Received: 21 April 2023 | Revised: 22 June 2023 | Accepted: 8 August 2023

DOI: 10.1002/prot.26582

RESEARCH ARTICLE

PROTEINS WILEY

Critical assessment of protein intrinsic disorder prediction (CAID) - Results of round 2

Alessio Del Conte[1] | Mahta Mehdiabadi[1] | Adel Bouhraoua[1] | Alexander Miguel Monzon[2] | Silvio C. E. Tosatto[1] | Damiano Piovesan[1]

[1]Department of Biomedical Sciences, University of Padova, Padova, Italy

[2]Department of Information Engineering, University of Padova, Padova, Italy

Correspondence
Silvio C. E. Tosatto, Department of Biomedical Sciences, University of Padova, via Ugo Bassi 58b, 35131 Padova, Italy.
Email: silvio.tosatto@unipd.it

Funding information
Horizon 2020, Grant/Award Numbers: 823886, 778247; Horizon Europe, Grant/Award Number: CA21160; ELIXIR

[Correction added after first online publication on 13 September 2023. Author name has been updated from "Alexander M. Mozon" to "Alexander Miguel Monzon"]

Abstract

Protein intrinsic disorder (ID) is a complex and context-dependent phenomenon that covers a continuum between fully disordered states and folded states with long dynamic regions. The lack of a ground truth that fits all ID flavors and the potential for order-to-disorder transitions depending on specific conditions makes ID prediction challenging. The CAID2 challenge aimed to evaluate the performance of different prediction methods across different benchmarks, leveraging the annotation provided by the DisProt database, which stores the coordinates of ID regions when there is experimental evidence in the literature. The CAID2 challenge demonstrated varying performance of different prediction methods across different benchmarks, highlighting the need for continued development of more versatile and efficient prediction software. Depending on the application, researchers may need to balance performance with execution time when selecting a predictor. Methods based on AlphaFold2 seem to be good ID predictors but they are better at detecting absence of order rather than ID regions as defined in DisProt. The CAID2 predictors can be freely used through the CAID Prediction Portal, and CAID has been integrated into OpenEBench, which will become the official platform for running future CAID challenges.

KEYWORDS

benchmarking, CAID, Critical assessment of protein intrinsic disorder prediction, intrinsic protein disorder

1 | INTRODUCTION

Intrinsically disordered proteins (IDPs) and regions (IDRs) display highly dynamic behavior by randomly sampling a vast array of conformations. This characteristic distinguishes them from protein switches, which, in response to a well-defined set of signals or stimuli, alternate between a limited number of conformations.[1]

Studying IDPs is challenging because experiments can only capture the average behavior of an ensemble.[2] For instance, NMR experiments provide conformational constraints but lack a description of the time dimension.[3] Additionally, some IDPs exhibit context-dependent behavior, displaying disorder under specific conditions such as the presence of a binding partner, changes in pH, etc.[4]

Predicting IDRs is problematic because protein dynamics cannot be described by a limited set of fixed conformations. The Critical Assessment of Protein Intrinsic Disorder Prediction (CAID)[5] focuses on analyzing the simpler problem of identifying positions within the protein sequence that have a propensity for being intrinsically disordered. In CAID, predictors are asked to provide a probability for each position, which can be turned into a binary prediction problem by applying a probability cutoff.

Over the years, numerous methods have been developed, varying based on implementation, training data, and purpose.[6] Some methods predict disordered regions derived from missing residues in X-ray experiments, while others attempt to capture context-dependent behavior (ID flavors), such as binding sites within IDRs. CAID not only assesses the accuracy of these methods but also evaluates their performance in terms of execution time and usability.

In this work we describe the results of the second round of CAID. The assessment includes an evaluation of the accuracy in predicting disorder and binding sites as well as a comparison of the execution time. The reference set is provided by the DisProt database[7] and includes a set of proteins for which disorder annotations were not previously available.

2 | MATERIALS AND METHODS

2.1 | Reference

There are various experimental techniques available to study the structural properties of proteins and ID within, each with its own biases and limitations. For instance, X-ray experiments tend to detect shorter IDRs because longer IDRs may be excised during construct preparation or hinder crystallization. On the other hand, circular dichroism can detect the absence of fixed structure of the whole protein, but lack residue resolution, applicable to full proteins. Therefore, IDRs should be confirmed by multiple lines of independent and diverse experimental evidence to increase their reliability. The DisProt database was chosen as the reference for structural disorder as in the first round of CAID because it contains a large number of manually curated disorder annotations at the protein level, with most residues annotated by more than one experiment. DisProt defines IDRs as regions of at least 10 residues that are likely to be associated with a biological function and excludes short loops connecting secondary structure elements. It also includes protein–protein interaction interfaces within disordered regions as a separate dataset, defined as "Binding" in CAID.

CAID benchmark proteins were obtained by calculating the delta between the public and private versions of the DisProt database as of November 20, 2022. Figure 1 panel A, shows how the reference datasets have been generated. Ideally, DisProt annotations would be complete, meaning that all disordered or binding regions present under physiological conditions would be annotated for each protein. In this scenario, all residues not annotated as disordered would be considered structured (negatives), while disordered residues would be considered positives. However, since not all IDRs are currently included in DisProt, the "Disorder-PDB" dataset was created, which only includes negatives among PDB Observed residues. This dataset is more conservative but considered more reliable as it excludes uncertain residues without any structural or disorder annotation. Additionally, since the availability of structural information in the Protein Data Bank (PDB) database[8] we excluded missing residues annotations in the default disorder benchmarking, "Disorder-NOX" dataset from here on. Indeed, missing residues are the type of ground truth data mainly used to train disorder prediction methods. Figure 1, panel B displays a Venn diagram of the number of proteins for the Disorder-NOX, Disorder-PDB, and Binding datasets.

