DeepPRObind: Modular deep learner that accurately predicts structure and disorder-annotated protein binding residues

Fuhao Zhang[1], Min Li[1*], Jian Zhang[2], Wenbo Shi[1], and Lukasz Kurgan[3*]

[1]Hunan Provincial Key Lab on Bioinformatics, School of Computer Science and Engineering, Central South University, Changsha, China, 410083.
[2]School of Computer and Information Technology, Xinyang Normal University, Xinyang, China, 464000
[3]Department of Computer Science, Virginia Commonwealth University, Richmond, VA, USA, 23284
Corresponding authors: *Min Li at limin@mail.csu.edu.cn and *Lukasz Kurgan at lkurgan@vcu.edu

Abstract

Current sequence-based predictors of protein-binding residues (PBRs) belong to two distinct categories: structure-trained vs. intrinsic disorder-trained. Since disordered PBRs differ from structured PBRs in several ways, including ability to bind multiple partners by folding into different conformations and enrichment in different amino acids, the structure-trained and disorder-trained predictors were shown to provide inaccurate results for the other annotation type. A simple consensus-based solution that combines structure- and disorder-trained methods provides limited levels of predictive performance and generates relatively many cross-predictions, where residues that interact with other ligand types are predicted as PBRs. We address this unsolved problem by designing a novel and fast deep-learner, DeepPRObind, that relies on carefully designed modular convolutional architecture and uses innovative aggregate input features. Comparative empirical tests on a low-similarity test dataset reveal that DeepPRObind generates accurate predictions of structured and disordered PBRs and low amounts of cross-predictions, outperforming a comprehensive collection of 12 predictors of PBRs. Given the relatively low runtime of DeepPRObind (40 seconds per protein), we further validate its results based on an analysis of putative PBRs in the yeast proteome, confirming that interactions in disordered regions are enriched among hub proteins. We release DeepPRObind as a convenient web server at https://www.csuligroup.com/DeepPRObind/.

Keywords: protein-protein interactions; protein-binding residues; prediction; deep learner; intrinsic disorder

1 Introduction

Protein-protein interactions (PPIs) are crucial for many cellular functions including regulation of cell cycle, signaling, and metabolism [1]. Knowledge of PPIs facilitates understanding of cellular processes, development of protein docking programs, construction of PPI networks, drug design efforts and exploration of molecular mechanisms that underpin certain diseases [2-6]. Information on PPIs is available in multiple databases including Protein Data Bank (PDB) [7] that provides atomic level details, DisProt [8] and BioLiP [9] that focus on the amino acid-level annotations, and STRING [10], mentha [11] and BioGRID [12] that annotate these interactions at the protein level. While billions of interactions were annotated at the protein level [10], more detailed annotations of interacting residues are available for a small fraction of them.

This large knowledge gap motivates development of computational predictors of protein-binding residues (PBRs) in protein sequences. Recent works categorize these tools into two classes: those developed using training datasets of structured protein-protein complexes (structure-trained predictors) vs. methods developed using annotations of intrinsically disordered PBRs (disorder-trained predictors) [13, 14]. The training and testing of the structure-trained methods rely on the data collected from PDB and BioLip while the disorder-trained predictors utilize data from DisProt. Recent surveys identify close to 20 structure-trained predictors [15-19]. In chronological order, they include SPPIDER [20], PSIVER [21], LORIS [22], SPRINGS [23], predictors by [24, 25], CRF-PPI [26], PPIS [27], SPRINT [28], iPPBS-Opt [29], SSWRF [30], methods by [31, 32], EL-SMURF [33], SCRIBER [34], DeepPPISP [35], and PROBselect [13], and DELPHI [36]. A similarly large number of the disorder-trained predictors was released [14, 37]. They include alpha-MoRFpred [38], ANCHOR [39, 40], retro-MoRFs [41], MoRFpred [42], MFSPSSMpred [43], DISOPRED3 [44], DisoRDPbind [45-47], MoRFChiBi [48],

MoRFCHiBiWeb [48], fMoRFPred [49], MoRDPred-plus [50], OPAL [51], ANCHOR2 [40], and OPAL+ [52]. Majority of these methods predict molecular recognition features (MoRFs), which are short sequence segments that undergo folding upon binding and that are typically embedded inside longer disordered regions [49]. ANCHOR, ANCHOR2 and DisoRDPbind predict a broader class of intrinsically disordered PBRs that are include longer regions. Intrinsically disordered PBRs are different from structured PBRs in several ways including the fact that the former may bind to several different proteins or peptides by folding into different conformations [53-55], have larger surface area and are enriched in the disorder-promoting amino acids [56], and are overrepresented among hub proteins [57, 58]. Given these differences, recent study demonstrates that the structure- and disorder-trained predictors provide inaccurate results for the other type of interaction, i.e., structured-trained methods perform poorly for the disordered PBRs and vice versa [13]. That study proposed a meta predictor-based solution by combining results of a well-performing disorder-trained predictor (DisoRDPbind) and a well-performing structure-trained method (SCRIBER) based on averaging their outputs [13]. While the resulting HybridPBRpred method offers relatively accurate results for structured and disordered PBRs, its predictive model is simple, leaving room to develop a more sophisticated design that would likely provide more accurate predictions. Moreover, HybridPBRpred and the disorder- and structure-trained methods were recently shown to cross-predict binding residues, i.e., they predict many residues that interact with other partner types (e.g., nucleic acids and small ligands) as PBRs and vice versa [13, 19].

We address these issues by developing a new tool, DeepPRObind, that aims to accurately predict structured and intrinsically disordered PBRs and to significantly reduce cross-predictions when compared to the current methods. DeepPRObind relies on a deep convolutional neural network that processes an information-rich profile derived from an input sequence using two empirically parametrized modules for the prediction of the disordered and the structured PBRs. The key innovations behind our model are the use of aggregate features that quantify relevant information at the sequence window and full sequence levels, an empirically crafted approach to combine results from the two modules, and relatively low runtime. We use a recently introduced test dataset [13], which shares <25% similarity to the training and validation proteins, to compare predictive performance and the cross-predictions generated by DeepPRObind, HybridPBRpred and a representative selection of the disorder-trained and structure-trained methods. Given the relatively low computational cost of DeepPRObind, we further validate its results using protein-level annotations of PPIs in the yeast proteome.

2 Materials and methods

2.1 Selection of representative predictors

We analyze relevant literature [14-19, 37] and perform PubMed search to identify disorder-trained and structure-trained predictors of PBRs for inclusion in the comparative study. We select predictors that are available as web servers and/or standalone code, produce results reasonably quickly to efficiently process the test dataset (<5 min per protein), and generate putative real-valued propensities of PBRs. The latter is necessary to generate metrics that can be adequately compared across methods, which we discuss in Section 2.3. Consequently, we identify a comprehensive collection of 12 predictors: SPPIDER [20], PSIVER [21], LORIS [22], SPRINGS [23], CRF-PPI [26], SSWRF [30], SPRINT [28], SCRIBER [34], DisoRDPbind [45], fMoRFpred [49], ANCHOR2 [40], and HybridPBRpred [13]. They include 8 structure-trained methods, 3 disorder-trained tools (ANCHOR2, DisoRDPbind and fMoRFpred), and HybridPBRpred – the only method that combines these two types of predictions. This collection of methods improves over the scope of a recent comparative study that focuses on empirical evaluation of 7 structure-trained methods [19]. Moreover, two of the disorder-trained methods that we include, ANCHOR2 and DisoRDPbind, secured the top two spots in the recent CAID community assessment of the prediction of disordered binding regions [59]. The 12 selected methods cover older and popular/highly cited tools, such as SPPIDER and PSIVER, and recently published methods, such as HybridPBRpred and ANCHOR2. These methods utilize a broad spectrum of predictive models, ranging from relatively simple Naïve Bayes (PSIVER) and regressors (LORIS, SCRIBER, DisoRDPbind, and ANCHOR2), through more complex models, such as random forest (CRF-PPI), to sophisticated solutions that involve support vector machines (SSWRF,

SPRINT, fMoRFpred), neural networks (SPPIDER and SPRINGS), and ensembles (SSWRF and HybridPBRpred).

