UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Head Office: Università degli Studi di Padova

Department of Chemical Sciences

Ph.D COURSE IN: Molecular Sciences
CURRICULUM: Pharmaceutical Sciences
CYCLE XXXIV

COMBINING COMPUTATIONAL TIME-INDEPENDENT WITH TIME-DEPENDENT APPROACHES IN DRUG DISCOVERY

Coordinator: Ch.mo Prof. Leonard Jan Prins
Supervisor: Ch.mo Prof. Stefano Moro

Ph.D Student: Giovanni Bolcato

ABSTRACT

Computer aided drug discovery (CADD) approaches have affirmed their role in many industrial and academic contexts as precious tools to rationalize and speed up the early stages of drug discovery pipelines. Starting from simple approaches focus only on the ligand properties and activities (ligand based methods), with the increasing number of available structural information regarding receptors a shift toward a new class of methods, called structure based approaches, occurred during the past 30 years.

This PhD thesis focus on this class of CADD approaches (homology modelling, molecular docking, molecular dynamics) and show different application of these methods in various contexts: from virtual screening to the elucidation of the protein-ligand recognition process at the atomistic level, from small soluble proteins to G-Protein coupled receptors.

Summary

INTRODUCTION

1 History of Drug Discovery.

1.1 Medicines: From plants to synthetic molecules.

For several thousand years different plant parts (roots, leaves, bark) have been used as medicines in all continents of the world. These plants were used to heal various diseases based on empirical observations of symptoms relieves like fever or pain.

Along with the use of the plant itself, the use of extracts (like dried extracts or infusions) has always been present, as evidence of the awareness raised towards the presence of active ingredients in the plant, for example, the use of extracts of *Ephedra sinica*as a stimulant can be dated up to 3000 B.C. in traditional Chinese medicine[1].

In the nineteenth century remarkable progress has been made with the isolation of the pure active ingredients from a plant: In 1804 Morphine was isolated from opium and in 1820 Quinine is isolated from cinchona bark. Along with this progress in 1828 is reported the first synthesis of an organic molecule, urea. Soon organic synthesis was applied for the synthesis of natural molecules, and in 1860 salicylic acid was synthesized for the first time and in the late 1800s, the extraction of this molecule from the natural source was replaced by its synthesis.

The real revolution in the history of medicines was the application of organic synthesis for the preparation of new molecules. That is, not the replication of molecules that are already present in nature but the creation of new molecular entities. Acetylsalicylic acid was first prepared in 1853 from salicylic acid, but while this is a semisynthetic derivate of a natural product, chloral hydrate can be considered the first synthetic drug (1832)[2].

While the application of organic synthesis for the creation of new molecules is a powerful concept, this new paradigm implies a new problem: which are the new molecules that have to be made? This question can be considered the central problem of modern Drug discovery.

1.2 From Serendipity to Rational Drug Discovery.

The drug discovery process can be ideally divided into three steps: Hit identification hit to lead optimization and lead optimization. A hit compound is a molecule that gives a positive response in a specific assay (an enzymatic inhibition assay for example), this molecule is optimized to improve not only the activity toward the target but also physicochemical properties, like solubility and lipophilicity, to give a lead compound. When more advanced studies on the lead compound have been performed, this is further optimized to improve properties like metabolic stability, toxicity, off-target effects.

During the 20th century many new hit compounds were often discovered by accident (often refers as serendipity in Drug Discovery) famous examples of these lucky events are the discovery of the first Penicillin and Librium. A huge progress was made with the setup of reliable and reproducible assays to test molecules for a specific purpose and the automation of this process, an approach known as High-throughput screening (HTS).

The optimization process of the hit compound (that could be a new molecule like Librium or a natural product like morphine) was then carried out by synthesizing and testing many different analogs. Soon it was clear that some trends were present in the variation of the measured activity of the analogs series: some modifications were not tolerated (like the basic nitrogen in opioids, that must be present), the nature (bulkiness, hydrophobicity, etc.) of a substituent in a certain position correlate quite well with the activity, the length of a linker between two regions of the molecule must be in a restricted range, the cyclization of a part of the molecule improved the activity and many other observations that are collectively called Structure-Activity Relationships (SARs).

The next step was the development of quantitative models to use these SARs (QSAR) in a perspective way, several equations were made for this purpose (or already existing models were applied to drug discovery): Hansch's equation, Hammet's equation, Taft's Equation[3].

These equations were limited to small variations in highly congeneric series and often works better in the interpolative validation rather than in the extrapolative predictions, nevertheless, with these first approaches it was clear that rationalizing the early stages of drug discovery campaigns could be useful and powerful. Since these first approaches were quite limited, the necessity of more advanced models was clear, and with this necessity, informatics makes its entrance in Drug Discovery.

The next generation of approaches were the so-called 3D-QSAR models. In these methods, several conformers are generated *in silico* and a multiple conformer alignment is performed (if a common scaffold is present this can be kept fixed and used to align all the conformers, for example). At this point a steric and electrostatic potential (in the classic COMFA approach) and eventually the hydrophobic potential (in the CoMSIA method) are calculated in a 3D grid surrounding the molecule, the obtained values are then correlated with the measured activity[4].

Other *in silico* approaches were developed not only for the lead optimization but also for the hit finding process. These methods start usually from an active compound and search similar molecules in virtual databases. To make this comparison, molecules can be represented in several

ways: using a set of molecular descriptors, using fingerprints (1D representation of the molecule obtained with different methods), representing the molecules with their 3D shape. More advanced methods can measure the similarity between molecules by comparing the electrostatic potentials exerted by these.

Another approach that can be used to search new hit compounds is pharmacophore screening. When some active molecules for a specific target are known, these can be aligned to search a common pharmacophore, a set of features (hydrogen bond donor and acceptor, aromatic ring, and so on) that are present in all the molecules. This pharmacophore is then used to screen large virtual databases, searching for new molecules that can fit the model.

While these methods have represented a great aid in the early stages of the many drug discovery processes, their main limitation was the high number of false positives, molecules that can fit very well a pharmacophore model or that are similar to a known active compound but when tested result not active. This high rate of false positives is mainly caused by neglecting the receptor role in the interaction.

2 Structure-Based Drug Discovery.

Starting from the late 90s the number of X-Ray, NMR, and electron microscopy structures in the Protein Data Bank (PDB[5]) has started to increase rapidly. In the beginning, the solved structures were mainly small soluble proteins, but in the 2000s many membrane proteins structures have been solved[6]. Often the protein's structure is solved as a complex with a small organic molecule, like drugs or other active compounds. This precious structural information has soon been used to develop the next generation of computational methods in drug discovery, usually called Structure-Based Drug Discovery approaches.

2.1 Molecular Docking

A molecular docking protocol is a computational tool that predicts the binding mode of a given molecule inside a site of a protein. It is made of a search algorithm that samples the conformational space accessible by the ligand, creating conformations that in the docking context are usually called poses, and of a scoring function, which evaluates the quality of such poses to assess which is the most probable (namely the one more similar to the real binding mode, the pose that would be observed in a crystal structure of the protein-ligand complex).

Note that while the aforementioned conformational search was operated on the ligand alone to find low energy and less constrained conformers, molecular docking can be thought of as a conformational search that takes into account (from a shape and interactive point of view) a protein environment surrounding the ligand.

Different scoring functions have been adopted for the evaluation of docking poses. Scoring functions can be divided into different families: 1. Force field-based scoring functions, where the energy of the system is evaluated using a force field, this model the energy of the system as a function of the sum of different functional terms and the values of the parameters used in these terms. 2. Empirical scoring functions[7,8] consist of different terms, each of which represents a different intermolecular interaction, every term is modeled using experimental values for that intermolecular interaction (so, for example, an angle of 180° and a length of 3Å are considered optimal for a hydrogen bond). The quality of the docking pose is evaluated by how far the system is from these experimental values and counting the number of positive interactions. 3. Knowledge-based scoring functions. These are based on statistical analyses on the most observed contacts between certain ligand's atom type and certain protein's atom type. The Docking poses that are more similar to what is statistically observed in high-quality X-ray databases are preferred.

Regarding the sampling of the conformational space, any type of search algorithm has been implemented in docking protocols over the years. Systematic search algorithms that exhaustively sample the conformational space defined by the degree of freedom of the ligand, heuristic and metaheuristic algorithms like Genetic algorithms[9,10] and Ant Colony Optimization algorithms[11,12].

One of the major limitations of docking is that the search algorithm samples the conformational space of the ligand while the protein is usually kept rigid (this approach is the so-called semi-flexible docking). However, there are some strategies to partially take into account the flexibility of the protein. The simplest way is to introduce some tolerance in the steric clashes between the protein and the ligand, with the logic that the protein can tolerate some steric clashes with the ligand adapting the binding site's residues[13]. A more sophisticated approach retains the flexibility of the side chains for the protein's residues in the binding site, the conformational space of these side chains is explored in a similar way to what is done for the ligand. Often only the position of some atoms of the side chains is optimized, these atoms are often polar hydrogens involved in hydrogen bonds. A different approach, called ensemble docking, involves the execution of docking run on different conformations for the protein (usually extracted from Molecular Dynamics trajectories), this way the flexibility of the protein is indirectly considered[14].

Molecular docking has been applied in both the three fundamental steps of drug design: the hit finding, the hit to lead optimization, and the lead optimization. The application of docking to find new hit compounds is called Docking-based virtual screening[15,16] or Structure-based virtual screening[17,18]. In this type of virtual screening, large virtual databases (up to billions of compounds[19]) are docked in the binding site of the protein of interest, and the poses are ranked according to the scoring function value. Note that while in the classic example discussed above the scoring function ranks the different poses of a molecule in virtual screening the scoring function is used to rank different poses of different molecules, with the aim to find the most promising molecules to test against the target.

Once a hit compound is identified usually several analogs are prepared and tested. Molecular Docking can be a useful tool to rationalize the observed SARs for a ligand series, from a Structure-based point of view. For this purpose, a common binding mode that can explain the observed SARs is searched. The decomposition of the scoring function value in the different constitutive terms (hydrogen bonds, electrostatics, etc.) can also help to understand the experimental observations.

Once a docking model has been validated it can be used to rational suggest modifications that can be made on the hit compound to improve the affinity.

It is well accepted that while usually, the conformational sampling of docking can often reproduce the crystallographic binding mode, scoring functions struggle to distinguish the native binding mode from wrong poses[20] (figure 1).

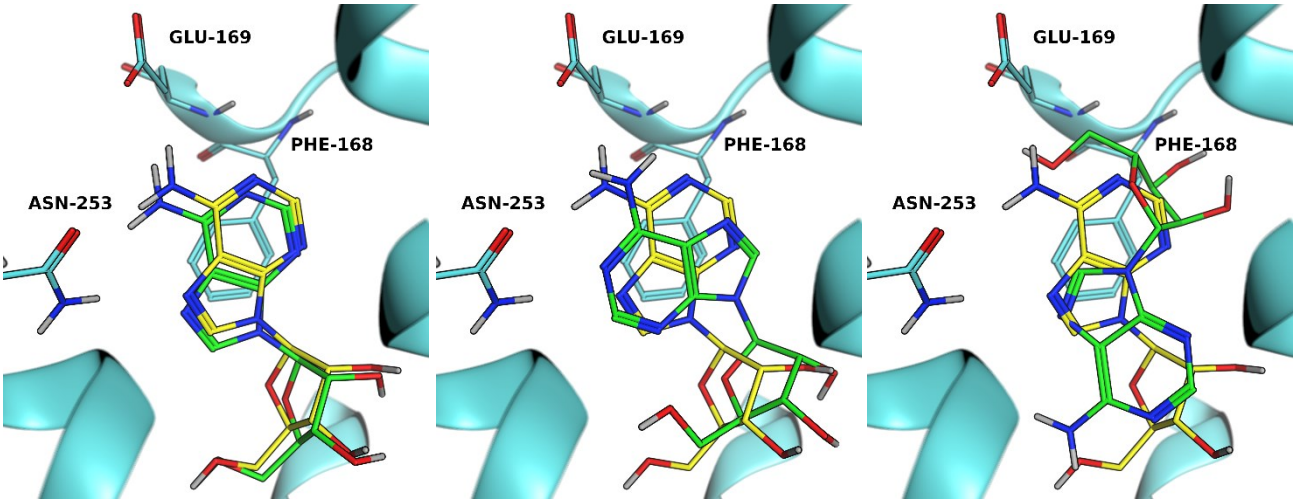


Figure 1. Three different Docking poses (green) of Adenosine and the crystallographic observed binding mode (yellow) in the orthosteric site of A_{2A} adenosine receptor. All the three poses are reasonable but only one is correct (close to the experimental observed one). Note that for each pose the interaction pattern observed in the crystal structure is preserved (two hydrogen bonds with Asn-253, electrostatic interaction with Glu-169, π-π stacking with Phe-168).

The reasons why often docking fails are many because several are the implicit approximations in the way docking models the protein-ligand problem. The role of water molecules is neglected and the protein is kept rigid or almost rigid, often only one or few tautomeric and charge states of the ligand are considered. Finally, the experimental values, like Kd, that we are trying to rationalize are the result of a dynamic process of association and dissociation of the ligand, while with docking all the considerations are made on the final bound state.

For these reasons, many approaches have been developed to improve docking performance.

2.2 Pharmacophore constraints in molecular docking

If for the target of interest more than one X-ray crystal structure of the protein-ligand complex is available, it is possible to model a pharmacophore hypothesis based on the commonly observed features, these structure-based pharmacophores are more reliable than the ligand-based models derived from multiple alignments of different conformers of a dataset of active molecules, especially if few active molecules are known[21,22]. Once this model has been validated it can be used to filter docking poses resulting from virtual screening, retaining only those that fit the pharmacophore model. Note that in this approach the scoring function values can also be neglected, being replaced by a knowledge-based selection criterion. Some docking software like MOE, Glide, and GOLD can also implement these pharmacophore constraints directly in the docking run, biasing the solutions toward the desired binding mode, if possible.

2.3 Consensus Docking

Consensus docking is an approach that tries to improve the performance of docking with a "wisdom of the crowd" logic. In this approach docking poses resulting from a virtual screening are rescored using several scoring functions and those with good values for the different metrics are prioritized[23]. Sometimes the consensus approach is also applied to the conformational search: the molecules that produce similar poses using different search algorithms are prioritized[24].

2.3 Water molecules in molecular docking

The importance that water molecules play in the ligand-receptor interaction[25,26]. In the ligand-receptor interaction perspective, water molecules can be divided into two categories: those which are displaced by the ligand during the recognition process, and those which are not displaced but can be stabilized by the ligand and participate in the complex formation as a third actor[27]. Like for the above-mentioned structure-based pharmacophores, also to model water molecules in a

protein's binding site, several low-resolution X-ray crystal structures are required. If these are available, several docking software can include water molecules in the calculation. While some docking software like GOLD automatically calculates the optimal orientation of the water molecules, several others require the user to manually set the proper orientation[28].

The performance of docking usually increases when water molecules are included, this can be due to two main reasons. For some targets, like the HIV protease[29], water molecules mediate the interaction of the ligand in the binding site of the protein, and the experimental binding mode cannot be reproduced without them. But it also must be remembered that by adding water molecules, we are also adding excluded volumes, so there is less accessible conformational space to be explored, the number of possible solutions Is therefore less.

There are some problems in the inclusion of water molecules in docking calculation for virtual screening purposes since different ligands can displace or not displace different solvent molecules and interact with these in different ways, so adopting a single structural configuration for the solvent can lead to misleading results in virtual screening.

2.4 The missing dimension: Molecular dynamics refinement of docking poses.

Molecular dynamics (often abbreviated in MD) is a class of computational simulations where the time evolution of a molecular system is analyzed. Starting from an initial atomic configuration the evolution of the system in time is simulated using Newton's equation of motion, which is integrated at every time step which are the short intervals in which time is discretized[30,31]. The length of the time step must assure an appropriate description of the fastest motion of the system, which is usually bonded vibration. The force acting on each particle is calculated using a force field and is then used to calculate the acceleration of the particle and so the position at the next time step. Many force fields have been developed for the simulation of biological systems: AMBER[32], CHARMM[33], OPLS[34] are the most used among many others.

Molecular dynamics simulations have been applied in recent years to refine docking poses[35,36]. Typically, the protein-docking pose complex is subjected to multiple MD simulations which are analyzed to assess the stability of the predicted binding mode. This stability is modeled as the root mean square fluctuation (RMSF) of the atomic positions during the trajectories, the persistence of particular interactions of interest (like the hydrogen bonds with hinge residues for a kinase inhibitor) can also be measured and used as a metric to evaluate the pose.

MD refinement of docking poses allows users to assess the severity of steric clashes between the protein and the ligand. While scoring functions values become quickly unfavorable in the presence of steric clashes, an MD refinement can be useful to understand if the protein can or not tolerate that steric clash by a conformational rearrangement and if the binding mode is preserved during this event.

The implementation of molecular dynamics simulations allows more accurate treatment of the solvent concerning what is described above and also solves some of the problems associated with these classic protocols. First, no “a priori” information regarding the position of water molecules is needed. Indeed, the molecular system is solvated and during the simulation, water molecules tend to stabilize in the same positions observed in X-ray crystal structures[37]. The problem of water molecules' orientation is also intrinsically solved since these tend to orient themselves dynamically to form interactions with the ligand and the protein[38].

MD refinement of docking poses allows a more complete and accurate vision of the protein-ligand complex, but some limitations are still present and this type of investigation still focuses only on the final bound state, while a full description of the molecular recognition process requires also an analysis of the (un)binding pathway.

3. Supervised Molecular Dynamics

One of the major limitations of classic Molecular Dynamics simulations is the low sampling of the potential energy surface described by the force field. This means that usually, the system ends up in local minima, separated by a high energy barrier from other minima, without a complete exploration of the energetic landscape. To overcome this problem different methods have been developed, which fall into two main categories: Enhanced sampling methods and Markov State Models methods.

Markov state models approaches[39,40] to treat an MD event as an ensemble of independent microstates and calculate a transition probability matrix that allows computing the probability that the system occupies a certain state and the probability that the system transitions in another state.

Enhanced sampling methods[41,42] are based on an alteration of the potential energy surface to escape from local minima.

SuMD[43,44,45] (Supervised Molecular Dynamics)is a Molecular Dynamics based approach that allows the investigation of molecular recognition events without altering the potential energy surface

with biases. The algorithm supervises the distance between the ligand's center of mass distance and the binding site's center of mass (binding site is defined as an ensemble of residues).

In a SuMD simulation a series of short classical MD simulations are performed (these are usually referred to as SuMD steps) and at the end of each of these small simulations, the distance between the two centers of mass is measured. If the ligand is approaching the binding site during the SuMD step, the simulation is prolonged by another SuMD step, otherwise, the simulation is restarted from the previous set of coordinates. So while enhanced sampling methods use an energetic bias to sample binding events (and other molecular events) SuMD adopts a pure geometric bias, without altering the potential energy surface.

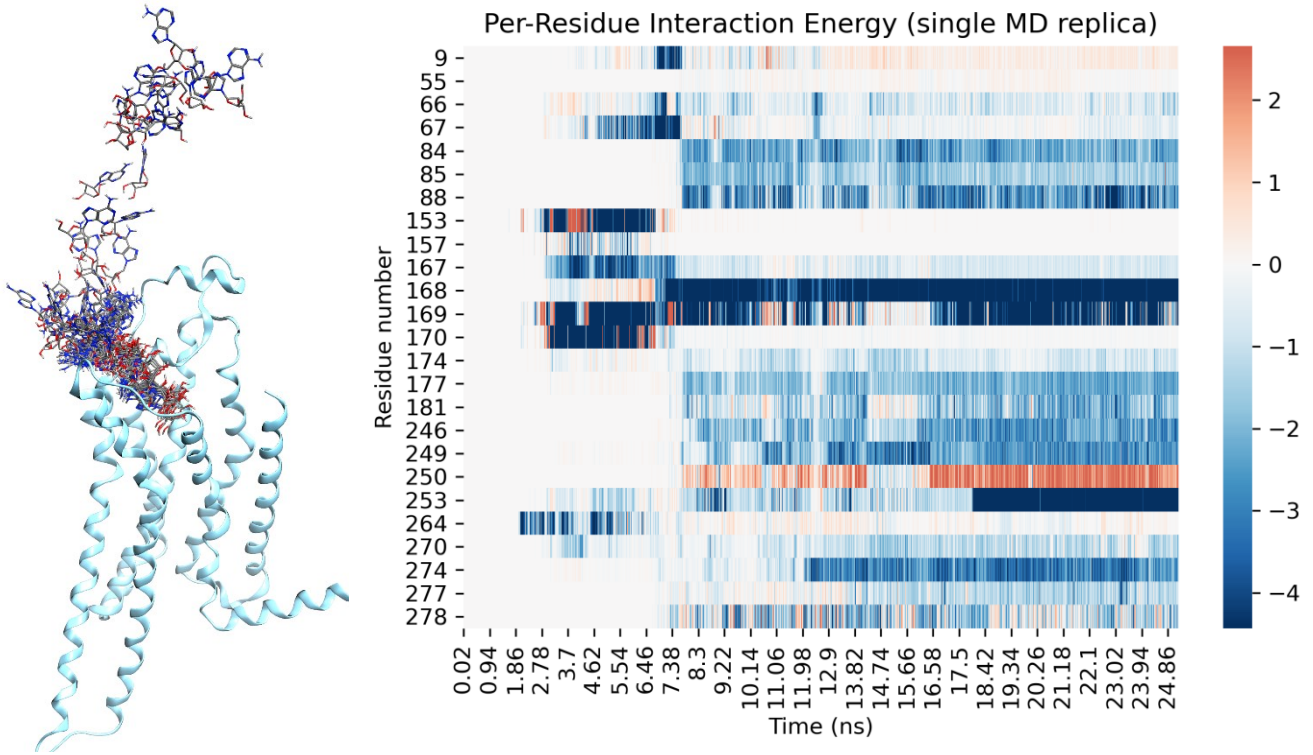


Figure 2.A possible binding trajectory of adenosine toward the orthosteric binding site of A_{2A} adenosine receptor,sampled SuMD simulation (on the left) and the per residue decomposition of the interaction energy as a function of the time, on the right.

SuMD simulation allows a more accurate and complete depiction of the binding event than molecular docking calculations. The molecular mechanism of the binding event (and also of the unbinding process[46]) can be analyzed at the atomistic level, gaining information on how the ligand is recruited from the bulk by the protein. These kind of information cannot be obtained by docking analysis, and are fundamental in the description of the molecular recognition process. For example, the importance of extracellular loops in ligand potency and selectivity has been proved for GPCRs[47].

In some cases, meta-stable binding sites can be observed along the binding pathway toward the final orthosteric site. In figure 2 is reported a per residue analysis of the interaction energy between adenosine and A_{2A} adenosine receptors during a SuMD simulation of the binding process. As it can be observed from this example, before the ligand reaches the final X-ray observed binding mode, interacting with residues 168, 169, and 253 (see figure 1), a meta-stable binding site is observed (between 2ns and 7ns). Here the ligand is interacting with some residues (153,157 and 170) that are not present after, in the final bound state.

Behind small molecules, SuMD has recently been applied to peptides[48] and fragments[38,49] which represent two difficult classes of ligands for molecular docking.

References

1. Dias, D. A., Urban, S. & Roessner, U. A Historical Overview of Natural Products in Drug Discovery. *Metabolites***2**, 303–336 (2012).

2. Jones, A. W. Early drug discovery and the rise of pharmaceutical chemistry. *Drug Test. Anal.***3**, 337–344 (2011).

3. Silakari, O. & Singh, P. K. QSAR: Descriptor calculations, model generation, validation and their application. in *Concepts and Experimental Protocols of Modelling and Informatics in Drug Design* 29–63 (Elsevier, 2021). doi:10.1016/B978-0-12-820546-4.00002-7.

4. Kim, K. H., Greco, G. & Novellino, E. A Critical Review of Recent CoMFA Applications. in *3D QSAR in Drug Design* 257–315 (Kluwer Academic Publishers). doi:10.1007/0-306-46858-1_16.

5. Berman, H. M. The Protein Data Bank: a historical perspective. *Acta Crystallogr. Sect. A Found. Crystallogr.***64**, 88–95 (2008).

6. Moraes, I., Evans, G., Sanchez-Weatherby, J., Newstead, S. & Stewart, P. D. S. Membrane protein structure determination — The next generation. *Biochim. Biophys. Acta - Biomembr.***1838**, 78–87 (2014).

7. 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).

8. Guedes, I. A., Pereira, F. S. S. & Dardenne, L. E. Empirical Scoring Functions for Structure-Based Virtual Screening: Applications, Critical Aspects, and Challenges. *Front. Pharmacol.***9**, (2018).

9. 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.

10. 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.

11. Korb, O., Stützle, T. & Exner, T. E. An ant colony optimization approach to flexible protein–ligand docking. *Swarm Intell.* (2007) doi:10.1007/s11721-007-0006-9.

12. 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.

13. Jiang, F. & Kim, S.-H. “Soft docking”: Matching of molecular surface cubes. *J. Mol. Biol.***219**, 79–102 (1991).

14. Amaro, R. E. *et al.*Ensemble Docking in Drug Discovery. *Biophys. J.***114**, 2271–2278 (2018).

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

16. Tuccinardi, T. Docking-Based Virtual Screening: Recent Developments. *Comb. Chem. High Throughput Screen.***12**, 303–314 (2009).

17. Li, Q. & Shah, S. Structure-Based Virtual Screening. in 111–124 (2017). doi:10.1007/978-1-4939-6783-4_5.

18. Pihan, E., Kotev, M., Rabal, O., Beato, C. & Diaz Gonzalez, C. Fine tuning for success in structure-based virtual screening. *J. Comput. Aided. Mol. Des.***35**, 1195–1206 (2021).

19. Bender, B. J. *et al.* A practical guide to large-scale docking. *Nat. Protoc.***16**, 4799–4832 (2021).

20. 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.

21. KB, S. *et al.* Structure based pharmacophore modelling approach for the design of azaindole derivatives as DprE1 inhibitors for tuberculosis. *J. Mol. Graph. Model.***101**, 107718 (2020).

22. Gaurav, A. & Gautam, V. Structure-based three-dimensional pharmacophores as an alternative to traditional methodologies. *J. Receptor. Ligand Channel Res.* 27 (2014) doi:10.2147/JRLCR.S46845.

23. 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).

24. Houston, D. R. & Walkinshaw, M. D. Consensus Docking: Improving the Reliability of Docking in a Virtual Screening Context. *J. Chem. Inf. Model.***53**, 384–390 (2013).

25. Roberts, B. C. & Mancera, R. L. Ligand−Protein Docking with Water Molecules. *J. Chem. Inf. Model.***48**, 397–408 (2008).

26. Wong, S. E. & Lightstone, F. C. Accounting for water molecules in drug design. *Expert Opin. Drug Discov.***6**, 65–74 (2011).

27. Barillari, C., Taylor, J., Viner, R. & Essex, J. W. Classification of Water Molecules in Protein Binding Sites. *J. Am. Chem. Soc.***129**, 2577–2587 (2007).

28. Verdonk, M. L. *et al.*Modeling Water Molecules in Protein−Ligand Docking Using GOLD. *J. Med. Chem.***48**, 6504–6515 (2005).

29. Suresh, C. H., Vargheese, A. M., Vijayalakshmi, K. P., Mohan, N. & Koga, N. Role of structural water molecule in HIV protease-inhibitor complexes: A QM/MM study. *J. Comput. Chem.***29**, 1840–1849 (2008).

30. Adcock, S. A. & McCammon, J. A. Molecular Dynamics: Survey of Methods for Simulating the Activity of Proteins. *Chem. Rev.***106**, 1589–1615 (2006).

31. Hollingsworth, S. A. & Dror, R. O. Molecular Dynamics Simulation for All. *Neuron***99**, 1129–1143 (2018).

32. 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.

33. 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.* (2010) doi:10.1002/jcc.21367.

34. 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).

35. Rastelli, G. & Pinzi, L. Refinement and Rescoring of Virtual Screening Results. *Front. Chem.***7**, (2019).

36. Kapla, J., Rodríguez-Espigares, I., Ballante, F., Selent, J. & Carlsson, J. Can molecular dynamics simulations improve the structural accuracy and virtual screening performance of GPCR models? *PLOS Comput. Biol.***17**, e1008936 (2021).

37. 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).

38. 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).

39. Husic, B. E. & Pande, V. S. Markov State Models: From an Art to a Science. *J. Am. Chem. Soc.***140**, 2386–2396 (2018).

40. Warfield, B. M. & Anderson, P. C. Molecular simulations and Markov state modeling reveal the structural diversity and dynamics of a theophylline-binding RNA aptamer in its unbound state. *PLoS One***12**, e0176229 (2017).

41. 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).

42. Bernardi, R. C., Melo, M. C. R. & Schulten, K. Enhanced sampling techniques in molecular dynamics simulations of biological systems. *Biochim. Biophys. Acta - Gen. Subj.***1850**, 872–877 (2015).

43. 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).

44. Salmaso, V., Sturlese, M., Cuzzolin, A. & Moro, S. Exploring Protein-Peptide Recognition Pathways Using a Supervised Molecular Dynamics Approach. *Structure* (2017) doi:10.1016/j.str.2017.02.009.

45. 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).

46. Deganutti, G., Moro, S. & Reynolds, C. A. A Supervised Molecular Dynamics Approach to Unbiased Ligand–Protein Unbinding. *J. Chem. Inf. Model.***60**, 1804–1817 (2020).

47. Nguyen, A. T. N. *et al.* Extracellular loop 2 of the adenosine A1 receptor has a key role in orthosteric ligand affinity and agonist efficacy. *Mol. Pharmacol.* (2016) doi:10.1124/mol.116.105007.

48. 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**, (2021).

49. 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.

Scientific Publications

Overview of the Scientific publications

1 Molecular Docking

1.1 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, 1168–1174 (2020).

In this work three different Molecular Docking protocols have been used (consensus docking logic) to identify two novel inhibitors of CK1δ, the library used was initially designed for other targets: the work can be considered an example of the so called scaffold repurposing approach.

1.2 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).

For this project the same approach mentioned above has been applied for the identification of novel chemotypes that can inhibit CK1δ, starting from a database of commercialy available fragments. The Docking poses have been refined using Molecular Dynamics simulations.

2 Molecular Dynamics

2.1 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).

Using HSP90 as a case study, in this work I studied the role of structural water molecules in the accuracy of Docking predictions and if the information derived from Molecular Dynamics simulations can replace these data when crystallographic solvent molecules are lacking in the crystal structure.

2.2 Bissaro, M., Bolcato, G., Deganutti, G., Sturlese, M., & Moro, S. (2019). Revisiting the Allosteric Regulation of Sodium Cation on the Binding of Adenosine at the Human A2A Adenosine Receptor: Insights from Supervised Molecular Dynamics (SuMD) Simulations. Molecules, 24(15), 2752.

In this work I used Supervised Molecular Dynamics simulations to investigate how the structural sodium ion in A2A Adenosine Receptor, could influence the molecular recognition of Adenosine.

2.3 Palazzotti, D. et al. Deciphering the molecular recognition mechanism of multidrug resistance staphylococcus aureus nora efflux pump using a supervised molecular dynamics approach. Int. J.

Mol. Sci. (2019)

Supervised Molecular Dynamics simulations have been applied here to elucidate the molecular recognition process of NorA substrates. NorA is an efflux pump involved in antibiotics resistance.

2.4 Bolcato, G., Bissaro, M., Deganutti, G., Sturlese, M. & Moro, S. New Insights into Key Determinants for Adenosine 1 Receptor Antagonists Selectivity Using Supervised Molecular Dynamics Simulations. Biomolecules 10, 732 (2020).

Since the selectivity profile of Adenonise receptor antagonsits has always been a difficult problem, in this work we studied if Supervised Molecular Dynamics simulations can help to solve this problem, taking advantage of the recently released X-Ray crystal structure of A1 Adenosine Receptor.

2.5 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, (2021).

In this work we extended the applicability of Supervised Molecular Dynamics to the investigation of the molecular recognition process of Macrocyclic peptides, using Thrombin as a case study.

2.6 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* cmdc.202100156 (2021) doi:10.1002/cmdc.202100156.

Since Fragment posing as always been a difficult task for classic Molecular Docking calculations, in this work we used Supervised Molecular Dynamics as a tool to perform dynamic posing of fragments molecules, using the several X-Ray crystal structures of SARS-CoV-2 M Pro available.

2.7 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).

2.8 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 this two works we applied Supervised Molecular Dynamics to investigate the molecular recognition process of clinical candidate SARS-CoV-2 M Pro inhibitors.

3 Methodological works

3.1 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, 3558 (2019).

To assess the reliability of common used Docking protocols, here we systematically benchmarked the accuracy prediction of several Docking softwares. We also analysed the data to understand if the performance of Docking can vary among different protein families.

3.2 Jiménez-Luna, J., Cuzzolin, A., Bolcato, G., Sturlese, M. & Moro, S. A Deep-Learning Approach toward Rational Molecular Docking Protocol Selection. Molecules 25, 2487 (2020).

In this methodological work we used a Deep learning based approach to predict the best Docking protocol and scoring function to use for a particular Protein-ligand complex.

Can We Still Trust Docking Results?

An Extension of the Applicability of DockBench on PDBbind Database

Giovanni Bolcato, Alberto Cuzzolin, Maicol Bissaro, Stefano Moro and Mattia Sturlese

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**, 3558 (2019).

Abstract

The number of entries in the Protein Data Bank (PDB) has doubled in the last decade, and it has increased tenfold in the last twenty years. The availability of an ever-growing number of structures is having a huge impact on the Structure-Based Drug Discovery (SBDD), allowing investigation of new targets and giving the possibility to have multiple structures of the same macromolecule in a complex with different ligands. Such a large resource often implies the choice of the most suitable complex for molecular docking calculation, and this task is complicated by the plethora of possible posing and scoring function algorithms available, which may influence the quality of the outcomes. Here, we report a large benchmark performed on the PDBbind database containing more than four thousand entries and seventeen popular docking protocols. We found that, even in protein families wherein docking protocols generally showed acceptable results, certain ligand-protein complexes are poorly reproduced in the self-docking procedure. Such a trend in certain protein families is more pronounced, and this underlines the importance in identification of a suitable protein–ligand conformation coupled to a well-performing docking protocol.

1. Introduction

Since its introduction in the early 1980s [1], molecular docking has served to aid medicinal computational chemists in optimizing the drug discovery process. Ten years later, due to methodological and technological advances, together with the increasing number of experimentally solved macromolecular structures, it became possible to process more and more molecules within a docking procedure, opening the era of Structure-Based Virtual Screening (SBVS) as a strategy in selecting appropriate compounds from large virtual libraries on the basis of good protein–ligand interaction patterns. [2] Thanks to molecular docking, Structure-Based Drug Discovery (SBDD) field has become very popular today. A docking protocol can be described as the combination of a search algorithm that samples the conformational space of a ligand, generating conformations for the

ligand itself (defined as poses) within a binding site, and a mathematical equation, called scoring function, which quantitatively evaluates the quality of such poses. The scoring function has always been the Achilles tendon of molecular docking due to the inaccuracy in quantified strength of the complex network of molecular interactions. Today, it is widely accepted that molecular docking has been outperformed by other structure-based in silico methodologies in investigating the stability and strength of the protein–ligand interaction, even though they are usually demanding techniques.[3] However, molecular docking still represents a valid technique in sampling the conformations of the ligand in a binding site in a very efficient manner—at a fraction of the computational cost of more accurate methods based for example on Molecular Dynamics. [4] To prove the extensive adoption of molecular docking in research, there are more than 50 docking software options listed up to date in the on-line Click2Drug repository. [5] It should also be considered that each docking software usually provides more than one scoring function in which performance ought to be evaluated in the protocol tuning step. This means that computational chemists have at their disposal a plethora of different protocols when they face a docking calculation and, more importantly, the success, for example, of a Virtual Screening (VS) campaign, strongly relies on the accuracy of the protocol employed to place and rank the conformation of candidates into a target binding site.[6] To further complicate matters, additional considerations need to be taken into account. In fact, more and more experimental structures are thankfully available, hence the range of possible combinations in protein conformation-docking protocol is growing in an unstoppable trend. It is, therefore, clear that a crucial step in SBVS is the selection of a proper docking protocol and an appropriate protein conformation.[7,8] To address this issue, we recently proposed a platform, DockBench, with the aim of simplifying the non-trivial task of automatically comparing the performance of different docking protocols in a self-docking exercise. The criteria of selection of the most appropriate protocol are based on geometrical and statistical basis evaluating few observables: the lowest and the average Root Main Square Deviation (RMSD) obtained for a pose of the ligand compared to its crystallographic pose and the protocol score. [9] In 2011, Plewczynski et al. reported a comparison among seven docking protocols on the PDBbind (http://www.pdbbind.org.cn) that, at that time, counted on 1300 structures. [8] Here, we report a large benchmark of 17 different docking protocols compared on the basis of the self-docking procedure on a dataset of 4169 protein–ligand complexes. The notable number of structures has offered the opportunity to evaluate the performance of molecular docking from different points of

view, underlining how the efficiency of docking protocols may vary depending on the nature of the protein family.

2. Results

The benchmark was performed on 4169 structures obtained from PDBbind, a free database of binding affinity data for biomolecular complexes including protein–ligand, nucleic acid-ligand, protein-nucleic acid, and protein–protein complexes.[10] The PDBbind "Refined set" is a subset of high-quality protein–ligand complex structures helpful for the validation of Docking protocols. All the structure needs to be processed prior to the docking calculation to keep only the protein and the ligand alone. This was necessary to simplify the execution on such a large set of complexes and protocols.The preparation of the data was accomplished by an automatic procedure based on the Molecular Operating Environment (MOE) suite for proteins and OpenEye toolkit for ligands (vide infra, see method section for details).[11,12] The benchmark execution was performed on all 17 protocols implemented in DockBench 1.0.6 based on seven different docking software options, each of which was coupled to different scoring functions whenever possible. The complete list of the protocols is reported in Table 1. The benchmark consisted of the execution of 70,873 single docking runs (4169 complexes; 17 protocols) distributed on a single server. The wall time necessary to perform all docking runs was approximately 72 h.

Program	Search Algorithm/ Placing Method	Scoring Function	Protocol Abbreviation
Autodock 4.2	Local Search	AutoDock SF	AUTODOCK-ls
	Lamarckian GA	AutoDock SF	AUTODOCK-lga
	Genetic Algorithm	AutoDock SF	AUTODOCK-ga
Vina 1.1.2	Monte Carlo + BFGS local search	Standard Vina SF	VINA-std
Glide 6.5	Glide Algorithm	Standard Precision	GLIDE-sp
GOLD 5.4.1	Genetic Algorithm	Goldscore	GOLD-goldscore
	Genetic Algorithm	Chemscore	GOLD-chemscore
	Genetic Algorithm	ASP	GOLD-asp
	Genetic Algorithm	PLP	GOLD-plp
MOE 2019.01	Triangle Matcher	London-dG	MOE-londondg
	Triangle Matcher	Affinity-dG	MOE-affinitydg
	Triangle Matcher	GBIVIWSA	MOE-gbiviwsa
PLANTS 1.2	ACO Algorithm	PLP	PLANTS-plp

	ACO Algorithm	PLP95	PLANTS-plp95
	ACO Algorithm	ChemPLP	PLANTS-chemplp
rDock 2013.1	Genetic Algorithm + Monte Carlo + Simplex minimization	Standard rDock master SF	RDOCK-std
	Genetic Algorithm + Monte Carlo + Simplex minimization	Standard rDock master SF + desolvation potential	RDOCK-solv

GA (Genetic Algorithm) BFGS (Broyden-Fletcher-Goldfarb-Shanno), ASP (Astex Statistical Potential), PLP (pairwise linear potential), ACO (Ant Colony Optimization)

Table 1 *List of docking protocols used in the benchmark.*

The automated analysis was based on the calculation of three scores: (i) RMSD minimum (RMSDmin), (ii) the RMSD average (RMSDave), (ii) the number of structure with RMSD lower than the (N(RSMD < R)), and a fourth score named Protocol Score Pscore that summarized the overall performance for a geometric point of view. The Pscore instead is defined as follows: One point is assigned to the protocols that have an RMSDave lower than the value of the crystallographic resolution, another point is assigned to the protocols producing at least 10 poses (50% of generated conformation) with an RMSD (compared to the crystallographic geometry) lower than the crystallographic resolution, and two points are assigned to protocols which fulfill both the previous conditions. The complete matrix of the results is available in supporting information. The observed RMSDmin values were in the range of 0.05 and 38.49 Å. High RMSDmin values are symptomatic for ligands placed far away from the native binding site. A possible explanation could be ascribed in having defined the pocket using a sphere with radius 15 Å. The radius was deliberately set large to give the possibility to be sufficiently broad for all the ligands in the dataset and may be problematic for docking of small ligands or in the case of multiple pockets closely located.

An interesting question we were considered was about the performance of docking protocols in different target families since, in PDBbind, many protein families are represented by several entries. The results were grouped on the basis of the protein in families (PF) using the Pfam (Protein Family) database families as definition.[13] For each complex, the PF Pfam code was retrieved for the protein chain and hence grouped. For many multi-domain proteins, a different Pfam code can be assigned depending on the domain solved in the structure; for instance, the proteins belonging to the family PF00069 (Protein Kinase) often contain domains labeled as PF02827 (Cyclic adenosine monophospate-dependent protein kinase inhibitor), PF00134 (Cyclin, N-terminal domain), PF02984 (Cyclin, C-terminal domain), and a few others. Some proteins cannot be classified in a single group, and therefore we merged those groups for analysis (for example, PF00183 and PF02518, Heat Shock

Protein 90, HSP90 and GHKL domain). To address this issue, we compared the docking performance by the Protocol Score (Pscore) for the major cluster to investigate whether the docking performances of the different protocols vary among the different protein families. Unexpectedly, the performance among different families showed a remarkable fluctuation (Table 2), with certain families having many protocols with Pscore > 1 on most of the complexes. It is interesting to note that, between the best performing group (PF00104) and the worst (PF00026) one, the percentage of protocols with Pscore > 1 showed a difference of an order of magnitude, 41.66%, and 4.37%, respectively. Among the best-performing ones, the families with good Pscore were: PF00104 (Hormone receptors), PF00497 (Bacterial extracellular solute-binding proteins, family 3), PF10613 (Ligated ion channel l-glutamate and glycine binding site), and PF01048 (Phosphorylase superfamily). All these families showed a Pscore > 1 in more than 29% of the docking runs.

Pfam Family	Protein Description	Size	Protocol Score P_{score}%				
			0	1	2	3	>1
PF00104	Ligand-binding domain of nuclear hormone receptor	85	59.34	10.24	26.57	4.84	41.66
PF00497	Bacterial extracellular solute-binding proteins, family 3	38	59.29	9.44	25.70	5.57	40.71
PF10613	Ligated ion channel L-glutamate- and glycine-binding site	83	67.97	6.80	20.55	4.68	32.03
PF01048	Phosphorylase superfamily	47	70.09	8.01	16.77	5.13	29.91
PF00102	Protein-tyrosine phosphatase	52	79.30	5.20	11.99	3.51	20.70
PF00069	Protein kinase domain	207	80.68	5.43	10.46	3.43	19.32
PF00061	Lipocalin/cytosolic fatty-acid binding protein family	49	82.11	4.08	10.80	3.00	17.88
PF02518 PF00183	Hsp90 protein and GHKL domain	89	82.74	5.35	8.26	3.64	17.25
PF07714	Protein tyrosine kinase	133	83.90	5.79	6.77	3.54	16.10
PF00089 PF14670 PF09396	Trypsin	330	85.54	4.65	6.84	2.96	14.45

PF00233	3’5’-cyclic nucleotide phosphodiesterase	37	87.92	3.82	5.41	2.86	12,08
PF00439	Bromodomain	112	90.02	2.89	4.67	2.42	9.98
PF00026	Eukaryotic aspartyl protease	73	90.49	3.14	4.11	2.26	9.51
PF00413	Matrixin	49	90.88	3.24	4.20	1.68	9.12
PF00077	Retroviral aspartyl protease	301	95.41	2.27	1.64	0.68	4.59
PF00194	Eukaryotic-type carbonic anhydrase	273	95.63	2.28	1.17	0.91	4.37

Pfam (Protein Family), Hsp90 (Heat shock protein 90).

Table 2 *Summary of benchmark results by Pfam families. Protocol scores are reported as percentage with respect to the total docking runs (Pscore%).*

On the other hand, we found that certain families had very poor results, with Pscore > 1 found below 10%; this is the case for PF00194 (Eukaryotic-type carbonic anhydrase), PF00077 (Retroviral aspartyl proteases), PF00413 (Matrixin), and PF00026 (Eukaryotic aspartyl protease). The trend observed for Pscore is also evident in RMSDave. The results for the most populated families are reported in Figure 1. The Pscores were reported as a heatmap to easily summarize the comparison of such a big matrix (higher scores highlight better protocol-complex couple). Numerical results are reported in the supplementary information. The results for the same families in terms of RMSDave are reported in Figure 2.

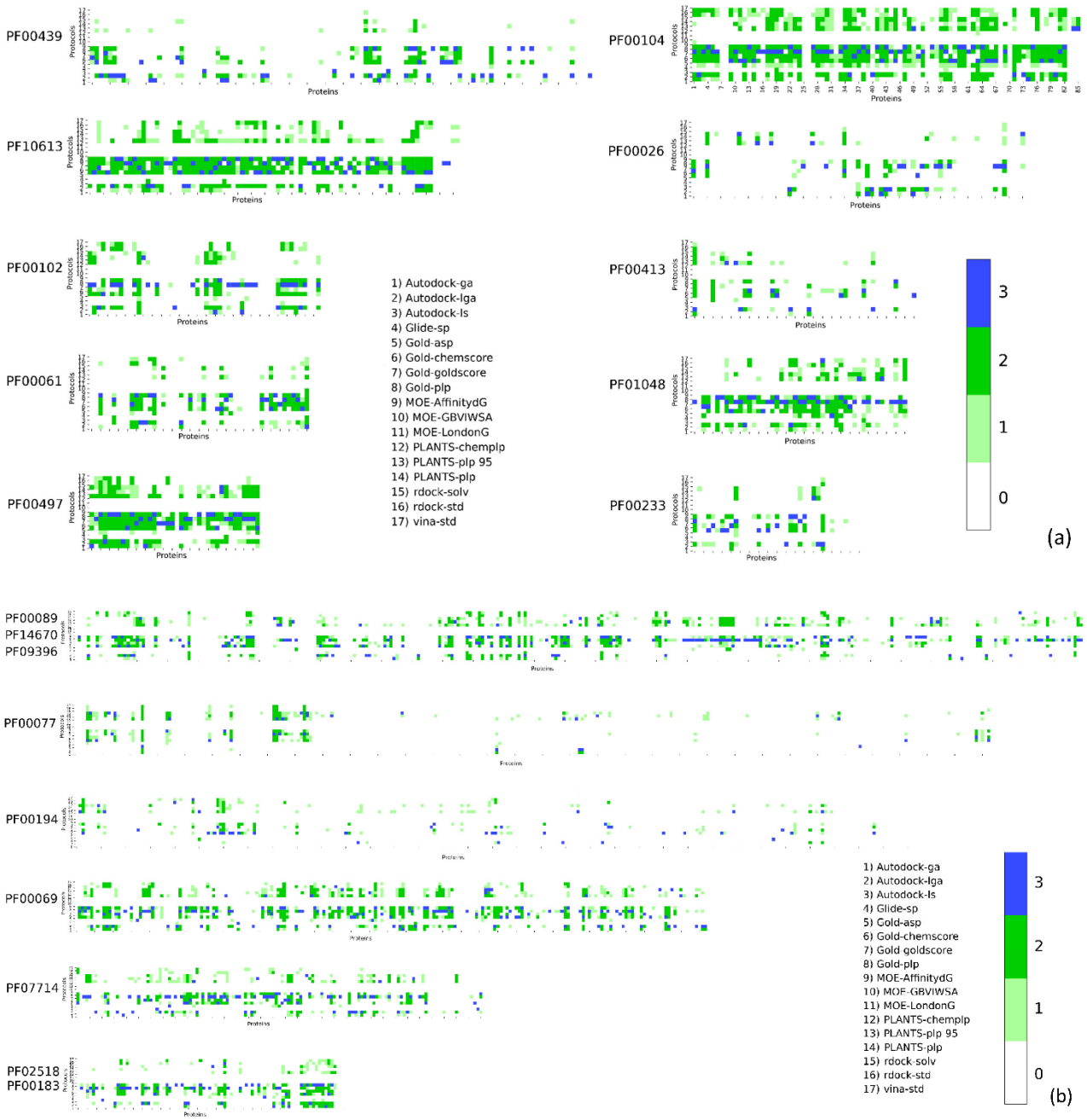


Figure 1 *DockBench Results divided by Pfam protein families. The heatmaps are color-coded according to the Pscore. The ten families in panel (a) are: PF00439, Bromodomain; PF10613, Ligated ion channel l-glutamate and glycine-binding site; PF00102, Protein tyrosine phosphatases; PF000061 Lipocalin; PF00497, Bacterial extracellular solute-binding proteins family 3; PF00104, Hormone receptors; PF00026, Eukaryotic aspartyl protease Peptidase M_10; PF01048, Phosphorylase superfamily; PF00233, 3'5'-cyclic nucleotide phosphodiesterases. The six families in panel (b) are: PF00089 Trypsin, PF14670 Coagulation Factor Xa inhibitory site, PF09396 Thrombin light chain, PF00077 Retroviral aspartyl proteases, PF00194 carbonic anhydrases, PF00069 protein kinase, PF07714 tyrosine kinase, PF02518 GHKL domain, and PF00183 HSP90.*

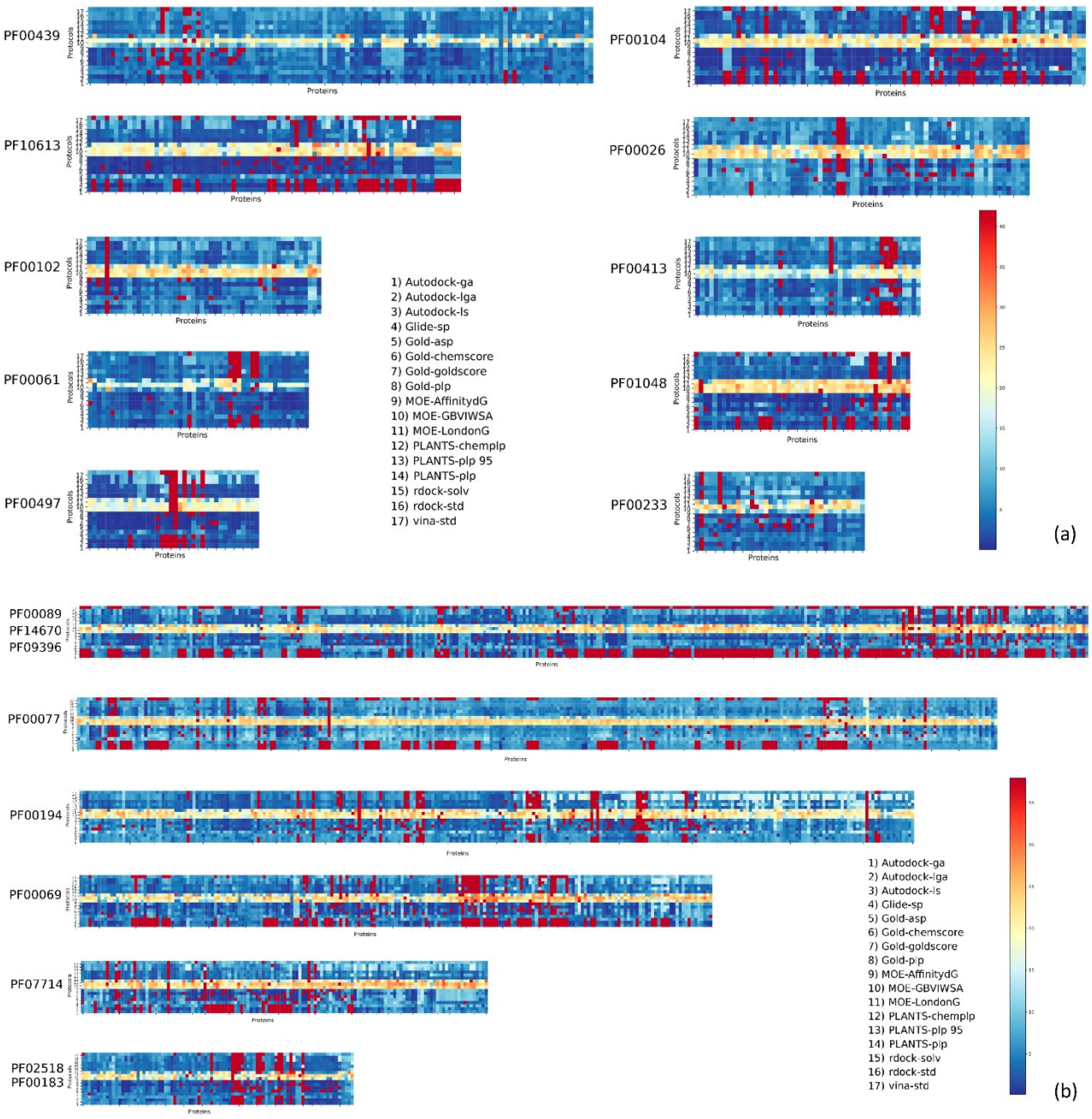


Figure 2 *DockBench Results divided by Pfam protein families. The heatmaps are color-coded according to the RMSDave. The ten families in panel (a) are: PF00439, Bromodomain; PF10613, Ligated ion channel l-glutamate and glycine-binding site; PF00102, Protein tyrosine phosphatases; PF000061 Lipocalin; PF00497, Bacterial extracellular solute-binding proteins family 3; PF00104, Hormone receptors; PF00026, Eukaryotic aspartyl protease Peptidase M_10; PF01048, Phosphorylase superfamily; and PF00233, 3'5'-cyclic nucleotide phosphodiesterases. In panel (b) the heatmaps are color-coded according to the Root Main Square Deviation (RMSD)ave. The six families are: PF00089 Trypsin, PF14670 Coagulation Factor Xa inhibitory site, PF09396 Thrombin light chain, PF00077 Retroviral aspartyl proteases, PF00194 carbonic anhydrases, PF00069 protein kinase, PF07714 tyrosine kinase, PF02518 GHKL domain, and PF00183 HSP90.*

A further aspect that was considered was the ability of the docking protocol in placing in the first position, according to their scoring function, the pose with the lowest RMSD. This aspect is particularly relevant because it indicates how the protocol is able to distinguish between different binding modes and, hopefully, prioritizing a binding mode close to the experimentally observed. In Figures S1 and S2, the heatmap plots reporting for the docking runs in which the best-scored pose is also the conformation with lowest RMSD. Unfortunately, in several cases, this simultaneous occurrence did not always guarantee the identification od near-native pose. Indeed, we observed for several cases where the lowest RMSD conformation was far from the experimentally solved one with RMSD values reaching values bigger than 10 Å. The RMSD value of the best conformations is reported on the heatmaps in Figures S3 and S4. Therefore, we performed further analysis focusing on investigation of when the best pose also had a low RMSD value but not necessarily the lowest values. We decided to set a threshold of 1.5 Å to define a near-native pose. In this way, we could highlight a protocol able to place a "good" pose as the first solution, even if potentially better conformation could be present among the 20 obtained.

