Probabilistic Annotations of Protein Sequences for Intrinsically Disordered Features

Nawar Malhis
nmalhis@msl.ubc.ca
Michael Smith Laboratories, University of British Columbia
Vancouver, BC V6T 1Z4, Canada

Abstract

This paper introduces a novel platform for IDR Probabilistic Annotation (IPA). The IPA platform now encompasses tools for predicting 'Linker' regions and 'nucleic', 'protein', and 'all' (protein or nucleic) IDR binding sites within protein amino acid sequences. Despite its simplicity and computational efficiency, results demonstrate that IPA performs competitively with leading tools in predicting 'protein' and 'all' IDR binding sites while considerably outperforming all tools in identifying Linker regions and nucleic binding sites. An important contribution of this work is the introduction of a new output paradigm for computational feature predictions. Traditional tools typically express predictions as scores, with higher values indicating greater probabilities. However, these scores lack true probabilistic meaning and interpretability, even derived from logistic regression models. This limitation arises primarily because training data priors differ from broader populations' unknown priors. This paper proposes applying a reverse Bayes Rule to logistic regression outputs, effectively normalizing for the priors in the training data. This adjustment produces scores representing actual probabilities, assuming 50% priors in the general population. Such scores are interpretable in isolation and enable comparability and integration across different tools, marking a significant step toward standardization in feature prediction methodologies.

Availability: orca.msl.ubc.ca/nmshare/ipa.tar.gz

1. Introduction

Intrinsically disordered protein regions (IDRs) are segments of proteins that lack a fixed or stable three-dimensional structure under physiological conditions [1][2][3]. They are rich in polar and charged amino acids (e.g., Serine, Glutamine, Lysine) and deficient in hydrophobic residues that drive folding in structured proteins [4]. IDRs are prevalent in eukaryotic proteomes, with an estimated 20% of eukaryotic protein residues being disordered; many are associated with diverse biological functions [5].

IDR interactions are characterized by high specificity and low affinity, i.e., reversible [6], making them suitable for signaling pathways and regulation processes, including allosteric regulation [7]. Upon binding to folded protein domains, IDRs adopt conformations complementary to their binding partners; they often fold upon binding, a phenomenon known as coupled folding and binding [8][9]. Interactions between IDRs and other IDR partners display diversity in conformational outcomes, broadly categorized into two types: 1) fuzzy interactions, in which IDRs retain a significant degree of conformational flexibility in the bound state [10]. 2) folding upon binding, where IDRs undergo coupled folding and binding, transitioning from their disordered state to a well-defined structure [11]. This flexibility allows IDRs to recruit various binding

partners, acting as hubs in protein interaction networks [12]. IDRs are often regulated by post-translational modification (PTM), such as phosphorylation, acetylation, and ubiquitination [13].

Annotating protein sequences with IDR interaction sites and other features is essential for understanding their roles in biological systems and their contribution to biological processes. Manually curated annotations sourced from experimental results are considered the highest-quality annotations. However, like other protein annotations, such annotations are limited and only cover a tiny proportion of the overall IDR features. Note that the percentage of the manually curated SwissProt sequences in the latest release of the UniProtKB database (2024_06) is 0.23% [14]. While SwissProt sequences are selected among those enriched in molecular functions, these sequences are not fully annotated. Thus, we can reasonably assume that the percentage of manually curated molecular function sites is comparable, i.e., ~1%. Consequently, understanding unannotated protein sequences relies heavily on high-throughput computational tools [15 … 22].

Various computational tools have been developed for annotating protein sequences within intrinsically disordered regions [23 .. 50]. Several attempts have been made to understand and assess the prediction quality of these tools [51 .. 57], most notably the community-based Critical Assessment of protein Intrinsic Disorder prediction (CAID)[51][56]. CAID utilizes newly annotated sequences from the DisProt database [58] to evaluate the predictions of IDR and IDR-binding sites. CAID test datasets are fair as they only include new manually curated annotations released after the training of the evaluated predictors. My only concern with respect to the CAID test datasets is that they do not account for the known false negative sites by limiting their annotations to newly unreleased ones. For example, if we extend the annotations of the 3,022 DisProt (2024_06) sequences by the annotations of the MobiDB-LIP [59] database, the number of protein-binding residues increases by more than a third from 68,925 to 82,394.

Furthermore, DisProt annotations used are limited to two states, True/False, e.g., protein binding vs. non-binding. Including a third 'unknown' state to represent potentially True residues can help reduce false negatives. Marking low-quality annotations of protein-binding IDR from the MobiDB-LIP database and 'ambiguous' annotations of the DisProt, as described in **Section 2.2**, converts the annotation of a total of 237,656 residues from non-protein-binding to 'unknown', i.e., decreases the number of the 'negative' state residues by ~14.5%.

Apart from the annotation, the non-standardized outputs of prediction tools can lead to a potential misinterpretation. Generally, most tools provide two output forms: an analog score and a binary prediction obtained by thresholding the analog score. The analog score is usually called a propensity score, and the definition of 'propensity', if provided, can vary, as it can be defined as probability [26] and used interchangeably or more broadly as a numerical value such that residues with higher propensities are more likely to be the targeted feature residues [49]. **Figure 1** illustrates the lack of uniformity in the distribution of IDR-protein binding propensity scores generated by three widely used prediction tools against the sequences in the TEST_24_06 dataset (section 2.2). This lack of uniformity limits the interpretation of these predictions to the relative likelihood of binding to protein targets, as the absolute score values provide little information. In addition, these predicted scores are not directly comparable. Binary predictions aim to simplify interpretation by offering definitive values. However, the absence of a standardized thresholding mechanism leads to different tools predicting significantly varying percentages of the input as positive residues, further complicating the interpretations. To overcome these issues, the authors of the CAID portal server [60] provided alternative threshold values, enabling the user to choose

to replace the predictor's default threshold with optimized thresholds corresponding to a selection of metrics.

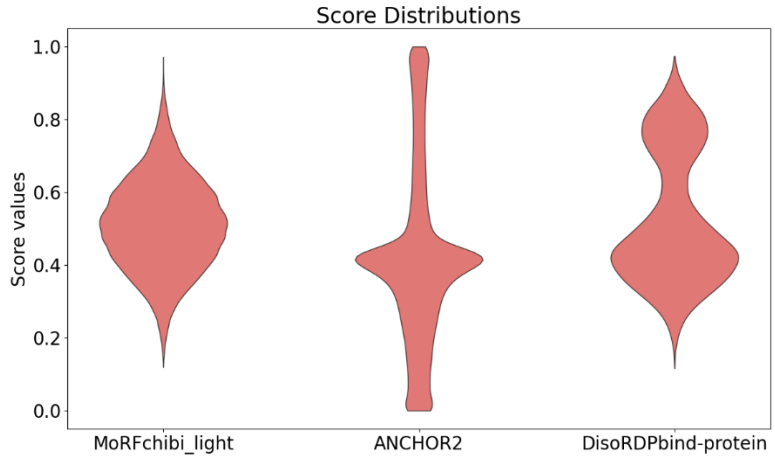

Figure 1: Violin plots for the propensity scores of MoRFchibi_light, ANCHOR2, and DisoRDPbind-protein against the TEST-24_04 dataset, reflecting substantial inconsistencies of these scores.

