DeepDISOBind: Accurate prediction of RNA, DNA and protein binding intrinsically disordered residues with deep multi-task learning

Fuhao Zhang[1], Bi Zhao[2], Wenbo Shi[1], Min Li[1*], and Lukasz Kurgan[2*]

[1]Hunan Provincial Key Lab on Bioinformatics, School of Computer Science and Engineering, Central South University, Changsha, 410083, China.

[2]Department of Computer Science, Virginia Commonwealth University, Richmond, VA, 23284, USA.

*corresponding authors

Min Li: Tel +86 073 188 879 560; Email limin@mail.csu.edu.cn

Lukasz Kurgan: Tel +1 804 827 3986; Email lkurgan@vcu.edu

ABSTRACT

Proteins with intrinsically disordered regions (IDRs) are common among eukaryotes. Many IDRs interact with nucleic acids and proteins. Annotation of these interactions is supported by computational predictors but to date only one tool that predicts interactions with nucleic acids was released and recent assessments demonstrate that current predictors offer modest levels of accuracy. We develop DeepDISOBind, an innovative deep multi-task architecture that accurately predicts DNA, RNA and protein binding IDRs from protein sequences. DeepDISOBind relies on an information-rich sequence profile that is processed by an innovative multi-task deep neural network where subsequent layers are gradually specialized to predict interactions with specific partner types. The common input layer links to a layer that differentiates protein and nucleic acids binding, which further links to layers that discriminate between DNA and RNA interactions. Empirical tests show that this multi-task design provides statistically significant gains in predictive quality across the three partner types when compared to a single-task design and a representative selection of the existing methods that cover both disorder- and structure-trained tools. Analysis of the predictions on the human proteome reveals that DeepDISOBind predictions can be encoded into protein-level propensities that accurately predict DNA and RNA binding proteins and protein hubs. DeepDISOBind's is available at https://www.csuligroup.com/DeepDISOBind/

Keywords: intrinsic disorder; protein-protein interactions; protein-nucleic acids interactions; deep learning.

Fuhao Zhang is a PhD student at the School of Computer Science and Engineering, Central South University, China. His research focuses on computational prediction and characterization of protein structure and function.

Bi Zhao earned PhD from University of South Florida in 2019 and currently is a postdoctoral fellow in the Computer Science department at the Virginia Commonwealth University. She spearheaded the development of multiple bioinformatics resources for protein disorder and disorder function prediction.

Wenbo Shi is a Master's student at the School of Computer Science and Engineering, Central South University, China, who specializes in the development of bioinformatics algorithms.

Min Li is the vice-Dean and Professor at the School of Computer Science and Engineering, Central South University, China. Her main research interests include bioinformatics and systems biology.

Lukasz Kurgan is a Fellow of AIMBE and the Robert J. Mattauch Endowed Professor of Computer Science at the Virginia Commonwealth University. His research work encompasses structural and functional characterization of proteins. He serves on the Editorial Board of Bioinformatics and as the Associate Editor-in-Chief of Biomolecules. Details about his research lab are at http://biomine.cs.vcu.edu/.

1 INTRODUCTION

Intrinsically disordered regions (IDRs) lack stable tertiary structures and form dynamic conformational ensembles under physiological conditions [1, 2]. Recent bioinformatics studies reveal that disorder is highly abundant in nature [3], with about 20% of residues in eukaryotic proteins estimated to be disordered [4]. Proteins with IDRs are involved in a variety of cellular functions [5, 6]. Many IDRs interact with partner molecules including

DNA, RNA and proteins [7-13]. More specifically, the version 8.1 of the DisProt database [14], the primary repository of the intrinsic disorder, includes 1,652 interacting IDRs, which constitute 42% of the IDRs annotated in this resource. Close to 90% (1,473 out of 1,652) of the interacting IDRs bind to proteins and nucleic acids. However, DisProt altogether covers only about 1700 proteins, while millions of protein sequences await annotation of the interacting IDRs.

Computational predictors of interacting IDRs assist with closing this huge and growing annotation gap [15]. Based on an extensive literature search [15-19] we identified 22 predictors of the interacting IDRs. Nearly all of them (19 out of 22) predict a subfamily of the protein-binding IDRs called molecular recognition features (MoRFs) [20]. MoRFs are short IDRs that undergo folding upon interaction with protein partners. Some of popular MoRF predictors include MoRFpred [21, 22], fMoRFpred [20], DISOPRED3 [23], MoRFCHiBi [24], MoRFCHiBiLight [25], OPAL+ (2018) [26] and SPOT-MoRF [27]. The other three methods, ANCHOR [28], DisoRDPbind [29, 30], and ANCHOR2 [31] predict a broad family of the protein-binding IDRs that encompasses MoRFs. Moreover, DisoRDPbind is the only current tool that predicts IDRs that interact with DNA and RNA. These tools are frequently used to guide experimental studies and reveal novel functional insights. Just as an example, DisoRDPbind was recently used to study the SARS-CoV-2 proteome [32], decipher functions of genes from animal pathogens [33], and investigate specific proteins, such as CS-like zinc finger (FLZ) [34], spindle-defective protein 2 (SPD-2) [35], Mixed Lineage Leukemia 4 (MLL4) [36], and heat shock factor 1 (Hsf1) [37], some of which are associated with cancers and neurodegenerative diseases. The importance of these predictors is further underscored by the fact that CAID (Critical Assessment of protein Intrinsic Disorder) experiment, which is an equivalent of CASP (Critical Assessment of protein Structure Prediction) but for the disordered proteins, included assessment of methods that predict interacting (in a partner-agnostic way) IDRs [38]. The top performing tools in the recent CAID were ANCHOR2, DisoRDPbind and MoRFCHiBiLight, but the organizers also noted that "*substantial room for improvement remains*" [38], suggesting the need to develop more accurate predictors of the interacting IDRs.

The methods that offer the most relevant and accurate predictions of the interacting IDRs, ANCHOR2 and DisoRDPbind, rely on relatively simple predictive models. DisoRDPbind utilizes logistic regression while ANCHOR2 uses biophysics-based scoring functions. Moreover, DisoRDPbind that predicts interactions with proteins, DNA and RNA applies three independent/concurrent regressors. This way, it misses the opportunity to model relations between the three types of interactions. For instance, residues that bind nucleic acids and proteins have higher relative solvent accessibility compared to the non-binding residues while the nucleic acids binding residues are often positively charged and more evolutionarily conserved than the protein binding residues [39]. The fact that DisoRDPbind is the only tool that predicts nucleic acid binding IDRs combined with modest accuracy of the current predictors of interacting IDRs motivate development of more accurate solutions.

Furthermore, we note that some protein and nucleic interacting residues are located in the structured protein regions. Numerous methods target prediction of the structured interacting regions and they rely on the training data extracted from Protein Data Bank [39-45]. Recently published structure-trained tools include SPRINT [46], SSWRF [47], EL-SMURF [48] and SCRIBER [49] that predict protein-binding residues; RNABindRPlus [50] and FastRNABindR [51] that predict RNA-binding residues; TargetDNA [52] and DNAPred [53] that predict DNA-binding residues; DRNApred [54], NCBRPred [55] and BindN+ [56] that predict interactions with RNA and with DNA; and ProNA2020 [57] and MTDsites [58] that identify protein, DNA and RNA interacting regions. Interestingly, recent study reveals that the structure-trained predictors of protein binding regions perform poorly when used to predict protein-binding IDRs [59]. We further investigate this finding by evaluating results produced by several recent and well-performing structure-trained predictors of the protein, DNA and RNA interacting residues on the corresponding disordered binding regions.

We introduce DeepDISOBind, a custom-designed multi-task deep neural network that accurately predicts DNA, RNA and protein-binding IDRs. Multi-task learning aims to improve predictive performance by using shared representations (i.e., common parts of the model) to predict related learning tasks (i.e., binding to different partners) [60, 61]. Recently, the multi-task models were shown to improve predictive quality for bioinformatics problems including prediction of cleavage sites [62] and inter-residue distances [63], when compared to the single-task models. We devise the multi-task architecture where subsequent layers progressively specialize to predict interactions with different partner types. We empirically compare this topology against a single-task implementation and a representative selection of the existing predictors. We compare DeepDISOBind against representative methods that predicts protein and nucleic acid binding IDRs as well as the structure-trained methods. We also assess the DeepDISOBind's predictions on the human proteome and release our tool as a convenient web-server.

2 METHODS

2.1 Datasets

We source the data for training and comparative assessment of our predictive model from DisProt [14]. DisProt annotates proteins with the experimentally validated IDRs, including IDRs that interact with proteins, DNA and RNA. We manually checked IDRs that were annotated in DisProt as nucleic acids, DNA and RNA binding using the underlying publication data listed in DisProt in order to classify them as DNA and/or RNA binding. This annotation work follows from parsing DisProt for a recent comparative survey [64]. We divide these proteins into three subsets that constitute training, validation and test datasets. We ensure that sequences in each dataset share low (<30%) similarity with the other datasets. We use training and validation datasets to design and optimize the predictive model and the set-aside (during design and optimization) test dataset to comparatively assess this model against other solutions. Using protocol from [64], we cluster the original set of proteins with CD-HIT [65] at 30% sequence similarity and we place the entire protein clusters into training, validation and test datasets. The test and combined training/validation datasets share similar size while the training dataset is set to be twice the size of the validation dataset. This procedure adheres to commonly used practice in this field [64] and ensures proper level of separation between the training/validation and test datasets (<30% sequence similarity). Detailed statistics, which cover distribution of RNA/DNA/protein binding residues in the three datasets, are shown in Table 1. The datasets, including annotations of the DNA, RNA and protein interacting IDRs, are freely available at https://www.csuligroup.com/DeepDISOBind/. We note that these datasets are larger than the datasets used to train and test DisoRDPbind [29] and on par with the size of datasets utilized in CAID [38].

Table 1. Summary of datasets.

