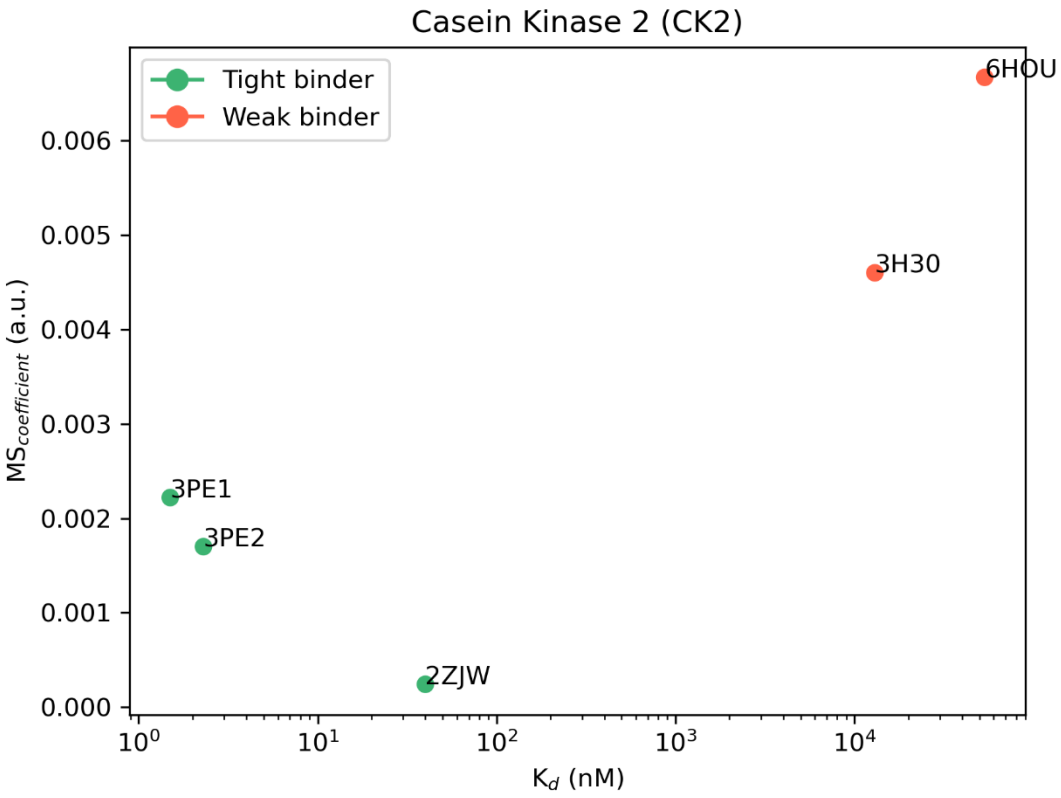


Figure 3. Aggregate results of the Thermal Titration Molecular Dynamics (TTMD) simulations performed on the five investigated CK2 complexes. On the horizontal axis, the experimentally determined affinity value (expressed as K_d) is reported, while on the vertical axis the average MS coefficient is indicated. Each dot is color-coded as green or red and classified as a tight or weak binder based on the MS cutoff value of 0.004. For the complex 2ZJW, two different ligand protonation states were considered, but only one aggregate result (the average of the two states) is reported in the plot.

As in the case of CK1δ, the investigated ligands show a different behavior during the TTMD simulations based on their experimental affinity value. The complexes characterized by a lower protein-ligand binding affinity (13 µM for 3H30, 53,4 µM for 6HOU), are also the ones characterized by the highest MS coefficient value (0.0046 and 0.0057, respectively). On the contrary, as observed for CK1δ, ligands with a binding affinity in the low nanomolar range are characterized by a conservation of the native binding mode throughout the simulation, resulting in a lower MS coefficient (below 0.003). Once again, an empirical

threshold MS value of 0.004 can be extracted from this test set and utilized to distinguish between tight and weak binders.

As can be noticed by the evolution of the interaction pattern between ligands and the binding pocket throughout the trajectories (Figures S6-S11), the ligands with the most stable binding mode are the ones that tightly interact with Lys68 and Ile174: this is particularly noticeable in the case of ellagic acid (2ZJW), for which two different protonation states have been simulated. In the neutral form, the ellagic acid binding mode is less stable throughout the trajectory, while in the monocharged form, the ellagic acid binding mode is very stable even at high simulation temperatures, due to a favorable interaction with Lys68. A comparison between the representative replicate for complexes 3PE2 and 6HOU is shown in Video S2.

Pyruvate Dehydrogenase Kinase 2 (PDK2)

Pyruvate Dehydrogenase Kinase 2 (PDK2) is a pivotal enzyme in cellular energy metabolism that has previously been implicated in cancer[95]. PDK2 is a member of the GHKL ATPase/kinase superfamily and exerts its activity by phosphorylating and regulating the pyruvate dehydrogenase complex (PDC), which is a central control point in cellular energy metabolism since it links glycolysis with the tricarboxylic acid cycle[96,97]. Due to its involvement in the regulation of the energetic metabolism of cells, it is a drug target both from a metabolic and an antitumoral perspective. At the present moment, 33 crystal structures of PDK2, among which several protein-ligand complexes can be found, are deposited in the Protein Data Bank, with the affinity of co-crystallized inhibitors ranging over six orders of magnitude, making it a suitable target for the application of our computational protocol.

The results of TTMD simulations performed on PDK2 crystal complexes are summarized in Table 4 and Figure 4, while a detailed analysis for a representative trajectory for each protein-ligand complex (the one highlighted in green in Table 4) is reported in Figures S12-S16 in the Supplementary Materials.

PDB ID	LIG ID	K_d(nM)	MS_{MD1}	MS_{MD2}	MS_{MD3}	MS_{MD4}	MS_{MD5}	$MS_{average}$
4MP2	PV1	3570	0.00603	0.01429	0.025	0.01111	0.00579	0.0105
4V25	SZ6	150	0.00216	0.00247	0.00194	0.00221	0.00332	0.0023
5J71	P35	110	0.00296	0.00296	0.00269	0.00243	0.00331	0.0029
5M4M	7FW	1	0.0028	0.0025	0.00286	0.00261	0.00318	0.0027
7EA0	W6P	958000	0.00347	0.00769	0.00667	0.00769	0.00665	0.0070

Table 4. Results for the Thermal Titration Molecular Dynamics (TTMD) simulations performed on the five investigated PDK2 complexes. For each protein-ligand complex, the PDB accession code, the ligand three-letter code, the experimentally determined affinity value, the MS coefficient for each simulation, and the average MS coefficient are reported. In each row, the lowest MS value is highlighted in red, while the highest value is highlighted in blue: both values were discarded for the calculation of the average MS coefficient reported in the last column. The most representative replicates, the one with the nearest MS coefficient to the average MS, is highlighted in green.

The analysis of results for the TTMD simulations performed on PDK2 protein-ligand complexes matches the ones already shown for CK1δ and CK2. Indeed, the ligands with the lowest binding affinity (3.57 µM for 4MP2, 958 µM for 7EA0), are the ones with the highest MS coefficient (0.0105 and 0.007, respectively), while on the contrary ligands characterized by a good binding affinity are also the ones characterized by the lowest MS coefficient (below 0.003). The same threshold value used for previous cases (MS < 0.004) can also be utilized in this case to distinguish between the weak and the tight binders.

Looking at the evolution of the interaction pattern of various ligands throughout the simulations, it can be noticed that tight binders are characterized by persistent attractive interactions with Asp290 and Thr354. These residues are buried within the binding pocket, which contributes to the persistence of their interaction with the ligand compared to other more solvent-exposed residues such as Asn255, Arg258, and Glu262, that instead seem to be less relevant in retaining the ligand within the binding site. Noticeably, in the case of complex 4MP2, a repulsive interaction with Asp290 is present at the beginning of the simulation, and this could be a possible explanation for the low persistence of the

native binding mode. Moreover, as can be noticed in Figures S14 and S15, in the case of complexes 5J71 and 5M4M, the fraction of the ligand which interacts with Asp290 and Thr354 barely moves from the starting position, fully retaining this interaction for the whole duration of the simulation, while partially losing the interactions with other more exposed residues which increases the ligand RMSD despite most of the binding determinants being conserved. A comparison between the representative replicate for complexes 4V25 and 4MP2 is illustrated in Video S3.