In Figures S5 and S6, the runs that fulfill such concurrence are reported. Again, the performance of docking protocols showed a very different performance depending on the protein family and, interestingly, in agreement with the Pscore trends. The Ligand-binding domain of nuclear hormone receptor (PF00194) showed in 50% of the runs RMSD < 1.5 Å for the first pose. The percentage of success is also remarkable for the Ligated ion channel l-glutamate- and glycine-binding site (PF10613), 49.3%; the Bacterial extracellular solute-binding proteins (PF00497), 47.6%; and Phosphorylase superfamily (PF01048), 41.7%. On the contrary, certain families performed poorly in this analysis, in particular, Eukaryotic-type carbonic anhydrase, which showed only a 10.8% (Table S1, on Supporting Material).

The factors that are so dramatically affecting the quality of the docking outputs among different families could be related to many variables. First, we address the possible different chemical natures of the ligands belonging to each protein family. To evaluate the ligand chemical space, several molecular descriptors were calculated, including weight, rotatable bonds, hydrogen bond acceptors, hydrogen bond donors, clogP, total polar surface area, and van der Waals volume. To reduce the number of the dimensions, and therefore make the distribution representable in a three-dimensional plot, a Principal Component Analysis (PCA) was performed. As can be seen in Figure 3, ligands of the different clusters do not seem to occupy a different portion of the chemical space. Hence, we then moved attention to possible players removed during the complex preparation,

considering that the poor performances of docking in the cluster PF00077 (Retroviral aspartyl proteases) and PF00439 (Bromodomain) could be eventually ascribed to the removal of the crystallographic waters. It was already reported that the binding mode for several ligands is mediated by a series of water molecules for bromodomains.[14]

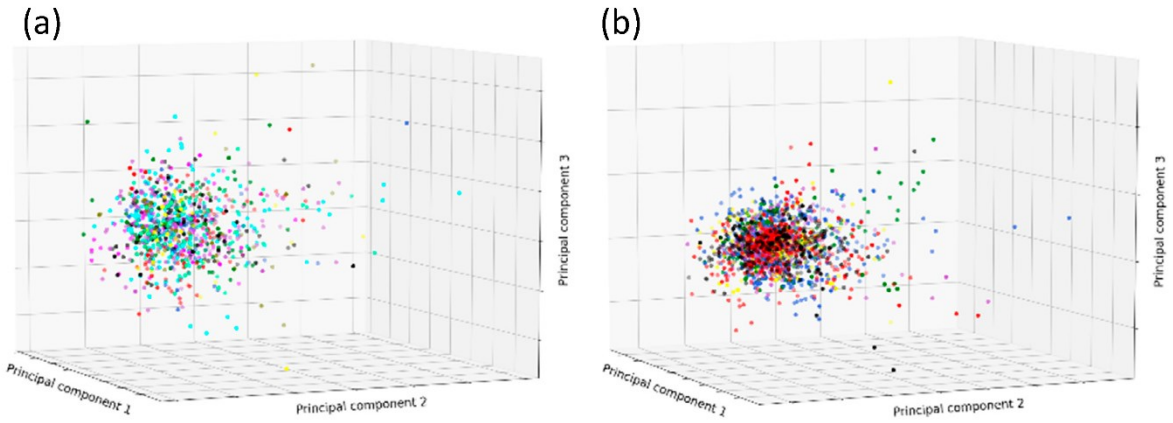


Figure 3 *Principal Component Analysis (PCA) analysis seven molecular descriptors for the groups of ligands on the base of protein families in Table 2. The PCA analysis of ligands from the protein families were split into two groups according to the same division on Figure 1b (a) and Figure 2b (b). The descriptors used in the analysis are weight, rotatable bonds, hydrogen bond acceptor, hydrogen bond donor, clogP, total polar surface area, and van der Waals volume.*

Similarly, in the performances observed for cluster PF00194 (Carbonic Anhydrases), a crucial aspect could be represented by the removal of the zinc ion from the binding sites. For this reason, we performed a further benchmark focused on this family, including the Zinc ion, employing the most promising protocols in the first benchmark, Plants- and Gold-based protocols. The comparison of the heatmaps of the Pscore reported in Figure 4 demonstrates that, despite the introduction of the Zinc ion, the trend of the Pscore improves only moderately. Surprisingly, the distribution of the high Pscore is different in the two benchmarks, suggesting that the Zinc ion introduction only improves for certain complex structures while getting worse for others

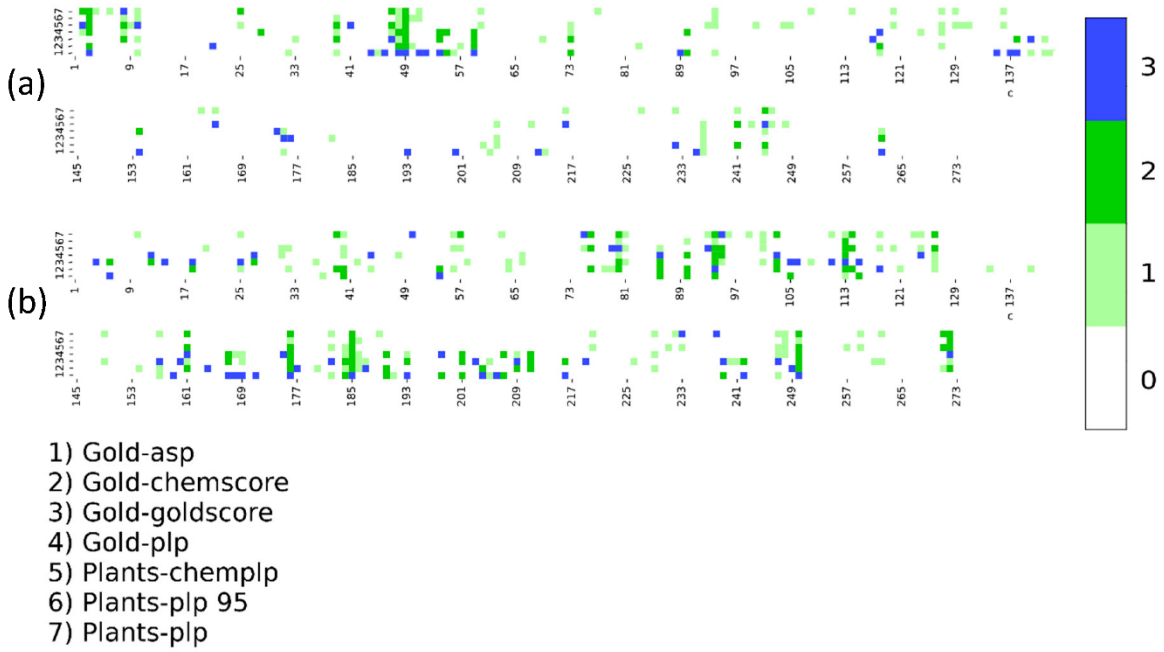


Figure 4 *Comparison between DockBench Results in terms of Protocol Score for cluster PF00194 (carbonic anhydrases) with (b) and without (a) the Zinc ion.*

3. Discussion

A computational chemist has to ask himself many of the right questions when facing molecular docking studies, and the answers are not univocal. Of course, the choice of the best performing protocol and, when multiple structures are available, of the target conformation is the most significant decision. However, the employment of molecular docking may have a different purpose, and a proficient protocol choice must consider such different use. If molecular docking is addressed in binding mode studies, the protocol performances should have the priority. At the same time, the choice of the protein target should depend on the similarity between the compounds to be studied and the ligand co-crystalized. When molecular docking is used in a VS campaign, more variables affect the selection, like the execution speed. The results obtained in this benchmark were obtained with parameter as close as possible to the default values resulting in very variable execution times. For instance, as already reported in previous Dockbench studies, certain protocols may require an order of magnitude of longer time in comparison to faster protocols. It is evident in the case of large libraries that this may represent a critical issue, hence protocols with similar outcomes in self-docking procedure where the choice can be influenced by the execution speed. In our benchmark, we observed, for example, in certain families of proteins, several protocols showing good performance, hence protocol selection may depend on the other factor. It is interesting to note that in the protein families in which molecular docking shows a good trend in reproducing the

experimental conformation, certain protein–ligand complexes are far from being predicted correctly, suggesting the importance of excluding them for docking simulations. Differently, other protein families are challenging targets in which the choice of the posing-scoring algorithm seems to be crucial, as well as the identification of the most suitable complex structure. The performance of such a challenging target should also point out the necessity to investigate the issues that are affecting the docking calculation, for instance, in considering the role of stable water molecules in the binding site or the role of a cofactor, flexible regions of the pocket, or other drawbacks of the system. This study may help the user approach a new target by molecular docking in identifying promising protocols and excluding problematic complex structures. In our opinion, the assessment of the suitable procedure should become a good practice also in light of the increasing number of entries available in the PDB and the advent of novel techniques like Cryo-EM and Solid-State NMR are wading the landscape of an experimentally solved target.

4. Materials and Methods

4.1. Database Preparation

The Refined-set of the PDBbind database was obtained from PDBbind web service (http://www.pdbbind.org.cn/).[10] This dataset is composed of 4463 protein–ligand complexes, and 4169 of them were used for this work. We excluded 294 structures containing peptide–protein complexes that are not particularly suitable for DockBench protocol since it used docking settings which were as close as possible to the default parameters provided by the developers of each software and mostly calibrated on small organic molecules typical of drug discovery. These 4169 complexes were prepared as described below.

The protein structures have been prepared using a Scientific Vector Language (SVL) script using the functions contained in MOE suite reproducing the protein preparation tool of MOE to fix crystal structures issues, such as prediction of coordinates of missing atoms of partially solved residues.[11] Co-crystallized solvent molecules and impurities (such as co-solvents) were removed, and only protein and ligand coordinates were retained. For all ligands, the most favorable ionic state was calculated with OpenEye tools fixpKa.[12] The partial charges were assigned with molcharge, also part of OpenEye toolkit.[12] Ligand geometries were minimized in the first step of DockBench with Openbabel routing using the MMFF94 force field.[15]

4.2. Benchmark: Software and Hardware

The benchmark was performed with DockBench 1.06 software, running on a single HP ProLiant server DL585G7, equipped with four AMD Opteron Processor 6282 servers, for a total of 64 CPU cores.[16,17] Docking protocol was executed according to the original implementation already reported.[16] All the 17 protocols from seven different software options (AutoDock 4.2.5.1 [18], Vina 1.1.2 [19], PLANTS 1.2 [20], rDOCK [21], Glide 6.5 [22], Gold 5.4.1 [23,24], and MOE 2019.01 [11]) were included in the benchmark and run on all 4169 protein–ligand complexes. Briefly, 20 poses were generated every single run. The binding site was defined using a sphere having a radius of 15 Å centered on the center of mass of the co-crystalized ligand present in the complex. An RMSD threshold set to a value of 1 Å value to define unique poses.

The analysis was performed with DockBench analyzer coupled to external Python and Bash script to manage the notable amount of data and to produce the plots. [25,26] The Pfam Protein family was retrieved for each protein using the RCSD PDB REST API service, while the Pfam Clan was obtained from Pfam REST API service. [13,27] Molecular descriptors were calculated using MOE suite.[11]

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. Horvath, D. A virtual screening approach applied to the search for trypanothione reductase inhibitors. *J. Med. Chem.* **40**, 2412–2423 (1997).

3. 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).

4. Moro, S., Sturlese, M., Ciancetta, A. & Floris, M. In silico 3D modeling of binding activities. in *Methods in Molecular Biology* **1425**, 23–35 (Humana Press Inc., 2016).

5. Directory of in silico Drug Design tools. Available at: http://www.click2drug.org/index.php#Docking. (Accessed: 27th September 2020)

6. Houston, D. R. & Walkinshaw, M. D. Consensus docking: Improving the reliability of docking in a virtual screening context. *J. Chem. Inf. Model.* **53**, 384–390 (2013).

7. Salmaso, V., Sturlese, M., Cuzzolin, A. & Moro, S. DockBench as docking selector tool: the lesson learned from D3R Grand Challenge 2015. *J. Comput. Aided. Mol. Des.* **30**, 773–789 (2016).

8. Plewczynski, D., Łaźniewski, M., Augustyniak, R. & Ginalski, K. Can we trust docking results? Evaluation of seven commonly used programs on PDBbind database. *Journal of Computational Chemistry* **32**, 742–755 (2011).

9. Ciancetta, A., Cuzzolin, A. & Moro, S. Alternative quality assessment strategy to compare performances of GPCR-ligand docking protocols: The human adenosine A2A receptor as a case study. *J. Chem. Inf. Model.* **54**, 2243–2254 (2014).

10. Wang, R., Fang, X., Lu, Y., Yang, C. Y. & Wang, S. The PDBbind database: Methodologies and updates. *J. Med. Chem.* **48**, 4111–4119 (2005).

11. ULC, C. C. G. Molecular Operating Environment (MOE). (2013).

12. OpenEye Scientific Software Inc. OEChem; OpenEye Scientific Software Inc.: Santa Fe, NM, USA, 2016.

13. El-Gebali, S. *et al.* The Pfam protein families database in 2019. *Nucleic Acids Res.* **47**, D427–D432 (2019).

14. Shadrick, W. R. *et al.* Exploiting a water network to achieve enthalpy-driven, bromodomain-selective BET inhibitors. *Bioorganic Med. Chem.* **26**, 25–36 (2018).

15. Halgren, T. A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem.* **17**, 490–519 (1996).

16. 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).

17. 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).

18. Morris, G. M. *et al.* Software news and updates AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J. Comput. Chem.* **30**, 2785–2791 (2009).

19. 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**, NA-NA (2009).

20. 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)* **4150 LNCS**, 247–258 (Springer Verlag, 2006).

21. Ruiz-Carmona, S. *et al.* rDock: A Fast, Versatile and Open Source Program for Docking Ligands to Proteins and Nucleic Acids. *PLoS Comput. Biol.* **10**, e1003571 (2014).

22. Halgren, T. A. *et al.* Glide: A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening. *J. Med. Chem.* **47**, 1750–1759 (2004).

23. 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.* **52**, 609–623 (2003).

24. Cambridge Crystallographic Data Centre. GOLD Suite, version 5.2; Cambridge Crystallographic Data Centre: Cambridge, UK, 2013..

25. Hunter, J. D. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).

26. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* (2011).

27. Berman, H. M. *et al.* The protein data bank. *Nucleic Acids Res.* **28**, 235–242 (2000).

Revisiting the Allosteric Regulation of Sodium Cation on the Binding of Adenosine at the Human A_{2A} Adenosine Receptor: Insights from Supervised Molecular Dynamics (SuMD) Simulations

Maicol Bissaro, **Giovanni Bolcato**, Giuseppe Deganutti, Mattia Sturlese, Stefano Moro

Bissaro, M., Bolcato, G., Deganutti, G., Sturlese, M., & Moro, S. (2019). Revisiting the Allosteric Regulation of Sodium Cation on the Binding of Adenosine at the Human A2A Adenosine Receptor: Insights from Supervised Molecular Dynamics (SuMD) Simulations. *Molecules*, *24*(15), 2752.

Abstract

One of the most intriguing findings highlighted from G protein-coupled receptors (GPCRs) crystallography is the presence, in many members of the class A, of a partially hydrated sodium ion in the middle of the seven transmembrane helices (7TM) bundle. In particular, the human adenosine A_{2A} receptor (A_{2A} AR) is the first GPCR in which a monovalent sodium ion was crystallized in a distal site from the canonical orthosteric one, corroborating, from a structural point of view, its role as a negative allosteric modulator. However, the molecular mechanism by which the sodium ion influences the recognition of the A_{2A} AR agonists is not yet fully understood. In this study, the supervised molecular dynamics (SuMD) technique was exploited to analyse the sodium ion recognition mechanism and how its presence influences the binding of the endogenous agonist adenosine. Due to a higher degree of flexibility of the receptor extracellular (EC) vestibule, we propose the sodium-bound A_{2A} AR as less efficient in stabilizing the adenosine during the different steps of binding.

1. Introduction

The human genome encodes more than 800 different G protein-coupled receptors (GPCRs), membrane proteins characterized by a distinctive seven transmembrane helices (7TM) architecture. This superfamily of receptors recognizes an enormous variety of extracellular signals (i.e. ions, neurotransmitters, peptides) and transmits the chemical information into the intracellular compartment, modulating many cellular activities.[1,2] This is achieved through the recruitment of different molecular effectors, such as G proteins, protein kinases, or β-arrestins. Given their crucial role at the cellular level, GPCRs represent an important family of therapeutic targets, and it is not surprising that more than 30% of the approved drugs act on at least one GPCR [3].

Adenosine receptors (ARs) are a family of class A GPCRs comprising four different subtypes, respectively, A_1, A_{2A}, A_{2B}, and A_3, all involved in purinergic signaling.[2] ARs recognize the extracellular nucleoside adenosine as the endogenous agonist, which, depending on the receptor subtype and tissue localization, affects and modulates different pathophysiological cellular conditions in a pleiotropic way. For example, purinergic signalling is involved in inflammation, cancer, neurodegeneration, and cardiovascular diseases.[4] The human A_{2A} AR subtype has been studied in depth both from a pharmacological and structural point of view. To date, 46 structures deposited in the Protein Data Bank (PDB) show the adenosine A_{2A} receptor (A_{2A} AR) in complex with both agonists (active and intermediate active states) and antagonists (inactive states).[5]

Interestingly, the A_{2A} AR was the first GPCR co-crystallised with a monovalent sodium ion, explaining from a structural point of view its negative allosteric effect. [6] In 1973, Pert and co-workers discovered how physiological concentrations of specific ions could decrease the opioid receptor affinity for agonists, without influencing the antagonists binding profile.[7,8] After this first body of evidence, the effect of the sodium ion (Na^+) was particularly investigated, leading to the discovery of at least 15 further GPCR subtypes sensible to its allosteric effect. Site-directed mutagenesis studies led to the identification of the conserved amino acid $D^{2.50}$ as a fundamental counterpart for sodium binding, later confirmed by the publication (2012) of the first high-resolution (1.8 Å) X-ray crystal structure of the A_{2A} AR.[6,9] In this structure, the Na^+ was located at the interface between TM2, TM3, and TM7, coordinated to five oxygen atoms belonging to the side chain of the conserved residues $D^{2.50}$, $S^{3.39}$ (the Ballesteros-Weinstein GPCRSs numbering is reported as superscript) and to an ordinate cluster of three water molecules. The negatively charged aspartic acid is conserved in over 90% of the class A GPCRs, thus suggesting an evolutive role in binding the monovalent ion.[10–12] As reported in Table 1, 34 GPCRs have been co-crystallized with a sodium ion, spanning members from three of the four branches in which the class A GPCRs are classified.

	Best Resolution (Å)	Number of Structures	Class A Branch
A_{2A} adenosine receptor	1.7	24	α
Protease-activated receptor 1	2.2	1	δ
Protease-activated receptor 2	2.8	2	δ
β_1 adrenergic receptor	2.1	3	α
D_4 dopamine receptor	2.1	1	α
Complement component 5a receptor 1	2.2	1	γ
δ opioid receptor	1.8	2	γ

Table 1 *Crystallographic structures of class A G protein-coupled receptors (GPCRs) deposited on the Protein Data Bank (PDB) and containing a sodium ion in the transmembrane helices (TM) region.*

A large body of structural evidence indicates that the sodium ion is detectable exclusively in the presence of antagonists, as all the GPCRs solved in the active state do not coordinate the cation. It follows that a receptor can exist in at least two conformational states, one able to bind the sodium ion and antagonists, the other with high affinity only for agonists. From a functional point of view, it has been proposed that the sodium stabilizes a specific conformation of the receptor and shifts the conformational equilibrium towards the inactive state.[13] In light of this, computational studies turned their attention to the influence of sodium ion coordination in the A_{2A} AR affinity for antagonists, focusing less on the structural basis of the sodium-bound receptor's inability to recognize agonists.[14] The sodium binding mechanism to 18 different GPCRs has been recently investigated through microsecond-scale molecular dynamics (MD) simulations.[15] Previous computational studies compared the allosteric binding site of the sodium ion in the A_{2A} AR inactive and active states, suggesting the latter conformation is characterized by an important reduction of the volume of the allosteric cavity, unfavourable to the ion coordination.[9,16] Although it is now widely accepted that the recognition of the sodium ion at its allosteric binding site occurs from the extracellular side, it is more complex to computationally describe how the sodium may dissociate and how the agonist can play a role in this process.[15] Recent scientific work has shown that Na^+ can leave the allosteric site either by translocating in the cytoplasmic side or by retracing the binding path towards the extracellular environment. Moreover, the protomeric state of the titratable residue $D^{2.50}$ seems to be determinant in controlling the Na^+ unbinding mechanism.[17–20] Further studies are therefore necessary to investigate, from a mechanistic point of view, the negative allosteric modulation of the sodium ion and attempt to understand how the stabilization of the inactive state of the receptor results, from a macroscopic point of view, in a decreased ability of the receptor to recognize an agonist.

In our laboratory we have implemented a computational method, named supervised molecular dynamics (SuMD), that enables the exploration of ligand-receptor recognition pathways in the nanosecond timescale.[21–23] The performance speedup is due to the combination of a tabu-like supervision algorithm on the ligand-receptor distance with classic MD simulation. SuMD enables the investigation of binding events independently from the ligand starting position, its chemical structure (small molecules or peptides), and the thermodynamic affinity.[21–23] In this work, we simulated and analysed the recognition between the sodium ion and the A AR, both in the inactive and intermediate-active conformations. SuMD simulations shed light on the molecular basis underneath the allosteric effect of the sodium ion from a site distinct from the orthosteric one,

allowing for a better understanding of how its presence perturbs the binding mechanism of the endogenous agonist adenosine.

2. Results

2.1. SuMD simulations of the sodium ion on the A_{2A} AR

As anticipated, SuMD simulations allow for the simulation of intermolecular recognition pathways in a very compressed time scale. However, this limits exploration to a limited subset of the complex GPCR conformational landscape during a single SuMD simulation. Considering also the lack of reliable structural information on the unbound (apo) state of the receptor, the experimentally-determined inactive (co-crystallised with the inverse agonist ZM241385) and intermediate active (co-crystallised with the adenosine) conformations of A_{2A} AR were retrieved from the PDB database (PDB codes: 4EIY and 2YDO, respectively) and prepared for the SuMD simulations, as described in the Materials and Methods section. In order to ensure the robustness of the results, five SuMD replicates for each state of the receptor were performed to simulate the recognition of the sodium ion. As far as we know, this is an expansion of the applicability domain of this MD method; previously it was only to small molecules and peptides. As reported in Table 2, a few nanoseconds were sufficient to sample a complete Na+ binding pathway during each repetition, instead of several microseconds as required by classical MD experiments.[15]

	A_{2A} AR inactive conformation			**A_{2A} AR intermediate active conformation**		
	SuMD time (ns)	*Reached the allosteric site*	*$RMSD_{min}$ (Å)*	*SuMD time (ns)*	*Reached the allosteric site*	*$RMSD_{min}$ (Å)*
Replicate 1	10.8	No	10.03	15.4	Yes	0.2
Replicate 2	23.6	Yes	0.17	12.0	Yes	0.1
Replicate 3	20.6	Yes	0.04	2.4	No	24.2
Replicate 4	20.8	Yes	0.18	15.6	Yes	0.1
Replicate 5	18.2	Yes	0.40	4.6	No	17.1

Table 2 *Supervised molecular dynamics (SuMD) simulations of the sodium ion performed on the inactive (left side) and intermediate active (right side) conformations of the adenosine A_{2A} receptor (A_{2A} AR). For each replica, the SuMD simulation time, the positive or negative outcome and the minimum RMSD ($RMSD_{min}$) reached by the sodium have been reported (the crystallographic structure 4EIY was used as a reference).*

On the inactive A_{2A} AR conformation, the cation reached the allosteric site (identified by the triad of residues $D^{2.50}$, $S^{3.39}$, and $N^{7.45}$) in four out of five SuMD replicates (low $RMSD_{min}$ values in Table 2), reproducing the experimental coordination with three water molecules (Figure 1, Video 1). Surprisingly, the sodium ion also reached the allosteric binding site during three out of the five SuMD

replicates of the receptor intermediate-active conformation, which has been suggested as the low-affinity state for the cation. In line with the results from a previous study, the active conformation of A_{2A} AR was able to bind the sodium only after a rearrangement of the TM domain (TMD), characterized by the increase of the distance between the TM2 and TM3, as well as the outward movement of the TM7 (Figure 1).[16] Of note, these are hallmarks of the inactivation process of GPCRs.[24]

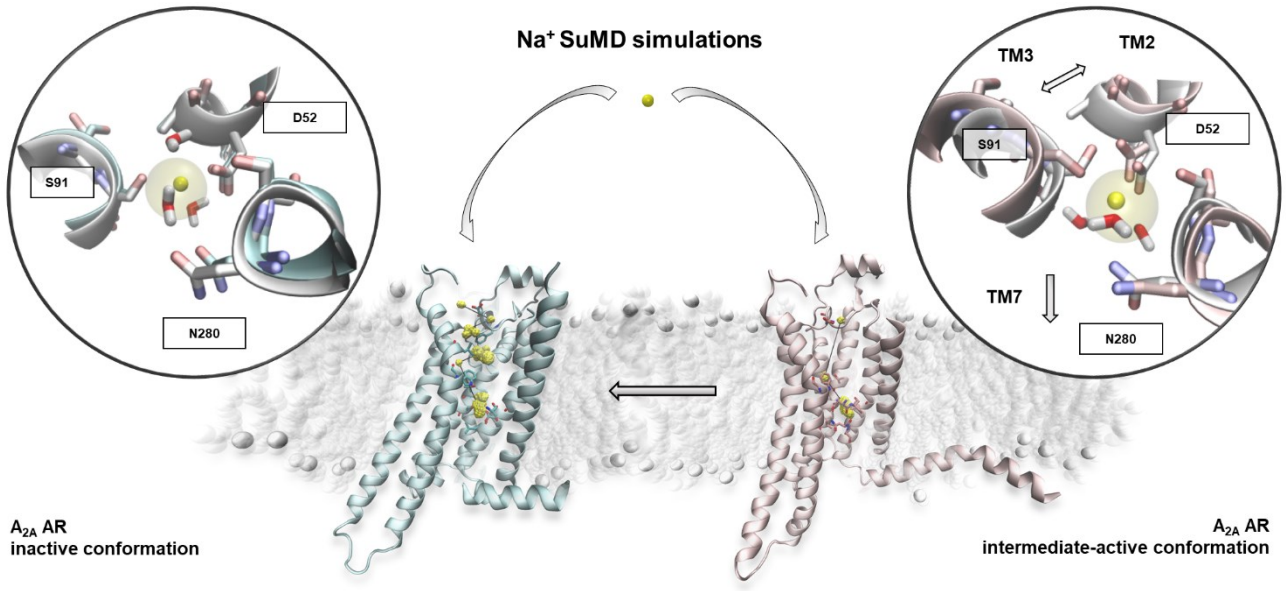


Figure 1 *The recognition pathway of Na^+ on the two relevant A_{2A} AR conformations. TM = transmembrane helices.*

On the left side of the panel in Figure 1, the inactive state of the receptor is reported, along with the sodium positions mainly occupied during the SuMD replicates (yellow dots). The ten most engaged residues are shown as a stick. Within the round box, a magnification shows the sodium allosteric site from a SuMD representative frame (cyan ribbon) and the crystallographic reference 4EIY (white ribbon). The cation reached the experimentally solved position (transparent van der Waals volume). On the right side, the intermediate active conformation of the A_{2A} AR is reported alongside the ion positions during binding (yellow dots). A SuMD final state (pink ribbon) and the crystallographic reference 2YDO (white ribbon) are compared in the magnification. The corresponding sodium location in the inactive structure 4EIY is showed as a transparent van der Waals volume. The receptor's structural changes upon sodium binding (indicated with arrows) can be summarised with an increase of the inter-helical distances in order to accommodate the cation.

To better analyse the sodium ion recognition against the two A_{2A} AR conformations, the SuMD trajectories were subjected to a clustering analysis using the DBSCAN algorithm (for details see the Materials and Methods section), which was able to geometrically map the regions of the receptor

in which the cation was stationed the most during its approach to the allosteric site (Figure 1, Figure S8.)[25] The clusters highlighted a binding mechanism articulated in three temporally consequent phases. During the first step, the sodium ion approached the vestibular region of the A_{2A} AR and interacted with negatively charged residues located at the second extracellular loop (ECL2). A strong electrostatic interaction was formed with $E169^{ECL2}$, before the breaking of the $E169^{ECL2}$–$H264^{ECL3}$ salt bridge.[26] Interestingly, in Replicate 1 (the only unproductive simulation of the active A_{2A} AR) the ion remained trapped in proximity to the ECL2 as strong interactions with $E169^{ECL2}$ were retained for the entire simulation. In the successive binding step, the sodium ion explored the orthosteric site and made interactions with residue $N253^{6.55}$, known to be fundamental for the binding of both agonists and antagonists. The final transition of the sodium to the allosteric site (step three) was controlled by the side chain rotameric state of the "toggle switch" $W246^{6.48}$ residue.[16,27] Although the sodium binding modes obtained from simulations on the two A_{2A} AR conformations were similar (Figure 1), the recognition mechanism of the sodium ion significantly diverged (Figure S8). On the active A_{2A} AR, indeed, the cation did not situate on the orthosteric site, putatively due to a different conformational state of the $W246^{6.48}$ side chain (which has been suggested as being able to modulate the communication between the orthosteric and allosteric sites).[28]

To investigate the reversibility of the sodium ion binding to the inactive A_{2A} AR, an unbiased MD simulation was performed from an SuMD replicate's final state (see the Materials and Methods section). As expected, in about 600 ns, a spontaneous unbinding event from the allosteric site was sampled (Figure S9).

SuMD simulation results suggested that in absence of the orthosteric ligand, the ion could spontaneously coordinate and stabilize the inactive conformation of the receptor (the receptor state also responsible for the antagonists and inverse agonists recognition). On the other hand, Na^+ was able to bind the active state of the receptor only after an adaptation of the allosteric binding site. Only that conformational population not bound to the sodium ion, in equilibrium with the previous one, could, therefore, be recognized by an agonist, ready to trigger the receptor activation process. In this way, we could give a molecular interpretation to the pharmacological meaning of the negative allosteric modulator attributed to the sodium ion.

To investigate the possible effects that these two different Na^+–A_{2A} AR complexes can trigger on the binding mechanism of the endogenous agonist, adenosine, further SuMD replicates were carried out and the results will be described in the next sections.

2.2. SuMD simulations of the adenosine on the intermediate-active, sodium-free, A_{2A}AR conformation

Ten SuMD replicates (Table 3) were performed using the A_{2A} AR coordinates in the intermediate-active conformation (PDB ID 2YDO). We define "productive" as a trajectory that resulted in the adenosine reaching the orthosteric site. The seven productive SuMD simulations were extended for a further 100 ns of unbiased MD simulation to evaluate the stability of the bound states sampled.

	SuMD time (ns)	*Reached orthosteric site*	*Adenosine binding mode*	*X-ray binding mode after 100 ns of MD*	*$RMSD_{min}$ (Å)*
Replicate 1a	7.2	No	No (Meta-binding site on ECL2)	-	14.3
Replicate 2a	31.8	Yes	No (Distorted binding mode)	Yes	0.4
Replicate 3a	40.8	Yes	Yes	Yes	0.4
Replicate 4a	32.4	Yes	No (Ribose Up)	Yes (Ribose *syn* conformation)	2.5
Replicate 5a	29.4	Yes	No (Ribose Up)	Yes (Ribose *syn* conformation)	2.7
Replicate 6a	15.6	No	No (Meta-binding site on ECL2)	-	12.2
Replicate 7a	28.2	Yes	No (Ribose Up)	Yes (Ribose *syn* conformation)	0.4
Replicate 8a	32.4	Yes	Yes (Ribose *syn* conformation)	Yes (Ribose *syn* conformation)	2.7
Replicate 9a	24.0	Yes	No (Distorted binding mode)	Yes (Ribose *syn* conformation)	2.3
Replicate 10a	10.2	No	No (Distorted binding mode)	-	15.3

Table 3 *Summary of the adenosine SuMD simulations performed on the A_{2A} AR intermediate-active conformation. For each replicate, the SuMD simulation time required, the positive or negative outcome, and the binding mode sampled at the end are reported along with the $RMSD_{min}$ (calculated using 2YDO as a reference). MD = molecular dynamics.*

We begin the description of the results from trajectories 1a, 6a, and 10a, in which the adenosine did not reach the orthosteric site (Table 3). Interestingly the ligand extensively sampled a metastable-binding site at the interface between ECL2 and ECL3, putatively representing an ancillary site of recognition besides the orthosteric one.[23,29] This intermediate binding mode was characterized by the polar interaction between the adenosine ribose moiety and the negatively charged residue $E169^{ECL2}$, as well as hydrophobic contacts with $M174^{5.35}$ and transient hydrogen bonds with residues at the ECL3 (Figure 2A). The interaction energy analysis (Figure S3) suggests that the stability of this metastable state is comparable with the adenosine in its crystallographic binding mode (Figure S4) and justifies the missed transition to the orthosteric site. The seven productive SuMD simulations (Table 3) allowed the adenosine to explore different conformations within the orthosteric site, including the crystallographic one. Trajectory 3a, indeed, was able to reproduce with great accuracy ($RMSD_{min}$ = 0.45 Å) the experimental binding mode (Video 3), with all the key interactions faithfully recovered (Figure 2B).[30] Interestingly, trajectories 2a, 4a, and 5a

described an alternative recognition mechanism, according to which the adenine ring of the agonist approaches the binding site, orienting the ribose moiety towards the extracellular (EC) receptor vestibule ("ribose-up" conformation). [31,32] These states were transient, as the classic MD simulations rapidly evolved towards the crystallographic binding mode, but without sampling the key hydrogen bond with residue S277$^{7.42}$ side chain (Figure 2C), due to the so-called syn conformation of the β-glycosidic bond (anti in the crystal structure).

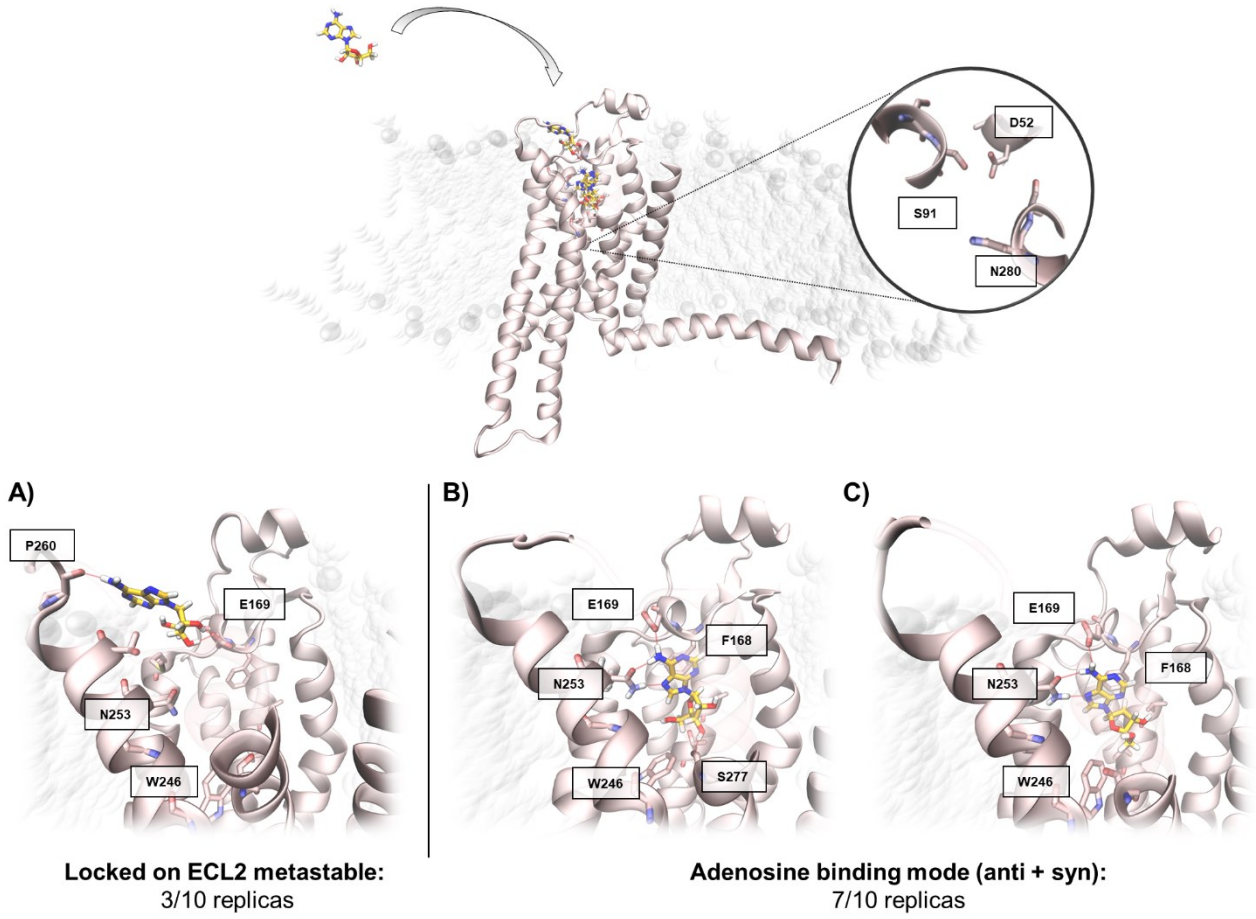


Figure 2 *Conformations sampled by the adenosine while recognizing the A_{2A} AR in the intermediate-active state. Top, the absence of a sodium ion in the allosteric binding site is highlighted. Panel A shows a representative adenosine binding mode in the extracellular loop 2 (ECL2) metastable binding. In panels B and C, the ribose in anti (B) and syn (C) conformation are reported. Only the syn orientation permits the hydrogen bonding with the residue S277$^{7.42}$.*

2.2. SuMD simulations of the adenosine on the inactive, sodium-bound, A_{2A} AR conformation

As anticipated, to verify the different adenosine propensities to recognize divergent A_{2A} AR conformational states, SuMD was performed on the inactive conformation of the receptor (PDB ID 4EIY), retaining the sodium in its allosteric site (Figure 3) but depleting the inverse agonist ZM241385. Consistently with the first part of this work, ten SuMD replicates were collected (as summarized in Table 4).

	SuMD time (ns)	*Reached the orthosteric site*	*Adenosine binding mode*	*X-ray binding mode after 100 ns of MD*	*$RMSD_{min}$ (Å)*
Replicate 1i	9.0	No	No (Meta-binding site on ECL2)	-	16.1
Replicate 2i	16.8	No	No (Meta-binding site on ECL2)	-	13.6
Replicate 3i	16.2	Yes	No (Receptor Vestibule)	No (Receptor Vestibule)	5.5
Replicate 4i	31.2	Yes	No (Receptor Vestibule)	No(Adenosine unbinding)	6.1
Replicate 5i	37.8	Yes	No (Receptor Vestibule)	No (Receptor Vestibule)	6.6
Replicate 6i	24.6	No	No (Meta-binding site on ECL2	-	15.4
Replicate 7i	7.8	No	No (Meta-binding site on ECL2)	-	14.7
Replicate 8i	7.8	No	No (Meta-binding site on ECL2)	-	13.8
Replicate 9i	8.4	Yes	No (Receptor Vestibule)	No (Adenosine unbinding)	7.9
Replicate 10i	45.6	Yes	No (Receptor Vestibule)	Yes	0.3

Table 4 *Summary of the adenosine SuMD simulations performed on the inactive conformation of the A2A AR. For each replicate, the SuMD simulation time required, the positive or negative outcome, and the binding mode sampled at the end is reported along with the RMSDmin (calculated using 2YDO as a reference).*

Unlike the intermediate-active conformation, on the inactive, sodium-coordinated A_{2A} AR just one replication out of ten resulted in the adenosine reproducing the experimental binding mode. Specifically, in half of the trajectories sampled (replicates 1i, 2i, 6i, 7i, and 8i in Table 4) adenosine did not reach the orthosteric site, but sampled the solvent-exposed metastable binding site at the interface between ECL2 and ECL3 (Figure 3A), again interacting with $E169^{ECL2}$ as reported in the previous section of the manuscript. The remaining five SuMD simulations were instead defined as quasi-productive, since the agonist reached the vestibular region of the orthosteric binding site without, however, reproducing the adenosine crystallographic pose. Lee and collaborators investigated, by means of classic MD simulation, the behaviour of adenosine within the inactive-state A_{2A} AR orthosteric site and pointed out the agonist's inability to maintain the original binding mode, thus corroborating our SuMD results.[33]

To evaluate the stability of the five quasi-productive SuMD final states (replicas 3i, 4i, 5i, 9i, and 10i) the trajectories were prolonged for 100 ns (unbiased MD). As reported in Figure 3B, during the extended trajectories, 3i and 5i the adenosine maintained its vestibular position. Trajectories 4i and 9i, on the other hand, were characterized by the spontaneous dissociation of the ligand, indicating a poor ligand stabilization (Figure 3C). Curiously, the extended trajectory of 10i was the only one during which the adenosine reached the experimental bound state ($RMSD_{min}$ = 0.3 Å Table 4, Figure 3D).

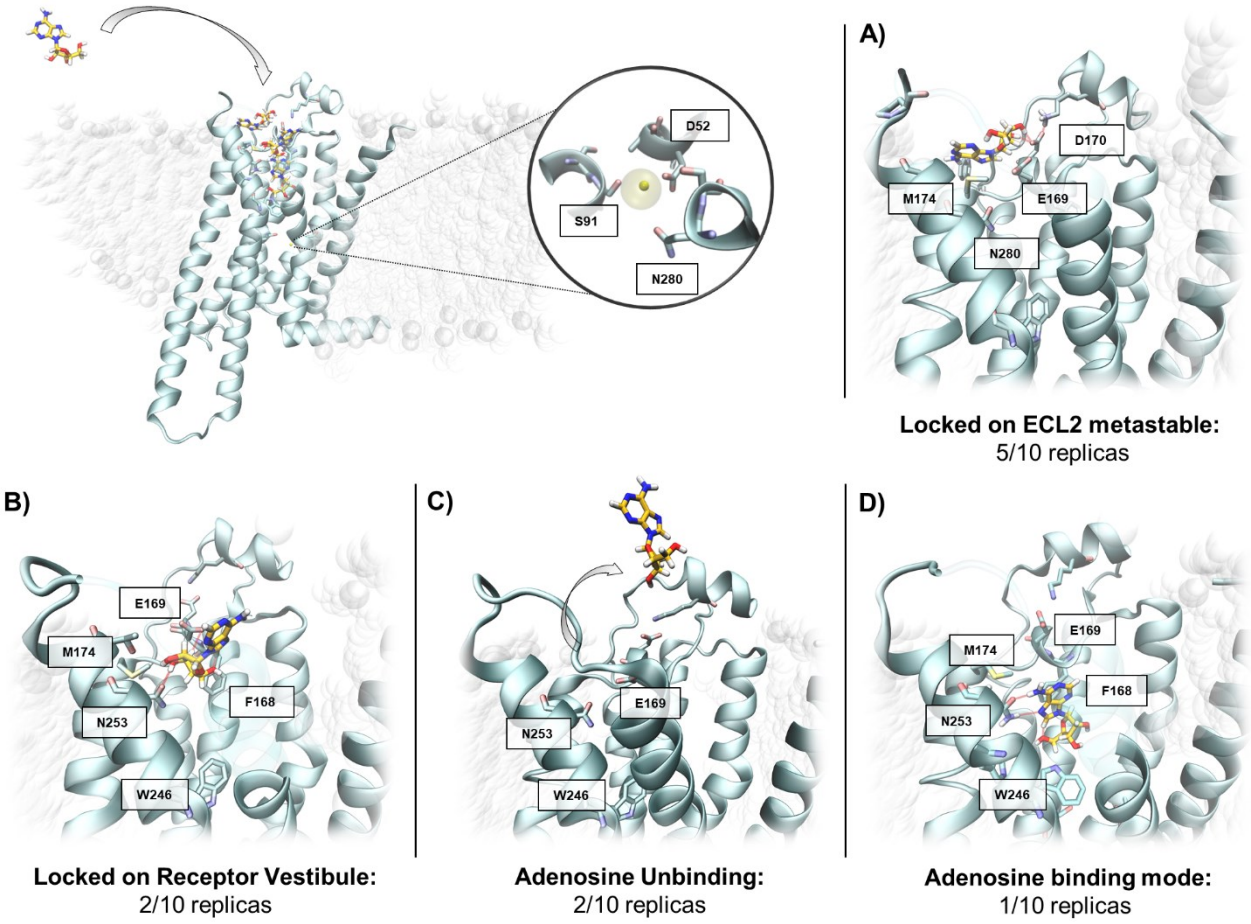


Figure 3 *Conformations sampled by the adenosine while recognizing the A_{2A} AR in the inactive state. Top, the presence of the sodium ion in the allosteric binding site is highlighted. In panel A, a representative adenosine binding mode in the ECL2 metastable binding site is depicted. In panel B, one of the different conformations sampled by the adenosine in the receptor vestibule is reported. Panel C summarizes the number of ligand unbinding events collected, starting from the vestibule region. Panel D represents the only SuMD simulation (Replica 10i) that showed an adenosine crystallographic binding mode.*

2.3. Insight on the role of the sodium ion in the recognition of A_{2A} AR agonists

In a schematic way, Na^+ coordination within the allosteric TMD allows for the discrimination of the two main conformational states of A_{2A} AR (i.e., active and inactive); it is capable of recognizing adenosine with antithetical efficiency, as suggested by the divergent binding frequencies sampled through the SuMD simulations. As highlighted in Figure S10, in the supplementary material, the limited structural differences between the two crystallographic conformations of the receptor would not be sufficient to explain, from a mechanistic point of view, the negative allosteric effect mediated by a sodium ion. Consequently, the use of techniques able to take into consideration the conformational plasticity associated with the receptor functionality is essential to realistically rationalize the role played by the monovalent ion. To decipher the molecular basis underneath such misleading outcomes described by the SuMD simulations (i.e., replicas 3a and 5i, sampled,

respectively, starting from the active and inactive receptor states), cumulative maps of the interatomic contacts between adenosine and A_{2A} AR binding site residues were graphically depicted, using polar diagrams. As reported in Figure 4, box A, the agonist's inability to reproduce the canonical experimental conformation in the receptor inactive state is accompanied by discrepancies in the adenosine recognition pathway, mainly at the level of TM1, TM2, and TM7. These differences, on the other hand, were not noticed during replicate 10i, the only productive trajectory sampled starting from the inactive state of the receptor in the presence of the sodium ion, as indicated in Figure 4, box B. These data further emphasize the importance of residues located in TM1, TM2, and TM7 for the correct molecular recognition process of agonists.

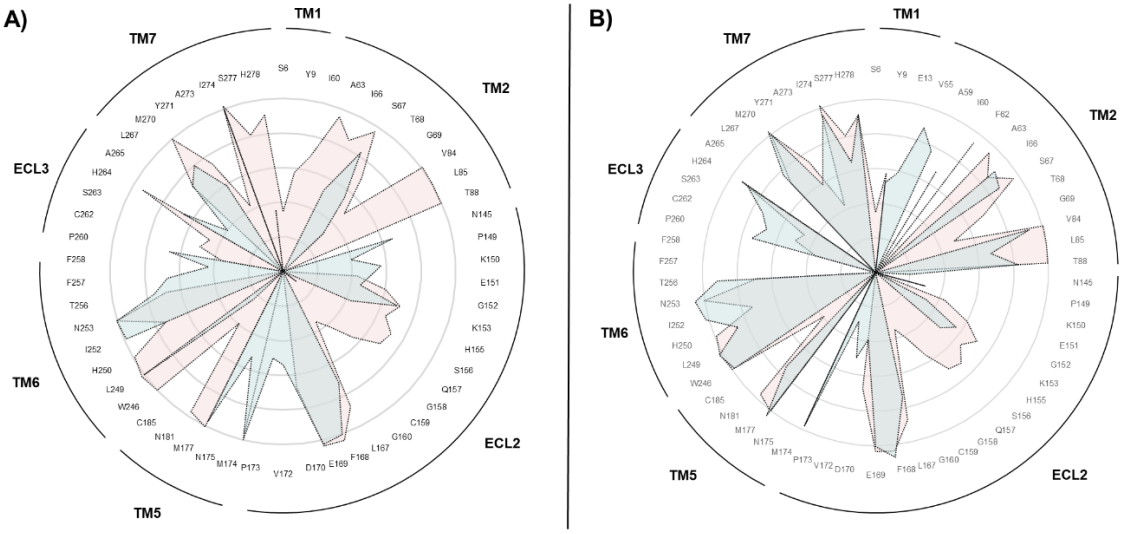


Figure 4 *The adenosine experienced different patterns of interactions during SuMD dynamic docking on the intermediate-active and inactive A_{2A} AR conformations. The adenosine-A_{2A} AR contacts are plotted as polar diagrams of overlapping data. In panel A replicate 3a (productive binding to the intermediate-active receptor state, pink) and Replicate 5i (quasi-productive to the inactive receptor state, cyan) are compared. In panel B, replicate 3a (productive binding to the intermediate-active receptor state, pink) is compared with replicate 10i (the only productive binding to the inactive receptor state, cyan).*

Deciphering the dynamics of the A_{2A} AR states is fundamental to interpreting the discrepant agonist recognition pathways. In a recent computational investigation, an increased flexibility of A_{2A} AR EC domains was described in the receptor inactive state, a phenomenon that is less relevant in the active conformation and thus could help in differentiating agonist binding mechansisms.[33] To verify if this evidence can be extrapolated from our SuMD simulations, the volume of the orthosteric binding site was dynamically monitored in the two aforementioned trajectories (replicates 3a and 5i). Interestingly, even if the starting volumes computed for the A_{2A} AR binding site on both

crystallographic structures taken under examination were quite similar (~ 250 Å^3), only a few ns of simulation were required to reveal the different evolutions of the two systems.

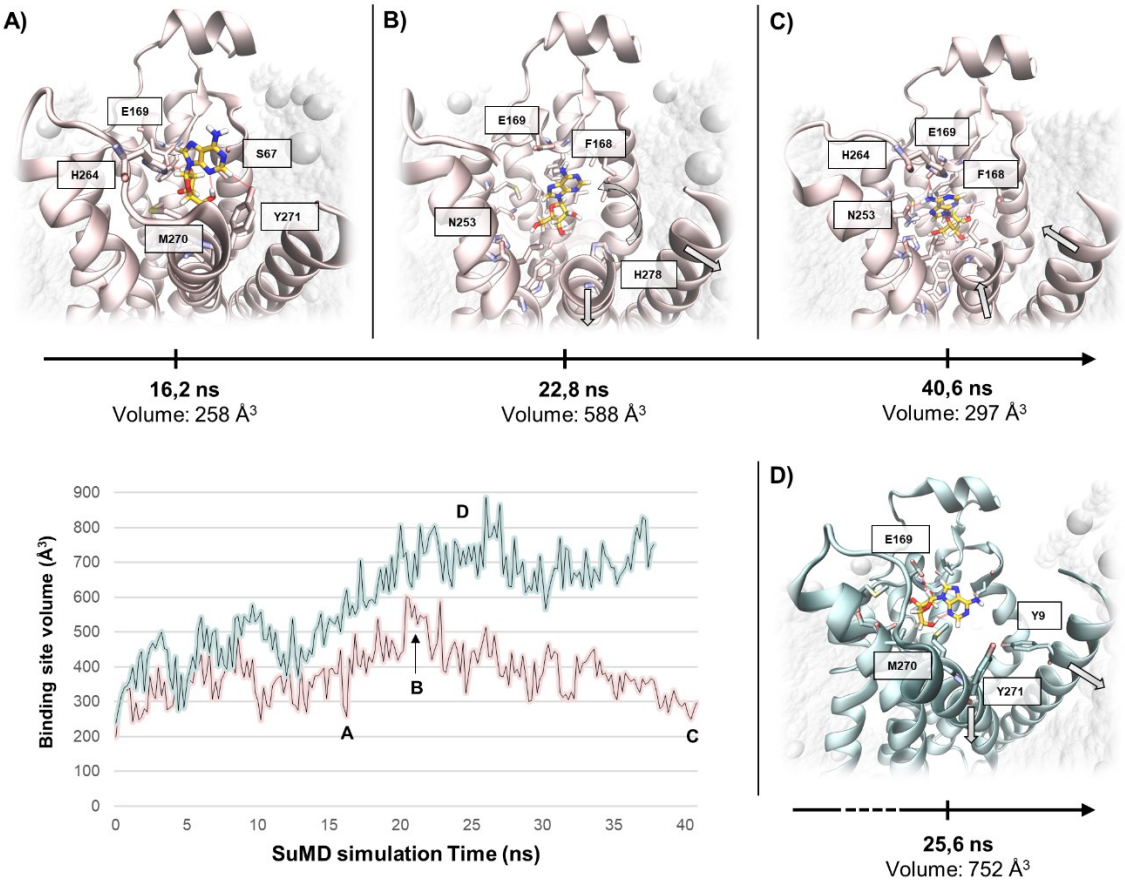


Figure 5 *The orthosteric site volumes change differently during SuMD simulations of the intermediate-active and the inactive A_{2A} ARs. Panels A, B, and C depict three snapshots from SuMD Replicate 3a, related to the key steps of the adenosine recognition. Initially, the agonist approaches the A_{2A} AR extracellular vestibule (A) and through a polar interactions network mediated by ECL3, TM2, and TM7 (whose overall organization is not perturbed with respect to the crystal structure), inserts the purine ring into the binding site (B). The adenosine is then able to reach the canonical binding mode (C) only when the cavity volume recedes toward the original value. On the A_{2A} AR inactive state, the binding site volume progressively increases due to the TM1 and TM7 outward movements (Panel D), making the agonist binding more difficult.*

On the intermediate-active conformation of the A_{2A} AR, adenosine approached the receptor, interacting with the vestibular region ECL2 (Figure 2A). The transition to the orthosteric binding site was mediated by a series of polar interactions with residues located at the ECL2, TM2, and TM7. In this phase, the compactness of the receptor orthosteric site was necessary for the productive adenosine recognitions, as indicated by the small fluctuation of the cavity volume (Figure 5A). From this standpoint, the accommodation of the adenosine in the orthosteric site required the first adaptation of the surrounding TM helices, as suggested by a transient increase in the volume up to a value of about 600 Å^3 (Figure 5B). Subsequently, the π-stabilizing interaction of the adenine

nucleus with the side chain of Phe168 compacts the structure of the recognition cavity, bringing its volume back to a value similar to the initial one (Figure5).