2.2 Datasets

We use a high-quality test dataset from a recent study of the disorder- and structure-trained predictors [13] to evaluate and compare predictive performance of DeepPRObind and the other 12 predictors. These test proteins and their annotations of binding were collected from PDB [7], BioLiP [9] and DisProt [8] databases. These data were clustered with the training datasets of the 12 predictors and only proteins that share <25% similarity to these training proteins are included in the test dataset [13]. The resulting test dataset is a balanced collection of 92 structured and 92 disordered proteins that features 9,442 PBRs and 3,477 residues that interact with other ligands. We use the latter residues to assess the cross predictions. We provide a more detailed breakdown of the number of binding residues across the structure-annotated and disorder-annotated proteins in Suppl. Table S1. We note that the size of this test dataset, which is 184 proteins, is in line with the test datasets used in related studies, including LORIS [22] and CRF-PPI [26] that used two test datasets with 72 and 164 proteins, and SSWRF [30], DisoRDPbind [45], SPPIDER [20], PSIVER [21], SPRINT [28], and HybridPBRpred [13] that were assessed on 164, 115, 149, 72, 80, and 184 test proteins, respectively.

We collect data that we use to empirically design DeepPRObind by combining training datasets of the 12 predictors and disordered proteins from DisProt [8]. We uniformly sample this large pool of proteins and remove those that share similarity with the test proteins. We combine these proteins with the 184 test proteins, cluster the resulting protein set with BLASTCLUST [60] at 25% similarity, remove the clusters that include test proteins, and retain one protein for each remaining cluster. The resulting protein set includes 1,190 structure-annotated and 680 disorder-annotated proteins. We select at random a balanced set of 210 structure-annotated and 210 disorder-annotated proteins that compose the validation dataset. The remaining 980 structure-annotated and 470 disorder-annotated proteins constitute the training dataset. We use the training and validation datasets to empirically design and optimize the DeepPRObind model, and subsequently compare the results produced by this optimized model with the other methods using the test dataset. Given the above clustering, proteins in the training and validation datasets share low (<25%) similarity with the test proteins.

We give detailed statistics concerning the number of PBRs, residues that interact with other ligands and the remaining (non-binding) residues for the training, validation and test datasets in Suppl. Table S1. The datasets, including annotations of interacting residues, are available at https://www.csuligroup.com/DeepPRObind/.

2.3 Evaluation criteria

The outputs of PBR predictors include putative real-valued propensities that quantify likelihood that a given residue binds proteins and binary values that classify residues as either PBRs or non-protein interacting. The binary predictions are typically derived from the propensities with a threshold, such that residues with propensities > threshold are categorized as PBRs, and otherwise they are assumed to be non-protein interacting. We assess binary predictions with a comprehensive set of metrics including F1, MCC, precision, specificity and recall. We standardize the binary predictions between considered predictors to ensure that they can be directly compared, i.e., they produce the same rates of predictions of PBRs. More specifically, we set the threshold such that specificity = 0.9. We evaluate the predicted propensities using two popular metrics [13, 61]: area under the ROC curve (AUC) and the area under the precision-recall curve (AUPRC). Moreover, we quantify the cross-predictions with area under cross-prediction curve (AUCPC), cross-prediction rate, area under the over-prediction curve (AUOPC) and over-prediction rate [13, 61]. Higher values of AUC and AUPRC indicate better predictive performance while lower values of AUCPC and AUOPC correspond to fewer cross-predictions and over-predictions. We define these metrics in the Supplement.

2.4 DeepPRObind predictor

DeepPRObind is a modular deep learner that generates propensity for protein binding for each residue in the input protein sequence. We summarize this predictive framework in Figure 1A. First, we produce two types of inputs from the protein sequences: the commonly used ***sequence profile*** and the new ***aggregate features*** (green block in

Figure 1A). The profile relies on fast and accurate tools that use the input sequence to derive key amino acid-level characteristic that are relevant to protein-protein interactions including conservation, propensity for binding, and putative solvent accessibility, secondary structure and intrinsic disorder [15]. The new aggregate features summarize characteristics relevant to protein binding at a local sequence window/segment and at the whole sequence levels, complementing the residue-level information from the sequence profile. Second, motivated by the distinct characteristics of the underlying binding sites for the structured vs. disordered protein binding sites [56], we use two deep convolutional network modules that we train from the structure-annotated and disorder-annotated data (yellow and orange blocks in Figure 1A). This allows us to optimize each network independently by evaluating whether they benefit from the inclusion of the new aggregate features and by optimizing the network size. Finally, we empirically design "combination layer" (blue block in Figure 1A) that outputs universal (i.e., annotation type agnostic) propensities for protein binding based on predictions of the two modules and the putative disorder content that we obtain with the popular IUPred2A method [40].

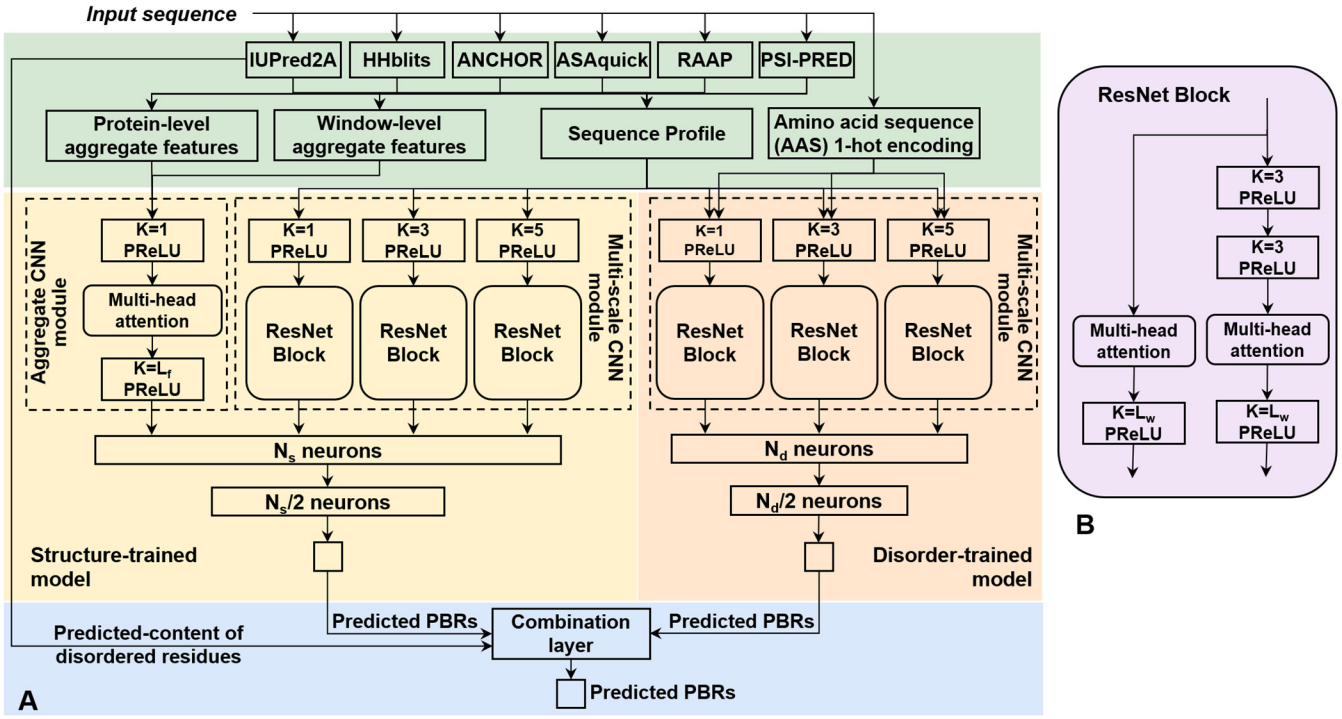

Figure 1. Architecture of DeepPRObind. Panel A is the overall topology. Panel B details the ResNet blocks from the deep convolutional neural network (CNN) modules. K is the kernel size in the CNN modules. L_f and L_w are the number of aggregate features and the window size, respectively. N_s and N_d are the numbers of neurons in the fully connected feed-forward layers of structure-trained and the disorder-trained networks, respectively, which we optimize empirically.