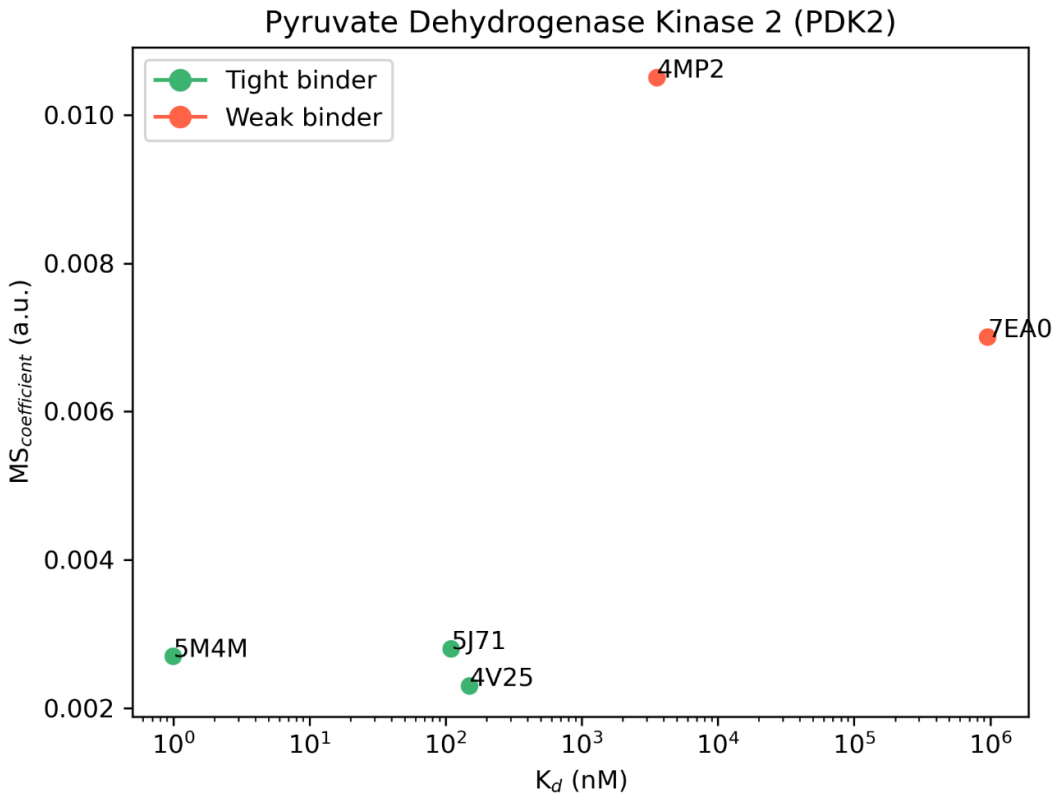


Figure 4. Aggregate results of the Thermal Titration Molecular Dynamics (TTMD) simulations performed on the five investigated PDK2 complexes. On the horizontal axis, the experimentally determined affinity value (expressed as K_d) is reported, while on the vertical axis the average MS coefficient is indicated. Each dot is color-coded as green or red and classified as a tight or weak binder based on the MS cutoff value of 0.004.

SARS-CoV-2 Main Protease (M^{pro})

SARS-CoV-2 is a betacoronavirus responsible for the COVID-19 pandemic which, to date, has caused the death of more than 6.5 million people around the world[98,99]. A pivotal enzyme in the virus' replication cycle is represented by its main protease (M^{pro}), a cysteine

peptidase that is involved in the proteolytic cleavage of the pp1a/pp1ab polyproteins into several mature nonstructural proteins[100,101]. Due to its crucial role in the ability of the virus to replicate itself, the main protease is a validated antiviral target[102] and, as such, has become the focus of several different drug discovery campaigns[103–105], leading to 613 experimentally solved structures deposited on the Protein Data Bank, a marketed drug (Paxlovid, therapeutic association of Nirmatrelvir and Ritonavir)[106,107] and several inhibitors, with affinity values ranging from low nanomolar to micromolar and beyond. Its pharmaceutical relevance and the abundance of structural data make the SARS-CoV-2 main protease a good target for the validation of the TTMD protocol. The results of TTMD simulations performed on SARS-CoV-2 M^{pro} crystal complexes are summarized in Table 5 and Figure 5, while a detailed analysis for a representative trajectory for each protein-ligand complex (the one highlighted in green in Table 5) is reported in Figures S17-S21 in the Supplementary Materials.

PDB ID	LIG ID	IC_{50}(nM)	MS_{MD1}	MS_{MD2}	MS_{MD3}	MS_{MD4}	MS_{MD5}	$MS_{average}$
6M2N	3WL	940	0.00714	0.00714	0.00557	0.00714	0.00909	0.0071
7LTJ	YD1	4200	0.00909	0.00642	0.00595	0.00667	0.00664	0.0066
7M8P	YSJ	20	0.00401	0.00336	0.00379	0.00317	0.00254	0.0034
7M91	YU4	25	0.00316	0.00437	0.00315	0.00313	0.00383	0.0034
7N44	06I	42	0.00325	0.00598	0.00398	0.00384	0.00383	0.0039

Table 5. Results for the Thermal Titration Molecular Dynamics (TTMD) simulations performed on the five investigated SARS-CoV-2 M^{pro} complexes. For each protein-ligand complex, the PDB accession code, the ligand three-letter code, the experimentally determined affinity value, the MS coefficient for each simulation, and the average MS coefficient are reported. In each row, the lowest MS value is highlighted in red, while the highest value is highlighted in blue: both values were discarded for the calculation of the average MS coefficient reported in the last column. The most representative replicates, the one with the nearest MS coefficient to the average MS, is highlighted in green.

The analysis of the TTMD simulations performed on M^{pro} protein-ligand complexes matches the ones already shown for the previous cases. Once again, the ligands

characterized by the lowest experimental binding affinity (0.94 µM for 6M2N, 4.2 µM for 7LTJ) are the ones associated with the highest MS coefficient (0.0071 and 0.0066, respectively). Accordingly, the ligands that present the highest binding affinity are associated with low MS coefficients, indicative of a persistent binding mode. Even for the SARS-CoV-2 M^{pro} it is possible to reutilize the previously determined threshold value (MS < 0.004) to separate the strong and the weak binders.

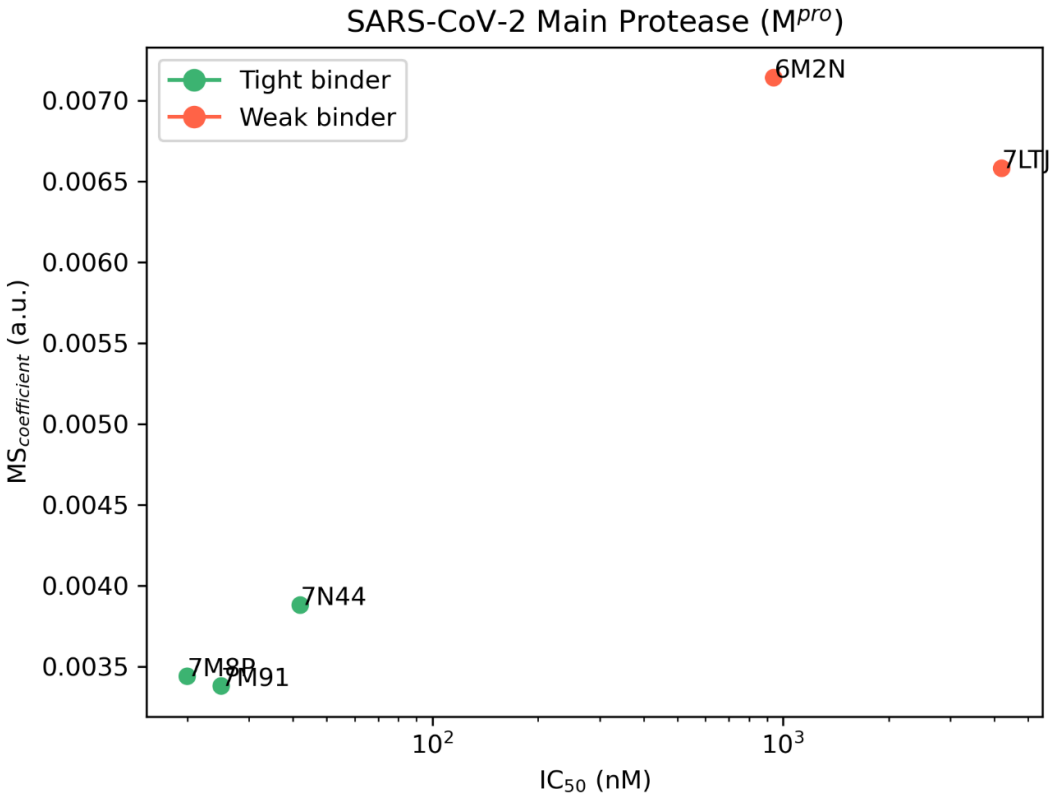


Figure 5. Aggregate results of the Thermal Titration Molecular Dynamics (TTMD) simulations performed on the five investigated SARS-CoV-2 M^{pro} complexes. On the horizontal axis, the experimentally determined affinity value (expressed as IC_{50}) is reported, while on the vertical axis the average MS coefficient is indicated. Each dot is color-coded as green or red and classified as a tight or weak binder based on the MS cutoff value of 0.004.

Regarding the evolution of the interaction pattern for the protein-ligand complexes that were investigated, it is possible to notice how the most persistent ligands are characterized by strong and stable interactions with key residues such as Met164-Glu166, located in a β-sheet that constitutes the central portion of the binding site lining several

subpockets that precede the catalytic dyad such as S1, S2 and S3, and residues Leu141-Cys145, which line the S1 subpocket and constitute the so-called oxyanion loop, a structure that portrays a crucial role in the catalytic cycle of the protease[101]. On the contrary, interactions with residues lining the S2 and S4 subpocket seem to not be pivotal for the interaction with the catalytic site. As can be seen in Figures S20-S21, for example, the ligands partially lose their interactions with residues Asp187-Gln192: this causes a slight increase in the ligand RMSD towards the end of the simulation without overall altering the conservation of the native binding mode, as depicted by the interaction fingerprint analysis. Interestingly, in the case of complex 6M2N, a repulsive interaction with Glu166 is present at the beginning of the simulation: considering the pivotal role that this residue portrays both in the dimerization process[108] (it forms a salt bridge through its side chain with the side chain of Ser1 of the second protomer) and in the binding of ligands, this could well explain the low persistence of the native binding mode for this ligand throughout the simulation. A comparison between the representative replicate for complexes 7LTJ and 7M91 is illustrated in Video S4.