The presence of the sodium ion within its putative binding site in the inactive A_{2A} AR conformation markedly altered the receptor flexibility. Indeed, during the first step of the simulation, the TM1 and TM7 moved outwards, progressively increasing the volume of the orthosteric site up to about 700 $Å^3$, not allowing the driving interactions to the bound final state to be established (Figure 5D). As previously described, the outward movement of segment TM7, combined with TM2 shifting from TM3, represents the key steps for Na^+ coordination in the active state of A_{2A} AR. It is reasonable to speculate that the presence of the monovalent ion in the middle of the 7TM bundle could be responsible for the greater flexibility of the extracellular portion of the receptor, allowing it to alter the dynamics of the TM2 and TM7, and thus the agonist binding mechanism.

4. Materials and Methods

4.1. General

MOE suite (Molecular Operating Environment, version 2018.0101) was exploited to perform most of the general molecular modelling operations, such as proteins and ligands preparation.[34] 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 ACEMD engine on a GPU cluster composed of 18 NVIDIA drivers, ranging from GTX 780 to Titan V.[35] For all the simulations, the CHARMM36/CHARMM general force field (CGenFF) combination was adopted.[36–38]

4.2. Systems preparation

Agonist and antagonist-bound complexes of A_{2A} AR were retrieved from the RCSB Protein Data Bank database (PDB ID 2YDO and 4EIY respectively) and handled by means of the MOE protein structure preparation tool.[6,30] Hydrogen atoms were assigned according to Protonate-3D, and any missing loop was modelled with the homology modelling protocol.[39] In the case of PDB ID 4EIY, the apocytochrome b562 (BRIL) inserted in the ICL3 was removed prior to protein preparation and loop modeling. Missing atoms in the side chains, as well as non-natural N-terminals and C-terminals, were rebuilt according to the CHARMM force field topology.[36] A_{2A} AR apo forms were obtained by simply deleting the orthosteric ligands from their respective complexes. Adenosine force field parameters were retrieved from the Paramchem web service, in concordance with CGenFF.[37,38]

4.4. Solvated System Setup and Equilibration

Systems were embedded in a 1-palmitoyl-2oleyl-sn-glycerol-3-phospho-choline (POPC) lipid bilayer, according to the pre-orientation provided by the Orientations of Proteins in Membrane (OPM) database and by using the VMD membrane builder plugin.[40,41] Lipids within 0.6 Å from the protein were removed and TIP3P model water molecules were added to solvate the systems by means of Solvate1.0.[42,43] Systems charge neutrality was reached by adding 100 Na^+ atoms and 111 Cl^+ counterions to a final concentration of 0.154 M (A_{2A} AR net charge was +11 for both the system-simulated 2YDO/4EIY). Equilibration was performed through a three-step procedure. In the first step, 1500 conjugate-gradient minimization steps were applied to reduce the clashes between proteins and lipids. Then, a 5 ns long MD simulation was performed in the NPT ensemble (Isothermal–isobaric statistic ensemble), with a positional constraint of 1 kcal mol^{-1} $Å^{-2}$ on ligand, protein, and lipid phosphorus atoms. During the second stage, 10 ns of MD simulation in the NPT ensemble were performed constraining all the protein and ligand atoms but leaving POPC residues free to diffuse in the bilayer. In the last equilibration stage, positional constraints were applied only to the ligand and protein backbone alpha carbons for a further 5 ns of MD simulation.

All the MD simulations were performed using the following protocols: an integration time step of 2 fs; a Berendsen barostat maintained the system pressure at 1 atm; a Langevin thermostat maintained the temperature at 310 K with a low dumping of 1 ps^{-1}; the M-SHAKE algorithm constrained the bond lengths involving hydrogen atoms.[44–46]

4.5. Supervised Molecular Dynamics (SuMD) Simulations

Supervised molecular dynamics (SuMD) simulations were exploited to sample and characterize the binding pathway of the Na^+ monovalent ion, as well to simulate the binding of the endogenous agonist adenosine to the two pharmacologically relevant A_{2A} AR conformations.[21–23,31] SuMD methodology reduces the timescale necessary 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 and centre of mass of the protein binding site, during unbiased MD simulations. A series of short unbiased MD simulations is performed, and after each simulation, the distance points collected 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 to the binding site. Otherwise, the simulation is restarted by randomly assigning the atomic velocities. The

length of each SuMD step in which the supervision is carried out was adapted relative to the nature of the ligand under investigation. In terms of the sodium ion, given its important diffusion rate, a 200 ps SuMD time window proved to be adequate to accurately describe the binding, whereas for adenosine, the classic SuMD time window of 600 ps, previously optimized and validated for small organic molecules, was set. Short simulations are perpetuated under supervision until the distance between the ligand and the binding site dropped below 5 Å, then the supervision was disabled, and a classical MD simulation was performed. In the present study, for the computation of the allosteric Na+ binding site centre of mass, residues D52, S91, and N280 were chosen; for the orthosteric A_{2A} AR binding site, residues N253, F168, H250, and H278 were selected.

In all SuMD productive replicates in which adenosine reached the orthosteric binding site, the final state evolution and stability was evaluated through the collection of a 100 ns long classical MD.

4.6. SuMD trajectory analysis

All the SuMD trajectories collected were analysed by an in-house tool written in tcl and python languages, as described in the original publication.[22] Briefly, dimension of each trajectory was reduced saving MD frames at a 20 ps interval, each trajectory was then superposed on the first-frame Cα carbon atoms of the A_{2A} AR and wrapped into an image of the system simulated under periodic boundary condition. In those cases where a reference was present, the RMSD of the ion or adenosine molecule was computed with respect to the experimental crystallographic complex (4EIY for sodium and 2YDO for adenosine). The RMSD values were plotted over time and reported in the movies present in the supplementary materials.

SuMD trajectories investigating the recognition pathway of sodium were furthermore geometrically analysed to identify significant populations of ion position, among the multitude of sampled data. Prody, a python framework for MD manipulation and analysis, was exploited to compute the pairwise root mean square deviations (RMSDs) of Na^+ atomic coordinates, during all replicates collected.[47] From each replicate, a square matrix of RMSDs was obtained (nf x nf), in which nf stands for the number of trajectory frames. Subsequently, DBSCAN, a density-based clustering algorithm, part of the scikit-learn python packages, was applied to cluster the different ion atomic positions and graphically represent them by exploiting VMD software.[25,41] The orthosteric binding site volume was dynamically monitored in the SuMD trajectories of adenosine recognition, collected starting from the two different A_{2A} AR conformations. POVME 2 python software was exploited to perform the calculation, after defining a spherical inclusion region cantered on agonist centroid coordinates and characterized by a 9 Å radius dimension.[48]

5. Conclusion

The molecular mechanism that triggers the negative allosteric modulation of the sodium ion on the A_{2A} AR agonists is not fully understood. X-ray structural studies have pointed out the presence of a binding site for the cation in the core of the TMD of the resting receptor (and many other class A GPCRs). However, the high degree of similarity with the intermediate-active (agonist-bound) state of the receptor (Figure S10) does not completely clarify the molecular basis of this effect. In this study, the SuMD technique was therefore employed to simulate the binding processes of the sodium ion and the endogenous agonist adenosine on these two different A_{2A} AR conformations (the intermediate-active and inactive one, respectively), in the attempt to retrieve mechanistic insight.

The Na^+, whose concentration in the extracellular environment is close to 140 mM, has a fundamental role in controlling the conformational landscape of the A_{2A} AR, characterized by few, highly populated, stable states. The most accepted model describes the sodium as capable of selectively binding only to the inactive-like receptor population. The macroscopic effect of this is a shift of the equilibrium towards the resting state of the receptor, and a decrease in affinity towards agonists. In keeping with this conformational selectivity as well with previous work, our simulations outlined the A_{2A} AR inactive structure as able to coordinate the sodium ion without any topological modification of the putative allosteric site.[10,16] On the other hand, during the simulated binding on the intermediate-active conformation, an increase of the inter-TM distances was necessary to accommodate the cation, possibly anticipating a receptor transition toward the inactive-state. The "toggle switch" $W246^{6.48}$ was pointed out as a possible gatekeeper of the sodium binding event. Interestingly, SuMD suggested different binding paths on the two A_{2A} AR states. It is intriguing to speculate that the inactive state of the receptor could selectively drive the binding of the sodium ion by putatively shaping the charge distribution of the meta-stable binding sites along the path.

During the successive SuMD simulations, the endogenous agonist showed a propensity to bind the sodium-free intermediate-active state of the receptor (Video 2). Indeed, seven simulations out of ten resulted in an orthosteric complex, while only one SuMD replica on the inactive structure was productive. We propose the different flexibilities of the extracellular side of the receptor (where the first interactions able to influence the agonists binding occur) as a driving force of this divergence. The presence of the sodium ion in its allosteric site possibly prevented the receptor from adapting to the incoming agonist, due to an opening up of the EC vestibule and, in turn, of the orthosteric site. As a partial confirmation of this, the TM1, TM2, and ECL2 formed less extensive contacts with

the adenosine in the inactive A_{2A} AR (Figure 5) due to the increased volume of the orthosteric site (Figure 4).

The speculative mechanism proposed in this work should be further investigated on other GPCRs.

References:

1. Wacker, D., Stevens, R. C. & Roth, B. L. How Ligands Illuminate GPCR Molecular Pharmacology. Cell 170, 414–427 (2017).

2. Vecchio, E. A. et al. New paradigms in adenosine receptor pharmacology: allostery, oligomerization and biased agonism. Br. J. Pharmacol. 175, 4036–4046 (2018).

3. 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).

4. Carpenter, B. & Lebon, G. Human Adenosine A2A Receptor: Molecular Mechanism of Ligand Binding and Activation. Front. Pharmacol. 8, 898 (2017).

5. Berman, H. M. et al. The protein data bank. Nucleic Acids Res. 28, 235–242 (2000).

6. Liu, W. et al. Structural basis for allosteric regulation of GPCRs by sodium ions. Science 337, 232–6 (2012).

7. Pert, C. B., Pasternak, G. & Snyder, S. H. Opiate agonists and antagonists discriminated by receptor binding in brain. Science (80-.). (1973). doi:10.1126/science.182.4119.1359

8. PERT, C. B. & SNYDER, S. H. Opiate Receptor Binding of Agonists and Antagonists Affected Differentially by Sodium. Mol. Pharmacol. 10, (1974).

9. Katritch, V. et al. Allosteric sodium in class A GPCR signaling. Trends Biochem. Sci. 39, 233–244 (2014).

10. Massink, A. et al. Sodium Ion Binding Pocket Mutations and Adenosine A 2A Receptor Function s. Mol. Pharmacol. Mol Pharmacol 87, 305–313 (2015).

11. Gao, Z.-G. & Ijzerman, A. P. Allosteric modulation of A2A adenosine receptors by amiloride analogues and sodium ions. Biochem. Pharmacol. 60, 669–676 (2000).

12. Ballesteros, J. A. & Weinstein, H. [19] Integrated methods for the construction of three-dimensional models and computational probing of structure-function relations in G protein-coupled receptors. Methods Neurosci. 25, 366–428 (1995).

13. Wootten, D., Christopoulos, A. & Sexton, P. M. Emerging paradigms in GPCR allostery: implications for drug discovery. Nat. Rev. Drug Discov. 12, 630–644 (2013).

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. Selvam, B., Shamsi, Z. & Shukla, D. Universality of the Sodium Ion Binding Mechanism in Class A G-Protein-Coupled Receptors. Angew. Chemie 130, 3102–3107 (2018).

16. Gutiérrez-de-Terán, H. et al. The Role of a Sodium Ion Binding Site in the Allosteric Modulation of the A2A Adenosine G Protein-Coupled Receptor. Structure 21, 2175–2185 (2013).

17. Vickery, O. N. et al. Intracellular Transfer of Na+ in an Active-State G-Protein-Coupled Receptor. Structure 26, 171-180.e2 (2018).

18. Hu, X. et al. Kinetic and thermodynamic insights into sodium ion translocation through the μ-opioid receptor from molecular dynamics and machine learning analysis. PLOS Comput. Biol. 15, e1006689 (2019).

19. Shang, Y. et al. Mechanistic Insights into the Allosteric Modulation of Opioid Receptors by Sodium Ions. Biochemistry 53, 5140–5149 (2014).

20. Fleetwood, O., Matricon, P., Carlsson, J. & Delemotte, L. Energy landscapes reveal agonist's control of GPCR activation via microswitches. bioRxiv 627026 (2019). doi:10.1101/627026

21. 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).

22. 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).

23. 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).

24. Latorraca, N. R., Venkatakrishnan, A. J. & Dror, R. O. GPCR Dynamics: Structures in Motion. Chem. Rev. 117, 139–155 (2017).

25. 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).

26. Segala, E. et al. Controlling the Dissociation of Ligands from the Adenosine A 2A Receptor through Modulation of Salt Bridge Strength. J. Med. Chem. 59, 6470–6479 (2016).

27. Pang, X., Yang, M. & Han, K. Antagonist binding and induced conformational dynamics of GPCR A 2A adenosine receptor. Proteins Struct. Funct. Bioinforma. 81, 1399–1410 (2013).

28. Yuan, S., Hu, Z., Filipek, S. & Vogel, H. W246 6.48 opens a gate for a continuous intrinsic water pathway during activation of the adenosine A2A receptor. Angew. Chemie - Int. Ed. 54, 556–559 (2015).

29. Igonet, S. et al. Enabling STD-NMR fragment screening using stabilized native GPCR: A case study of adenosine receptor. Sci. Rep. 8, 8142 (2018).

30. Lebon, G. et al. Agonist-bound adenosine A2A receptor structures reveal common features of GPCR activation. Nature 474, 521–525 (2011).

31. 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).

32. 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).

33. Lee, S., Nivedha, A. K., Tate, C. G. & Vaidehi, N. Dynamic Role of the G Protein in Stabilizing the Active State of the Adenosine A2A Receptor. Structure 27, 703-712.e3 (2019).

34. ULC, C. C. G. Molecular Operating Environment (MOE). (2013).

35. 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).

36. Huang, J. & MacKerell, A. D. CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data. J. Comput. Chem. 34, 2135–2145 (2013).

37. Vanommeslaeghe, K. & MacKerell, A. D. Automation of the CHARMM General Force Field (CGenFF) I: Bond Perception and Atom Typing. J. Chem. Inf. Model. 52, 3144–3154 (2012).

38. Vanommeslaeghe, K., Raman, E. P. & MacKerell, A. D. Automation of the CHARMM General Force Field (CGenFF) II: Assignment of Bonded Parameters and Partial Atomic Charges. J. Chem. Inf. Model. 52, 3155–3168 (2012).

39. Labute, P. Protonate 3D: assignment of macromolecular protonation state and geometry. Chem. Comput. Gr. Inc (2007).

40. Lomize, M. A., Lomize, A. L., Pogozheva, I. D. & Mosberg, H. I. OPM: Orientations of Proteins in Membranes database. Bioinformatics 22, 623–625 (2006).

41. Humphrey, W., Dalke, A. & Schulten, K. VMD - Visual Molecular Dynamics. J. Mol. Graph. 14, 33–38 (1996).

42. 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).

43. Grubmuller, H.; Groll, V. Solvate 1.0.

44. Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A. & Haak, J. R. Molecular dynamics with coupling to an external bath. J. Chem. Phys. 81, 3684–3690 (1984).

45. 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).

46. Essmann, U. et al. A smooth particle mesh Ewald method. J. Chem. Phys. 103, 8577–8593 (1995).

47. Bakan, A., Meireles, L. M. & Bahar, I. ProDy: protein dynamics inferred from theory and experiments. Bioinformatics 27, 1575–1577 (2011).

48. Durrant, J. D., Votapka, L., Sørensen, J. & Amaro, R. E. POVME 2.0: An Enhanced Tool for Determining Pocket Shape and Volume Characteristics. J. Chem. Theory Comput. 10, 5047–5056 (2014).

Deciphering the Molecular Recognition Mechanism of Multidrug Resistance Staphylococcus Aureus NorA Efflux Pump Using a Supervised Molecular Dynamics Approach

Deborah Palazzotti, Maicol Bissaro, **Giovanni Bolcato**, Andrea Astolfi, Tommaso Felicetti, Stefano Sabatini, Mattia Sturlese, Violetta Cecchetti, Maria Letizia Barreca, Stefano Moro

Palazzotti, D. *et al.* Deciphering the molecular recognition mechanism of multidrug resistance staphylococcus aureus nora efflux pump using a supervised molecular dynamics approach. *Int. J. Mol. Sci.* (2019)

Abstract

The use and misuse of antibiotics has resulted in critical conditions for drug-resistant bacteria emergency, accelerating the development of antimicrobial resistance (AMR). In this context, the co-administration of an antibiotic with a compound able to restore sufficient antibacterial activity may be a successful strategy. In particular, the identification of efflux pump inhibitors (EPIs) holds promise for new antibiotic resistance breakers (ARBs). Indeed, bacterial efflux pumps have a key role in AMR development; for instance, NorA efflux pump contributes to Staphylococcus aureus (*S. aureus*) resistance against fluoroquinolone antibiotics (e.g., ciprofloxacin) by promoting their active extrusion from the cells. Even though NorA efflux pump is known to be a potential target for EPIs development, the absence of structural information about this protein and the little knowledge available on its mechanism of action have strongly hampered rational drug discovery efforts in this area. In the present work, we investigated at the molecular level the substrate recognition pathway of NorA through a Supervised Molecular Dynamics (SuMD) approach, using a NorA homology model. Specific amino acids were identified as playing a key role in the efflux pump-mediated extrusion of its substrate, paving the way for a deeper understanding of both the mechanisms of action and the inhibition of such efflux pumps.

1. Introduction

Antimicrobial resistance (AMR) is a complex global health challenge, mainly resulting from the excessive use and abuse of antimicrobial agents in humans and animals. [1] Indeed, over the years, the microbial world has developed the molecular tools to drive resistance and evade antibiotic action via (i) alteration of targeted site, (ii) enzymatic drug inactivation/modification, (iii) decreased uptake or enhanced efflux of the drug, and (iv) biofilm formation [2].

In this context, Staphylococcus aureus represents the most dangerous superbug among Gram-positive organisms due to its ability to develop resistance to a wide range of compounds [3]. S. aureus possesses several efflux pumps belonging to different families able to extrude a wide array of common antibacterial drugs [4]. NorA is a multidrug resistance (MDR) efflux pump, well-studied since 1986 when it was isolated from the urine of a patient treated with norfloxacin (NOR).

NorA was thus the first chromosomally-encoded S. aureus MDR pump to be identified: it is codified by norA gene and expressed in 43% of bacterial strains [5]. From a structural point of view, NorA is a single-chain transmembrane protein of 42,385 kDa composed of 388 amino acids. It belongs to the Major Facilitator Superfamily (MFS) consisting of 12 hydrophobic transmembrane (TM) α-helices with the N- and C-terminal domains that are placed in the cytoplasmic side, connected by hydrophilic loops and arranged as pseudo-twofold symmetry [6,7]. Unfortunately, little is known about the mechanism of efflux, except that it works by using the proton that allows the entry of a proton-coupled to the extrusion of the drug from the cell. Indeed, NorA is classified as a drug/H^+ antiporter. NorA overexpression is associated with drug resistance. In particular, NorA is a promiscuous efflux pump involved in quinolones and fluoroquinolones (such as ciprofloxacin—CPX) resistance [8], but also in the extrusion of other natural and synthetic structurally unrelated compounds (e.g., quaternary ammonium compounds and antiseptics, phenothiazines and thioxanthenes, totarol, ferruginol, carnosic acid, ethidium bromide (EtBr), tetraphenylphosphonium, rhodamine, acridine, and biocides)[8.]

To date, several scientific efforts have been made to identify efflux pump inhibitors (EPIs) with the final aim to counteract the S. aureus resistance mechanism and restore bacterial susceptibility to antibiotic action [8–16]. Even though structure information of different drug/H^+ antiporter are publicly available [17–20] in the RCSB Protein Data Bank (PDB) [21], unfortunately, neither 3D structures of NorA have been made public nor computational studies have been reported to understand the recognition mechanism between the efflux pump and the substrate.

Against this backdrop, the aim of the present work was to explore at the molecular level the possible recognition pathway and interactions between NorA efflux pump and its substrate CPX by using a Supervised Molecular Dynamics approach (SuMD) [22]. In brief, a SuMD simulation is composed of a number of consecutive short unbiased MD simulations (600 ps) in which a supervision strategy, based on a tabu search-like strategy, is applied at the end of each simulation. The supervised variable is the distance between the ligand and protein binding site center of mass (dcm (L-R)). In few words, if this distance is likely to be shortened during the simulation, the MD simulation is

prolonged, otherwise, it is stopped, and the simulation is restarted from the previous set of coordinates. The supervision is maintained until the protein-ligand distance reaches a pre-set threshold value, then the simulation proceeds as a conventional unbiased MD simulation.
SuMD aided for the first time the recognition pathway of the efflux pump NorA, with the substrate CPX giving interesting information about the sites explored during its trajectory prior to extrusion toward the periplasmatic side.

2. Results and Discussion

2.1. Prediction and Assessment of the NorA 3D Structure

First, four bioinformatics tools—I-TASSER [23,24], SWISS-MODEL [25], RaptorX [26] and Phyre2 [27]—were used to generate NorA efflux pump homology models (Table S1; Supplementary Materials). Overall, two different conformations of NorA were obtained as output: an outward conformation (Cout) with an opening toward the periplasmic side, and an inward conformation (Cin) with an opening toward the cytoplasmic side. Given our main interest in the molecular recognition mechanisms underneath the interactions between a substrate and the transporter immediately antecedent to its extrusion, we decided to focus our subsequent studies on the predicted inward conformations. The different software used provided us with three Cin models using three different templates. Indeed I-TASSER, RaptorX, and Phyre2 produced Cin models built based on the MSF E. coli MdfA transporter (PDB: 4ZOW) [19], the MFS proton-dependent oligopeptide transporters (POTs) of E. coli (PDB: 4IKV) [28] and the MSF E. coli MdfA transporter (PDB ID 4ZP0), respectively (Table S1; Supplementary Materials). Interestingly, in the 4ZOW crystal structure, the MdfA efflux pump is co-crystallized with its substrate chloramphenicol (CLM).
The quality of the models was assessed on the basis of the geometry using MOE suite [29] and the Qualitative Model Energy ANalysis (QMEAN) value (Table S1). Model's evaluation was also performed according to the Root Mean Square Deviations (RMSD). Noteworthy the RaptorX model RMSD was 3.234 Å, while the I-TASSER Cin RMSD on the template was 1.7 Å (Figure S1) and Phyre2 Cin RMSD was 0.35 Å. All models were good and reliable from a geometric point of view, quality of prediction and RMSD compared to the templates. To support the choice of our model we also carried out a sequence alignment with MOE suite to evaluate the sequence similarity and identity between the template and the NorA model generated. Although the percentage of sequence identity was low for all models, the choice of the model generated from MdfA was strongly

supported by similarity (Figure S2). We chose the I-TASSER model Cin because, at the same quality, it was built on a crystal in which the substrate was present [19].

2.2. MdfA Template and NorA Comparison

Aside from predicting homology models, the Phyre2 web portal provided useful information to better understand the evolutionary correlation between NorA and MdfA. Indeed, even though the identity similarity percentage predicted by Phyre2 between the crystal structure of MdfA (PDB ID 4ZP0) and the NorA model was very low (11% sequence identity), the associated confidence score, obtained by alignment of the sequence, was equal to 100%. Phylogenetically NorA and MdfA are strictly related: they belong in fact to the same transporter superfamily, MFS. Moreover, they also belong to the subfamily of drug/H^+ antiporters. Thus, with a high degree of confidence, the software considered these two transporters as analogs, therefore hypothesizing a possible conserved transport mechanism. Since the two structures were closely phylogenetically linked, and therefore perform the same function, it is assumed that the folding and the generated structure can be reliable.

2.3. Biological Assay of CLM on NorA

While it is well known that CLM is a substrate of MdfA, there is no information in the literature about the possible role of CLM as a NorA substrate. As mentioned before, the superimposition between the generated NorA homology model and 4ZOW (MdfA co-crystallized with CLM) suggested a very close structure organization (RMSD of 1.7 Å). In 4ZOW structure, CLM performed two key interactions with Asn33 and Asp34. However, the visual inspection of the NorA amino acids corresponding to these two MdfA acidic residues highlighted the presence of Ile19 and Gly20 (Supplementary Materials, Figure S3).

Thus, we supposed that NorA could not extrude CLM. In order to have some experimental evidence on this topic, we evaluated the CLM minimum inhibitory concentration (MIC) on two different S. aureus strains, one of which was wild-type (SA-1199-norA wt) and the other one overexpressing the norA gene and also possessing an A116E GrlA substitution (SA-1199B-norA+), which is a known fluoroquinolones target [30]. The obtained results showed that CLM had the same MIC values (4µg/mL) against the two used strains, thus highlighting that this compound could retain its antibacterial effect regardless of the NorA efflux pump overexpression. Indeed, MIC values of CPX and EtBr, known NorA substrates, appeared significantly different against SA-1199 and SA-1199B (Table 1). This data clearly demonstrated that CLM is not a NorA substrate.

2.3. Refinement of the NorA Predicted Model Using MD

The chosen homology model (i.e., I-TASSER Cin) was embedded in a 1-palmitoyl-2-oleyl-glycerol-3-phospho-choline (POPC) bilayer (Figure 1a) and subjected to MD simulations of 500 ns for structural refinement. All the subsequent analyses performed have been conducted in parallel using three different systems, i.e., (i) NorA homology model, (ii) MdfA in complex with CLM (PDB ID 4ZOW) and (iii) MdfA apo. The latter system was used as a reference structure. As highlighted by Figure 1b, the RMSD value of C_α showed good model stability for NorA. RMSD quickly reached a maximum value of approximately 4 Å, which remained steady and constant during the dynamic simulation time. Since the analyzed structure was a homology model, the value obtained, and in particular, the stability achieved can be considered good enough to validate the model. In addition, comparing the RMSDs trends for NorA and MdfA (Figure S4, Supplementary Materials), it was remarkable that the generated homology model became stable after 60 ns of the simulation time and seemed even more stable than the MdfA crystallographic structure. In accordance with the interval time within which SuMD samples binding events, the model can be considered stable. The most significant residue fluctuations occurred at the level of the loop connecting helix 6 and helix 7 and of the C-term and N-term domains (Figure 1c,d).

Furthermore, to evaluate possible conformational changes during the NorA MD simulation, we clustered the MD conformations using the density-based algorithm DBSCAN [32]. Although the whole MD protein conformations during the trajectory could be divided into two main clusters, we considered only the first cluster for its higher density. Indeed, the first cluster was populated by 4928 protein conformations out of a total of 5000. The centroid conformation of this cluster was then selected for the structural analysis.

The available biological data showed that CPX and CLM were endowed with different specificity for the MdfA and NorA efflux pumps. In particular, while CPX was a substrate of both proteins [30,33], no substrate activity against NorA was observed in our assays for CLM.

Thus, we planned differently in silico approaches (including SuMD simulations) to get insights about the different behavior of the two ligands on MdfA and NorA efflux pumps.

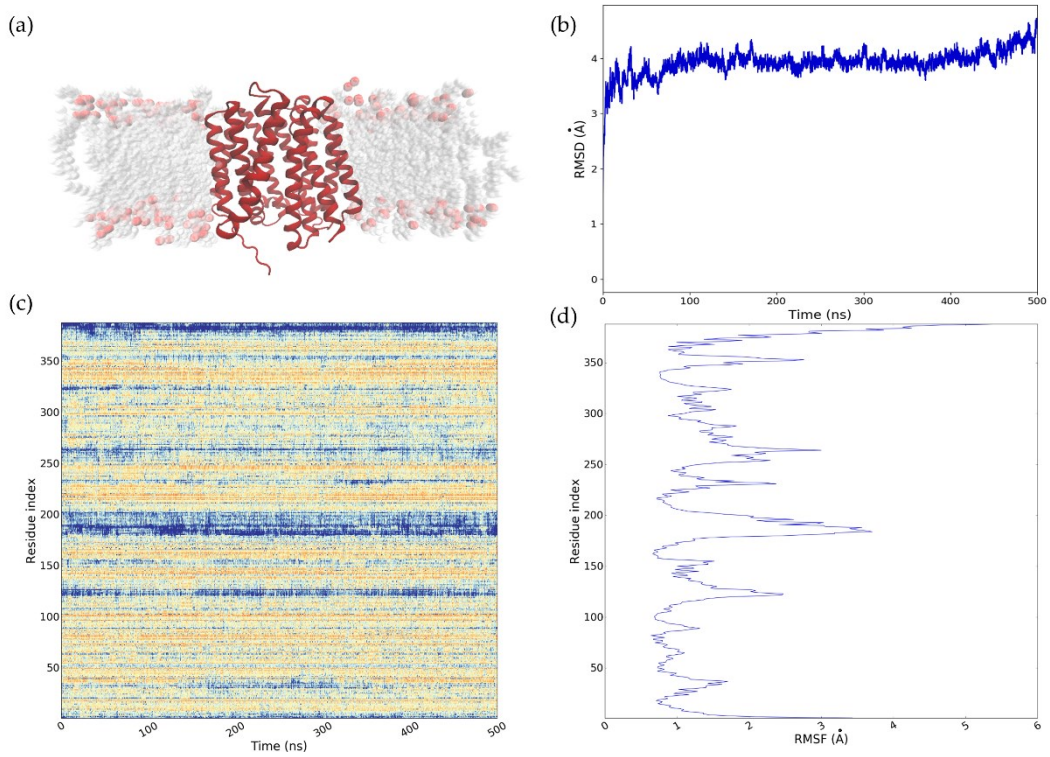


Figure 1 *(a) NorA homology model embedded in POPC bilayer. (b) Calculated RMSD graph of 500 ns of MD simulation of multidrug resistance S. aureus NorA efflux pump. Time (ns) is plotted on the x-axis and RMSD (Å) on the y-axis. (c) RMSF fluctuation during the MD simulation time of the NorA model. Depending on the intensity of the fluctuation, the color ranges from yellow (low RMSF) to blue, for higher values. (d) RMSF of the protein residues.*

2.4. Binding Site Definition and Preliminary Docking

Since the SuMD approach requires the binding site knowledge to address the ligand in the right direction, we performed preliminary docking study only to assess the ability of the two ligands to be hosted into a specific pocket of MdfA and NorA. In a first analysis, we observed whether the crystallographic binding site was translatable into NorA (Figure S5). However, as we had no crystallographic information on NorA, we decided to explore further sites within the pump. Indeed, while for MdfA the binding site was defined by some of the twelve residues that showed interactions to the ligand with the crystallographic ligand (Tyr30, Asn33, Asp34 and 236), to identify the NorA putative binding site, the cluster centroid belonging to the most populated conformation was submitted to SiteMap tool [34] in Maestro suite. The highest-ranked binding site (SiteScore = 1.119) was selected as putative NorA binding site and in particular Ile23, Pro24, Pro27, Tyr225, Ser226, and Gly348 were set as binding site residues. Some of the selected binding site residues are in agreement with some of those residues chosen in previous studies [35]. This site was located more outwards than the CLM binding site. First, three different docking programs (i.e., Glide [36], PLANTS [37] and GOLD [38])

were explored with the aim to identify the best performing method in reproducing the crystallographic binding mode of CLM into MdfA (Supplementary Materials, Table S2). Glide turned out to be the best protocol in generating the correct CLM pose on the basis of the obtained RMSD and E_rvdw (i.e., van der Waals interaction energy) values calculated for each pose. Second, the same Glide protocol was applied to dock CPX against the experimental MdfA pocket, and both CLM and CPX against the hypothesized NorA binding site. The gained results suggested that the two compounds could potentially be hosted in the defined efflux pump binding sites.

2.6. Substrate Binding Simulations Using SuMD

2.6.1. General Overview of SuMD Analysis

As already anticipated, in this work the SuMD approach has been applied to MdfA and NorA proteins. Depending on the substrate (CLM or CPX) and on the protein (MdfA or NorA) used in the experiments, four complexes divided into two subsets (A and B) have been subjected to SuMD simulations (as summarized in Table 2). Using the binding site residues previously highlighted, different SuMD simulations were planned. The preliminary docking results suggested that CLM and CPX were potentially able to fit the cavity of the analyzed proteins (i.e., MdfA and NorA).

System	Replica	Outcome	Time (ns)	Best $d_{cm(L-R)}$ Å
		Subset A		
CLM-MdfA	1	productive	32	3.1
CLM-MdfA	2	productive	36	2.9
CLM-MdfA	3	non productive	13	23.4
CLM-MdfA	4	productive	37	3.1
CLM-NorA	1	productive	16	34.4
CLM-NorA	2	productive	58	0.4
CLM-NorA	3	productive	47	1.7
CLM-NorA	4	productive	14	30.7
		Subset B		
CPX-MdfA	1	productive	56	3.6
CPX-MdfA	2	non productive	23	16.3
CPX-MdfA	3	productive	44	28
CPX-MdfA	4	productive	47	2.9
CPX-NorA	1	non productive	16	25.5
CPX-NorA	2	non productive	19	26.3
CPX-NorA	3	non productive	26	26.3
CPX-NorA	4	productive	73	3.4

Table 2 *SuMD replicas results summary.*

SuMD replicas of the two studied systems (i.e., S. aureus NorA and E. coli MdfA) provided some interesting information about the molecular recognition mechanisms and the kinetic processes underlying the interaction between these efflux pumps and their substrates.

First, a self-recognition SuMD simulation of the CLM into MdfA was performed to validate the applicability of the in-silico technique. Indeed, this work represents the first example of SuMD applied to efflux pumps. In total, four replicas were performed, and in three of them, CML was able to reach the defined orthosteric site. We refer to these replicas as productive replicas. It is worth noting that in one of the three productive replicas, this approach was able to reproduce the crystallographic binding mode of CLM. Indeed, the RMSD between the experimental and the SuMD pose of CLM was 1.77 Å, underling that the used technique worked pretty well in identifying the correct CLM pose on MdfA, also considering that the crystallographic resolution is 2.4 Å.

The analysis of the SuMD results for CLM on NorA protein showed that a binding event was observed in two replicas. However, the in-silico results were not supported by the previously obtained biological assays, which showed that CLM was not a substrate of NorA efflux pump. However, it should be noted that the simulation data only indicated that CLM could be able to enter and reach the binding site (Figure S6). Thus, the fact that geometrically and energetically CLM could be hosted inside NorA did not mean that it had to be extruded at all. For instance, NorA binding by CLM could be compatible with the inhibitory activity of this compound, but unfortunately, no information is available in the literature about this topic to validate or not the hypothesis. The data obtained left the way open for this scenario.

Second, we focused our attention on the recognition of CPX on MdfA and NorA. Both for CPX on MdfA and NorA, the SuMD simulation needed a remarkable number of SuMD steps. In the case of CPX on MdfA to sample a binding event within the orthosteric pocket, we had to increase the number of tries that the system could do before reaching the binding site. This behavior could be explained, considering that during the path the ligand was able to reach a meta-binding state, which was characterized by lower energy compared to that of the final state (Figure S7).

When analyzing the NorA case study results, we observed this behavior again; indeed, in one productive replica, the final state reached by CPX at the defined site was energetically less stable than the ligand pose found at the meta-binding site identified during the path. Moreover, the channel of the pump is rich in charged amino acids. Overall, these observations once more suggested that the kinetic process that allows the substrate to reach the binding is hard and harsh.

2.6.2. SuMD Validation: MdfA-CLM Recognition Pathway

The CLM was at first positioned 62 Å far away from the MdfA experimental canonical binding site defined by four residues (Tyr30, Asn33, Asp34, and Leu236) (dcm (L-R) = 62 Å). The whole recognition pathway can also be appreciated in this case by browsing Movie S1. The centers of mass distance (dcm (L-R)) quickly decreased from the initial 62 Å to about 30 Å during the first 2 ns of the SuMD simulation, as shown in the Dynamic Total Interaction Energy plot (Figure 2d). At this point, CLM established the first contacts with the protein by the "electrostatic recruiters" Arg131 (TM4) and Lys346 (TM10) located at the protein entry. Subsequently, the ligand was stabilized between the two residual recruiters and its center of mass was located at about 20 Å away from the orthosteric site. This recognition mechanism was clearly evident in the Interaction Energy Landscape (Figure 2b) in which there was the first region of minimum; the energy dropped from −40 kcal/mol to −70 kcal/mol. The substrate remained in this position for about 10 ns. Arg131 turned to have a key role in the molecular recognition mechanism, contributing to the binding events with cumulative energy of around −10,000 kcal/mol (Figure 2c). Later, the interaction between CLM and the two mentioned residues stopped and the ligand moved again along the trajectory pathway to penetrate the transporter. The protein region involved in this prolonged interaction could be defined as a meta-binding binding site, as revealed by the stability of MMGBSA energy values (see movie S1). A meta-binding site is a sort of stopover with enough residence time, which breaks the progressive and continual approach of the ligand. At this point, CLM orientation changed and reached a deeper position inside the canonical binding site, through a horizontal placement, where it makes contacts first with Tyr30 at 16 ns and then with Asp34. Noteworthy, this latter residue showed strong participated in the stabilization of CLM into the canonical binding site by interacting with the substrate OH groups. During the SuMD simulation, CLM was able to reach the orthosteric site in a conformation very close to the crystallographic one, as reported in Figure 2a, where the RMSD reached a minimum value of 1.77 Å at 17 ns. The geometric reproduction of the binding mode can also be observed from Movie S2. The predominant energy role of the amino acids mentioned above can be better understood by looking at the graph of the Total Interactions Energy (Figure 2c). Indeed, the cumulative interaction energy between residue Asp34 and the two ligand oxygens reached the value of −30,000 kcal/mol. Although this residue established contacts with the ligand until the end of the simulation at 37 ns, the substrate changed its binding mode during the interaction time.

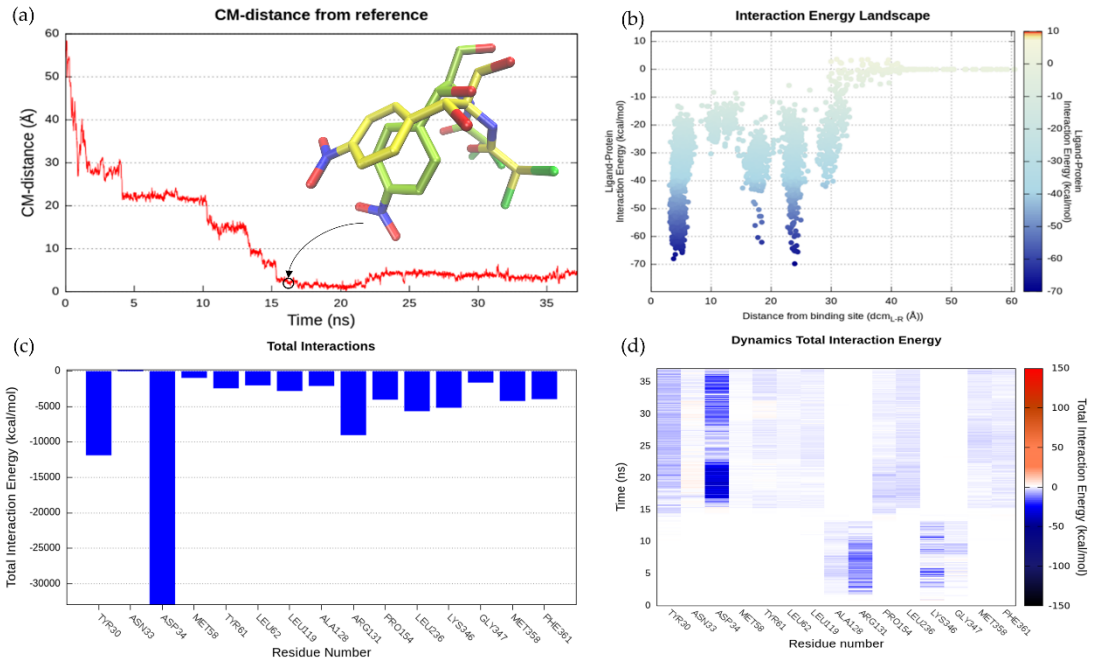


Figure 2 *SuMD MdfA-CLM recognition pathway analysis. (a) CM-distance between the ligand and the reference binding site calculated as RMSD of simulated position (light green) against the experimental (i.e., crystallographic) one (yellow). (b) Interaction Energy Landscape. (c) Total Interaction energy plot. (d) Dynamics Total Interaction Energy for each ligand-interacting residue.*

To identify the possible CLM recognition sites during the SuMD trajectory, we performed a clustering analysis using DBSCAN (Figure S8). DBSCAN algorithm enables to identify clusters of ligand conformations during the SuMD trajectory, highlighting which regions were most explored by the ligand. Each sphere represents a population of ligand clustered conformations and the sphere radius in relation to the cluster population is set. According to what can be deduced from the analysis carried out, the CLM seemed to have a fairly immediate recognition pathway. Indeed, the clustering analysis identified two main steps characterizing the recognition process: the electrostatic recruitment by the vestibular region residues (Arg131 and Lys346), followed by a rapid transition to the crystallographic binding site, where the higher conformation cluster was identified and retained until the end of the simulation.

2.6.3. NorA-CPX Recognition Pathway

In the starting geometry, the ligand was placed at a distance of 84 Å far away from the postulated canonical binding site. As depicted in Figure 3d and shown in Movie S3, the first interaction between the ligand and the protein occurred after 3 ns of productive trajectory, involving the Lys127 side chain. The distance between the ligand and protein centers of mass then rapidly decreased from 84 Å to about 40 Å (Figure 3a). This region of first recognition was very rich in positively and negatively

charged residues that slowed down the entry of the ligand such as Lys127 Lys264, Glu385, and Asn319 (Figures 3 and 4). This behavior was expected, considering the CPX zwitterionic nature. Indeed, the compound was almost always stabilized in the pump vestibular region by the Lys127 side chain that had strong interactions with the carboxylate group of CPX. As Figure 3a shows, the ligand persisted in this first recognition site until 13 ns. The residence time of the ligand in this region was also supported by the energy interaction of the ligand-protein complex, which reached a value of -300 kcal/mol when the distance between the two centers of mass was between 30 Å and 40 Å (Figure 3b). Therefore, this region was considered a meta-binding site, a key region for the passage of the ligand inside the protein. Subsequently, CPX shifted deeper into the protein by losing the interaction with Lys127, but maintaining the interaction with Asn319. After about 15 ns, the carbonyl group of CPX established again an H-bond with Lys127, whereas the carboxylic group acquired interaction with Tyr131. This binding mode was also stabilized by Ser318. In this second site, the CPX binding mode changed. Indeed, while previously the protonated amine group was located towards the cytoplasmic side, it was now oriented towards the inner periplasmic side of the protein. The ligand was here stabilized by Ser318 and the hydrophobic component had a role in the orientation exploited by Met109. This was another site explored by the ligand, although at a low energy level of −150 kcal/mol (Figure 3b). The arene-H interaction with Thr314 also contributed to the CPX orientation, and this contact was retained until about 30 ns when the distance between the two mass centers was 20 Å. At 30 ns, the ligand again changed its conformation, establishing interaction with Gln51 at the level of the protonated amine. This binding pose was preserved up to about 36 ns, after which CPX began to interact with Arg310. The substrate carboxylic group was engaged in contacts with Arg310, Ser133, and Asn137, while the protonated piperidine nitrogen interacted with Gln51. As can be observed by the IE landscape (Figure 3b) this ligand conformation occurred at about 18 Å of distance from the binding site and was characterized by the energy of −150 kcal/mol. This kind of interactions was retained until 38 ns. Then Gln51 interacted with the carboxylic group of CPX. At about 10 Å from the orthosteric pocket, its orientation was strongly stabilized by Arg310 and Glu222 until 49 ns. The relevance of these two residues was also supported by the histogram of Total Interactions Energy (Figure 3c) and by the Dynamics Total Interaction Energy (Figure 4d). Indeed, Glu222 and Arg310 had total interaction energies of −100,000 kcal/mol and −150,000 kcal/mol, respectively. In addition, this conformation was stabilized by the π-π stacking interaction with Phe140. Finally, CPX shifted to the orthosteric binding site losing the interaction with Arg310 at around 52 ns. The CPX established π-π stacking interaction with Phe303.

This orientation was stabilized by Arg310 interaction and the arene-H interaction between the cyclopropyl and the aromatic moiety of the Tyr225. The minimum value of distance observed was 3.6 Å and this conformation persists until the end of the SuMD simulation at 73 ns.

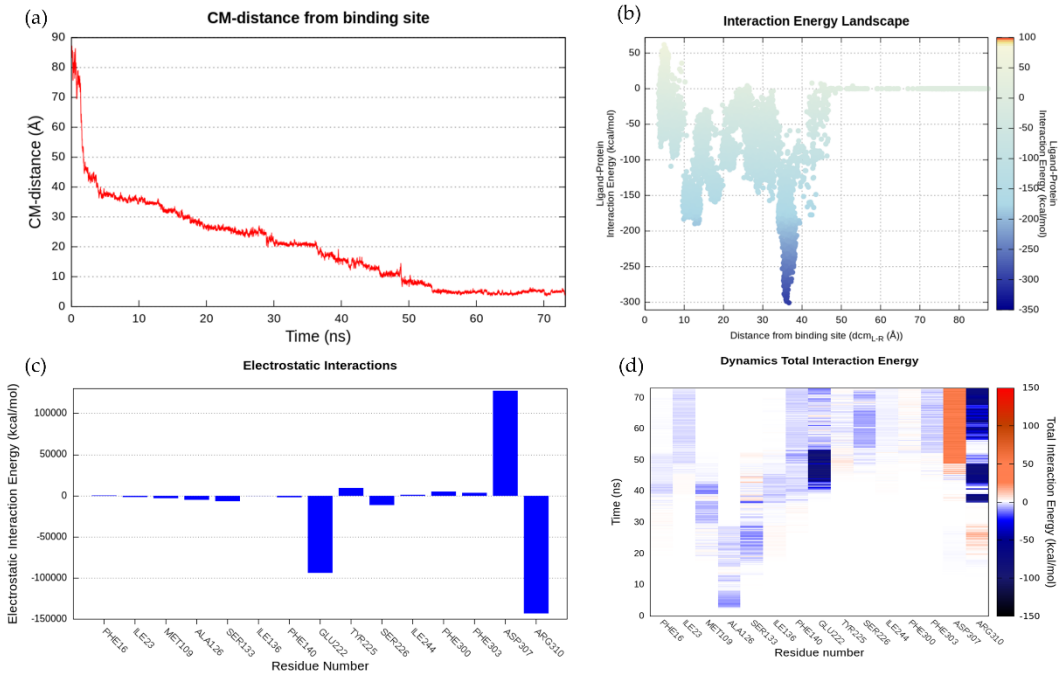


Figure 3 *SuMD NorA-CPX recognition pathway analysis. (a) CM-distance between the ligand and the binding site. (b) Interaction Energy Landscape. (c) Total Interaction energy plot. (d) Dynamics Total Interaction Energy for each ligand-interacting residue.*

To reveal the most crucial binding sites, a clustering analysis was performed (Movie S4). As Figure 3a shows, two meta-binding sites were identified at about 35 Å distance. As we previously highlighted, several charge residues hosted in this site. Subsequently, the ligand shifted at a distance of about 27 Å from the center of mass of the protein binding site. At this level, we found a third populated site formed by 940 conformations, where the ligand was stationed for a fairly long time (Figure 4a). CPX presented a conformation with the carboxylic group faced towards the cytoplasmic side, while the protonated amine group was directed towards the pump channel all the time. This cluster of conformations was stabilized by Ser318 whose side chain was hydrogen bonded to the CPX carboxylate, and by Met109, Thr314, Ile136, Ser133, Arg310, Phe129, Ala126, which contributed with the van der Waals component (Figure 4b). A further cluster of 653 conformations was found at 18 Å from the binding site cavity, as shown in Figure 4c. Here, the CPX binding mode is characterized by the interaction with Gln51. A small relatively sparsely populated cluster (233 conformations) was identified immediately above the previous one, where this time CPX made polar contact with Gln51 at the level of its carboxyl group. The next cluster identified was that represented in Figure 4d populated by 348 conformations. In this pose, the substrate was firmly stabilized by the

two charged residues, Arg310 and Glu222. The CPX final state, identified by the fourth group of conformations, was broadly explored and widely populated (1980 conformations) (Figure 4d).

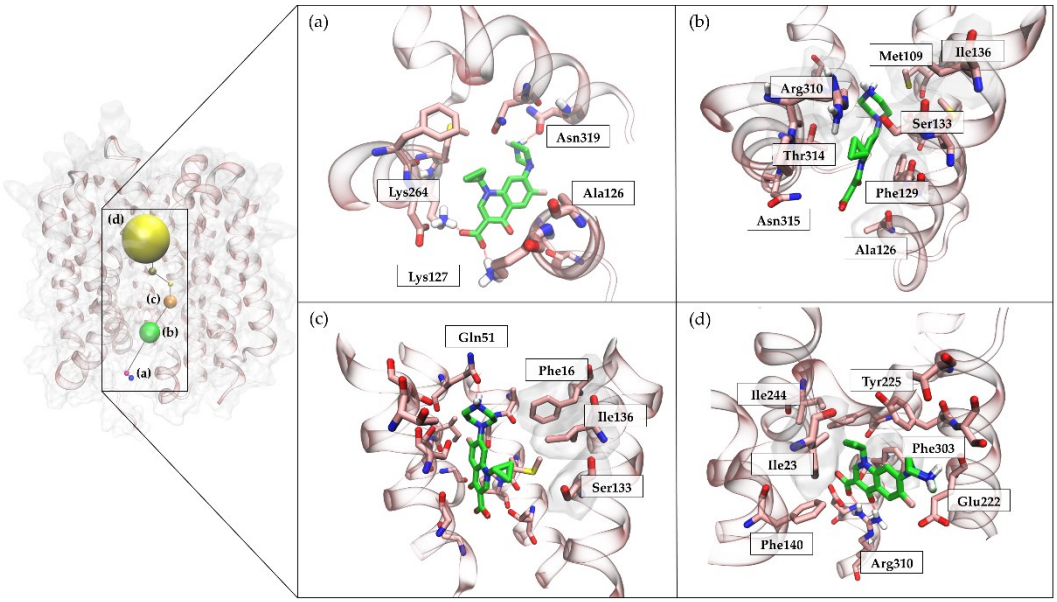


Figure 4 *Clustering analysis of CPX recognition pathway during a SuMD trajectory. (a) CPX binding mode in the first recognition site. The ligand establishes interactions with Ala126, Lys127, Lys264 and Asn319. (b) Panel b shows the interaction between CPX and NorA protein during its trajectory. CPX interacts with Met109, Ala126, Phe129, Ser133, Ile136, Arg310, Thr314, Asn315. (c) In cluster c, the ligand interacts with Phe16, Gln51; a hydrophobic contribute comes from Ser133 and Ile136 residues. (d) CPX is hosted in the orthosteric binding site. This is also the most populated cluster. CPX mostly establish contacts with Ile23, Phe140, Glu222, Tyr225, Ile244, Phe 303, Arg310.*

3. Materials and Methods

3.1. General

All simulations were performed on a hybrid CPU/GPU cluster. MD and SuMD simulations were carried out with the ACEMD [39] engine on a GPU cluster provided of 18 NVIDIA graphics cards, whose models include GTX 780 to Titan V. Before running MD and SuMD simulations, the following preliminary phases were carried out: (i) protein modeling, (ii) protein-ligand system preparation, (iii) ligand parametrization, and (iv) solvated system setup and equilibration. The protocol based on the CHARMM36/CHARMM general force field (CGenFF) force fields combinations was adopted for transmembrane systems.

3.2. Protein Modeling: Preparation of the NorA Target

Quinolone resistance protein NorA amino acid sequence was downloaded in the FASTA format from the UniProtKB database (Uniprot: P0A0J4) [40] and submitted to the different software employed for the 3-D protein structure prediction. Towards this aim, we used I-TASSER [23], SWISS-MODEL [25], RaptorX web server [41] and Phyre2 server [27]. The quality of the NorA 3-D structure models was assessed analyzing the Ramachandran plot generated by MOE suite and QMEANBrane [42]. Model 2 was then refined with MOE Geometry tool. The refined structure was aligned and superimposed on the MdfA crystal structure in the Orientations of Proteins in Membranes (OPM) database [43].

3.3. MdfA Crystal Structure Preparation

Protein-ligand complex of E.coli was retrieved from the RCSB PDB database (PDB: 4ZOW) [21]. The protein structure to be used as template was prepared with the protein preparation tool as implemented in MOE [29]: hydrogen atoms were added to the complex, and appropriate ionization states were assigned by means of the Protonate-3D tool. Missing atoms in protein side chains were built according to the CHARMM36 force field topology. Missing loops were modeled by the default homology modeling protocol implemented in the MOE protein preparation tool. Non-natural N-terminal and C-terminal domains were capped to mimic the previous residue.

3.4. Ligand Preparation

The investigate substrates CLM and CPX are small organic molecules. The substrates were designed using MOE software, after which the partial charges were assigned, followed by a minimization step using the MMFF94 force field. The ligands parameters were achieved from the Paramchem service [44] (CGenFF). Using these initial parameters, we subjected each ligand to 150 ns of preliminary MD simulation. Since the ligands' behavior observed during the simulation was consistent, we decided to use these parameters for SuMD simulation.

3.5. Molecular Docking Experiments

The molecular docking experiments were performed using three different docking protocols: PLANTS [37], GLIDE [36] and GOLD [38]. Starting from the crystal structure, the grid was centered on the center of mass of the co-crystallized ligand (CLM). The grid center was -17.2275, 13.7332 and 24.7736 to x, y and z-axis for all the used protocols. The docking space was defined as a cubic box (22 Å side), with a nested cubic box (10 Å) defining the region where the centroid of the ligand had to be located using Glide. In GOLD and PLANTS protocols, the grid is a sphere with a radius set at 12

Å. Docking on NorA was performed using GLIDE as the best protocol selected. The grid coordinates for NorA model were -28.845, 15.2, 29.82 to x, y and z-axis. Each docking protocol generated 20 poses per ligand. The RMSD and the E_rvdw were calculated with MOE tool.