In this paper, I present a platform for predicting annotations related to disordered protein regions from protein sequences. Because the prior for each category is unknown, i.e., the percentage of positive in the general population, I designed IPA predictions to generate, to the degree possible, probability values assuming 50% priors. Consequently, the scores of IPA predictions can be understood individually, not just as relative to other scores. And because IPA probabilities are simple and easily interpretable, IPA outputs do not include binary values. When binary predictions are needed, users can threshold IPA scores appropriately to their needs. At this point, IPA generates predictions for IDR protein-binding sites, IDR nucleic-binding sites, and IDR all-binding sites, where 'all' refers to nucleic or amino acids. In addition, I also created a second flavor of IPA that utilizes AlphaFold-2 predicted structures, thus called IPA-AF2, to predict protein-binding sites and Linker regions.

But before going into the details of IPA's predictions, let's first define IDR binding sites in general. For an IDR site to bind to any partner, the targeted partner must be available within the vicinity of the site, and other environmental conditions must be favorable, e.g., organism, subcellular location, temperature, pressure, phosphorylation state, etc. However, tools are optimized solely based on the protein's annotated amino acid sequences, and binding predictions are based solely on these sequences. Therefore, the best we can achieve is to predict sites with the potential of binding if everything else is available. Consequently, homologous sequences will have similar predictions, even though they might not exist within the correct species or the intended binding favorable condition. In other words, from the computational viewpoint, homologs to a protein binding site are protein binding sites regardless of their actual molecular function.

2. Results

Even though the scores generated by many tools are outputs of logistic regression models, these scores are not likely to represent actual probabilities because of the annotation incompleteness of the training

datasets, and the composition of positives and negatives in these datasets are not expected to represent the unknown composition of the general population. Plus, and most importantly, the percentages of 'positive' in the training datasets, i.e., the priors, do not reflect those percentages in the general populations, which are also unknown. To reduce the effect of the annotation incompleteness and compositions to the degree possible, I assembled training and testing datasets by integrating the annotations of multiple databases into three-state annotation data. High-quality annotations represent the 'Positive' state, low-quality the 'Unknown' state, and the remaining residues are in the 'Negative' state. To resolve the unknown priors issue, I applied a reversed Bayes rule to the IPA outputs generated using logistic regression models to factor out the priors of the training data. Thus, IPA's outputs are probabilities, assuming 50% priors.

2.1 Training datasets:

I downloaded all 2,845 sequences of then (March 2024) the latest release, 2023_12, of the DisProt database and selected protein-binding (GO:0005515), nucleic acid, or just nucleic binding (GO:0003676, GO:0003723, and GO:0003677), and Linker regions (Intrinsically Disordered Proteins Ontology, IDPO:00502). I complemented the protein-binding annotations of the DisProt sequences with manually curated annotations from the IDEAL [61] and the DIBS [62] databases. I assembled a dataset with three-state annotations: manually curated annotations are in the 'positive' class, DisProt Ambiguous annotations are assigned to the 'unknown' class, and the remaining residues are in the 'negative' class. I used CH-HIT [63] to cluster sequences at 90% identity and only retained the cluster centers. The annotations of sequences at 100% identity are transferred to cluster centers. I used HAC [64] to mask, i.e., convert to 'unknown', conflicting annotations separately in each category. For training cross-validation (N=4), I divided the sequences for each of the 'linker', 'protein', 'nucleotides', and 'all' (protein and nucleic binding) IDR-binding annotations at random into four groups, such that the number of binding/Linker residues and the number of binding-sites and linker-regions are about equal in each group. Finally, for each of the four groups, I doubled (for 'linker', 'protein', and 'all') and tripled (for nucleic) its sequences by adding random sequences from the remaining DisProt sequences.

2.2 Test dataset:

For the test dataset, I include all descendants of each GO term for each annotation category to increase the data size. First, I included all of the 185 new DisProt release 2024_06 sequences that are not in the 2023_12 release, **Supplementary Table S1** (DSET1). To eliminate potential data contamination related to test sequences homologous to those used in training, I used CD-HIT to filter out test sequences with more than 30% identity to those used in training, **Supplementary Table S1** (DSET2). Then, I extend the annotations of these sequences from the MobiDB-LIP database. MobiDB integrates manually curated IDR-binding annotations from FuzDB [10], IDEAL [61], DIBS [62], MFIB [11], ELM [65], and DisProt databases, homologous to these annotations, as well as high-throughput annotations derived from PDB structures by Flipper [66]. In line with my earlier discussion, MobiDB considers the quality of homology annotations comparable to those curated and much higher than that of the derived. Thus, I included all the curated and homology annotations for DIBS, IDEAL, and MFIB for protein binding and masked out the derived annotations. I also masked out the curated and the homology for FuzDB and ELM as they include non-protein binding sites, **Supplementary Table S1** (DSET3). While CD-HIT excluded sequences with more than 30% identity to the training data, short regions homolog to the training sequences can still exist in the test data. Thus, I used HAM [64] to mask out homolog regions equal to or longer than ten residues with an

identity to the training sequences higher than 80%. **Supplementary Table S1** (DSET4). The resulting test dataset is small and insufficient to provide quality evaluation for the five IPA predictors, especially the linker regions and the nucleotide-binding sites. Thus, I also included all new 2024_06 nucleic-binding, protein-binding, and linker regions annotations to sequences in the 2023_12 release after masking out the remaining residues in these sequences; see an example in **Figure 1**.

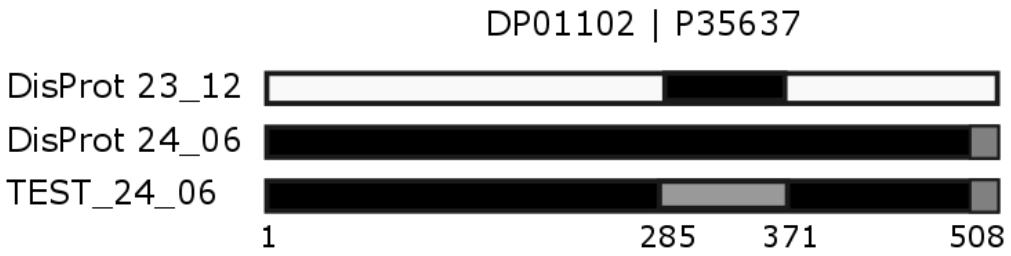

Figure 1: Annotation of nucleic-binding IDR residues in the DisProt DP01102 (UniProt P35637). DP01102 residues 285 to 370 were annotated as a nucleic-binding IDR site in the DisProt release 23_12 (black). DP01102 annotation is revised in the 24_06 release to include all 526 residues. However, residues from 508 to 526 are not confirmed to be disordered under physiological conditions. Thus, in the 24_06 release, residues from 1 to 507 are IDR-nucleic-binding (black), and 508 to 526 are 'unknown' (gray). TEST-24_06 nucleic-binding residues for DP01102 are those in the 24_06 release and not in the 23_12 release. Thus, nucleotide-binding residues are from 1 to 284 and 371 to 507; the remaining are 'unknown' (gray). 'Black' represents nucleic binding, 'gray' represents unknown, and 'white' represents non-nucleic binding.