Discussion

The Thermal Titration Molecular Dynamics (TTMD) method is an alternative protocol for the qualitative estimation of the protein-ligand complex stability based on the persistence of the native binding mode throughout a series of molecular dynamics simulations performed at progressively increasing temperatures. To evaluate the protocol capabilities, we performed four different case studies on an equal number of pharmaceutically relevant test cases, i.e. protein kinase CK1δ, protein kinase CK2, pyruvate dehydrogenase kinase 2, and SARS-CoV-2 main protease (M^{pro}). Despite its simplicity, the TTMD workflow was able to correctly discriminate between tight binders (with affinity values in the low nanomolar range) and weak binders (the ones with affinity values superior to the micromolar threshold) by applying an appropriate MS coefficient cutoff. This classification was performed on ligands with different scaffolds, and different interaction features, making its application interesting in real-world drug-discovery campaigns, where compounds from different chemical classes are usually identified in the

first stages and are then subjected to an iterative optimization of their binding affinity towards the target of interest through chemical modification of their structures.

Contrary to most protocols that aim at predicting or estimating the drug-target residence time or other proxy values for the ligand affinity, TTMD does not require simulating the full unbinding event. Although this results in a rawer prediction compared to other similar protocols, this approach has two major advantages. The first one is that the simulation time is limited and can be accurately estimated right from the start. This is particularly useful in the case of a batch application of the protocol across a library of different compounds resulting from a screening campaign or an optimization process. Moreover, it facilitates the automatization of the process and its incorporation into existing drug-discovery pipelines. The second advantage is that by determining a relative metric (the MS coefficient) rather than an absolute one (i.e., the time required to observe the unbinding), there is no need for the definition of an arbitrary cutoff value for the detection of the unbinding event. Concerning this, most protocols exploit geometric descriptors such as the distance between the center of masses of the ligand and the binding site, or the distance between the ligand and the protein to define whether the ligand detached from the binding site[43,46]. This poses the problem of choosing the right distance because, in the case of deep and buried binding sites, the chosen cutoff value could not consider the whole unbinding process, leading once again to an underestimation of the residence time. On the contrary, arbitrarily increasing the distance could elongate the simulation time without improving the prediction accuracy. Furthermore, using an interaction fingerprint-based metric instead of the standard RMSD for monitoring the evolution of the binding mode throughout the simulation results in lower sensitivity toward the chemical structure of the ligand: as highlighted in some of our trajectory analyses (Figures S14 and S15, for example), the presence of some ligand moieties that do not directly interact with the binding pocket or are slightly solvent exposed, leading to a less stable interaction with the target, could lead to an increase in the ligand RMSD without compromising the key binding determinants of the compounds. This could lead to a false perception of the unbinding event, causing errors in the evaluation of the persistence of the receptor-ligand complex, especially if very low cutoffs are utilized, such is the case in

some studies in which classic MD simulations are used as a way to refine docking results and distinguish between native-like poses and decoys[109,110].

Another major advantage is related to the accessibility of the method. First, although the protocol in its current form exploits the ACEMD3 program to run molecular dynamics simulations, it can be easily and readily adapted to be utilized with any other major molecular dynamics engine such as OpenMM, GROMACS, or AMBER. Secondly, compared to other approaches TTMD is easier to implement. For example, contrary to metadynamics-based approaches, where the choice of the collective variable to monitor is not trivial[38], in the case of TTMD the user only needs to choose a temperature ramp where that ensures the conservation of the protein fold by monitoring a simple geometric descriptor such as the protein backbone RMSD. Although some attempts at optimizing the temperature ramp to decrease the simulation time without reducing the accuracy of the method are already going on in our laboratory, the temperature ramp proposed in this article should represent a good starting point for the third-party implementation of the method. Theoretically, increasing the simulation time for each “TTMD step” should provide an increase in the resolution of the technique, but would also result in an increased computational effort. On the contrary, reducing the simulation time for each TTMD step would reduce the computational effort, making the protocol more affordable, especially for those setups where a large number of different ligands is evaluated at a given time, but would also flatten the difference in MS coefficient between the ligands, thus decreasing the sensitivity of the technique. One possible solution could be to use different simulation times at different temperatures, for example simulating longer steps at lower temperatures and shorter ones at higher temperatures. The pool of test cases provided in this article should, in principle, aid the user in the choice of a non-default temperature ramp, since the user could compare the results of its custom temperature ramp with the ones originally obtained and evaluate on its own the performances of a different ramp.

Other than estimating the protein-ligand binding affinity, the TTMD protocol could be easily adapted to perform mechanistic evaluations on the unbinding event, by appropriately tuning the temperature ramp and the simulation time to carry on the

simulation until the native binding mode is completely lost. Although in the current form this protocol is not specifically designed for this purpose, it can be already used to discriminate different protein-ligand interactions based on their effect on the binding affinity. This indication could be very useful in the generation and refinement of pharmacophore filters, which are commonly used to reduce the false positive rates in docking-based virtual screening campaigns.

Another possible application of the TTMD protocol could be to distinguish the native binding pose from a set of decoy ones. This could be particularly useful in the case of fragment compounds, which can have several plausible binding modes, and are usually evolved into mature ligands by rationally modifying their scaffold to expand on their binding determinants without altering the existing interaction profile. Work in this sense is already going on in our laboratory and will be the scope of a future paper.

The last element that needs to be addressed in the nearest future is the applicability of the method to membrane proteins: so far, the protocol has been applied only to globular targets, but a wide variety of pharmaceutically relevant targets are membrane systems. Membrane systems are intrinsically more complicated to manage, because other than monitoring the protein fold throughout the simulations one has to decide how to manage the membrane. A possible solution could be to remove the membrane and treat the protein as soluble, possibly with the implementation of restraints on the atomic positions of atoms outside the binding site[43]. Evaluations in this sense are already going on in our laboratory to tune the protocol to be utilized also for this class of targets.

Conclusions

In this scientific work, we presented the first application of Thermal Titration Molecular Dynamics (TTMD), an alternative protocol for the qualitative estimation of protein-ligand complex stability by monitoring the conservation of the native ligand binding mode throughout a series of classic molecular dynamics simulations performed at progressively increasing temperatures through a scoring function based on protein-ligand interaction fingerprints. Four different test cases regarding the application of the technique on three different pharmaceutically relevant targets were presented. For each case, TTMD

simulations were able to distinguish between tight (low nanomolar) from weak (micromolar) binders. The simplicity of the protocol, particularly regarding the choice of user-defined parameters to run the simulations, the agnosticism concerning the selection of the molecular dynamics engine, and the limited simulation time make it a viable choice for various medicinal chemistry projects, especially as a screening tool in the early stages of drug discovery campaigns. Further work is needed to extend the applicability domain of the technique to membrane proteins, and evaluations in this sense are already going on in our laboratory.

References

(1) Ehrlich, P. Address in Pathology, ON CHEMIOTHERAPY. *Br Med J* **1913**, *2*, 353–359. https://doi.org/10.1136/BMJ.2.2746.353.

(2) Langley, J. N. On the Reaction of Cells and of Nerve-Endings to Certain Poisons, Chiefly as Regards the Reaction of Striated Muscle to Nicotine and to Curari. *J Physiol* **1905**, *33*, 374. https://doi.org/10.1113/JPHYSIOL.1905.SP001128.

(3) Kaufmann, S. H. E. Paul Ehrlich: Founder of Chemotherapy. *Nat Rev Drug Discov* **2008**, *7*, 373–373. https://doi.org/10.1038/nrd2582.

(4) Maehle, A. H. "Receptive Substances": John Newport Langley (1852–1925) and His Path to a Receptor Theory of Drug Action. *Med Hist* **2004**, *48*, 153. https://doi.org/10.1017/S0025727300000090.

(5) Bennett, M. R. The Concept of Transmitter Receptors: 100 Years On. *Neuropharmacology* **2000**, *39*, 523–546. https://doi.org/10.1016/S0028-3908(99)00137-9.

(6) Gourley, D. R. Isolation and Characterization of Membrane Drug Receptors. *Prog Drug Res* **1976**, *20*, 323–346. https://doi.org/10.1007/978-3-0348-7094-8_10/COVER/.

(7) Rang, H. P. The Receptor Concept: Pharmacology's Big Idea. *Br J Pharmacol* **2006**, *147*, S9–S16. https://doi.org/10.1038/SJ.BJP.0706457.

(8) Borea, P. A.; Varani, K.; Gessi, S.; Gilli, P.; Dalpiaz, A. Receptor Binding Thermodynamics as a Tool for Linking Drug Efficacy and Affinity. *Farmaco* **1998**, *53*, 249–254. https://doi.org/10.1016/S0014-827X(98)00017-2.

(9) Kairys, V.; Baranauskiene, L.; Kazlauskiene, M.; Matulis, D.; Kazlauskas, E. Binding Affinity in Drug Design: Experimental and Computational Techniques. *https://doi.org/10.1080/17460441.2019.1623202* **2019**, *14*, 755–768. https://doi.org/10.1080/17460441.2019.1623202.