3.6. Solvated System Setup and Equilibration

Four systems composed by the combination of the two analyzed proteins (MdfA and NorA) and the two designed substrates (CLM and CPX) were then prepared. Then, the position of the ligands was manually assigned. To avoid protein-ligand long-range interactions in the starting geometry, CLM and CPX was positioned 62 Å away from the MdfA transporter atom and 84 Å away from the NorA efflux pump atoms, respectively. Transmembrane proteins were embedded in a POPC lipid bilayer, according to the suggested orientation reported in the OPM database. Initial POPC atoms were placed through the VMD membrane builder plugin [45], and lipids within 0.6 Å from amino acid atoms were removed. The membrane used in all the simulations has a dimension of 120Å x 120Å. The systems were solvated with TIP3P water using the program Solvate 1.0 [46] and neutralized by Na^+/Cl^-counterions to a final concentration of 0.154 M. The systems were then equilibrated through three main steps of molecular dynamics to equilibrate them. In the first stage, after 1500 steps of minimization to allow the system to reduce the clashes between proteins and lipids, 5 ns of MD simulation (2,500,000 steps) were performed in the NPT ensemble, restraining ligand, protein atoms and phosphorousof phospholipid by a positional constraint of 1 kcal mol^{-1} $Å^{-2}$. The temperature was maintained at 310 K using a Langevin thermostat with low damping constant of 1 ps^{-1}. The pressure was maintained at 1 atm using a Berendsen barostat; bond lengths involving hydrogen atoms were constrained using the M-SHAKE algorithm with an integration time step of 2 fs. In the second stage, applying the restraints only to the protein and to the ligand and keeping the conditions of constant pressure and temperature (NPT), the temperature was set at 310 K and the pressure at 1 atm, and 10 ns of MD were performed. Then, the last equilibration step included 20 ns of MD simulation and the only restraints left were on the α carbon of amino acids and on the ligand. The stability of the cell volume and POPC area per lipid headgroup during the simulation were evaluated using a script that relies on VMD and GridMAT-MD, a tool for calculating bilayer parameters (Figure S9) [47]. In according with GridMAT-MD values, the area per lipid headgroup ranged from 63 to 70.

3.7. Molecular Dynamics (MD) Simulations.

MD simulations of 500 ns of the systems (MdfA and NorA Cin, both without substrates and MdfA in complex with CLM) were performed using ACEMD engine with a time step of 2 fs. The MD trajectory was stridden at 5000 frames. The protein RMSD and RMSF were computed on the protein C_α using VMD trajectory tool. The MD conformations were then clustered using the density-based clustering DBSCAN, setting the RMSD threshold to 2 and the minimum number of protein conformations that could generate a cluster to 30. The cluster centroid was selected using a script based on Numpy [48] and MDTraj python library [49].

3.8. Supervised Molecular Dynamics (SuMD).

Each SuMD simulation is composed of a number of consecutive short unbiased MD simulations (600 ps, editable by the user) in which a supervision strategy, based on a tabu search-like strategy, is applied at the end of each simulation. The supervised variable is the distance between the ligand and protein binding site center of mass it is maintained until the protein-ligand distance reaches a preset threshold value (5 Å in this case study). Then the simulation proceeds as a conventional unbiased MD simulation. For a more detailed description of the SuMD analyzer, Salmaso et al. provide all the necessary information [50]. Four simulations were carried out for each system, starting from the same initial geometry is based on the subset. The more significant replicas are described in the results and discussion section.

3.9. Analysis of pepSuMD Trajectories

All the trajectories generated by pepSuMD [50] were analyzed by an in-house script written in tcl and python, that makes use of Numpy [48] and ProDy modules [51]. The analyses were then performed on the whole trajectories. In brief, the single SuMD step trajectories were stridden, by a user-defined value (here 10), superposed on the first frame C_α carbon atoms of the target protein, wrapped and merged. The in-house script computes several sides of the SuMD simulation performed. It analyzes the geometry, such as the distance between the ligand and the binding site center of mass and the protein RMSD, the ligand-target interaction energy estimation during the recognition process plotted on the Interaction Energy Landscape plots. This analysis also calculates all the established interactions between the protein and the ligand. The clustering analysis was performed using the density-based clustering algorithm DBSCAN, setting the RMSD threshold to 1.75 Å and the minimum number of protein conformations that could generate the cluster to 200 for MdfA-CLM and NorA-CPX system. Representations of the molecular structures were prepared with VMD [45].

3.10. Microbiological Assays

The strains of S. aureus employed were SA-1199 (wt) and SA-1199B (overexpressing norA and also possessing an A116E Grla mutation). The MIC of the CLM was determined by microdilution technique according to CLSI guidelines [52].

4. Conclusions

This work investigated at the molecular level the substrate recognition pathway of NorA through a Supervised Molecular Dynamics (SuMD) approach, using NorA homology models. In this work, different NorA homology models' structural quality assessment and validation was carried out. These analyses allowed the selection of a NorA model built based on the MSF E. coli MdfA transporter, showing an inward conformation with an opening toward the cytoplasmic side as the best starting point for further studies. Notably, the antibiotic CPX is a substrate of both NorA and MdfA efflux pumps, while CLM is a specific substrate of MdfA, as confirmed by our biological experiments.

With this information in hand, a series of SuMD simulations were planned in an attempt to investigate the molecular basis of NorA substrate recognition. To test the ability of the chosen technique in studying these protein systems (i.e., efflux pumps), CLM was used as internal control given that a co-crystal structure between this substrate and MdfA was available.

The obtained results on the MdfA-CLM system supported the choice of the SuMD methodology to study the substrate recognition by efflux pumps. Indeed, CLM was able to reach the orthosteric site in a very close orientation (RMSD of 1.77 Å) with respect to the crystallographic position.

Interesting results were also obtained from the NorA-CPX SuMD simulations. In one of the five replicas, CPX was able to reach the orthosteric site. Additionally, in three replicas, CPX explored a meta-binding state where the strong electrostatic interaction seemed to be critical. This meta-binding site, placed at the interface between the protein and the cytoplasm, could work as the first recognition site for CPX, which was then oriented and released into the protein cavity. During its trajectory, CPX explored several recognition sites, establishing interaction with Lys127, Lys264, Met109, Ser133, Ile136 Gln51, Arg310, Glu222, Phe303, and Tyr225.

To correctly interpret the obtained results, we have to keep in mind that the efflux pumps (e.g., MdfA and NorA) are promiscuous proteins. They are involved in the extrusion of structurally unrelated chemical compounds. Thus, different pathways of substrates recognition can be used on the basis of the specific substrate chemical structure. The internalization of a molecule into the

efflux pump cavity cannot ensure the activation of the protein conformational change required to have the substrate extrusion. Thus, a binding event cannot always correspond to an extrusion event. In this context, the present work provides a solid homology structural model and an accurate technique (i.e., SuMD) that could rationally aid the comprehension of both the molecular mechanisms of action and inhibition of NorA efflux pump.

References

1. WHO. ANTIMICROBIAL RESISTANCE Global Report on Surveillance. (2014).

2. Blair, J. M. A., Webber, M. A., Baylay, A. J., Ogbolu, D. O. & Piddock, L. J. V. Molecular mechanisms of antibiotic resistance. Nature Reviews Microbiology vol. 13 42–51 (2015).

3. Bagnoli, F., Rappuoli, R. & Grandi, G. Staphylococcus aureus: Microbiology, Pathology, Immunology, Therapy and Prophylaxis. (2017). doi:10.1007/978-3-319-72063-0.

4. Costa, S. S., Viveiros, M., Amaral, L. & Couto, I. Multidrug Efflux Pumps in Staphylococcus aureus: an Update. Open Microbiol. J. 7, 59–71 (2013).

5. Ubukata, K., Itoh-Yamashita, N. & Konno, M. Cloning and expression of the norA gene for fluoroquinolone resistance in Staphylococcus aureus. Antimicrob. Agents Chemother. 33, 1535–1539 (1989).

6. Pao, S. S., Paulsen, I. A. N. T. & Saier, M. H. Amr000001. Microbiol. Mol. Biol. Rev. 62, 1–25 (2009).

7. Yoshida, H., Bogaki, M., Nakamura, S., Ubukata, K. & Konno, M. Nucleotide sequence and characterization of the Staphylococcus aureus norA gene, which confers resistance to quinolones. J. Bacteriol. 172, 6942–6949 (1990).

8. Neyfakh, A. A., Borsch, C. M. & Kaatz, G. W. Fluoroquinolone resistance protein NorA of Staphylococcus aureus is a multidrug efflux transporter. Antimicrobial Agents and Chemotherapy vol. 37 128–129 (1993).

9. Felicetti, T. et al. 2-Phenylquinoline S. aureus NorA Efflux Pump Inhibitors: Evaluation of the Importance of Methoxy Group Introduction. J. Med. Chem. 61, 7827–7848 (2018).

10. Sabatini, S. et al. Investigation on the effect of known potent: S. aureus NorA efflux pump inhibitors on the staphylococcal biofilm formation. RSC Adv. 7, 37007–37014 (2017).

11. Holler, J. G., Slotved, H.-C., Mølgaard, P., Olsen, C. E. & Christensen, S. B. Chalcone inhibitors of the NorA efflux pump in Staphylococcus aureus whole cells and enriched everted membrane vesicles. Bioorg. Med. Chem. 20, 4514–4521 (2012).

12. Fontaine, F. et al. First identification of boronic species as novel potential inhibitors of the Staphylococcus aureus NorA efflux pump. J. Med. Chem. 57, 2536–2548 (2014).

13. Schmitz, F. J. et al. The effect of reserpine, an inhibitor of multidrug efflux pumps, on the in-vitro activities of ciprofloxacin, sparfloxacin and moxifloxacin against clinical isolates of Staphylococcus aureus. J. Antimicrob. Chemother. 42, 807–810 (1998).

14. Sabatini, S. et al. Evolution from a Natural Flavones Nucleus to Obtain 2-(4-Propoxyphenyl)quinoline Derivatives As Potent Inhibitors of the S. aureus NorA Efflux Pump. J. Med. Chem. 54, 5722–5736 (2011).

15. Sabatini, S., Kaatz, G. W., Rossolini, G. M., Brandini, D. & Fravolini, A. From phenothiazine to 3-phenyl-1,4-benzothiazine derivatives as inhibitors of the Staphylococcus aureus NorA multidrug efflux pump. J. Med. Chem. 51, 4321–4330 (2008).

16. Brincat, J. P. et al. Discovery of novel inhibitors of the NorA multidrug transporter of staphylococcus aureus. J. Med. Chem. 54, 354–365 (2011).

17. Yin, Y., He, X., Szewczyk, P., Nguyen, T. & Chang, G. Structure of the multidrug transporter EmrD from Escherichia coli. Science (80-.). 312, 741–744 (2006).

18. Kang, X. et al. Structure of the YajR transporter suggests a transport mechanism based on the conserved motif A. Proc. Natl. Acad. Sci. 110, 14664–14669 (2013).

19. Zhao, Y. Y. et al. Substrate-bound structure of the E. coli multidrug resistance transporter MdfA. Cell Res. 25, 1060–1073 (2015).

20. Nagarathinam, K. et al. Outward open conformation of a Major Facilitator Superfamily multidrug/H+ antiporter provides insights into switching mechanism. Nat. Commun. 9, (2018).

21. Berman, H. M. The Protein Data Bank http://www.rcsb.org/pdb/. Nucleic Acids Res. 28, 235–242 (2000).

22. 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.

23. Zhang, Y. I-TASSER server for protein 3D structure prediction. BMC Bioinformatics 9, (2008).

24. Roy, A. et al. The I-TASSER Suite: protein structure and function prediction. Nat. Methods 12, 7–8 (2014).

25. Bienert, S. et al. SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res. 46, W296–W303 (2018).

26. Peng, J. & Xu, J. Raptorx: Exploiting structure information for protein alignment by statistical inference. Proteins Struct. Funct. Bioinforma. 79, 161–171 (2011).

27. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. E. The Phyre2 web portal for protein modeling, prediction and analysis. Nat. Protoc. 10, 845–58 (2015).

28. ISHITANI, R., DOKI, S., KATO, H. E. & NUREKI, O. Structural Basis for Dynamic Mechanism of Proton-Coupled Symport by the Peptide Transporter POT. Seibutsu Butsuri 54, 085–090 (2014).

29. Chemical Computing Group (CCG) | Computer-Aided Molecular Design.

30. Kaatz, G. W. & Seo, S. M. Mechanisms of fluoroquinolone resistance in genetically related strains of Staphylococcus aureus. Antimicrob. Agents Chemother. 41, 2733–2737 (1997).

31. Singh, S. et al. Boeravinone B, A novel dual inhibitor of nora bacterial efflux pump of Staphylococcus aureus and Human P-Glycoprotein, reduces the biofilm formation and intracellular invasion of bacteria. Front. Microbiol. 8, (2017).

32. 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).

33. Edgar, R. & Bibi, E. MdfA, an Escherichia coli Multidrug Resistance Protein with an Extraordinarily Broad Spectrum of Drug Recognition. JOURNAL OF BACTERIOLOGY vol. 179 (1997).

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

35. Kalia, N. P. et al. Capsaicin, a novel inhibitor of the NorA efflux pump, reduces the intracellular invasion of Staphylococcus aureus. J. Antimicrob. Chemother. 67, 2401–2408 (2012).

36. 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).

37. 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.

38. 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.

39. 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).

40. Bateman, A. et al. UniProt: A hub for protein information. Nucleic Acids Res. 43, D204–D212 (2015).

41. Wang, H. et al. Template-based protein structure modeling using the RaptorX web server. Nat. Protoc. 7, 1511–1522 (2012).

42. Studer, G., Biasini, M. & Schwede, T. Assessing the local structural quality of transmembrane protein models using statistical potentials (QMEANBrane). Bioinformatics 30, 505–511 (2014).

43. Lomize, M. A., Lomize, A. L., Pogozheva, I. D. & Mosberg, H. I. OPM: Orientations of proteins in membranes database. Bioinformatics 22, 623–625 (2006).

44. 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. (2010) doi:10.1002/jcc.21367.

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

46. Grubmüller, H., Groll, V. Max Planck Institute for Biophysical Chemistry. Solvate. (2016).

47. Allen, W. J., Lemkul, J. A. & Bevan, D. R. GridMAT-MD: A grid-based membrane analysis tool for use with molecular dynamics. J. Comput. Chem. (2009) doi:10.1002/jcc.21172.

48. van der Walt, S., Colbert, S. C. & Varoquaux, G. The NumPy Array: A Structure for Efficient Numerical Computation. Comput. Sci. Eng. 13, 22–30 (2011).

49. McGibbon, R. T. et al. MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. Biophys. J. 109, 1528–32 (2015).

50. 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).

51. Bakan, A., Meireles, L. M. & Bahar, I. ProDy: Protein dynamics inferred from theory and experiments. Bioinformatics 27, 1575–1577 (2011).

52. CLSI. National Committee for Clinical Laboratory Standards. 2000. Performance standard for antimicrobial susceptibility testing. Document M100–S10. Natl. Comm. Clin. Lab. Stand. Wayne, Pa, USA. (2017).

Scaffold repurposing of in-house chemical library toward the identification of new Casein kinase 1 δ inhibitors

*Eleonora Cescon, **Giovanni Bolcato**, Stephanie Federico, Maicol Bissaro, Alice Valentini, Maria Grazia Ferlin, Gianpiero Spalluto, Mattia Sturlese, Stefano Mor*

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**, 1168–1174 (2020).

Abstract

Recent studies have highlighted the key role of Casein kinase 1 δ (CK1δ) in the development of several neurodegenerative pathologies, such as Alzheimer's disease (AD), Parkinson's disease (PD) and Amyotrophic Lateral Sclerosis (ALS). So far, CK1δ inhibitors are non-covalent ATP competitive ligands and no drugs are currently available for this molecular target: hence the interest in developing new CK1δ inhibitors. The study aims to identify new inhibitors able to bind the enzyme, by a dual approach in silico/in vitro, the virtual screening has been performed on an in-house chemical library, which was previously designed and synthesized for other targets, the work can, therefore, be seen in the scaffold repurposing logic. The proposed strategy has led to the identification of two hits, having a novel scaffold in the landscape of CK1δ's inhibitors and with an activity in the micromolar range.

1. Introduction

The development of novel therapeutic approaches for the treatment of neurodegenerative diseases is still a great challenge. The discovery of the CK1 isoforms involvement in the development of neurodegenerative disorders has paved the way for the development of CK1 inhibitors. In particular, the physiopathological role of CK1 isoform δ in neurodegenerative diseases like Alzheimer's disease (AD), Parkinson's disease (PD) and Amyotrophic Lateral Sclerosis (ALS) has encouraged the research for innovative therapeutic approaches.

The protein kinase CK1 isoform δ is encoded by the gene CSNK1D which is located on chromosome 17 (chromosomal localization 17q25). CK1δ human gene was characterized as a sequence of 1245 nucleotides which is transcribed into a 49 kDa protein consisting of 415 amino acids. The poor substrate specificity of CK1 family members is supported by the fact that nearly 140 substrates are reported in the literature.[1] CK1δ is an acidotropic protein kinase, which means it recognizes

substrates containing acidic or phosphorylated amino acid residues. The canonical consensus sequence for CK1 is:(P)S/T-X-X-S/T. Where (P)S/T indicates a phosphorylated serine or threonine residues. Nevertheless, CK1 can also phosphorylate the target if there is an N-terminal cluster of acidic residues or acidic amino acids in the N-3 position. This allows CK1 to play the role of priming kinase activating the substrate for other kinases. Also, non-canonical sequences are recognized by CK1 such as the SLS motif.[2]

CK1 family members have several effectors able to modulate their expression and activity. X-ray studies demonstrate that the formation of homodimers could have a negative regulatory effect on CK1δ activity.[3,4] Moreover, post-translational modifications as phosphorylation are involved in the regulation of CK1 activity. Ser318, Thr323, Ser328, Thr329, Ser331, and Thr337 are the main residues subjected to autophosphorylation. In addition to autophosphorylation, CK1δ is phosphorylated by other kinases including PKA, Akt, CLK2 (CDC-like kinase), PKC isoform α and Chk1.[2, 5, 6] Several studies have also underlined the importance of compartmentalization and subcellular localization in CK1 activity regulation. The subcellular localization of the kinases is mostly regulated by binding to intracellular structures or protein complexes.[7, 8] Dysregulations in expression or activity of CK1δ have been observed in cancer as well as in different neurodegenerative disorders like AD, PD, and ALS.

CK1δ appears to be involved in different stages of AD development. The residues Ser202/Thr205 and Ser396/Thr404 have been identified as CK1δ phosphorylation sites on the Tau protein.[9, 10] Furthermore, CK1 family is overexpressed in Alzheimer's disease and CK1 isoforms colocalize with granulovacuolar degeneration bodies in AD hippocampus.[11] As concerns PD, it has been demonstrated that CK1 isoforms constitutively phosphorylate α–synuclein at Ser129. This suggests that CK1 mediated phosphorylation of the protein can play a key role in PD development.9 Moreover, recent studies have demonstrated that CK1δ phosphorylates many different sites of TAR (TransActive Response) DNA-binding Protein 43 (TDP-43) in vitro.[12] TDP-43 was identified as the major component of ALS protein aggregates and it is responsible for the onset and progression of ALS. As a consequence, the identification of potent and selective inhibitors of CK1δ may provide an innovative therapeutic strategy for ALS.[13]

Initiating hit identification campaigns by using chemical scaffolds from an in-house library designed for other indications (scaffold repurposing) can speed up drug discovery in several therapeutic areas.[13-14] Additionally, in silico approaches for the discovery of new kinase ligands is now mainly

structure-driven, with the determination of the X-ray of several hundred structures of kinase-ligand complexes. Structure comparisons have been widely used to identify the most common and stabilizing interaction networks between ligands and their corresponding kinase binding sites. Regarding specifically CK1δ, nowadays 19 unique protein–inhibitor complexes are available from the Protein Data Bank (PDB). In parallel, docking-based virtual screening (DBVS) has extensively and successfully used to identify potential hit compounds.[14]

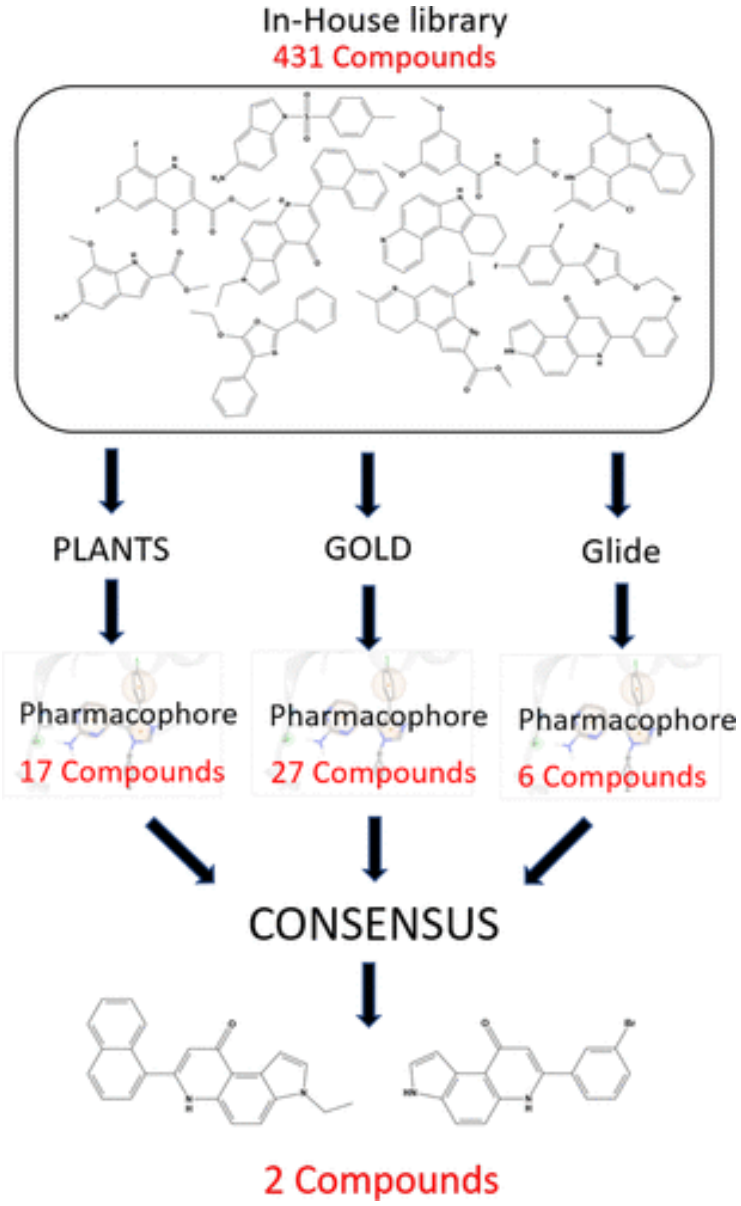


Figure 1 *Workflow for hit compounds identification*

Following this approach, in this work, we have performed a DBVS of an in-house chemical library composed by 431 compounds synthesized over more than thirty years of research in the field of oncology and directed to the inhibition of several molecular targets such as topoisomerase 1 and 2, aromatase, and tubulin. In particular, our computational pipeline was based on a combination of a

canonical DBVS followed by a pharmacophore-driven filtering process of all obtained docking poses, as summarized in Figure 1. The primary goal of this study is to verify if in our in-house library there were some ligands characterized by a scaffold not yet used among the already known inhibitors of CK1, and that was therefore susceptible to a later phase of optimization. After the preliminary in silico screening, the most promising candidates have been undergone to the in vitro tests to confirm whether they have shown a detectable inhibition of CK1δ activity. Interestingly, we have identified two hit compounds, that share the pyrrolo[3,2-f]quinolinone moiety as key-scaffold, that are able to inhibit CK1δ activity in the micro-molar range. This repurposed scaffold is now subject to further study for the construction of focused libraries for the necessary phase of optimization of its pharmacodynamic and pharmacokinetic properties.

2. Materials and Methods

2..1 Preparation of the Virtual database for the Docking Protocol calculation

The preparation of the in-house chemical library for the DBVS consisted in the enumeration of the tautomeric state and selection of the most stable one (when more than one tautomeric state is possible), the generation of the three-dimensional coordinates, the assignation of the correct ionization state for a given pH and the calculation of the atomic partial charges.

The Tautomers application, which is included in the OpenEye toolkit QUACPAC, enumerates the most reasona-ble tautomeric forms of the molecule. Subsequently, the FixpKa program (also included in the Openeye toolkit QUACPAC) can be used to assign the most probable mole-cule ionization state for pH 7.4. The 3D conformations were generated by Corina Classic. [15, 16] To determine the partial charges of each compound, the Molcharge applica-tion (also included in the Openeye toolkit QUACPAC) in accordance with the MMFF94 force field was used.

2.2 Selection of the best Protocols trough DockBench and Virtual Screening

All 19 Holo Crystal Structures of CK1δ were retrieved from the Protein Data Bank (PDB). These structures were prepared with MOE Structure Preparation tool.[17] If more than one chain is reported in the crystal data file, the best-solved chain was selected. The highest occupancy alternative for each residue with alternate locations was selected. The system was protonated with the Proto-nate3D tool (which assigns the most probable protonation state at selected pH) using the AMBER 10 force field. The partial charges of the system (protein and ligand) were calculated and the hydrogen atoms were minimized. The co-crystalized ligands were saved in a separate database for

the following analyses while the protein structures were saved after removing ions, solvent or other molecules used to obtain the crystal formation. This procedure is speeded up by the use of a platform called DockBench.[18] The software is based on a self-docking analysis. Briefly, each co-crystallized ligand is docked using the docking protocols and the ability of each protocol in reproducing the pose of the crystallographic complex is evaluated. For each structure-docking protocol pair, minimum (RMSDmin) and average RMSD (RMSDave) values with respect to the X-ray binding mode were calculated. Twenty poses for each molecule were generated and analyzed. The VS was performed using GOLD (Scoring Function: Goldscore), PLANTS (Scoring Function: chemplp), and Glide (standard precision). The results were evaluated using a consensus strategy.

2.3 Interaction Energy Fingerprint (IEF)

The per residue analysis was performed using the software MOE (Molecular Operating Environment)[17] and the SVL programming language. The electrostatic interaction energies were measured through the Coulombic function, and they were expressed in kcal/mol, while the hydrophobic contribution resulted from the contact surfaces analysis performed by MOE and are associated to a dimensionless score. To rationalize the binding mode of each compound, the interaction energy values can be translated into heat maps called Interaction Energy Fingerprint (IEF).

2.4 Generation of the Pharmacophore model

The conformation originated from docking were further filtered by a pharmacophore model. The alignment and the superimposition of CK1δ crystal structures have allowed a comparison between different ligands and the detection of common interaction features. The identification of the main features to build the pharmacophore model for CK1δ ligands has required a visual investigation of the protein-ligand crystallographic complexes in addition to information from the previous IEF analysis. The pharmacophoric query design and the consequential search was performed using the MOE pharmacophore modeling tools.[17]

2.5 CK1δ activity assay

Compounds were evaluated towards CK1δ (full length, ThermoFisher) with the KinaseGlo® luminescence assay (Promega) slightly modifying a procedure reported in literature.[13] In detail, luminescent assays were performed in black 96-well plates, using the following buffer: 50 mM HEPES (pH 7.5),1 mM EDTA, 1 mM EGTA, and 15 mM magnesium acetate. Compound PF-670462

(IC50 = 7.7 nM) was used as positive control for CK1δ[19] while DMSO/buffer solution was used as 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 (26 nM) of enzyme solution were added to the 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 1-2%. After 10 minutes of incubation at 30 °C the enzymatic reactions were stopped with 40 µL of KinaseGlo ® reagent (Promega). Luminescence signal (relative light unit, RLU) was recorded after 10 minutes at 30 °C using Tecan Infinite M100. For IC_{50} determination, ten different inhibitor concentrations ranging from 100 and 0.026 µM were used. IC_{50} values are reported as means ± standard errors of three independent experiments. Data were analyzed using GraphPad Prism software (version 8.0).

3. Results and Discussion

The first step of our work was the identification of a suit-able docking protocol on which to base the DBVS of our in-house library. To this purpose, we performed a benchmark of the 17 docking protocols applied to 19 lig-and- CK1δ complexes. This procedure was speeded up by the use of a platform for a self-docking comparison called DockBench.[18] The results of the DockBench Analysis are visualized through the use of heatmap plots. In each plot, the vertical axis shows the docking protocols while the horizontal axis represents the protein-ligand complexes. A color code, from blue to red, displays the RMSD value. The plot in figure 2 summarizes the minimum value of RMSD ($RMSD_{MIN}$) calculated for each docking protocol on each protein-ligand complex; blue spots represent low RMSD values while red ones indicate higher values. The average RMSD value ($RMSD_{AVE}$) of poses generated by each docking protocol for each protein-ligand complex was also considered (Figure 2, right plot) reporting a sim-ilar profile to $RMSD_{MIN}$. According to these metrics, the crystal structure selected for the subsequent molecular docking analyses was 3UZP since it has resulted in one of the protein structures for which molecular docking better reproduces the crystal structure pose with different protocols.

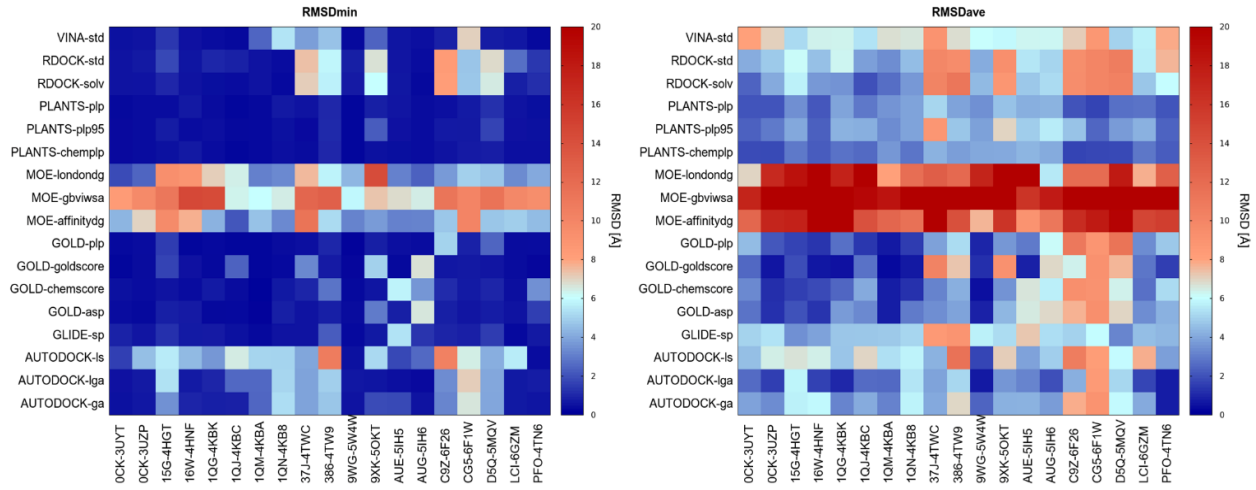


Figure 2 *Heatmaps summarizing the performances of molecular docking benchmark in the self-docking procedure. In panel A, the RMSD lower value obtained by each Docking Protocol (y-axis) for each Protein-ligand complex (x-axis). In panel B, the RMSD average value obtained by each Docking Protocol (y-axis) for each Protein-Ligand complex (x-axis).*

The comparison of the different docking protocols on the complex 3UZP revealed that several different algo-rithms were able to nicely reproduce the experimental geometries showing $RMSD_{MIN}$ below 0.55 Å (Table 1).

	$RMSD_{MIN}$	**$RMSD_{AVE}$**
GOLD - Goldscore	0.29 Å	0.54 Å
PLANTS - Chemplp	0.35 Å	1.73 Å
GLIDE - SP	0.55 Å	5.33 Å

Table 1

En-couraged by these performances, we decided to maximize the conformational sampling by using three different docking protocols in the Virtual Screening: GOLD[20] coupled to Goldscore Scoring Function, PLANTS[21,22] coupled to Chemplp Scoring Function[23] and Glide-sp[24]. This strategy, usually named consensus docking[25], is a method to improve the reliability of docking results, it consists in the parallel use of several docking protocols based on different search algorithms, and in the interpolation of the results of these. In this view, the selection of the protocols not only satisfies the benchmark results but also respects the fundamental requirement to have an orthogonal search algorithm. Indeed, PLANTS relies on an Ant Colony Optimization algo-rithm for the search algorithm, GOLD on a Genetic Algo-rithm and Glide on a systematic search. Ten poses for each molecule of the chemical library were hence calcu-lated generating a total of 12930 ligand conformations.

To analyze the VS output instead of using a classical scor-ing function we adopted a geometrical based method based on a structure-based pharmacophore developed on the same dataset of CK1δ

holo-complexes used in the previous benchmark. The alignment and the superimposi-tion of CK1δ crystal structures have allowed a compari-son between different ligands and the detection of com-mon interaction features to build the pharmacophore model. In addition, a qualitative analysis of the molecular interaction was carried out by considering the Interaction Energy Fingerprints (IEF) of the 19 ligands in our dataset (figure 3).

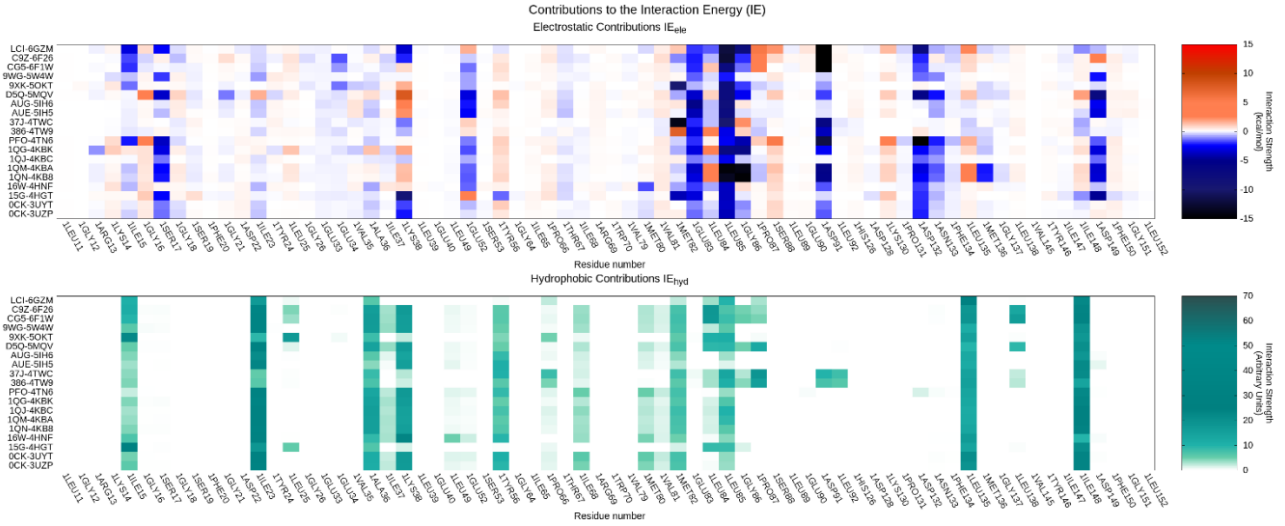


Figure 3 *Interaction Energy Fingerprint (IEF). Per residue Electrostatic (upper plot) and the hydrophobic contribution (lower plot) interaction for each crystallographic ligand (reported on the y-axis) of CK1δ. For Electrostatic interaction the colorimetric scale is blue to red while for the hydrophobic contribution it is white to green.*

By coupling the geometrical alignment and IEFs it was confirmed the relevance of interactions with the hinge region of the kinase. In particular, Leu 85 plays a key role in establishing two hydrogen bonds with most of the co-crystalized ligands. The hypothesis of the Leu 85 key role is strongly supported by studies reported in literature.[26–28] For this reason, the H-bond interaction with the backbone of this residue has been included in the pharmacophore model. In addition, the superimposition of the compounds revealed the presence of aromatic moieties for most structures; their presence guarantees a strong hydrophobic contribution as con-firmed in the hydrophobic fingerprint (figure 3). All these analyses were summarized in a pharmacophore having 5 features: two hydrogen bonds (one acceptor and one donor) and three hydrophobic ones (figure 4).

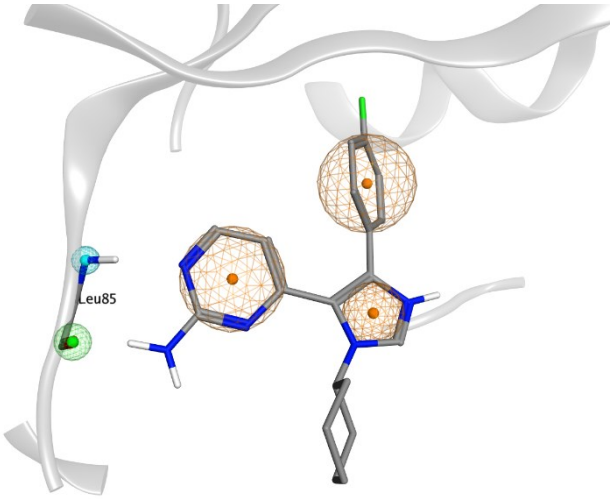


Figure 4 *The Pharmacophoric model superposed to the crystallographic complex ligand 0CK-CK1δ (PDB ID: 3UZP). The orange sphere represents an aromatic feature, while the blue and the green ones indicate respectively the presence of a Hydrogen Bond Donor (HBD) and a Hydrogen Bond Acceptor (HBA) acceptor mediating the interaction with Leu 85.*

To filter out the conformations obtained from the DBVS the following criterion was used: only the poses that sat-isfy at least three features of the pharmacophore model were retained including the mandatory presence of at least one donor/acceptor feature. The pharmacophoric filter was applied to each docking protocol separately in order to obtain three independent lists. Only the mole-cules that satisfy the pharmacophore model in each Dock-ing Protocol were retained. In this way, we were able to select two molecules: compound 1 and 2 (figure 5). For compound 2, the bromophenyl group fills the hydropho-bic pocket formed between the sidechains of Lys 38, Met 80 and the gatekeeper residue Met 82, while the pyrrolo-quinolinone scaffold occupies the outer portion of the binding site. The carbonyl oxygen of the pyrrolo-quinolinone portion maintains the recurrent interaction with the hinge region, especially with the backbone NH of Leu 85. In addition, π - CH interactions occur between the pyridone and pyrrole moieties and non-polar amino acids as Ile 15 and Ile 23.

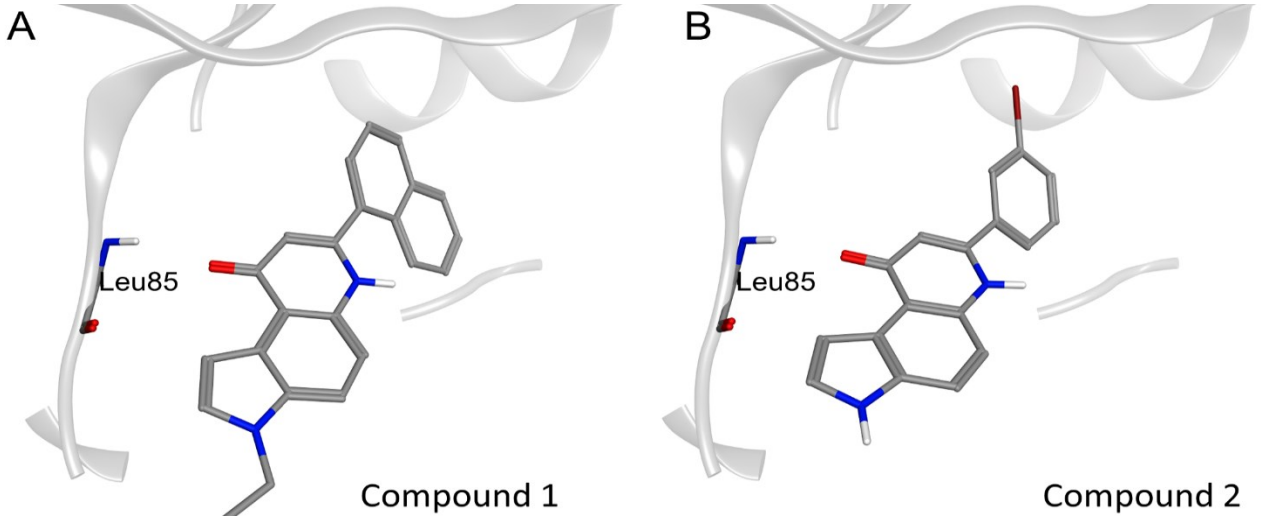


Figure 5 *The resulting pose for compound 1 (panel A) and 2 (panel B). CK1δ Binding site is reported using the ribbon representation (light gray). The key residue Leu 85 in the hinge region is explicated by stick representation*

For compound 1, the hydrogen bond with Leu 85 is con-served as well as the CH – π interaction with Ile 23. The hydrophobic pocket is widely occupied by the naphthyl group while the ethyl-substituted pyrrole is faced out-ward. In figure 6 are reported the IEF of the two compounds while in the supplementary material are reported two comparison between the electrostatic interaction of Compound 1 and Compound 2 and the crystallographic ligand 0CK.

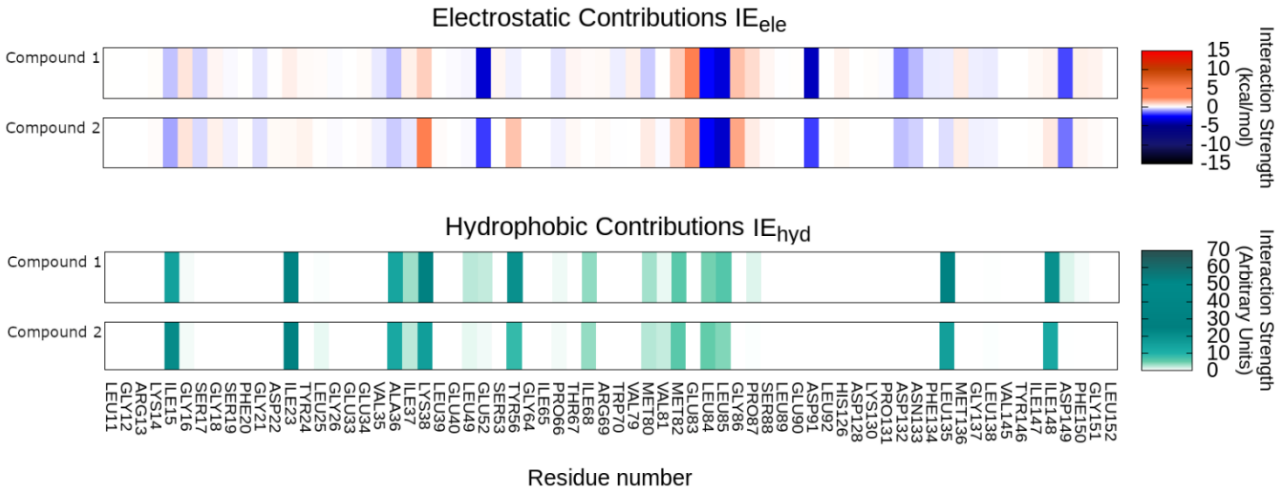


Figure 6 *Interaction Energy Fingerprint (IEF) for compound 1 and compound 2. Per residue Electrostatic (upper plot) and the hydrophobic contribution (lower plot). For Electrostatic interaction the colorimetric scale is blue to red while for the hydrophobic contibution it is white*

To verify the accuracy of the results obtained by our computational pipeline, the two selected candidates were tested using a conventional in vitro kinase activity inhibi-tory assay.

The IC_{50} values against CK1δ were of 15.22 ± 2.71 μM for compound 1 and 12.95 ± 3.21 μM for compound 2, respec-tively (Figure 3 and 4 on SI). Despite the two selected mol-ecules showed an

inhibitory effect of CK1δ activity in the micro-molar range, it is worth to underline that they were initially designed for completely different targets and, consequently, the repurposing aim of a novel scaffold can be considered as achieved. In fact, pyrrolo[3,2-f]quinolinone represents a novel scaffold for designing new CK1δ inhibitors. It is interesting to note how the strategy of in-house chemical library repurposing can be now particularly useful to cherry-pick from the library the closest analogs to our hit for developing a very pre-liminary structure-activity-relationship useful to quickly investigate the role of certain molecular decoration. However, as already anticipated, this repurposed scaffold is now subject to further study for the construction of fo-cused libraries for the necessary phase of optimization of its pharmacodynamic and pharmacokinetic properties. Interestingly, during the writing of this work a new CK1δ crystal has been released (PDB code: 6RCH) co-crystallized with a ligand having a naphthyl substituent positioned like the one suggested by us.

Concluding, the preliminary results here described sup-porting the fact that the suggested computational pipeline could represent an alternative valuable strategy to effi-ciently analyze the unexplored chemical space.

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 MAY, 1–33 (2014).

3. 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

4. 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

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 715, (Elsevier B.V, 2019).

9. Perez, D. I., Gil, C. & Martinez, A. Protein kinases CK1 and CK2 as new targets for neurodegenerative diseases. Medicinal Research Reviews 31, 924–954 (2011).

10. 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

11. 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

12. 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. 291, 5473–83 (2016).

13. Salado, I. G. et al. Protein kinase CK-1 inhibitors as new potential drugs for amyotrophic lateral sclerosis. J. Med. Chem. 57, 2755–2772 (2014).

14. Cozza, G. et al. Identification of novel protein kinase CK1 delta (CK1δ) inhibitors through structure-based virtual screening. Bioorganic Med. Chem. Lett. (2008). doi:10.1016/j.bmcl.2008.08.072

15. 3D Structure Generator CORINA Classic, Molecular Networks GmbH, Nuremberg, Germany, www.mn-am.com. Date Accessed 2020-04-24.

16. 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

17. Chemical Computing Group ULC, Molecular Operating Environment (MOE), 2019.01. 1010 Sherbrooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, 2019.

18. 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).

19. Bettayeb, K. et al. CR8, a potent and selective, roscovitine-derived inhibitor of cyclin-dependent kinases. Oncogene 27, 5797 (2008).

20. 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).

21. Korb, O., Stützle, T. & Exner, T. E. An ant colony optimization approach to flexible protein–ligand docking. Swarm Intell. (2007). doi:10.1007/s11721-007-0006-9

22. 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) 4150 LNCS, 247–258 (Springer Verlag, 2006).

23. Korb, O., Stützle, T. & Exner, T. E. Empirical scoring functions for advanced Protein-Ligand docking with PLANTS. J. Chem. Inf. Model. (2009). doi:10.1021/ci800298z

24. Halgren, T. A. et al. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening. J. Med. Chem. 47, 1750–1759 (2004).

25. Houston, D. R. & Walkinshaw, M. D. Consensus docking: Improving the reliability of docking in a virtual screening context. J. Chem. Inf. Model. 53, 384–390 (2013).

26. Mente, S. et al. Ligand-protein interactions of selective casein kinase 1δ inhibitors. J. Med. Chem. 56, 6819–6828 (2013).

27. Wager, T. T. et al. Identification and Profiling of a Selective and Brain Penetrant Radioligand for in Vivo Target Occupancy Measurement of Casein Kinase 1 (CK1) Inhibitors. ACS Chem. Neurosci. 8, 1995–2004 (2017).

28. García-Reyes, B. et al. Discovery of Inhibitor of Wnt Production 2 (IWP-2) and Related Compounds As Selective ATP-Competitive Inhibitors of Casein Kinase 1 (CK1) δ/ϵ. J. Med. Chem. 61, 4087–4102 (2018).

New Insights into Key Determinants for Adenosine 1 Receptor Antagonists Selectivity Using Supervised Molecular Dynamics Simulations

Giovanni Bolcato, Maicol Bissaro, Giuseppe Deganutti, Mattia Sturlese, Stefano Moro

Bolcato, G., Bissaro, M., Deganutti, G., Sturlese, M. & Moro, S. New Insights into Key Determinants for Adenosine 1 Receptor Antagonists Selectivity Using Supervised Molecular Dynamics Simulations. *Biomolecules* **10**, 732 (2020).

Abstract

Adenosine receptors (ARs), like many other Gprotein-coupledreceptors (GPCRs), are targets of primary interest indrug design. However, one of the main limits for the development of drugs for this class of GPCRs is the complex selectivity profile usually displayed by ligands. Numerous efforts have been made for clarifying the selectivity of ARs, leading to the development of many ligand-based models. The structure of the AR subtype A1 (A_1 AR) has been recently solved, providing important structural insights. In the present work, we rationalized the selectivity profile of two selective A_1 AR and A_{2A} AR antagonists, investigating their recognition trajectories obtained by Supervised Molecular Dynamics from an unbound state and monitoring the role of the water molecules in the binding site.

1. Introduction

Adenosine receptors (ARs) are class A G protein-coupled receptors (GPCRs) that bind the endogenous agonist adenosine. ARs are composed of four subtypes: A_1, A_{2A}, A_{2B}, A_3. While A_1 and A_3ARs (which share 49% of a sequence identity) are preferentially coupled to $G_{\alpha i}$ proteins and therefore inhibit the adenylatecyclase, A_{2A} and A_{2B} ARs (sharing 59% of a sequence identity) stimulate this enzyme, as being coupled to $G_{\alpha s}$ proteins [1]. Several ARs antagonists are in clinical trials for various diseases. With regards to A_{2A}AR, istradefylline has been recently approved for Parkinson's disease (NCT02610231) [2], PBF-509 is in phase I/II trials for non-small cell lung cancer (NCT02403193), and CPI-144 is in phase I trials for various cancer types (NTC02655822). PBF-680, on the other hand, is the only A_1 AR antagonist in clinical phase II, for the treatment of Asthma (NTC02635945)[3].

One of the difficulties during the development of ARs agonists and antagonists as therapeutic agents is the poor selectivity between differentreceptors subtypes [4]. For this reason, many efforts have

been made to elucidate the molecular basis of ARs ligands selectivity, and several structure-activity relationship (SAR) models have been developed for selective ligands of all the four subtypes [1,2,13,14,5–12]. With the increasing availability of structural information (mainly from mutagenesis, X-ray, and cryo-EM approaches [15]) in the last few years, light has been shed on the origin of selectivity on ARs. Recently, the A_1AR inactive (PDB, Protein Data Bank, code 5UEN [16] and 5N2S [17]) and active (PDB code 6D9H [18]) structures have been solved. Interestingly, in [17], Cooke and colleagues obtained the X-ray crystal structure of A_1 and A_{2A} Ars, in a complex with the same xanthine ligand PSB36, providing insight about the selectivity. There are several structural differences between A_1 AR and A_{2A} AR. The second extracellular loop (ECL2), in particular, is more folded in A_1 AR and orients perpendicularly to the plane of the membrane, while in A_{2A} AR it forms a longer helix, which is parallel to the lipid bilayer. This difference is probably due to the presence of two disulfide bonds uniquely present in A_{2A} AR. Indeed, the bond between Cys71 and Cys159 anchors ECL2 to ECL1, while the bond between Cys74 and Cys146 tethers TM3 to ECL2. The class A conserved disulfide bonds between Cys80 and Cys169 is present in both the two subtypes. These differences in the disulfide bonds likely contribute to the outward movement of the top of transmembrane helix 2 (TM) in the inactive A_1AR (Figure 1). Further divergence involves TM7, which is shifted outward compared to A_{2A} AR, due to the shorter ECL3, and TM6 slightly shifted inward in A_1 AR. These rearrangements, in turn, affect the orthosteric site of A_1 AR, which is wider than A_{2A} AR. Interestingly, the key residues in the orthosteric site of the two receptors are conserved and drive the same binding mode of the antagonist PSB36 (Figure 1). More precisely, the xanthine scaffold forms two hydrogen bonds with Asn254 (A_1 AR, Asn253 in A_{2A} AR) and a π- π stacking with Phe171 (A_1 AR, Phe168 in A_{2A} AR). Nevertheless, Asn254 (A_1 AR) is located in the binding site deeper than Asn253 in A_{2A} AR, and the xanthine ligand is consequently positioned deeply in the orthosteric site (Figure 1).

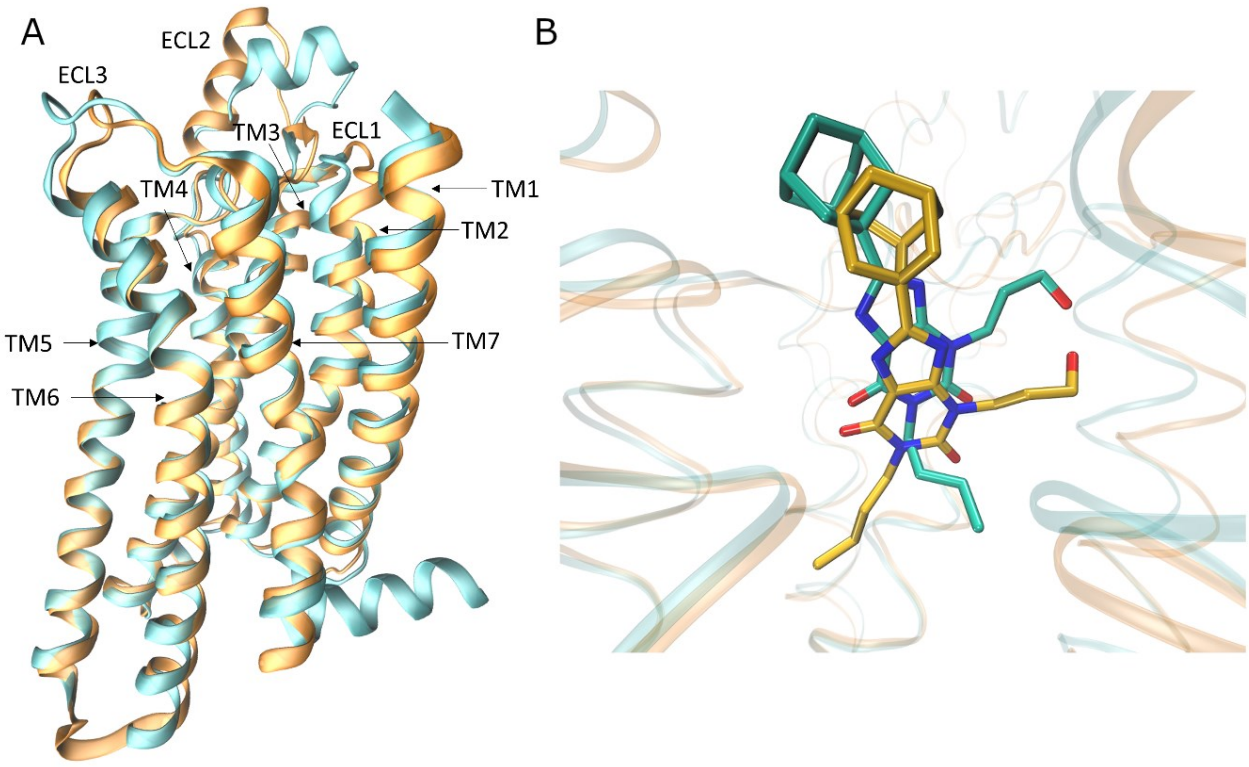


Figure 1 *Comparison of A1AR and A2AAR structures (A): Superposition of the crystal structure of the inactive A1 (orange) and A2A (cyan) adenosine receptors (Ars) (PDB code 5N2S and 5N2R respectively). (B) Superposition of the same xanthine ligand PSB36 in the two aforementioned crystal structures.*

Despite the huge help provided from high-resolution structural biology techniques, certain selectivity profiles cannot be only rationalized by the mere coordinates of bound state or "final" state. Ligand recognition is an articulated mechanism in which many variables may play a relevant role and over the last few years, there has been rising attention in the understanding of binding kinetics at GPCRs and its determinant role to successfully target this class of proteins [19].