The resulting TEST_24_06 dataset includes 184 sequences, 42 annotated with protein binding sites, 11 with nucleic binding sites, 48 with all binding sites, and 14 with linker regions. **Supplementary Table S1** (TEST_24_06). IPA-AF2 predictors require structures generated by AlphaFold-2 from the AlphaFold database; thus, they can only process a subset of the TEST-24_06 sequences, TEST-24_06-AF2. **Table 1**.

Table 1: The Test-24_06 dataset. Test-24_06-AF2 is a subset of the TEST-24_06 sequences with structures generated by AlphaFold-2 in the AlphaFold database.

		Total sequences	Annotated sequences	Target sites	Non-target residues	Target residues	Masked-out residues
Test-24_06	**P_bind**	184	42	55	68,719	3,873	28,742
	N_bind	184	11	13	88,279	1,183	11,872
	LIP	184	48	65	68,445	4,696	28,193
Test-24_06-AF2	**P_bind**	147	35	47	58,985	3,436	13,123
	LIP	147	40	55	56,567	4,220	11,757
	Linker	144	14	18	67,519	619	584

2.3 Comparison with existing predictors

I contrasted the predictions of each one of the IPA predictors with the leading relevant tools available against the TEST-24_06 dataset. **Figure 2** and **Supplementary Figure S1** show the ROC curves for the IPA

and IPA-AF2 predictions compared to other tools. **Table 2** and **Supplementary Table S2** show that IPA and IPA-AF2 predictors outperformed or matched others in AUC values.

Table 2: The AUC values for the IPA predictions compared to other tools. The highest AUC value and those within one AUC point are in bold for each column. (*) The DisoFLAG predictor does not generate predictions for nucleic-binding but generates predictions for RND-binding (RB) and DNA-binding (DB); thus, I averaged the scores of RB and DB as DisoFLAG-nucleic. It also does not generate predictions for all-binding; thus, I averaged the scores of DisoFLAG-nucleic and DisoFLAG-PB as DisoFLAG-all.

	TEST-24_06 dataset			TEST-24_06-AF2 dataset		
	All-bind (LIP)	Nucleotide-bind	Protein-bind	All-bind (LIP)	Protein-bind	Linker
IPA	**0.851**	**0.873**	**0.844**	**0.848**	**0.840**	
IPA-AF2					**0.850**	**0.861**
DisoFLAG	0.833*	0.743*	**0.843**	0.813*	0,831	
ENSHROUD	0.826	0.767	0.812	0.829	0.816	0.752
AlphaFold_binding				0.831		
DeepDISOBind	0.754	0.654	0.717	0.771	0.731	
APOD						0.786
MoRFchibi_Light			0.790		0.781	
MoRFchibi_web			0.756		0.745	
ANCHOR2			0.749		0.778	
DisoRDPbind			0.755		0.768	
OPAL			0.759		0.727	
ProBiPred		0.599	0.400		0.365	

Even though IPA-nucleic, with an AUC=0.873, is ahead of other tools and more than 9 points higher than the second tool, DisoFLAF-RB (AUC=0.782), the zoomed-in lower left corner of the ROC curve shows that the IPA-nucleic curve is flat with a TPR=0 for FPR up to 0.0161. That is 1,425 residues annotated as non-nucleic-binding from 13 protein sequences in the Test-24_06 dataset predicted by IPA with a probability of nucleic-binding higher than its prediction of any residue annotated as nucleic-binding, **Table 3**. Even though the test dataset is relatively small, with only 13 binding sites in 11 protein sequences, given the overall IPA-nucleic AUC of 0.873, such a high number of false positives was unanticipated. At first, I contributed this to the UniProt P35637's 421 residues annotated as non-binding in the IPA-nucleic training data, which represents ~35% of the 1,183 test nucleic-binding residues, **Figure 1**. However, after investigating these 13 proteins, and in light of our earlier estimation that ~1% of molecular function sites are annotated, I concluded it is reasonable to assume that many of these 1,425 residues are unannotated nucleic-binding residues. Only six of these 13 proteins in **Table 3** have the UniProt maximum annotation score of 5. Even though UniProt's maximum annotation score does not imply that proteins are fully annotated, UniProt gave three of these six a molecular function of RNA/DNA binding, which accounts for 456 out of the 599 high-score residues in these six proteins, i.e., 76%.

IPA predictions are fast; using an Intel Core i9, 8/16 cores/threads CPU with 40G DDR4 RAM, and an RTX 3080 GPU machine, it took IPA 4 hours, 9 minutes, and 1 second to process all of the SwissProt (2024_06) 572,619 sequences and generate nucleotide, protein, and all binding predictions. On average, 2,300 sequences per minute or 38.3 sequences per second. When using the CPU to run IPA on the same

machine, processing these SwissProt sequences took 11 hours, 45 minutes, and 39 seconds. That is 812 sequences per minute or 13.5 sequences per second.

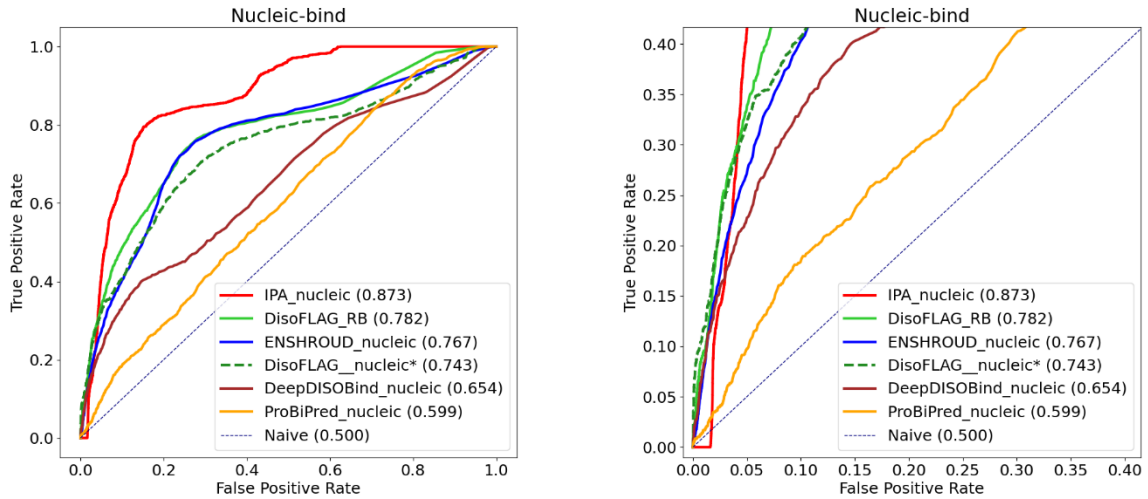

Figure 2: Comparing the full ROC curves (left) of IPA-Nucleic bind predictions to that of DisoFLAG-RB, DisoFLAG-nucleic*, ENSHROUD_nucleic, DeepDISOBind_nucleic, and ProBiPred_nuclic, and a zoom on the lower left corner of these curves (right). (*) The DisoFLAG predictor does not generate predictions for nucleic-binding but generates predictions for RND-binding (RB) and DNA-binding (DB); thus, I averaged the scores of RB and DB as DisoFLAG-nucleic.