In CAID2 we added a new category "Linker" which includes 40 proteins. Linkers are defined in DisProt as unstructured regions, providing separation and permitting movement between adjacent functional regions, for example, structured domains or disordered motifs. In contrast to CAID1, for the Binding dataset we considered only proteins with at least one binding region.

The total number of positive residues is 31 315 (19.5%), 37 072 (28.3%), 8209 (12.2%), and 2023 (5.4%) for the Disorder-NOX, Disorder-PDB, Binding, and Linkers datasets, respectively (Figure 1, panel D).

When missing residues are excluded, as in the Disorder-NOX, the number of target proteins decreases, but the average number of positive residues per protein is higher (Figure 1, panel F). This is due to the larger size of disorder regions detected by experimental techniques other than X-ray.

On the other hand, the Disorder-PDB dataset has a slightly higher relative disorder content (Figure 1, panel G). This can be attributed to most of the PDB constructs being fragments. Probably a large fraction of excluded positions, which are the negatives in Disorder-NOX, are bona fide disorder regions corresponding to missing annotations in DisProt.

Figure 1, panel E provides the fraction of residues covered by each experimental technique in the entire protein set. Circular dichroism covers approximately 80% of disordered residues, followed by "author statements," which refers to annotations where the authors mention a disordered region citing other works or because the disorder state is a well-established piece of knowledge.

The ID targets in CAID come from a diverse range of organisms, with the majority from eukaryotes, followed by bacteria and viruses and bacteria, and zero archaea (Figure 1, panel C). The target proteins are different from both those in the previous DisProt release and PDB construct sequences (PDB SEQRES field), with mean local sequence identity of 22.8% and 31.6%, respectively, when aligned with the Smith–Waterman algorithm (data not shown).

2.2 | Containers and predictions

In the CAID2 challenge, we collected a total of 46 different software programs and containerized (explained below) them for standardization of input and output data. Some of these software programs generated multiple outputs, resulting in a total of 71 different predictor "flavors," corresponding to the different variations of the predictor. The containerization of the softwares was done using Singularity (https://sylabs.io) containers to standardize the input and output data and ensure reproducible results. By containerizing the software, we

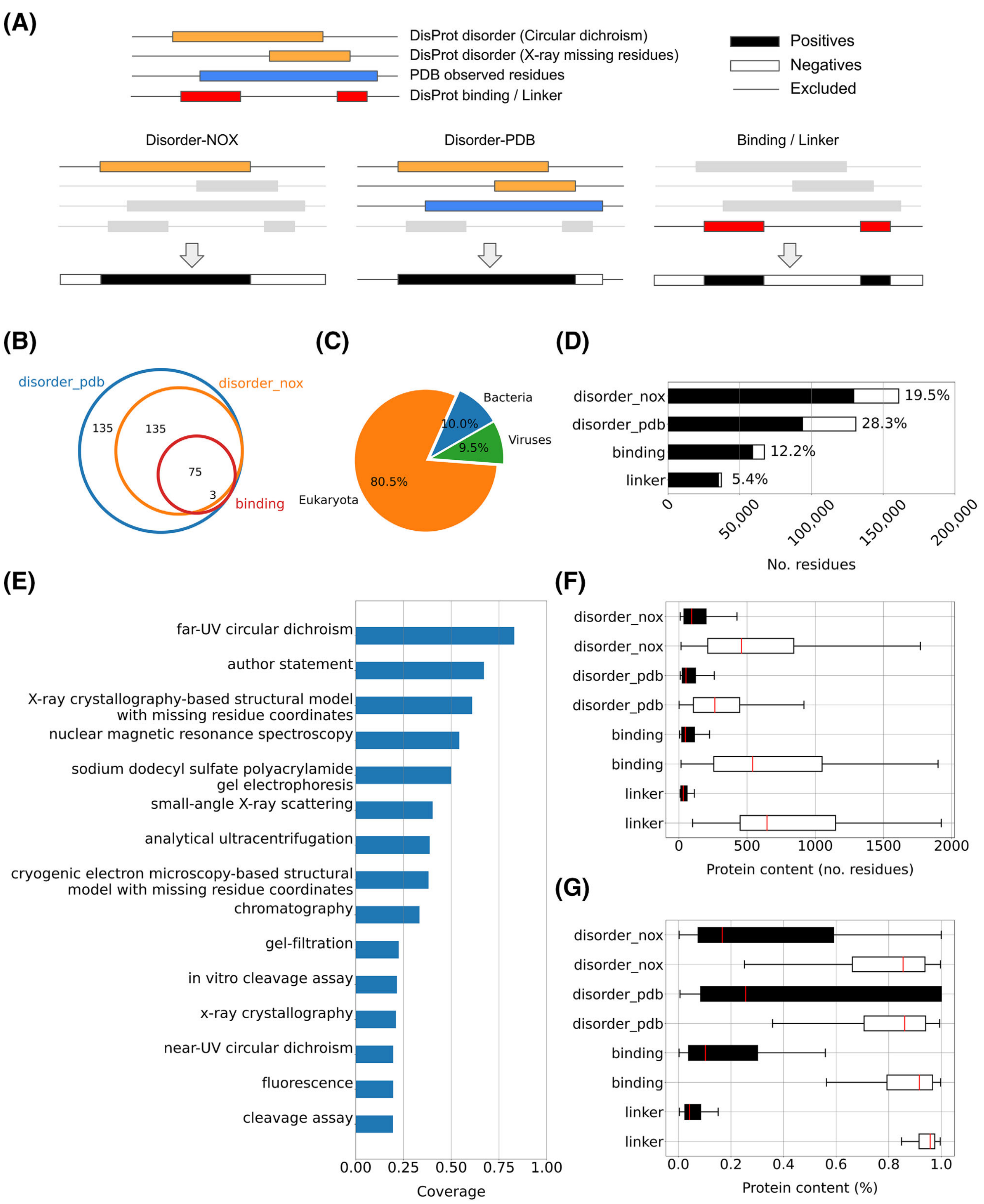

FIGURE 1 CAID2 dataset statistics. (A) The benchmarking dataset. (B) A Venn diagram of proteins in the benchmarking datasets. (D) The distribution of proteins across the three primary domains of life for the whole set of proteins. (D) Residue classification; the fraction of positive residues is reported at the top of the bars. (E) The fraction of disordered residues covered by specific experimental evidence. It is important to note that the same region can be identified by multiple experiments, and only the top 15 experimental methods are reported. The distribution of positive and negative classes at the protein level at the level of residues (F) and as percentage normalized by protein length (G).