Dataset	Number of proteins	Number of disordered residues				Number of all
		Protein-binding	DNA-binding	RNA-binding	All disordered	residues
Training	238	15,341 (14.5%)	2,913 (2.7%)	1,437 (1.4%)	27,304 (25.9%)	105,601
Validation	118	6,464 (14.7%)	1,284 (2.9%)	608 (1.4%)	11,716 (26.8%)	43,776
Test	394	17,540 (8.4%)	2,377 (1.1%)	1,518 (0.7%)	46,041 (22.2%)	207,743

2.2 Evaluation criteria

DeepDISOBind and other related tools produce putative propensities for the disordered DNA, RNA and protein binding interactions for each residue in the input protein sequences. These real-valued propensities are accompanied by binary predictions, i.e., residues are classified as either DNA/RNA/protein-interacting or non-DNA/RNA/protein-interacting. The binary predictions are derived from the propensities by thresholding, i.e., residues with propensities > threshold are assumed to interact while the remaining residues are assumed not to interact. Following related works [29, 59], we calibrate the thresholds for all considered predictors such that their binary predictions produce to the same specificity = 0.8. Specificity is the rate of predictions of the interacting residues among the native non-interacting residues. We select 0.8 since it approximates the combined rate of the interacting residues across the three partner types. This calibration facilitates direct comparison of the binary predictions across different methods. Moreover, Table 1 reveals that the rates of the DNA and RNA interacting residues are much smaller than the rates of the protein interacting residues. Thus, we further calibrate the evaluation between the three partner types by randomly undersampling the non-binding residues when evaluating performance for the RNA and DNA interactions, so that their rate is the same as for the protein interactions. We assess the binary predictions with two popular metrics: F1 = (2*TP)/(2*TP+FN+FP) and sensitivity = TP/(TP+FN), where TP is the number of correctly predicted protein/RNA/DNA interacting residues, TN is the number of correctly identified non-protein/RNA/DNA-interacting residues, FN is the number of protein/RNA/DNA-interacting residues incorrectly predicted as non-interacting, and FP is the number of the non-interacting residues incorrectly predicted as protein/RNA/DNA-interacting. We assess the predicted propensities with a commonly used AUC (area under the receiver operating characteristics (ROC) curve) that plots sensitivity against FPR = FP/ (FP+TN). Higher values of the three metrics (F1, sensitivity and AUC) indicate better predictive quality. In addition, since some residues interact with more than one partner, we evaluate predictors that provide protein-, DNA-, and RNA-binding predictions with the macro-average and micro-average metrics that are used in related multi-label predictions studies [66-68]:

$$Micro - Sensitivity = \frac{TP_{avg}}{TP_{avg}+FN_{avg}},\ Micro - F1 = \frac{2*TP_{avg}}{2*TP_{avg}+FP_{avg}+FN_{avg}},$$

$$Macro - Sensitivity = \frac{1}{N}\sum\frac{TP_i}{TP_i+FN_i},\ Macro - F1 = \frac{2}{N}\frac{\sum\frac{TP_i}{TP_i+FN_i}*\sum\frac{TP_i}{TP_i+FP_i}}{\sum\frac{TP_i}{TP_i+FN_i}+\sum\frac{TP_i}{TP_i+FP_i}}$$

where TP_{avg} is the average number of correctly identified protein-, DNA-, and RNA-interacting residues, FN_{avg} is the average number of protein/RNA/DNA-interacting residues incorrectly predicted as non-interacting, FP_{avg} is the average number of the non-interacting residues incorrectly identified as protein/DNA/RNA-interacting, TP_i is the number of correctly predicted protein, DNA or RNA binding residues, FN_i is the number of protein/RNA/DNA-interacting residues incorrectly predicted as non-interacting, and FP_i is the number of incorrectly identified as protein/DNA/RNA interactions, and *i* represents RNA, DNA and protein interaction labels.

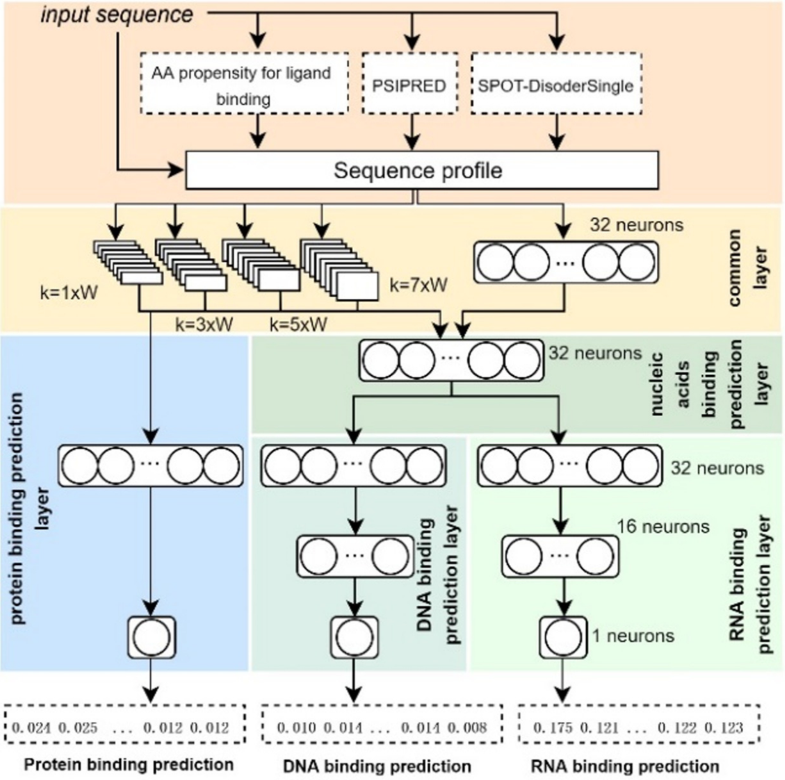

Figure 1. The multi-task topology of the DeepDISOBind predictor.

2.3 The DeepDIOSBind predictor

DeepDISOBind is a multi-task deep neural network that concomitantly predicts IDRs that interact with proteins, DNA and RNA (Figure 1). We use a custom-defined sequence profile that is extracted directly from the protein sequence as the input. Subsequent layers of the DeepDISOBind's network progressively specialize to predict interactions with different partner types. Correspondingly, the network is composed of five major elements (Fig. 1): the common layer, the nucleic acid binding layer, the protein binding layer, the DNA binding layer, and the RNA binding layer. Following, we provide a more detailed description of the sequence profile and network topology.

Sequence profile. Inspired by other recent models in this area [23, 27, 29, 69], the input protein sequence is first converted into a multi-dimensional profile. The profile covers the sequence itself together with relevant sequence-derived structural and functional properties that include relative amino acid propensities (RAAP) for ligand binding and predicted secondary structure and disorder. We use the one-hot encoding to represent the sequence. More specifically, each amino acid in the input sequence is represented by the 20-dimensional vector where the position of the corresponding amino acid type is set to 1 while the other positions are set to 0. Moreover, we compute the maximum, minimum, and average of the sequence embedding vectors that are defined in [70]. Inspired by recent studies that introduce novel predictors of the protein binding residues from structured/ordered proteins [49, 71], we use RAAP for ligand binding. These scores are derived empirically from binding data and quantify propensities of each amino acid type to bind a specific type of ligand. We use the five RAAP scales for the protein and nucleic acid binding that were introduced in Table 3 in ref. [39]. Finally, we use popular and fast predictors of the secondary structure, the single-sequence version of PSIPRED [72], and of the intrinsic disorder, SPOT-Disorder Single [73]. PSIPRED generates the 3-state secondary structures (helix, strand and coil), which we represent with the one-hot encoding. SPOT-Disorder Single produces real-valued propensities and binary predictions of disorder. Altogether, the profile includes 33 dimensions: 20 for one-hot encoding of sequence + 3 sequence embedding values + 5 RAAP values + 3 secondary structure predictions + 2 disorder predictions. Similar to the other solutions in this area [27, 29, 69, 73-75], we use sliding windows to predict the interaction propensity for the residues in the middle of the window. We pad the windows at the sequence termini with zeros.

Architecture of the DeepDISOBind network. The underlying idea is to initially model a generic set of interacting residues and progressively specialize the network to more specific interacting partners. To this end, the partner-agnostic common layer (yellow block in Fig. 1) links to layers that discriminate protein and nucleic acids binding (blue and green blocks in Fig. 1), while the latter layer further connects to layers that distinguish between DNA and RNA interactions.

The first, common layer consists of convolutional neural network (CNN) and feed-forward neural network (FNN) modules. The CNN module is composed of four different kernels that differ in size (k = 1, 3, 5 and 7). The variable kernel size designs were shown to be effective to reproduce the sequential nature of the protein sequences by accommodating for varying sizes of the residue neighborhoods, leading to improvements in predictive performance when compared to more traditional network architectures [70, 76-78]. We use 8 channels for each kernel that are followed by ReLU activation units and a 1D max-pooling layer. We utilize the 1D max-pooling layer to reduce the dimension of the latent feature spaces before they are passed to the subsequent layers. Since the CNN module focuses specifically on local information (in a small sequence neighborhood around the predicted residue), we supplement it with the FNN module that extracts information from a larger window. This module uses a layer of n = 32 ReLU activation units that work in parallel to the CNN module. The outputs of the CNN and FNN modules are combined and fed into the subsequent FNN layers that aim to specialize the latent feature space produced in the common layer to specific types of interactions. We use four of these layers. First, the common layer is linked to the protein binding and the nucleic acid binding layers. Next, the nucleic acid binding layer is linked to the DNA-binding and RNA-binding layers. We fix the sizes of the protein, DNA and RNA layers to n = 32 units, and we add additional sub-layers (smaller by a factor of 2) into the DNA and RNA layers. Consequently, RNA and DNA elements consist of two fully connected sub-layers, n = 32 and n/2 = 16 units. The latter is motivated by the fact that DNA and RNA interactions are harder to differentiate compared to the nucleic acids and protein interactions [39]. Finally, the output layer that generates putative propensities for disordered RNA, DNA and protein interactions consists of 3 neurons implemented with the sigmoid transfer function.

Learning of the multi-task network requires a more specialized strategy compared to classical single-task networks. This is because some of the tasks (interactions) could be easier to optimize compared to the other tasks. This can be solved by relative weighting between tasks. We use a recently proposed tuning that relies on estimating uncertainty of each task [79]. Under this approach, if the performance of two tasks improves and the reduction of the other task gets worse by no more than ε (we set ε to a small value of 0.1), then we continue training the model. Otherwise, we stop the training process. Moreover, we adopt early stopping approach to avoid overfitting the training dataset.