In the present study, we used supervised molecular dynamics (SuMD) simulations to shed light on the molecular basis of the selectivity of three different ligands to A_1 AR and A_{2A} AR, not only considering the bound states, but also the possible different recognition mechanism preceding the final orthosteric site and the role of the solvent. We focus our attention on three antagonists: the A_{2A} AR selective antagonists Z48 (K_i 16.9 nM in A_{2A} AR and 1345.7 nM in A_1 AR) [20]; the A_1 AR selective antagonist LC4 (K_i 16,800 nM in A_{2A} AR and 89 nM in A_1 AR) 21; and the nonselective antagonist caffeine (Figure 2). SuMD [22,23] is a molecular dynamics (MD) approach that allows for the study of molecular recognition processes in a fully atomistic way, in the nanosecond timescale, without introducing any energetic biases.

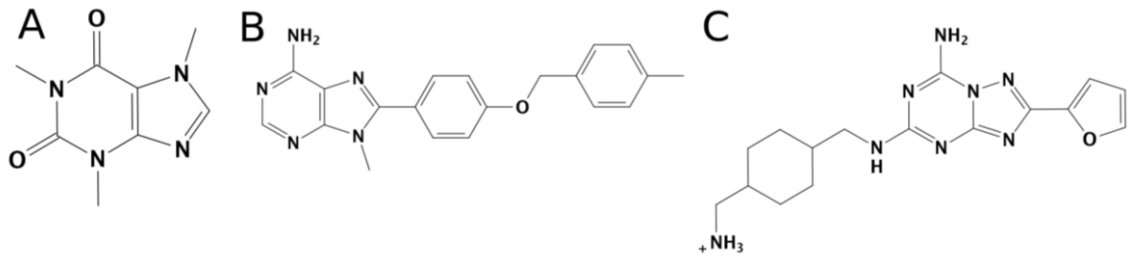


***Figure 2**. The three ligands considered in the present study. (A) Caffeine, a non-selective ARs antagonist. (B) LC4, an A_1 AR selective antagonist. (C) Z48, an A_{2A} AR selective antagonist.*

2. Materials and methods

2.1 System Setup

The crystal structures of the two receptors were retrieved from PDB (the PDB code is 5N2S for A_1 AR and 5NM4 for A_{2A} AR). Systems preparation was performed using a Molecular Operating Environment (MOE)) suite (Chemical Computing Group ULC, Molecular Operating Environment (MOE), 2019.01. 1010 Sherbrooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, 2019) [24] for protein preparation (removal of crystallographic water molecules, ions, and other solvent molecules, selection of the highest occupancy for each residue, assignment of the correct protonation state at pH 7.4). Systems preparation for the molecular dynamics simulations was carried out using VMD [25]. The protein was explicitly solvated in a water box with the borders placed at a distance of 15 Å from any protein atom, the water model used was TIP3P 26. The system charge was neutralized to a concentration of 0.154 M using Na^+/Cl^-. The lipid bilayer consisted of phosphatidylcholine (POPC) units. The sodium ion within the TMD allosteric site of A_{2A} AR was retained, and it was also placed by superposition in A_1 AR.

2.2. Equilibration of the System

All the simulations were performed with a CHARMM36 force field [27] and using ACEMD2 [28]. Ligands parameters were retrieved from Paramchem [29], a web interface for the assignment of parameters based on the CGenFF [30] force field. The system energy was minimized in 1500 steps using the conjugate-gradient method, then the equilibrationof the system was done in four steps. The first one consisted of 5 ns of NPT simulation with harmonic positional constraints of 1 kcal $mol^{-1}Å^{-2}$ on each atom of the protein and the lipid bilayer. The second step consisted in 10 ns of NPT simulation with harmonic positional constraints of 1 kcal $mol^{-1}Å^{-2}$ only on eachprotein atom and on the phosphorus atom of the POPC units, the third step consisted in 5 ns of NPT simulation with harmonic

positional constraints of 1 kcal $mol^{-1}Å^{-2}$ only on the alpha carbons of the protein, and the last step consisted in 50 ns of NVT simulation without any constraints.

For the productive simulations, the temperature was maintained at 310 K using the Langevin thermostat (company, city, city abbreviation if USA, country), with a low dumping of 1 ps^{-1}. The pressure was set at 1 atm using the Berendsen barostat [31]. The particle-mesh Ewald (PME) method was used to calculate the electrostatic interactions with a 1 Å grid [32]. A 9.0 Å cutoff was applied for long-term interactions. The M-SHAKE algorithm was applied to constrain the bond lengths involving hydrogen atoms.

At the end of the equilibration, several parameterswere calculated to assess the stability of the system:the root mean square deviation (RMSD) of the alpha carbons of the protein ,the root mean square fluctuation (RMSF) of each protein residue, the volume of the cell (which should tend to a plateau in the NPT ensemble), and the area per lipid (APL) for each membrane layer (calculated using GridMAT-MD [33]). We also computed the volume of the orthosteric siteduring the equilibration using POVME [34]. Figure S1 (A_{2A} AR) and Figure S2 (A_1 AR) report the analysis performed during the equilibration of the system. In both the two systems, the protein reached a stable conformation (RMSD of the protein C_α (panel A, figure S1 and S2) stably below 2 Å for A_{2A} AR and below 3 Å for A_1 AR). The volume of the orthosteric site reached a plateau in both cases (Panel B, figure S1 and S2). The most flexible parts of the protein, as expected, are the loops. Indeed, the RMSF of these regions is higher than the TMs (panels C and D, figure S1 and S2).

As shown in Figure S5, the orthosteric site of A_1 AR appears to be deeper than A_{2A} AR, due to a cleft between TM5 and TM6. The APL and the volumetric analysis are reported in Figure S3 (A_1 AR) and Figure S4 (A_{2A} AR). For both systems, the cell volume reached stable values.

2.3. Supervised Molecular Dynamics Simulations

SuMD [22,23] is a molecular dynamics (MD) approach that allows for the investigation of molecular recognition processes in a fully atomistic way, in the nanosecond timescale, without introducing any energetic bias (Figure 3). Ligands were placed 35 Å away from the protein. Each SuMD step was set to 600 ps. During each SuMD step, the distance between the center of mass of the binding site (defined by a series of residues) and the center of mass of the ligand is monitored. These data are then fitted, and if the slope of the interpolating linear function is negative, then the coordinates and the velocities are used for the successive time window, otherwise, the last time window is simulated again reassigning the velocities (this reassignment of the velocities is intrinsic in the use of Langevin thermostat). If the condition fails 30 consecutive times, then the simulation is stopped. Otherwise,

the algorithm continues until the distance between the two centers of mass is below the threshold of 5Å; at this point, the supervision is turned off and 30 short classical MD simulations are performed, switching on the supervision if the distance between the centroids becomes greater than 5Å. At the end of the SuMD process, the trajectory is prolonged for 25 ns of classical MD simulation.

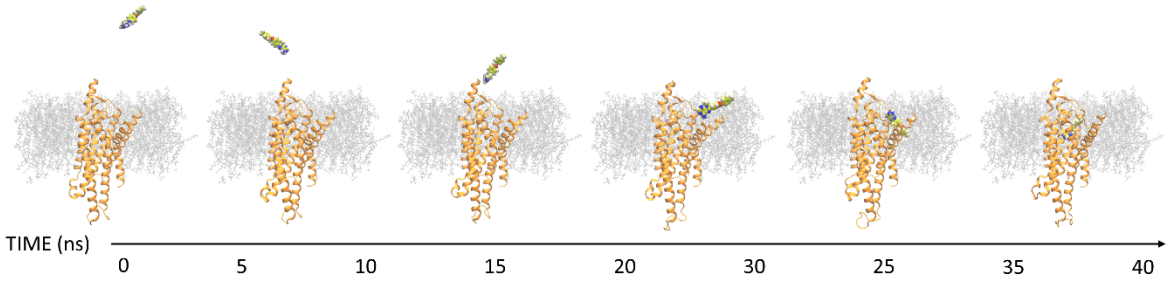


Figure 3 *Representation of a binding event sampled by a supervised molecular dynamics (SuMD) simulation. After each reported step, the distance between the ligand and the binding site decreases. In less than (merged) 50 ns, a binding event was sampled.*

Twenty simulations were performed for each of the six systems (Z48/A_1 AR, Z48/A_{2A} AR, LC4/A_1 AR, LC4/A_{2A} AR, Caffeine/A_1 AR, Caffeine/A_{2A} AR). Only the simulations that sampled the ligand reaching the orthosteric site and interacting with the classic fingerprint of these class of ligands (Figure S6) were here reported (e.g., one replica for each system, excepted LC4/A_1 AR for which two replicas were analyzed). For each system, the reasons for failure are similar. In most cases, the ligands interact strongly with the residues of the ECLs, and do not reach the binding site. In a few cases, the ligands only partially reach the orthosteric site. Finally, in other rare cases, the ligands get stuck between the protein and the membrane.

2.4. Trajectories Analysis

The SuMD trajectories were analyzed using anin-house python tool (described in [35]) that provides information on the geometry and the energetic of the system. The output consists of a per-residue analysis of the electrostatic and van Der Waals contributions to the protein-ligand interactions; a representation of the distance between the center of mass of the ligand and the center of mass of the binding site as a function of time; a global energetic evaluation of the system as a function of the aforementioned distance. This analysis allowscomparing the energetic profile of two systems both in a general way and, through the per-residue analysis (i.e., it is possible to evaluate which ligand has better interaction with some specific amino acids of interest over the time).

Three replicas of 50 ns were performed on the apo form of A_1 AR and A_{2A} AR, after the equilibration stage described before. These three replicas were merged and analyzed by AquaMMapS [36].

3. Results

3.1. SuMD Binding of the A_1 AR Nonselective Antagonist Caffeine

According to the SuMD simulations (Figures S7,S8; Video 1 for A_1 AR and Video 2 for A_{2A} AR in the Supplementary Information), the nonselective antagonist caffeine establishes intermediate interactions with the extracellular vestibule of A_1 AR and A_{2A} AR, before reaching the orthosteric site. The most stable bound configurations sampled on A_1 AR and A_{2A} AR differed for theorientation of the xanthine ring (Figure S9). On A_{2A} AR, caffeine pointed the N7-methyl toward Asn253, while in A_1 AR it was rotated by 180°, with the N3 methylin the proximity of Asn254. Notably, both these two conformations have been experimentally observed in X-ray crystal structuresof A_{2A} AR [17]. Fluctuations of the ligand in both orthosteric sites and the transient nature of the interactions are easily depicted in the energy interaction landscapes (Panels A,S7,S8) in which the points are particularly scattered over the distance between the centers of mass of the ligand and the orthosteric site.

3.2. SuMD Binding of the A_{2A} AR Selective Antagonist Z48

Figures S10,S11 report the analysis of Z48/A_1 AR and Z48/A_{2A} AR, including 25 ns of classic MD simulations performed at the end of each SuMD simulation (Videos 3,4 in the Supplementary Information). For what concerns the SuMD trajectory of Z48 on A_1 AR, notable electrostatic repulsion took place between the ligand and Lys168, Lys173, and Lys265, before the ligand reached the orthosteric site, as clearly observable in the per-residue electrostatic interaction energy plot (Panel C,S10). These residues are positioned on the ECLs (Lys168 and Lys173 on ECL2, and Lys265 on ECL3). As a result, in A_1 AR, Z48 did not adopt the binding fingerprint of the ARs antagonists (e.g., only one hydrogen bond with Asn254 out of two was formed). Moreover, this binding mode was unstable over the 25 ns of classic MD simulations. Interestingly, the terminal amine group of the ligand strongly interacts with Glu170 (Leu167 in A_{2A} AR). Figure 4 reports the binding modes of Z48 at the end of the SuMD simulation on A_1 AR (the binding mode at the end of the 25 ns of classic MD simulation is shown in Figure S12) and A_{2A} AR (at the end of the 25 ns of classic MD simulation), respectively. Moving to the binding simulations of Z48 toA_{2A} AR, two SuMD replicas led to the classic binding mode of the ARs antagonists. The π πstacking with Phe168 was present, along with the hydrogen bonds with Asn253 and with Glu169. Both binding modes were stable during 25 ns of classic MD simulations (Figure S11, Panel C,D, Figure 4). No significant protein-ligand electrostatic

repulsions were observed. The narrow funnel-like interaction energy profile of the ligand in the orthostatic also suggested a good complementarity of the ligand in the pocket, and a rapid reaching of a stable bound state.

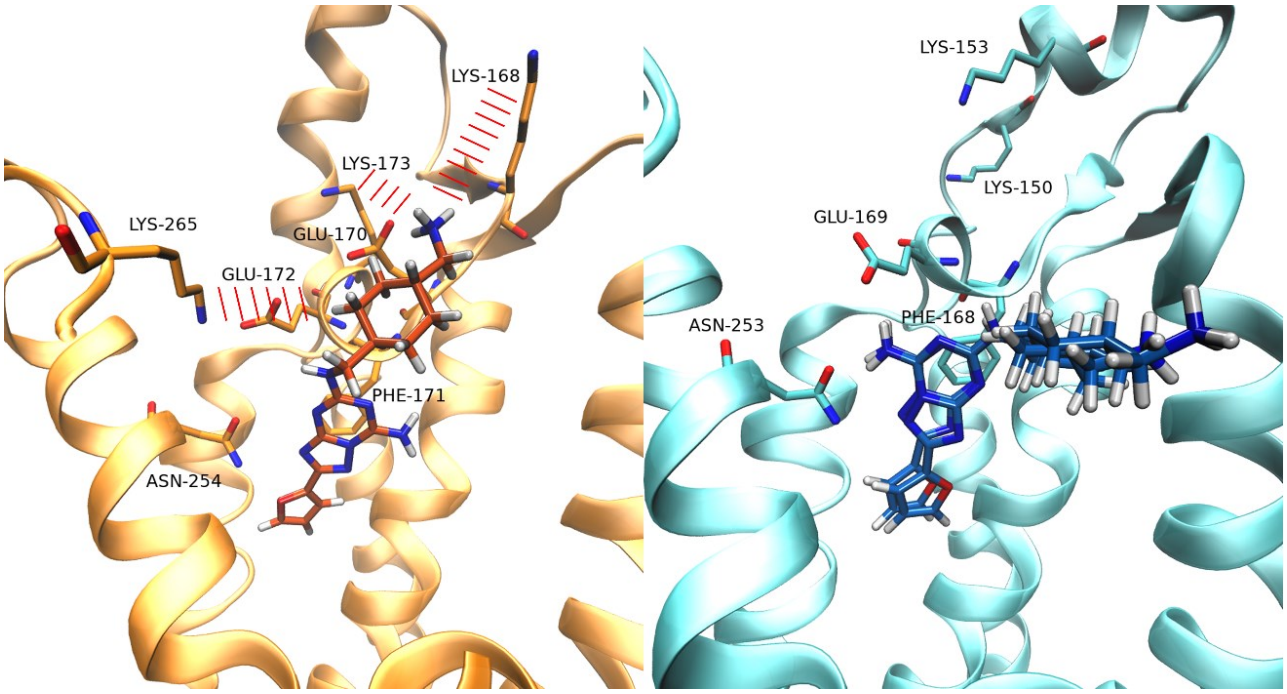


Figure 4 *Binding mode of Z48 within the binding site of A_1 AR on the left-hand side, and within the binding site of A_{2A} AR on the right-hand side (superposition of the two simulations analyzed). For A_1 AR, the binding mode is reported at the end of the SuMD simulation. For A_{2A} AR, the two poses at the end of the classic MD simulation are reported.*

3.3. SuMD Binding of the A_1 AR Selective Antagonist LC4

With regards to LC4, two SuMD replicas led to different binding modes of LC4 into the A_1 AR orthosteric site. During Replica 1, the ligand formed a complex characterized by the xanthine scaffold positioned in the orthosteric binding site, and the N8-substituent pointed outward the receptor (Figure 5). A hydrogen bond with Asn254 and hydrophobic contacts with Phe171 occurred (Figure 6A). Interestingly, the LC4 oxygen atom in the N8 substituent interacted with and stabilized water molecules in the proximity of the Phe171 backbone, a hydrated spot on ECL3, according to the AquaMMapS analysis (Figure 6A).

In the case of A_{2A} AR, the xanthine scaffold reached the orthosteric site rotated by 180° compared to the binding mode adopted in A_1 AR (Figure 6B) and with an unfavorable geometry for hydrogen bonding with Asn253. In both of these two binding modes (Figure S13 and Figure S14), LC4 did not interact withconserved glutamate Glu172 (A_1 AR) or Glu169 (A_{2A} AR).

During SuMD Replica 2 on A_1 AR, LC4 experienced a two-step binding (Figure 6). First, the antagonist entered the orthosteric site pointing the methylphenyl group into a hydrophobic pocket located

between TM2 and TM3. This cryptic pocket is not present in A_{2A} AR in light of bulkier residues and a higher degree of packing between the helixes (Figure 6B). From this metastable configuration, LC4 moved deeper into the orthosteric site and engaged Asn254 and Glu172 in hydrogen bonds (Figure 6). In this binding mode, the ligand inserted the N8-substituent inside a further cryptic pocket between TM5 and TM6 (Figure 6C,D), which is delimited by the "toggle switch" residue Trp247 [15], and the residues Ile95 and Phe253, being part of the conserved class A structural motif PIF [37]. Notably, this hydrophobic sub-pocket was occupied by likely "unhappy" water molecules during simulations of the apo-A_1 AR (Figure 6A). The video of these three simulations can be found in Supplementary Materials, Video 5 and Video 6 for the two simulations of LC4/A_1 AR, and Video 7 for the simulation of LC4/A_{2A} AR.

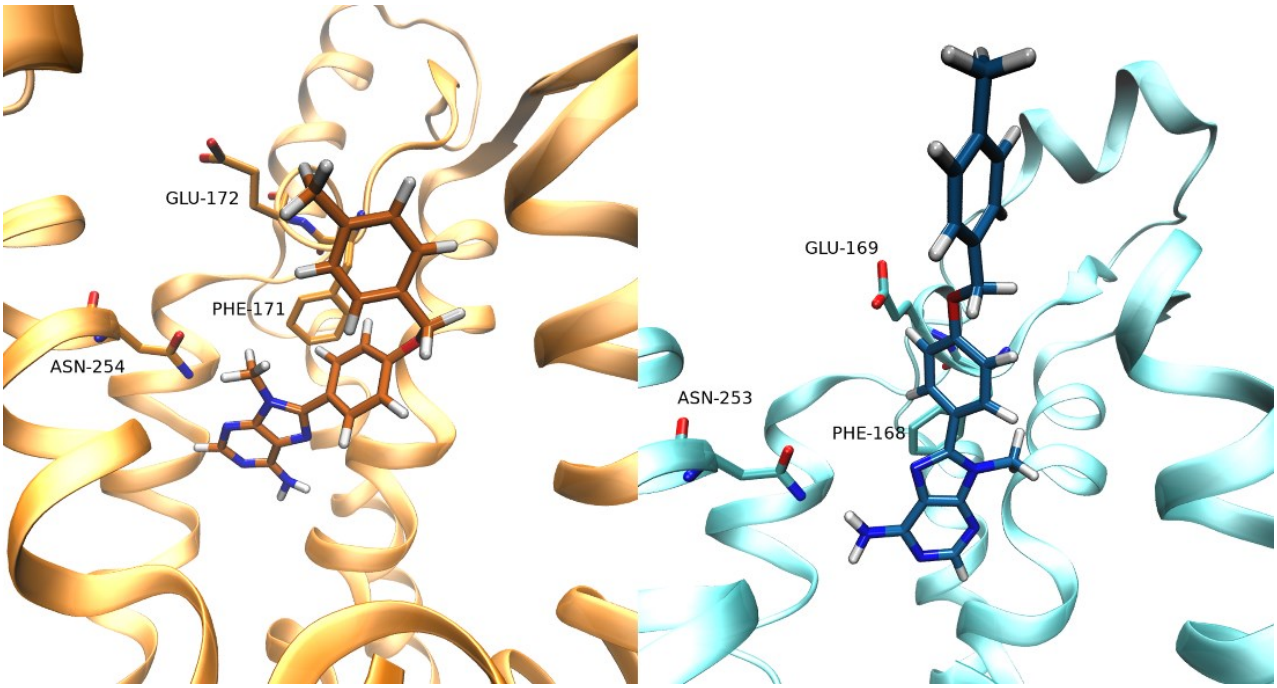


Figure 5 *Binding mode of ligand LC4. A_1 AR on the left and A_{2A} AR on the right.*

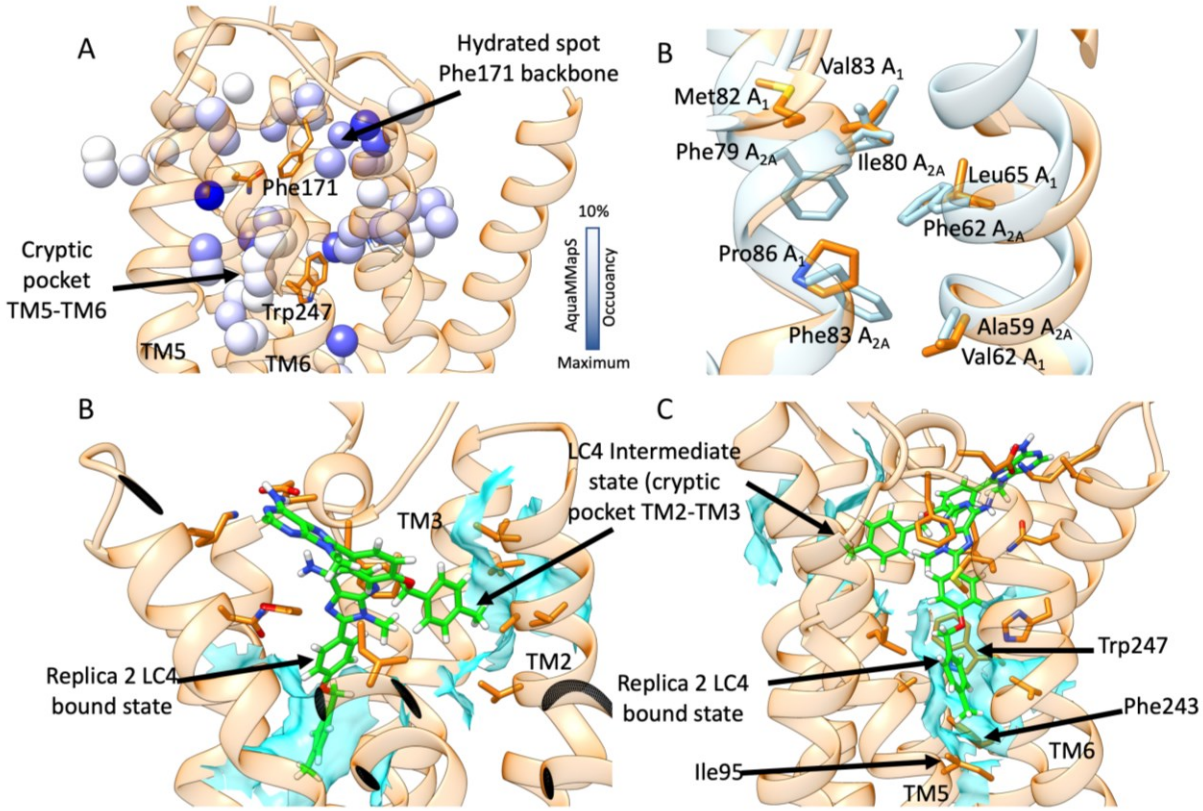


Figure 6 *(A) Hydrated spots within A_1 AR with AquaMMapS occupancy > 10%. The cryptic pocket between TM5 and TM6, as well as the spot nearby the Phe171 backbone, are indicated; (B) comparison between A_1 AR (orange) and A_{2A} AR(cyan) at the level of TM2 and TM3—a sub-pocket formed only in A_1 AR during MD simulations, due to different residues and interhelical packing; (C) and (D) two side views of the two superimposed binding steps of LC4 (green stick) to A_1 AR. Hydrophobic contacts are shown as cyan transparent surfaces.*

4. Discussion

Here we present results from SuMD simulations performed to shed light on the selectivity displayed by LC4 and Z48, two antagonists of A_1 AR and A_{2A} AR, respectively. The nonselective antagonist caffeine was also dynamically docked to the two ARs subtype. Caffeine, which is in a weak binder (micromolar range [38]) of all the ARs, experienced more than one binding mode, in line with our previous simulations [22] and experimental observation [17].

SuMD binding of the selective A_{2A} AR antagonist Z48 on A_1 AR and A_{2A} AR suggested different interaction patterns along the pathways. Z48 experienced unfavorable electrostatic interactions between positively charged A_1 AR residues Lys168, Lys173 (ECL2), and Lys265 (ECL3), and the ligand charged amine on the N8-substituent. Such transitory states did not take place during binding to A_{2A} AR, as no significant electrostatic repulsions were computed at the extracellular vestibule. Interestingly, A_{2A} AR bears Ala265 instead of Lys265 on the ECL3, while Lys168 and Lys173 (ECL2) are farther from the binding site, compared to A_1 AR. This is consistent with mutagenesis studies that demonstrated the importance of these lysine residues for the binding of several A_1 AR ligands

[40,41]. On A_{2A} AR, Z48 reached the orthosteric site producing the classic interactions fingerprint of the ARs antagonists. On A_1 AR, on the other hand, the ligand sampled a different binding mode. Moreover, the A_1 AR residue Glu170 (Val167 in A2AAR) strongly interacted with the charged terminal amine of the ligand, stabilizing this alternative binding mode. Z48 was proposed to bind to A_{2A} AR overcoming low enthalpy transition state(s) [40]. From this standpoint, the unfavorable electrostatic interactions with ECL2 could implicate a slower binding toA_1 AR, and therefore a kinetic selectivity for A_{2A} AR.

SuMD simulations of the antagonist LC4 proposed two possible binding modes that could drive selectivity. It is possible that the ligand binds A_1 AR and A_{2A} AR with the same conformation, but differently interacting with water molecules nearby ECL3. From this standpoint, and in analogy with [41], we propose these "happy" water molecules contribute to the ligand stabilization in A_1 AR (Figure S16), but not in A_{2A} AR (Figure S17). An alternative and unique LC4 binding mechanism was sampled only on A_1 AR, with a metastable state before the final complex formation. Along this pathway, two cryptic hydrophobic pockets (between TM2 and TM3 and between TM5 and TM6) allowed the N8-substituent of the ligand to correctly orient first, and then engage key residues for the receptor activation (the "toggle switch" Trp247, Ile95, and Phe253, which are part of the conserved class A motif PIF). Notably, the cryptic pocket between TM2 and TM3 has recently been proposed as a determinant for A_1AR selectivity displayed by the triazolotriazine antagonist LUF5452 42

5. Conclusions

Understanding the selectivity of GPCRs ligands is an important task in drug design. This study supports the emerging idea that selectivity is driven by a plethora of phenomena, other than the protein-ligand interactions in the bound state. Receptor-ligands recognitions are multistep events modulated by intermediate interaction along with the (un)binding paths. This picture may be further complicated by the presence of stable water molecules, which can have a tremendous impact on stabilizing or destabilizing an orthosteric complex. To consider different aspects that may affect the selectivity on A_1 AR and A_{2A} AR, we used SuMD simulations to investigate the recognition of thee different antagonists. Overall, our results suggest that kinetic selectivity may favor the binding of Z48 to A_{2A} AR over LC4. A different scenario was observed for A_1 AR, the recognition trajectories highlighted the key role of water molecules in the binding mode of LC4, which is favored by two hidden sub pockets within A_1 AR.

References

1. Müller, C. E. & Jacobson, K. A. Recent developments in adenosine receptor ligands and their potential as novel drugs. *Biochimica et Biophysica Acta - Biomembranes* (2011) doi:10.1016/j.bbamem.2010.12.017.

2. Navarro, G., Borroto-Escuela, D. O., Fuxe, K. & Franco, R. Purinergic signaling in Parkinson's disease. Relevance for treatment. *Neuropharmacology* (2016) doi:10.1016/j.neuropharm.2015.07.024.

3. Gao, Z. G. & Jacobson, K. A. Purinergic signaling in mast cell degranulation and asthma. *Frontiers in Pharmacology* (2017) doi:10.3389/fphar.2017.00947.

4. Chen, J. F., Eltzschig, H. K. & Fredholm, B. B. Adenosine receptors as drug targets-what are the challenges? *Nat. Rev. Drug Discov.* (2013) doi:10.1038/nrd3955.

5. Moro, S., Gao, Z. G., Jacobson, K. A. & Spalluto, G. Progress in the pursuit of therapeutic adenosine receptor antagonists. *Medicinal Research Reviews* (2006) doi:10.1002/med.20048.

6. 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* (2010).

7. Kiesman, W. F., Elzein, E. & Zablocki, J. A1 adenosine receptor antagonists, agonists, and allosteric enhancers. *Handbook of Experimental Pharmacology* (2009) doi:10.1007/978-3-540-89615-9_2.

8. Cristalli, G., Müller, C. E. & Volpini, R. Recent developments in Adenosine A2A receptor ligands. *Handbook of Experimental Pharmacology* (2009) doi:10.1007/978-3-540-89615-9_3.

9. Manera, C. & Saccomanni, G. A2A Receptor Ligands: Past, Present and Future Trends. *Curr. Top. Med. Chem.* (2010) doi:10.2174/156802610791268765.

10. Baraldi, P. G., Tabrizi, M. A., Fruttarolo, F., Romagnoli, R. & Preti, D. Recent improvements in the development of A2B adenosine receptor agonists. *Purinergic Signalling* (2009) doi:10.1007/s11302-009-9140-8.

11. Ortore, G. & Martinelli, A. A2B Receptor Ligands: Past, Present and Future Trends. *Curr. Top. Med. Chem.* (2010) doi:10.2174/156802610791268747.

12. Jacobson, K. A. *et al.* Medicinal chemistry of the A3 adenosine receptor: Agonists, antagonists, and receptor engineering. *Handbook of Experimental Pharmacology* (2009) doi:10.1007/978-3-540-89615-9_5.

13. Müller, C. E. & Jacobson, K. A. Xanthines as adenosine receptor antagonists. *Handbook of Experimental Pharmacology* (2011) doi:10.1007/978-3-642-13443-2_6.

14. Schenone, S., Brullo, C., Musumeci, F., Bruno, O. & Botta, M. A1 Receptors Ligands: Past, Present and Future Trends. *Curr. Top. Med. Chem.* (2010) doi:10.2174/156802610791268729.

15. Jespers, W. *et al.* Structural Mapping of Adenosine Receptor Mutations: Ligand Binding and Signaling Mechanisms. *Trends in Pharmacological Sciences* (2018) doi:10.1016/j.tips.2017.11.001.

16. Glukhova, A. *et al.* Structure of the Adenosine A1 Receptor Reveals the Basis for Subtype Selectivity. *Cell* (2017) doi:10.1016/j.cell.2017.01.042.

17. Cheng, R. K. Y. *et al.* Structures of Human A1 and A2A Adenosine Receptors with Xanthines Reveal Determinants of Selectivity. *Structure* (2017) doi:10.1016/j.str.2017.06.012.

18. Draper-Joyce, C. J. *et al.* Structure of the adenosine-bound human adenosine A1 receptor-Gi complex. *Nature* (2018) doi:10.1038/s41586-018-0236-6.

19. Sykes, D. A., Stoddart, L. A., Kilpatrick, L. E. & Hill, S. J. Binding kinetics of ligands acting at GPCRs. *Molecular and Cellular Endocrinology* vol. 485 9–19 (2019).

20. Federico, S. *et al.* Synthesis and biological evaluation of a new series of 1, 2, 4-triazolo[1, 5-α]-1, 3, 5-triazines as human a 2a adenosine receptor antagonists with improved water solubility. *J. Med. Chem.* (2011) doi:10.1021/jm101349u.

21. Lambertucci, C. *et al.* New 9-methyl-8-(4-hydroxyphenyl)adenine derivatives as A1 adenosine receptor antagonists. *Collect. Czechoslov. Chem. Commun.* (2011) doi:10.1135/cccc2011091.

22. 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).

23. 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).

24. Chemical Computing Group ULC, Molecular Operating Environment (MOE), 2019.01. 1010 Sherbrooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, 2019.

25. Humphrey, W., Dalke, A. & Schulten, K. VMD: Visual molecular dynamics. *J. Mol. Graph.* (1996) doi:10.1016/0263-7855(96)00018-5.

26. 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.

27. Brooks, B. R. *et al.* CHARMM: The biomolecular simulation program. *J. Comput. Chem.* (2009) doi:10.1002/jcc.21287.

28. 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.

29. https://cgenff.umaryland.edu/. https://cgenff.umaryland.edu/.

30. 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.* (2010) doi:10.1002/jcc.21367.

31. Berendsen, H. J. C., Postma, J. P. M., Van Gunsteren, W. F., Dinola, A. & Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* (1984) doi:10.1063/1.448118.

32. Essmann, U. *et al.* A smooth particle mesh Ewald method. *J. Chem. Phys.* **103**, 8577–8593 (1995).

33. Allen, W. J., Lemkul, J. A. & Bevan, D. R. GridMAT-MD: A grid-based membrane analysis tool for use with molecular dynamics. *J. Comput. Chem.* (2009) doi:10.1002/jcc.21172.

34. Wagner, J. R. *et al.* POVME 3.0: Software for Mapping Binding Pocket Flexibility. *J. Chem. Theory Comput.* (2017) doi:10.1021/acs.jctc.7b00500.

35. Salmaso, V., Sturlese, M., Cuzzolin, A. & Moro, S. Exploring Protein-Peptide Recognition Pathways Using a Supervised Molecular Dynamics Approach. *Structure* (2017) doi:10.1016/j.str.2017.02.009.

36. 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* (2018) doi:10.1002/cmdc.201700564.

37. Wacker, D. *et al.* Structural features for functional selectivity at serotonin receptors. *Science (80-.).* (2013) doi:10.1126/science.1232808.

38. Ishiyama, H., Ohshita, K., Abe, T., Nakata, H. & Kobayashi, J. Synthesis of eudistomin D analogues and its effects on adenosine receptors. *Bioorganic Med. Chem.* (2008) doi:10.1016/j.bmc.2008.01.041.

39. Nguyen, A. T. N. *et al.* Extracellular loop 2 of the adenosine A1 receptor has a key role in orthosteric ligand affinity and agonist efficacy. *Mol. Pharmacol.* (2016) doi:10.1124/mol.116.105007.

40. Dawson, E. S. & Wells, J. N. Determination of amino acid residues that are accessible from the ligand binding crevice in the seventh transmembrane-spanning region of the human A1 adenosine receptor. *Mol. Pharmacol.* (2001) doi:10.1124/mol.59.5.1187.

41. Deganutti, G. *et al.* Impact of protein–ligand solvation and desolvation on transition state thermodynamic properties of adenosine A2A ligand binding kinetics. *Silico Pharmacol.* (2017) doi:10.1007/s40203-017-0037-x.

42. Bortolato, A., Tehan, B. G., Bodnarchuk, M. S., Essex, J. W. & Mason, J. S. Water network perturbation in ligand binding: Adenosine A2A antagonists as a case study. *J. Chem. Inf. Model.* (2013) doi:10.1021/ci4001458.

43. Mattedi, G., Deflorian, F., Mason, J. S., De Graaf, C. & Gervasio, F. L. Understanding Ligand Binding Selectivity in a Prototypical GPCR Family. *J. Chem. Inf. Model.* (2019) doi:10.1021/acs.jcim.9b00298.

A Deep-Learning Approach toward Rational Molecular Docking Protocol Selection

José Jiménez-Luna, Alberto Cuzzolin, **Giovanni Bolcato**, Mattia Sturlese and Stefano Moro

Jiménez-Luna, J., Cuzzolin, A., Bolcato, G., Sturlese, M. & Moro, S. A Deep-Learning Approach toward Rational Molecular Docking Protocol Selection. Molecules 25, 2487 (2020).

Abstract

While a plethora of different protein–ligand docking protocols have been developed over the past twenty years, their performances greatly depend on the provided input protein–ligand pair. In this study, we developed a machine-learning model that uses a combination of convolutional and fully connected neural networks for the task of predicting the performance of several popular docking protocols given a protein structure and a small compound. We also rigorously evaluated the performance of our model using a widely available database of protein–ligand complexes and different types of data splits. We further open-source all code related to this study so that potential users can make informed selections on which protocol is best suited for their particular protein–ligand pair.

1. Introduction

Molecular docking is nowadays a common approach in a computational drug discovery pipeline[1,2]: knowing a good approximation to the crystal pose of a ligand can provide medicinal chemists with new ideas for lead optimization that could potentially accelerate structure-based drug design. A docking protocol can be described as the combination of a search algorithm that samples the conformational space of a ligand within a binding site and a scoring function, which quantitatively evaluates the accuracy of such poses. While in many cases the conformational search operated by docking protocols is effective in producing the correct pose for a ligand (i.e., the crystallographic pose is generally reproduced within reasonable accuracy), scoring functions often fail in ranking them (i.e., the crystallographic pose often is usually not the one with the best score)[3]. Given that the choice of the scoring function considerably affects results, and, to rationalize protocol choice, the comparison of the performance of different protocols is commonly performed in the early stages of docking studies. In particular, the DockBench platform[4] was recently developed with the aim to facilitate protocol selection. The aforementioned platform presents a benchmark of different docking protocols in a self-docking routine, whose goal is to reproduce the pose of a ligand with a

known co-crystal: the ability of each protocol in producing the crystallographic pose being measured in terms of their Root Mean Square Deviation (RMSD).

In particular, the average and the lowest RMSD (RMSDave and RMSDmin) of the generated poses are reported, as well as the number of poses with a lower RMSD than the X-ray resolution of the corresponding crystal (nRMSD)[5]. The success of introducing a benchmarking procedure in molecular docking campaigns has been reported in several blind challenges[6,7]. This approach has been shown to be particularly useful when multiple protein–ligand complexes are available for the same target, making protein conformation choice a further variable to be considered. An ideal docking scoring function would produce the lowest RMSDave and RMSDmin metrics, leading to a better reproduction of the crystallographic pose. Motivated by this and the previously mentioned challenges, in the work presented here, we try to address the following two questions: 1. Given a particular docking protocol, would it be possible to know a priori which protein–ligand pairs will result in the best docking pose? 2. Is there a preferable way of choosing the best docking protocol for an arbitrary ligand rather than selecting the one that reproduces the best self-docking pose for a particular proteins structure? Applications of Deep Learning (DL) in drug discovery have become ubiquitous in the last few years, as these methods have shown promise in relevant problems such as property prediction[8,9,10,11,12,13], compound retrosynthesis[14], de-novo drug design[15,16], and reaction prediction[17], among many others. In the context of molecular docking, DL approaches have been investigated to replace classical scoring functions, showing moderate success[18,19], but still far behind the accuracy provided by standard docking procedures. Partially due to this fact, in this study, we explored the potential of DL approaches to both select the best possible docking protocol given a protein–ligand pair and to provide insight into which protein–ligand pairs will result in a better pose given a docking protocol. We performed an exhaustive evaluation of the proposed methodology using the diverse and well-known PDBbind protein–ligand database[20] and different data splits to conclude that the approach is able to help users make informed docking modeling choices. We furthermore open-source all our production and evaluation code so that the community can either use our models or reproduce the results presented in this work easily.

2. Results and Discussion

We prepared the protein–ligand refined set of the PDBbind database[21] (v.2017) according to the workflow previously described in the DockBench suite (see Sections 3.1 and 3.2). With these data, we used the aforementioned software to generate docking results for 14 different well-known

commercial and open-source protocols (see Section 3.3). A combination of 3D-convolutional and fully connected neural networks (see Section 3.5) was used as our main model alongside a voxelized representation of the protein pocket and a mixture of extended connectivity fingerprints[22] and two-dimensional descriptors for the ligand (see Section 3.4). The proposed model was trained to predict three quantities of interest (RMSDave,RMSDmin, and nRMSD) with the goal of determining which protein–ligand pairs work better under specific docking protocols (i.e., our first research question). We furthermore used four different evaluation data splits (see Section 3.6) to understand under which circumstances the models here presented perform optimally. For each docking protocol (see Section 3.3), we present results on the evaluation of the predicted RMSDave, RMSDmin, and nRMSD against the molecular docking results, using the root mean squared error (RMSE) and Pearson's correlation coefficient R metrics (Table 1 and Tables S1 and S2).

Protocol	RMSE	Pearson's *R*	RMSE	Pearson's *R*	RMSE	Pearson's *R*	RMSE	Pearson's *R*
	Random		Ligand Scaffold		Protein Classes		Protein Classes Balanced	
autodock-ga	1.60 (±0.08)	0.74 (±0.03)	1.34 (±0.26)	0.38 (±0.21)	1.76 (±0.09)	0.60 (±0.05)	1.48 (±0.04)	0.73 (±0.02)
autodock-lga	2.01 (±0.08)	0.65 (±0.03)	1.82 (±0.41)	0.30 (±0.20)	2.20 (±0.13)	0.57 (±0.05)	1.89 (±0.03)	0.70 (±0.02)
autodock-ls	2.04 (±0.09)	0.50 (±0.04)	1.79 (±0.18)	0.50 (±0.14)	2.02 (±0.05)	0.41 (±0.04)	1.93 (±0.03)	0.46 (±0.02)
glide-sp	2.79 (±0.18)	0.52 (±0.05)	3.34 (±0.55)	0.14 (±0.14)	2.84 (±0.38)	0.44(±0.07)	2.34 (±0.12)	0.64 (±0.03)
gold-asp	2.43 (±0.10)	0.68 (±0.02)	2.50 (±0.58)	0.50 (±0.21)	2.52 (±0.21)	0.64 (±0.14)	2.08 (±0.08)	0.78 (±0.01)
gold-chemscore	2.59 (±0.14)	0.62 (±0.03)	2.74 (±0.39)	0.37(±0.19)	2.62 (±0.12)	0.61 (±0.03)	2.25 (±0.13)	0.73 (±0.02)
gold-goldscore	2.47 (±0.10)	0.52 (±0.03)	2.44 (±0.72)	0.53 (±0.29)	2.49 (±0.19)	0.51 (±0.06)	2.12 (±0.14)	0.66 (±0.03)
gold-plp	2.49 (±0.15)	0.66 (±0.03)	2.53 (±0.52)	0.32 (±0.22)	2.57 (±0.27)	0.62 (±0.06)	2.14 (±0.05)	0.76 (±0.01)
plants-chemplp	2.55 (±0.17)	0.44 (±0.02)	2.68 (±0.99)	−0.02 (±0.06)	2.55 (±0.24)	0.56 (±0.23)	2.23 (±0.13)	0.58 (±0.02)
plants-plp95	3.04 (±0.09)	0.42 (±0.02)	3.16 (±0.89)	−0.12 (±0.07)	3.08 (±0.23)	0.40 (±0.03)	2.58 (±0.22)	0.57 (±0.04)
plants-plp	2.75 (±0.17)	0.43 (±0.02)	2.76 (±0.58)	0.09 (±0.37)	2.79 (±0.27)	0.41 (±0.28)	2.44 (±0.10)	0.54 (±0.02)
rdock-solv	3.95 (±0.23)	0.35 (±0.26)	3.58 (±0.34)	0.09 (±0.08)	3.73 (±0.48)	0.42 (±0.09)	3.33 (±0.22)	0.54 (±0.18)
rdock-std	3.92 (±0.05)	0.35 (±0.25)	3.62 (±0.43)	0.08 (±0.46)	3.71 (±0.41)	0.42 (±0.09)	3.23 (±0.19)	0.56 (±0.03)
vina-std	2.23 (±0.03)	0.40 (±0.03)	2.30 (±0.15)	0.19 (±0.38)	2.35 (±0.16)	0.33 (±0.06)	1.97 (±0.12)	0.69 (±0.05)
Average	2.63 (±0.63)	0.52 (±0.11)	2.62 (±0.71)	0.24 (±0.16)	2.66 (±0.57)	0.50 (±0.12)	2.29 (±0.48)	0.64 (±0.10)

Table 1. Predictive performance for RMSDave (±1 std.) per docking protocol, for each of the four splits considered.

We first focus on the comparison between the random and ligand scaffold splits, arguably the most commonly used evaluation procedures in other chemoinformatics ML-based studies. Results for the random split show moderately good results, with some docking protocols showing average correlations over 0.6 (autodock-ga, autodock-lga, gold-asp, gold-chemscore, and gold-plp), suggesting that for those it is easier to predict which ligands will result in a better docking pose. On the other hand, results are significantly worse for the ligand-scaffold-based split for most protocols, which suggests that it is significantly harder for the model to distinguish which compounds outside the training set chemical manifold will result in a better docking result. This conclusion is in line with other works, where random-split-based results were significantly better than those provided by more sophisticated alternatives, such as the ligand-scaffold-based one[13,23,24]. Given that docking is inherently a structure-based problem, we also decided to explore model performance under

different protein-dependent splits. The first protein-based split separates samples into different non-overlapping PFAM clusters (here named protein classes), showing a similar performance to the random split, albeit slightly inferior, suggesting that, while protein information plays a role, wider sampling of ligand chemistry space during training may have a more relevant impact. In the last type of split we evaluated, we sampled for training a percentage of complexes belonging to each protein family (protein classes balanced): our reasoning was that having a more homogeneous sampling of protein space would show a significant performance improvement. Further evaluation was considered to tackle our second research question, the capability of the proposed model to choose the optimal docking protocol given a particular protein–ligand pair. Results can be consulted in Table 2 and Tables S3 and S4 as well as in Figure 1, where we draw similar conclusions as in the protocol-centric evaluation, with the proposed model performing worse in the ligand scaffold split scenario than in the others. Furthermore, in Figure 2, we consider the distribution of the experimental RMSDmin, RMSDave, and nRMSD values had we followed the recommendations of the proposed model, with the intent of investigating whether in fact it produces protocol selections that may improve docking errors. For both RMSDmin and RMSDave values, the protocol with the minimum predicted value was selected, while for nRMSD the maximum was chosen—and then their corresponding experimental values were analyzed. With the exception of the ligand scaffold scenario, the decisions undertaken by the proposed model produce the lowest mean RMSDmin and RMSDave, and the highest nRMSD values compared to the rest of the protocols. Additional significance analyses were performed with a unilateral two-sample Mann–Whitney test. Using a significance level of $\alpha = 0.01$, we can conclude that the procedure here proposed results in significantly lower RMSDave values than the rest of the protocols in all the evaluation scenarios, with the notable exception of gold-goldscore, where no statistical conclusion could be drawn in any direction either in the Molecules 2020, 25, 2487 4 of 12 random, ligand scaffold, and protein clases splits. Interestingly, in the balanced protein split scenario, our approach manages to significantly outperform the aforementioned protocol.

Split Type	Pearson's R	RMSE
random	0.54 ($\pm$0.01)	2.47 ($\pm$0.05)
ligand scaffold	0.47 ($\pm$0.07)	2.58 ($\pm$0.64)
protein classes	0.56 ($\pm$0.04)	2.33 ($\pm$0.21)
protein classes balanced	0.65 ($\pm$0.01)	1.98 ($\pm$0.07)

Table 2. Ligand-centric evaluation (RMSDave, ±1 std.) for the four different proposed split types in this study.

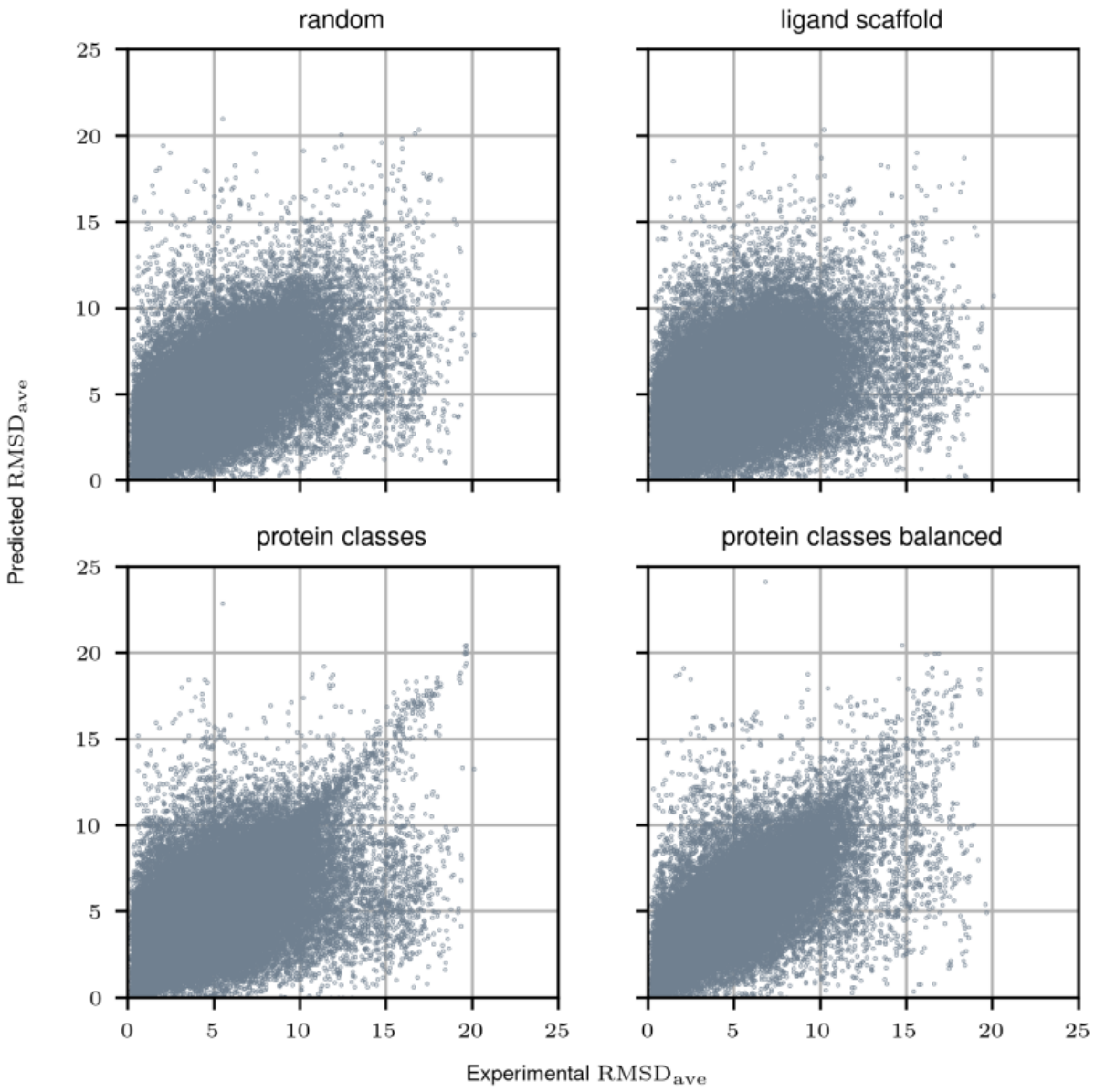


Figure 1. Ligand-centric RMSDave evaluation merging all protocols and for all different types of proposed splits

Overall results suggest that the proposed model provides better suggestions if both ligand chemistry and protein families are not significantly far from the training set manifold. We also investigated disaggregated performance for the 30 most populated PFAM families in our dataset (Figure 3 and Figure S1), to find similar conclusions to the previous evaluations. The results show that the model performs similarly well for the most populated families, and particularly for those splits that more uniformly sample protein space (i.e., the random and protein classes balanced), again highlighting the importance of structure-based models.

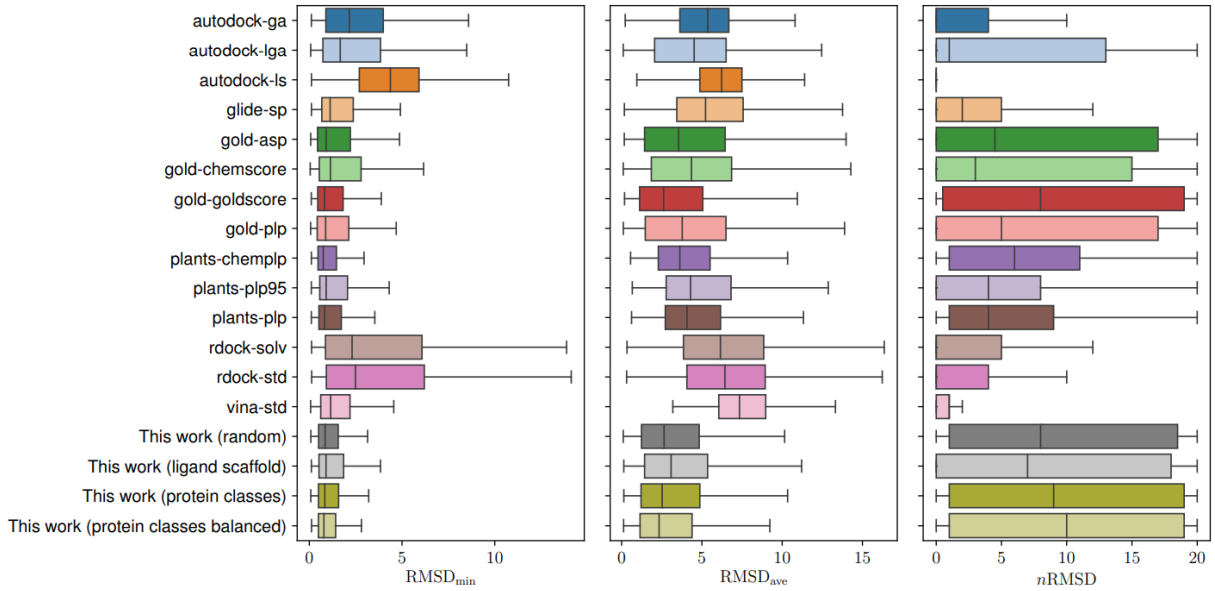


Figure 2. Distribution of RMSDmin, RMSDave, and nRMSD values in a self-docking scenario using the PDBbind v.2017 database of cocrystals, for all the protocols described in Table 3, and the approach proposed in this work under different evaluation scenarios

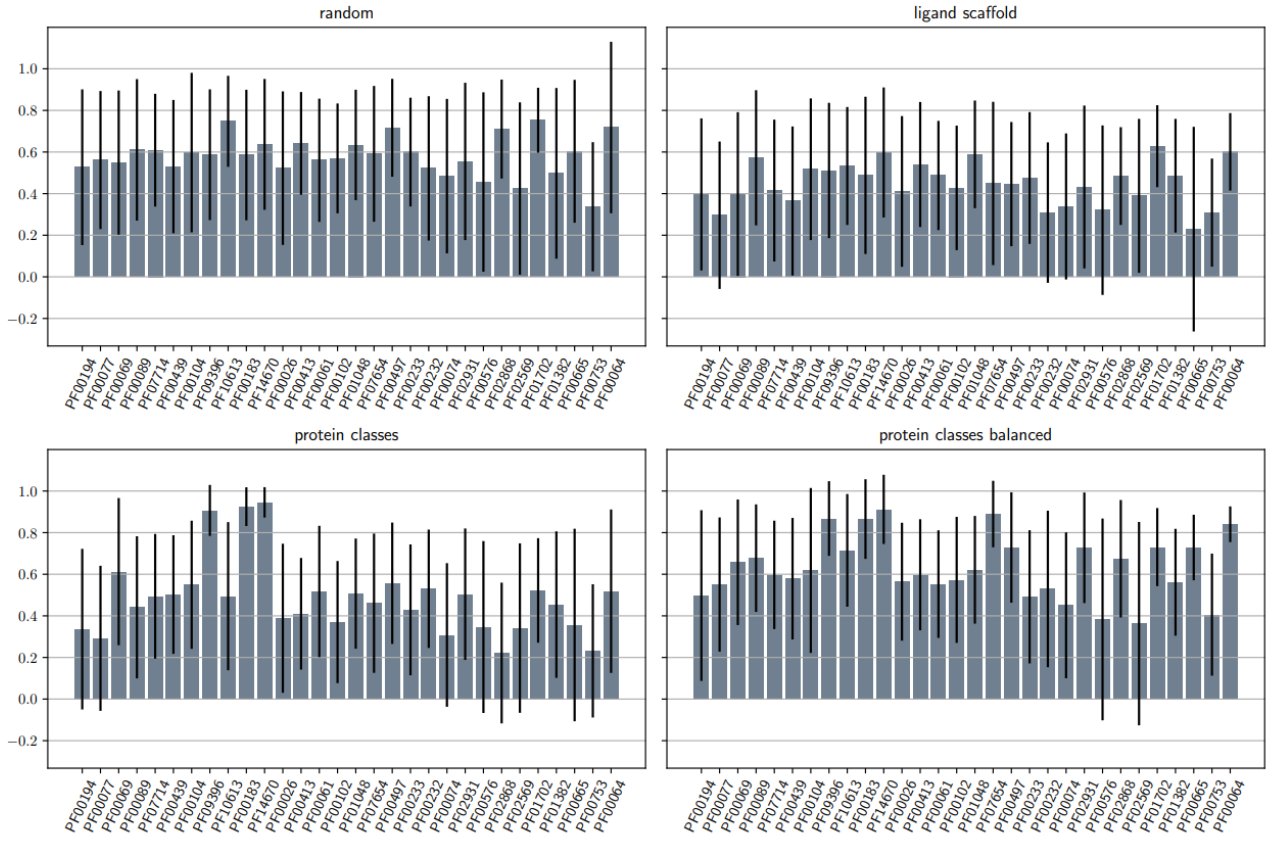


Figure 3. Average Pearson's R correlation coefficient for the RMSDave metric for all types of splits disaggregated into the 30 most populated PFAM families in the PDBbind refined dataset.