eliminated the need for manual installation on each machine and ensured that the software runs consistently across different machines. Additionally, containerizing the predictors enabled us to package all the necessary software and dependencies together, making it easier to deploy and update the predictors. Before and after executing the predictor, scripts were run to standardize the input and output of the container, creating an interface with the predictor software. The predictor received a FASTA file with multiple sequences as input, and generated one or multiple outputs per sequence while recording the execution time for each.

To execute all the 46 different predictor softwares of the CAID2 challenge, we utilized a cluster and tuned each software program's usage of RAM and CPU cores. Different software programs have different needs in terms of RAM and CPU, so we evaluated the capabilities of each software in using multiple CPUs and the maximum amount of RAM it used during execution. For each predictor, we set a maximum runtime of 4 h per protein sequence and limited the number of CPU cores to an upper limit of 24 and the amount of RAM to an upper limit of 47GB.

The protocol implemented for managing predictions in the CAID2 challenge is identical to that employed in the CAID1 challenge.

2.3 | AlphaFold disorder and binding prediction

To establish a comparison baseline, we also incorporated the AlphaFold-disorder package,[9] which utilizes AlphaFold2 predicted structures publicly available in databases[10,11] to infer predictions for disorder and binding. Unlike using the sequence, the AlphaFold-disorder method takes the protein structure predicted by AlphaFold2 as input. The structure was obtained directly from the AlphaFold Protein Structure Database (AlphaFoldDB)[11] by searching the UniProtKB[12] accession number. However, if the protein sequence is not present in UniProtKB, no structure can be downloaded, thus the prediction is unavailable.

There are three types of AlphaFold-disorder predictions: (i) AlphaFold-pLDDT, which is the 1 − pLDDT score, (ii) AlphaFold-rsa, which predicts the relative solvent accessibility (RSA)[13] over a local window centered on the residue to be predicted, and (iii) AlphaFold-binding, which identifies regions with high RSA and pLDDT. The specific formula used for AlphaFold-disorder predictions is outlined in reference 9.

2.4 | Evaluation

For each residue in the input sequence, the predictors produce a set of pairs consisting of a score and a state. These scores are expressed as decimal numbers, while the states are binary labels that anticipate whether the residue is structured or disordered. If the scores are not available, the states are utilized as scores. In situations where the states are absent, they are computed by applying a threshold to the scores. The default threshold is established by the states; if the method's authors do not specify a threshold, a value of 0.5 is used. This guarantees that the default threshold approximations are accurate for any score distribution. To provide 1000 possible thresholds, the prediction scores are rounded to the third decimal place. In CAID1, $F_{\max}$ and AUC were the primary evaluation criteria utilized, where $F_{\max}$ denotes the highest point on the precision–recall curve and AUC represents the area under the receiver operating characteristic (ROC) curve. In CAID2, we reported the average precision score (APS), which is calculated as the arithmetic mean of the precision values along the precision–recall curve. APS is proportional to the area under the precision–recall curve, and compared to $F_{\max}$, it better captures a method's ability to prioritize disordered regions. In order to guide the interpretation of the results each method is accompanied by the fraction of predicted targets (coverage). Indeed some methods crashed unexpectedly or did not provide an output in a reasonable time (see Section 2.2).

In the Shuffled dataset baseline, the reference is randomized at the dataset level, whereas the Random predictor is a completely random method that does not utilize any prior information.

2.5 | Execution time

The execution times of the predictors were recorded for each input sequence, and to minimize overhead, the start time was set after container initialization and just before executing the predictor software. This was achieved by incorporating a custom script into the container that managed input and output data, recorded execution times, and executed the prediction software.

For certain predictors, additional precomputed inputs such as PSI-BLAST[14] search results against UniRef90,[15] HHblits,[16] or SPIDER2[17] were required along with the input sequence in FASTA format. To save time, we precomputed these inputs for all sequences and provided them as input to the predictors that needed them. The time taken to generate these inputs was also recorded and added to the total processing time for a sequence when necessary.

Regarding the execution times of AlphaFold-disorder, as we did not run AlphaFold2 ourselves but used structures from the AlphaFoldDB (alphafold.ebi.ac.uk),[11] hence we do not have complete information on the execution time for this method. However, for a fair comparison in Figure 4, we executed AlphaFold2[18] locally on a CPU-based machine (without using GPUs) to predict the structure of a 1000-residue long sequence while recording the execution time. These structures were then used to run the AlphaFold-disorder package, and its runtime was added to the total.

3 | RESULTS AND DISCUSSION

Similar to the first round, the second round of CAID was organized with participants submitting their implemented prediction software to the assessors, who ran the packages and generated predictions for a set of proteins where disorder annotations were not previously available. However, in CAID2, software methods were encapsulated into standardized software containers.

The task of an ID predictor is to assign a score to each residue for its propensity to be intrinsically disordered at any stage of the protein's life, given a protein sequence. Another class of predictors is trained to identify binding sites within IDRs. We present an evaluation of the accuracy of prediction methods and a comparison between the accuracy and software runtimes, which directly impact their suitability for large-scale analyses.