Table 3: A list of the 13 proteins with high probabilities of IPA nucleic-binding predictions sorted in descending order first by UniProt annotation score and then the number of high-scored residues.

DisProt ID	UniProt Accession	UniProtKB Status	Annotation score	UniProt Molecular Function	High-score residue count	Protein length
DP02348	Q8N5F7	reviewed	5	DNA & RNA binding	301	415
DP03988	Q86WX3	reviewed	5	RNA binding	125	136
DP04024	P63040	reviewed	5		95	134
DP03888	P61956	reviewed	5	RNA binding	30	95
DP04118	O43665	reviewed	5		29	181
DP00253	P26645	reviewed	5		19	309
DP04006	Q9RV58	reviewed	2		111	298
DP03916	Q3E7B7	reviewed	2		68	68
DP03982	Q7Y1A0	unreviewed	2		200	210
DP04008	Q8S8Z3	unreviewed	1		168	226
DP04007	G5EFU3	unreviewed	1		140	733
DP04009	O64820	unreviewed	1		84	92
DP03923	Q9R825	unreviewed	1		55	150

3. Methods

IPA predictors employ a novel window_in/window_out approach. Each predictor utilizes an ensemble of four simple convolutional networks, CNN, models.

3.1 CNN model structure

Each CNN consists of three 1D convolutional layers separated by two 1D average pooling layers, followed by two fully connected layers. Thus, the model structure looks like this:

- 1D convolutional layer, with an input size of 510, kernel size CKZ, and depth K.
- ReLU activation function.
- 1D averagePooling layer, with a kernel size equal to APKZ.
- 1D convolutional layer, with a kernel size equal to CKZ and depth (K + 1 * D_inc).
- ReLU activation function.
- 1D averagePooling layer, with a kernel size equal to APKZ.
- 1D convolutional layer, with a kernel size equal to CKZ and depth (K + 2 * D_inc).
- ReLU activation function.
- Fully connected layer 1 (FC1), with an output size FC1Z.
- ReLU activation function.
- Fully connected layer 2 (FC2), with an output size of 10.

Each input sequence is padded at each side by 250 'blank' residues, where a 'blank' residue is a hypothetical residue with the values of all its features equal to zero. Then, the first input window starts from the first actual residue at position p=1; we substitute each residue from (p - 250) to (p + 9 + 250) by its K=5 for IPA-nucleic, K=7 features for IPA protein and all, and K=9 features for the IPA-AF2 predictors. The ten outputs of the second fully connected layer, FC2, are associated with input residues positions (p) to (p + 9). Then, we increase p by 10 to generate the second input window and continue repeating these steps until we cover all the sequence residues.

3.2 Input Features

Input features are divided into three categories: two affinity features, five DisProt features (**F5** in short), and two alphaFold-2 features. IPA nucleic only utilizes the F5 features, IPA (protein and all) utilize the affinity and the F5 features, and IPA-AF2 utilizes all nine. Affinity features are derived from P, a 20 x 20 energy predictor matrix [67]. For each position p between 1 and the last in the sequence, one of the affinity features is the sum of the energy in P between the residue at p and the w1=10 residues on each side of p. The second is the sum of this energy between w1 and w2=w1+25, from p + w1 to p + w2 on each p side. The **F5** features are the frequency of each of the 20 standard amino acids in each of the five classes in the DisProt/CAID datasets: IDR, nucleic binding (N_bind), protein binding (P_bind), and 'Linker' in the DisProt data and 'PDB' annotation in the combined CAID1 and CAID2 sequences, divided by the frequency of that amino acid in the entire dataset, **Supplementary Table S3**. Finally, the two AlphaFold-2 features are those used by the AlphaFold-binding predictor [48] and collected from the AlphaFold-2 generated structures; the alphaFold-2 relative solvent accessibility, RSA, and the AlphaFold-2 prediction accuracy score, pLDDT.

3.3 Hyperparameters optimization and the final probability predictions

For all 20 CNN models, four models in each of the five predictors, I used a BCEWithLogitsLoss, thus generating probability outputs. For each model separately, I used a cross-validation grid search to optimize the size of the convolutional kernels (CKZ), the size of the pooling kernels (APKZ), depth increases (D_inc), the output size of the first fully connected layer (FC1Z), and regularization parameters. This results in CKZ between 7 and 15 for all models, APKZs are 7 or 15, D_incs are 2 or 10, and FC1Zs are between 100

and 150. A complete list of values for CKZ, APKZ, D_inc, and FC1Z values for all 20 CNN models is available in **Supplementary Table S4**.

For each predictor, first, I applied a reversed Bayes Rule to the probability output to factor out the priors of its training data; thus, the scores for all four outputs were aligned, and then I averaged these four scores to produce the predictor probability output.

Discussion

This paper presents a new platform for the probability annotation of intrinsically disordered proteins (IPA). The IPA platform includes tools designed to predict nucleotide, protein, and all binding sites within intrinsically disordered regions (IDRs) using only amino acid sequences. A separate tool, IPA-AF2, also leverages AlphaFold-2 structural models to predict protein binding sites and linker regions. To evaluate the performance of IPA, I constructed a test dataset that is completely homology-free to IPA training data. Results demonstrate that despite relying solely on amino acid sequences and achieving high computational efficiency, IPA favorably matched leading tools in predicting protein and all IDR binding sites, and it significantly outperforms in identifying nucleic binding sites and linker regions.

Most importantly, I introduced a new output paradigm form for computational predictions of features. Currently, the output of tools predicting features is expressed as a score, where higher values indicate a greater probability. Some tools rely on heuristics to generate these scores [48][37], while others heuristically redistribute them to fit some arbitrarily selected distribution [31][34]. Although most are in the range between zero to one, these scores are not actual probabilities even when logistic regression is used, as the priors of the training data do not reflect the unknown priors of the general population. As a result, these scores are not comparable across different tools. They are also meaningless independently and can only be interpreted by comparing them to others of the same tool. As an alternative, I am suggesting the application of a reverse Bayes Rule to the output of logistic regression models to factor out the priors in the training data. As a result, the final prediction scores will reflect probabilities, assuming 50% positive in the general population. Thus, these scores are individually interpretable, and if adopted by other tools, the output of these tools can be compared and integrated.