(10) Pan, A. C.; Borhani, D. W.; Dror, R. O.; Shaw, D. E. Molecular Determinants of Drug–Receptor Binding Kinetics. *Drug Discov Today* **2013**, *18*, 667–673. https://doi.org/10.1016/J.DRUDIS.2013.02.007.

(11) Copeland, R. A. The Drug–Target Residence Time Model: A 10-Year Retrospective. *Nature Reviews Drug Discovery 2015 15:2* **2015**, *15*, 87–95. https://doi.org/10.1038/nrd.2015.18.

(12) Copeland, R. A.; Pompliano, D. L.; Meek, T. D. Drug-Target Residence Time and Its Implications for Lead Optimization. *Nat Rev Drug Discov* **2006**, *5*, 730–739. https://doi.org/10.1038/nrd2082.

(13) Swinney, D. C. The Role of Binding Kinetics in Therapeutically Useful Drug Action. *Curr Opin Drug Discov Devel* **2009**, *12*, 31–39.

(14) Zhang, R.; Monsma, F. The Importance of Drug-Target Residence Time. *Curr Opin Drug Discov Devel* **2009**, *12*, 488–496.

(15) Lu, H.; Tonge, P. J. Drug–Target Residence Time: Critical Information for Lead Optimization. *Curr Opin Chem Biol* **2010**, *14*, 467–474. https://doi.org/10.1016/J.CBPA.2010.06.176.

(16) Hill, T. L. Effect of Rotation on the Diffusion-Controlled Rate of Ligand-Protein Association. *Proc Natl Acad Sci U S A* **1975**, *72*, 4918. https://doi.org/10.1073/PNAS.72.12.4918.

(17) Guo, D.; Hillger, J. M.; Ijzerman, A. P.; Heitman, L. H. Drug-Target Residence Time—A Case for G Protein-Coupled Receptors. *Med Res Rev* **2014**, *34*, 856–892. https://doi.org/10.1002/MED.21307.

(18) Cusack, K. P.; Wang, Y.; Hoemann, M. Z.; Marjanovic, J.; Heym, R. G.; Vasudevan, A. Design Strategies to Address Kinetics of Drug Binding and Residence Time. *Bioorg Med Chem Lett* **2015**, *25*, 2019–2027. https://doi.org/10.1016/J.BMCL.2015.02.027.

(19) Bernetti, M.; Cavalli, A.; Mollica, L. Protein–Ligand (Un)Binding Kinetics as a New Paradigm for Drug Discovery at the Crossroad between Experiments and Modelling. *Medchemcomm* **2017**, *8*, 534–550. https://doi.org/10.1039/C6MD00581K.

(20) Meyer-Almes, F. J. Kinetic Binding Assays for the Analysis of Protein–Ligand Interactions. *Drug Discov Today Technol* **2015**, *17*, 1–8. https://doi.org/10.1016/J.DDTEC.2015.08.004.

(21) Bernetti, M.; Masetti, M.; Rocchia, W.; Cavalli, A. Kinetics of Drug Binding and Residence Time. *https://doi.org/10.1146/annurev-physchem-042018-052340* **2019**, *70*, 143–171. https://doi.org/10.1146/ANNUREV-PHYSCHEM-042018-052340.

(22) Hulme, E. C.; Trevethick, M. A. Ligand Binding Assays at Equilibrium: Validation and Interpretation. *Br J Pharmacol* **2010**, *161*, 1219–1237. https://doi.org/10.1111/J.1476-5381.2009.00604.X.

(23) Sridharan, R.; Zuber, J.; Connelly, S. M.; Mathew, E.; Dumont, M. E. Fluorescent Approaches for Understanding Interactions of Ligands with G Protein Coupled Receptors. *Biochim Biophys Acta* **2014**, *1838*, 15–33. https://doi.org/10.1016/J.BBAMEM.2013.09.005.

(24) Ilien, B.; Franchet, C.; Bernard, P.; Morisset, S.; Weill, C. O.; Bourguignon, J. J.; Hibert, M.; Galzi, J. L. Fluorescence Resonance Energy Transfer to Probe Human M1 Muscarinic Receptor Structure and Drug Binding Properties. *J Neurochem* **2003**, *85*, 768–778. https://doi.org/10.1046/J.1471-4159.2003.01717.X.

(25) Stoddart, L. A.; Johnstone, E. K. M.; Wheal, A. J.; Goulding, J.; Robers, M. B.; MacHleidt, T.; Wood, K. V.; Hill, S. J.; Pfleger, K. D. G. Application of BRET to Monitor Ligand Binding to GPCRs. *Nat Methods* **2015**, *12*, 661–663. https://doi.org/10.1038/NMETH.3398.

(26) Núñez, S.; Venhorst, J.; Kruse, C. G. Target–Drug Interactions: First Principles and Their Application to Drug Discovery. *Drug Discov Today* **2012**, *17*, 10–22. https://doi.org/10.1016/J.DRUDIS.2011.06.013.

(27) Patching, S. G. Surface Plasmon Resonance Spectroscopy for Characterisation of Membrane Protein-Ligand Interactions and Its Potential for Drug Discovery. *Biochim Biophys Acta* **2014**, *1838*, 43–55. https://doi.org/10.1016/J.BBAMEM.2013.04.028.

(28) Millet, O.; Bernadó, P.; Garcia, J.; Rizo, J.; Pons, M. NMR Measurement of the off Rate from the First Calcium-Binding Site of the Synaptotagmin I C2A Domain. *FEBS Lett* **2002**, *516*, 93–96. https://doi.org/10.1016/S0014-5793(02)02508-5.

(29) Gronewold, T. M. A.; Baumgartner, A.; Hierer, J.; Sierra, S.; Blind, M.; Schäfer, F.; Blümer, J.; Tillmann, T.; Kiwitz, A.; Kaiser, R.; Zabe-Kühn, M.; Quandt, E.; Famulok, M. Kinetic Binding Analysis of Aptamers Targeting HIV-1 Proteins by a Combination of a Microbalance Array and Mass Spectrometry (MAMS). *J Proteome Res* **2009**, *8*, 3568–3577. https://doi.org/10.1021/PR900265R.

(30) Burnouf, D.; Ennifar, E.; Guedich, S.; Puffer, B.; Hoffmann, G.; Bec, G.; Disdier, F.; Baltzinger, M.; Dumas, P. KinITC: A New Method for Obtaining Joint Thermodynamic and Kinetic Data by Isothermal

Titration Calorimetry. *J Am Chem Soc* **2012**, *134*, 559–565. https://doi.org/10.1021/JA209057D/SUPPL_FILE/JA209057D_SI_001.PDF.

(31) Li, D.; Chen, L.; Wang, R.; Liu, R.; Ge, G. Synergetic Determination of Thermodynamic and Kinetic Signatures Using Isothermal Titration Calorimetry: A Full-Curve-Fitting Approach. *Anal Chem* **2017**, *89*, 7130–7138. https://doi.org/10.1021/ACS.ANALCHEM.7B01091/ASSET/IMAGES/LARGE/AC-2017-01091U_0005.JPEG.

(32) Acker, M. G.; Auld, D. S. Considerations for the Design and Reporting of Enzyme Assays in High-Throughput Screening Applications. *Perspect Sci (Neth)* **2014**, *1*, 56–73. https://doi.org/10.1016/J.PISC.2013.12.001.

(33) Bruce, N. J.; Ganotra, G. K.; Kokh, D. B.; Sadiq, S. K.; Wade, R. C. New Approaches for Computing Ligand-Receptor Binding Kinetics. *Curr Opin Struct Biol* **2018**, *49*, 1–10. https://doi.org/10.1016/J.SBI.2017.10.001.

(34) Nunes-Alves, A.; Kokh, D. B.; Wade, R. C. Recent Progress in Molecular Simulation Methods for Drug Binding Kinetics. *Curr Opin Struct Biol* **2020**, *64*, 126–133. https://doi.org/10.1016/J.SBI.2020.06.022.

(35) De Vivo, M.; Masetti, M.; Bottegoni, G.; Cavalli, A. Role of Molecular Dynamics and Related Methods in Drug Discovery. *J Med Chem* **2016**, *59*. https://doi.org/10.1021/acs.jmedchem.5b01684.

(36) Pan, A. C.; Xu, H.; Palpant, T.; Shaw, D. E. Quantitative Characterization of the Binding and Unbinding of Millimolar Drug Fragments with Molecular Dynamics Simulations. *J Chem Theory Comput* **2017**, *13*, 3372–3377. https://doi.org/10.1021/ACS.JCTC.7B00172/SUPPL_FILE/CT7B00172_SI_002.ZIP.

(37) Schuetz, D. A.; de Witte, W. E. A.; Wong, Y. C.; Knasmueller, B.; Richter, L.; Kokh, D. B.; Sadiq, S. K.; Bosma, R.; Nederpelt, I.; Heitman, L. H.; Segala, E.; Amaral, M.; Guo, D.; Andres, D.; Georgi, V.; Stoddart, L. A.; Hill, S.; Cooke, R. M.; De Graaf, C.; Leurs, R.; Frech, M.; Wade, R. C.; de Lange, E. C. M.; IJzerman, A. P.; Müller-Fahrnow, A.; Ecker, G. F. Kinetics for Drug Discovery: An Industry-Driven Effort to Target Drug Residence Time. *Drug Discov Today* **2017**, *22*, 896–911. https://doi.org/10.1016/J.DRUDIS.2017.02.002.