We empirically investigate the impact of the selection of the hyperparameter n (size of the FNN modules in the common, protein, nucleic acids, DNA and RNA layers) on the predictive performance. We consider networks with n = 16 (small size), n = 32 (medium), n = 64 (large) and n = 256 (very large). We summarize the corresponding topologies in Supplementary Table S1. We also empirically compare learning of the complete networks with the dropout learning [80] across the different network sizes. We set the dropout rate to 0.2. The dropout is meant to prevent overfitting, which would be apparent if the dropout-based learned networks would provide superior results. We compare the results on the validation dataset across different network sizes and when learning with and without the dropout on the training dataset in Supplementary Table S2. The average (across the three interaction types and three training runs) AUC ranges between 0.759 (small network with dropout) and 0.791 (medium network without dropout). Similarly, the average F1 varies between 0.238 (small network with dropout) and 0.271 (medium network without dropout). We observe that the averaged AUC and F1 scores are highly correlated (Pearson correlation of 0.95), which means that the considered networks produce high-quality propensities that are used to generate similarly accurate binary predictions. The medium size networks produce slightly better results than the small and large networks. Further increasing the size to the very large does not improve over the large-size networks. This means that the medium size networks are sufficiently large for this prediction. Lastly, we find that use of dropout does not lead to improvements. This together with the observation that modest-sized network produces the best results and outperforms the very large network suggest that our design does not overfit the training dataset. Consequently, we implement DeepDISOBind based on the medium network size (n = 32) and using training without dropout.

We also compare the above architecture that combines CNN and FNN modules with a design that relies on the graph neural network (GNN). GNNs were recently used in related projects that target prediction of protein-protein interactions at the protein level [81] and protein-protein interactions at the residue level from protein structure [82]. The corresponding underlying graphs represents the protein-protein interaction networks and the spatial arrangement of amino acids in the protein structures. We use the graph to represent our input protein sequence, and more specifically the sequential nature of connection between the residues in the input sliding window. The architecture of the GNN model draws from the best-performing medium size CNN/FNN network (i.e., DeepDISOBind) where we replace the CNN-based common layer with two graph convolutional layers, where nodes correspond to amino acids linked by peptide bonds, and we retain the other layers. Table S2 compares the results produced by this GNN model with the DeepDISOBind. The average AUC and F1 of the GNN-based design are modestly lower than the results produced by the CNN-based DeepDISOBind; AUC of 0.756 vs. 0.791 and F1 of 0.234 vs. 0.271. This could be explained by the fact that the underlying graph is rather simple as it can only represent corrections between residues in the protein sequence, compared to the CNN architecture that models these sequential relations more effectively. The more successful application of GNNs for the above-mentioned prediction of protein-protein interaction networks and protein-protein interactions from protein structure stems from a more informative structure of the corresponding graphs.

Table 2. Ablation analysis for the DeepDISOBind predictor on the test dataset. We compare the complete DeepDISOBind model against 10 versions where we remove specific parts of the sequence profile (v1 to v7) and where we implement the model as the combination of three single-task networks (versions v8, v9 and v10). Supplementary Tables S3 and S4 define further details. The profile includes amino acid sequence (AAS), relative amino acid propensity for binding (RAAP), putative secondary structure (PSS), and putative intrinsic disorder (PID). Sensitivity and F1 are calibrated to the same specificity = 0.8. The last set of columns shown in bold font shows the average values over the three types of the partner molecules.

Ablation design	Model	Protein interactions			RNA interactions			DNA interactions			**Average**		
		AUC	Sensitivity	F1	AUC	Sensitivity	F1	AUC	Sensitivity	F1	**AUC**	**Sensitivity**	**F1**
Exclusion of inputs from the profile	DeepDISOBind	0.77	0.60	0.31	0.75	0.61	0.32	0.74	0.47	0.26	**0.75**	**0.56**	**0.30**
	v1 (excludes AAS)	0.75	0.55	0.29	0.74	0.52	0.28	0.70	0.44	0.24	**0.73**	**0.50**	**0.27**
	v2 (excludes PID)	0.74	0.53	0.28	0.69	0.43	0.24	0.72	0.46	0.25	**0.72**	**0.47**	**0.26**
	v3 (excludes RAAP)	0.77	0.55	0.29	0.67	0.46	0.25	0.73	0.40	0.22	**0.72**	**0.47**	**0.25**
	v4 (excludes AAS and RAAP)	0.76	0.56	0.30	0.68	0.40	0.22	0.70	0.46	0.25	**0.71**	**0.47**	**0.26**
	v5 (excludes PSS and PID)	0.72	0.53	0.28	0.68	0.45	0.25	0.70	0.42	0.23	**0.70**	**0.47**	**0.25**
	v6 (excludes RAAP, PSS and PID)	0.71	0.45	0.25	0.67	0.38	0.21	0.72	0.43	0.24	**0.70**	**0.42**	**0.23**
	v7 (excludes AAS, PSS and PID)	0.69	0.47	0.25	0.72	0.51	0.28	0.68	0.48	0.26	**0.70**	**0.49**	**0.26**
Single-task prediction	v8 (single-task prediction of protein-binding)	0.75	0.51	0.27	N/A	N/A	N/A	N/A	N/A	N/A	**0.72**	**0.46**	**0.25**
	v9 (single-task prediction of RNA-binding)	N/A	N/A	N/A	0.72	0.44	0.24	N/A	N/A	N/A			
	v10 (single-task prediction of DNA-binding)	N/A	N/A	N/A	N/A	N/A	N/A	0.70	0.44	0.24			

Table 3. Comparative assessment on the test dataset. The binary predictions use thresholds that equalize specificity to 0.8 across the methods to allow for direct comparisons (details in Section 2.2). + means that DeepDISOBind is statistically significantly better (*p*-value<0.05). = means that the difference between DeepDISOBind and another predictor is not significant (*p*-value≥0.05). The best results for each column are shown in bold font.

Predictive target	Method	Protein binding			RNA binding			DNA binding			Multi-label macro average		Multi-label micro average	
		AUC	Sensitivity	F1	AUC	Sensitivity	F1	AUC	Sensitivity	F1	Sensitivity	F1	Sensitivity	F1
Protein, DNA and RNA binding residues	DeepDISOBind	**0.771**	**0.595**	**0.313**	**0.746**	**0.611**	**0.320**	**0.736**	**0.472**	**0.255**	**0.559**	**0.297**	**0.580**	**0.305**
	Single-task predictor (combination of v8, v9, v10)	0.746+	0.516+	0.277+	0.725+	0.446+	0.243+	0.697+	0.443+	0.242+	0.468+	0.254+	0.503+	0.271+
	DisoRDPbind	0.727+	0.456+	0.249+	0.594+	0.364+	0.202+	0.671+	0.452+	0.246+	0.426+	0.234+	0.457+	0.248+
	MTDsites	0.576+	0.304+	0.173+	0.677+	0.479+	0.258+	0.675+	0.253+	0.242+	0.406+	0.225+	0.322+	0.182+
	ProNA2020	0.398+	0.205+	0.120+	0.468+	0.193+	0.08+	0.551+	0.441+	0.187+	0.215+	0.132+	0.204+	0.120+
Protein binding residues	ANCHOR2	0.719+	0.501+	0.270+										
	MoRFChibiLight	0.735+	0.502+	0.271+										
	SCRIBER	0.684+	0.423+	0.232+										
DNA and RNA binding residues	BindN+				0.685+	0.473+	0.257+	0.615+	0.331+	0.187+				
	NCBRPred				0.662+	0.455+	0.243+	0.617+	0.367+	0.205+				
DNA binding residues	TargetDNA							0.580+	0.274+	0.157+				
RNA binding residues	RNABindRPlus				0.576+	0.336+	0.186+							

3 RESULTS

3.1 Ablation analysis of the network design

DeepDISOBind relies on two major elements: the multi-element sequence profile and the multi-task architecture. We investigate the relation between the specific formulation of these elements and the resulting predictive performance. We run ablation analysis where we measure predictive performance when removing certain parts of the profile and when we implement the topology as the collection of three single-task networks. The corresponding 10 versions of the predictive model are defined in Supplementary Tables S3 (modifications of the sequence profile) and S4 (modifications of the topology).

We summarize the results of the ablation analysis on the test dataset in Table 2. The top portion of the Table 2 focuses on the sequence profile and reveals that all major parts of this profile that we employ provide useful information for the predictive model. More specifically, removal of the sequence, putative disorder or binding propensities (versions v1, v2 and v3) leads to a substantial drop in predictive performance from 0.75 to between 0.72 and 0.73 in the average AUC and from 0.56 to between 0.47 and 0.50 in the average sensitivity; we average over the three partner types. Removal of two or more parts of the profile (versions v4, v5, v6 and v7) further deteriorates the performance, with the average AUC dropping to between 0.70 and 0.71. Interestingly, the v7 model that relies solely on the amino acid level propensities for binding (5-dimensional RAAP input) is comparable to the v6 model that uses the protein sequence (23-dimensional AAS input), where both models secure the average AUC of 0.7. This shows that the RAAP scores provide a high-quality reduced representation of the sequence for the purpose of the prediction of the protein and nucleic acids interactions. Supplementary Figure S1A provides the corresponding ROC curves. The curves demonstrate that DeepDISOBind offers particularly strong improvements over the models that exclude certain types of inputs for the low values of FPR (false positive rate) < 0.3 (in Supplementary Figure S1B). The increase in the sensitivity at the same FPR can be as high as 7% when compared to the best input-reduced version. We argue that predictions with the low FPRs are more practical than the predictions with higher FPRs, given our imbalanced dataset where only about 20% of residues are interacting. In other words, FPRs of over 0.3 would correspond to substantial overprediction of interactions. Altogether, these results indicate that all elements of the sequence profile contribute to the quality of predictions produced by the DeepDISOBind model.