References

1. Dunker AK, Silman I, Uversky VN, Sussman JL (December 2008). "Function and structure of inherently disordered proteins". Current Opinion in Structural Biology. 18 (6): 756–64. doi:10.1016/j.sbi.2008.10.002. PMID 18952168.
2. Oldfield CJ, Uversky VN, Dunker AK, et al. Chapter 1 - introduction to intrinsically disordered proteins and regions. In: Salvi N (ed). Intrinsically Disordered Proteins. Academic Press, Cambridge, Massachusetts, 2019, 1–34.
3. Oldfield CJ, Dunker AK. Intrinsically disordered proteins and intrinsically disordered protein regions. Annu Rev Biochem. 2014; 83:553–84. pmid:24606139.
4. Uversky VN. Intrinsically disordered proteins from A to Z. Int J Biochem Cell Biol. 2011 Aug;43(8):1090-103. doi: 10.1016/j.biocel.2011.04.001. Epub 2011 Apr 8. PMID: 21501695.
5. Peng Z, Yan J, Fan X, et al. Exceptionally abundant exceptions: comprehensive characterization of intrinsic disorder in all domains of life. Cell Mol Life Sci 2015;72(1):137–51.
6. Dunker, A. K., Brown, C. J., Lawson, J. D., Iakoucheva, L. M. & Obradovic, Z. (2002). Intrinsic disorder and protein function. Biochemistry, 41, 6573– 6582.
7. Kamerlin SC, Warshel A (May 2010). "At the dawn of the 21st century: Is dynamics the missing link for understanding enzyme catalysis?". Proteins. 78 (6): 1339–75. doi:10.1002/prot.22654. PMC 2841229. PMID 20099310.
8. Mohan A, Oldfield CJ, Radivojac P, et al. Analysis of molecular recognition features (MoRFs). J Mol Biol 2006;362:1043–59.
9. Vacic V, Oldfield CJ, Mohan A, et al. Characterization of molecular recognition features, MoRFs, and their binding partners. J Proteome Res 2007;6:2351–66.
10. Hatos A, Monzon AM, Tosatto SCE, Piovesan D, Fuxreiter M. FuzDB: a new phase in understanding fuzzy interactions. Nucleic Acids Res. 2022 Jan 7;50(D1):D509-D517. doi: 10.1093/nar/gkab1060. PMID: 34791357; PMCID: PMC8728163.
11. Fichó E, Pancsa R, Magyar C, Kalman ZE, Schád É, Németh BZ, Simon I, Dobson L, Tusnády GE. MFIB 2.0: a major update of the database of protein complexes formed by mutual folding of the constituting protein chains. Nucleic Acids Res. 2024 Nov 11:gkae976. doi: 10.1093/nar/gkae976. Epub ahead of print. PMID: 39526403.
12. Haynes C, Oldfield CJ, Ji F, et al. Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes. PLoS Comput Biol 2006; 2 (8):e100.
13. Collins MO, Yu L, Campuzano I, Grant SG, Choudhary JS (July 2008). "Phosphoproteomic analysis of the mouse brain cytosol reveals a predominance of protein phosphorylation in regions of intrinsic sequence disorder". Molecular & Cellular Proteomics. 7 (7): 1331–48. doi:10.1074/mcp.M700564-MCP200. PMID 18388127. S2CID 22193414
14. UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Res. 2023 Jan 6;51(D1):D523-D531. doi: 10.1093/nar/gkac1052. PMID: 36408920; PMCID: PMC9825514.
15. Bhardwaj, T., Saumya, K. U., Kumar, P., Sharma, N., Gadhave, K., Uversky, V. N., & Giri, R. (2020). Japanese encephalitis virus – exploring the dark proteome and disorder–function paradigm. The FEBS Journal, 287(17), 3751-3776. https://doi.org/10.1111/febs.15427
16. Sun X, Malhis N, Zhao B, Xue B, Gsponer J, Rikkerink EHA. Computational Disorder Analysis in Ethylene Response Factors Uncovers Binding Motifs Critical to Their Diverse Functions. International Journal of Molecular Sciences. 2020; 21(1):74. https://doi.org/10.3390/ijms21010074
17. Kuechler, E. R., Budzyńska, P. M., Bernardini, J. P., Gsponer, J., & Mayor, T. (2020). Distinct Features of Stress Granule Proteins Predict Localization in Membraneless Organelles. Journal of Molecular Biology, 432(7), 2349-2368.
18. Hassan, S. S., Attrish, D., Ghosh, S., Choudhury, P. P., Uversky, V. N., Aljabali, A. A., Lundstrom, K., Uhal, B. D., Rezaei, N., Seyran, M., Pizzol, D., Adadi, P., Soares, A., Abd El-Aziz, T. M., Kandimalla, R., Tambuwala, M. M., Azad, G. K., Sherchan, S. P., Baetas-da-Cruz, W., . . . Brufsky, A. M. (2021). Notable sequence homology of the

ORF10 protein introspects the architecture of SARS-CoV-2. International Journal of Biological Macromolecules, 181, 801-809. https://doi.org/10.1016/j.ijbiomac.2021.03.199