3. Materials and Methods

In this section, we first describe the preprocessing procedure for the complexes considered in this study as well as the docking simulation setup. We then describe the two different types of features used and the proposed neural-network architecture. Finally, we discuss technical training details as well as the evaluation procedure undertaken.

3.1. Datasets

The complexes considered for this study were retrieved from the 2017 version of the PDBbind database[21] In particular, we focused on its refined set, that we recently used for a large docking benchmarking campaign[25]. It consists of 4463 protein–ligand complexes, although 294 protein–peptide complexes were excluded as they were not considered in the original DockBench study, resulting in a final dataset of 4169 complexes. Docking settings were selected so as to match as close as possible the default parameters provided by the developers of each protocol for the handling of small organic molecules.

3.2. Complex Preparation.

The proteins in the complexes were prepared according to a protocol previously reported[25]. Structures were processed using an internal workflow written in Scientific Vector Language (SVL), based on the protein preparation tool included in the MOE molecular suite[26]. First, crystal structural issues such as missing atoms and partially solved residues were fixed, hydrogen atoms were added and protonation states for all titrable residues were computed. Finally, solvent molecules and impurities (e.g., co-solvents) were removed. An additional preparation step for the ligands was taken, in which the most favorable ionic state was calculated and partial charges of atoms were assigned. Towards this end, we take advantage of two tools provided by the OpenEye toolkit: fixpKa and molcharge[27]. Finally, ligand geometries were minimized before docking using Open Babel's[28] routing and the MMFF94 force field[29].

3.3. Data Generation

The docking simulation and consequent data generation were performed via the DockBench software (version 1.06), which automates docking simulations and evaluates protocol performances in reproducing ligand conformations in the crystal structure. We included 14 docking protocols from six different software alternatives: AutoDock 4.2.5.1[30], Vina 1.1.2[31], PLANTS 1.2[32], rDOCK[33], Glide

6.5[34], and Gold 5.4.1[35]. For each of the included protocols, we defined the binding site as a sphere of a 15Å radius centered at the center of mass of the co-crystalized ligand, and we generated 20 poses with an RMSD separation of at least 1Å. In the case of both Autodock and Vina, since they do not support spheric site definition, the cube side is scaled to r $(((4\pi/3)*r))^{1/2}$ to maintain comparable volumes with the protocols adopting parallelepiped-shaped cavity definitions, where r is the sphere radius. In addition, in the case of Vina, to guarantee that at least 20 poses were returned, we modified the "maximum energy difference" argument. Description of the protocols, as well as their search algorithms and scoring functions can be found in Table 3. We studied three different and complementary evaluation values for prediction as described in the DockBench suite: the minimum RMSD (RMSDmin), the average RMSD (RMSDave) and the number of poses with an RMSD lower than the resolution of their corresponding crystal structures (nRMSD). Box plots detailing the distribution of these values are available in Figure 2, where we can clearly highlight that some protocols (e.g., gold-asp, gold-goldscore, gold-plp, or glide-sp) display consistent accuracy in many benchmark scenarios, while others (e.g., rdock-solv and autodock-lga) display a higher error variability depending on the input.

Score	Search Algorithm	Scoring Function	Protocol Abbrv.
Autodock 4.2	Local search Lamarckian GA GA	Autodock SF	autodock-ls autodock-lga autodock-ga
Glide 6.5	Glide algorithm	Standard precision	glide-sp
GOLD 5.4.1	GA	ASP Chemscore Goldscore PLP	gold-asp gold-chemscore gold-goldscore gold-plp
PLANTS 1.2	ACO algorithm	ChemPLP PLP PLP95	plants-chemplp plants-plp plants-plp95
rDock 2013.1	GA + MC + Simplex minimization	rDock master SF rDock master SF + desolvation	rdock-std rdock-solv
Vina 1.1.2	MC + BFGS local search	Vina SF	vina-std

GA (Genetic Algorithm), MC (Monte Carlo), BFGS (Broyden–Fletcher–Goldfarb–Shanno), ASP (Astex Statistical Potential), PLP (Pairwise Linear Potential), ACO (Ant Colony Optimization).

Table 3. Docking protocols, search algorithms, and scoring functions considered in this study

3.4. Descriptor Calculation

We take a structure-based approach to represent proteins, deciding to use 3D-voxel descriptors[36,37] that capture the influence of each atom to each voxel of the grid via a pair correlation function n(r) that depends on their euclidean distance r and the Van der Waals radius rvdw of the first:

$$n(r) = 1 - \exp\left(-\left(\frac{r_{\mathrm{vdw}}}{r}\right)^{12}\right). \quad (1)$$

We used the voxelization routines available in the HTMD python framework for molecular modeling[38], which computes eight different pharmacophore-like properties: hydrophobic, aromatic, hydrogen-bond acceptor and donor, positive and negative ionizable, and metallic and total excluded volume. A 24 Å3 array was computed and centered on the center of mass of the co-crystalized ligand, with a resolution of 1 Å. For the ligands, we used Extended Connectivity Fingerprints (ECFP4)[22] with a size of 1024 bits and a radius of 2 bonds as well as a set of 183 physical-chemical descriptors available in the RDKit software[39].

3.5. Neural Network Architecture

A Neural Network (NN) architecture usually takes an array-based input and performs several transformations to obtain another array-based output[40]. Depending on the nature of the input array, some architectures are more appropriate than others. For instance, when the input represents a spatial arrangement (e.g., an image or the 3D-voxel representation described here), a convolutional neural network (CNN) is a typical choice, whereas a fully forward neural network (FNN) is more suitable for a one-dimensional vector, such as a chemical fingerprint[41]. In this study, we designed a specific neural network that takes advantage of both CNN and FNN architectures so as to handle both input types appropriately.

We designed a two-legged neural network that takes protein voxels and ligand fingerprints as inputs separately (Figure 4). Protein voxels pass through five convolutional layers with a rectified linear unit activation function and then they are flattened into a one-dimensional vector. In parallel, ligand descriptors are fed to three consecutive linear layers again with the ReLU activation function. Then, the outputs of both legs are concatenated into a single vector of size 1024. A batch normalization layer[42] is then applied to this hidden protein–ligand representation and three different output linear

layers with ReLU activation function are computed, corresponding to each of the three metrics used Molecules 2020, 25, 2487 8 of 12 by DockBench: RMSDmin, RMSDave and nRMSD. For the first two RMSD-based outputs, we used a standard mean-squared-error loss, while, for nRMSD, we use a Poisson negative log-likelihood loss function, defined by:

$$\ell(y, \hat{y}) = \hat{y} - y \log(\hat{y}) + \log(y!), \tag{2}$$

where y and yˆ are true and predicted values, respectively. We consider the unweighted sum of these three objectives for loss minimization.

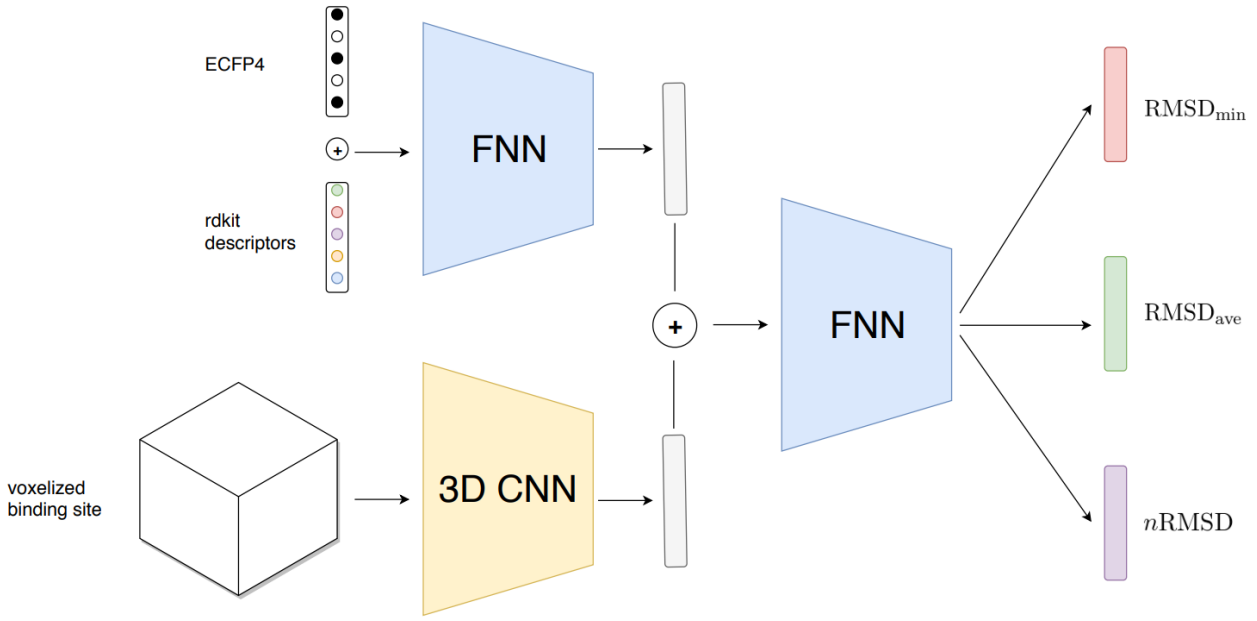


Figure 4. Schema of the proposed architecture in this work. A fully connected neural network handles ECFP4 fingerprints and descriptors computed from RDKit while a 3D-convolutional neural network processes a voxelized representation of the protein binding site. Latent space from both inputs is then concatenated and fed into further fully connected layers that predict the three outputs of interest per docking protocol.

3.6. Training and Validation

We used a k-fold cross-validation scheme (k = 5) to estimate model performance under different split scenarios: for each split, a model is trained on k − 1 non-overlapping subsets and evaluated on the remaining one. Furthermore, we decided to investigate the dependency of the performance with respect to the composition of the chosen subsets. For this reason, we considered four different sampling procedures, each representing a particular application scenario: (i) a completely random

split; (ii) a ligand-scaffold-based split where compounds are grouped according to a k-means clustering of the ligands' ECFP4 fingerprints[43]; (iii) a protein-based split based on non-overlapping PFAM families[44]; and (iv) a balanced protein-class-based split, where we randomly sample 20% of the validation complexes from each PFAM family. In each of the splits, we trained the model for 200 epochs using the Adam optimizer[45] (β1 = 0.99, β2 = 0.999) with a starting learning rate of 10−3 coupled with an exponential learning rate scheduler (γ = 0.95) and a batch size of 32 samples. Data augmentation was performed during training by applying random rotations to the protein pocket coordinates using the geometric center of the ligand as point of reference.

3.7. Implementation and Code Availability

The final production model as well as code to train it and replicate all results and analyses in this paper are openly available on a GitHub repository (github.com/cuzzo87/CNN_DockBench) under an AGPLv3 license. Users can easily use production model scripts to run predictions for their protein–ligand pairs. Our model was implemented in Python using PyTorch (version 1.0)[46] as our main tensor manipulation and automatic differentiation library. While GPU support is not needed Molecules 2020, 25, 2487 9 of 12 for the replication of our work, as well as its production usage, it is strongly recommended, as it can substantially accelerate computations.

4. Conclusions

In this study, we developed a deep-learning-based pipeline for the informed selection of a particular molecular docking protocol, given a protein–ligand pair, and the elucidation of which protein–ligand pairs result in a better pose with a predefined docking algorithm. In conclusion, we believe that we successfully managed to answer both of those research questions. First, we show that it is possible to predict which protein–ligand pairs produce the best poses given a particular docking protocol, although results greatly vary depending on the latter. Interestingly, some protocols (autodock-ga, autodock-lga, gold-asp, and gold-plp) show easier predictability across different data splits than others (plants-plp95, plants-plp, rdock-solv, and rdock-std). We also show that it is certainly possible to predict which docking protocols are better suited for a given protein–ligand pair using the proposed model, although predictive performance greatly depends on the type of the evaluation split taken. Specifically, performance on the random and balanced protein classes splits is undoubtedly superior to that on the ligand scaffold split in most of our evaluations. In addition, we measured the distribution of several relevant docking-related metrics according to the suggestions

of the proposed methodology, to find that these are consistently better than other existing individual protocols under most circumstances. In general, the results presented in this work highlight the usefulness of the presented methodology, but also show that its performance greatly varies depending on the type of evaluation split taken, suggesting that its prospective applicability may differ depending on how close both protein and ligand queries are to the training set manifold. Along those lines, we believe that future interested users in the proposed approach should take these points into consideration before evaluation or re-training of the neural network on their own data. Additionally, while we thoroughly benchmarked our model, all the evaluations presented here are retrospective per se. Future blind structure-based evaluations, such as the ones proposed by the D3R Grand Challenges[47,48,49], would provide excellent opportunities to evaluate approaches similar to the one proposed here prospectively. Methodology-wise, there are several interesting directions for future research regarding neural network architectural design. In particular, it is a well-known issue that 3D-convolutional neural networks are not rotationally equivariant[50] (i.e., the output of the network varies if the coordinates of the protein are rotated), a desirable characteristic when modeling atomistic systems. While this issue is mitigated in the current work through data augmentation, recent approaches such as SE(3) equivariant neural networks[51] bear promise towards solving this issue. On the ligand side, graph convolutions[52] are a family of approaches that are displaying good results in a variety of tasks relevant to drug discovery, such as property prediction[11,12,53] or compound generation[54]. How these approaches would perform in the task proposed here remains a topic for further exploration. Finally, while we firmly believe that future-generation docking protocols will more tightly incorporate machine-learning elements into their pipelines[18,19] (e.g., by the design of more efficient search algorithms or scoring functions[55,56]), we think that the approach proposed in this paper represents a novel research direction that will drive structure-based drug design researchers towards more rational existing docking protocol choices. Hence, with the intent of improving research reproducibility and lowering accessibility barriers, we have open-sourced all evaluation and deployment code as well as trained models related to this work.

References

1. 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. 2004 311* **3**, 935–949 (2004).

2. Sousa, S. F., Fernandes, P. A. & Ramos, M. J. Protein–ligand docking: Current status and future challenges. *Proteins Struct. Funct. Bioinforma.* **65**, 15–26 (2006).

3. 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).

4. 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. *Mol. 2015, Vol. 20, Pages 9977-9993* **20**, 9977–9993 (2015).

5. Ciancetta, A., Cuzzolin, A. & Moro, S. Alternative quality assessment strategy to compare performances of GPCR-ligand docking protocols: The human adenosine A2A receptor as a case study. *J. Chem. Inf. Model.* **54**, 2243–2254 (2014).

6. 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).

7. Salmaso, V., Sturlese, M., Cuzzolin, A. & Moro, S. DockBench as docking selector tool: the lesson learned from D3R Grand Challenge 2015. *J. Comput. Aided. Mol. Des.* **30**, 773–789 (2016).

8. Dahl, G. E., Jaitly, N. & Salakhutdinov, R. Multi-task Neural Networks for QSAR Predictions. (2014).

9. Wallach, I., Dzamba, M. & Heifets, A. AtomNet: A Deep Convolutional Neural Network for Bioactivity Prediction in Structure-based Drug Discovery. (2015).

10. Mayr, A., Klambauer, G., Unterthiner, T. & Hochreiter, S. DeepTox: Toxicity prediction using deep learning. *Front. Environ. Sci.* **3**, 80 (2016).

11. Wu, Z. *et al.* MoleculeNet: a benchmark for molecular machine learning. *Chem. Sci.* **9**, 513–530 (2018).

12. Feinberg, E. N. *et al.* PotentialNet for Molecular Property Prediction. *ACS Cent. Sci.* **4**, 1520–1530 (2018).

13. Jiménez-Luna, J. *et al.* DeltaDelta neural networks for lead optimization of small molecule potency. *Chem. Sci.* **10**, 10911–10918 (2019).

14. Segler, M. H. S., Preuss, M. & Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nat. 2018 5557698* **555**, 604–610 (2018).

15. Gómez-Bombarelli, R. *et al.* Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.* **4**, 268–276 (2018).

16. Skalic, M., Jiménez, J., Sabbadin, D. & De Fabritiis, G. Shape-Based Generative Modeling for de Novo Drug Design. *J. Chem. Inf. Model.* **59**, 1205–1214 (2019).

17. Segler, M. H. S. & Waller, M. P. Neural-Symbolic Machine Learning for Retrosynthesis and Reaction Prediction. *Chem. – A Eur. J.* **23**, 5966–5971 (2017).

18. Ragoza, M., Turner, L. & Koes, D. R. Ligand Pose Optimization with Atomic Grid-Based Convolutional Neural Networks. (2017).

19. Gentile, F. *et al.* Deep Docking - a Deep Learning Approach for Virtual Screening of Big Chemical Datasets. *bioRxiv* 2019.12.15.877316 (2019) doi:10.1101/2019.12.15.877316.

20. Liu, Z. *et al.* PDB-wide collection of binding data: current status of the PDBbind database. *Bioinformatics* **31**, 405–412 (2015).

21. Wang, R., Fang, X., Lu, Y. & Wang, S. The PDBbind Database: Collection of Binding Affinities for Protein−Ligand Complexes with Known Three-Dimensional Structures. *J. Med. Chem.* **47**, 2977–2980 (2004).

22. Rogers, D. & Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **50**, 742–754 (2010).

23. Sheridan, R. P. Time-split cross-validation as a method for estimating the goodness of prospective prediction. *J. Chem. Inf. Model.* **53**, 783–790 (2013).

24. Jiménez, J. *et al.* PathwayMap: Molecular Pathway Association with Self-Normalizing Neural Networks. *J. Chem. Inf. Model.* **59**, 1172–1181 (2019).

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**, 3558 (2019).

26. Vilar, S., Cozza, G. & Moro, S. Medicinal Chemistry and the Molecular Operating Environment (MOE): Application of QSAR and Molecular Docking to Drug Discovery. *Curr. Top. Med. Chem.* **8**, 1555–1572 (2008).

27. QUACPAC 2.1.1.0: OpenEye Scientific Software, Santa Fe, NM. http://www.eyesopen.com.

28. O'Boyle, N. M., Morley, C. & Hutchison, G. R. Pybel: A Python wrapper for the OpenBabel cheminformatics toolkit. *Chem. Cent. J.* **2**, 1–7 (2008).

29. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94 - Halgren - 1996 - Journal of Computational Chemistry - Wiley Online Library. https://onlinelibrary.wiley.com/doi/10.1002/(SICI)1096-987X(199604)17:5/6%3C490::AID-JCC1%3E3.0.CO;2-P.

30. Automated docking of flexible ligands: Applications of autodock - Goodsell - 1996 - Journal of Molecular Recognition - Wiley Online Library. https://onlinelibrary.wiley.com/doi/10.1002/(SICI)1099-1352(199601)9:1%3C1::AID-JMR241%3E3.0.CO;2-6.

31. 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**, 455–461 (2010).

32. Korb, O., Stützle, T. & Exner, T. E. Empirical scoring functions for advanced Protein-Ligand docking with PLANTS. *J. Chem. Inf. Model.* (2009) doi:10.1021/ci800298z.

33. Li, L., Chen, R. & Weng, Z. RDOCK: Refinement of rigid-body protein docking predictions. *Proteins*

Struct. Funct. Bioinforma. **53**, 693–707 (2003).

34. 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).

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

36. Jiménez, J., Doerr, S., Martínez-Rosell, G., Rose, A. S. & De Fabritiis, G. DeepSite: protein-binding site predictor using 3D-convolutional neural networks. *Bioinformatics* **33**, 3036–3042 (2017).

37. Ragoza, M., Hochuli, J., Idrobo, E., Sunseri, J. & Koes, D. R. Protein-Ligand Scoring with Convolutional Neural Networks. *J. Chem. Inf. Model.* **57**, 942–957 (2017).

38. Doerr, S., Harvey, M. J., Noé, F. & De Fabritiis, G. HTMD: High-Throughput Molecular Dynamics for Molecular Discovery. *J. Chem. Theory Comput.* **12**, 1845–1852 (2016).

39. Landrum, G. RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling.

40. Lecun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nat. 2015 5217553* **521**, 436–444 (2015).

41. Goodfellow, I.; Bengio, Y.; Courville, A. Deep Learning. *MIT Press* (2016).

42. Ioffe, S. & Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. (2015).

43. Schultes, S. *et al.* Ligand efficiency as a guide in fragment hit selection and optimization. *Drug Discovery Today: Technologies* vol. 7 e157–e162 (2010).

44. Bateman, A. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **32**, D138–D141 (2004).

45. Kingma, D. P. & Ba, J. L. Adam: A Method for Stochastic Optimization. *3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc.* (2014).

46. Paszke, A. *et al.* PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Adv. Neural Inf. Process. Syst.* **32**, (2019).

47. Gathiaka, S. *et al.* D3R grand challenge 2015: Evaluation of protein–ligand pose and affinity predictions. *J. Comput. Aided. Mol. Des.* **30**, 651–668 (2016).

48. Gaieb, Z. *et al.* D3R Grand Challenge 2: blind prediction of protein–ligand poses, affinity rankings, and relative binding free energies. *J. Comput. Aided. Mol. Des.* **32**, (2018).

49. Gaieb, Z. *et al.* D3R Grand Challenge 3: blind prediction of protein–ligand poses and affinity rankings. *J. Comput. Aided. Mol. Des.* **33**, 1–18 (2019).

50. Cohen, T. S., Geiger, M., Köhler, J. & Welling, M. Spherical CNNs. *6th Int. Conf. Learn. Represent. ICLR 2018 - Conf. Track Proc.* (2018).

51. Thomas, N. *et al.* Tensor field networks: Rotation- and translation-equivariant neural networks for 3D point clouds. (2018).

52. Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O. & Dahl, G. E. Neural Message Passing for Quantum Chemistry. 1263–1272 (2017).

53. Kearnes, S., McCloskey, K., Berndl, M., Pande, V. & Riley, P. Molecular graph convolutions: moving beyond fingerprints. *J. Comput. Aided. Mol. Des.* **30**, 595–608 (2016).

54. Jin, W., Barzilay, R. & Jaakkola, T. Junction Tree Variational Autoencoder for Molecular Graph Generation. 2323–2332 (2018).

55. Morrone, J. A., Weber, J. K., Huynh, T., Luo, H. & Cornell, W. D. Combining docking pose rank and structure with deep learning improves protein−ligand binding mode prediction over a baseline docking approach. *J. Chem. Inf. Model.* **60**, 4170–4179 (2020).

56. Wang, X., Terashi, G., Christoffer, C. W., Zhu, M. & Kihara, D. Protein docking model evaluation by 3D deep convolutional neural networks. *Bioinformatics* **36**, 2113–2118 (2020).

Comparing Fragment Binding Poses Prediction Using HSP90 as a Key Study: When Bound Water Makes the Difference

Giovanni Bolcato, Maicol Bissaro, Mattia Sturlese and Stefano Moro

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).

Abstract:

Fragment-Based Drug Discovery (FBDD) approaches have gained popularity not only in industry but also in academic research institutes. However, the computational prediction of the binding mode adopted by fragment-like molecules within a protein binding site is still a very challenging task. One of the most crucial aspects of fragment binding is related to the large amounts of bound waters in the targeted binding pocket. The binding affinity of fragments may not be sufficient to displace the bound water molecules. In the present work, we confirmed the importance of the bound water molecules in the correct prediction of the fragment binding mode. Moreover, we investigate whether the use of methods based on explicit solvent molecular dynamics simulations can improve the accuracy of fragment posing. The protein chosen for this study is HSP-90.

1. Introduction

Fragment-Based Drug Discovery[1] (FBDD) is an ensemble of approaches used in the early stages of drug candidates identification which consists in the screening of small molecules, typically with a molecular weight below 250-300 Da and a logP value below 3 (these empirical criteria are known as the "rule of three"[2]). FBDD approaches have gained popularity not only in industry, but also in academic research institutes, speeding up the hit-to lead-process and showing an interesting success rate.

Generally, fragment screens lead to the identification of a subset of hit-fragments having an affinity range from µM to mM to the target. However, their identification only represents the beginning of an iterative optimization process to turn a weak fragment into a mature high-affinity lead[3]. One of the challenging aspects of FBDD is the detection of such weak binders commonly achieved by high-sensitivity biophysical techniques, such as isothermal titration calorimetry (ITC), surface plasmon resonance (SPR), thermal shift assay, nuclear magnetic resonance (NMR), and X-ray crystallography (XRC), with only the last two methodologies able to provide structural information.

An alternative to a biophysical method to quickly selecta putative binder from a chemical library is represented by in silico strategies in particular when the target structure is available. The virtual

screening of fragments is typically a challenging task; mostly due to the weak performance of scoring functions used to discriminate native from non-native poses[34].

Most of the scoring functions indeed were mainly developed on high-affinity ligands while fragments are more prone to experience less stable binding states or in certain cases multiple binding modes. For certain targets, the situation can be further complicated by the presence of stable water molecules within the binding site. In fact, for those cases, the understanding of the fragment-target recognition is not only depending on the mere shape or electrostatic and chemical complementarity to its target but also the presence of stable solvent molecules. The presence of stable water molecules can be considered if high-resolution crystallographic structures are available or by computational methodologies investigating the position or the thermodynamic profile of explicit water molecules in protein hot-spots such as 3D-RISM[5], AquaMMapS[6], GIST[7], JAWS[8], SZMAP[9], WaterAlignment[10], WATCLUST[11], WATERDOCK[12], Water FLAP[13], WaterMap[14], and WATsite[15]. Other tools can support the user in selecting those waters that are more stable in high-resolution structures like HINT[16], pyWATER[17], ProBiS H2O[18],WaterScore[19]. It should be noted that the stability of the water network within the protein binding site could be similar to that of weak fragments[20] and, taking this concept to the extreme, a stable water molecule could be considered similar to a very low molecular weight fragment[21]. In this scenario, it is clear that whenever a computational approach is adopted to predict the fragment binding mode an appropriate investigation about the role of the water molecules within the binding site is necessary. From a historical point of view, the first structure-based approach aimed to consider the explicit presence of a water molecule within the binding site was molecular docking[22]. The presence of a stable molecule mediating the ligand interaction may have a great impact on the quality of the pose prediction. Nevertheless, appropriate knowledge about which water molecules to be included is required. The rise of molecular dynamics (MD) strategies which include explicit water offers further alternatives to docking like investigating the stability of a predicted pose along with the monitoring of the water molecules. Novel strategies developed from MD allow investigating the small molecule recognition from the target unbound state with direct observation of the water molecules displaced during the ligand association. One example of these simulations is the supervised molecular dynamics (SuMD)[23] which allows us to follow the molecular association in the nanosecond scale without introducing forces or energetical bias.

The present work aims to compare the performance of different methodologies to face the problem of studying the binding mode of fragments in the challenging scenario of a binding site in which stable water molecules are present and play a pivotal role in their stabilization.

Our comparison embraced four different computational approaches: (i) molecular docking without explicit solvent molecules, (ii) molecular docking with highly conserved water molecules, (iii) molecular docking (without solvent) followed by MD simulation in explicit solvent, and (iv) SuMD starting from the unbound state in explicit solvent. For this study, we have chosen the crystal structures of the loop-in N-terminal domain of Heat.

Shock Protein 90 (N-HSP90) cocrystallized with a low molecular weight ligand (MW < 175 Da). Those crystal structures correspond to the PDB codes: 2JJC, 2WI2, 2YE4, 2YE5, 2YE6, 2YEA, 2YEB, 2YEC, 2YED, 3B24, 4FCP (The structure of the corresponding ligands is reported in Figure 1). We also decided to focus only on the loop-in structures since all the structures of N-HSP90 in apo-form have this specific conformation. HSP90 is a molecular chaperone involved in the maturation of several other proteins and it is a target for the development of chemotherapy agents in many types of cancer. The N-terminal domain of HSP90 binds ATP, essential for the activity of HSP90 [2425]. The choice of N-HSP90 as a case study is based not only because several high-resolution crystal structures are available for this protein, both in the apo form and in complex with low molecular weight ligand (allowing an accurate study of the structural water molecules), but also due to the well-known role that the solvent plays in the binding between the protein and its ligands[26,27,28]. Interestingly, as pointed out in[29], fragments bind HSP90 through a network of conserved water molecules that mediate the interaction with the protein (in particular with Asn51, Ser52, Asp93, and Gly97). The importance of these structural water molecules in the design of HSP90 inhibitors has been proven[28].

2. Results and Discussion

2.1. AquaMMapS Simulation Results

Since our comparison includes also MD-based strategies, we first investigated if the conditions used for the MD simulations and the force field chosen were appropriate to simulate the correct behavior of water. To address this issue, we performed an MD simulation of the HSP90 in the apostate and subjected to AquaMMaps[6] to assess if the regions with predicted stationary water molecules were in agreement with the position of those having low B-factor observed in X-ray structures.

AquaMMaps is a software that, through a posteriori analysis of water molecule trajectories during explicit solvent molecular dynamics simulations can calculate for each space region an occupancy value that expresses the ratio between the time during which a water molecule is located in that region during the dynamics and the total time of the simulation.

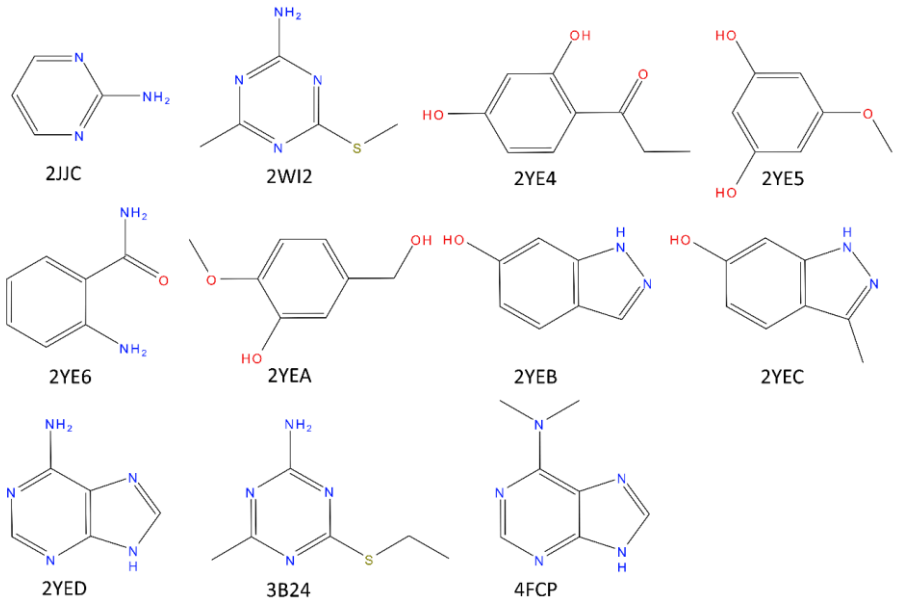


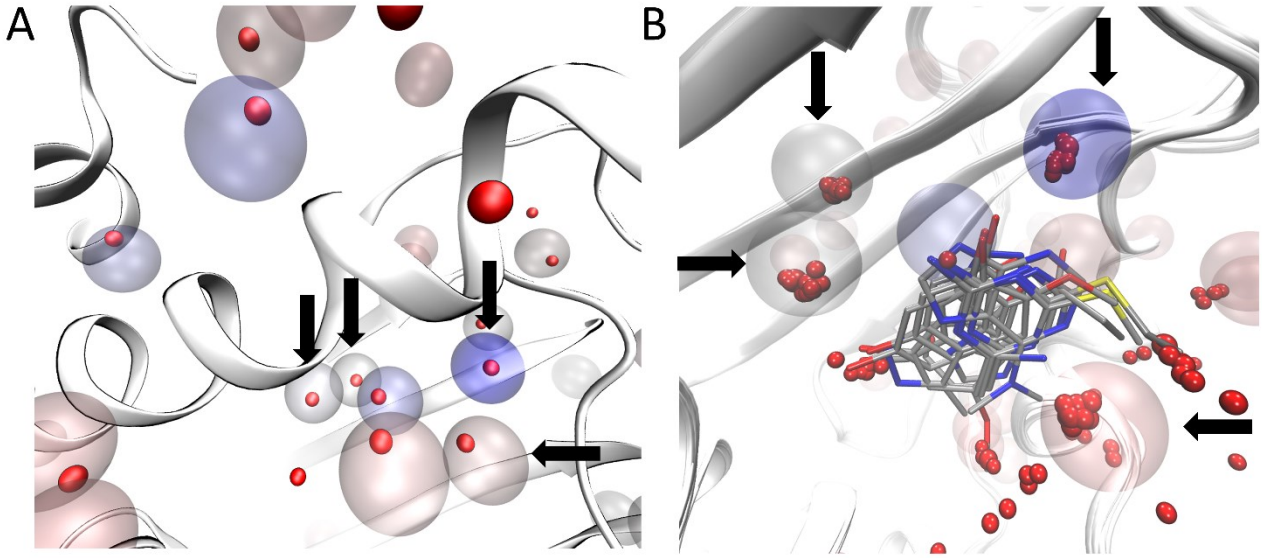
Figure 1. Structure of the crystal ligands bound to N-HSP90 in the structures used for the present work. All the ligands have a molecular weight below 175 Da. Only for complexes 2WIC, 2YE4, and 2YE6 affinity data was reported on literature (2WI2IC_{50} = 350 µM; 2YE4,IC_{50} = 570 µM; 2YE6, IC_{50} = 4000 µM; 3B24, K_d = 42 µM).

The AquaMMapS analysis was performed for five replicates of 100 ns for a total simulation time of 500 ns. These five replicates were merged and submitted to the AquaMMapS analysis. As reported in Figure 2, a good agreement as observed between the AquaMMapS cells with a $\%O_{RMSF}$ value greater than 25 (see Materials and Methods for a detailed explanation of AquaMMapS and its outputs) and the crystallographic waters with a B-factor below 25, especially within the binding site. The crystal structure employed for these simulations is 5J2V, this is one of the several structures of N-HSP90 in apo form.

2.2. Docking, MD Post-Docking,and SuMD Simulation Results

Four different approaches were set up to assess their ability in reproducing the crystallographic structure by evaluating the RMSD: (i) molecular docking without explicit solvent molecules, (ii) molecular docking with highly conserved water molecules, (iii) molecular docking (without solvent) followed by MD in explicit solvent (post-docking MD), and (iv) SuMD starting from the unbound state in fully explicit solvent.

The docking simulations have been performed using GOLD 5.4.1 with the Chemscore scoring function. We identified this protocol by comparing the performance of 17 different docking protocols over all the available ligand-HSP90 complexed reported in the protein data bank (see details supporting information and Figure S1). Starting from this benchmark, we focused our attention on the results for the 11 test cases selected in this work. Among all others, Gold-based protocols outperformed the other ones. Finally, we restricted the comparison on the two scoring functions on which the implementation of water molecules has been reported: goldscore and chemscore[30]. Besides the identical dockbench cumulative score of the two (both scored 5, see table S1 on SI), chemscore showed better results in terms of minimum RMSD averaged on the 11 structures (goldscore:2.64Å; chemscore: 1.64 Å) and hence selected for docking calculation. In the first approach (i), the crystallographic ligands have been docked within the binding site of N-HSP90 without any water molecule. On the contrary, in the second method, the same docking protocol was implemented to include four water molecules placed in the binding site. These four water molecules have been chosen as they appear to be highly conserved in all the structures employed in this work. To address the challenging issue in identifying which water molecules to include in the calculation among those experimentally reported on the test set, we decided to adopt pyWATER



tool[17]. This method identifies stable water molecules by a consensus strategy through a cluster-based approach. For each docking protocol, (i) and (ii), three poses have been generated for each ligand.

Figure 2. Identification of water molecules within the binding site of N-HSP90 (PDB ID: 5J2V). Panel (**A**): The red opaque spheres correspond to the oxygen atom of the crystallographic water with a low B-factor (below 25) while the transparent spheres are the cell predicted by AquaMMapS with a high $\%O_{RMSF}$ value (above 25), the color of the AquaMMapS cell refers to the $\%O_{RMSF}$ value of that cell (the $\%O_{RMSF}$ value increases from red to blue). As it can

be observed, there is a good agreement between the high occupancy AquaMMapS cells and the low B-factor crystallographic water molecules. Panel (**B**): Superposition of the 11 crystals structures used in this work. The Four highly conserved water molecules identified by pyWATER are marked by the black arrows. Each highly conserved water molecule corresponds to a high $\%O_{RMSF}$ value AquaMMapS cell (the cells displayed have all an $\%O_{RMSF}$ value greater than 25).

The post-docking MD strategy started with the best pose obtained from the docking without solvent. The pose was hence equilibrated in a fully explicit solvent simulation box and the system was finally refined by classical MD for 25 ns. This approach aimed to observe if the role of the solvent missing in the docking calculation could be eventually restored by an a posteriori strategy. The advantage of this strategy is that a priori information about the stable water molecules is not required. On the contrary, the drawback of an a posteriori strategy could be eventually the steric hindrance of the ligand placed in the binding site that could obstruct the correct placing of the waters. In light of this hypothesis, the fourth protocol was based on a more demanding strategy simulating the recognition event of a fragment from the unbound state by using SuMD. In this protocol, the fragment was placed 30Å away from its HSP90 binding site. In this way, the binding site is fully solvated by explicit water and the ligand needs to displace them during the recognition. To better understand the SuMD methodology, a recognition trajectory for the complex HSP90–2-pyrimidinamine (PDB ID: 2JJC) is represented in Video-S1. The fragment nicely displaced the solvent in the HSP90 cavity but the regions characterized by stable water molecules are not explored where the water molecules are retained and they mediate the interaction with the fragment and HSP90 in a very similar way to the experimentally solved complex. The comparison between the first pose for the fourprotocols is reported in figure 3. To go into more detailed comparison, in Figure 4 (panel B), the RMSD value of each pose for each ligand is reported, the poses are ordered according to their docking score and to their MMGBSA value (for docking-based and MD-based, respectively). The RMSD values were further used to measure the ability four protocols in geometrically reproducing the experimental complex; in panel A we reported the relationship between the fraction of poses reproduced (below a certain RMSD) to the RMSD. We performed this analysis for both for the first pose and the top-three poses. Ideally, the sooner the profile reaches the top of the fraction of poses reproduced respect to RMSD better the protocol is performing. It should be noted that for the considered complexes we observed that the poses with RMSD lower than 2 Å showed also a correct pattern of interaction with the target, in particular presenting the key interaction with Asp93.

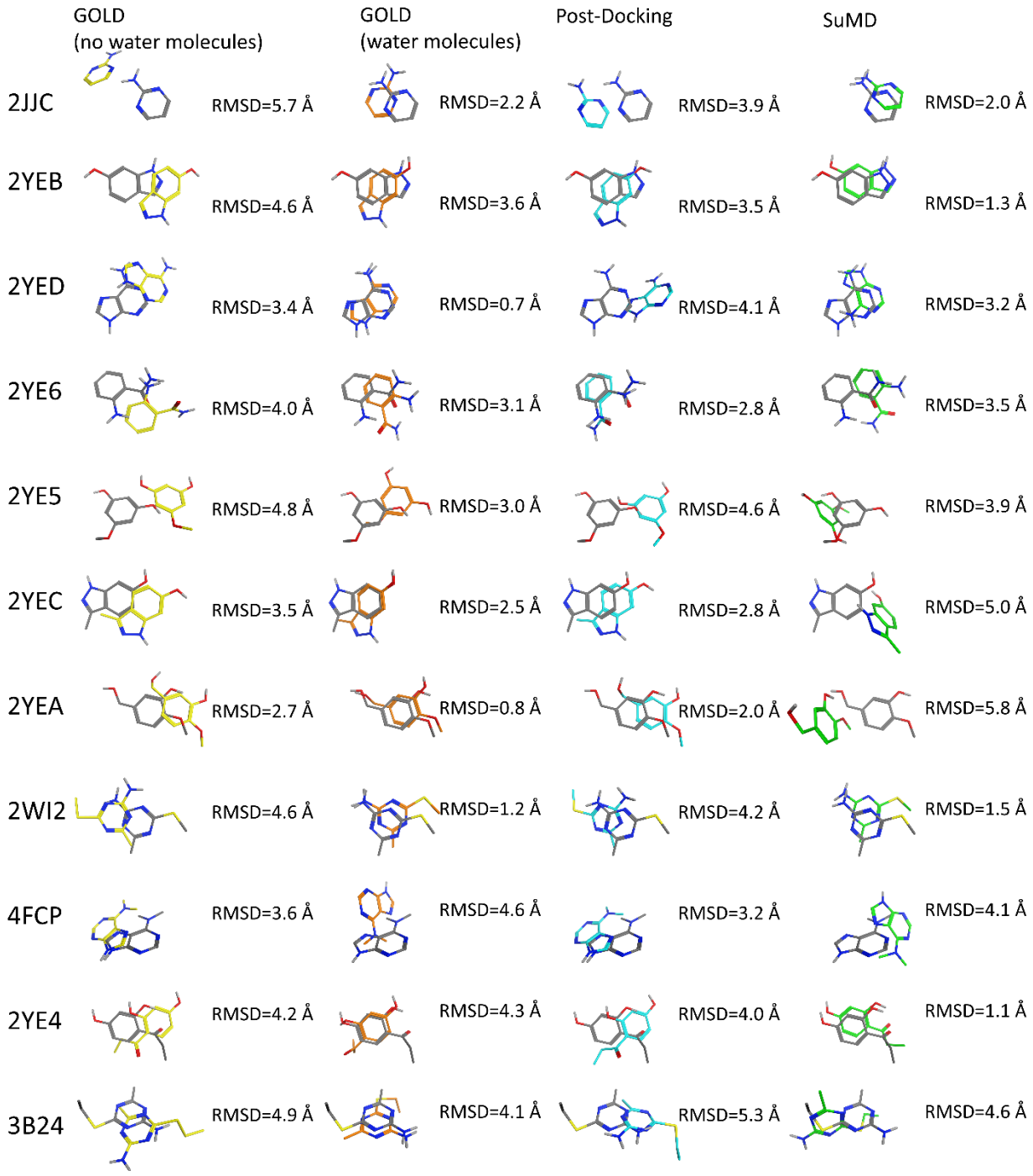


Figure 3. First pose comparison between the different methods (yellow: GOLD without water molecules. Orange: GOLD with water molecules. Blue: Post-Docking. Green: SuMD) for each fragment. In gray is reported the crystallographic pose.

The most clear results is about the performance of the docking protocol without water molecules which results are poor; this can be expected since this is a challenging scenario for a docking protocol not only for the low molecular weight of the ligands but also for the use of the apo form of the protein. The best scoring poses for this protocol report high RMSD values spanning from 2.7 to 5.7 Å; in this range of RMSD the specific fragment-protein molecular interactions observed in the experimentally solved complex are lost. Instead, a dramatic increase in the performance of docking

posing is observed simply retaining the four aforementioned water molecules in the calculation. When the four water molecules are taken into account the performance notably increases, and most importantly the binding mode of the fragments 2YED, 2YEA, 2WI2 is correctly predicted (the RMSD of best pose for these complexes drops below 1.2 Å).

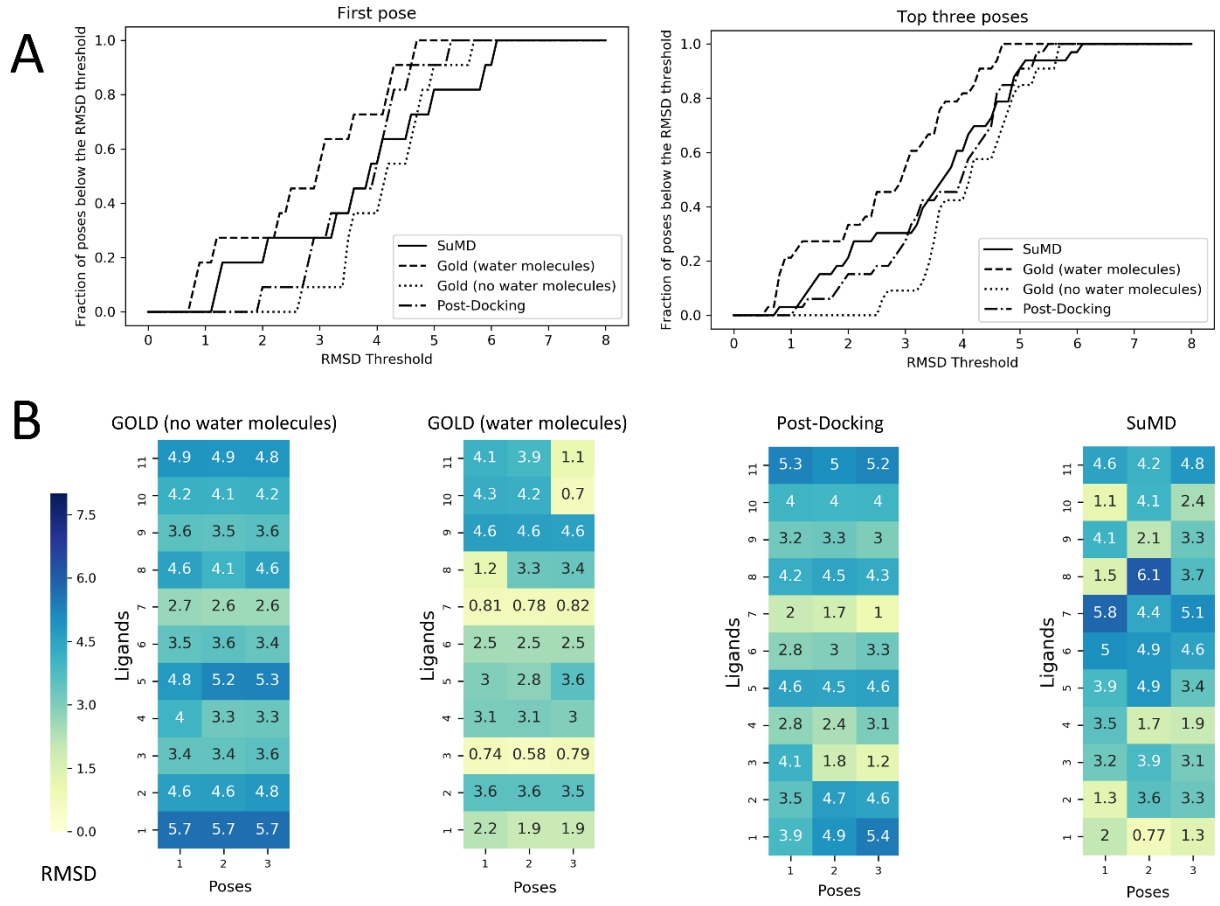


Figure 4. Performance comparison between Docking with and without the consideration of conserved water molecules within the binding site and the two MD based approach: post-Docking and SuMD. These comparisons are made calculating the RMSD values of the predicted poses respect to the crystallographic ones. In Panel (**A**) the fraction of poses below an RMSD threshold are displayed as a function of the threshold itself, this analysis is reported both for the first pose (the one with the best score value for Docking and with the best MMGBSA value for SuMD) and for the top three poses (always the three with the best score or MMGBSA values).In Panel (**B**) the RMSD values of each pose are reported for each ligand as heatmaps. The poses are ordered from left to the right according to their score values for Docking and to their MMGBSA values for SuMD and for post-Docking (so pose 1 has a better score/MMGBSA value in comparison with pose 2 and so on), the heatmap is colored according to the RMSD value which is reported in each grid box. To better compare docking strategies, (i) and (ii), which resulted in three poses, also for MD-based protocol (the post-docking MD and SuMD), three poses for each fragment have been selected by adopting a clustering strategy to select significant representative fragment conformation among the trajectory frames and ranked by the MMGBSA method.

The performance of molecular docking without explicit water molecules is enhanced when coupled with a post-docking refinement of the best scoring pose. The refinement of the docking pose lead to a lower RMSD value for every fragment except for 3B24. It should be noted that despite the

improvement in terms of RMSD only in one case—2YEA—in which the hydrogen bond with the Asp93 is restored, was the correct binding mode recovered. Despite the improvement due to the refinement procedure the results do not reach the quality of the docking protocol with the explicit water molecules.

Among the MD-based protocols, SuMD (iv) outperformed post-docking MD both in terms of RMSD and in the binding mode sampling. In Video-S2 (Supplementary Material) the superposed trajectories of SUMD runs are reported to highlight the association process of the fragments to the fully solvated HSP90 binding site. The performance of SuMD in terms of RMSD seems to be superior also to the molecular docking without water molecules (i) but slightly below to GOLD retaining the four crystallographic water molecules (ii). The poses of 2JJC, 2YEB, 2WI2, 2YE4 have been predicted by SuMD with an RMSD value below 2Å with respect to the crystallographic pose and most importantly for those fragments, the binding mode is correctly predicted by SuMD. It is interesting to note that despite the lower performance respect to the docking with the explicit water molecules (ii) from a geometric point of view (i.e., RMSD), SuMD (iv) is slightly better in reproducing the correct binding mode: four complexes for SuMD while only three complexes GOLD with explicit solvent (ii). A further notable observation is that all the four protocols failed to reproduce the poses for 2YE6, 2YEC, 4FCP, and 3B24. This observation indicates that fragment pose prediction still represents a challenging task, even for advanced structure-based approaches. Also, we observed that the docking differently performed depending on the test case (i.e., the complexes correctly reproduced are different for each protocol), suggesting that it is difficult to have a clear picture in the identification of the most appropriate protocol a priori.

The case of 3B24 and 2WI2 is particularly interesting. The two fragments differ only in a methylene group: 3B24 presents an ethyl group attached to the sulfur atom while 2WI2 has a methyl group. The crystallographic binding mode of the two fragments is very well conserved but the pose prediction performance of the different protocols is quite different. In the case of 2WI2, both molecular docking with water molecules (ii) and SuMD (iv) reproduced the crystallographic pose with an RMSD tolerance of 1.2 Å and 1.5 Å, respectively. On the contrary, in the case of3B24, all protocols fail in the pose prediction with RMSD values over 4 Å. Surprisingly, the affinity reported for 3B34 is particularly high (K_D = 42 μM) for such a small fragment and the resolution of the complex is higher than 2WI2 (1.70 Å and 2.09 Å, respectively).

It is clear that the molecular docking in presence of defined and explicit water molecules outperformed in terms of RMSD the other approaches followed by a more time-demanding SuMD

method that on the contrary did not require a priori information about the stable solvent molecules. We observed that for the correctly predicted case SuMD not only nicely reproduced the bound-state geometries, but the stable hydration sites were also retained (Video-S1). This aspect represents the most notable advantage of SuMD and, in perspective, it could be particularly relevant for all those cases in which a few information is available about the role of the solvent in mediating the ligand-protein interaction. MD-based refinement slightly improves the performance of the docking without water molecules but is not able to balance the performances neither of docking with water molecules nor SuMD.

A further aspect that should be considered in the comparison of those methodologies is the different calculation time required. While for molecular docking a single run can be performed on the order of minutes, MD-based approaches are more demanding and to complete a SuMD simulation usually requires around a dozen hours to complete on a modern GPU card. On the same hardware, a post-docking MD refinement can be easily achieved within a couple of hours. Finally, it should be also considered that the four different protocols present a different level of complexity. Undeniably, molecular docking protocol is easier to set up in comparison to molecular dynamics, and as a consequence, it is suitable for a larger number of users.

3. Materials and Methods

3.1. System Preparation and MD Setup

System preparation has been performed using the Molecular Operating Environment (MOE) suite [31] for what concerns the protein preparation (removing the crystallographic water molecules, ions, and other solvents, selecting of the highest occupancy alternate for each residue, assigning the correct protonation state at pH 7.4 to all atoms). The system preparation for the Molecular Dynamics Simulations has been carried out using AmberTools14[32,33] for what concerns the simulations performed with the ff14SB force field. The protein was explicitly solvated in a water box with the borders placed at a distance of 15 Å from any protein atom, the water model used was TIP3P[34]. The system charge was neutralized to a concentration of 0.154 M using Na^+/Cl^-.

Molecular dynamics simulations have been performed using ACEMD[35]. The system energy was minimized in 500 steps using the conjugate-gradient method, then, during the equilibration stage, two simulations have been done. The first consisted of 0.1 ns of NVT simulation with harmonic positional constraints of 1 kcal mol^{-1} $Å^{-2}$ on each protein atom. The second consisted of 0.5 ns of NPT simulation with harmonic positional constraints of 1 kcal mol^{-1} $Å^{-2}$ only on the α-carbons of the

protein. The simulations consist of 100 ns NVT simulations (temperature 310 K, timestep 2 fs), the last 50 ns of these simulations were submitted to the AquaMMapS analysis.

3.2. AquaMMapS

AquaMMapSis a software aimed to identify hydration sites at the protein-solvent interface in which water molecules show a high-occupancy rate during an MD simulation. Briefly, the tool performs a grid-based analysis of the frequency of occupation of the water molecules. The size is chosen to accommodate one water molecule per cell at most. For each cell of the grid two data are computed: an occupancy value that expresses the ratio between the number of frames during which that cell has been occupied by a water molecule and the total number of frames (%Oall), and an occupancy value that expresses instead the ratio between the number of frames during which that cell has been occupied by a stationary water molecule (i.e., water molecules with an RMSF below1.4Å) and the total number of frames ($\%O_{RMSF}$), so if a cell has an $\%O_{RMSF}$ of 25%, this means that during the simulation this cell has been occupied by a stationary water molecule for 25% of the frames.