(38) Tiwary, P.; Limongelli, V.; Salvalaglio, M.; Parrinello, M. Kinetics of Protein-Ligand Unbinding: Predicting Pathways, Rates, and Rate-Limiting Steps. *Proc Natl Acad Sci U S A* **2015**, *112*, E386–E391. https://doi.org/10.1073/PNAS.1424461112/SUPPL_FILE/PNAS.1424461112.SAPP.PDF.

(39) Bortolato, A.; Deflorian, F.; Weiss, D. R.; Mason, J. S. Decoding the Role of Water Dynamics in Ligand-Protein Unbinding: CRF1R as a Test Case. *J Chem Inf Model* **2015**, *55*, 1857–1866. https://doi.org/10.1021/ACS.JCIM.5B00440/ASSET/IMAGES/LARGE/CI-2015-00440A_0006.JPEG.

(40) Pietrucci, F.; Marinelli, F.; Carloni, P.; Laio, A. Substrate Binding Mechanism of HIV-1 Protease from Explicit-Solvent Atomistic Simulations. *J Am Chem Soc* **2009**, *131*, 11811–11818. https://doi.org/10.1021/JA903045Y/SUPPL_FILE/JA903045Y_SI_001.PDF.

(41) Sun, H.; Li, Y.; Shen, M.; Li, D.; Kang, Y.; Hou, T. Characterizing Drug-Target Residence Time with Metadynamics: How to Achieve Dissociation Rate Efficiently without Losing Accuracy against Time-Consuming Approaches. *J Chem Inf Model* **2017**, *57*, 1895–1906. https://doi.org/10.1021/ACS.JCIM.7B00075/ASSET/IMAGES/LARGE/CI-2017-00075P_0004.JPEG.

(42) Bussi, G.; Laio, A. Using Metadynamics to Explore Complex Free-Energy Landscapes. *Nature Reviews Physics 2020 2:4* **2020**, *2*, 200–212. https://doi.org/10.1038/s42254-020-0153-0.

(43) Mollica, L.; Decherchi, S.; Zia, S. R.; Gaspari, R.; Cavalli, A.; Rocchia, W. Kinetics of Protein-Ligand Unbinding via Smoothed Potential Molecular Dynamics Simulations. *Scientific Reports 2015 5:1* **2015**, *5*, 1–12. https://doi.org/10.1038/srep11539.

(44) Schuetz, D. A.; Bernetti, M.; Bertazzo, M.; Musil, D.; Eggenweiler, H. M.; Recanatini, M.; Masetti, M.; Ecker, G. F.; Cavalli, A. Predicting Residence Time and Drug Unbinding Pathway through Scaled Molecular Dynamics. *J Chem Inf Model* **2019**, *59*, 535–549. https://doi.org/10.1021/ACS.JCIM.8B00614/ASSET/IMAGES/LARGE/CI-2018-00614Y_0008.JPEG.

(45) Bernetti, M.; Rosini, E.; Mollica, L.; Masetti, M.; Pollegioni, L.; Recanatini, M.; Cavalli, A. Binding Residence Time through Scaled Molecular Dynamics: A Prospective Application to HDAAO Inhibitors. *J Chem Inf Model* **2018**, *58*, 2255–2265. https://doi.org/10.1021/ACS.JCIM.8B00518/ASSET/IMAGES/MEDIUM/CI-2018-00518D_M001.GIF.

(46) Kokh, D. B.; Amaral, M.; Bomke, J.; Grädler, U.; Musil, D.; Buchstaller, H. P.; Dreyer, M. K.; Frech, M.; Lowinski, M.; Vallee, F.; Bianciotto, M.; Rak, A.; Wade, R. C. Estimation of Drug-Target Residence Times by τ-Random Acceleration Molecular Dynamics Simulations. *J Chem Theory Comput* **2018**, *14*, 3859–3869. https://doi.org/10.1021/ACS.JCTC.8B00230/ASSET/IMAGES/LARGE/CT-2018-00230J_0003.JPEG.

(47) Nunes-Alves, A.; Kokh, D. B.; Wade, R. C. Ligand Unbinding Mechanisms and Kinetics for T4 Lysozyme Mutants from TRAMD Simulations. *Curr Res Struct Biol* **2021**, *3*, 106–111. https://doi.org/10.1016/J.CRSTBI.2021.04.001.

(48) Kokh, D. B.; Wade, R. C. G Protein-Coupled Receptor-Ligand Dissociation Rates and Mechanisms from TrAMD Simulations. *J Chem Theory Comput* **2021**, *17*, 6610–6623. https://doi.org/10.1021/ACS.JCTC.1C00641/SUPPL_FILE/CT1C00641_SI_002.MP4.

(49) Berman, H. M. The Protein Data Bank. *Nucleic Acids Res* **2000**, *28*, 235–242. https://doi.org/10.1093/nar/28.1.235.

(50) *Molecular Operating Environment (MOE), 2019.01; Chemical Computing Group ULC, 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, 2021.* https://www.chemcomp.com/Research-Citing_MOE.htm (accessed 2021-01-19).

(51) Long, A.; Zhao, H.; Huang, X. Structural Basis for the Interaction between Casein Kinase 1 Delta and a Potent and Selective Inhibitor. *J Med Chem* **2012**, *55*, 956–960. https://doi.org/10.1021/JM201387S.

(52) *RCSB PDB - 4TN6: CK1d in complex with inhibitor*. https://www.rcsb.org/structure/4TN6 (accessed 2022-06-23).

(53) Ursu, A.; Illich, D. J.; Takemoto, Y.; Porfetye, A. T.; Zhang, M.; Brockmeyer, A.; Janning, P.; Watanabe, N.; Osada, H.; Vetter, I. R.; Ziegler, S.; Schöler, H. R.; Waldmann, H. Epiblastin A Induces Reprogramming of Epiblast Stem Cells into Embryonic Stem Cells by Inhibition of Casein Kinase 1. *Cell Chem Biol* **2016**, *23*, 494–507. https://doi.org/10.1016/j.chembiol.2016.02.015.

(54) Halekotte, J.; Witt, L.; Ianes, C.; Krüger, M.; Bührmann, M.; Rauh, D.; Pichlo, C.; Brunstein, E.; Luxenburger, A.; Baumann, U.; Knippschild, U.; Bischof, J.; Peifer, C.; Koch, P.; Laufer, S. Optimized 4,5-Diarylimidazoles as Potent/Selective Inhibitors of Protein Kinase CK1δ and Their Structural Relation to P38α MAPK. *Molecules* **2017**, *22*. https://doi.org/10.3390/MOLECULES22040522.

(55) Sekiguchi, Y.; Nakaniwa, T.; Kinoshita, T.; Nakanishi, I.; Kitaura, K.; Hirasawa, A.; Tsujimoto, G.; Tada, T. Structural Insight into Human CK2α in Complex with the Potent Inhibitor Ellagic Acid. *Bioorg Med Chem Lett* **2009**, *19*, 2920–2923. https://doi.org/10.1016/J.BMCL.2009.04.076.

(56) Raaf, J.; Brunstein, E.; Issinger, O. G.; Niefind, K. The CK2α/CK2β Interface of Human Protein Kinase CK2 Harbors a Binding Pocket for Small Molecules. *Chem Biol* **2008**, *15*, 111–117. https://doi.org/10.1016/J.CHEMBIOL.2007.12.012.

(57) Battistutta, R.; Cozza, G.; Pierre, F.; Papinutto, E.; Lolli, G.; Sarno, S.; Obrien, S. E.; Siddiqui-Jain, A.; Haddach, M.; Anderes, K.; Ryckman, D. M.; Meggio, F.; Pinna, L. A. Unprecedented Selectivity and Structural Determinants of a New Class of Protein Kinase CK2 Inhibitors in Clinical Trials for the Treatment of Cancer. *Biochemistry* **2011**, *50*, 8478–8488. https://doi.org/10.1021/BI2008382.

(58) Cozza, G.; Zonta, F.; Dalle Vedove, A.; Venerando, A.; Dall'Acqua, S.; Battistutta, R.; Ruzzene, M.; Lolli, G. Biochemical and Cellular Mechanism of Protein Kinase CK2 Inhibition by Deceptive Curcumin. *FEBS J* **2020**, *287*, 1850–1864. https://doi.org/10.1111/febs.15111.

(59) Tso, S. C.; Qi, X.; Gui, W. J.; Wu, C. Y.; Chuang, J. L.; Wernstedt-Asterholm, I.; Morlock, L. K.; Owens, K. R.; Scherer, P. E.; Williams, N. S.; Tambar, U. K.; Wynn, R. M.; Chuang, D. T. Structure-Guided Development of Specific Pyruvate Dehydrogenase Kinase Inhibitors Targeting the ATP-Binding Pocket. *Journal of Biological Chemistry* **2014**, *289*, 4432–4443. https://doi.org/10.1074/JBC.M113.533885.

(60) Moore, J. D.; Staniszewska, A.; Shaw, T.; D'Alessandro, J.; Davis, B.; Surgenor, A.; Baker, L.; Matassova, N.; Murray, J.; Macias, A.; Brough, P.; Wood, M.; Mahon, P. C. VER-246608, a Novel Pan-Isoform ATP Competitive Inhibitor of Pyruvate Dehydrogenase Kinase, Disrupts Warburg Metabolism and Induces Context-Dependent Cytostasis in Cancer Cells. *Oncotarget* **2014**, *5*, 12862–12876. https://doi.org/10.18632/ONCOTARGET.2656.