19. Hou, C., Li, Y., Wang, M. et al. Systematic prediction of degrons and E3 ubiquitin ligase binding via deep learning. BMC Biol 20, 162 (2022). https://doi.org/10.1186/s12915-022-01364-6.
20. Mignon, J., Mottet, D., Leyder, T., Uversky, V. N., Perpète, E. A., & Michaux, C. (2022). Structural characterisation of amyloidogenic intrinsically disordered zinc finger protein isoforms DPF3b and DPF3a. International Journal of Biological Macromolecules, 218, 57-71. https://doi.org/10.1016/j.ijbiomac.2022.07.102
21. Flynn AJ, Miller K, Codjoe JM, King MR, Haswell ES. Mechanosensitive ion channels MSL8, MSL9, and MSL10 have environmentally sensitive intrinsically disordered regions with distinct biophysical characteristics in vitro. Plant Direct. 2023 Aug 3;7(8):e515. doi: 10.1002/pld3.515. PMID: 37547488; PMCID: PMC10400277.
22. Juniku B, Mignon J, Carême R, Genco A, Obeid AM, Mottet D, Monari A, Michaux C. Intrinsic disorder and salt-dependent conformational changes of the N-terminal region of TFIP11 splicing factor. Int J Biol Macromol. 2024 Oct;277(Pt 3):134291. doi: 10.1016/j.ijbiomac.2024.134291. Epub 2024 Jul 30. PMID: 39089542.
23. Dosztányi Z, Mészáros B, Simon I. ANCHOR: web server for predicting protein binding regions in disordered proteins. Bioinformatics. 2009 Oct 15;25(20):2745-6. doi: 10.1093/bioinformatics/btp518. Epub 2009 Aug 28. PMID: 19717576; PMCID: PMC2759549.
24. Mészáros B, Simon I, Dosztányi Z. Prediction of protein binding regions in disordered proteins. PLoS Comput Biol. 2009 May;5(5):e1000376. doi: 10.1371/journal.pcbi.1000376. Epub 2009 May 1. PMID: 19412530; PMCID: PMC2671142.
25. Xue B, Dunker AK, Uversky VN. Retro-MoRFs: identifying protein binding sites by normal and reverse alignment and intrinsic disorder prediction. Int J Mol Sci. 2010 Sep 29;11(10):3725-47. doi: 10.3390/ijms11103725. PMID: 21152297; PMCID: PMC2996789.
26. Miri Disfani F, Hsu WL, Mizianty MJ, et al. MoRFpred, a computational tool for sequence-based prediction and characterization of short disorder-to-order transitioning binding regions in proteins. 2012; Bioinformatics, 28(12): i75-i83.
27. Khan W, Duffy F, Pollastri G, et al. Predicting binding within disordered protein regions to structurally characterised peptide-binding domains. PLoS One. 2013 Sep 3;8(9):e72838. doi: 10.1371/journal.pone.0072838. PMID: 24019881; PMCID: PMC3760854.
28. Fang C, Noguchi T, Tominaga D, et al. MFSPSSMpred: identifying short disorder-to-order binding regions in disordered proteins based on contextual local evolutionary conservation. BMC Bioinformatics. 2013 Oct 4;14:300. doi: 10.1186/1471-2105-14-300. PMID: 24093637; PMCID: PMC3853019.
29. Malhis N, and Gsponer J. Computational Identification of MoRFs in Protein Sequences. Bioinformatics (2015) 31 (11): 1738-1744. pmid: 25637562.
30. Jones DT, Cozzetto D. DISOPRED3: precise disordered region predictions with annotated protein-binding activity. Bioinformatics. 2015 Mar 15;31(6):857-63. doi: 10.1093/bioinformatics/btu744. Epub 2014 Nov 12. PMID: 25391399; PMCID: PMC4380029.
31. Malhis N, Wong TCE, Nassar R, et al. Computational Identification of MoRFs in Protein Sequences Using Hierarchical Application of Bayes Rule. PLOS ONE (2015), DOI: 10.1371/journal.pone.0141603. pmid: 26517836.
32. Yan J, Dunker AK, Uversky VN, et al. Molecular recognition features (MoRFs) in three domains of life. Mol Biosyst. 2016 Mar;12(3):697-710. doi: 10.1039/c5mb00640f. PMID: 26651072.
33. Peng Z, Kurgan LA, 2015. High-throughput prediction of RNA, DNA and protein binding regions mediated by intrinsic disorder. Nucleic Acids Research, 43(18): e121.
34. Malhis N, Jacobson M, and Gsponer J. MoRFchibi SYSTEM: Software Tools for the Identification of MoRFs in Protein sequences. Nucleic Acids Research (2016), doi: 10.1093/nar/gkw409. Pmid: 27174932.
35. Sharma R, Kumar S, Tsunoda T, et al. Predicting MoRFs in protein sequences using HMM profiles. BMC Bioinformatics. 2016 Dec 22;17(Suppl 19):504. doi: 10.1186/s12859-016-1375-0. PMID: 28155710; PMCID: PMC5259822.

36. Sharma R, Bayarjargal M, Tsunoda T, et al. MoRFPred-plus: Computational Identification of MoRFs in Protein Sequences using Physicochemical Properties and HMM profiles. J Theor Biol. 2018 Jan 21;437:9-16. doi: 10.1016/j.jtbi.2017.10.015. Epub 2017 Oct 16. PMID: 29042212.
37. Bálint Mészáros and others, IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding, Nucleic Acids Research, Volume 46, Issue W1, Jul 2 2018, Pages W329–W337, https://doi.org/10.1093/nar/gky384.
38. Sharma R, Raicar G, Tsunoda T, et al. OPAL: prediction of MoRF regions in intrinsically disordered protein sequences. Bioinformatics. 2018 Jun 1;34(11):1850-1858. doi: 10.1093/bioinformatics/bty032. PMID: 29360926.
39. Sharma R, Sharma A, Raicar G, et al. OPAL+: Length-Specific MoRF Prediction in Intrinsically Disordered Protein Sequences. Proteomics. 2019 Mar;19(6):e1800058. doi: 10.1002/pmic.201800058. Epub 2018 Nov 2. PMID: 30324701.
40. Fang C, Moriwaki Y, Tian A, et al. Identifying short disorder-to-order binding regions in disordered proteins with a deep convolutional neural network method. J Bioinform Comput Biol. 2019 Feb;17(1):1950004. doi: 10.1142/S0219720019500045. PMID: 30866736.
41. He H, Zhao J, Sun G. Computational prediction of MoRFs based on protein sequences and minimax probability machine. BMC Bioinformatics. 2019 Oct 28;20(1):529. doi: 10.1186/s12859-019-3111-z. PMID: 31660849; PMCID: PMC6819637.
42. He H, Zhao J, Sun G. Prediction of MoRFs in Protein Sequences with MLPs Based on Sequence Properties and Evolution Information. Entropy (Basel). 2019 Jun 27;21(7):635. doi: 10.3390/e21070635. PMID: 33267349; PMCID: PMC7515128.
43. Hanson J, Litfin T, Paliwal K, Zhou Y. Identifying molecular recognition features in intrinsically disordered regions of proteins by transfer learning. Bioinformatics. 2020 Feb 15;36(4):1107-1113. doi: 10.1093/bioinformatics/btz691. PMID: 31504193.
44. He H, Zhou Y, Chi Y, He J. Prediction of MoRFs based on sequence properties and convolutional neural networks. BioData Min. 2021 Aug 14;14(1):39. doi: 10.1186/s13040-021-00275-6. PMID: 34391457; PMCID: PMC8364704.
45. Hu G, Katuwawala A, Wang K, Wu Z, Ghadermarzi S, Gao J, Kurgan L. flDPnn: Accurate intrinsic disorder prediction with putative propensities of disorder functions. Nat Commun. 2021 Jul 21;12(1):4438. doi: 10.1038/s41467-021-24773-7. PMID: 34290238; PMCID: PMC8295265.
46. Zhang F, Zhao B, Shi W, Li M, Kurgan L. DeepDISOBind: accurate prediction of RNA-, DNA- and protein-binding intrinsically disordered residues with deep multi-task learning. Brief Bioinform. 2022 Jan 17;23(1):bbab521. doi: 10.1093/bib/bbab521. PMID: 34905768.
47. Katuwawala A, Zhao B, Kurgan L. DisoLipPred: accurate prediction of disordered lipid-binding residues in protein sequences with deep recurrent networks and transfer learning. Bioinformatics. 2021 Dec 22;38(1):115-124. doi: 10.1093/bioinformatics/btab640. PMID: 34487138.
48. Piovesan D, Monzon AM, Tosatto SCE. Intrinsic protein disorder and conditional folding in AlphaFoldDB. Protein Sci. 2022 Nov;31(11):e4466. doi: 10.1002/pro.4466. PMID: 36210722; PMCID: PMC9601767.
49. Peng Z, Li Z, Meng Q, Zhao B, Kurgan L. CLIP: accurate prediction of disordered linear interacting peptides from protein sequences using co-evolutionary information. Brief Bioinform. 2023 Jan 19;24(1):bbac502. doi: 10.1093/bib/bbac502. PMID: 36458437.
50. Pang Y, Liu B. DisoFLAG: accurate prediction of protein intrinsic disorder and its functions using graph-based interaction protein language model. BMC Biol. 2024 Jan 2;22(1):3. doi: 10.1186/s12915-023-01803-y. PMID: 38166858; PMCID: PMC10762911.
51. Necci M, Piovesan D; CAID Predictors; DisProt Curators; Tosatto SCE. Critical assessment of protein intrinsic disorder prediction. Nat Methods. 2021 May;18(5):472-481. doi: 10.1038/s41592-021-01117-3. Epub 2021 Apr 19. PMID: 33875885; PMCID: PMC8105172.
52. Chen R, Li X, Yang Y, Song X, Wang C, Qiao D. Prediction of protein-protein interaction sites in intrinsically disordered proteins. Front Mol Biosci. 2022 Sep 30;9:985022. doi: 10.3389/fmolb.2022.985022. PMID: 36250006; PMCID: PMC9567019.