Interactive figures with the curves for all methods are available at URL: https://caid.idpcentral.org/challenge#Benchmarking.

A brief description of the majority of methods present in this assessment can be found at the following URL: https://caid.idpcentral.org/overview.

3.1 | Disorder prediction performance

Precision–recall and ROC curves of top methods for all three benchmarks are shown in Figure 2. The legend reflects the ranking based on average precision score (APS) for the precision–recall, and the area under the ROC (AUC) for the ROC curves. In all plots is also reported the coverage for each method, indicating the fraction of targets for which an output was generated. We also evaluated the performance considering the subset of proteins predicted by all methods (not shown) but the ranking did not change.

Precision–recall curves provide insight into how the precision correlates with the recall. Methods where the prediction monotonically decreases with an increase of the recall, indicate that the prediction score well correlates with the precision, that is, high scoring sites are more likely to be disordered in the reference.

In the Disorder-NOX benchmark, there is a distinction between methods that exhibit a steep precision–recall curve and methods that maintain a constant, usually lower, precision, in particular at low recall (Figure 2, panel A). This phenomenon is also evident, albeit less prominently, when comparing ROC curves (Figure 2, panel B), particularly with regard to the true positive rate in the region where the false positive rate is low. Methods with a high precision at low recall, are the

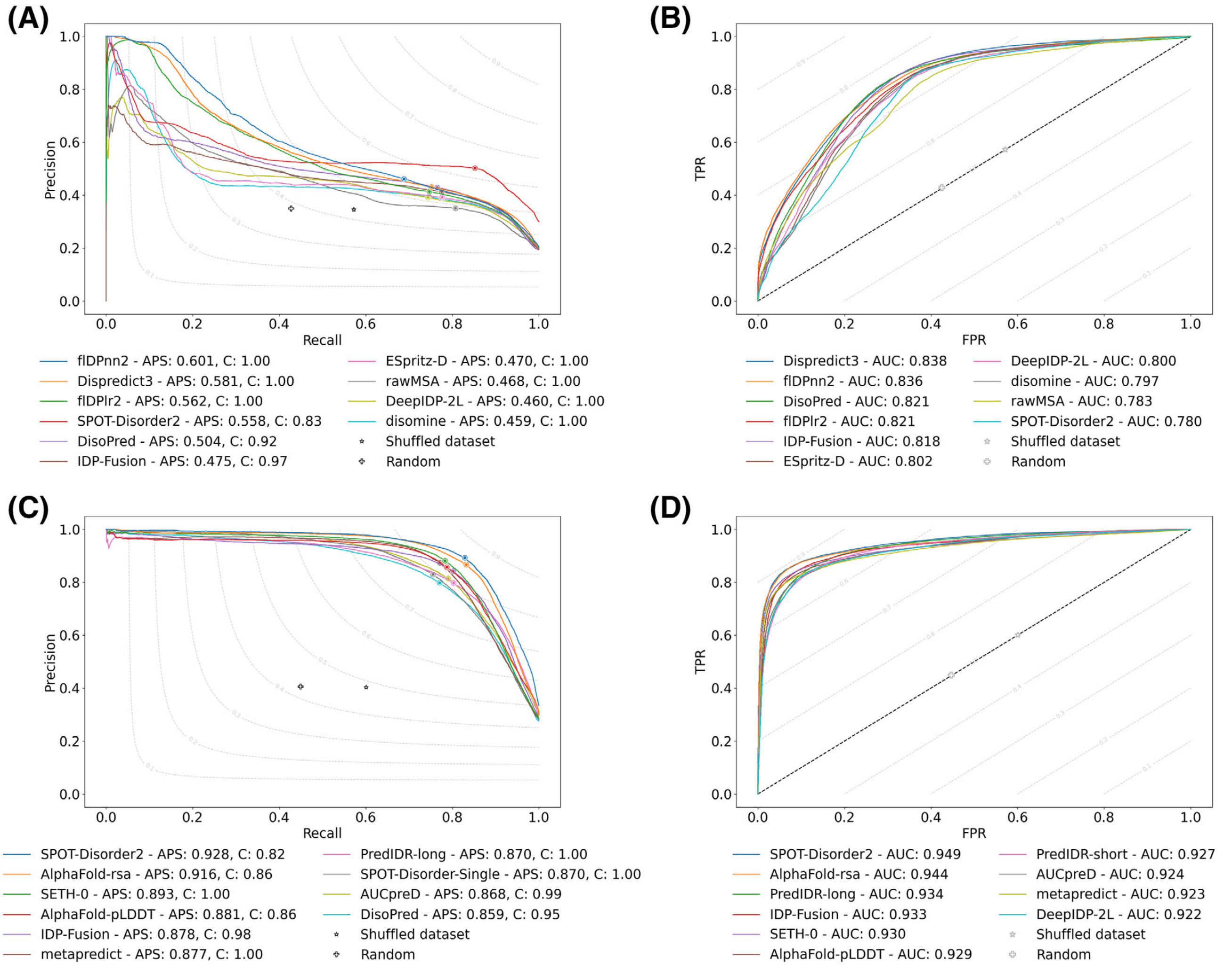

FIGURE 2 Disorder prediction evaluation for the 10 top-ranking methods. Left (A, C), precision–recall curves. Right (B, D), receiver operating characteristic (ROC) curves. The evaluation is reported for the Disorder-NOX (panels A and B, $n = 210$) and Disorder-PDB (panels C and D, $n = 348$) reference sets. In the legend "C" represents the coverage, that is, the fraction of predicted proteins. The points highlighted in panels A and C represent the $F_{\max}$.

top-ranking when considering the AUC and have a higher APS, where the APS provides a good approximation of the area under the precision–recall curve.