We also study benefits of the application of the multi-task architecture by comparing it with the implementation that combines three single-task networks that use corresponding subsets of the layers from the original network and the same complete sequence profile (Supplementary Table S4). We summarize these results in the bottom section of Table 2 (versions v8, v9 and v10). Each of the three single-task models underperforms when compared with DeepDISOBind. More specifically, the AUC for the prediction of the disordered protein interactions drops from 0.77 (DeepDISOBind) to 0.75 (single-task deep network), for the RNA interactions drops from 0.75 to 0.72 and for the DNA interactions decreases from 0.74 to 0.70. Moreover, average (over the three types of interactions) F1 and sensitivity (measured as the same specificity = 0.8) are reduced from 0.30 and 0.56 to 0.25 and 0.46, respectively, when comparing the multi-task and the single-task networks. This suggests that the use of the multi-task design leads to substantial improvements in the predictive performance across the three types of the interactions. This conclusion is in agreement with literature that similarly demonstrates that the multi-task learning improves over the single-task learning in a generic machine learning setting [61, 83], as well as when applied to bioinformatics problems [62, 63, 84]. We note that the multi-task learning was not previously used for the prediction of the disordered protein-protein and protein-nucleic acids interactions.

3.2 Comparative assessment of predictive performance between DeepDISOBind and related methods

We compare results produced by DeepDISOBind with other relevant and representative methods that predict protein, DNA and RNA interactions from protein sequences. These methods include the only other tool that predicts disordered protein, DNA and RNA interactions, DisoRDPbind [29], and two popular and accurate predictors of the disordered protein interactions, ANCHOR2 [31] and MoRFCHiBiLight [25]. These methods secured the top three spots in the assessment of the prediction of interacting IDRs in the recent CAID experiment [38]. We also include a comprehensive selection of the structure-trained predictors including SCRIBER [49], which predicts protein-binding residues and which was recently shown to outperform other structure-trained predictors of protein interacting residues [59]; RNABindRPlus [50] that was ranked as the best tool in the recent assessment of the structure-trained predictors of the RNA interactions [45]; TargetDNA [52], one of the most accurate and popular predictors of the DNA interactions in the structured regions [85]; two representative structure-trained methods that predict DNA and RNA binding regions, popular BindN+ [56] that was shown to provide strong predictive performance in comparative surveys [39, 43] and one of the most recent methods, NCBRPred [55]; and two structure-trained methods which target prediction of protein, DNA and RNA interactions, ProNA2020 that was released in 2020 [57] and MTDsites that was published in 2021 [58]. The latter two methods offer the same scope of predictions as DeepDISOBind and DisoRDPbind, but they address predictions for structured rather than disordered regions. We use the author-provided webservers or implementations to make the predictions for these ten tools: DisoRDPbind, ANCHOR2, MoRFCHiBiLight, SCRIBER, RNABindRPlus, TargetDNA, BindN+, NCBRPred, ProNA2020 and MTDsites.

We compare results produced by DeepDISOBind with the ten representative tools and our implementation that is based on the single-task networks on the test dataset in Table 3. We empirically assess whether DeepDISOBind offers statistically significant improvements over the other solutions that are robust across different datasets. We bootstrap 50% of the test proteins 50 times, and compare the corresponding results with the *t*-test (for normal measurements) or with the Wilcoxon test (otherwise). We test normality with the Kolmogorov-Smirnov test at the *p*-value of 0.05. Similar tests were done in related comparative studies [45, 64, 86]. Table 3 reveals that DeepDISOBind consistently secures the best predictive performance across the three binding partner types and the three metrics of performance. Moreover, the improvements in AUC, sensitivity and F1 are statistically significant compared to each of the ten other methods for the predictions of protein, DNA and RNA interactions (*p*-value < 0.05).

The average AUC, sensitivity and F1 (computed over the three interactions) for DeepDISOBind are 0.75, 0.56 and 0.30, compared the other three tools that provide the same scope of predictions that covers protein, DNA and RNA interactions: DisoRDPbind (0.66, 0.42 and 0.23), ProNA2020 (0.47, 0.28 and 0.13) and MTDsites (0.64, 0.35 and 0.22). The corresponding ROC curves for these four predictors are separated by a relatively wide margin (Supplementary Figure S2). We also assess multi-label predictions for these four methods using the macro-average and micro-average metrics (Table 3). Consistent with the single-label assessment, DeepDISOBind outperforms the other three predictors by securing macro F1 of 0.30, macro sensitivity of 0.56, micro F1 of 0.30 and micro sensitivity of 0.58. These results are statistically better than the results of the three other methods (*p*-value < 0.05), with the second-best DisoRDPbind that obtains macro F1 of 0.23, macro sensitivity of 0.43, micro F1 of 0.25 and micro sensitivity of 0.46. The predictive performance of MTDsites and ProNA2020 is worse than DeepDISOBind and DisoRDPbind since the two former methods are trained using structured proteins. The lower predictive quality of these tools for the prediction of interactions in the IDR is in agreement with similar observations in a recent comparative survey of the disorder-trained and structure-trained predictors of protein binding residues [59].

For the disordered protein interactions prediction, the sensitivity of DeepDISOBind is better by (0.595-0.456)/0.456 = 30.5%, 190.2%, 95.7%, 40.5%, 18.8%, and 18.5% when compared with DisoRDPbind, ProNA2020, MTDsites, SCRIBER, ANCHOR2, and MoRFChibiLight, respectively. This means that DisoRDPbind correctly identifies at least 18.5 % more interacting residues at the same false positive rate, i.e., we fix specificity at 0.8 for all methods, which corresponds to 0.2 false positive rate. Similarly, for the RNA interactions, DeepDISOBind's sensitivity is better by 67.9%, 216.5%, 27.5%, 34.3%, 29.2% and 81.8% when contrasted with DisoRDPbind, ProNA2020, MTDsites, NCBRPred, BindN+ and RNABindRPlus, respectively. The improvements in the sensitivity for the DNA interaction predictions are at 4.4%, 7.0%, 86.5%, 28.6%, 42.6% and 72.3% when compared against DisoRDPbind, ProNA2020, MTDsites, NCBRPred, BindN+ and TargetDNA, respectively. Similar observations are true when using the F1 and AUC metrics.

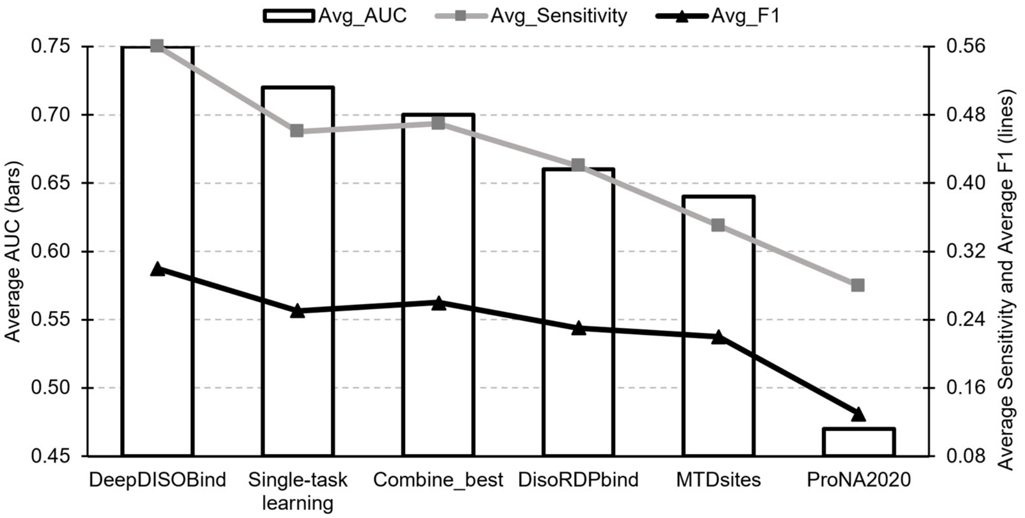

Figure 2. Comparison of the predictive performance on the test dataset between DeepDISOBind, MTDsites, ProNA2020, the single-task network (combination of the v8, v9 and v10 networks), the Combine_best approach which uses the best method for each interaction type selected across the six predictors (i.e., MoRFChibiLight for the protein interactions, BindN+ for the RNA interactions, and MTDsites for the DNA interactions), and DisoRDPbind. We quantify the predictive performance with the average (over the three interaction types) values of AUC (bars and vertical axis on the left), F1 and sensitivity (lines and vertical axis on the right).

Figure 2 offers a more direct approach to compare DeepDISOBind with the state-of-the-art. We compare the average values of the AUC (bars), sensitivity (gray line) and F1 (black line) computed over the three interaction types. The comparison includes DeepDISOBind, the single-task network (combination of the v8, v9 and v10 networks), the "combine best" approach which uses the best method (i.e., having highest AUC) for each interaction type selected across the ten predictors (i.e., MoRFChibiLight for the protein interactions, BindN+ for the RNA interactions, and MTDsites for the DNA interactions), DisoRDPbind which is the only other disorder-trained predictor with the same scope as DeepDISOBind, and MTDsites and ProNA2020 which are the two recently published structure-trained methods that predict protein, DNA and RNA interacting residues. Firstly, we note a substantial and statistically significant (see Table 3) improvement when contrasting the multi-task (DeepDISOBind) vs. single-task solutions across the three metrics (p-value < 0.05). Secondly, DeepDISOBind improves against the combination of the best current methods by a large and statistically significant margin (0.75 vs. 0.70 in AUC, 0.56 vs. 0.41 in sensitivity, and 0.30 vs. 0.26 in F1). Thirdly, DeepDISOBind and the single-task networks outperform DisoRDPbind, primarily because the latter relies on simpler logistic regression models that are applied utilizing the single-task architecture. Lastly, DeepDISOBind improves over ProNA2020 and MTDsites because the latter two are trained on the structured proteins.