53. Han B, Ren C, Wang W, Li J, Gong X. Computational Prediction of Protein Intrinsically Disordered Region Related Interactions and Functions. Genes (Basel). 2023 Feb 8;14(2):432. doi: 10.3390/genes14020432. PMID: 36833360; PMCID: PMC9956190.
54. Basu S, Kihara D, Kurgan L. Computational prediction of disordered binding regions. Comput Struct Biotechnol J. 2023 Feb 10;21:1487-1497. doi: 10.1016/j.csbj.2023.02.018. PMID: 36851914; PMCID: PMC9957716.
55. Kurgan L, Hu G, Wang K, Ghadermarzi S, Zhao B, Malhis N, Erdős G, Gsponer J, Uversky VN, Dosztányi Z. Tutorial: a guide for the selection of fast and accurate computational tools for the prediction of intrinsic disorder in proteins. Nat Protoc. 2023 Nov;18(11):3157-3172. doi: 10.1038/s41596-023-00876-x. Epub 2023 Sep 22. PMID: 37740110.
56. 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 Dec;91(12):1925-1934. doi: 10.1002/prot.26582. Epub 2023 Aug 25. PMID: 37621223.
57. Malhis, N, Gsponer, J. Computational Prediction of Linear Interacting Peptides. Springer Natures. Methods in Molecular Biology book, "Prediction of Protein Secondary Structure" second edition, chapter 14. (2024)
58. Aspromonte MC, Nugnes MV, Quaglia F, Bouharoua A; DisProt Consortium; Tosatto SCE, Piovesan D. DisProt in 2024: improving function annotation of intrinsically disordered proteins. Nucleic Acids Res. 2024 Jan 5;52(D1):D434-D441. doi: 10.1093/nar/gkad928. PMID: 37904585; PMCID: PMC10767923.
59. Piovesan D, Del Conte A, Mehdiabadi M, Aspromonte MC, Blum M, Tesei G, von Bülow S, Lindorff-Larsen K, Tosatto SCE. MobiDB in 2025: Integrating ensemble properties and function annotations for intrinsically disordered proteins (2024) Nucleic Acid Research.
60. Del Conte A, Bouhraoua A, Mehdiabadi M, Clementel D, Monzon AM; CAID predictors; Tosatto SCE, Piovesan D. CAID prediction portal: a comprehensive service for predicting intrinsic disorder and binding regions in proteins. Nucleic Acids Res. 2023 Jul 5;51(W1):W62-W69. doi: 10.1093/nar/gkad430. PMID: 37246642; PMCID: PMC10320102.
61. S. Fukuchi, T. Amemiya, S. Sakamoto, Y. Nobe, K. Hosoda, Y. Kado, S. D. Murakami, R. Koike, H. Hiroaki and M. Ota., "IDEAL in 2014 illustrates interaction networks composed of intrinsically disordered proteins and their binding partners", 2014, Nucleic Acids Res., 42, D320-D325., PMID: 24178034, DOI: 10.1093/nar/gkt1010.
62. Schad E, Fichó E, Pancsa R, Simon I, Dosztányi Z, Mészáros B. DIBS: a repository of disordered binding sites mediating interactions with ordered proteins. Bioinformatics. 2018 Feb 1;34(3):535-537. doi: 10.1093/bioinformatics/btx640. PMID: 29385418; PMCID: PMC5860366.
63. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012 Dec 1;28(23):3150-2. doi: 10.1093/bioinformatics/bts565. Epub 2012 Oct 11. PMID: 23060610; PMCID: PMC3516142.
64. Malhis, N. 2024. Pre-processing annotated homologous regions in protein sequences concerning machine-learning applications. bioRxiv doi: 10.1101/2024.10.25.620288.
65. Kumar M, Gouw M, Michael S, Sámano-Sánchez H, Pancsa R, Glavina J, Diakogianni A, Valverde JA, Bukirova D, Čalyševa J, Palopoli N, Davey NE, Chemes LB, Gibson TJ. (2020) Nucleic Acids Res., Jan, 2020.
66. Monzon AM, Bonato P, Necci M, Tosatto SCE, Piovesan D. FLIPPER: Predicting and Characterizing Linear Interacting Peptides in the Protein Data Bank. J Mol Biol. 2021 Apr 30;433(9):166900. doi: 10.1016/j.jmb.2021.166900. Epub 2021 Feb 27. PMID: 33647288.
67. Dosztányi, Z., et al. 2005 The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. J. Mol. Biol.347827–839.

Supplement

Probabilistic Annotations of Protein Sequences for Intrinsically Disordered Features