3.3. Molecular Docking

A benchmark of 17 different docking protocols over 200 HSP90-ligand x-ray complexes was performed using DockBench[36] to select the most suitable protocol (details are provided in the Supplementary Material). The results were in agreement with previously reported docking studies on HSP90[37] .

GOLD 5.4.1 was used as docking engine and coupled to the scoring function Chemscore. GOLD is a flexible docking protocol that relies on a genetic algorithm for the pose generation while Chemscore is an empirical Scoring Function[38].

The center of the binding site has been defined by superposing all the structure on 2JJC and using the center of mass of its crystallographic ligand. Two docking runs have been performed, one with the inclusion of water molecules within the binding site and one without those water molecules. For each run three poses for each fragment have been generated with an RMSD clustering value of 2 Å. In the docking run with the inclusion of the water molecules, these have always been present (*on* option), the position of the oxygen atom is fixed while the position of the hydrogen is optimized by GOLD (*spin* option).

The clustering analysis on the holo loop-in crystal structures of HSP90 has been performed using the tool PyWATER[17]. PyWATER works aligning a series of protein crystal structures of interest and

performing a clustering analysis on the crystallographic water molecules to identify the most conserved water molecules among the different crystals.

Four highly conserved water molecules have been detected and retained in the protein structure for the Docking calculation. The four water molecules correspond to residues 2078,2082, 2164, and 2166 in the PDB entry 2JJC. The orientation of the water molecules is optimized by GOLD for each case.

3.4. SuMD Simulations, Post-Docking Simulations, and Pose Selection

SuMD[39,40] is a method based on MD aimed to investigate molecular recognition events without energetic biases. Briefly, the algorithm relies on the supervision of the ligand-protein center of mass distance during consecutive small classical MD simulation. The supervision algorithm acts at the end of each small simulation, named SuMD step: if this distance is likely to be shortened during the SuMD step, the simulation is prolonged by a further SuMD step, otherwise, it is stopped, and the simulation is restarted from the previous set of coordinates. In this work, fragments were placed 30 Å away from the protein. Each SuMD step was set to 300 ps The default settings were maintained except the maximum number of consecutive rejected SuMD step that was set to 30. At the end of the SuMD process, the simulation has been extended for 25 ns of classical MD.

The three conformations reported for each fragment have been selected as follows. For each case study, ten SuMD simulations have been performed and only the simulations which led to a binding event have been retained (so only the simulation in which the 30 classical steps of MD below 5 Å has been performed). These trajectories have been aligned on the same reference and merged. The position of the ligand in the merged trajectory has been clustered using Scikit-learn [41]. First, all the sets of coordinates of the ligand (each set is composed of the coordinates of the ligand in a frame of the trajectory) identified as noise by the OPTICS algorithm have been discarded, then all the remaining set of coordinates have been clustered using K-means. The number of clusters has been set to three, in analogy to the three poses obtained in the docking calculation, and to facilitate the comparison with this. For each cluster, the set of coordinates identified as a centroid has been selected as representative of that cluster and then the three centroids obtained for each fragment have been ranked according to their MMGBSA value.

For what concerns the post-docking refinement, three simulations of 25ns for each fragment have been performed on the best pose resulting from the docking calculation with GOLD (without water molecules within the binding site). The simulations have been performed with the same conditions

used for the SuMD simulations. The three trajectories for each fragment have been aligned and merged, then three poses have been extracted as described above for SuMD.

4. Conclusions

The results of the present work emphasize once again the importance of taking into account structural water molecules in the prediction of fragment binding modes. We have focused our investigation on small molecular weight ligands for which molecular docking protocols usually exhibit poorer performance than with classical high-affinity ligands. As expected, the docking simulation carried out without any structural water resulted in poor results; this observation was in agreement with our previous observation that for HSP90 the absence of stable water molecules deteriorates the docking performances even for ligand with stronger affinity [37]. On the contrary, when the conserved water molecules within the binding site are retained or more sophisticated methods like SUMD have used the performances increase dramatically.

For the protein under investigation, N-HSP90, several crystal structures are available, so the identification and the placement of conserved water molecules is an easy task. In this scenario, molecular docking with specific water mediating the interaction remains the best choice both in terms of computational effort and in geometrical terms, but in the worst-case scenario where no information about water molecules is available, like in the case of low-resolution XRC structures or for NMR-based ones, the use of explicit solvent MD simulation can be useful to fill this gap from several points of view. First, one could place stable water molecules using MD-based tools, several tools are already designed with this aim. A further possibility is investigating the ligand recognition process starting from a distant position of the ligand; the role of stable water molecules could be restored since the ligand will need to displace most of the solvent present in the binding site but maintain those water molecules that guarantee a more stable interaction or that are less prone to be displaced. Finally, a slight improvement of pose prediction could be obtained by performing the post-docking refinement of the docking pose, the results are better than those obtained with Docking when no water molecules are considered.

References

1. Ress, D. C., Congreve, M., Murray, C. W. & Carr, R. Fragment-based lead discovery. *Nature Reviews Drug Discovery* (2004) doi:10.2174/978160805201110703010105.

2. Jhoti, H., Williams, G., Rees, D. C. & Murray, C. W. The 'rule of three' for fragment-based drug discovery: Where are we now? *Nature Reviews Drug Discovery* (2013) doi:10.1038/nrd3926-c1.

3. 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.

4. Grove, L. E., Vajda, S. & Kozakov, D. Computational Methods to Support Fragment-based Drug Discovery. in (2016). doi:10.1002/9783527683604.ch09.

5. Sindhikara, D. J. & Hirata, F. Analysis of biomolecular solvation sites by 3D-RISM theory. *J. Phys. Chem. B* **117**, 6718–6723 (2013).

6. 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* (2018) doi:10.1002/cmdc.201700564.

7. Ramsey, S. *et al.* Solvation thermodynamic mapping of molecular surfaces in ambertools: GIST. *J. Comput. Chem.* **37**, 2029–2037 (2016).

8. Michel, J., Tirado-Rives, J. & Jorgensen, W. L. Prediction of the water content in protein binding sites. *J Phys Chem B* **113**, 13337–13346 (2009).

9. Snyder, P. W. *et al.* Mechanism of the hydrophobic effect in the biomolecular recognition of arylsulfonamides by carbonic anhydrase. *Proc. Natl. Acad. Sci.* **108**, 17889–17894 (2011).

10. Brill, D., Giles, J. B., Haworth, I. S. & Nakano, A. WaterAlignment: Identification of displaced water molecules in molecular docking using Jonker and Volgenant shortest path augmentation for linear assignment. *Comput. Phys. Commun.* **244**, 324–328 (2019).

11. Ló Pez, E. D. *et al.* WATCLUST: a tool for improving the design of drugs based on protein-water interactions. doi:10.1093/bioinformatics/btv411.

12. Ross, G. A., Morris, G. M. & Biggin, P. C. Rapid and accurate prediction and scoring of water molecules in protein binding sites. *PLoS One* **7**, (2012).

13. Pastor, M., Cruciani, G. & Watson, K. A. A strategy for the incorporation of water molecules present in a ligand binding site into a three-dimensional quantitative structure - Activity relationship analysis. *J. Med. Chem.* **40**, 4089–4102 (1997).

14. Wang, L., Berne, B. J. & Friesner, R. A. Ligand binding to protein-binding pockets with wet and dry regions. *Proc Natl Acad Sci USA* **108**, 1326–1330 (2011).

15. Hu, B. & Lill, M. A. WATsite: Hydration site prediction program with PyMOL interface. *J. Comput. Chem.* **35**, 1255–1260 (2014).

16. Kellogg, G. E. & Chen, D. L. The importance of being exhaustive. Optimization of bridging structural water molecules and water networks in models of biological systems. *Chem. Biodivers.* **1**, 98–105 (2004).

17. Patel, H., Grüning, B. A., Günther, S. & Merfort, I. PyWATER: a PyMOL plug-in to find conserved water molecules in proteins by clustering. *Bioinformatics* (2014) doi:10.1093/bioinformatics/btu424.

18. Jukič, M., Konc, J., Gobec, S. & Janežič, D. Identification of Conserved Water Sites in Protein Structures for Drug Design. *J. Chem. Inf. Model.* **57**, 3094–3103 (2017).

19. García-Sosa, A. T., Mancera, R. L. & Dean, P. M. WaterScore: A novel method for distinguishing between bound and displaceable water molecules in the crystal structure of the binding site of protein-ligand complexes. *J. Mol. Model.* **9**, 172–182 (2003).

20. Aldeghi, M. *et al.* Large-scale analysis of water stability in bromodomain binding pockets with grand canonical Monte Carlo. *Commun. Chem.* **1**, 1–12 (2018).

21. O'Reilly, M. *et al.* Crystallographic screening using ultra-low-molecular-weight ligands to guide drug design. *Drug Discovery Today* vol. 24 1081–1086 (2019).

22. Villacanas, O., Madurga, S., Giralt, E. & Belda, I. Explicit Treatment of Water Molecules in Protein-Ligand Docking. *Curr. Comput. Aided-Drug Des.* **5**, 145–154 (2009).

23. 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).

24. Hong, D. S. *et al.* Targeting the molecular chaperone heat shock protein 90 (HSP90): Lessons learned and future directions. *Cancer Treatment Reviews* (2013) doi:10.1016/j.ctrv.2012.10.001.

25. Trepel, J., Mollapour, M., Giaccone, G. & Neckers, L. Targeting the dynamic HSP90 complex in cancer. *Nature Reviews Cancer* (2010) doi:10.1038/nrc2887.

26. Davies, N. G. M. *et al.* Targeting conserved water molecules: Design of 4-aryl-5-cyanopyrrolo[2,3-d] pyrimidine Hsp90 inhibitors using fragment-based screening and structure-based optimization. *Bioorganic Med. Chem.* (2012) doi:10.1016/j.bmc.2012.08.050.

27. Haider, K. & Huggins, D. J. Combining solvent thermodynamic profiles with functionality maps of the Hsp90 binding site to predict the displacement of water molecules. *J. Chem. Inf. Model.* (2013) doi:10.1021/ci4003409.

28. Kung, P. P. *et al.* Design strategies to target crystallographic waters applied to the Hsp90 molecular chaperone. *Bioorganic Med. Chem. Lett.* (2011) doi:10.1016/j.bmcl.2011.04.130.

29. Roughley, S. D. & Hubbard, R. E. How well can fragments explore accessed chemical space? A case study from heat shock protein 90. *Journal of Medicinal Chemistry* (2011) doi:10.1021/jm200350g.

30. Verdonk, M. L. *et al.* Modeling water molecules in protein-ligand docking using GOLD. *J. Med. Chem.* (2005) doi:10.1021/jm050543p.

31. Chemical Computing Group ULC, Molecular Operating Environment (MOE), 2019.01. 1010 Sherbrooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, 2019.

32. https://ambermd.org/index.php.

33. 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.

34. 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.

35. 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.

36. 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* (2015) doi:10.3390/molecules20069977.

37. Salmaso, V., Sturlese, M., Cuzzolin, A. & Moro, S. DockBench as docking selector tool: the lesson learned from D3R Grand Challenge 2015. *J. Comput. Aided. Mol. Des.* **30**, 773–789 (2016).

38. 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.

39. 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).

40. 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).

41. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* (2011).

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.

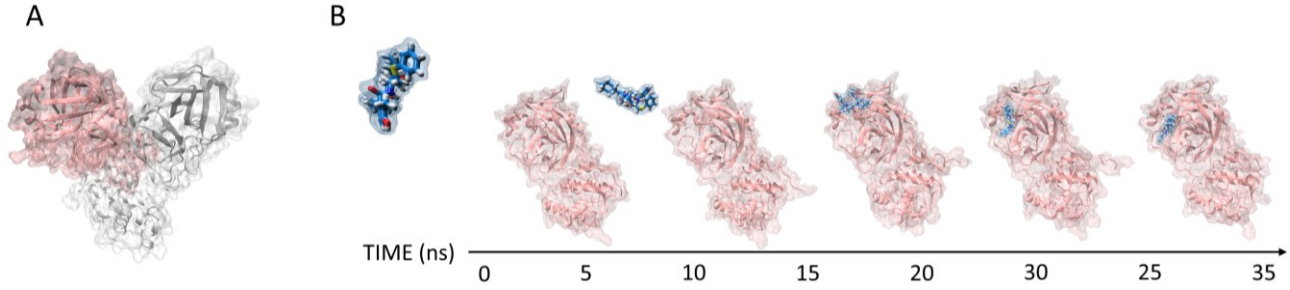


Figure 1. *The crystallographic structure of SARS-CoV-2 C30 Endopeptidase exploited in our computational investigation (PDB ID : 6LU7) is reported in Panel A. The two different monomers composing the homodimeric proteases are depicted using different colors (i.e. pink and white respectively for monomer 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.

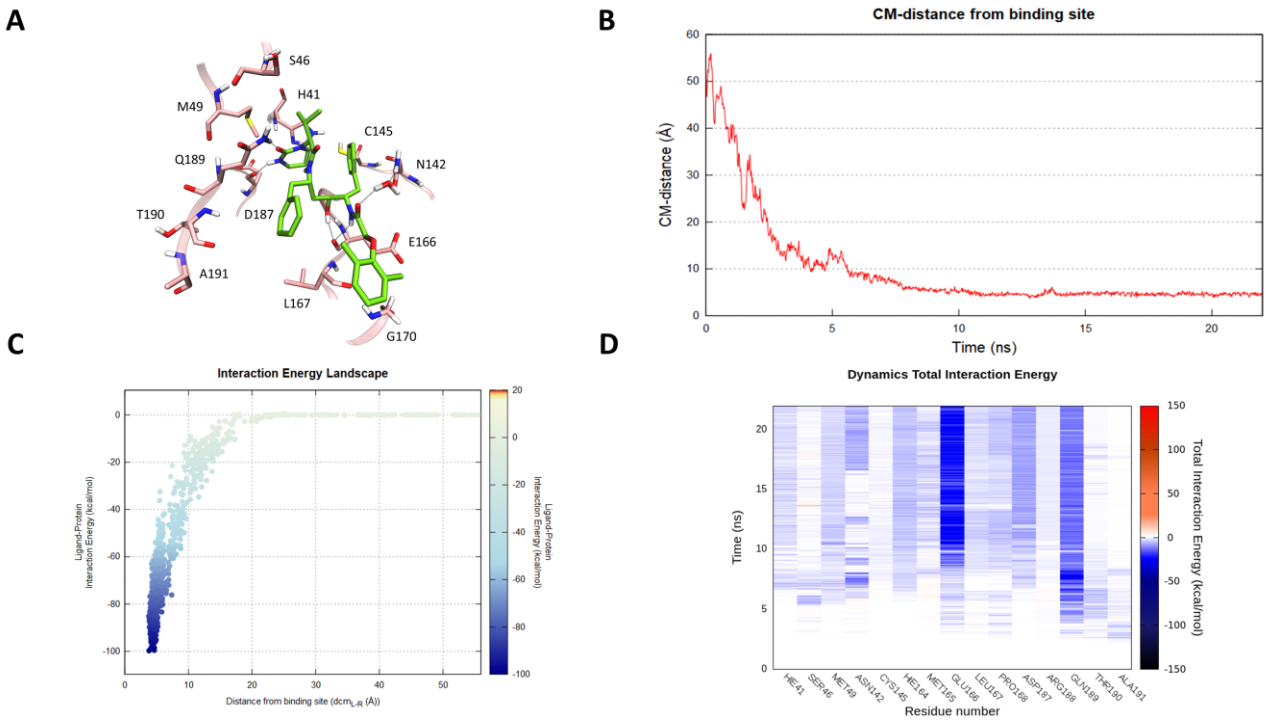


Figure 2: This panel summarizes the recognition pathway of Lopinavir against the SARS-CoV-2 main protease. (A) Lopinavir conformation sampled in the last frame of the SuMD trajectory (green-colored molecule). The residues surrounding the binding site are reported in pink color. (B) Distance between the 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).

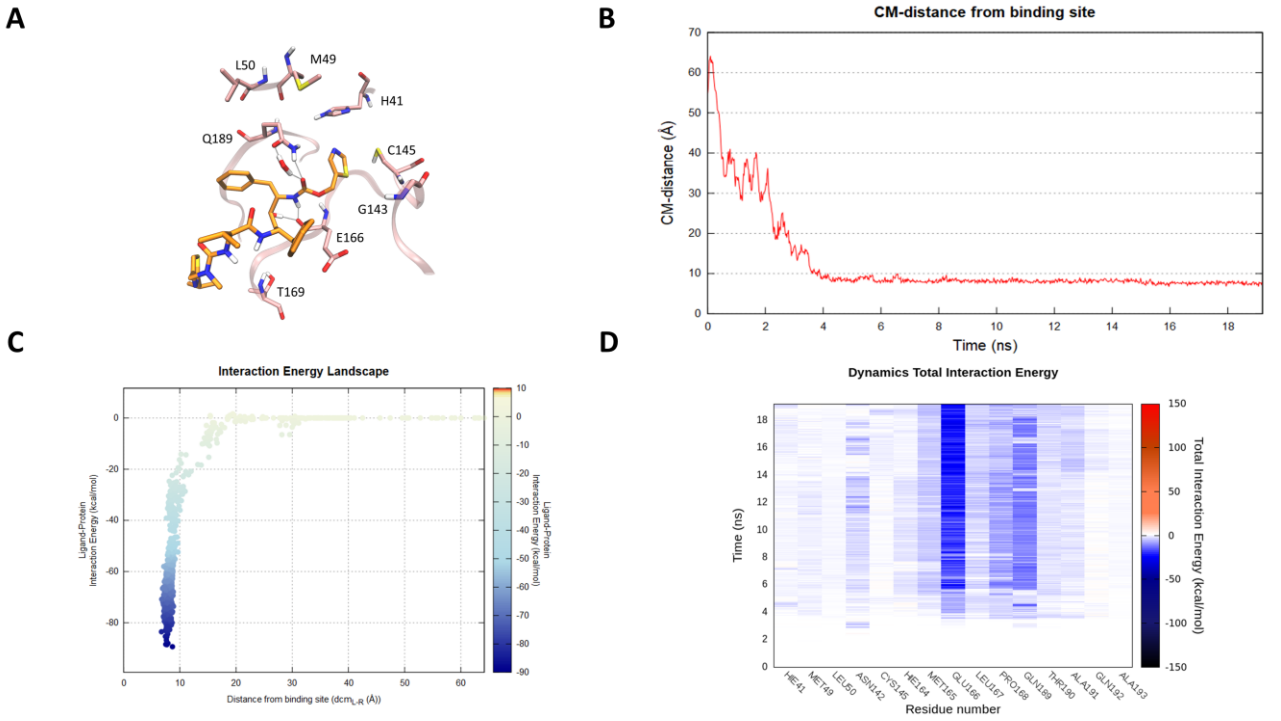
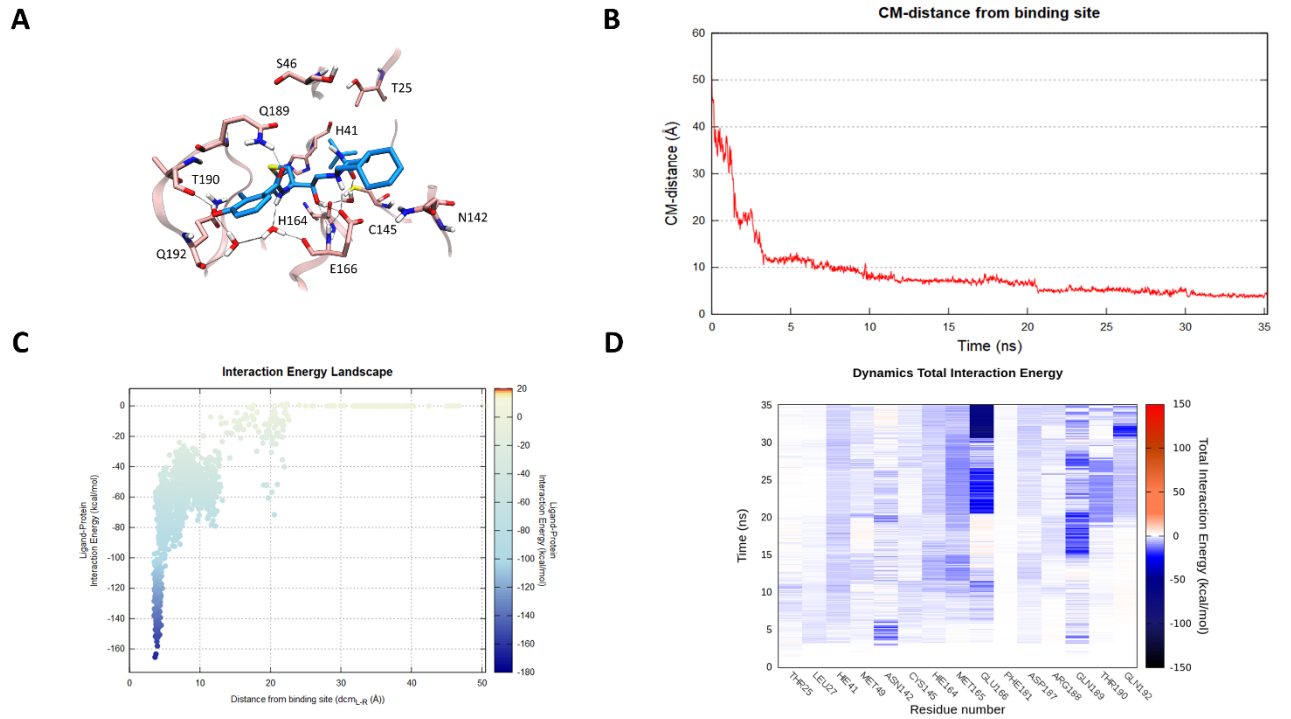


Figure 3. *This panel summarizes the recognition pathway of Ritonavir against the SARS-CoV-2 main protease. (A) Ritonavir conformation sampled in the last frame of the SuMD trajectory (orange-colored molecule). The residues surrounding the binding site are reported in pink color. (B) Distance between 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 diseasesSARS-CoV, MERS-CoV and ultimatelySARS-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-2C30

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 Carlobarostat[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]. 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[10]. 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.

In this perspective computational investigation, we exploited Supervised Molecular Dynamics (SuMD)[11], 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[12,13], 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[14].

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[15]) was used, exploiting an 8 CPU (Intel Xeon E5-1620 3.50 GHz) Linux Workstation. Molecular Dynamics simulations were carried out

with ACEMD[16] (version 3.3.0), which is based upon OpenMM[17] (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[18] 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[19].

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[20] 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[21] 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.

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[14]. 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[12]) showed that this residue plays a key role in linking the dimer interface with the substrate-binding site[22]. 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[23]. Asn142 and Gln189, located on opposite sides at the

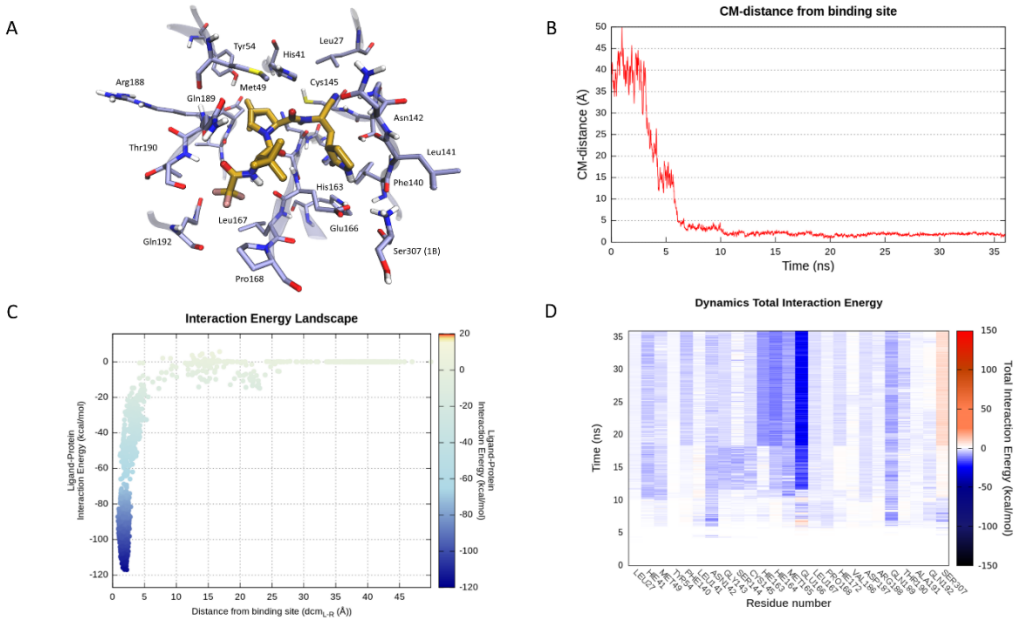


Figure 1. This panel encompasses the recognition pathway between PF-07321332 and the SARS-CoV-2 main protease predicted by SuMD. **(A)** PF-07321332 conformation within the binding site, sampled in the last SuMD trajectory frame (orange). Binding site residues within 4 Å of the ligand are depicted in 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.

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[24]. Moreover, this interaction is conserved across all possible substrate peptide crystal structures, where the interacting group is the sidechain of the Glutamine P1 residue[25].

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.

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).

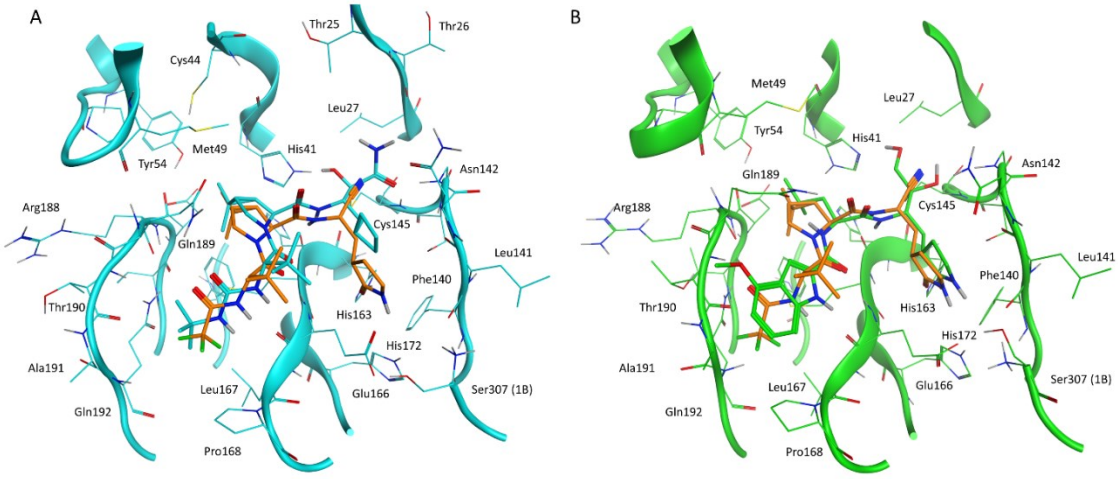


Figure 2. This panel illustrates the similarities between PF-07321332 conformation in the final SuMD trajectory frame and the crystallographic complexes of two structurally related covalent inhibitors of SARS-CoV-2 M^{pro}: Boceprevir and PF-00835231 (active metabolite of PF-07304814). **(A)** superposition 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[26]. 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[27]. 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 MPro 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)[25], 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 (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 (2020).

4. Heymann, D. L. & Shindo, N. COVID-19: what is next for public health? *The Lancet* vol. 395 542–545 (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. *N. Engl. J. Med.* **383**, 1813–1826 (2020).

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

8. Viveiros Rosa, S. G. & Santos, W. C. Clinical trials on drug repositioning for COVID-19 treatment. *Rev. Panam. Salud Publica/Pan Am. J. Public Heal.* **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. Pfizer unveils its oral SARS-CoV-2 inhibitor. https://cen.acs.org/acs-news/acs-meeting-news/Pfizer-unveils-oral-SARS-CoV/99/i13.

11. 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).

12. 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).

13. 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.

14. 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.

15. 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.

16. 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).

17. Eastman, P. *et al.* OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLOS Comput. Biol.* **13**, e1005659 (2017).

18. QUACPAC 2.0.1.2: OpenEye Scientific Software, Santa Fe, NM. https://www.eyesopen.com/.

19. Sadowski, J., Gasteiger, J. & Klebe, G. Comparison of Automatic Three-Dimensional Model Builders Using 639 X-ray Structures. *J. Chem. Inf. Comput. Sci.* **34**, 1000–1008 (1994).

20. 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).

21. Bakan, A., Meireles, L. M. & Bahar, I. ProDy: Protein dynamics inferred from theory and experiments. *Bioinformatics* **27**, 1575–1577 (2011).

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. 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 (2020).

24. 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. *Sci. Rep.* **11**, 7429 (2021).

25. 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.

26. 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).

27. 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 Prepr. Serv. Biol.* 2020.09.12.293498 (2020) doi:10.1101/2020.09.12.293498.

Inspecting the Mechanism of Fragment Hits Binding on SARS-CoV-2 Mpro 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* cmdc.202100156 (2021) doi:10.1002/cmdc.202100156.

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 HTSuMD 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

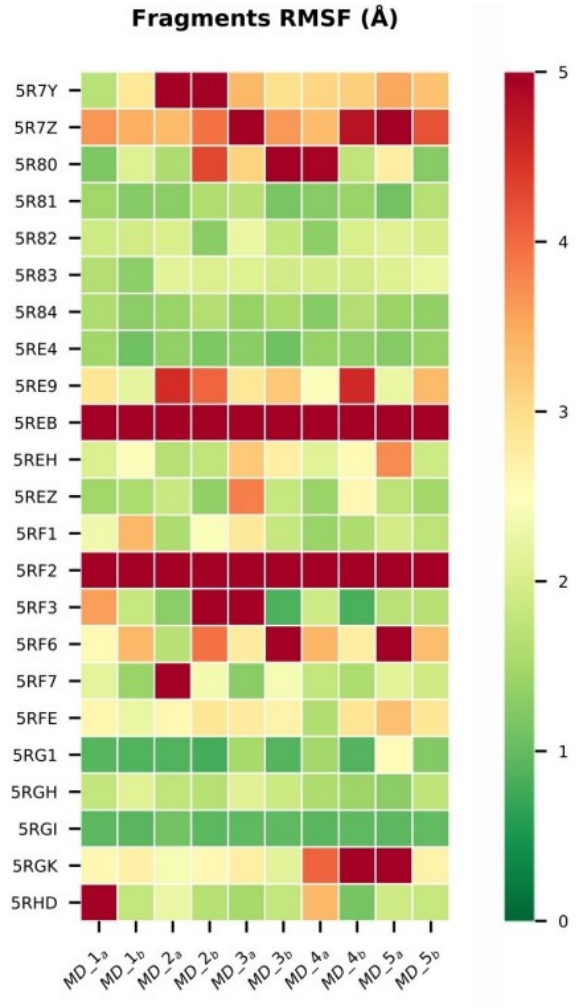
extensively applied also in the FBDD field, providing an atomistic insight on the fragment-receptor binding mechanisms, with a femtosecond temporal resolution[15]. From this perspective, we recently developed HT-SuMD, a computational protocol exploiting supervised MD simulations to perform the screening 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 3- C-like main protease (Mpro) 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 Mpro 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 Mpro 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 Mpro 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, Mpro catalytic cleft is easily accessible from the bulk solvent and hence suitable to SuMD studies, as recently demonstrated for a couple of Mpro inhibitors[23]. The complexity, as well as the plasticity of the Mpro 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 Mprocomplexes, 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 Mpro. To characterize the geometric

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

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. 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 RMSFavg, 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 Mpro (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 Mpro 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. More specifically, for each fragment, the minimum RMSD (RMSDmin) and the average RMSD (RMSDavg) values for the best cluster, i. e. the cluster closer to the crystallographic reference, were reported then comparing the

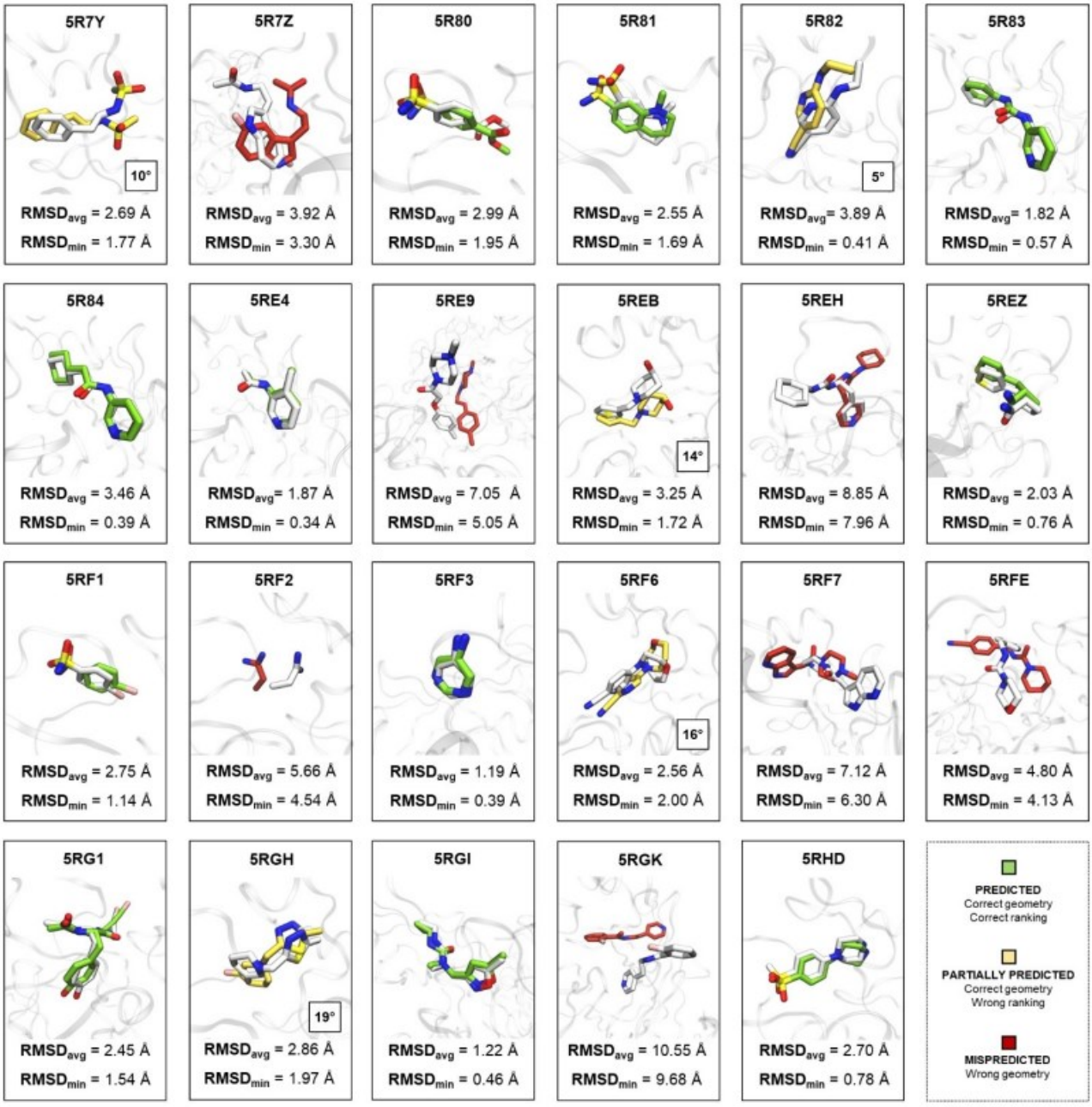


Figure 2. The results of the HT-SuMD posing protocol have been herein summarized. For each of the 23 fragments investigated the cluster of ligand conformations closes to the experimentally solved binding mode was reported, measuring the accuracy of the prediction through the RMSDavg and RMSDmin values 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.

predicted binding mode with the experimental one. The fragment posing exercise was considered correctly achieved if the RMSDmin 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 RMSDmin 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.

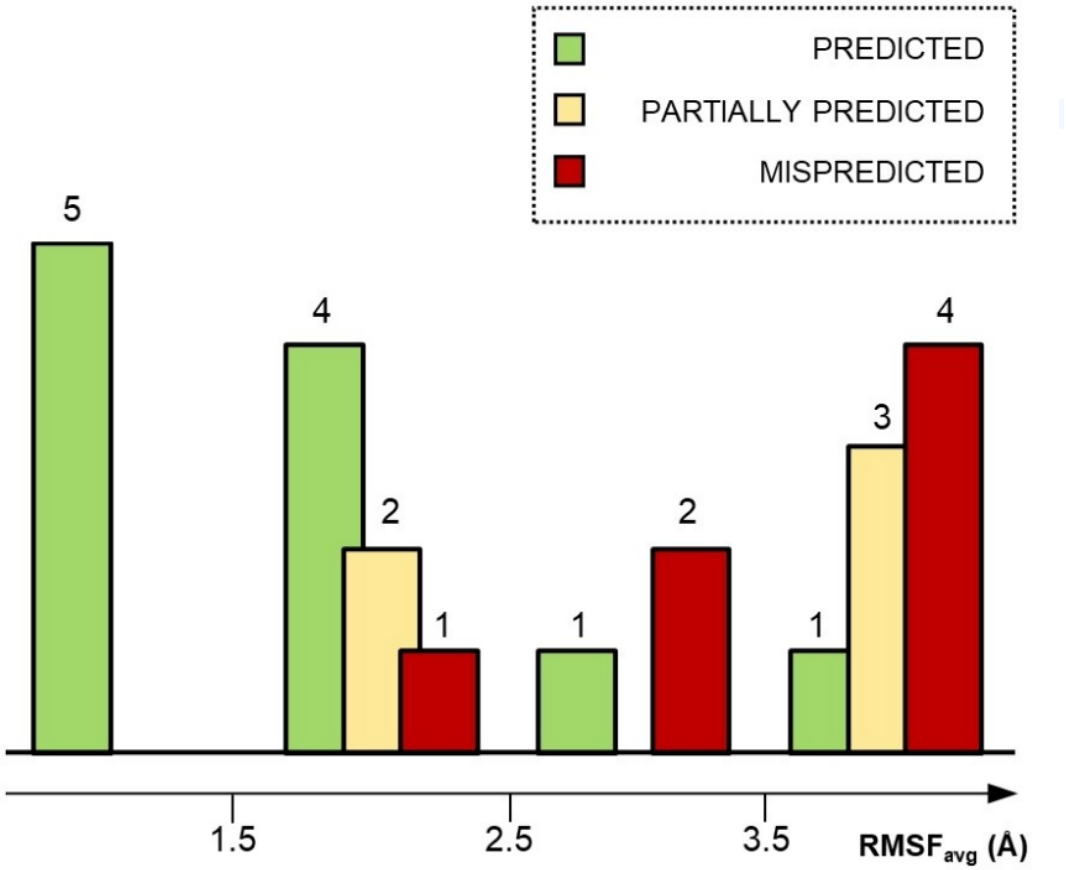


Figure 3. HT-SuMD predictions have been analyzed and related to the average fragment pose stability values (RMSFavg) showed by each original crystallographic complex during the MD-based characterization study

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]. HT-SuMD outcomes have therefore been compared, as reported in Figure 3, with the average values of atomic coordinates fluctuation (i. e. RMSFavg) 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 RMSFavg 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.

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 Mpro fragments previously investigated through HT-SuMD, the binding mode with the lowest MM/ GBSA score was blindly selected, regardless of whether or not 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 RMSFSuMD 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 RMSFSuMD, 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.

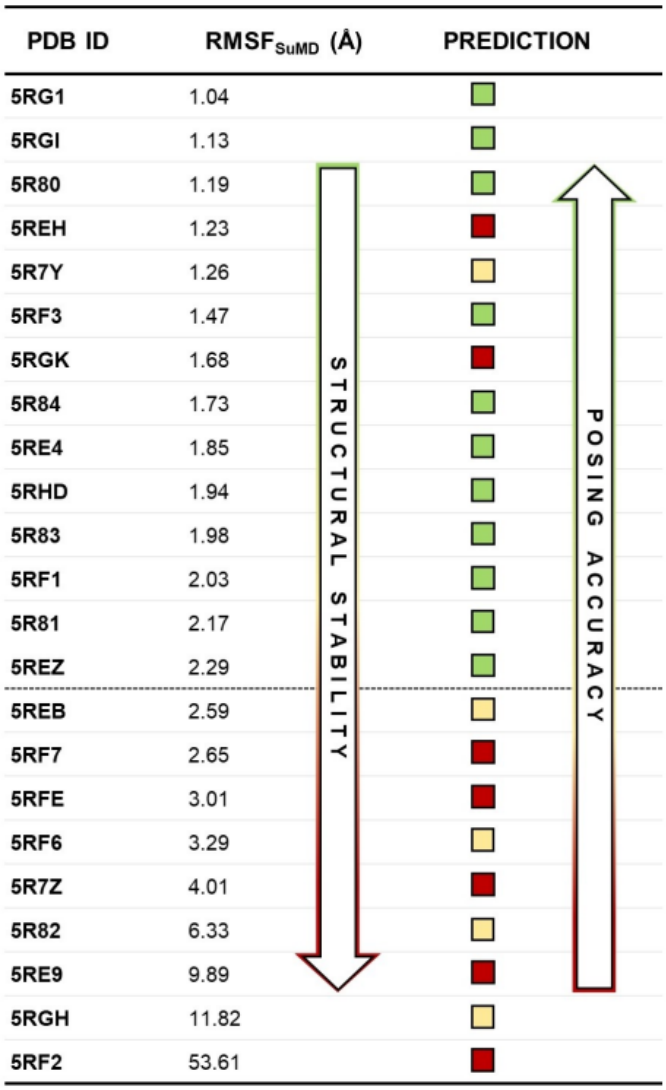


Figure 4. HT-SuMD predicted binding modes (i. e. the cluster of fragments conformations characterized by the lowest MM/GBSA value) have undergone an MD-based refinement step. The fragment poses stability of each prediction, measured as the RMSFSuMD, has been exploited to rank HT-SuMD results, allowing 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 ofHT-SuMD posing prediction

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 SARSCoV-2 Mpro, 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.

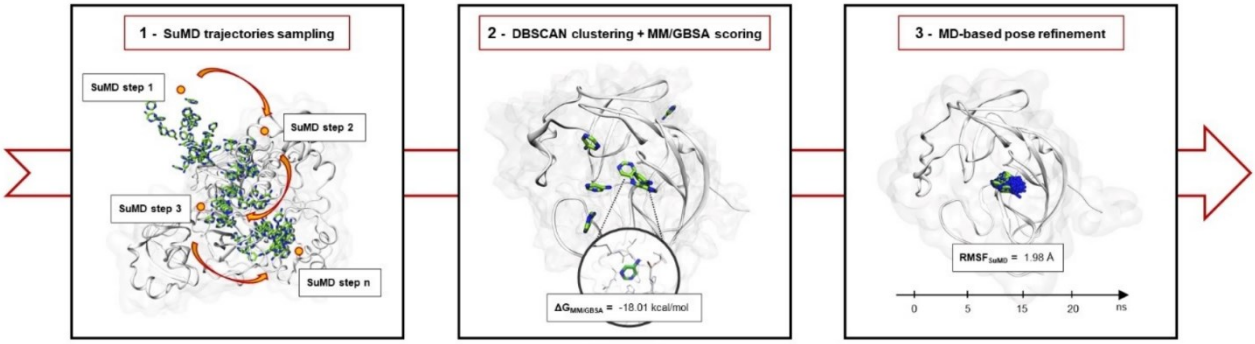


Figure 5. HT-SuMD protocol for the posing of fragments mainly consists in three operative steps, that are respectively summarized in this graphical workflow. In detail, supervised MD simulations are exploited to sample multiple binding trajectories for all the fragments analyzed (1), then DBSCAN clustering 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 RMSFSuMD 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 RMSFavg. 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 RMSFSuMD, 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 RMSFSuMD.

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).

Shedding light on the molecular recognition of sub-kilodalton macrocyclic peptides on thrombin by Supervised Molecular Dynamics

Mahdi Hassankalhori, **Giovanni Bolcato**, Maicol Bissaro, Mattia Sturlese, Stefano Moro

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, (2021).

Abstract

Macrocycles are attractive structures for drug development due to their favorable structural features, potential in binding to targets with flat featureless surfaces, as well as their ability to disrupt protein-protein interactions. Moreover, large novel highly diverse libraries of low molecular weight macrocycles with therapeutically favorable characteristics have been recently established. Considering the mentioned facts, having a validated fast, and accurate computational protocol for studying the molecular recognition and binding mode of this interesting new class of macrocyclic peptides deemed to be helpful as well as insightful in the quest of accelerating drug discovery. To that end, the ability of the in-house supervised molecular dynamics protocol called SuMD in the reproduction of X-ray crystallography final binding state of a macrocyclic non-canonical tetrapeptide—from a novel library of 8988 sub-kilodalton macrocyclic peptides—in thrombin active site was successfully validated. A comparable binding mode with the minimum root-mean-square deviation (RMSD) of 1.4 Å at simulation time point 71.6 nanoseconds was achieved. This method validation study extended the application domain of SuMD sampling method for computationally cheap, fast but accurate, and insightful macrocycle-protein molecular recognition studies.

1 Introduction

The ever-increasing expeditious development of computer hardware, software, and algorithms have positively contributed to many domains of research such as drug design. The developed computational methods, namely molecular docking, and molecular dynamics (MD) simulations, to name but two, greatly reduce the time and cost of drug development, in a way that *in silico* modeling tools are highly utilized in the research ambit of drug discovery[1,2,3]. Particularly, the investigation of binding mode, following the steps of varied ligand-target recognition pathways, as well as exploring their interactions have been claimed to be the area of impressive application of MD computational protocols[3].

Molecular dynamics simulations are considered an endorsed computational method in which by integrating the numerical solution of the Newton equation of motion, the time-dependent evolution of a molecular system can be revealed and described. However, obtaining a complete molecular recognition trajectory leading to binding, from the unbound to the bound state, is a rare event, and to capture moments of importance, therapeutically speaking, via free classical molecular dynamics approach requires a long microsecond timescale and therefore massive computing resources even with the novel GPU-based protocols[4,5,6].

Our in-house alternative MD approach, compared to the classical method, named supervised molecular dynamics (SuMD), improves the efficiency of sampling a binding event and decreases the simulation time from a microsecond (μs) to a nanosecond (ns) timescale[7]. To do that, it applies a tabu-like algorithm to monitor the distance between the ligand center of mass and the target binding site center of mass during a short classical MD simulation; only productive simulations in terms of reducing this distance are considered productive. Despite the exploration of the recognition event, SuMDhas been previously proved to be able to reproduce the experimental bound state of several various kinds of complexes with great geometric accuracy. Its already validated application domain covers the molecular recognition simulation of small molecules, natural linear peptides, most classic peptidomimetics, and nucleic acidss[8].

Among different classes of compounds, macrocycles are attractive structures for drug development, due to their potential in binding to “undruggable with canonical small molecules or proteins”[9]. Macrocyclic peptides represent an efflorescing class of molecules potentially targeting numerous disease-related protein targets otherwise intractable via established pharmacological approaches[10]. Several remarkable characteristics can be considered for this class of molecules. First, compared to linear peptides, they are relatively stable and less prone to protease degradation. The cyclization also confers advantages such as having a compromised state between a flexible and preorganized structure required for dynamic interactions with protein targets with a conformational bias; a reduced binding entropy cost can be imagined compared to their linear counterparts[11]. However, it is worth to mention that due to the reduced accessible conformational states, shifting the structure—upon macrocyclization—toward states that can anticipate bioactivity for a specific target binding site is consequential, because otherwise the non-bioactive conformation stabilization can slow down the binding. Therefore, identification of highly populated conformations of macrocycles is of significance when it comes to drug design[12]. Moreover, it has been shown that macrocyclic peptides are capable of selectively bind to relatively shallow, flat, and featureless protein surfaces

often involved in clinically important protein−protein interactions (PPI), in a fashion similar to antibody-based therapeutics and conversely to small molecules which generally need a pocket to bind[13,14]. Furthermore, thanks to their amino acid composition, a low innate toxicity is anticipated which is of advantage as therapeutic modalities. Being synthetically accessible makes possible lead optimization attempts and altering biophysical properties in terms of binding affinity and specificity, proteolytic stability, and/or solubility improvement for a particular purpose. A variety of macrocyclization reactions has been devised over the years and now different topologies can be easily synthetically available[14]. However, this interesting class of molecules has been underrepresented in numbers and diversity in the available libraries[9]. In recent years, innovative approaches evolved for further development of cyclic peptides, like generating and screening large combinatorial cyclic peptide libraries using *in vitro* display. These attempts have increased the availability and potential screening of ten to hundreds of thousands up to 1 trillion compounds or more highly diverse macrocycles with extraordinary target affinity, selectivity, and bioactivity[13,9,10,15]. In a recent research project of Kale et al., via novel thiol-to-amine cyclization reactions, they introduced a strategy that enables the generation of high yield purification-free large library of diverse macrocycles to screen for various targets in an efficient, relatively small-effort manner. Generating a library containing 8988 macrocycles of sub-kilodalton molecular weight (ideal for addressing the lingering challenge of macrocycles) and screening of this library against thrombin and other homologous targets identified a potent selective thrombin inhibitor called P2 (Ki = 42 ± 5 nM)[9].

Given the emerged perspective stemming from all referred above, having a reasonably fast and accurate computational method like SuMD for studying the molecular recognition pathway and reproduction of experimentally comparable binding mode of this promising macrocyclic class of peptides is deemed significant. With that intent, through this study, the ability of the SuMD protocol in the reproduction of X-ray crystallography final bound state of the candidate macrocyclic peptide P2 as a potent thrombin inhibitor was evaluated.

P2 is a tetra-peptide composed of “Glycine”-“L-beta-homoproline”-“Arginine”-“Cysteine” cyclized with a linker of di-bromomethyl benzyl and an N-(2-(hydroxymethyl)benzyl) substituent coming from an additional reaction of the linker (Figure 1). P2 is proved to be a highly selective inhibitor for thrombin with a snug fit of the specific backbone to the target while did not show any considerable inhibition for other homologous structurally and functionally similar proteases such as activated protein C (APC), tissue plasminogen activator (tPA), to name but two[9]. A representation of thrombin

in complex with P2 structure is shown in Figure 1. During library screening, another macrocycle called P1, with a similar structure as P2 and merely lacking hydroxymethyl-benzyl moiety showed three orders lower inhibition constant than P2[9]. Given that and the fact of not being available any experimental reported binding state for P1, the idea to try simulating a probable binding mode of P1 in addition and possibly hypothesizing the inhibition potency difference through our *in silico* studies was emerged.

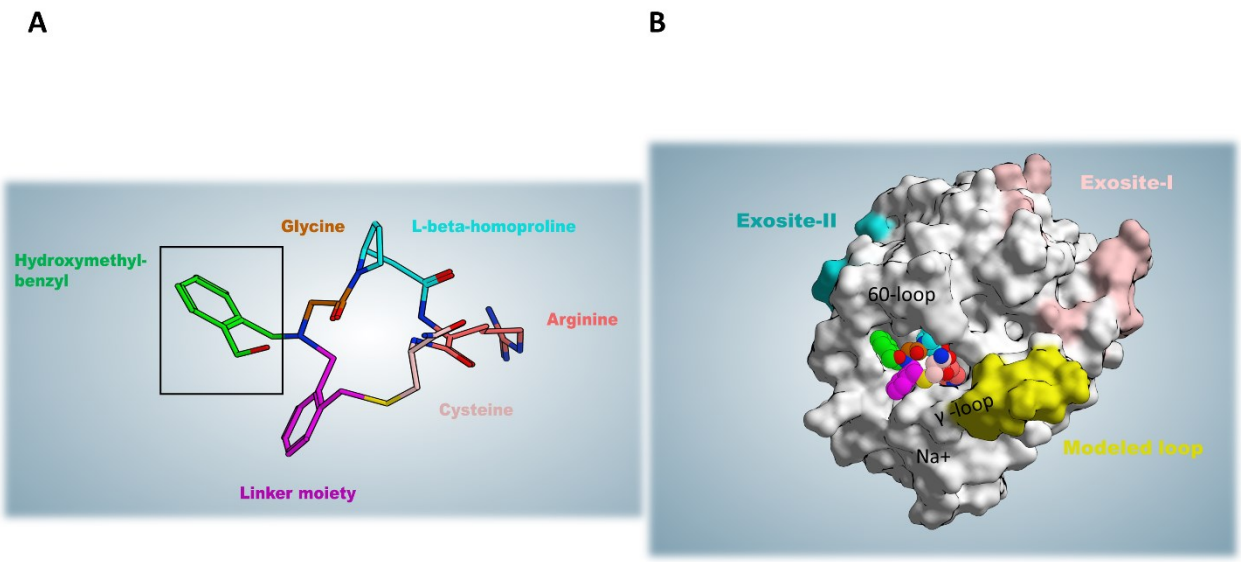


Figure 1. **(A)** The structure of P2 is shown; the hydroxymethyl-benzyl moiety(in green) that is lacking in the P1 macrocycle is highlighted by a black frame. **(B)** Thrombin in complex with P2; Thrombin structural determinants for its function and client recognition are also reported.

The protein target in this study, thrombin, is a typical trypsin-like serine protease and the final generated protease during blood coagulation cascade. It is worth raising the point that distinct structural features are present in this single protease for the recognition ability of different substrates in a specific manner[16,17]. As reported in figure 1, the walls of a deep active site cleft—often referred to as canyon—are formed by the 2 insertion loops known as the 60-loop and γ-loop. The upper 60-loop, is a rigid, hydrophobic cap over the active site, while the more hydrophilic and flexible γ-loop is situated at the downside of the cleft. A constricted access to the catalytic site of thrombin is provided only to proteins with long, flexible substrate loops[17]. The substrate recognition within the active site of thrombin occurs thanks to favorable interactions between the P1 residue (according to the Schechter and Berger nomenclature of amino acid residues around the substrate scissile bond[18]) and the deep acidic S1 pocket (Asp189, Ser190, Gly219), as well as presence of hydrophobic/aromatic residues N-terminal to P1 occupying S2 pocket (Tyr60A and Trp60D as the main residues), and S3 (the aryl-binding pocket composed of Trp215, Leu99, Ile174)[19,17]. Apart from

the active site, three other regions are involved in the diverse specific recognition of different substrates. There are two electropositive exosites, termed anion binding exosites (ABE), and a sodium-binding site. All-natural thrombin substrate directly or via cofactor mediation establish contacts with at least one exosite and usually both; this represents the prerequisite event to form initial stable complex conformation needed for the peptide bond cleavage[17,16,20]. Sodium-binding site, 15 Å away from the catalytic triad (His57, Asp102, Ser195), with Na^+ coordinated to the main chain oxygens of Arg221a and Lys224 and four conserved water molecules, is considered as another allosteric activity modulator site of this protease, helping the maintenance of the hemostatic balance. Upon binding to sodium, thrombin shifts toward a conformation known as 'fast conformation' able to cleave all procoagulant substrates such as fibrinogen and protease-activated receptors more readily. On the other hand, in the Na-unbound 'slow' state, the protein C anticoagulant pathway is preferentially activated. Under physiologic conditions, the 140 mmol/L Na^+ concentration in the blood would not saturate the site, and a present 2:3 ratio between slow: fast accounts for optimal allosteric regulation of anticoagulant: procoagulant activities and hemostasis[6,16,21].