(61) Tso, S. C.; Lou, M.; Wu, C. Y.; Gui, W. J.; Chuang, J. L.; Morlock, L. K.; Williams, N. S.; Wynn, R. M.; Qi, X.; Chuang, D. T. Development of Dihydroxyphenyl Sulfonylisoindoline Derivatives as Liver-Targeting Pyruvate Dehydrogenase Kinase Inhibitors. *J Med Chem* **2017**, *60*, 1142–1150. https://doi.org/10.1021/ACS.JMEDCHEM.6B01540.

(62) Brough, P. A.; Baker, L.; Bedford, S.; Brown, K.; Chavda, S.; Chell, V.; D'Alessandro, J.; Davies, N. G. M.; Davis, B.; Le Strat, L.; Macias, A. T.; Maddox, D.; Mahon, P. C.; Massey, A. J.; Matassova, N.; McKenna, S.; Meissner, J. W. G.; Moore, J. D.; Murray, J. B.; Northfield, C. J.; Parry, C.; Parsons, R.; Roughley, S. D.; Shaw, T.; Simmonite, H.; Stokes, S.; Surgenor, A.; Stefaniak, E.; Robertson, A.; Wang, Y.; Webb, P.; Whitehead, N.; Wood, M. Application of Off-Rate Screening in the Identification of Novel Pan-Isoform Inhibitors of Pyruvate Dehydrogenase Kinase. *J Med Chem* **2017**, *60*, 2271–2286. https://doi.org/10.1021/ACS.JMEDCHEM.6B01478.

(63) Akaki, T.; Bessho, Y.; Ito, T.; Fujioka, S.; Ubukata, M.; Mori, G.; Yamanaka, K.; Orita, T.; Doi, S.; Iwanaga, T.; Ikegashira, K.; Hantani, Y.; Nakanishi, I.; Adachi, T. Fragment-Based Lead Discovery to Identify Novel Inhibitors That Target the ATP Binding Site of Pyruvate Dehydrogenase Kinases. *Bioorg Med Chem* **2021**, *44*, 116283. https://doi.org/10.1016/j.bmc.2021.116283.

(64) Su, H. xia; Yao, S.; Zhao, W. feng; Li, M. jun; Liu, J.; Shang, W. juan; Xie, H.; Ke, C. qiang; Hu, H. chen; Gao, M. na; Yu, K. qian; Liu, H.; Shen, J. shan; Tang, W.; Zhang, L. ke; Xiao, G. fu; Ni, L.; Wang, D. wen; Zuo, J. ping; Jiang, H. liang; Bai, F.; Wu, Y.; Ye, Y.; Xu, Y. chun. Anti-SARS-CoV-2 Activities in Vitro of Shuanghuanglian Preparations and Bioactive Ingredients. *Acta Pharmacol Sin* **2020**, *41*, 1167–1177. https://doi.org/10.1038/S41401-020-0483-6.

(65) Clyde, A.; Galanie, S.; Kneller, D. W.; Ma, H.; Babuji, Y.; Blaiszik, B.; Brace, A.; Brettin, T.; Chard, K.; Chard, R.; Coates, L.; Foster, I.; Hauner, D.; Kertesz, V.; Kumar, N.; Lee, H.; Li, Z.; Merzky, A.; Schmidt, J. G.; Tan, L.; Titov, M.; Trifan, A.; Turilli, M.; van Dam, H.; Chennubhotla, S. C.; Jha, S.; Kovalevsky, A.; Ramanathan, A.; Head, M. S.; Stevens, R. High-Throughput Virtual Screening and Validation of a SARS-CoV-2 Main Protease Noncovalent Inhibitor. *J Chem Inf Model* **2022**, *62*, 116–128. https://doi.org/10.1021/ACS.JCIM.1C00851.

(66) Deshmukh, M. G.; Ippolito, J. A.; Zhang, C. H.; Stone, E. A.; Reilly, R. A.; Miller, S. J.; Jorgensen, W. L.; Anderson, K. S. Structure-Guided Design of a Perampanel-Derived Pharmacophore Targeting the SARS-CoV-2 Main Protease. *Structure* **2021**, *29*, 823-833.e5. https://doi.org/10.1016/J.STR.2021.06.002.

(67) Zhang, C. H.; Spasov, K. A.; Reilly, R. A.; Hollander, K.; Stone, E. A.; Ippolito, J. A.; Liosi, M. E.; Deshmukh, M. G.; Tirado-Rives, J.; Zhang, S.; Liang, Z.; Miller, S. J.; Isaacs, F.; Lindenbach, B. D.; Anderson, K. S.; Jorgensen, W. L. Optimization of Triarylpyridinone Inhibitors of the Main Protease of SARS-CoV-2 to Low-Nanomolar Antiviral Potency. *ACS Med Chem Lett* **2021**, *12*, 1325–1332. https://doi.org/10.1021/ACSMEDCHEMLETT.1C00326.

(68) Cozza, G.; Bonvini, P.; Zorzi, E.; Poletto, G.; Pagano, M. A.; Sarno, S.; Donella-Deana, A.; Zagotto, G.; Rosolen, A.; Pinna, L. A.; Meggio, F.; Moro, S. Identification of Ellagic Acid as Potent Inhibitor of Protein Kinase CK2: A Successful Example of a Virtual Screening Application. *J Med Chem* **2006**, *49*, 2363–2366. https://doi.org/10.1021/JM060112M/SUPPL_FILE/JM060112MSI20060306_104657.PDF.

(69) Cozza, G.; Gianoncelli, A.; Bonvini, P.; Zorzi, E.; Pasquale, R.; Rosolen, A.; Pinna, L. A.; Meggio, F.; Zagotto, G.; Moro, S. Urolithin as a Converging Scaffold Linking Ellagic Acid and Coumarin Analogues: Design of Potent Protein Kinase CK2 Inhibitors. *ChemMedChem* **2011**, *6*, 2273–2286. https://doi.org/10.1002/CMDC.201100338.

(70) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual Molecular Dynamics. *J Mol Graph* **1996**, *14*, 33–38. https://doi.org/10.1016/0263-7855(96)00018-5.

(71) Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. The Amber Biomolecular Simulation Programs. *J Comput Chem* **2005**, *26*, 1668–1688. https://doi.org/10.1002/jcc.20290.

(72) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. Ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from Ff99SB. *J Chem Theory Comput* **2015**, *11*, 3696–3713. https://doi.org/10.1021/ACS.JCTC.5B00255.

(73) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and Testing of a General Amber Force Field. *J Comput Chem* **2004**, *25*, 1157–1174. https://doi.org/10.1002/JCC.20035.

(74) Jakalian, A.; Jack, D. B.; Bayly, C. I. Fast, Efficient Generation of High-Quality Atomic Charges. AM1-BCC Model: II. Parameterization and Validation. *J Comput Chem* **2002**, *23*, 1623–1641. https://doi.org/10.1002/jcc.10128.

(75) Jorgensen, W. L.; Tirado-Rives, J. The OPLS Potential Functions for Proteins. Energy Minimizations for Crystals of Cyclic Peptides and Crambin. *J Am Chem Soc* **1988**, *110*, 1657–1666. https://doi.org/10.1021/JA00214A001/ASSET/JA00214A001.FP.PNG_V03.

(76) Davidchack, R. L.; Handel, R.; Tretyakov, M. V. Langevin Thermostat for Rigid Body Dynamics. *J Chem Phys* **2009**, *130*, 234101. https://doi.org/10.1063/1.3149788.

(77) Kräutler, V.; Van Gunsteren, W. F.; Hünenberger, P. H. A Fast SHAKE Algorithm to Solve Distance Constraint Equations for Small Molecules in Molecular Dynamics Simulations. *J Comput Chem* **2001**, *22*, 501–508. https://doi.org/10.1002/1096-987X.

(78) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. A Smooth Particle Mesh Ewald Method. *J Chem Phys* **1998**, *103*, 8577. https://doi.org/10.1063/1.470117.

(79) Faller, R.; De Pablo, J. J. Constant Pressure Hybrid Molecular Dynamics–Monte Carlo Simulations. *J Chem Phys* **2001**, *116*, 55. https://doi.org/10.1063/1.1420460.

(80) Harvey, M. J.; Giupponi, G.; De Fabritiis, G. ACEMD: Accelerating Biomolecular Dynamics in the Microsecond Time Scale. *J Chem Theory Comput* **2009**, *5*, 1632–1639. https://doi.org/10.1021/ct9000685.

(81) Eastman, P.; Swails, J.; Chodera, J. D.; McGibbon, R. T.; Zhao, Y.; Beauchamp, K. A.; Wang, L.-P.; Simmonett, A. C.; Harrigan, M. P.; Stern, C. D.; Wiewiora, R. P.; Brooks, B. R.; Pande, V. S. OpenMM 7: Rapid Development of High Performance Algorithms for Molecular Dynamics. *PLoS Comput Biol* **2017**, *13*, e1005659. https://doi.org/10.1371/journal.pcbi.1005659.

(82) Michaud-Agrawal, N.; Denning, E. J.; Woolf, T. B.; Beckstein, O. MDAnalysis: A Toolkit for the Analysis of Molecular Dynamics Simulations. *J Comput Chem* **2011**, *32*, 2319–2327. https://doi.org/10.1002/JCC.21787.