Nawar Malhis

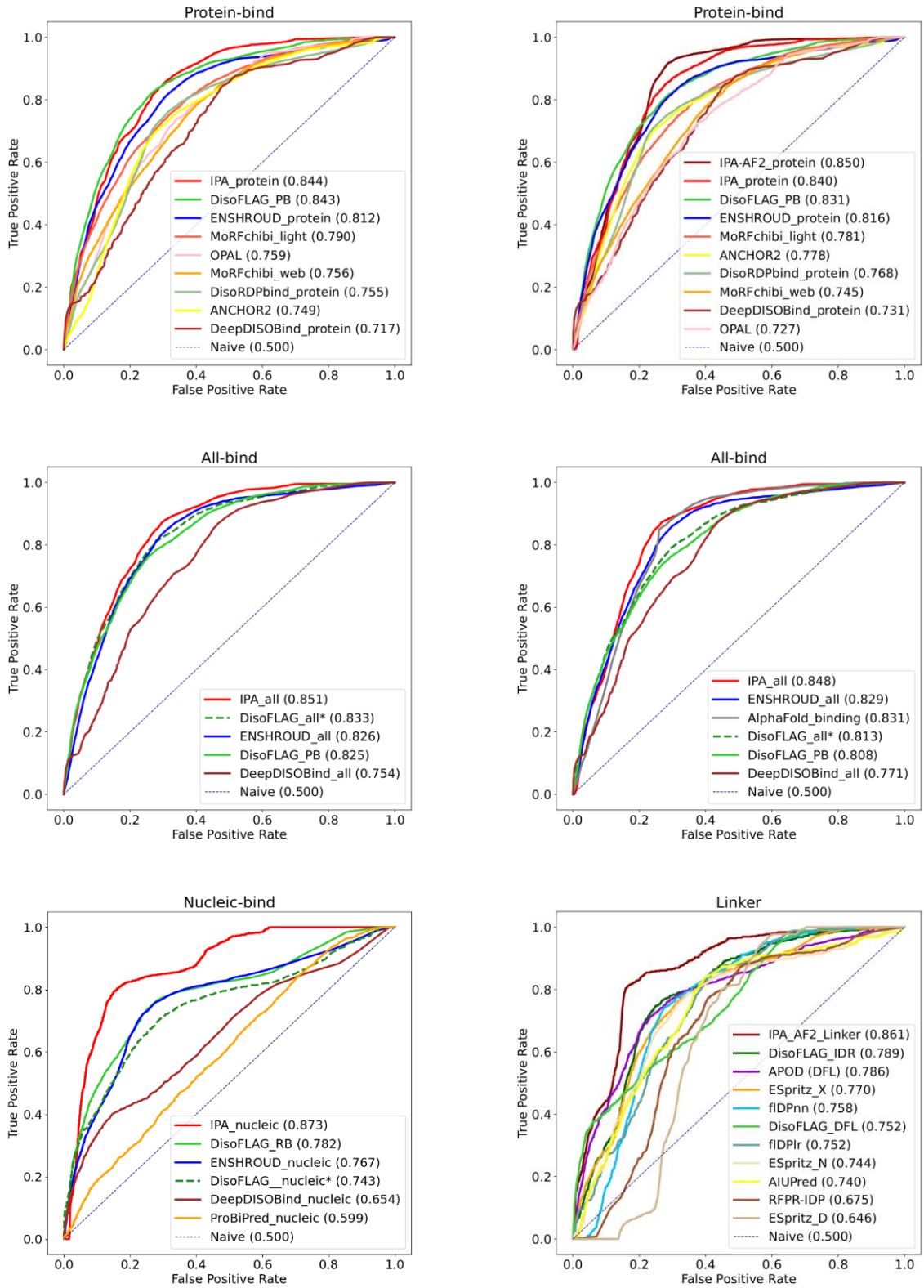

Figure S1: Compared to other tools, the ROC curves for the IPA protein, nucleic, and all binding predictions and IPA-AF2 protein-binding and Linker regions predictions.

Table S1: Steps used to assemble the Test-24_06 dataset. Test-24_06-AF2 is a subset of the TEST-24_06 sequences with structures generated by AlphaFold-2 in the AlphaFold database.

		Total sequences	Annotated sequences	Target sites	Non-target residues	Target residues	Masked-out residues
DSET1 = DP24_06 - DP23_12	**P_bind**	185	25	28	96,714	2,910	0
	N_bind	185	5	5	99,212	412	0
	LIP	185	28	31	96,470	3,154	0
	Linker	185	14	19	98,981	643	0
DSET2 = DSET1 after excluding homologs with > 30% identity using CD-Hit	**P_bind**	164	19	22	80,966	2,224	8,505
	N_bind	164	5	5	88,324	412	2,959
	LIP	164	22	25	80,722	2,468	8,505
	Linker	164	12	16	86,543	559	4,593
DSET3 = Including MobiDB annotations into DSET2	**P_bind**	164	26	33	68,778	2,515	20,402
	N_bind	164	5	5	88,324	412	2,959
	LIP	164	40	51	67,264	4,147	20,284
	Linker	164	12	16	86,543	559	4,593
DSET4 = DSET3 after processing with HAM	**P_bind**	164	26	33	68,719	2,515	20,461
	N_bind	164	5	5	88,279	412	3,004
	LIP	164	40	51	67,060	4,147	20,488
	Linker	164	12	16	86,521	559	4,615
Test-24_06 = DSET4 + DP23_12 sequences with new annotations	**P_bind**	184	42	55	68,719	3,873	28,742
	N_bind	184	11	13	88,279	1,183	11,872
	LIP	184	48	65	68,445	4,696	28,193
	Linker	184	14	18	86,521	619	14,194
Test-24_06-AF2 (limited to AlphaFold2 structures)	**P_bind**	147	35	47	58,985	3,436	13,123
	LIP	147	40	55	56,567	4,220	11,757
	Linker	144	14	18	67,519	619	584

Table S2: IPA-AF2_Linker predictor outperformed other Linker predictors (in bold) and IDR predictors in AUC values against the TEST-24_06-AF2 dataset.

	AUC
IPA_AF2_Linker	**0.861**
DisoFLAG_IDR	0.789
APOD	**0.786**
ESpritz_X	0.770
flDPnn	0.758
DisoFLAG_DFL	**0.752**
flDPlr	0.752
ESpritz_N	0.744
AIUPred	0.740
RFPR-IDP	0.675
ESpritz_D	0.646

Table S3: The **F5** features for each of the 20 standard amino acids. For each amino acid, these features are the frequency of that amino acid in each of the five classes: IDR, nucleic binding (N_bind), protein binding (P_bind), and 'Linker' in the DisProt 2023_12 data and 'PDB' annotation in the combined CAID1 and CAID2 sequences, divided by the frequency of that amino acid in the entire dataset and then multiplied by 100.

Amino Acid	IDR	Linker	N_bind	P_bind	PDB
A	108.776	105.711	96.964	103.172	93.301
C	51.852	39.502	51.072	59.233	117.768
D	112.017	109.835	97.596	113.697	101.451
E	124.133	114.155	104.308	118.074	96.901
F	69.997	59.429	73.528	74.956	115.953
G	116.862	121.453	111.039	111.067	94.527
H	83.513	79.084	91.107	80.180	98.521
I	66.664	67.150	81.632	73.633	121.547
K	119.192	102.862	158.014	100.009	101.104
L	71.549	65.828	66.448	83.231	109.540
M	82.179	92.349	77.203	84.958	100.098
N	97.746	117.902	114.751	100.493	103.319
P	133.391	147.432	110.458	126.119	78.845
Q	112.595	123.824	113.074	114.435	97.290
R	89.594	91.773	127.314	100.377	102.433
S	117.347	138.093	116.409	122.400	82.040
T	106.789	108.159	96.611	104.118	92.888
V	81.305	72.932	73.531	81.647	106.683
W	62.060	42.601	53.834	65.289	121.847
Y	69.566	50.065	85.432	72.148	119.810

Table S4: The complete list of values for convolutional kernel sizes (CKZ), average pooling kernel sizes (APKZ), depth increase (D_inc), and the output size of the first fully connected layer (FC1Z) values for all 20 CNN models

Predictor	Model CV num	CKZ	APKZ	D_inc	FC1Z
IPA-Protein	1	10	15	4	100
	2	15	15	6	100
	3	7	15	4	150
	4	15	15	4	100
IPA-Nucleic	1	10	15	8	150
	2	7	10	4	100
	3	7	15	6	100
	4	15	7	4	100
IPA-All	1	10	10	10	130
	2	10	10	10	130
	3	10	10	8	130
	4	10	10	8	110
IPA-AF2-Protein	1	7	10	6	100
	2	10	15	4	150
	3	7	15	4	100
	4	15	10	4	150
IPA-AF2-Linker	1	10	10	2	100
	2	10	15	4	150
	3	7	7	4	150
	4	15	10	4	100