As already noted in CAID1, the precision–recall and ROC curves and method's ranking were substantially different when predictors were tested on the Disorder-PDB dataset (Figure 2, panels C and D), which does not contain uncertain residues, as opposed to the Disorder dataset. This reference pertains to an application that aims to predict disordered protein fragments by using examples in the PDB, which is distinct from the prediction of functional IDRs that involves assessing their biophysical characteristics.

When the precision is constant in the precision–recall curve, it indicates there is a high fraction of false positives even when the prediction score is high. This can be explained by the fact that methods poorly discriminate disordered regions, despite capturing most of them when decreasing the prediction score threshold, or by the fact that disorder annotation in DisProt is incomplete and part of the false positives are simply mislabeled as negatives in the reference. On the other end, methods that perform particularly well at low recall could overfit DisProt features.

Both AlphaFold-disorder methods, which are based on pLDDT and RSA, perform particularly well in the Disorder-PDB benchmark (Figure 2, panels C and D), but less well in the Disorder-NOX (Figure 2, panels A and B) and they are not in the top 10 when considering APS. It seems that neither the original AlphaFold2 prediction score (pLDDT), nor the RSA values derived from the predicted structure, are not able to correctly prioritize disordered positions as defined in the Disorder-NOX reference.

Other methods behave exactly the opposite. For example, Dispredict3, which has the best AUC in Disorder-NOX, performs poorly in the Disorder-PDB benchmark, with a 5% lower AUC compared with the best method. Notably, none of the methods perform well on both references indicating they represent slightly different problems and methods are not designed to be generic.

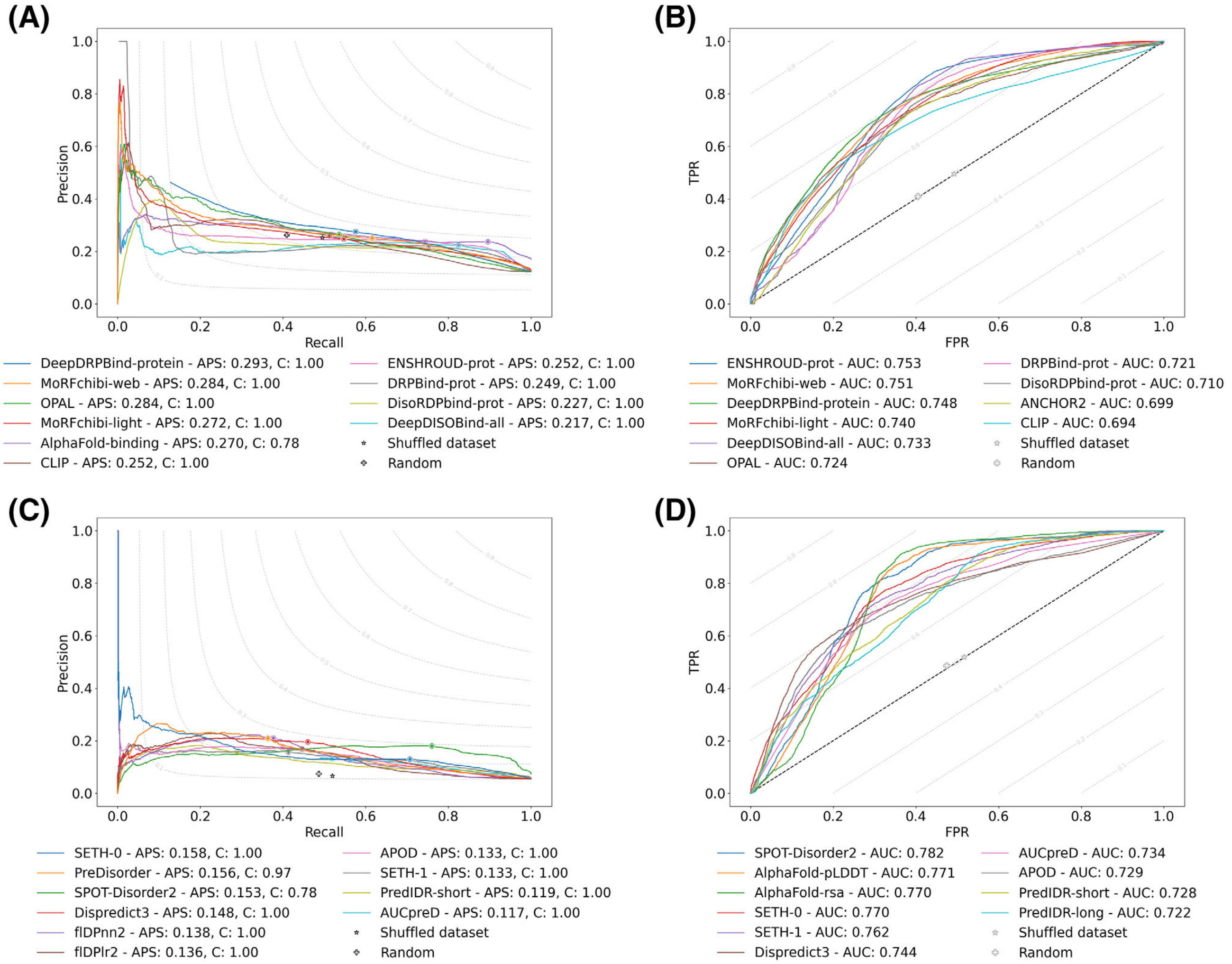

FIGURE 3 Binding and linker prediction evaluation for the 10 top-ranking methods. Left (A, C), precision–recall curves. Right (B, D), receiver operating characteristic (ROC) curves. The evaluation is reported for the Binding (panels A and B, $n = 78$) and Linker (panels C and D, $n = 40$) reference sets. In the legend "C" represents the coverage, that is, the fraction of predicted proteins. The points highlighted in panels A and C represent the $F_{\max}$.