Finally, we investigate impact of similarity between the test proteins and the proteins that were used to train PSIPRED and SPOT-Disorder-Single methods, which we utilize to derive inputs for DeepDISOBind (Figure 1). We collect and combine the training datasets of these two predictors. Next, we align each test protein to every training protein with BLASTp [87] to annotate regions in the test proteins that share similarity>30%. Finally, we retest the predictive performance of DeepDISOBind and the other predictors of protein, DNA and RNA binding residues on the test proteins when excluding the similar regions. We summarize these results in Supplementary Table S5. DeepDISOBind secures results that are on average very similar to the results on the complete test dataset, with the average AUC (over the protein, DNA and RNA predictions) of 0.752 vs. 0.751 and the average F1 of 0.295 vs. 0.296. Moreover, DeepDISOBind's predictions consistently maintain statistically significant advantage over the results of the other ten predictors (p-value<0.05). Altogether, the results on the complete test dataset and the sequence regions that share low similarity to the training data of PSIPRED and SPOT-Disorder-Single are similar. This could be explained by the fact that we use the single-sequence version of PSIPRED and the inherently single-sequence SPOT-Disorder-Single. Both methods do not use sequence alignment, thus minimizing the likelihood of over-fitting training datasets [72, 73]. To sum up, the empirical analysis demonstrates that DeepDISOBind provides accurate predictions of the disordered protein, DNA and RNA-interactions.

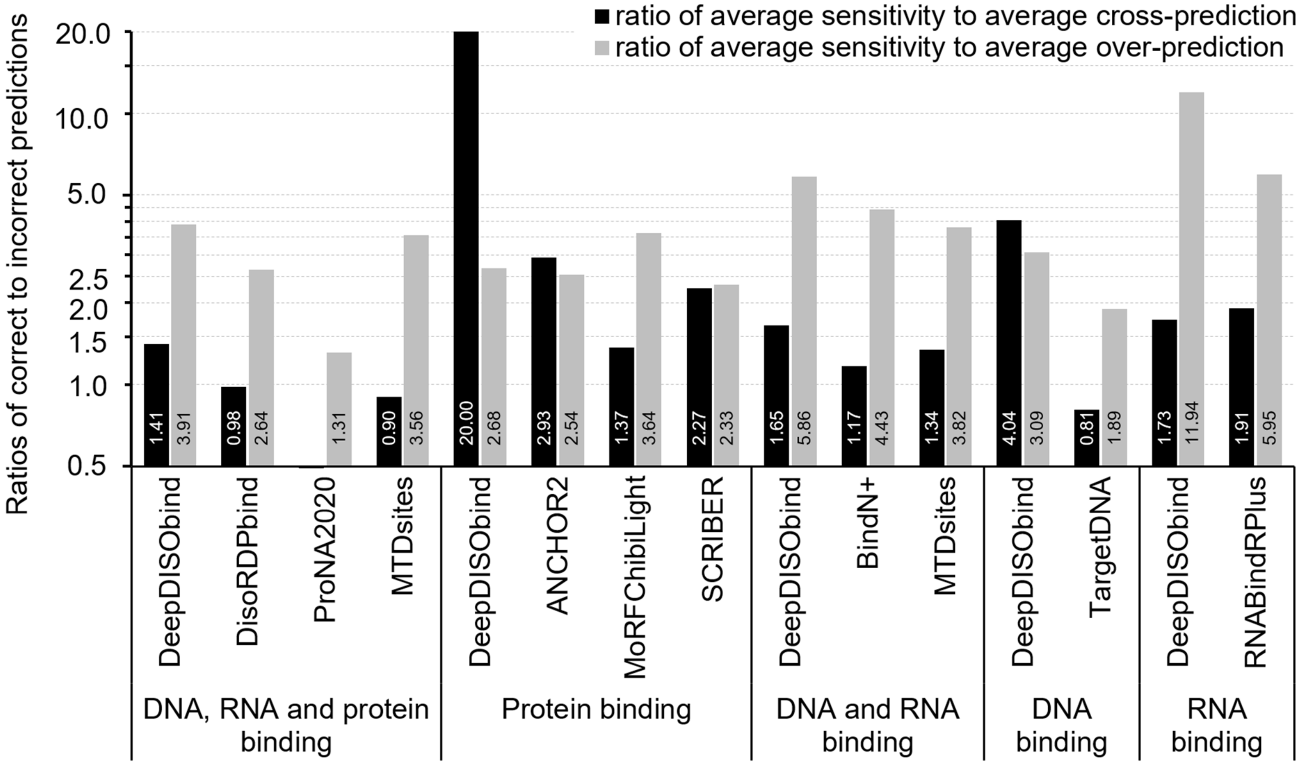

Figure 3. Comparison of the ratios of the average sensitivity (over the three interaction types) to the average cross-prediction and over-prediction rates on the test dataset. Larger ratios indicate higher quality predictions. The predictions rely on thresholds that equalize the number of the predicted binding residues with the number of native binding residues for each predictor. Predictors are grouped by the scope of their predictions, as described on the *x*-axis, where DeepDISOBind's predictions are limited to the predictions of the other methods in the same group.

3.3 Assessment of cross-predictions and over-predictions

The binding residues share certain characteristics, such as high levels of evolutionary conservation and high solvent accessibility, irrespective of the type of their binding partners. This may lead to a substantial amount of cross-predictions, measured as the fraction of residues that bind a given partner type that are predicted to interact with another ligand type, e.g., protein binding residues predicted as DNA or RNA binding residues. Recent studies have found that majority of methods that predict interacting residues for the structured regions generate substantial amounts of cross-predictions, which in some cases are as high as their sensitivity that quantifies the rate of correct predictions [42, 43, 59, 88]. Correspondingly, we assess the cross predictions and over-predictions (fraction of non-binding residues predicted to interact with a given partner type) for DeepDISOBind and the other ten considered here predictors. Figure 3 quantifies the average (over the different partner types) ratios of sensitivity (rate of correct predictions) to the cross-prediction and over-prediction rates (rates of incorrect predictions) on the test dataset; ratios > 1 denote methods for which the rate of the correct predictions is higher than the rate of over- or cross-predictions. We normalize rate of predictions of binding residues across predictors to allow for side-by-side comparisons of the ratios across methods, i.e., we equalize the number of the predicted protein/DNA/RNA binding residues to the number of the native protein/DNA/RNA binding residues. We provide the complete set of results including cross-prediction rates, over-prediction rates and sensitivity values for each partner type (protein, DNA and RNA) in Supplementary Table S6. We compare DeepDISOBind to the other methods using the same set of predictions, e.g., we compare DeepDISOBind's predictions of proteins binding residues to the SCRIBER's, ANCHOR2's and MoRFChibiLight's results that also predict only the protein binding residues. The ratios to the over-predictions are relatively high across all methods, ranging between 1.89 for TargetDNA and 11.94 for DeepDISOBind's prediction of the RNA binding residues (gray bars in Fig. 3). This means that relatively few non-binding residues are predicted to bind. We also observe that DeepDISOBind generates the highest/best ratios to the cross-predictions across all scenarios, except when compared for the RNA binding prediction with RNABindRPlus where both methods achieve good results, 1.73 and 1.91 (black bars in Figure 3). Moreover, the DeepDISOBind's ratios are always > 1, which means that that its rates of correct prediction of binding residues outperform the rates of the cross-predictions. We observe that relatively few protein binding residues are incorrectly predicted as RNA binding (7%) or DNA binding (11%), compared to the corresponding average sensitivity (26%). Overall, when making predictions of the protein, DNA and RNA binding, the DeepDISOBind's ratio to cross-predictions equals 1.41. This means that its average rate of correct predictions is 40% higher than the rate of the cross-predictions, which is substantially better than the 0.98, 0.46, and 0.90 ratios secured by DisoRDPbind, ProNA2020, and MTDsites.

3.4 Assessment of predictions in the human proteome

We assess DeepDISOBind's predictions on the proteome scale. While the coverage of the residue/region-level annotations is limited at this scale, we can obtain a comprehensive set of experimental annotations at the protein level. We evaluate DeepDISOBind's predictions of the disordered DNA and RNA binding proteins in one of the most-comprehensively annotated proteomes, the human proteome. To do that, we collect disordered human proteins that are annotated to interact with DNA and with RNA, as well as the human proteins that are unlikely to interact with the nucleic acids. First, we collect the human proteome from UniProt version 2019_09 [89] and remove partial sequences that we identify based on the *Sequence status* term *Fragment*. This produces 43,789 protein sequences. Second, we annotate the DNA interacting proteins by combining data from a comprehensive collection of relevant databases including 3D-footprint [90], CIS-BP [91], JASPAR [92], HumanTF2 [93], SMiLE-seq [94], animalTFDB [95] and the gene ontology (GO) terms [96] in UniProt. We also annotate the RNA binding proteins based on the data from ATtRACT [97], RBPDB [98], and the GO terms in UniProt. We map proteins in these diverse resources into the human set based on the UniProt's accession numbers. This results in 2,379 DNA-binding and 2,371 RNA-binding proteins, which is in line with related studies [99]. We identify the disordered subset of these proteins using the popular VSL2B predictor [100]. This method is different than the SPOT-Disorder-Single predictor used in DeepDISOBind and offers high-quality predictions of the disordered proteins [38, 64]. We annotate a given DNA/RNA binding protein as disordered if its putative disorder content > 0.2. Consequently, we identify 1,739 and 1,711 disordered DNA and RNA interacting proteins, respectively. Third, we derive proteins that are unlikely to interact with the nucleic acids. We select the human proteins that share < 30% sequence similarity with the annotated DNA- and RNA-binding proteins, which we quantify with BLASTp [87, 101]. This results in the set of 24,435 proteins. Finally, we convert the residue/regions-level propensities produced by DeepDISOBind into protein-level propensities of the disordered RNA and DNA interactions. Since typically only a small portion of the amino acids interact with the nucleic acids, we compute average of the highest 5% of the residue-level propensities produced by DeepDISOBind for a given protein to quantify the protein-level propensities. We emphasize that this approach does not validate correctness of the positions of the predicted binding residues in the protein sequence (which we assess in Sections 3.2 and 3.3) but rather the ability to quantify propensity for binding at the whole protein level. We assess these protein-level predictions of the DNA and RNA interacting proteins in the human proteome with the ROC curves and the corresponding AUC scores (Figure 4). DeepDISOBind secures AUC of 0.72 and 0.82 for the prediction of the human RNA and DNA interacting proteins, respectively, which is consistent with the results on the test dataset.