2.4.1 Sequence profile, amino acids-based features and aggregate features

The sequence profile is a collection of residue-level characteristics that are relevant to protein binding [15], and which are presented to the deep network using a sliding window. This type of input is commonly used by related deep network methods [62, 63]. We ensure that these characteristics are produced by fast and accurate methods, so that the overall runtime does not exceed 1 minute. The profile includes relative solvent accessibility (RSA) predicted with ASAquick [64]; secondary structure predicted by the single sequence-based version 3.2 of PSI-PRED [65]; putative intrinsic disorder and disordered binding regions produced by IUPred2A and ANCHOR [40], respectively; evolutionary conservation (ECO) calculated with HHblits [66] based on the uniprot20_2015_06 database; and relative amino acid propensity (RAAP) for binding to proteins, DNA, and RNA protein computed using the approach described in ref. [15]. We provide a detailed description of the sequence profile in Suppl. Table S2. Some of the related deep network-based solutions also directly use the amino acid sequence (AAS) as the input [67]. We encode the amino acids via the 1-hot scheme, which is a 21-dimensional

binary vector where 20 positions denote the 20 amino acid types and the last position encodes for undetermined types, such as X. We study empirically whether inclusion of the AAS input would benefit our model.

Recent works in related areas show that application of the protein-level and window-level features benefits prediction of disordered linkers when using support vector machines [68, 69]. The protein-level features average the relevant residue-level characteristics that are included in the sequence profile over the whole protein chain. These features include the average putative RSA, average conservation, average putative disorder and disordered binding propensities, average propensities for putative helix, coil and strand conformations, average RAAP values and amino acid composition of the sequence. The window-level features contrast average characteristics of the neighboring residues in the sequence (middle half of the sliding window) against the flanking regions (the two adjacent regions that cover quarter of the window at each of its termini), which is motivated by the use of such approach in a couple of related studies [42, 69]. These features intend to identify regions in the input sequence that are different from the flanking regions, e.g., regions that have high putative solvent accessibility and conservation surrounded by putative buried residues that have relatively lower conservation, which may suggest a higher likelihood for binding. The window-level features consider the differences in conservation, putative RSA, putative disorder and disordered binding propensities, propensities for putative helix, coil and strand conformations, RAAP values and amino acid composition. We enumerate and describe the aggregate protein-level and window-level features in Suppl. Table S2. We empirically investigate whether these aggregate features would benefit predictions made by the deep network modules.

As Figure 1A shows, we feed the input protein sequences into the IUpred2A, HHblits, ANCHOR, ASAquick and PSI-PRED methods. Next, we combine their outputs to generate sequence profile, separately compute AAS from the input sequence, and calculate the aggregate features, which we subsequently input into the neural network modules.

2.4.2 Design of the deep convolutional networks

Motivated by the diversity of the key characteristics of the structured vs. disordered binding sites (e.g., disordered sites have larger surface area and are enriched in the disorder-promoting amino acids) [56], we design and individually optimize disorder-trained and the structure-trained deep networks. We also formulate and empirically optimize a combination layer that merges the two predictions to obtain an annotation type-agnostic prediction of PBRs.

We design the deep networks by relying on two convolutional neural network (CNN) modules: the multi-scale CNN module that uses the sequence profile and AAS as its input, and the aggregate CNN module that processes the two types of the aggregate features (Figure 1A). The multi-scale CNN module uses several different kernel sizes (1, 3 and 5) followed by PReLU activation units to capture relations between neighboring residues in the input sequence using different sizes of these neighborhoods. Similar approach was shown to be successful in related predictions from protein and nucleic acid sequences [67, 70]. Next, we pass the outputs of PReLU units into ResNet blocks (Figure 1B), which produce information-rich latent spaces by utilizing shortcut connections and multi-head attention units. The multi-head attention units, which we use in both the multi-scale and the aggregate CNN modules, aim to identify latent features that are useful for prediction of PBRs. The attention is calculated by using the following formulas:

$$Q_i = XW_i^Q, K_i = XW_i^K, V_i = XW_i^V \quad (1)$$

$$attention\ weight(Q_i, K_i) = softmax\left(\frac{Q_i K_i^T}{\sqrt{d_k}}\right) \quad (2)$$

$$head_i = attention\ weight(Q_i, K_i)V_i \quad (3)$$

$$multi\ head(Q, K, V) = concactinate(head_1, \dots, head_i, \dots, head_n)W^O \quad (4)$$

where X is the input matrix; $i = 1, 2, \dots, n$ is the head index for a design with n heads; d_{in} is the number of columns of X; $d_k = d_{in}/n$; and $\mathrm{W_i^Q} \in \mathrm{R^{d_{in} \times d_k}}$, $\mathrm{W_i^K} \in \mathrm{R^{d_{in} \times d_k}}$, $\mathrm{W_i^V} \in \mathrm{R^{d_{in} \times d_k}}$, and $W^O \in R^{d_{in} \times d_{in}}$ are matrices that are trained from the training dataset. First, for each head, the three embedding matrices (Q_i, K_i, V_i) are calculated by the dot product of X with W_i^Q, W_i^K and W_i^V, respectively; see formula (1). Next, Q_i and K_i are used

to calculate attention weight by formula (2), where d_k represents the number of columns of the Q_i and K_i. Scaling by $\sqrt{d_k}$ aims to ensure stability of the softmax function gradient. The attention weights are used to identify input features that are more valuable for the prediction of PBRs. Next, the output of $head_i$ is computed by the dot product between the attention weights and V_i; see formula (3). Finally, the results from the n heads are concatenated and multiplied by the trained matrix W^O to obtain outputs; see formula (4). Next, we reduce the size of the resulting latent feature space using a convolutional layer with the kernel size equal to the feature space size (L_w), which is followed by the PReLU activation units. This approach performs pooling using the entire latent space, improving over a typically used 1D max-pooling that instead relies on the max values.

In contrast to the multi-scale CNN module, the aggregate CNN module does not utilize kernels sizes > 2 since the aggregate features are not sequential (i.e., they represent averaging of different characteristics over the whole sequence and sliding windows). Thus, we use a convolutional layer with kernels of size =1 and multi-head attention layer to extract latent features. First, we obtain L_f by 32 embedding matrix by processing the L_f = 68 aggregate features with a CNN layer that includes 32 kernels of size 1. This step could be seen as equivalent to using a fully connected layer with 32 neurons. Next, we feed the embedding matrix into the attention layer with 8 heads (default number of heads) that aims to select elements that are more useful for the prediction of PBRs. Each head processes outputs from the 32 kernels and generates inputs to the subsequently used CNN layer with one kernel of size L_f = 68. This layer generates a 32-dimensional output vector, where each value is based on one of the columns from the attention matrix. This design reduces the number of connections compared to a fully connected feed-forward design and the use of the attention layer facilitates identification of parts of the corresponding feature space that are relevant to prediction of PBRs.

We pass the pooled outputs generated by the multi-scale and aggregate CNNs into two fully connected feed-forward layers, one for the structure-trained and one for the disorder-trained model. We empirically optimize sizes of these two layers that consist of N_s and $N_s/2$ neurons for the structure-trained network and N_d and $N_d/2$ neurons for the disorder-trained network. We consider three network sizes: small with N=32, medium with N=64 and large with N=128. In addition, we consider eight different topologies that we summarize in Suppl. Table S3. They cover all combinations of corresponding inputs and matching network designs, where multi-scale CNN modules process the sequence profile and/or AAS features and aggregate CNN modules process the protein-level aggregate features, the window-level aggregate features or both types of the aggregate features together.

We train the CNNs modules using Pytorch with the Adam optimizer and binary cross-entropy loss function. We optimize the learning rate and batch size parameters separately for the structure-trained module using the structure-annotated training proteins and for the disorder-trained module using the disorder-annotated training proteins. We use a grid search based on the training dataset. Consequently, we set the learning rates of the structure-trained and disorder-trained CNNs modules to 0.0003 and 0.0002, respectively. The batch sizes of these two modules are both set to 1024.