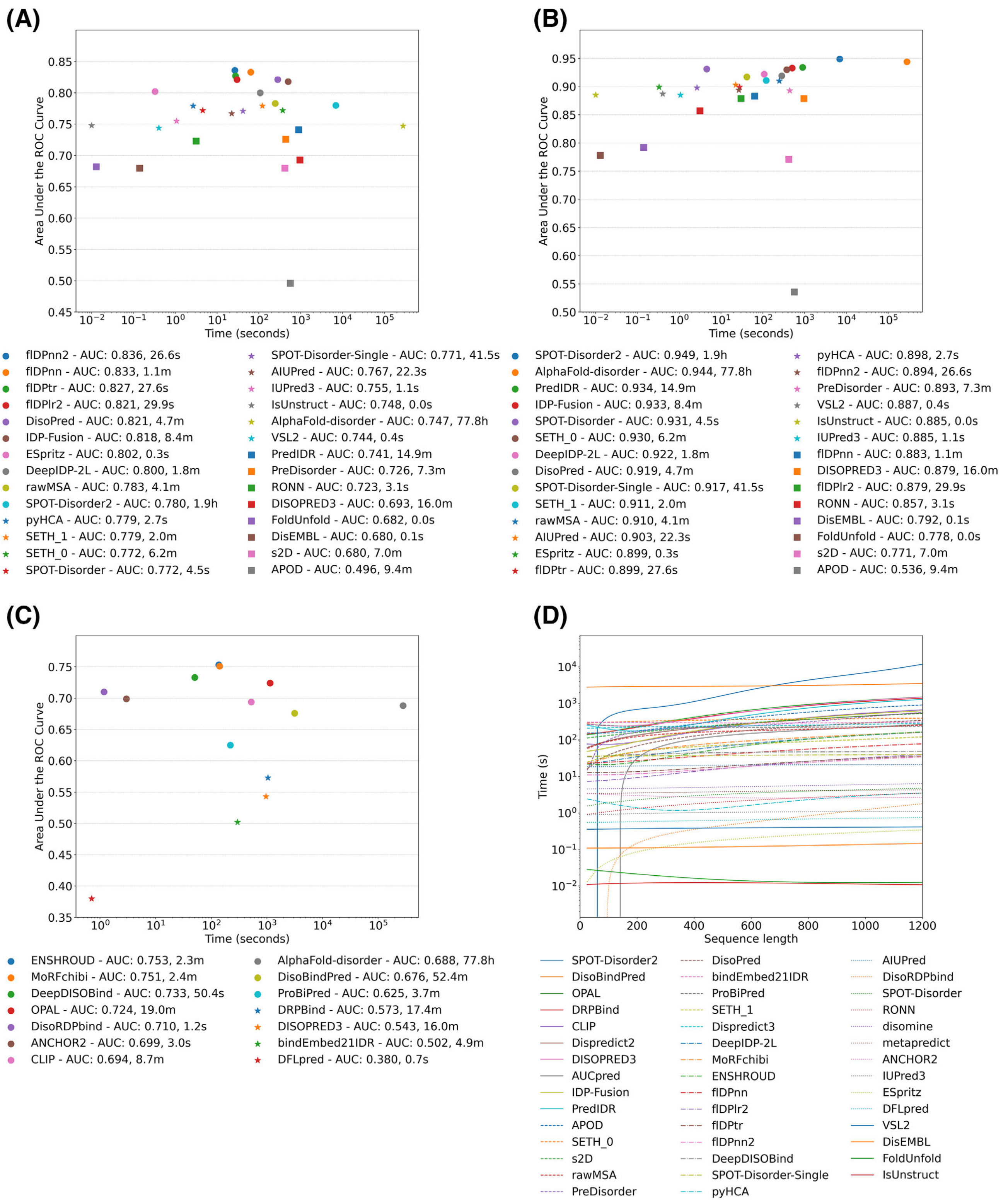

FIGURE 4 The AUC (area under the receiver operating characteristic [ROC] curve) and prediction time in seconds calculated for a sequence of 1000 residues length are compared in panels A, B, and C for the Disorder-NOX, Disorder-PDB, and Binding reference sets, respectively. The predictors are considered as a single package, without differentiation between flavors. For these panels, in the legend is reported the AUC (A) and the time. The displayed AUC represents the best result obtained for each reference set. Panel D provides the execution time (in seconds) of the predictor software in relation to the sequence length.

The definition of the Disorder-PDB dataset coincides with the definition of the DisProt-PDB dataset in CAID1, therefore, it is possible to make a direct comparison. For example, considering those methods that were evaluated in both CAID editions and with 100% coverage, such as AUCpreD,[19] we noted an improvement of 3% and 1.8% in terms of F_{max} and AUC, respectively (Figure 2, panels C and D). This indicates that in CAID2 the reference is closer to predictions compared with CAID1.

Considerations about methods performance in relation to the type of input and execution time are provided in Section 2.5.

3.2 | Prediction of disordered binding sites and linker regions

In addition to the evaluation of binding sites within IDRs, in CAID2 we included the assessment of the prediction of linkers, that is, disordered regions connecting two structured domains. A growing number of methods are designed to predict nucleic acid binding (DNA, RNA, or both), a dedicated nucleic acid binding benchmark would have included only nine proteins, so we decided to include all of them in the general binding assessment. Since the limited number of methods designed to predict linker regions, we considered all disorder predictors for the Linker assessment.

In contrast to CAID1, we built the binding reference considering only proteins with at least one binding region, therefore, it is difficult to make a fair comparison with CAID1. Despite the higher class imbalance compared to the disorder references, in CAID2 the binding sites cover the 12.2% of the residues, where it was 6.3% in CAID1. In general, methods seem to perform better both in terms of $F_{\max}$ and AUC, but as observed for disorder prediction, the method's ranking for the two metrics is substantially different (Figure 3, panels A and B). Notably, AlphaFold-binding, which prioritizes regions with high structural propensity (pLDDT) and high relative solvent accessibility (RSA), has the best $F_{\max}$ and a good APS but falls in the 13th position when considering AUC. Three methods, MoRFchibi-web[20] and DeepDRPBind-protein, seem to better capture the reference definition. They are in the top three in terms of both APS and AUC, but also have a better true positive rate at low false positive rate and reach higher precision when the prediction score is high.