Moreover, motivated by the discussion in Section 3.3, we evaluate the potential for cross-predictions of the protein-level scores. We group the considered human proteins into four sets: (1) proteins that bind DNA and do not bind RNA; (2) proteins that bind RNA and do not bind DNA; (3) proteins that bind both RNA and DNA; (4) proteins that do not bind neither DNA nor RNA. Next, we compare the protein-level scores for DNA and RNA interactions that we extract from DeepDISOBind's predictions (i.e., average of the highest 5% of the residue-level propensities) inside the sets 1, 2 and 3 to study the cross-prediction. We utilize the pairwise *t*-test (for normal measurements) or the Wilcoxon test (otherwise), where we test normality with the Kolmogorov-Smirnov test at the 0.05 *p*-value. The protein-level DNA binding propensities are higher than the protein-level RNA propensities within protein set 1 and the difference is statistically significant (*p*-value < 0.01). Similarly, the protein-level RNA binding propensities are significantly higher than the corresponding DNA propensities for the set 2 (*p*-value < 0.01). Interestingly, the protein-level RNA and DNA binding propensities are not significantly different for the protein set 3 (*p*-value = 0.66). These results suggest that the predictions of DeepDISOBind that we aggregate at the protein-level successfully differentiate between DNA and RNA binding proteins. Finally, we further examine the accuracy of the prediction of the DNA and RNA binding proteins by comparing the protein-level DNA binding propensities between sets 1 and 4, and the protein-level RNA binding propensities between sets 2 and 4. In both cases the protein-level propensities for DNA and RNA interactions are higher in the sets 1 and 2, respectively, when compared with the set 4 and these differences are statistically significant (p-value < 0.01).

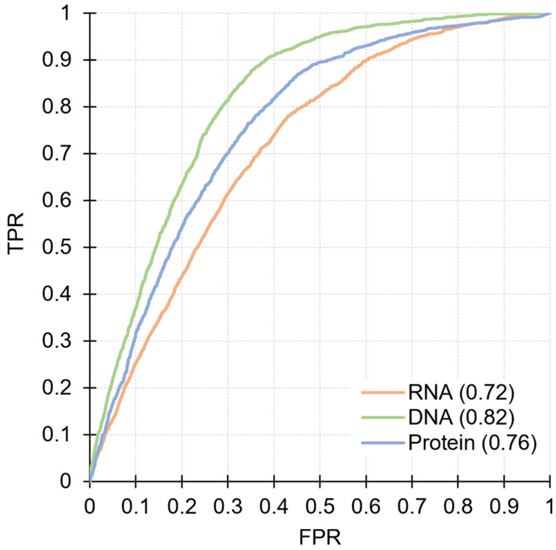

Figure 4. ROC curves for the DeepDISOBind's prediction of the DNA interacting proteins (green line) and the RNA interacting proteins (orange line) in the human proteome. The blue ROC curve is for the scenario where DeepDISOBind classifies disordered human hub proteins (proteins that interact with many proteins) vs. human proteins that interact with a few protein partners.

We also assess whether DeepDISOBind accurately predicts the disordered protein interactions. Since majority of human proteins interact with proteins and thus it would be virtually impossible to reliably identify non-protein-binding proteins, we use DeepDISOBind's predictions to differentiate between disordered hub proteins [102], which interact with many protein partners, and proteins that interact with relatively few proteins. This is motivated by the finding that the human hub proteins are enriched in the intrinsic disorder [103, 104]. First, we collected a comprehensive set of protein-protein interaction (PPI) annotations in the human proteome from the mentha resource, which combines data from several relevant source databases [105]. Second, we process the corresponding set of 17,598 protein-interacting proteins to extract the highly promiscuous hub proteins (25% of proteins with the highest PPI counts) and proteins that interact with a few protein partners (25% that interact with the smallest number of proteins). The same as for the assessment of the DNA/RNA interactions, we use VSL2B to identify a subset of the disordered hub proteins. Finally, we convert the residue/regions-level protein-binding propensities produced by DeepDISOBind into the protein-level propensities of the disordered proteins interactions using the same approach as for the assessment of the nucleic acids binding proteins. Blue ROC curve in Figure 4 quantifies the predictive quality of DeepDISOBind applied to differentiate between the disordered hubs and the proteins that interact with few proteins. DeepDISOBind obtains AUC of 0.76, which is similar to the results on the test dataset. Altogether, these results suggest that the outputs produced by DeepDISOBind can be converted into protein-level scores that correctly predict disordered RNA, DNA and protein-interacting proteins. The DeepDISOBind's predictions for the human proteins are available at https://www.csuligroup.com/DeepDISOBind/.

3.5 Case study

We illustrate DeepDISOBind's predictions on one of the test proteins, the silent information regulator Sir3p from budding yeast (DisProt: DP00533; UniProt: P06701). Sir3p is involved in the initiation, propagation and compaction of the silenced chromatin [106, 107]. Sir3p has a long IDR (positions 216-549) that interacts with RAP1p [108], RAD7p [109] and Sir4p coiled-coil domain [110]. This DNA and protein interacting IDR is flanked by structured regions that extend to the termini.

The case study aims to visualize the putative propensities and binary predictions produced by DeepDISOBind and the other methods that we cover in Section 3.2 and Table 3. This example is not intended to quantify or compare the predictive performance. Supplementary Figure S3A reveals that the disorder-trained predictors (ANCHOR2, MoRFCHiBiLight, DisoRDBbind and DeepDISOBind) correctly identify this IDR as interacting with proteins. MoRFCHiBiLight slightly overpredicts protein interacting regions in the structured domain at the N-terminus. On the other hand, the structure-trained predictors of the protein-binding residues, such as SCRIBER, ProNA2020, and MTDsites, miss this binding region. This can be explained by the fact that they target prediction of protein binding in structured regions. Supplementary Figure S3B shows that while most methods (except for ProNA2020) correctly predict DNA interactions in this IDR, DisoRDPbind, TargetDNA, BindN+, MTDsites, and NCBRPred overpredict DNA interactions outside this region. TargetDNA, BindN+, MTDsites, and NCBRPred were designed to identify interactions in the structured regions and this is likely why they make more predictions at both structured termini. Moreover, predictions from DeepDISOBind, DisoRDPbind, ProNA2020, and RNABindRPlus suggests that this protein is unlikely to interact with RNA while BindN+, NCBRPred, and MTDsites predict multiple RNA binding regions (Supplementary Figure S3C). This can be again attributed to the fact that BindN+, NCBRPred, and MTDsites aim to predict RNA interactions in the structured regions.

3.6 DeepDISOBind webserver

DeepDISOBind is available as a user-friendly webserver at https://www.csuligroup.com/DeepDISOBind/. With the user's convenience in mind, we make predictions on the server side and process up to 20 proteins in a single request. The only required input are the FASTA-formatted protein sequences. Users can opt to provide an email address where we send links to the results when predictions are completed. Predictions take about 30 seconds for an average-size sequence. The server outputs numeric propensities for the protein, RNA and DNA interactions and the three corresponding binary predictions for each residue in the input chain(s). We also provide putative propensities and binary annotations of disorder generated by SPOT-Disorder-Single. The results are available in three convenient formats: 1) parseable text file that can be downloaded from a request-specific URL; 2) color-coded (to ease identification of interacting residues) table in the browser window; and 3) an interactive graphical format in the browser window. We will store these predictions for at least one month. The graphical format allows users to select predictions of specific interactions, identify propensity scores, amino acid type and position on mouse hover, and zoom on a specific protein segment. Users should employ the putative propensities as a measure of confidence, i.e., residues predicted with higher values of propensity are more likely to interact with the corresponding partner. Moreover, the binary predictions can be used identify the putative protein, RNA and DNA binding residues when assuming low false positive rate at 0.2; we use the same calibration in Tables 2 and 3, and Supplementary Tables S5 and S6.

Importantly, DeepDISOBind targets prediction of IDRs that interact with proteins, DNA and RNA and, by design, is not going to produce reliable predictions for the structured regions. Thus, predictions of the interacting residues for the structured regions, which can be identified with the help of the SPOT-Disorder-Single's predictions, should be pursued with the structure-trained methods. Recent surveys of the structure-trained predictors can be used to identify suitable methods [39-45].

4 DISCUSSION

IDRs interact with a variety of partner molecules including nucleic acids and proteins. The availability of experimental data for hundreds of interacting IDRs gave rise to the development of machine learning models that learn from these data to predict these interactions for the millions of unannotated protein chains. However, only one such tool is available for the prediction of disordered interactions with the nucleic acids and the recent CAID experiment concludes that new and more accurate predictors of the interacting regions are needed [38]. To this end, we develop DeepDISOBind, a novel multi-task deep learner that provides accurate predictions of the DNA, RNA and protein binding IDRs. We empirically demonstrate that our selection of the predictive inputs and the multi-task design of DeepDISOBind's model contribute to its predictive performance. Side-by-side evaluation on an independent (low similarity) test dataset reveals that DeepDISOBind offers statistically significant improvements over the single-task topology and a representative collection of 10 existing tools that cover both disorder-trained and structure-trained methods. These improvements are consistent across the three interactions types. Evaluation on the human proteome shows that DeepDISOBind accurately identifies hubs and DNA- and RNA-binding proteins. We provide a convenient webserver at https://www.csuligroup.com/DeepDISOBind/. This webserver allows for batch predictions, performs calculations on the server side, and provides results in multiple formats, including an interactive graphical visualization.

KEY POINTS

- CAID experiment shows that current predictors of disordered regions interacting with nucleic acids and proteins offer modest levels of predictive accuracy
- DeepDISOBind uses an innovative deep multi-task architecture to accurately predict DNA, RNA and protein binding disordered regions from protein sequences
- DeepDISOBind's predictions outperform results of current disorder- and structure-trained methods across the interactions with DNA, RNA and protein partners
- DeepDISOBind accurately identifies protein hubs and DNA- and RNA-binding proteins in the human proteome
- DeepDISOBind’s webserver is available at https://www.csuligroup.com/DeepDISOBind/

FUNDING

The National Natural Science Foundation of China (61832019); The 111 Project (B18059); and the Hunan Provincial Science and Technology Program (2019CB1007); The Robert J. Mattauch Endowed Chair funds.

Conflict of interest statement. None declared.