We compare results of the corresponding 24 configurations (3 network sizes and 8 topologies) for the structure- and disorder-trained networks on the validation dataset in Suppl. Table S4. We note a modest impact of the network size on the predictive performance. When using the typically applied sequence profile, the AUCs range between 0.684 and 0.688 for the structure-trained network, and between 0.817 and 0.823 for the disorder-trained network. This suggests that large networks are not necessary to solve this problem. The impact of the new aggregate features is more substantial. Using the middle size network for the structure-trained model (i.e., the best size for this model), inclusion of the aggregate features and aggregate CNN module improves AUC from 0.684 (best design without the aggregate features) to 0.698 (best design with the aggregate features). The structure-trained network that combines the protein-level and window-level aggregate features secures AUC = 0.698, which is slightly better than when using the protein-level aggregate features alone (AUC = 0.694) and window-level aggregate features alone (AUC = 0.692). This suggests that both window and protein-level features are helpful for the predictions for the structured proteins, although the improvement from combining them is relatively small. We believe that the aggregate characteristics help to separate regions and sequence that bind proteins from those that do not. Similarly, addition of the AAS features as the input for the small size disorder-trained network (i.e., the

best network size for this model) produces a considerable increase in AUC from 0.823 to 0.840. To compare, adding aggregate features for the disorder-trained network generates a smaller increase, from 0.823 to 0.837. This can be explained by the fact that disordered regions are known to possess a strong amino acid bias, being depleted in Cysteines and aromatics and enriched in polar and charged residues [71-73], which is not necessarily reflected at the whole sequence level.

To summarize, the selected structure-trained and disorder-trained models are different in multiple aspects including inputs (aggregate features and sequence profile for the structure-based model vs. AAS features and sequence profile for the disorder-based model), topology (aggregate and multi-scale CNNs for the structure-based model vs. only multi-scale CNN for the disorder-based model) and network size (smaller fully connected layers with $N_d = 32$ for the disorder-based model vs. larger fully connected layers with $N_s = 64$ for the structure-based model). Figure 1A summarizes these differences.

We re-tested the corresponding best network sizes, i.e., medium size for the structure-trained CNN and small size disorder-trained CNN, on the test set (Suppl. Table S5). The corresponding results confirm that the structure-trained network benefits from the aggregate features and aggregate CNN module (AUC of 0.706 vs. 0.688) and the best results for disorder-trained network is when using AAS (AUC of 0.849 vs. 0.836). Moreover, using both types of aggregate features for the structure-trained network (AUC of 0.706) is better than using the window-based features separately (AUC = 0.696) and protein-based features separately (AUC = 0.696), confirming that both types of aggregate features are useful. This demonstrates that our observations are robust across multiple datasets that are characterized by the low sequence similarity. The bottom line is that inclusion of the new to this area inputs and corresponding extension of the network topology produces noticeable and consistent improvements in predictive quality.

2.4.3 Design of the combination layer

After developing the structure- and disorder-trained CNNs modules, we rationally design and empirically compare several implementations of the combination layer (blue block in Figure 1A). This layer uses 3 inputs: results from the structure-trained CNN, outputs of the disorder-trained CNN, and putative protein-level disorder content. We use the latter to suggests which of the two inputs carries more weight, i.e., a high putative disorder content would indicate that the disorder-trained CNN is likely to provide a more suitable input. We use the popular and fast IUPred2A [40] to predict the content, which is defined as fraction of disordered residues in a protein sequence. IUPred2A was shown to produce accurate content predictions [74, 75] and is available to our model since we use it to generate a part of the sequence profile.

The first design relies on a simple feed-forward network that takes a sliding window of size w of the predictions from the structure-trained and the disorder-trained CNNs and the putative content as the inputs, which are processed by a single layer of $2*w+1$ neurons, i.e., the number of neurons matches the number of inputs. This layer is connected to an output neuron that produces the real-valued propensity for protein binding. We train this network on the combined set of the structure-annotated and the disorder-annotated training proteins by freezing the disorder-trained and the structure-trained CNNs. We use learning rate = 0.0003 and batch size = 256 that we determine based on the grid search on the training dataset. We use an early stop condition to minimize overfitting. We stop training when the AUC on the validation set does not improve.

The second design relies on a more complex transformer encoder network with the same inputs and sliding window of size w as the first design. This network has 3 encoding layers followed by a feed-forward layer with 32 neurons which feeds into the output neuron. Each encoding layer includes a multi-head self-attention layer with 3 heads and a feed-forward layer with 64 neurons. The size of each head in the multi-head self-attention layer is the same as the window of size w. Similar to the feed-forward design, we train this transformer network on the structure-annotated and the disorder-annotated training proteins by freezing the disorder-trained and the structure-trained transformers. We apply the learning rate of 0.0003 and batch size of 256 that we determine based on the grid search on the training dataset, and we use the early stop condition.

Finally, we craft rules that rely on the background knowledge to implement the combination layer. These rules do not utilize a training process. We formulate these rules to ensure that they merge PBRs predicted by the two CNN models which target different structural context (structured vs. disordered regions). To do that, we first binarize the predicted propensities using thresholds that maximize F1 value on the validation dataset. We use the binary predictions to differentiate how the underlying propensities are combined together, where we aim to produce larger combined propensity when either and/or both of the two models predict putative PBRs. Moreover, we use IUPred2A's predictions to rationally select a more suitable model when we do not attempt to combine their predictions, e.g., when neither predicts PBRs. We choose the structure-trained model when the putative disorder content produced by IUPred2A is relatively low (>0.13, which is the average disorder content in the training dataset), and otherwise we select the disorder-trained dataset. We devise four rules that gradually consider more information:

- CombinationRule1 picks higher of two putative propensities when the residue is predicted as PBR by either model and lower when neither model predicts PBRs
- CombinationRule2 chooses the propensity predicted by one of the two models based on the putative disorder content from IUPred2A
- CombinationRule3 selects higher of two putative propensities when the residue is predicted as PBR by both models and lower when neither model predicts PBRs. If one model predicts PBR then we select the score based on the disorder content prediction from IUPred2A
- CombinationRule4 uses higher of two putative propensities when the residue is predicted as PBR by either model, and otherwise it selects the score based on the disorder content prediction from IUPred2A

Suppl. Table S6 compares results produced by the feed-forward networks, the transformer networks and the four rules on the validation dataset. We consider and empirically compare several window sizes w = {3, 5, 7, 9, 15 and 25} for the two networks. The results of the transformer networks are more accurate than the feed-forward networks, irrespective of the window size. This is expected since transformers consider the sequence order. Moreover, we find that the predictions from the transformer networks slightly benefit from larger window sizes, with the best results for w = 15 and 25. These designs are often among the top three results shown in bold font in Suppl. Table S6. The best rule is the CombinationRule4, which outperforms the other 3 rules of thumb. Our analysis also highlights the value of selecting models based on the putative disorder content, given the relatively low quality of the predictions produced with the CombinationRule1. Interestingly, the predictive quality of the CombinationRule4 is marginally better than the predictions from the transformer network: overall AUC = 0.792 vs. 0.791; AUC for disorder-annotated proteins = 0.703 vs. 0.696; AUC for the structure-annotated proteins = 0.822 vs. 0.823. This rule also generates predictions with the lowest amount of cross-predictions (the lowest AUCPC = 0.314) and over-predictions (the lowest AUOPC = 0.199). This suggests that carefully crafted rules that rely on the background knowledge could rival deep networks that are agnostic to the underlying characteristics of the problems. Given the slightly more favorable results produced by the CombinationRule4 and the fact that it is much faster to compute compared to processing the data through a transformer network, we select this rule to implement DeepPRObind. We re-tested these results on the test set (Suppl. Table S7) and found that the above observations are consistent across both datasets, further justifying our design choice.

3 Results

3.1 Comparative assessment

We compare DeepPRObind against the 12 disorder- and structure-trained predictors on the test dataset. We collect predictions from the 12 methods using web servers or implementations provided by their authors. We also consider combining results of the best predictors to evaluate whether they could outperform DeepPRObind. We utilize the four combination rules to ensemble the best structure-trained method, SCRIBER, with the best disorder-trained method, DisoRDPbind. These are the methods that secure the highest AUC scores for the structure-annotated and disorder-annotated proteins on the test dataset, respectively (Suppl. Table S8). We note that our comparison includes HybridPBRpred, which is another combination/meta-predictor. Since we have ~14%

of PBRs in our dataset, we set the binary predictions across all methods to specificity = 0.9. This allows us to directly compare the binary metrics (F1, precision, recall, and MCC) across methods. We evaluate significance of the differences in predictive performance between DeepPRObind and the other predictors, which quantifies robustness of the improvements offered by our method by sampling different protein sets drawn from the test dataset. We adopt procedure applied in recent related studies [13, 19, 34], where we compare results over ten protein sets which represent 50% of the test proteins that we draw at random. We use the Kolmogorov-Smirnov test at 0.05 significance level to check if a given set of measurements is normal. We use the t-test to quantify significance of differences for normal measurements, otherwise we use the Wilcoxon rank sum test.