Linkers are likely to carry specific structural features, being characterized by an extreme extended conformation compared to other IDRs and some methods, for example, APOD,[21] have been developed to specifically detect linker regions. The performance of disorder predictors over the Linker dataset is reported in Figure 3, panels C and D. The linker dataset is even smaller and more unbalanced than the binding reference. Despite the poor performance, all methods have an AUC well above random, indicating linker regions encompass some properties learned by the predictors.

3.3 | Software evaluation

The speed of a predictor is an important factor, particularly when analyzing large-scale genomic data. A predictor that can quickly and accurately identify IDRs in a genome-scale dataset can greatly facilitate the analysis of protein function and regulation. The quality of IDR prediction can be evaluated in various ways depending on the specific application. The fraction of disorder is an important metric when the goal is to understand the overall complexity of an organism or to identify proteins with a high degree of disorder. On the other hand, the exact position of the IDR in the sequence is crucial for applications such as protein structure prediction or drug design.

The execution times of disorder predictors (Figure 4, panels A and B) have very different scales and are inversely proportional to their performance, with the best methods requiring more computational time. However, the AUC gain is marginal compared to the magnitude of computation required to improve the performance. As an example, ESpritz[22] achieves an AUC that is approximately 3% lower than the top-performing method, but it only takes a fraction of the computation time, with two orders of magnitude less required. For Binding methods there is not such a relationship between the execution time and the quality of the prediction (Figure 4, panel C).

Figure 4 panel D, shows the execution time in relation to the sequence length. The collected data points were fitted using a polynomial function of degree five through a simple linear regression method to generate trendlines. These statistics explain well the limitations and the computational complexity of the various methods.

4 | CONCLUSIONS

ID is a complex phenomenon that covers a continuum between fully disordered states and folded states with long dynamic regions. Therefore, it is difficult to define a ground truth that fits all ID aspects, moreover ID can be context dependent and proteins can undergo order-to-disorder transitions depending on specific conditions, for example, binding a partner molecule.

Disorder prediction was previously assessed in CASP, but this was abandoned due to the lack of good reference data. Specifically, from CASP5 to CASP10 disorder was defined from missing residues in X-ray experiments which are usually short and represent only a portion of the non-resolvable structure.[23–28] To overcome this problem, in CAID we leverage the annotation provided by the DisProt database which stores the position of IDRs along the protein sequence when there is experimental evidence in the literature. DisProt is manually curated and spans more than 50 biochemical methods which provide orthogonal measures to X-ray data. In addition, DisProt annotates binding sites inside IDRs and flags those IDRs connecting two structured domains as linkers.

ID prediction was evaluated in two different ways. The first assessment hypothesizes DisProt annotations are complete and everything not annotated as IDR is negative. In the second scenario the fraction of negative residues not observed in PDB structures are excluded. For binding regions the problem of missing annotations is even more relevant. In CAID2 for the binding and linker benchmarks we considered only proteins with at least one annotated binding or linker region, respectively.

None of the methods performed well on both the Disorder-NOX and Disorder-PDB references, indicating that they represent slightly different problems and methods are not designed to be generic. The AUC and the APS give the same ranking while considering the same reference, instead when considering the $F_{\max}$ on Disorder-NOX notably SPOT-Disorder2[29] and AlphaFold-rsa stand out. The two methods probably suffer a poor prioritization of the DisProt regions but there exists an optimal cutoff that maximizes the trade-off between precision and recall over other methods. Methods with top AUC in Disorder-NOX have been likely trained on DisProt data, whereas other methods are instructed to predict absence of order, that is, learning the PDB complement.

Comparing the results of CAID1 and CAID2 in the Disorder-PDB, there is an overall improvement in performance. Since some methods are the same, the improvement can be explained by differences in references.

Methods in general seem to perform better in terms of $F_{\max}$ and AUC for binding site prediction. MoRFchibi-web[20] and DeepDRPBind-protein seem to capture the reference definition better, as they are in the top-three in terms of both APS and AUC.

Despite the poor performance of disorder predictors on the Linker dataset, all methods have an AUC well above random, indicating that linker regions encompass some properties learned by the predictors.

Depending on the application, researchers may need to balance performance with execution time when selecting a predictor. The execution times of disorder predictors are inversely proportional to their performance. However, the AUC gain is marginal compared to the magnitude of computation required to improve the performance. For Binding methods, there is no such relationship between the execution time and the quality of the prediction. For example, ESpritz[22] achieves an AUC that is approximately 3% lower than the top-performing method but requires significantly less computation time.

In summary, the CAID2 challenge demonstrates the varying performance of different prediction methods across different benchmarks and highlights the need for continued development of more versatile and efficient prediction software.

Most of the CAID2 predictors can be freely used through the CAID Prediction Portal available at https://caid.idpcentral.org/.

Finally, CAID has been integrated into OpenEBench (https://openebench.bsc.es/) which will become the official platform for running future CAID challenges.

AUTHOR CONTRIBUTIONS

Alessio Del Conte: Investigation; software; methodology; validation; conceptualization; writing – original draft. **Mahta Mehdiabadi:** Methodology; validation; software. **Adel Bouhraoua:** Software; methodology; validation. **Alexander Miguel Monzon:** Conceptualization; methodology; supervision. **Silvio C. E. Tosatto:** Conceptualization; funding acquisition; writing – original draft; methodology; writing – review and editing; supervision; investigation. **Damiano Piovesan:** Conceptualization; investigation; writing – original draft; writing – review and editing; methodology; validation; software; formal analysis; project administration; supervision.