(83) Gowers, R. J.; Linke, M.; Barnoud, J.; Reddy, T. J. E.; Melo, M. N.; Seyler, S. L.; Domański, J.; Dotson, D. L.; Buchoux, S.; Kenney, I. M.; Beckstein, O. MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations. *Proceedings of the 15th Python in Science Conference* **2016**, 98–105. https://doi.org/10.25080/MAJORA-629E541A-00E.

(84) Wójcikowski, M.; Zielenkiewicz, P.; Siedlecki, P. Open Drug Discovery Toolkit (ODDT): A New Open-Source Player in the Drug Discovery Field. *J Cheminform* **2015**, *7*, 26. https://doi.org/10.1186/s13321-015-0078-2.

(85) Pavan, M.; Menin, S.; Bassani, D.; Sturlese, M.; Moro, S. Implementing a Scoring Function Based on Interaction Fingerprint for Autogrow4: Protein Kinase CK1δ as a Case Study. *Front Mol Biosci* **2022**, *0*, 629. https://doi.org/10.3389/FMOLB.2022.909499.

(86) Phillips, J. C.; Hardy, D. J.; Maia, J. D. C.; Stone, J. E.; Ribeiro, J. V.; Bernardi, R. C.; Buch, R.; Fiorin, G.; Hénin, J.; Jiang, W.; McGreevy, R.; Melo, M. C. R.; Radak, B. K.; Skeel, R. D.; Singharoy, A.; Wang, Y.; Roux, B.; Aksimentiev, A.; Luthey-Schulten, Z.; Kalé, L. V.; Schulten, K.; Chipot, C.; Tajkhorshid, E. Scalable Molecular Dynamics on CPU and GPU Architectures with NAMD. *J Chem Phys* **2020**, *153*, 044130. https://doi.org/10.1063/5.0014475.

(87) Knippschild, U.; Gocht, A.; Wolff, S.; Huber, N.; Löhler, J.; Stöter, M. The Casein Kinase 1 Family: Participation in Multiple Cellular Processes in Eukaryotes. *Cell Signal* **2005**, *17*, 675–689. https://doi.org/10.1016/j.cellsig.2004.12.011.

(88) Xu, P.; Ianes, C.; Gärtner, F.; Liu, C.; Burster, T.; Bakulev, V.; Rachidi, N.; Knippschild, U.; Bischof, J. Structure, Regulation, and (Patho-)Physiological Functions of the Stress-Induced Protein Kinase CK1 Delta (CSNK1D). *Gene* **2019**, *715*.

(89) Perez, D. I.; Gil, C.; Martinez, A. Protein Kinases CK1 and CK2 as New Targets for Neurodegenerative Diseases. *Med Res Rev* **2011**, *31*, 924–954. https://doi.org/10.1002/MED.20207.

(90) Cescon, E.; Bolcato, G.; Federico, S.; Bissaro, M.; Valentini, A.; Ferlin, M. G.; Spalluto, G.; Sturlese, M.; Moro, S. Scaffold Repurposing of In-House Chemical Library toward the Identification of New Casein Kinase 1 δ Inhibitors. *ACS Med Chem Lett* **2020**, *11*. https://doi.org/10.1021/acsmedchemlett.0c00028.

(91) Bolcato, G.; Cescon, E.; Pavan, M.; Bissaro, M.; Bassani, D.; Federico, S.; Spalluto, G.; Sturlese, M.; Moro, S. A Computational Workflow for the Identification of Novel Fragments Acting as Inhibitors of the Activity of Protein Kinase CK1δ. *Int J Mol Sci* **2021**, *22*, 9741. https://doi.org/10.3390/ijms22189741.

(92) Venerando, A.; Ruzzene, M.; Pinna, L. A. Casein Kinase: The Triple Meaning of a Misnomer. *Biochem J* **2014**, *460*, 141–156. https://doi.org/10.1042/BJ20140178.

(93) Meggio, F.; Pinna, L. A. One-Thousand-and-One Substrates of Protein Kinase CK2? *FASEB J* **2003**, *17*, 349–368. https://doi.org/10.1096/FJ.02-0473REV.

(94) Borgo, C.; D'Amore, C.; Sarno, S.; Salvi, M.; Ruzzene, M. Protein Kinase CK2: A Potential Therapeutic Target for Diverse Human Diseases. *Signal Transduction and Targeted Therapy 2021 6:1* **2021**, *6*, 1–20. https://doi.org/10.1038/s41392-021-00567-7.

(95) Atas, E.; Oberhuber, M.; Kenner, L. The Implications of PDK1–4 on Tumor Energy Metabolism, Aggressiveness and Therapy Resistance. *Front Oncol* **2020**, *10*, 2835. https://doi.org/10.3389/FONC.2020.583217/BIBTEX.

(96) Zhang, S.; Hulver, M. W.; McMillan, R. P.; Cline, M. A.; Gilbert, E. R. The Pivotal Role of Pyruvate Dehydrogenase Kinases in Metabolic Flexibility. *Nutr Metab (Lond)* **2014**, *11*, 1–9. https://doi.org/10.1186/1743-7075-11-10/FIGURES/2.

(97) Patel, M. S.; Korotchkina, L. G. Regulation of the Pyruvate Dehydrogenase Complex. *Biochem Soc Trans* **2006**, *34*, 217–222. https://doi.org/10.1042/BST20060217.

(98) Wang, C.; Horby, P. W.; Hayden, F. G.; Gao, G. F. A Novel Coronavirus Outbreak of Global Health Concern. *The Lancet* **2020**, *395*, 470–473. https://doi.org/10.1016/S0140-6736(20)30185-9.

(99) *COVID Live - Coronavirus Statistics - Worldometer*. https://www.worldometers.info/coronavirus/ (accessed 2022-10-06).

(100) Dai, W.; Zhang, B.; Jiang, X. M.; Su, H.; Li, J.; Zhao, Y.; Xie, X.; Jin, Z.; Peng, J.; Liu, F.; Li, C.; Li, Y.; Bai, F.; Wang, H.; Cheng, X.; Cen, X.; Hu, S.; Yang, X.; Wang, J.; Liu, X.; Xiao, G.; Jiang, H.; Rao, Z.; Zhang, L. K.; Xu, Y.; Yang, H.; Liu, H. Structure-Based Design of Antiviral Drug Candidates Targeting the SARS-CoV-2 Main Protease. *Science (1979)* **2020**, *368*, 1331–1335. https://doi.org/10.1126/SCIENCE.ABB4489/SUPPL_FILE/ABB4489_DAI_SM.PDF.

(101) Fornasier, E.; Macchia, M. L.; Giachin, G.; Sosic, A.; Pavan, M.; Sturlese, M.; Salata, C.; Moro, S.; Gatto, B.; Bellanda, M.; Battistutta, R. A New Inactive Conformation of SARS-CoV-2 Main Protease. *Acta Crystallogr D Struct Biol* **2022**, *78*, 363–378. https://doi.org/10.1107/S2059798322000948.

(102) Ullrich, S.; Nitsche, C. The SARS-CoV-2 Main Protease as Drug Target. *Bioorg Med Chem Lett* **2020**, *30*, 127377. https://doi.org/10.1016/J.BMCL.2020.127377.

(103) Consortium, T. C. M.; Achdout, H.; Aimon, A.; Bar-David, E.; Barr, H.; Ben-Shmuel, A.; Bennett, J.; Bilenko, V. A.; Bilenko, V. A.; Boby, M. L.; Borden, B.; Bowman, G. R.; Brun, J.; BVNBS, S.; Calmiano, M.; Carbery, A.; Carney, D.; Cattermole, E.; Chang, E.; Chernyshenko, E.; Chodera, J. D.; Clyde, A.; Coffland, J. E.; Cohen, G.; Cole, J.; Contini, A.; Cox, L.; Cvitkovic, M.; Dias, A.; Donckers, K.; Dotson, D. L.; Douangamath, A.; Duberstein, S.; Dudgeon, T.; Dunnett, L.; Eastman, P. K.; Erez, N.; Eyermann, C. J.; Fairhead, M.; Fate, G.; Fearon, D.; Fedorov, O.; Ferla, M.; Fernandes, R. S.; Ferrins, L.; Foster, R.; Foster, H.; Gabizon, R.; Garcia-Sastre, A.; Gawriljuk, V. O.; Gehrtz, P.; Gileadi, C.; Giroud, C.; Glass, W. G.; Glen, R.; Glinert, I.; Godoy, A. S.; Gorichko, M.; Gorrie-Stone, T.; Griffen, E. J.; Hart, S. H.; Heer, J.; Henry, M.; Hill, M.; Horrell, S.; Huliak, V. D.; Hurley, M. F. D.; Israely, T.; Jajack, A.; Jansen, J.; Jnoff, E.; Jochmans, D.; John, T.; Jonghe, S. de; Kantsadi, A. L.; Kenny, P. W.; Kiappes, J. L.; Kinakh, S. O.; Koekemoer, L.; Kovar, B.; Krojer, T.; Lee, A.; Lefker, B. A.; Levy, H.; Logvinenko, I. G.; London, N.; Lukacik, P.; Macdonald, H. B.; MacLean, B.; Malla, T. R.; Matviiuk, T.; McCorkindale, W.; McGovern, B. L.; Melamed, S.; Melnykov, K. P.; Michurin, O.; Mikolajek, H.; Milne, B. F.; Morris, A.; Morris, G. M.; Morwitzer, M. J.; Moustakas, D.; Nakamura, A. M.; Neto, J. B.; Neyts, J.; Nguyen, L.; Noske, G. D.; Oleinikovas, V.; Oliva, G.; Overheul, G. J.; Owen, D.; Pai, R.; Pan, J.; Paran, N.; Perry, B.; Pingle, M.; Pinjari, J.; Politi, B.; Powell, A.; Psenak, V.; Puni, R.; Rangel, V. L.; Reddi, R. N.; Reid, S. P.; Resnick, E.; Ripka, E. G.; Robinson, M. C.; Robinson, R. P.; Rodriguez-Guerra, J.; Rosales, R.; Rufa, D.; Saar, K.; Saikatendu, K. S.; Schofield, C.; Shafeev, M.; Shaikh, A.; Shi, J.; Shurrush, K.; Singh, S.; Sittner, A.; Skyner, R.; Smalley, A.; Smeets, B.; Smilova, M. D.; Solmesky, L. J.; Spencer, J.; Strain-Damerell, C.; Swamy, V.; Tamir, H.; Tennant, R.; Thompson, W.; Thompson, A.; Tomasio, S.; Tsurupa, I. S.; Tumber, A.; Vakonakis, I.; Rij, R. P. van; Vangeel, L.; Varghese, F. S.; Vaschetto, M.; Vitner, E. B.; Voelz, V.; Volkamer, A.; Delft, F. von; Delft, A. von; Walsh, M.; Ward, W.; Weatherall, C.; Weiss, S.; White, K. M.; Wild, C. F.; Wittmann, M.; Wright, N.; Yahalom-Ronen, Y.; Zaidmann, D.; Zidane, H.; Zitzmann, N. Open Science Discovery of Oral Non-Covalent SARS-CoV-2 Main Protease Inhibitor Therapeutics. *bioRxiv* **2022**, 2020.10.29.339317. https://doi.org/10.1101/2020.10.29.339317.