Table 1, which gives results on the test dataset, reveals that DeepPRObind provides the best predictions across all metrics, with the differences that are statistically significant (p-value < 0.05). The DeepPRObind's AUC, AUPRC and F1 are 0.808, 0.457 and 0.454, respectively, compared with 0.772, 0.337 and 0.359 for the second-best HybridPBRpred, the only other method that was designed to cover prediction for the structured and disordered proteins. This corresponds to the improvement by (0.808-0.772)/0.772 = 5%, (0.457-0.337)/0.337 = 36%, and (0.454-0.359)/0.359 = 26%, respectively. The best combination result that relies on the CombinationRule1 to predict PBRs is only modestly lower that HybridPBRpred (AUC = 0.758), while at the same time substantially improving over the results of its input predictors, DisoRDPbind (AUC = 0.699) and SCRIBER (AUC = 0.692). The relatively large and statistically significant difference between the results of this combination approach and the better results of DeepPRObind (AUC = 0.758 vs. 0.808; p-value < 0.05) can be attributed to the use of a well-designed deep network model and new feature types. Suppl. Figure S1, which gives the ROC and precision-recall curves, confirms that DeepPRObind performs substantially better than the selected best-performing current methods including HybridPBRpred, SCRIBER and DisoRDPbind. The improvements are particularly large when FPR values are small, < 0.1, which is arguably the most useful part of the predictions where methods do not overpredict the amount of PBRs. Some predictors provide results that are close to random levels, with AUC<0.55 and MCC<0.05. This can be explained by the fact that they were designed a long time ago using limited amount of training data (PSIVER and SPPIDER) and because they address a related but different prediction target, with SPRINT predicting interactions with peptides and fMoRFpred predicting binding for short MoRFs [49].

Table 1. Comparative assessment on the test set. The binary predictions use thresholds that equalize specificity to 0.9 across methods to allow for side-by-side comparisons. + means that DeepPRObind is statistically significantly better than another method at p-value<0.05. The best results for each column are shown in bold font. Predictors are sorted by their AUCs.

Methods	AUC	AUPRC	F1	Precision	Recall	MCC
DeepPRObind	**0.808**	**0.457**	**0.454**	**0.448**	**0.460**	**0.355**
HybridPBRpred	0.772+	0.337+	0.359+	0.377+	0.343+	0.251+
CombinationRule1 using SCRIBER and DisoRDPbind	0.758+	0.336+	0.381+	0.393+	0.371+	0.274+
CombinationRule3 using SCRIBER and DisoRDPbind	0.756+	0.334+	0.392+	0.401+	0.383+	0.286+
CombinationRule4 using SCRIBER and DisoRDPbind	0.741+	0.324+	0.380+	0.393+	0.369+	0.274+
CombinationRule2 using SCRIBER and DisoRDPbind	0.735+	0.320+	0.373+	0.396+	0.352+	0.269+
DisoRDPbind	0.699+	0.312+	0.381+	0.394+	0.369+	0.274+
SCRIBER	0.692+	0.294+	0.334+	0.363+	0.310+	0.227+
ANCHOR	0.658+	0.271+	0.338+	0.358+	0.320+	0.228+
LORIS	0.592+	0.195+	0.192+	0.230+	0.165+	0.078+
SSWRF	0.588+	0.181+	0.156+	0.194+	0.130+	0.041+
SPRINGS	0.587+	0.198+	0.213+	0.253+	0.184+	0.100+
CRFPPI	0.572+	0.183+	0.170+	0.208+	0.143+	0.055+
fMoRFpred	0.524+	0.168+	0.144+	0.181+	0.119+	0.027+
SPPIDER	0.477+	0.156+	0.163+	0.200+	0.137+	0.046+
PSIVER	0.473+	0.144+	0.105+	0.134+	0.086+	-0.017+
SPRINT	0.403+	0.133+	0.095+	0.123+	0.078+	-0.027+

We also evaluate results separately for the structure-annotated and the disorder-annotated test proteins (Supp. Table S8). DeepPRObind provides the highest values for all metrics across both subsets of proteins when compared to the 12 predictors and the four results based on the four combination rules. The improvements are always statistically significant for the disorder-annotated proteins. We observe that DeepPRObind provides higher values which are not statistically significant for a few metrics on the structure-annotated proteins when compared with SCRIBER and hybridPBRpred. However, this is compensated by much larger and statistically significant differences for the disorder-annotated test proteins. We note that the disorder-trained predictors, DisoRDPbind and ANCHOR (excluding fMoRFpred that targets prediction of short MoRF regions), outperform the structure-trained predictors on the disorder-annotated test proteins. The results reverse on the structure-trained proteins where the structure-trained methods, such as SCRIBER, CRFPPI, SSWRF and LORIS produce more accurate results than the disorder-trained methods. These results are in agreement with the recently published study [13], which motivated the release of HybridPBRpred and the development of DeepPRObind. Overall, we conclude that DeepPRObind secures accurate and the best results on the test dataset, irrespective of the type of the PBR annotations.

3.2 Evaluation of cross-predictions and over-predictions

Recent studies demonstrate that current sequence-based predictors frequently predict residues that do not bind proteins as PBRs [13, 19]. Consequently, we analyze the cross-predictions (i.e., residues that bind other types of ligands, such as nucleic acids and small ligands, predicted as PBRs) and over-predictions (i.e., residues that do not interact identified as PBRs) using AUCPC and AUOPC, respectively. Table 2 provides these values for the 13 predictors on the test dataset. Higher AUCPC and AUOPC values denote higher rates of cross- and over-predictions; they are defined in the Supplement. Results show that DeepPRObind produces the lowest amounts of both cross-predictions and over-predictions. The improvements over the 12 current methods are relatively large and statistically significant (p-value < 0.05). The second-best HybridPBRpred secures AUCPC of 0.343 and AUOPC of 0.220 vs. 0.297 and 0.185 for DeepPRObind. We directly compare the corresponding cross-prediction and over-prediction curves for the top four methods that secure AUCPC ≤ 0.4 and AUOPC ≤ 0.3 in Figure 2. We find that the DeepPRObind's curves are particularly favorable when TPRs are < 0.3, leading to high quality prediction of PBRs coupled with near-zero rates of the cross-predictions and over-predictions.

Table 2. Comparative assessment of the cross-predictions and the over-predictions on the test dataset. + means that DeepPRObind is statistically significantly better than another method at *p*-value<0.05. The best results for each column are shown in bold font. Predictors are sorted by their AUCPC values, where lower values denote higher quality predictions

Methods	AUCPC	AUOPC
SPRINT	0.580+	0.598+
SPPIDER	0.555+	0.520+
PSIVER	0.519+	0.528+
CRFPPI	0.502+	0.423+
FMoRFpred	0.485+	0.475+
SSWRF	0.464+	0.408+
SPRINGS	0.446+	0.410+
ANCHOR	0.445+	0.335+
LORIS	0.436+	0.406+
disoRDPbind	0.401+	0.294+
SCRIBER	0.376+	0.303+
hybridPBRpred	0.343+	0.220+
DeepPRObind	**0.297**	**0.185**

Suppl. Table S9 summarizes results for the subsets of the disorder annotated and the order annotated proteins in the test dataset. The results are consistent with the overall results from Table 2. One exception is SCRIBER that secures the lowest rates of cross-predictions for the structure-annotated proteins. We note that SCRIBER was specifically designed to reduce the cross-predictions [34] but the caveat is that this strong result comes with relatively high cross- and over-prediction rates for the disorder annotated proteins. DeepPRObind is the only

accurate predictor that generates similarly low AUCPC values across the two subsets of proteins, i.e., 0.294 for the structure annotated vs. 0.296 for the disorder annotated proteins. In the nutshell, we find that DeepPRObind generates relatively low amounts of the cross-predictions and over-predictions.