ACKNOWLEDGMENTS

This publication is part of a project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements No. 778247 and No. 823886; and from the ML4NGP CA21160 project supported by COST (European Cooperation in Science and Technology) under Horizon Europe. This work was also supported by ELIXIR, the research infrastructure for life-science data.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in CAID Portal at https://caid.idpcentral.org/challenge#Data.

CODE AVAILABILITY

The code to produce references and dataset statistics is available in the caid2-reference GitHub repository at URL https://github.com/BioComputingUP/caid2-reference. Some of the CAID2 methods are available as web services in the CAID Prediction Portal at URL https://caid.idpcentral.org/. Results of the CAID assessment can be fully reproduced by downloading the code and following the instructions in the CAID GitHub repository at https://github.com/BioComputingUP/CAID (v2 branch).

ORCID

Damiano Piovesan https://orcid.org/0000-0001-8210-2390

REFERENCES

1. Wright PE, Dyson HJ. Intrinsically disordered proteins in cellular signalling and regulation. *Nat Rev Mol Cell Biol*. 2015;16:18-29.
2. Felli IC, Pierattelli R. *Intrinsically Disordered Proteins Studied by NMR Spectroscopy*. Springer; 2015.
3. Sormanni P, Piovesan D, Heller GT, et al. Simultaneous quantification of protein order and disorder. *Nat Chem Biol*. 2017;13:339-342.
4. Dyson HJ, Wright PE. Coupling of folding and binding for unstructured proteins. *Curr Opin Struct Biol*. 2002;12:54-60.
5. Necci M, Piovesan D, CAID Predictors, DisProt Curators, Tosatto SCE. Critical assessment of protein intrinsic disorder prediction. *Nat Methods*. 2021;18:472-481.
6. Liu Y, Wang X, Liu B. A comprehensive review and comparison of existing computational methods for intrinsically disordered protein and region prediction. *Brief Bioinform*. 2019;20:330-346.
7. Quaglia F, Mészáros B, Salladini E, et al. DisProt in 2022: improved quality and accessibility of protein intrinsic disorder annotation. *Nucleic Acids Res*. 2022;50:D480-D487.
8. PDBe-KB consortium. PDBe-KB: collaboratively defining the biological context of structural data. *Nucleic Acids Res*. 2022;50:D534-D542.
9. Piovesan D, Monzon AM, Tosatto SCE. Intrinsic protein disorder and conditional folding in AlphaFoldDB. *Protein Sci*. 2022;31:e4466.
10. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021;596:583-589.
11. Varadi M, Anyango S, Deshpande M, et al. AlphaFold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res*. 2022;50:D439-D444.
12. UniProt Consortium. UniProt: the universal protein knowledgebase in 2023. *Nucleic Acids Res*. 2023;51:D523-D531.
13. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*. 1983;22:2577-2637.
14. Altschul SF, Madden TL, Schäffer AA, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. 1997;25:3389-3402.

15. Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH, UniProt Consortium. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*. 2015;31:926-932.
16. Steinegger M, Meier M, Mirdita M, Vöhringer H, Haunsberger SJ, Söding J. HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics*. 2019;20:473.
17. Yang Y, Heffernan R, Paliwal K, et al. SPIDER2: a package to predict secondary structure, accessible surface area, and Main-chain torsional angles by deep neural networks. *Methods Mol Biol*. 2017;1484:55-63.
18. Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold: making protein folding accessible to all. *Nat Methods*. 2022;19:679-682.
19. Wang S, Ma J, Xu J. AUCpreD: proteome-level protein disorder prediction by AUC-maximized deep convolutional neural fields. *Bioinformatics*. 2016;32:i672-i679.
20. Malhis N, Jacobson M, Gsponer J. MoRFchibi SYSTEM: software tools for the identification of MoRFs in protein sequences. *Nucleic Acids Res*. 2016;44:W488-W493.
21. Peng Z, Xing Q, Kurgan L. APOD: accurate sequence-based predictor of disordered flexible linkers. *Bioinformatics*. 2020;36:i754-i761.
22. Walsh I, Martin AJM, Di Domenico T, Tosatto SCE. ESpritz: accurate and fast prediction of protein disorder. *Bioinformatics*. 2012;28:503-509.
23. Melamud E, Moult J. Evaluation of disorder predictions in CASP5. *Proteins*. 2003;53(Suppl 6):561-565.
24. Jin Y, Dunbrack RL. Assessment of disorder predictions in CASP6. *Proteins*. 2005;61(Suppl 7):167-175.
25. Bordoli L, Kiefer F, Schwede T. Assessment of disorder predictions in CASP7. *Proteins*. 2007;69(Suppl 8):129-136.
26. Noivirt-Brik O, Prilusky J, Sussman JL. Assessment of disorder predictions in CASP8. *Proteins*. 2009;77(Suppl 9):210-216.
27. Monastyrskyy B, Fidelis K, Moult J, Tramontano A, Kryshtafovych A. Evaluation of disorder predictions in CASP9. *Proteins*. 2011;79(Suppl 10):107-118.
28. Monastyrskyy B, Kryshtafovych A, Moult J, Tramontano A, Fidelis K. Assessment of protein disorder region predictions in CASP10. *Proteins*. 2014;82(Suppl 2):127-137.
29. Hanson J, Paliwal KK, Litfin T, Zhou Y. SPOT-Disorder2: improved protein intrinsic disorder prediction by ensembled deep learning. *Genomics Proteomics Bioinformatics*. 2019;17:645-656.

How to cite this article: Conte AD, Mehdiabadi M, Bouhraoua A, Miguel Monzon A, Tosatto SCE, Piovesan D. Critical assessment of protein intrinsic disorder prediction (CAID) - Results of round 2. *Proteins*. 2023;1-10. doi:10.1002/prot.26582