2 MATERIALS AND METHODS

2.1 Computational Study Infrastructure

This project was carried out a hybrid GPU-CPU Linux cluster of 280CPU-cores and 30 NVIDIA graphic cards.

2.2 Structure Preparation

To begin with the simulation, the three-dimensional coordinates of the crystal structure of thrombin bound to P2 macrocycle (PDB ID: 6GWE) were retrieved from RCSB Protein Data Bank (PDB) with a resolution of 2.3 Å[9]. Then, using MOE suite version 2019.01[22], the structure was checked and modeled (via loop modeler plugin) for the missing loop, 3D protonated and energy minimized regarding the energy of the added hydrogens and their positions. For this study one of each unique chain which is chain A with 257 residues and chain B with 30 residues in their sequence, in addition to the sodium ion bound to the chain A sodium binding loop was kept. The modeled 8-residue missing loop between Glu146 and Gly150 amino acid sequence comprised of TWTANVGK.

2.3 Solvated system setup and Equilibration

All MD simulations were carried out using AMBERTools14. To parametrize the ligand, Antechamber tool[23] in conjunction with General Amber Force Field (GAFF) was utilized to classify atom and bond types, assign charges, and estimate force field parameters. The charge method AM1-BCC of GAFF which is semi-empirical was used in this study. The solvation box with charge neutrality and physiological ionic strength (0.154 M in Na^+ and Cl^- ions), as well as complex system parameters and topology files, were prepared using tLEaP[24]. Protein and water were represented by Amber ff14SB[25] and TIP3P[26] models respectively in the prepared system. In all SuMD replicas, simulation starts with ligand located 40 Å far from the orthosteric active site at time zero, which is a distance bigger than the electrostatic cut-off term used in the simulation (9 Å with Amber force field), to avoid premature interaction during the initial phases of SuMD simulations.

All simulation systems were energy minimized through two equilibration steps. Considering 2 fs as a time step equal to the vibrational frequency of bonds, 500,000 steps (1 ns) of NVT in addition to 500,000 steps (1 ns) of NPT simulations were carried out. Gradual reduction of harmonic positional constraints by a force constant of 5 kcal mol^{-1} $Å^{-2}$ was applied in both steps. In the first equilibration, ions (except bounded Na^+ in the sodium-binding loop) and water were kept free, while protein and ligand atoms were constrained. However, in the second equilibration, the constraints were kept only on the alpha carbons of the protein, as well as ligand atoms and the loop sodium. In both steps, the temperature was maintained at 310 K by a Langevin thermostat with low damping of 0.1 $ps^{-1,}$ and in the second NPT step, the pressure was maintained at 1 atm by a Berendsenbarostat[27] as well. To calculate electrostatic interactions with a cubic spline interpolation and a 9.0 Å cutoff for Lennard–Jones interactions, the Particle-mesh Ewald (PME) method was utilized[28].

2.4 Supervised Molecular Dynamics (SuMD) production

The SuMD simulations were done in NVT conditions with the temperature equal to 310 K, while the pressure of the system was free to change. To perform a supervised MD, the topology and coordinates of the last frame of the second equilibration phase were used as the starting point. In the configuration file of SuMD, three selected amino acid residues Glu97A, Gly219, Cys191 that whose center of mass (CM) approximately define the binding site CM were inputted. SuMD applies a dynamic selection on the indicated residues position to calculate the center of mass of the binding site. MOE suite was used to determine the center of the mass of the co-crystallized ligand regarded as the center of the mass of the thrombin active site to be then visually selecting a combination of residues that their center of mass could represent the approximate position of the binding site guiding the supervision. Each SuMD replica was produced on a graphics machine using ACEMD3[29]

as the MD engine. The length of the SuMD steps for SuMD replicas was set either to 600 ps or 1 ns time window.

2.5 Free (unsupervised) Classical Molecular Dynamics (cMD) production

For each cMD, after system preparation and equilibration steps, ACEMD3[29] engine was used with the same settings, except for the simulation length, of the cMD simulation in each SuMD step.

2.6 Visualization of the MD trajectories

Visual Molecular Dynamics (VMD)[30] and MOE suite[22] were utilized during this project for molecular visualization and analysis of the trajectories.

2.7 Trajectory versus Trajectory RMSD calculation

Using MDAnalysis[31,32], a matrix of frames related to the cMD of reference against frames of each SuMD replica was set for comparative RMSD calculation. Then via Seaborn python library[33], a heat map of the resulted RMSD calculation was illustrated (Figure 2).

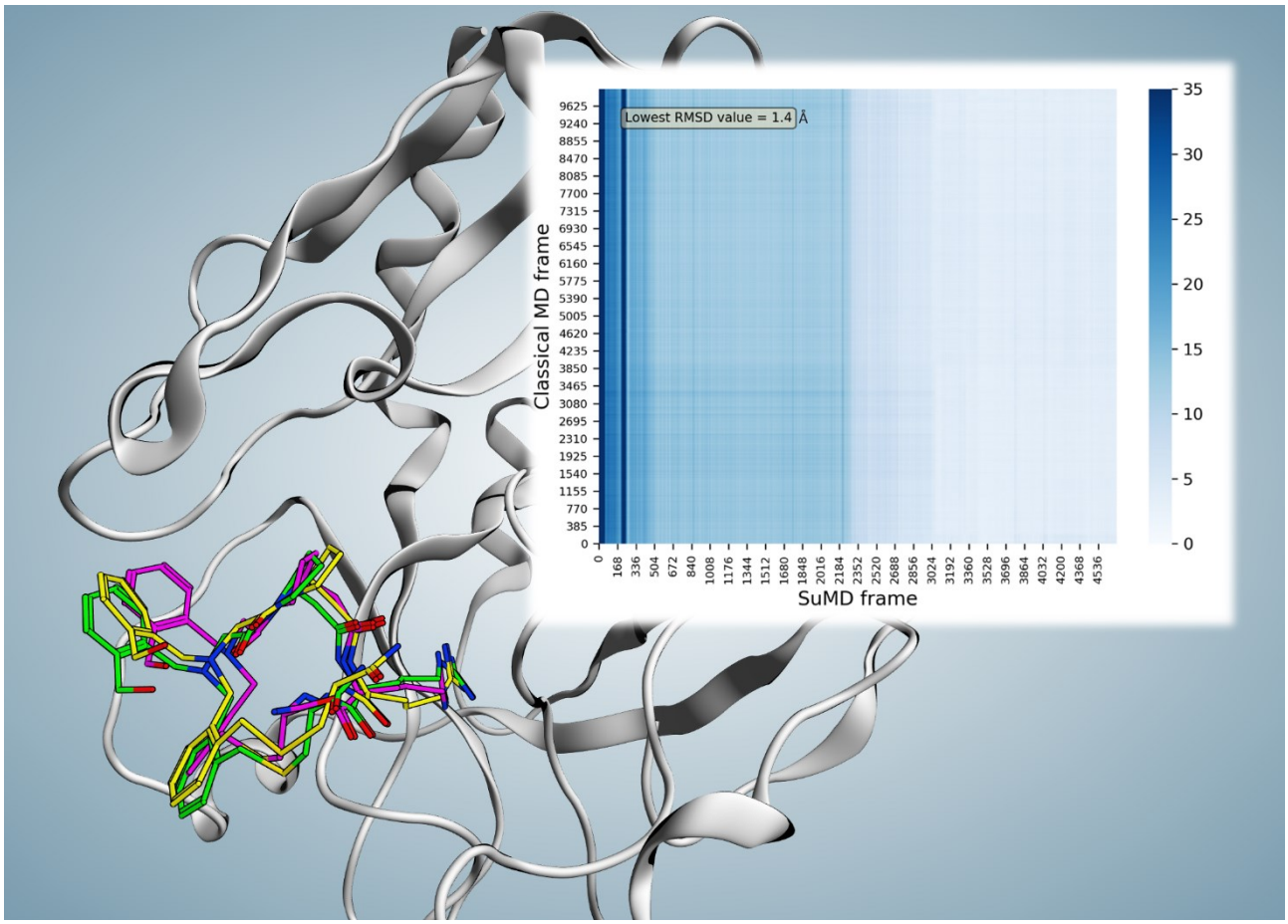


Figure 2. Superposition of the P2 reported X-ray crystallography conformation (magenta), the frame number 102 (2.04 ns) of reference cMD(green), and the frame number 3579 (71.56 ns) of replica 74 (yellow; the frame with lowest RMSD value compared to both reported binding conformation and the parallel trajectory analysis reference resulted in frame). The minimum obtained RMSD compared to the reported binding conformation showed a value of 2.27 Å at time point

71.56 ns. However, considering observed instability of the reported conformation, parallel frame RMSD calculation of replica 74 trajectory versus reference cMD trajectory was performed which resulted in the minimum RMSD value of 1.4 Å at 71.56 ns for the frame number 3579 (the same frame with the lowest RMSD value compared to the reported binding conformation). Heat map of the parallel trajectory RMSD analysis of replica 74 versus reference cMD is shown in the top right.

2.8 MM-GBSA Energetic Profile Analysis and Clustering

All total free energy calculations in this work were computed using the MMPBSA.py tool[34] using GB-OBC(II) Born solvation model and no entropy calculation. To identify other energetically favorable binding sites and elucidate P2 ligand−protein recognition scenario, the trajectories of 99 SuMD replicas of P2 were first solvent-dried, aligned, merged and ten times strided as input for positional clusterization. To do so, ligand sets of coordinates (each set of coordinates corresponds to the ligand conformation in a frame) after discarding noise sets considering cosine similarity value of 0.01 were clusterized using OPTICS algorithm of Scikit-learn[35]. Thereafter, given MM-GBSA value of the included ligand coordinates in each cluster, the representative ligand conformation with the most favorable energetic value was selected for the corresponding cluster.

3 Results

3.1 Study principal outcomes

This study was conducted aiming to extend the application domain of our molecular dynamics supervision method for studies related to models of sub-kilodalton macrocyclic peptide-protein binding event. As a case study, the SuMD ability to reproduce the X-ray crystallography bound state of P2 macrocyclic peptide to thrombin was evaluated. To that end, 99 SuMD simulations were performed starting from an unbound state obtained by separating P2 from its binding site by around 40 Å. Among 99 SuMD replicas, 84 trajectories finished with the ligand arriving in the proximity of the binding site and its sub-pockets with different binding orientations and conformations, while 15 trajectories ended with ligand stopping over a varied site categorized as “failed” based on SuMD termination criteria (far from the binding site). Overall, five trajectories concluded with the ligand reached the narrow S1 pocket (guanidinium moiety entering S1), all below 100 nanoseconds of SuMD-productive simulation time.

To better compare the SuMDresults with the experimental structure, the X-ray crystallography complex (reference) was subjected to 200 ns of cMD, allowing to have both systems in similar conditions: equilibrated and relaxed in a fully explicit solvent environment. In fact, during the initial 4 ns of the cMDa fluctuation within the range of experimental resolution (2.3 Å) was observed while,

after 4 ns, a more significant shift of the macrocycle occurred as confirmed by a drop of its RMSD values to above 3 Å and below 5.66 Å until the end (200 ns) was detected(Video-S1). The mean calculated RMSD during this trajectory was 3.57 ± 0.47 Å(Figure-S1). Those RMSDvalues highlight a discrepancy between the experimental bound conformation and the one assumed once the system is equilibrated in fully explicit solvent suggesting that the cMD could represent a more adequate comparison for SuMD. Indeed, we performed a frame-to-frame analysis of the SuMD trajectory versus cMD trajectory monitoring the ligand RMSD. Among all replicas, replica 74 is deemed as the best-produced binding event trajectory for P2. This SuMD simulation (replica 74) with 94 nanoseconds duration reproduced a possible binding event trajectory with the most comparability to the X-ray crystallography binding mode. The minimum obtained RMSD compared to the X-ray conformation showed a value of 2.27 Å at time point 71.56 ns. However, considering thecMD trajectory as reference the minimum RMSD value of 1.4 Å at the same time-point(71.56 ns, frame number 3579) versus frame number 102 (2.04 ns) of reference cMD (Figure 2).

The simulation started with P2 located 40 Å far from the orthosteric active site (AS) at time zero (**Video-S2**), and then upon approaching toward the AS the first stable binding occurred from time point 5.5 ns until 43 ns with a mean MM-GBSA free energy (ΔG) of -27.1 kcal/mol. As this stopover had enough residence time to break the progressive and continual approach of the ligand, it can be defined as a meta-stable binding site. Afterward, for around 4 ns from time point 48 ns, another stable contact near the active site (ΔG = -18.9 kcal/mol) was seen as the ligand was transitioning to the active site area. Then, from 54 ns a favorable orientation of P2 facilitated the entrance of the fundamental guanidinium moiety to the S1 pocket. From that point, an initial evolution of the final binding state phase was followed by fluctuating but stable similar conformations until the end. The mean total ΔG of the last 22.44 ns (from the $RMSD_{min}$ frame until the end) resulted in a value of -27.3 kcal/mol compared to the calculated mean total ΔG value of -32.2 kcal/mol in the same duration (22.44 ns) of reference cMD, showing similar with no meaningful MM-GBSA difference (ΔΔG < 5 kcal/mol) energetic profile. A figure reporting some of the most relevant ligand P2 conformations during this binding trajectory is present in Figure 3. To evaluate the P2 flexibility we calculated its RMSD during the best 5 trajecories; the ligand fluctuates until 5.7 Å, suggesting that a certain flexibility is explored during the recognition (Figure S2). To identify significant states during the P2 recognition and their corresponding meta-stable binding sites, all the conformations sampled during all the SuMD trajectories were geometrically clustered resulting in seven clusters (Figure 4). All the clusters showed a favorable average of MM-GBSA binding free energy value compared to

the calculated value for the reference binding conformation in the canonical binding site (active site) (Table 1). This outcome suggests that multiple energetically favorable binding patches on the thrombin surface for P2. Among all the clusters, the seventh cluster comprised of 2-3 times higher number of frames with the minimum average free energy value of -33 kcal/mol. The position of this

Cluster No.	Number of included frames	Average MM-GBSA total ΔG kcal/mol (rounded)
1	459	-22
2	309	-28
3	460	-27
4	356	-32
5	438	-26
6	275	-28
7	890	-33

Table 1. Size and energetic analysis of all the clusters obtained during P2 SuMD simulations. Several conformation showed a favorable MM-GBSA binding free energy value suggesting multiple energetically favorable binding states on thrombin surface for P2.

cluster population was identified near to the exosite II. Given high population of the seventh cluster, highly favorable binding free energy value—closely comparable to the canonical binding site—of this positional cluster near exosite II which is considered as an important contact point for natural thrombin substrates to form an initial stable complex conformation required for the peptide bond cleavage; it can be hypothesized that thrombin inhibition by P2 might be resulted from dual-site inhibition i.e. allosterically preventing the selective stable recognition of substrates in addition to occupying the orthosteric proteolysis site and thus be a potent thrombin inhibitor.

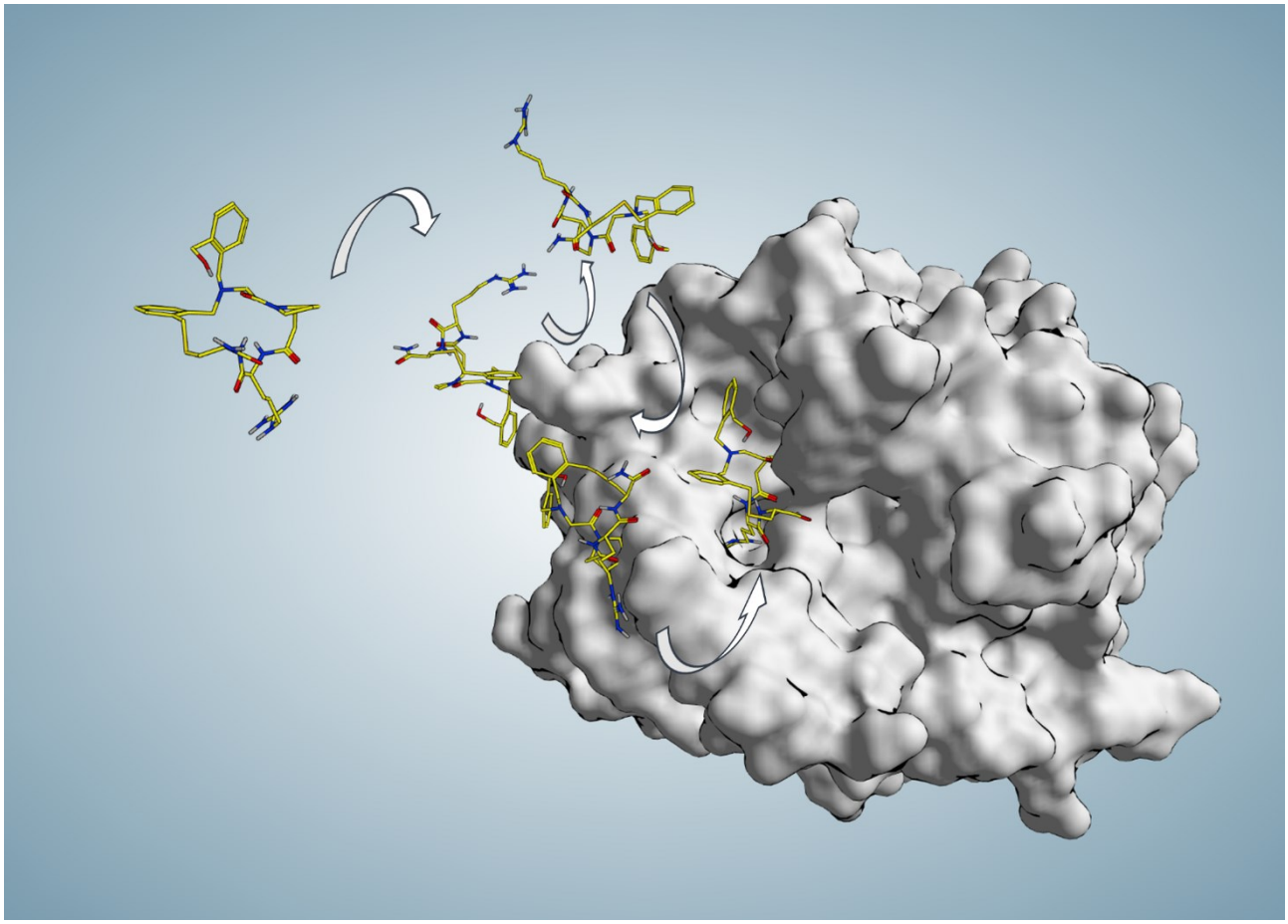


Figure 3. This figure shows some representative P2 poses along the binding trajectory (94 ns) produced in SuMD replica 74 during which P2 start approaching the active site from 30 Å far from any protein atom at time zero and reaches the binding site and S1 pocket in an experimentally comparable binding mode ($RMSD_{min}$=1.4 Å at 71.56 ns).

3.2 Elucidation of the role hydroxymethyl-benzyl-moiety

AsP1 shares the same structure of P2 except for the presence of a hydroxymethyl-benzyl structure on the latter, it could be hypothesized that a similar binding mode and orientation of the ligand with guanidinium moiety entering the S1 pocket and the macrocycle occupying the rest of the active site. For further witnessing of SuMDhelpful implication in depicting the molecular basis of the recognition of this class of compounds, we investigate the hydroxymethyl-benzyl role that leads to an increased inhibition activity (three orders of magnitude); SuMD simulations for P1were

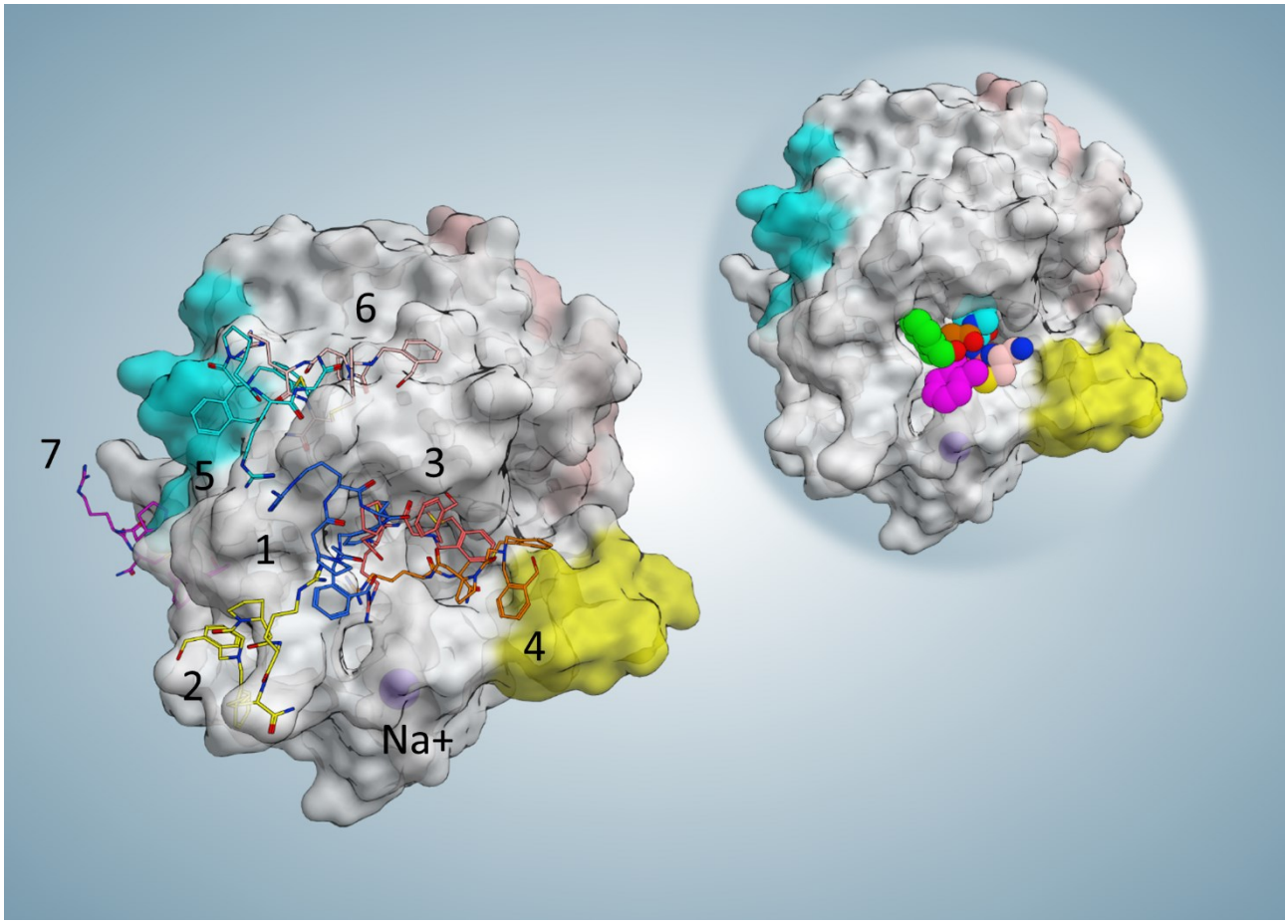


Figure 4.This figure shows the representative frame (minimum MM-GBSA free binding energy in each cluster) and the position of each cluster. For the collective illustration of all representing poses on one protein surface, the molecular surface of the reference PDB was selected to be shown here. On the top right reference, complex is shown to indicate the active site position.

additionally performed until reaching a representative replica (Video-S3) in which P1 fully enters the active site S1pocket to establish a salt bridge with Asp189. After 18 replicas (in 16 replicas P1 reached the binding site in different final binding modes among which 6 replicas had the supposed orientation; one of those had expected orientation while entered the S1 pocket), we obtained a possible binding trajectory in which during 34.84 ns of simulation, P1 reached the active site with guanidinium partly inside S1 pocket as supposed. For further evolution and reaching the most stable conformation, the simulation continued with 50 ns of cMD. After that, a stable conformation was achieved; having a salt bridge with Asp189 and contacting four of the same reference P2 interacting residues(Asp189, Cys220, Gly216, Glu217) (Figure 5).

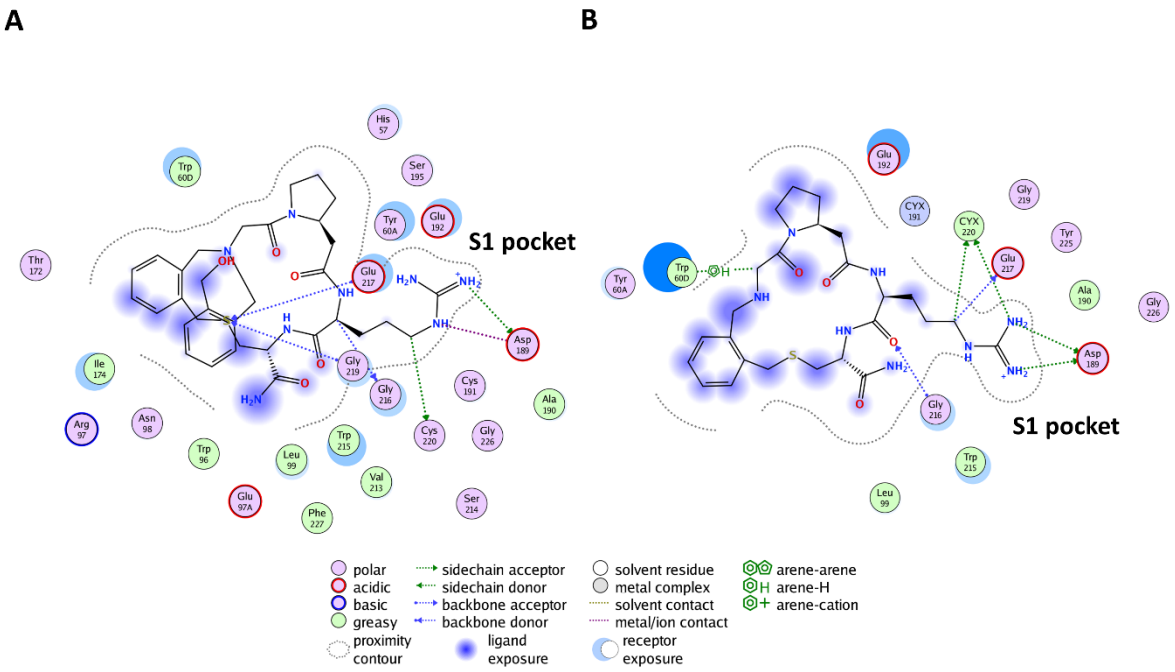


Figure 5. The reported binding mode interactions of P2 **(A)** and the interaction panel of P1 simulated final stable conformation **(B).**

To compare P1 and P2 from an energetic point of view, the mean total MM-GBSA binding free energy during 50 ns of cMD trajectories were taken into consideration. The calculated total ΔG_{P1}= -20.2 kcal/mol and total ΔG_{P2}= -29.3 kcal/mol, show a more favorable energy profile in P2 as expected. To be confident about compared value correctly associated with the final evolved stable P1 binding conformation, similar total ΔG_{P1}= -20.37 kcal/mol was obtained for the last 6 ns of the P1 continued cMD($RMSD_{last\ 6\ ns}$=1.4 ± 0.4 Å). The energy landscape of P1 and P2 trajectories (Figure 6) indicates a similar profile characterized by a large number of energetically stable frames when the distance between the centers of mass (dcm_{L-R}) is in the range of 3-7.5 Å. This observation suggests that many ligand states, even if they present different binding modes, contribute to a stable protein-ligand association. The presence of metastable binding sites far from the active site (dcm_{L-R}> 10 Å)is slightly more pronounced in the representative trajectory of P2 where three transient spikes are evident at dcm_{L-R} 9 Å, 15 Å, and 20 Å.

Additionally, to compare P1 and P2 structural characteristics and their possible effects on each ligand dynamics and binding during SuMD condition, the representative replica of P2 and the P1were selected for further analysis. Given the experimental final binding state of P2, an internal hydrogen bond (2.15 Å/H-O) between the hydrogen of the hydroxyl group of hydroxymethyl-benzyl moiety and the nearby carbonyl group of the macrocycle ring can be seen. This hydrogen bond during produced binding event trajectory (replica 74) sustains an average value of 2.72 Å (H-O).

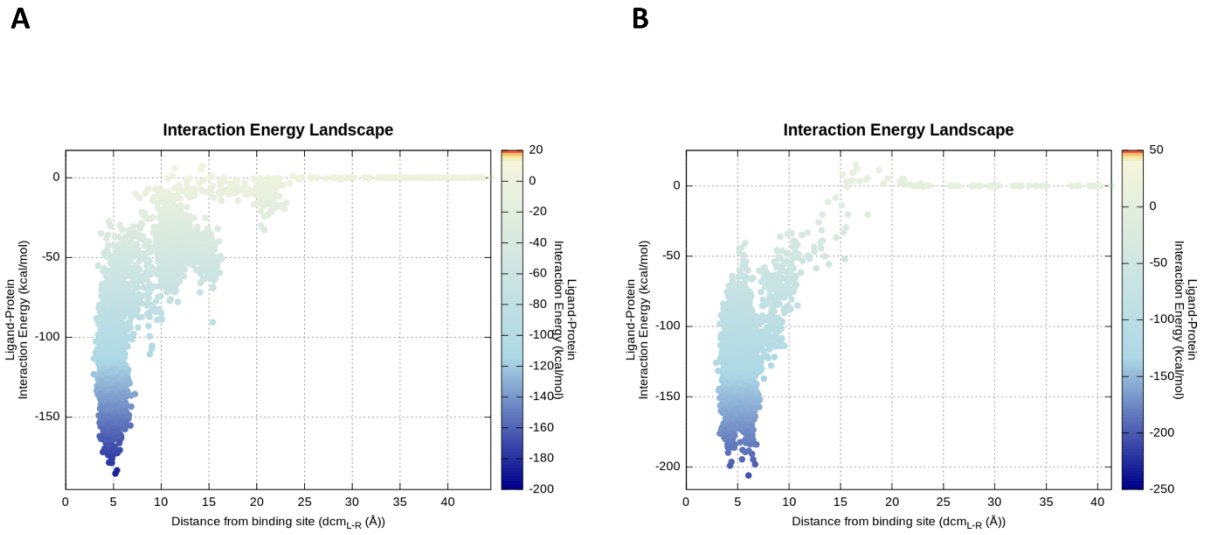


Figure 6. Energy landscape of P2 **(A)** and P1 **(B)** representative produced trajectories. The interaction energy calculation is based on *mdenergy* function of VMD[36] and plotted via in-house pepSuMD analyzer[3]. **(A)**along this trajectory 2 to 3 local minima can be seen which correspond to meta-stable binding sites for P2. **(B)** P1 directly goes to the canonical active site during this representative trajectory.

Considering that, it could be hypothesized that this present internal bond thanks to hydroxymethyl-benzyl moiety which is absent in P1, contributes to a less flexible structure and a biased maintained conformation necessary for the observed favorable snug-fit binding. To corroborate this idea, the average RMSD of mutual macrocycle ring of P2 and P1 during the time, in addition to the RMSD of the whole structure of each ligand along their representative SuMD trajectory was calculated. For this RMSD calculation, all frames of each representative replica were aligned on the comprising atoms of the mutual ring of the corresponding replica first frame separately. The calculations obtained in this way indicate the flexibility of the mutual ring and each ligand and not the ligand transition during their molecular dynamics trajectory. The achieved values of the mutual ring and the whole ligand in P2 trajectory were respectively four times and 2.7 times less than calculated values for P1 (average RMSD $_{ring/P2}$= 0.41 ± 0.15 Å, average RMSD $_{ligand/P2}$= 1.62 ± 0.4 Å; average RMSD $_{ring/P1}$= 1.68 ± 0.22 Å, average RMSD $_{ligand/P1}$= 4.43 ± 0.56 Å). Thus, as expected, this result can quantitatively show a more biased stable conformation for P2 during time compared to P1.

4 Discussion

In this study, the ability of SuMD protocol in the reproduction of X-ray crystallography final binding state of the candidate macrocyclic tetrapeptide P2—from a novel library of 8988 sub-kilodalton macrocyclic peptides—bound to thrombin to inhibit its activity, was successfully investigated (minimum RMSD of 1.4 Å at 71.56 ns). The outcomes reported that more than 80 percent of

trajectories reached the canonical binding surface in varied conformations below or around a hundred nanoseconds, and near five percent mimic the experimentally-solved final bound state for this class of macrocyclic peptides to a challenging target, characterized by a narrow active site cleft and deep significant-for-activity sub-pocket (S1). These results reiterated and extended SuMD high value as a computational protocol to explore the recognition pathway. Additionally, based on the observations, SuMD can be regarded as an insightful tool in terms of meta-stable binding sites identification, as well as the binding mode and molecular recognition pattern elucidation of sub-kilodalton macrocyclic peptides (with different scaffold than natural peptides or small molecules) to a protein target with relatively low computational expense. Therefore, this study further validated and expanded the applicability of SuMD as a valuable protocol in studying varied molecular complex recognition.

The main advantages of the method used in this work are being able to correctly parametrize ligand P2 of this class of macrocyclic peptides with a general amber force field (GAFF) similar to small molecules and thus no need for tailored parametrization due to the presence of unnatural amino acids and linkers; as well as the possibility to simulate the trajectory of a binding event in nanosecond timescale thanks to SuMD. Consider that the association event starting from an unbound state is arare event to be observed by cMD without the implementation of a enhanced sampling strategy. For instance, in Video-S4 a comparative cMD starting from the same state of SuMD is reported; during the 900 ns of simulation P2 never approached thrombin, confirming the different sampling rate of the two methods. Given that, the opportunity of performing an efficient high-throughput molecular dynamics study of the remaining macrocyclic peptides of the same class, after further optimization and validation can be envisioned. Therefore, the prospective use of this study findings would be toward using SuMD to perform high throughput molecular dynamics studies of other available macrocyclic peptides of the same class, enjoying highly diverse scaffold, to find probable hit candidates for various protein targets of interest and predict their binding mode as an adjunct predictive and screening tool, similarly to what recently reported for fragments[37]; narrowing down the requirement of going through experimental structural studies for each molecular complex of interest. On the contrary, a particular attention should be paid to the starting conformation of the macrocycle that could affect the recognition sampling since their flexibility could be rather pronounced. Using specific methods (e.g. low-mode MD) to preproccess novel ligand for selecting at least one of few adequate starting conformation in solution. It should be also considered that particularly flexible sub-kDa macrocycle could present more issue in sampling the bound

conformation during the recognition. Anyway, all of these prospective enhancements would lead to the main goal of achieving computationally cheap molecular dynamics study methods with ever-increasing power in predicting experimental-equivalent final binding states and recognition of key elements and patterns of complexes.

References:

1. Lin, X., Li, X. & Lin, X. A Review on Applications of Computational Methods in Drug Screening and Design. *Molecules* **25**, 1375 (2020).

2. Muegge, I., Bergner, A. & Kriegl, J. M. Computer-aided drug design at Boehringer Ingelheim. *J. Comput. Aided. Mol. Des.* **31**, 275–285 (2017).

3. 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).

4. Buch, I., Giorgino, T. & De Fabritiis, G. Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations. *Proc. Natl. Acad. Sci.* **108**, 10184–10189 (2011).

5. Dror, R. O. *et al.* Pathway and mechanism of drug binding to G-protein-coupled receptors. *Proc. Natl. Acad. Sci.* **108**, 13118–13123 (2011).

6. Kahler, U., Kamenik, A. S., Kraml, J. & Liedl, K. R. Sodium-induced population shift drives activation of thrombin. *Sci. Rep.* **10**, 1086 (2020).

7. 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).

8. 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).

9. Kale, S. S. *et al.* Thiol-to-amine cyclization reaction enables screening of large libraries of macrocyclic compounds and the generation of sub-kilodalton ligands. *Sci. Adv.* **5**, (2019).

10. Passioura, T. The Road Ahead for the Development of Macrocyclic Peptide Ligands. *Biochemistry* **59**, 139–145 (2020).

11. Giordanetto, F. & Kihlberg, J. Macrocyclic Drugs and Clinical Candidates: What Can Medicinal Chemists Learn from Their Properties? *J. Med. Chem.* **57**, 278–295 (2014).

12. Kamenik, A. S., Lessel, U., Fuchs, J. E., Fox, T. & Liedl, K. R. Peptidic Macrocycles - Conformational Sampling and Thermodynamic Characterization. *J. Chem. Inf. Model.* **58**, 982–992 (2018).

13. Deyle, K., Kong, X.-D. & Heinis, C. Phage Selection of Cyclic Peptides for Application in Research and Drug Development. *Acc. Chem. Res.* **50**, 1866–1874 (2017).

14. Vinogradov, A. A., Yin, Y. & Suga, H. Macrocyclic Peptides as Drug Candidates: Recent Progress and Remaining Challenges. *J. Am. Chem. Soc.* **141**, 4167–4181 (2019).

15. Taylor, R. D., Rey-Carrizo, M., Passioura, T. & Suga, H. Identification of nonstandard macrocyclic peptide ligands through display screening. *Drug Discov. Today Technol.* **26**, 17–23 (2017).

16. Huntington, J. A. How Na+ activates thrombin – a review of the functional and structural data. *Biol. Chem.* **389**, (2008).

17. HUNTINGTON, J. A. Molecular recognition mechanisms of thrombin. *J. Thromb. Haemost.* **3**, 1861–

1872 (2005).

18. Schechter, I. & Berger, A. On the size of the active site in proteases. I. Papain. *Biochem. Biophys. Res. Commun.* **27**, 157–162 (1967).

19. He, L.-W., Dai, W.-C. & Li, N.-G. Development of Orally Active Thrombin Inhibitors for the Treatment of Thrombotic Disorder Diseases. *Molecules* **20**, 11046–11062 (2015).

20. Chahal, G., Thorpe, M. & Hellman, L. The Importance of Exosite Interactions for Substrate Cleavage by Human Thrombin. *PLoS One* **10**, e0129511 (2015).

21. Di Cera, E. Thrombin Interactions. *Chest* **124**, 11S-17S (2003).

22. Chemical Computing Group ULC, Molecular Operating Environment (MOE), 2019.01. 1010 Sherbrooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, 2019.

23. 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).

24. Case, D. A. *et al.* The Amber biomolecular simulation programs. *J. Comput. Chem.* **26**, 1668–1688 (2005).

25. 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).

26. 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).

27. Berendsen, H. J. C., Postma, J. P. M., Van Gunsteren, W. F., Dinola, A. & Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* (1984) doi:10.1063/1.448118.

28. Essmann, U. *et al.* A smooth particle mesh Ewald method. *J. Chem. Phys.* **103**, 8577–8593 (1995).

29. 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.

30. Humphrey, W., Dalke, A. & Schulten, K. VMD: Visual molecular dynamics. *J. Mol. Graph.* (1996) doi:10.1016/0263-7855(96)00018-5.

31. Gowers, R. *et al.* MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations. in 98–105 (2016). doi:10.25080/Majora-629e541a-00e.

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. Waskom, M. seaborn: statistical data visualization. *J. Open Source Softw.* **6**, 3021 (2021).

34. Miller, B. R. *et al.* MMPBSA.py: An Efficient Program for End-State Free Energy Calculations. *J Chem Theory Comput* **8**, 3314–3321 (2012).

35. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* (2011).

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

37. 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.

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 [34]. 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 (TransActivate 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 non 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

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.

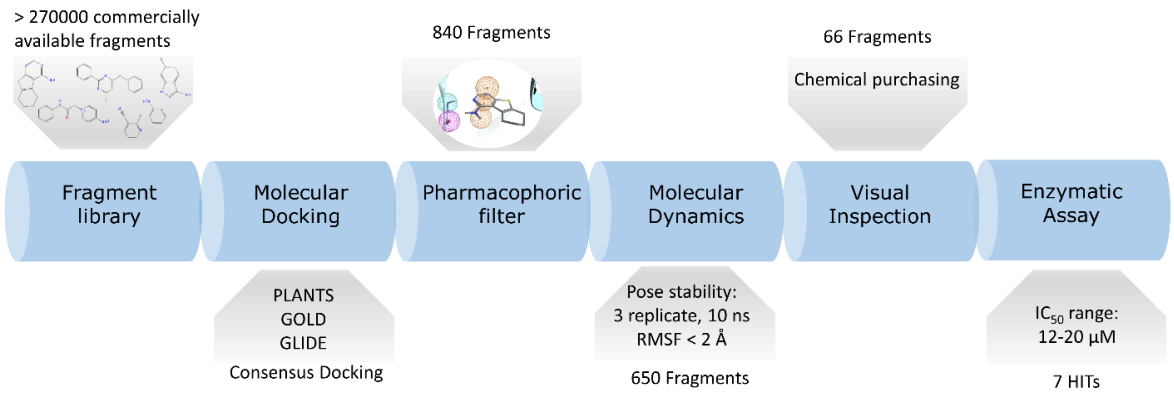


Figure 1. Schematic representation of the workflow adopted in the present work. First the fragments are retrieved from several vendors libraries. After proper preparation, the database is docked using three different docking protocols. the resulting poses have been filtered using a pharmacophore model 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 trough visual inspection.

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 Ck1d 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.

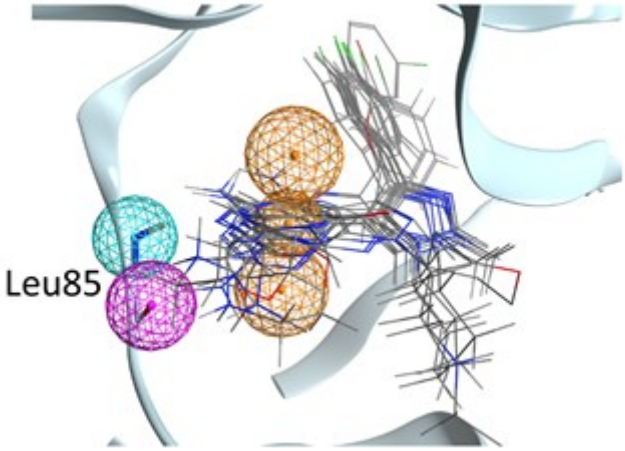


Figure 2. Representation of the pharmacophore model used in the present work. Some representative crystallographic ligands are displayed (not all for clarity). The Pharmacophore model is formed by an aromatic ring (the three orange spheres define the position and its orientation) and two hydrogen bonds 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 > 2Å, 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

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.

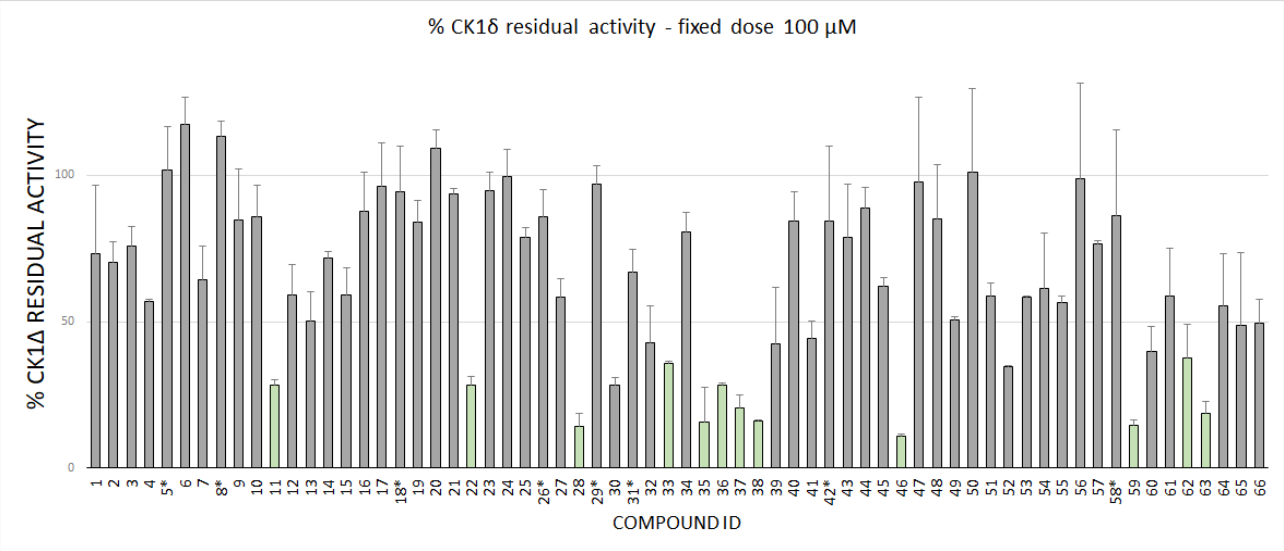


Figure 3. CK1δ residual activity at a concentration of 100 µM of the ligand under examination.The molecules marked with a star has been tested at 50 µM due to solubility issues.

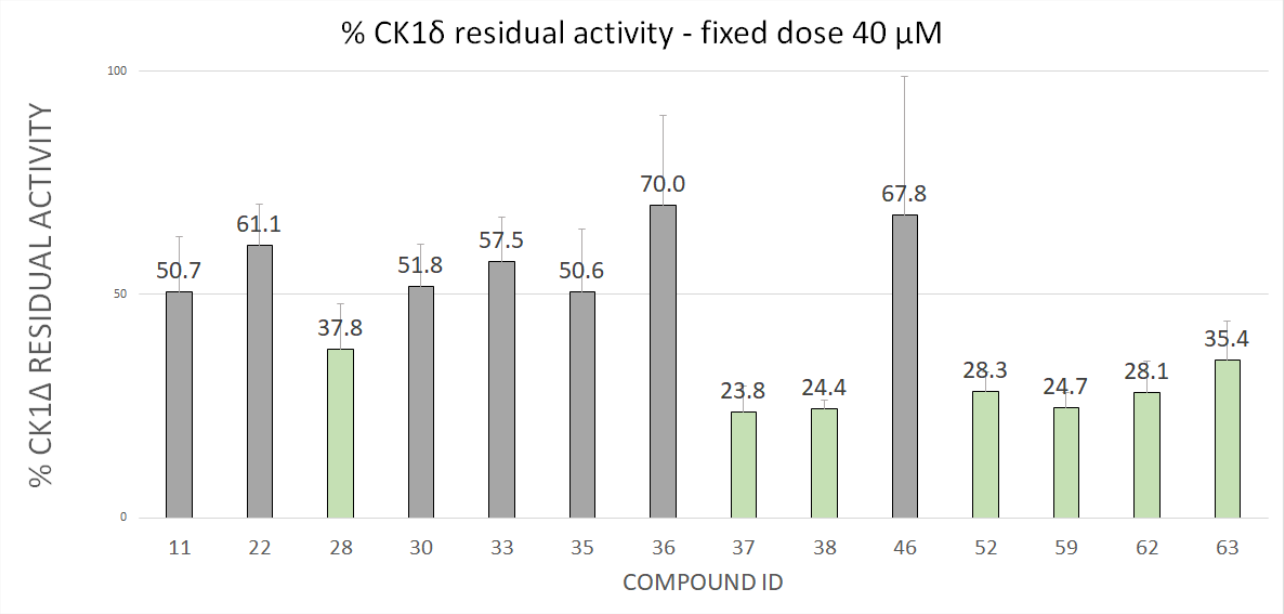


Figure 4. CK1δ Residual activity at a concentration of 40 µM of the ligands that showed a residual activity of less than 40% at 100 µM.

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).

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)

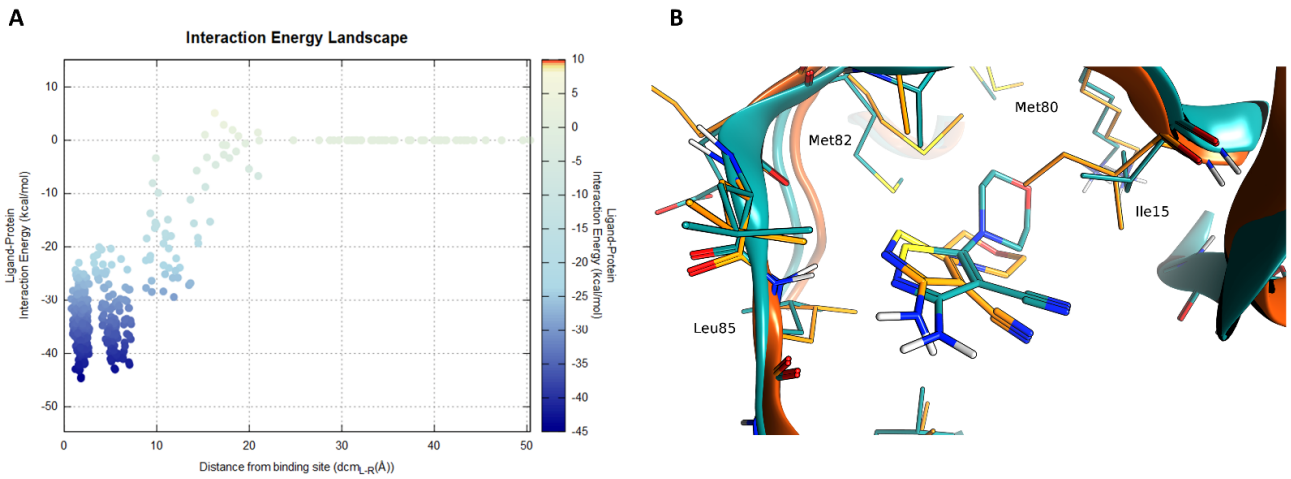
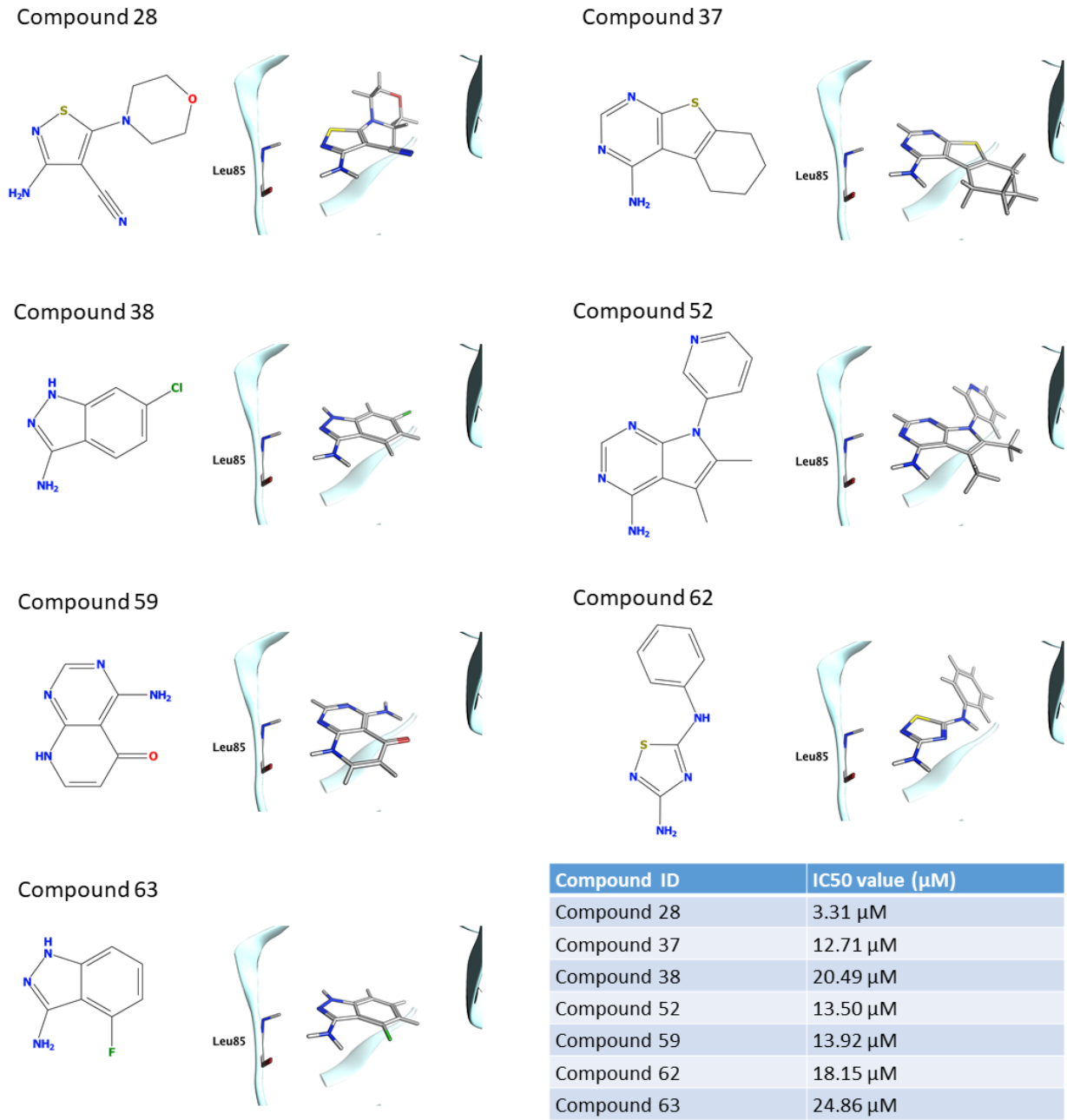


Figure 6. SuMD simulation of compound **28.** In panel (**A)** the interaction energy landscape is reported for the recognition trajectory displaying the ligand–protein interaction energy plotted against the distances between the protein–ligand center of mass. In panel (**B)**, the superposition of the VS-pose (cyan) for compound **28** against the lowest energy frame from the SuMD trajectory (orange).

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. 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].



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 IC50 value is reported. The value of IC50 is based on the average of three independent measurements.

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 hwas 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 15Å 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 1ns 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 1ns 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 2Å 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 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.

Conclusions

In the Ph.D project discussed in the present thesis several structure based computational approaches have been discussed and their application in different contexts have been shown.
These approaches include classic methods like Molecular Docking, that have been extensively analysed and applied for Virtual screening of Fragment molecules and for scaffold repurposing in order to find new kinase inhibitors, while these approaches still have a relevant role in Computational Aided Drug Discovery (CADD), more advanced tools to investigate in detail the ligand-receptor recognition are needed. In this perspective Supervised Molecular Dynamics simulations (SuMD) applications in CADD have been studied.
SuMD has been applied for the prediction of Fragment molecules binding mode, studying the importance of structural water molecules, to assess if the explicit solvent full atomistic conditions used in the SuMD simulation can compensate for the lack of structural information regarding the solvent molecules (using HSP90 as a case study). Other target investigated for the applicability of SuMD for fragment posing include SARS-CoV-2 main protease.
SuMD application has also been studied to elucidate the selectivity profile in a case study of Adenosine Receptors antagonists, where the final bound alone state cannot explain the observed differences, indeed some strong interactions between the antagonists and the receptor extracellular loops can be observed before the ligands reach the orthosteric binding site.
Other case studies include the elucidation of the recognition pathway for a macrocyclic peptide ligand that bind thrombin, for several inhibitors of the SARS-Cov-2 main protease and the NorA efflux pump.