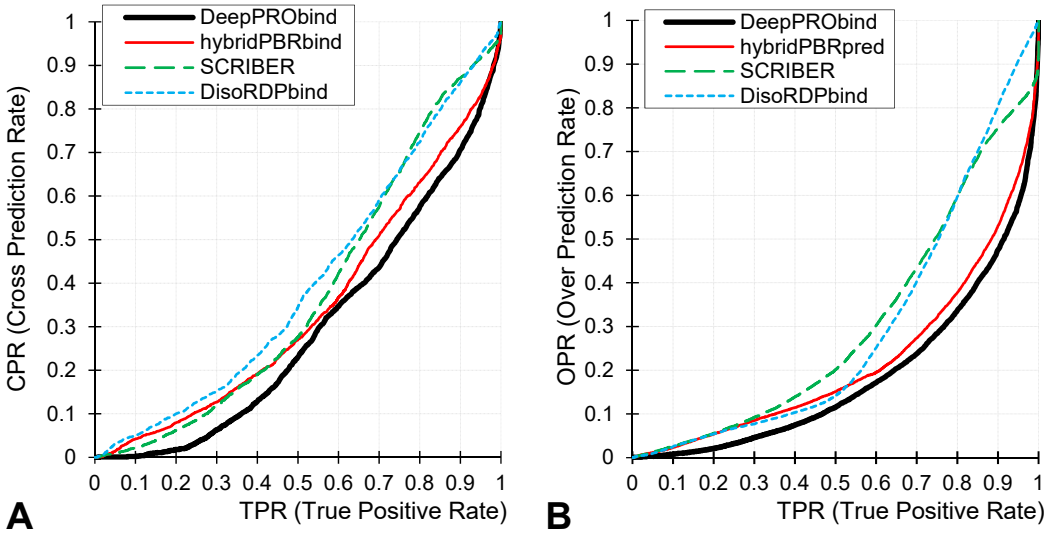

Figure 2. Cross-prediction curves (Panel A) and over-prediction curves (Panel B) on the test dataset for DeepPRObind (black line) and three other methods with the lowest AUCPC and AUOPC on the test dataset: hybridPBRpred (red), SCRIBER (green), and DisoRDPbind (blue).

3.3 Analysis of PPIs in the yeast proteome

We apply DeepPRObind to analyze PPIs in yeast, which is a commonly used model organism that has one of the most complete PPI networks. We collect PPIs from the mentha resource [11], which is updated weekly and combines manually curated PPIs from several interaction databases from the IMEx (International Molecular Exchange) consortium. Several past studies have found that intrinsically disordered proteins are enriched among the hub proteins [57, 58, 76], which are highly connected in the PPI networks. This can be explained by the fact that disordered proteins are involved in one-to-many and in many-to-one interactions [54, 55] and since disordered regions may serve as scaffolds to assemble groups of interacting proteins [77, 78]. We empirically investigate whether the amounts of the putative PBRs generated by the disorder-trained vs. structure-trained models from DeepPRObind agree with these observations.

We collect PPIs from mentha and obtain sequences for the corresponding 6,063 yeast proteins from UniProt [79]. We release DeepPRObind's predictions for the yeast proteins at https://www.csuligroup.com/DeepPRObind/. Next, we sort the yeast proteins by the number of their interaction partners and compare the content of putative PBRs (fraction of predicted PBRs in a protein sequence) and the average propensities produced by the structure-trained and the disorder-trained models from DeepPRObind between the most-connected proteins (i.e., hubs) and the least-connected proteins. Since there is no clear consensus on how to define hubs (i.e., how many PPIs a given protein has to be involved in to be classified as a hub), we repeat this analysis by comparing the top 10%, 20% and 30% of most connected proteins against the same fraction of the least-connected proteins. We report results in Table 3. We find a consistent pattern where DeepPRObind identifies statistically significantly more disordered putative PBRs among the hub proteins and significantly more structured putative PBRs among the least connected proteins. This pattern holds true irrespective of how hubs are defined (top 10, 20 or 30%) and whether we measure the content or average propensities. This result is consistent with the findings in the literature [57, 58, 76], where the disordered proteins were found to be over-represented among the hubs. This provides further support for the claim that DeepPRObind produces accurate predictions.

Table 3. Analysis DeepPRObind's predictions in the yeast proteome. We compare the content of putative PBRs (fraction of PBRs in protein sequences) and average per-protein propensities generated by the disorder-trained and the structure-trained modules of the DeepPRObind model between the most-connected and the least-connected proteins in the PPI network of yeast. We report median [25th percentile, 75th percentile] for each protein set. We evaluate statistical significance of differences in the content and propensities between the most connected and the least connected proteins using the procedure described in Section 3.1. + means that the difference is statistically significant at p-value<0.05. Higher content/propensity is denoted by bold font

Selection of the most connected/hub protein vs. the least connected proteins		Predicted content of PBRs		Average scores of predicted PBRs	
		Structure-based model	Disorder-based model	Structure-based model	Disorder-based model
Top 10% vs. bottom 10%	Most-connected	0.01[0.00, 0.05] +	0.13[0.01, 0.31] +	0.14[0.11, 0.26] +	0.29[0.15, 0.50] +
	Least-connected	**0.54[0.01, 0.98]**	0.04[0.00, 0.20]	**0.54[0.31, 0.75]**	0.17 [0.12, 0.24]
Top 20% vs. bottom 20%	Most-connected	0.01[0.00, 0.07] +	**0.13[0.02, 0.32] +**	0.18[0.11, 0.28] +	**0.29 [0.15, 0.50] +**
	Least-connected	**0.18[0.03, 0.82]**	0.03[0.00, 0.20]	**0.35[0.22, 0.63]**	0.17[0.11, 0.28]
Top 30% vs. bottom 30%	Most-connected	0.02[0.00, 0.08] +	**0.12[0.02, 0.32] +**	0.19[0.12, 0.29] +	**0.29[0.15, 0.51] +**
	Least-connected	0.10[0.02, 0.58]	0.03[0.00, 0.20]	0.31[0.19, 0.54]	0.17[0.11, 0.32]

3.4 Case study

We illustrate DeepPRObind's predictions using one of the test proteins, structured protein YbaA from *Shigella Flexneri* (UniProt ID: P0AAQ9). This case study is not meant to represent a broader comparative analysis that we already cover in Sections 3.1 and 3.2, but instead aims to visualize and compare outputs generated by DeepPRObind and selected best other methods. We select methods that secure high AUCs in Table 1: ANCHOR, SCRIBER, DisoRDPbind and HybridPBRpred. This protein forms a dimer where residues identified with black markers in Suppl. Fig S2 compose the corresponding interface (PBD ID: 2OKQ). ANCHOR and DeepRDPbind do not predict PBRs, i.e., propensities that they generate are below the threshold which is used to identify putative PBRs (Suppl. Figure S2). This can be explained by the fact that they target predictions of disordered PBRs while this protein is structured. The structure-trained SCRIBER correctly identifies majority of native PBRs, although it over-predicts PBRs. Both methods capable of finding structured and disordered PBRs, HybridPBRpred and DeepPRObind, also correctly predict many of the native PBRs. HybridPBRpred, which is a consensus of SCRIBER and DisoRDPbind, identifies a subset of PBRs that are predicted by SCRIBER. This stems from the fact that this consensus relies on a simple combination of the two predictions. In contrast, DeepPRObind produces a prediction that identifies a larger fraction of native PBRs and where these predictions (particularly residues with high putative propensities > 0.4) are better aligned with the high-density regions of binding residues, which are localized at both termini and in the middle of the sequence. This shows that while disorder-trained methods target predictions for the disordered proteins, DeepPRObind makes predictions that identify PBRs in the structured proteins and which are different than the outputs from HybridPBRpred and SCRIBER.