REFERENCES

1. Lieutaud, P., et al., *How disordered is my protein and what is its disorder for? A guide through the "dark side" of the protein universe.* Intrinsically Disord Proteins, 2016. **4**(1): p. e1259708.
2. Oldfield, C.J., et al., *Introduction to intrinsically disordered proteins and regions*, in *Intrinsically Disordered Proteins*, N. Salvi, Editor. 2019, Academic Press. p. 1-34.
3. Xue, B., A.K. Dunker, and V.N. Uversky, *Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life.* J Biomol Struct Dyn, 2012. **30**(2): p. 137-49.
4. Peng, Z., et al., *Exceptionally abundant exceptions: comprehensive characterization of intrinsic disorder in all domains of life.* Cell Mol Life Sci, 2015. **72**(1): p. 137-51.
5. Dunker, A.K., et al., *Function and structure of inherently disordered proteins.* Curr Opin Struct Biol, 2008. **18**(6): p. 756-64.
6. Xie, H., et al., *Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions.* J. Proteome Res., 2007. **6**(5): p. 1882-98.
7. Wang, C., V.N. Uversky, and L. Kurgan, *Disordered nucleiome: Abundance of intrinsic disorder in the DNA- and RNA-binding proteins in 1121 species from Eukaryota, Bacteria and Archaea.* Proteomics, 2016. **16**(10): p. 1486-98.
8. Meng, F., et al., *Compartmentalization and Functionality of Nuclear Disorder: Intrinsic Disorder and Protein-Protein Interactions in Intra-Nuclear Compartments.* Int J Mol Sci, 2016. **17**(1).
9. Wu, Z., et al., *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**(19 Pt A): p. 2561-9.
10. Varadi, M., et al., *Functional Advantages of Conserved Intrinsic Disorder in RNA-Binding Proteins.* PLoS One, 2015. **10**(10): p. e0139731.
11. Dyson, H.J., *Roles of intrinsic disorder in protein-nucleic acid interactions.* Mol Biosyst, 2012. **8**(1): p. 97-104.
12. Vacic, V., et al., *Characterization of molecular recognition features, MoRFs, and their binding partners.* J Proteome Res, 2007. **6**(6): p. 2351-66.
13. Peng, Z., et al., *A creature with a hundred waggly tails: intrinsically disordered proteins in the ribosome.* Cell Mol Life Sci, 2014. **71**(8): p. 1477-504.
14. Hatos, A., et al., *DisProt: intrinsic protein disorder annotation in 2020.* Nucleic Acids Res, 2020. **48**(D1): p. D269-D276.
15. Katuwawala, A., S. Ghadermarzi, and L. Kurgan, *Computational prediction of functions of intrinsically disordered regions*, in *Progress in Molecular Biology and Translational Science*, V.N. Uversky, Editor. 2019, Academic Press. p. 341-369.
16. Meng, F., V.N. Uversky, and L. Kurgan, *Comprehensive review of methods for prediction of intrinsic disorder and its molecular functions.* Cell Mol Life Sci, 2017. **74**(17): p. 3069-3090.
17. Varadi, M., et al., *Computational approaches for inferring the functions of intrinsically disordered proteins.* Front Mol Biosci, 2015. **2**: p. 45.

18. Katuwawala, A., et al., *Computational Prediction of MoRFs, Short Disorder-to-order Transitioning Protein Binding Regions.* Comput Struct Biotechnol J, 2019. **17**: p. 454-462.
19. Barik, A. and L. Kurgan, *A comprehensive overview of sequence-based protein-binding residue predictions for structured and disordered regions*, in *Protein Interactions*. 2020. p. 33-58.
20. Yan, J., et al., *Molecular recognition features (MoRFs) in three domains of life.* Mol Biosyst, 2016. **12**(3): p. 697-710.
21. Oldfield, C.J., V.N. Uversky, and L. Kurgan, *Predicting functions of disordered proteins with MoRFpred*, in *Computational Methods in Protein Evolution*. 2019, Springer. p. 337-352.
22. Disfani, F.M., 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**(12): p. i75-i83.
23. Jones, D.T. and D. Cozzetto, *DISOPRED3: precise disordered region predictions with annotated protein-binding activity.* Bioinformatics, 2015. **31**(6): p. 857-63.
24. Malhis, N. and J. Gsponer, *Computational identification of MoRFs in protein sequences.* Bioinformatics, 2015. **31**(11): p. 1738-1744.
25. Malhis, N., M. Jacobson, and J. Gsponer, *MoRFchibi SYSTEM: software tools for the identification of MoRFs in protein sequences.* Nucleic Acids Res, 2016.
26. Sharma, R., et al., *OPAL+: Length-Specific MoRF Prediction in Intrinsically Disordered Protein Sequences.* Proteomics, 2018. **1800058**: p. 1800058.
27. Hanson, J., et al., *Identifying molecular recognition features in intrinsically disordered regions of proteins by transfer learning.* Bioinformatics, 2020. **36**(4): p. 1107-1113.
28. Mészáros, B., I. Simon, and Z. Dosztányi, *Prediction of Protein Binding Regions in Disordered Proteins.* PLoS Comput Biol, 2009. **5**(5): p. e1000376.
29. Peng, Z. and L. Kurgan, *High-throughput prediction of RNA, DNA and protein binding regions mediated by intrinsic disorder.* Nucleic Acids Res, 2015. **43**(18): p. e121.
30. Oldfield, C.J., Z. Peng, and L. Kurgan, *Disordered RNA-Binding Region Prediction with DisoRDPbind.* Methods Mol Biol, 2020. **2106**: p. 225-239.
31. Mészáros, B., G. Erdős, and Z.J.N.a.r. Dosztányi, *IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding.* 2018. **46**(W1): p. W329-W337.
32. Giri, R., et al., *Understanding COVID-19 via comparative analysis of dark proteomes of SARS-CoV-2, human SARS and bat SARS-like coronaviruses.* Cell Mol Life Sci, 2020.
33. Oliva Chavez, A.S., et al., *Mutational analysis of gene function in the Anaplasmataceae: Challenges and perspectives.* Ticks Tick Borne Dis, 2019. **10**(2): p. 482-494.
34. Jamsheer, K.M., et al., *The FCS-like zinc finger scaffold of the kinase SnRK1 is formed by the coordinated actions of the FLZ domain and intrinsically disordered regions.* J Biol Chem, 2018. **293**(34): p. 13134-13150.
35. Murph, M., S. Singh, and M. Schvarzstein, *The Centrosomal Swiss Army Knife: A combined in silico and in vivo approach to the structure-function annotation of SPD-2 provides mechanistic insight into its functional diversity.* bioRxiv, 2021: p. 2021.04.22.441031.
36. Szabo, B., et al., *Disordered Regions of Mixed Lineage Leukemia 4 (MLL4) Protein Are Capable of RNA Binding.* Int J Mol Sci, 2018. **19**(11).
37. Pujols, J., et al., *The Disordered C-Terminus of Yeast Hsf1 Contains a Cryptic Low-Complexity Amyloidogenic Region.* Int J Mol Sci, 2018. **19**(5).
38. Necci, M., et al., *Critical assessment of protein intrinsic disorder prediction.* Nat Methods, 2021. **18**(5): p. 472-481.
39. Zhang, J., Z. Ma, and L. Kurgan, *Comprehensive review and empirical analysis of hallmarks of DNA-, RNA- and protein-binding residues in protein chains.* Brief Bioinform, 2019. **20**(4): p. 1250-1268.
40. Xue, L.C., et al., *Computational prediction of protein interfaces: A review of data driven methods.* FEBS Lett, 2015. **589**(23): p. 3516-26.
41. Esmaielbeiki, R., et al., *Progress and challenges in predicting protein interfaces.* Brief Bioinform, 2016. **17**(1): p. 117-31.
42. Zhang, J. and L. Kurgan, *Review and comparative assessment of sequence-based predictors of protein-binding residues.* Brief Bioinform, 2018. **19**(5): p. 821-837.
43. Yan, J., S. Friedrich, and L. Kurgan, *A comprehensive comparative review of sequence-based predictors of DNA- and RNA-binding residues.* Brief Bioinform, 2016. **17**(1): p. 88-105.
44. Miao, Z. and E. Westhof, *A Large-Scale Assessment of Nucleic Acids Binding Site Prediction Programs.* PLoS Comput Biol, 2015. **11**(12): p. e1004639.
45. Wang, K., et al., *Comprehensive Survey and Comparative Assessment of RNA-Binding Residue Predictions with Analysis by RNA Type.* International Journal of Molecular Sciences, 2020. **21**(18): p. 6879.