(104) Luttens, A.; Gullberg, H.; Abdurakhmanov, E.; Vo, D. D.; Akaberi, D.; Talibov, V. O.; Nekhotiaeva, N.; Vangeel, L.; Jonghe, S. de; Jochmans, D.; Krambrich, J.; Tas, A.; Lundgren, B.; Gravenfors, Y.; Craig, A. J.; Atilaw, Y.; Sandström, A.; Moodie, L. W. K.; Lundkvist, Å.; Hemert, M. J. van; Neyts, J.; Lennerstrand, J.; Kihlberg, J.; Sandberg, K.; Danielson, U. H.; Carlsson, J. Ultralarge Virtual Screening Identifies SARS-CoV-2 Main Protease Inhibitors with Broad-Spectrum Activity against Coronaviruses. *J Am Chem Soc* **2022**, *144*, 2905–2920. https://doi.org/10.1021/JACS.1C08402.

(105) Ghahremanpour, M. M.; Tirado-Rives, J.; Deshmukh, M.; Ippolito, J. A.; Zhang, C. H.; Cabeza De Vaca, I.; Liosi, M. E.; Anderson, K. S.; Jorgensen, W. L. Identification of 14 Known Drugs as Inhibitors of the Main Protease of SARS-CoV-2. *ACS Med Chem Lett* **2020**, *11*, 2526–2533. https://doi.org/10.1021/ACSMEDCHEMLETT.0C00521/SUPPL_FILE/ML0C00521_SI_002.XLSX.

(106) Owen, D. R.; Allerton, C. M. N.; Anderson, A. S.; Aschenbrenner, L.; Avery, M.; Berritt, S.; Boras, B.; Cardin, R. D.; Carlo, A.; Coffman, K. J.; Dantonio, A.; Di, L.; Eng, H.; Ferre, R.; Gajiwala, K. S.; Gibson, S. A.; Greasley, S. E.; Hurst, B. L.; Kadar, E. P.; Kalgutkar, A. S.; Lee, J. C.; Lee, J.; Liu, W.; Mason, S. W.; Noell, S.; Novak, J. J.; Obach, R. S.; Ogilvie, K.; Patel, N. C.; Pettersson, M.; Rai, D. K.; Reese, M. R.; Sammons, M. F.; Sathish, J. G.; Singh, R. S. P.; Steppan, C. M.; Stewart, A. E.; Tuttle, J. B.; Updyke, L.; Verhoest, P. R.; Wei, L.; Yang, Q.; Zhu, Y. An Oral SARS-CoV-2 M pro Inhibitor Clinical Candidate for the Treatment of COVID-19. *Science (1979)* **2021**, *374*, 1586–1593. https://doi.org/10.1126/science.abl4784.

(107) Pavan, M.; Bolcato, G.; Bassani, D.; Sturlese, M.; Moro, S. Supervised Molecular Dynamics (SuMD) Insights into the Mechanism of Action of SARS-CoV-2 Main Protease Inhibitor PF-07321332. *J Enzyme Inhib Med Chem* **2021**, *36*, 1646–1650. https://doi.org/10.1080/14756366.2021.1954919.

(108) Cheng, S. C.; Chang, G. G.; Chou, C. Y. Mutation of Glu-166 Blocks the Substrate-Induced Dimerization of SARS Coronavirus Main Protease. *Biophys J* **2010**, *98*, 1327–1336. https://doi.org/10.1016/j.bpj.2009.12.4272.

(109) Proctor, E. A.; Yin, S.; Tropsha, A.; Dokholyan, N. V. Discrete Molecular Dynamics Distinguishes Nativelike Binding Poses from Decoys in Difficult Targets. *Biophys J* **2012**, *102*, 144–151. https://doi.org/10.1016/J.BPJ.2011.11.4008.

(110) Liu, K.; Kokubo, H. Exploring the Stability of Ligand Binding Modes to Proteins by Molecular Dynamics Simulations: A Cross-Docking Study. *J Chem Inf Model* **2017**, *57*, 2514–2522. https://doi.org/10.1021/ACS.JCIM.7B00412/SUPPL_FILE/CI7B00412_SI_001.PDF.

CONCLUSIONS AND FUTURE PERSPECTIVES

The present work was conceived amid the COVID-19 pandemic, one of the hardest striking illnesses that the human genre has faced in a long period. To alleviate the socio-economic burden that the global society had to sustain, an unprecedented cooperative effort was carried out by scientists from all over the world to accelerate the usually long time required to identify a cure for emerging diseases.

In a constantly evolving scenario, where decisions must be taken promptly to avoid unnecessary collateral damage associated with the pandemic, computer simulations played a pivotal role in providing quick and accurate responses to the ever-emerging doubts of the scientific community. Specifically, in the field of drug discovery various existing computational strategies were successfully exploited to guide the rational development of efficient treatments against COVID-19, other than excluding harmful or ineffective ones.

The remarkable amount of structural information rendered available to scientists, coupled with highly efficient structural prediction methodologies, enabled the exploitation of state-of-the-art structure-based drug discovery protocols. Molecular docking, a central player within SBDD drug discovery, played a prominent role also in several campaigns, aimed at targeting different viral targets with different therapeutic tools, ranging from small organic molecules to macromolecules such as peptides, monoclonal antibodies, and aptamers.

Despite its indisputable usefulness, docking on its own was not sufficient, therefore several complementary techniques like pharmacophore modeling, molecular dynamics, and binding free energy estimation methods were often exploited to create unique pipelines.

Although the combination of various existing tools proved useful in identifying several different active molecules, some of which even advanced to clinical stage experimentation or even reached approval by regulatory agencies, various works highlighted how there is still much room for improvement of already established pipelines. The centrality of molecular docking causes most structure-based campaigns to focus only on the final state of the ligand-receptor recognition process, disregarding the

long process that precedes the bound state, other than the time-dependent evolution of such a complex. Techniques such as molecular dynamics, which could be readily applied to mechanistic binding studies, are not routinely utilized within most SBDD campaigns yet due to the long simulation times that are required to obtain meaningful pieces of information, which hamper the high-throughput nature of modern days drug discovery pipelines.

To bridge the gap between the existence of techniques that more thoroughly describe the binding event and their actual implementation in drug discovery projects, we developed two MD-based enhanced sampling protocols. The first protocol is an evolved version of Supervised Molecular Dynamics (SuMD), an established methodology for studying receptor-ligand association processes, applied to the study of RNA-protein interactions. The centrality of RNA-based therapeutics, of which aptamers represent the most prominent example, led us to expand on the applicability domain of the technique and make it suitable also for the study of these systems. This work will not only impact the discovery of RNA-based therapeutic or diagnostic tools aimed toward COVID-19 but would most likely become a central point of other projects soon, thanks to the increasing relevance of RNA within the drug discovery environment.

The second protocol, instead, is a brand-new methodology named Thermal Titration Molecular Dynamics (TTMD), an MD-based enhanced sampling protocol for studying receptor-ligand dissociation processes. The marked interest of medicinal chemists in the determination and prediction of binding kinetic parameters that sparked in the last fifteen years render TTMD a potentially precious tool for drug development, thanks to its ease of use compared to similar existing protocols, and to its feasible simulation times.

Overall, the present work constitutes the basis for the development of efficient therapeutics against COVID-19 and for future similar scenarios, thanks to the structural and methodological information collected. Furthermore, the development of novel CADD methodologies represents a good starting point for their broader application, especially for systems that have not yet been thoroughly disclosed. Further development of the TTMD protocol to extend its applicability domain to systems involving membrane protein or nucleic acids, and different purposes such as the docking pose selection are already

going on within the Molecular Modeling Section laboratory and will be part of future scientific works.