3.5 Web server

We provide a free and convenient web server that implements DeepPRObind at https://csuligroup.com/DeepPRObind/. The web server supports batch predictions of up to 20 protein sequences in one request. The input protein sequence(s) should be formatted using FASTA format and can be entered via a text box or a text file. Predictions are done on the server side and they take about 40 seconds for an average size protein sequence (~200 amino acids in length). This low runtime is a consequence of using fast tools to produce inputs (IUPred2A, ANCHOR, ASAquick, single-sequence version of PSIPRED and HHblits) and a relatively small deep network, which reduces the time needed to process predictions. Users can optionally provide email address where we send a notification once the predictions are completed. The outputs include the putative propensities and binary predictions of PBRs for each residue in the input protein sequences. We provide these predictions in three complementary formats: 1) a parsable text file that can be downloaded from the URL provided by the web server; 2) webpage that tabularizes results per residues and annotates putative PBRs in green color; 3) an interactive graphical format that visualizes propensities and binary predictions in the format similar to Suppl. Figure S2. The graphical plot can be adjusted to zoom in and out on specific sequence regions and

provides useful information, such as numeric values of propensity and residue positions, on mouse hover. We will store the user-generated predictions for at least one month on the web server.

4 Summary

Predictors of PBRs in protein sequences are categorized into two major classes: structure-trained vs. disorder-trained [13, 14]. Since disordered PBRs are different from structured PBRs in several key ways [54, 56, 57], a recent study finds that the structure-trained and the disorder-trained predictors provide inaccurate predictions for the other annotation type [13]. The existing consensus-based solution, HybridPBRpred [13], relies on a simple model that limits its predictive performance and results in substantial amounts of cross-predictions and over-predictions. To this end, we introduce a modern deep learner, DeepPRObind, that benefits from a carefully designed modular convolutional architecture. We empirically demonstrate that the two innovations that underlie this architecture, i.e., use of aggregate feature and hand-crafted approach to combine results from modules, produce substantial improvements in predictive performance (see Sections 2.4.2. and 2.4.3). Using a recently introduced low-similarity test set [13], we show that DeepPRObind generates the most accurate predictions of structured and disordered PBRs and significantly reduces cross-predictions when compared to the comprehensive collection of 12 current predictors of PBRs that include HybridPBRpred. Given the relatively low runtime of DeepPRObind (~40 seconds per average size protein), we further validate results generated by DeepPRObind based on a comparative analysis of putative structured and disordered PBRs in the yeast proteome. We show that the predicted disordered PBRs are significantly enriched among hub proteins, which agrees with published observations [57, 58, 76] and further demonstrates value of the DeepPRObind predictor. We release a free and user-friendly web server that implements DeepPRObind at https://www.csuligroup.com/DeepPRObind/.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (61832019), 111 Project (B18059), and the Hunan Provincial Science and Technology Program (2019CB1007 and 2021RC0048). L.K. was supported in part by the Robert J. Mattauch Endowed Chair funds.

References

[1] Braun P, Gingras ACJP. History of protein–protein interactions: From egg-white to complex networks. 2012;12:1478-98.

[2] Athanasios A, Charalampos V, Vasileios TJCdm. Protein-protein interaction (PPI) network: recent advances in drug discovery. 2017;18:5-10.

[3] Kuzmanov U, Emili AJGm. Protein-protein interaction networks: probing disease mechanisms using model systems. 2013;5:1-12.

[4] Uversky VN. Wrecked regulation of intrinsically disordered proteins in diseases: pathogenicity of deregulated regulators. Front Mol Biosci. 2014;1:6.

[5] Vakser IA. Protein-protein docking: from interaction to interactome. Biophys J. 2014;107:1785-93.

[6] Scott DE, Bayly AR, Abell C, Skidmore J. Small molecules, big targets: drug discovery faces the protein-protein interaction challenge. Nature Reviews Drug Discovery. 2016;15:533-50.

[7] wwPDB consortium. Protein Data Bank: the single global archive for 3D macromolecular structure data. Nucleic Acids Res. 2019;47:D520-D8.

[8] Hatos A, Hajdu-Soltesz B, Monzon AM, Palopoli N, Alvarez L, Aykac-Fas B, et al. DisProt: intrinsic protein disorder annotation in 2020. Nucleic Acids Res. 2020;48:D269-D76.

[9] Yang J, Roy A, Zhang Y. BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions. Nucleic Acids Res. 2013;41:D1096-103.

[10] Szklarczyk D, Gable AL, Nastou KC, Lyon D, Kirsch R, Pyysalo S, et al. The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids Research. 2021;49:D605-D12.

[11] Calderone A, Castagnoli L, Cesareni G. mentha: a resource for browsing integrated protein-interaction networks. Nat Methods. 2013;10:690-1.

[12] Oughtred R, Rust J, Chang C, Breitkreutz BJ, Stark C, Willems A, et al. The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. Protein Sci. 2021;30:187-200.

[13] Zhang J, Ghadermarzi S, Kurgan L. Prediction of protein-binding residues: dichotomy of sequence-based methods developed using structured complexes versus disordered proteins. Bioinformatics. 2020;36:4729-38.

[14] Barik A, Kurgan L. A comprehensive overview of sequence-based protein-binding residue predictions for structured and disordered regions. Protein Interactions2020. p. 33-58.

[15] Zhang J, Ma Z, Kurgan L. Comprehensive review and empirical analysis of hallmarks of DNA-, RNA- and protein-binding residues in protein chains. Brief Bioinform. 2019;20:1250-68.
[16] Aumentado-Armstrong TT, Istrate B, Murgita RAJAfMB. Algorithmic approaches to protein-protein interaction site prediction. 2015;10:1-21.
[17] Xue LC, Dobbs D, Bonvin AM, Honavar VJFl. Computational prediction of protein interfaces: A review of data driven methods. 2015;589:3516-26.
[18] Esmaielbeiki R, Krawczyk K, Knapp B, Nebel J-C, Deane CMJBib. Progress and challenges in predicting protein interfaces. 2016;17:117-31.
[19] Zhang J, Kurgan L. Review and comparative assessment of sequence-based predictors of protein-binding residues. Brief Bioinform. 2018;19:821-37.
[20] Porollo A, Meller JJPS, Function,, Bioinformatics. Prediction-based fingerprints of protein–protein interactions. 2007;66:630-45.
[21] Murakami Y, Mizuguchi KJB. Applying the Naïve Bayes classifier with kernel density estimation to the prediction of protein–protein interaction sites. 2010;26:1841-8.
[22] Dhole K, Singh G, Pai PP, Mondal SJJotb. Sequence-based prediction of protein–protein interaction sites with L1-logreg classifier. 2014;348:47-54.
[23] Singh G, Dhole K, Pai PP, Mondal S. SPRINGS: prediction of protein-protein interaction sites using artificial neural networks. PeerJ PrePrints; 2014.
[24] Wang DD, Wang R, Yan HJN. Fast prediction of protein–protein interaction sites based on extreme learning machines. 2014;128:258-66.
[25] Geng H, Lu T, Lin X, Liu Y, Yan FJBri. Prediction of protein-protein interaction sites based on naive Bayes classifier. 2015;2015.
[26] Wei Z-S, Yang J-Y, Shen H-B, Yu D-JJIton. A cascade random forests algorithm for predicting protein-protein interaction sites. 2015;14:746-60.
[27] Liu G-H, Shen H-B, Yu D-JJTJomb. Prediction of protein–protein interaction sites with machine-learning-based data-cleaning and post-filtering procedures. 2016;249:141-53.
[28] Taherzadeh G, Yang Y, Zhang T, Liew AWC, Zhou YJJocc. Sequence-based prediction of protein–peptide binding sites using support vector machine. 2016;37:1223-9.
[29] Jia J, Liu Z, Xiao X, Liu B, Chou K-CJM. iPPBS-Opt: a sequence-based ensemble classifier for identifying protein-protein binding sites by optimizing imbalanced training datasets. 2016;21:95.
[30] Wei Z-S, Han K, Yang J-Y, Shen H-B, Yu D-JJN. Protein–protein interaction sites prediction by ensembling SVM and sample-weighted random forests. 2016;193:201-12.
[31] Tahir M, Hayat MJAiim. Machine learning based identification of protein–protein interactions using derived features of physiochemical properties and evolutionary profiles. 2017;78:61-71.
[32] Guo H, Liu B, Cai D, Lu TJIJoML, Cybernetics. Predicting protein–protein interaction sites using modified support vector machine. 2018;9:393-8.
[33] Wang X, Yu B, Ma A, Chen C, Liu B, Ma QJB. Protein–protein interaction sites prediction by ensemble random forests with synthetic minority oversampling technique. 2019;35:2395-402.
[34] Zhang J, Kurgan L. SCRIBER: accurate and partner type-specific prediction of protein-binding residues from proteins sequences. Bioinformatics. 2019;35:i343-i53.
[35] Zeng M, Zhang F, Wu F-X, Li Y, Wang J, Li MJB. Protein–protein interaction site prediction through combining local and global features with deep neural networks. 2020;36:1114-20.
[36] Li Y, Golding GB, Ilie LJB. DELPHI: accurate deep ensemble model for protein interaction sites prediction. 2021;37:896-904.
[37] Katuwawala A, Peng Z, Yang J, Kurgan L. Computational Prediction of MoRFs, Short Disorder-to-order Transitioning Protein Binding Regions. Comput Struct Biotechnol J. 2019;17:454-62.
[38] Cheng Y, Oldfield CJ, Meng J, Romero P, Uversky VN, Dunker AKJB. Mining α-helix-forming molecular recognition features with cross species sequence alignments. 2007;46:13468-77.
[39] Dosztányi Z, Mészáros B, Simon IJB. ANCHOR: web server for predicting protein binding regions in disordered proteins. 2009;25:2745-6.
[40] Meszaros B, Erdos G, Dosztanyi Z. IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. Nucleic Acids Res. 2018;46:W329-W37.
[41] Xue B, Dunker AK, Uversky VNJIjoms. Retro-MoRFs: identifying protein binding sites by normal and reverse alignment and intrinsic disorder prediction. 2010;11:3725-47.
[42] Disfani FM, Hsu WL, Mizianty MJ, Oldfield CJ, Xue B, Dunker AK, et al. MoRFpred, a computational tool for sequence-based prediction and characterization of short disorder-to-order transitioning binding regions in proteins. Bioinformatics. 2012;28:i75-83.
[43] Fang C, Noguchi T, Tominaga D, Yamana HJBb. MFSPSSMpred: identifying short disorder-to-order binding regions in disordered proteins based on contextual local evolutionary conservation. 2013;14:1-14.
[44] Jones DT, Cozzetto DJB. DISOPRED3: precise disordered region predictions with annotated protein-binding activity. 2015;31:857-63.
[45] Peng Z, Kurgan L. High-throughput prediction of RNA, DNA and protein binding regions mediated by intrinsic disorder. Nucleic Acids Res. 2015;43:e121.
[46] Peng Z, Wang C, Uversky VN, Kurgan L. Prediction of Disordered RNA, DNA, and Protein Binding Regions Using DisoRDPbind. Methods Mol Biol. 2017;1484:187-203.
[47] Zhao B, Katuwawala A, Oldfield CJ, Dunker AK, Faraggi E, Gsponer J, et al. DescribePROT: database of amino acid-level protein structure and function predictions. Nucleic Acids Res. 2021;49:D298-D308.