46. Taherzadeh, G., et al., *Sequence-based prediction of protein–peptide binding sites using support vector machine.* Journal of computational chemistry, 2016.
47. Wei, Z.-S., et al., *Protein–protein interaction sites prediction by ensembling SVM and sample-weighted random forests.* Neurocomputing, 2016. **193**: p. 201-212.
48. Wang, X., et al., *Protein-protein interaction sites prediction by ensemble random forests with synthetic minority oversampling technique.* Bioinformatics, 2018.
49. Zhang, J. and L. Kurgan, *SCRIBER: accurate and partner type-specific prediction of protein-binding residues from proteins sequences.* Bioinformatics, 2019. **35**(14): p. i343-i353.
50. Walia, R.R., et al., *RNABindRPlus: A Predictor that Combines Machine Learning and Sequence Homology-Based Methods to Improve the Reliability of Predicted RNA-Binding Residues in Proteins.* Plos One, 2014. **9**(5).
51. El-Manzalawy, Y., et al., *FastRNABindR: Fast and Accurate Prediction of Protein-RNA Interface Residues.* PLoS One, 2016. **11**(7): p. e0158445.
52. Hu, J., et al., *Predicting protein-DNA binding residues by weightedly combining sequence-based features and boosting multiple SVMs.* IEEE/ACM transactions on computational biology and bioinformatics, 2016. **14**(6): p. 1389-1398.
53. Zhu, Y.H., et al., *DNAPred: Accurate Identification of DNA-Binding Sites from Protein Sequence by Ensembled Hyperplane-Distance-Based Support Vector Machines.* Journal of Chemical Information and Modeling, 2019. **59**(6): p. 3057-3071.
54. Yan, J. and L. Kurgan, *DRNApred, fast sequence-based method that accurately predicts and discriminates DNA- and RNA-binding residues.* Nucleic Acids Res, 2017. **45**(10): p. e84.
55. Zhang, J., Q. Chen, and B. Liu, *NCBRPred: predicting nucleic acid binding residues in proteins based on multilabel learning.* Brief Bioinform, 2021.
56. Wang, L., et al., *BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features.* BMC Syst Biol, 2010. **4 Suppl 1**: p. S3.
57. Qiu, J., et al., *ProNA2020 predicts protein-DNA, protein-RNA, and protein-protein binding proteins and residues from sequence.* J Mol Biol, 2020. **432**(7): p. 2428-2443.
58. Sun, Z., et al., *To improve the predictions of binding residues with DNA, RNA, carbohydrate, and peptide via multi-task deep neural networks.* IEEE/ACM Trans Comput Biol Bioinform, 2021. **PP**.
59. Zhang, J., S. Ghadermarzi, and L. Kurgan, *Prediction of protein-binding residues: dichotomy of sequence-based methods developed using structured complexes versus disordered proteins.* Bioinformatics, 2020. **36**(18): p. 4729-4738.
60. Zhang, Y. and Q. Yang, *An overview of multi-task learning.* National Science Review, 2018. **5**(1): p. 30-43.
61. Caruana, R., *Multitask learning.* Machine Learning, 1997. **28**(1): p. 41-75.
62. Singh, D., D.S. Sisodia, and P. Singh, *Compositional framework for multitask learning in the identification of cleavage sites of HIV-1 protease.* Journal of Biomedical Informatics, 2020. **102**.
63. Wu, T.Q., et al., *DeepDist: real-value inter-residue distance prediction with deep residual convolutional network.* Bmc Bioinformatics, 2021. **22**(1).
64. Katuwawala, A. and L. Kurgan, *Comparative Assessment of Intrinsic Disorder Predictions with a Focus on Protein and Nucleic Acid-Binding Proteins.* Biomolecules, 2020. **10**(12).
65. Fu, L., et al., *CD-HIT: accelerated for clustering the next-generation sequencing data.* Bioinformatics, 2012. **28**(23): p. 3150-2.
66. Li, M., et al., *Automated ICD-9 Coding via A Deep Learning Approach.* IEEE/ACM Trans Comput Biol Bioinform, 2019. **16**(4): p. 1193-1202.
67. Gao, J., et al., *PSIONplus(m) Server for Accurate Multi-Label Prediction of Ion Channels and Their Types.* Biomolecules, 2020. **10**(6).
68. Long, W., Y. Yang, and H.B. Shen, *ImPLoc: a multi-instance deep learning model for the prediction of protein subcellular localization based on immunohistochemistry images.* Bioinformatics, 2020. **36**(7): p. 2244-2250.
69. Fang, C., et al., *Identifying short disorder-to-order binding regions in disordered proteins with a deep convolutional neural network method.* J Bioinform Comput Biol, 2019. **17**(1): p. 1950004.
70. Zhang, F., et al., *A deep learning framework for gene ontology annotations with sequence - and network-based information.* IEEE/ACM Trans Comput Biol Bioinform, 2020. **PP**.
71. Zhang, F., et al., *PROBselect: accurate prediction of protein-binding residues from proteins sequences via dynamic predictor selection.* Bioinformatics, 2020. **36**(Supplement_2): p. i735-i744.
72. McGuffin, L.J., K. Bryson, and D.T. Jones, *The PSIPRED protein structure prediction server.* Bioinformatics, 2000. **16**(4): p. 404-5.

73. Hanson, J., K. Paliwal, and Y. Zhou, *Accurate Single-Sequence Prediction of Protein Intrinsic Disorder by an Ensemble of Deep Recurrent and Convolutional Architectures.* J Chem Inf Model, 2018. **58**(11): p. 2369-2376.
74. Hanson, J., et al., *SPOT-Disorder2: Improved Protein Intrinsic Disorder Prediction by Ensembled Deep Learning.* Genomics Proteomics Bioinformatics, 2020.
75. Hanson, J., et al., *Improving protein disorder prediction by deep bidirectional long short-term memory recurrent neural networks.* Bioinformatics, 2017. **33**(5): p. 685-692.
76. Zeng, M., et al., *Protein-protein interaction site prediction through combining local and global features with deep neural networks.* Bioinformatics, 2020. **36**(4): p. 1114-1120.
77. Li, F., et al., *DeepCleave: a deep learning predictor for caspase and matrix metalloprotease substrates and cleavage sites.* Bioinformatics, 2020. **36**(4): p. 1057-1065.
78. Shen, Z., S.P. Deng, and D.S. Huang, *RNA-Protein Binding Sites Prediction via Multi Scale Convolutional Gated Recurrent Unit Networks.* IEEE/ACM Trans Comput Biol Bioinform, 2020. **17**(5): p. 1741-1750.
79. Kendall, A., Y. Gal, and R. Cipolla. *Multi-task learning using uncertainty to weigh losses for scene geometry and semantics*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.
80. Srivastava, N., et al., *Dropout: A Simple Way to Prevent Neural Networks from Overfitting.* Journal of Machine Learning Research, 2014. **15**: p. 1929-1958.
81. Yang, F., et al., *Graph-based prediction of Protein-protein interactions with attributed signed graph embedding.* BMC Bioinformatics, 2020. **21**(1): p. 323.
82. Yuan, Q., et al., *Structure-aware protein–protein interaction site prediction using deep graph convolutional network.* Bioinformatics, 2021.
83. Maurer, A., M. Pontil, and B. Romera-Paredes, *Sparse coding for multitask and transfer learning*, in *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*. 2013, JMLR.org: Atlanta, GA, USA. p. II–343–II–351.
84. Concu, R. and M.N.D.S. Cordeiro, *Alignment-Free Method to Predict Enzyme Classes and Subclasses.* International Journal of Molecular Sciences, 2019. **20**(21).
85. Nguyen, B.P., et al., *iProDNA-CapsNet: identifying protein-DNA binding residues using capsule neural networks.* Bmc Bioinformatics, 2019. **20**(1).
86. Monastyrskyy, B., et al., *Assessment of protein disorder region predictions in CASP10.* Proteins, 2014. **82 Suppl 2**: p. 127-37.
87. Camacho, C., et al., *BLAST+: architecture and applications.* BMC Bioinformatics, 2009. **10**: p. 421.
88. Su, H., et al., *Improving the prediction of protein-nucleic acids binding residues via multiple sequence profiles and the consensus of complementary methods.* Bioinformatics, 2019. **35**(6): p. 930-936.
89. UniProt, C., *UniProt: a worldwide hub of protein knowledge.* Nucleic Acids Res, 2019. **47**(D1): p. D506-D515.
90. Contreras-Moreira, B., *3D-footprint: a database for the structural analysis of protein-DNA complexes.* Nucleic Acids Res, 2010. **38**(Database issue): p. D91-7.
91. Weirauch, M.T., et al., *Determination and inference of eukaryotic transcription factor sequence specificity.* Cell, 2014. **158**(6): p. 1431-1443.
92. Khan, A., et al., *JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework.* Nucleic Acids Res, 2018. **46**(D1): p. D260-D266.
93. Jolma, A., et al., *DNA-dependent formation of transcription factor pairs alters their binding specificity.* Nature, 2015. **527**(7578): p. 384-+.
94. Isakova, A., et al., *SMiLE-seq identifies binding motifs of single and dimeric transcription factors.* Nat Methods, 2017. **14**(3): p. 316-322.
95. Hu, H., et al., *AnimalTFDB 3.0: a comprehensive resource for annotation and prediction of animal transcription factors.* Nucleic Acids Res, 2019. **47**(D1): p. D33-D38.
96. Blake, J.A. and M.A. Harris, *The Gene Ontology (GO) Project: Structured Vocabularies for Molecular Biology and Their Application to Genome and Expression Analysis.* Current Protocols in Bioinformatics, 2008. **23**(1): p. 7.2.1-7.2.9.
97. Giudice, G., et al., *ATtRACT-a database of RNA-binding proteins and associated motifs.* Database (Oxford), 2016. **2016**.
98. Cook, K.B., et al., *RBPDB: a database of RNA-binding specificities.* Nucleic Acids Res, 2011. **39**(Database issue): p. D301-8.
99. Chowdhury, S., J. Zhang, and L. Kurgan, *In Silico Prediction and Validation of Novel RNA Binding Proteins and Residues in the Human Proteome.* Proteomics, 2018: p. e1800064.

100. Obradovic, Z., et al., *Exploiting heterogeneous sequence properties improves prediction of protein disorder.* Proteins, 2005. **61 Suppl 7**: p. 176-82.
101. Hu, G. and L. Kurgan, *Sequence Similarity Searching.* Curr Protoc Protein Sci, 2019. **95**(1): p. e71.
102. Patil, A., K. Kinoshita, and H. Nakamura, *Hub promiscuity in protein-protein interaction networks.* Int J Mol Sci, 2010. **11**(4): p. 1930-43.
103. Haynes, C., et al., *Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes.* Plos Computational Biology, 2006. **2**(8): p. 890-901.
104. Hu, G., et al., *Functional Analysis of Human Hub Proteins and Their Interactors Involved in the Intrinsic Disorder-Enriched Interactions.* Int J Mol Sci, 2017. **18**(12).
105. Calderone, A., L. Castagnoli, and G. Cesareni, *mentha: a resource for browsing integrated protein-interaction networks.* Nat Methods, 2013. **10**(8): p. 690-1.
106. Georgel, P.T., et al., *Sir3-dependent assembly of supramolecular chromatin structures in vitro.* Proc Natl Acad Sci U S A, 2001. **98**(15): p. 8584-9.
107. McBryant, S.J., V.H. Adams, and J.C. Hansen, *Chromatin architectural proteins.* Chromosome Res, 2006. **14**(1): p. 39-51.
108. Liu, C. and A.J. Lustig, *Genetic analysis of Rap1p/Sir3p interactions in telomeric and HML silencing in Saccharomyces cerevisiae.* Genetics, 1996. **143**(1): p. 81-93.
109. Paetkau, D.W., et al., *Interaction of the yeast RAD7 and SIR3 proteins: implications for DNA repair and chromatin structure.* Genes Dev, 1994. **8**(17): p. 2035-45.
110. Chang, J.F., et al., *Structure of the coiled-coil dimerization motif of Sir4 and its interaction with Sir3.* Structure, 2003. **11**(6): p. 637-49.