[48] Malhis N, Jacobson M, Gsponer JJNar. MoRFchibi SYSTEM: software tools for the identification of MoRFs in protein sequences. 2016;44:W488-W93.
[49] Yan J, Dunker AK, Uversky VN, Kurgan L. Molecular recognition features (MoRFs) in three domains of life. Mol Biosyst. 2016;12:697-710.
[50] Sharma R, Bayarjargal M, Tsunoda T, Patil A, Sharma AJJotb. MoRFPred-plus: computational identification of MoRFs in protein sequences using physicochemical properties and HMM profiles. 2018;437:9-16.
[51] Sharma R, Raicar G, Tsunoda T, Patil A, Sharma AJB. OPAL: prediction of MoRF regions in intrinsically disordered protein sequences. 2018;34:1850-8.
[52] Sharma R, Sharma A, Raicar G, Tsunoda T, Patil AJP. OPAL+: length-specific MoRF prediction in intrinsically disordered protein sequences. 2019;19:1800058.
[53] Dyson HJ, Wright PEJCoisb. Coupling of folding and binding for unstructured proteins. 2002;12:54-60.
[54] Hsu WL, Oldfield CJ, Xue B, Meng J, Huang F, Romero P, et al. Exploring the binding diversity of intrinsically disordered proteins involved in one-to-many binding. Protein Sci. 2013;22:258-73.
[55] Uversky VN. Analyzing IDPs in Interactomes. Methods Mol Biol. 2020;2141:895-945.
[56] Wu Z, Hu G, Yang J, Peng Z, Uversky VN, Kurgan L. In various protein complexes, disordered protomers have large per-residue surface areas and area of protein-, DNA- and RNA-binding interfaces. FEBS Lett. 2015;589:2561-9.
[57] Haynes C, Oldfield CJ, Ji F, Klitgord N, Cusick ME, Radivojac P, et al. Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes. 2006;2:e100.
[58] Hu G, Wu Z, Uversky VN, Kurgan L. Functional Analysis of Human Hub Proteins and Their Interactors Involved in the Intrinsic Disorder-Enriched Interactions. Int J Mol Sci. 2017;18.
[59] Necci M, Piovesan D, Predictors C, DisProt C, Tosatto SCE. Critical assessment of protein intrinsic disorder prediction. Nat Methods. 2021;18:472-81.
[60] Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. 1997;25:3389-402.
[61] Zhang J, Kurgan LJB. SCRIBER: accurate and partner type-specific prediction of protein-binding residues from proteins sequences. 2019;35:i343-i53.
[62] 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.
[63] Hanson J, Litfin T, Paliwal K, Zhou Y. Identifying molecular recognition features in intrinsically disordered regions of proteins by transfer learning. Bioinformatics. 2020;36:1107-13.
[64] Zhang T, Faraggi E, Xue B, Dunker AK, Uversky VN, Zhou Y. SPINE-D: accurate prediction of short and long disordered regions by a single neural-network based method. J Biomol Struct Dyn. 2012;29:799-813.
[65] McGuffin LJ, Bryson K, Jones DTJB. The PSIPRED protein structure prediction server. 2000;16:404-5.
[66] Remmert M, Biegert A, Hauser A, Söding JJNm. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. 2012;9:173-5.
[67] Li F, Chen J, Leier A, Marquez-Lago T, Liu Q, Wang Y, et al. DeepCleave: a deep learning predictor for caspase and matrix metalloprotease substrates and cleavage sites. Bioinformatics. 2020;36:1057-65.
[68] Peng Z, Xing Q, Kurgan L. APOD: accurate sequence-based predictor of disordered flexible linkers. Bioinformatics. 2020;36:i754-i61.
[69] Meng F, Kurgan L. DFLpred: High-throughput prediction of disordered flexible linker regions in protein sequences. Bioinformatics. 2016;32:i341-i50.
[70] Zeng HY, Edwards MD, Liu G, Gifford DK. Convolutional neural network architectures for predicting DNA-protein binding. Bioinformatics. 2016;32:121-7.
[71] Campen A, Williams RM, Brown CJ, Meng J, Uversky VN, Dunker AK. TOP-IDP-scale: a new amino acid scale measuring propensity for intrinsic disorder. Protein Pept Lett. 2008;15:956-63.
[72] Yan J, Cheng J, Kurgan L, Uversky VN. Structural and functional analysis of "non-smelly" proteins. Cell Mol Life Sci. 2020;77:2423-40.
[73] Zhao B, Kurgan L. Compositional Bias of Intrinsically Disordered Proteins and Regions and Their Predictions. Biomolecules. 2022;12.
[74] Katuwawala A, Oldfield CJ, Kurgan L. Accuracy of protein-level disorder predictions. Brief Bioinform. 2020;21:1509-22.
[75] Walsh I, Giollo M, Di Domenico T, Ferrari C, Zimmermann O, Tosatto SC. Comprehensive large-scale assessment of intrinsic protein disorder. Bioinformatics. 2015;31:201-8.
[76] Ekman D, Light S, Bjorklund AK, Elofsson A. What properties characterize the hub proteins of the protein-protein interaction network of Saccharomyces cerevisiae? Genome Biology. 2006;7.
[77] Xue B, Romero PR, Noutsou M, Maurice MM, Rudiger SG, William AM, Jr., et al. Stochastic machines as a colocalization mechanism for scaffold protein function. FEBS Lett. 2013;587:1587-91.
[78] Uversky VN. The multifaceted roles of intrinsic disorder in protein complexes. FEBS Lett. 2015;589:2498-506.
[79] UniProt C. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res. 2021;49:D480-D9.