JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

Cite This: *J. Chem. Inf. Model.* 2018, 58, 2369−2376

pubs.acs.org/jcim

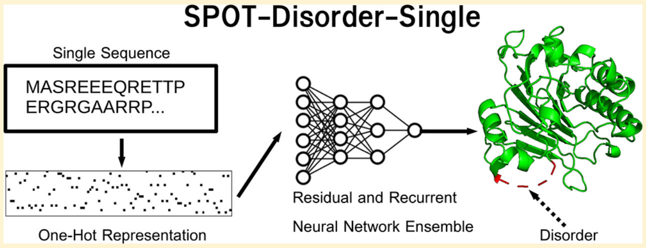
Accurate Single-Sequence Prediction of Protein Intrinsic Disorder by an Ensemble of Deep Recurrent and Convolutional Architectures

Jack Hanson,[†] Kuldip Paliwal,*[,†] and Yaoqi Zhou*[,‡]

[†]Signal Processing Laboratory, Griffith University, Brisbane, Queensland 4122, Australia

[‡]Institute for Glycomics and School of Information and Communication Technology, Griffith University, Southport, Queensland 4222, Australia

ABSTRACT: Recognizing the widespread existence of intrinsically disordered regions in proteins spurred the development of computational techniques for their detection. All existing techniques can be classified into methods relying on single-sequence information and those relying on evolutionary sequence profiles generated from multiple-sequence alignments. The methods based on sequence profiles are, in general, more accurate because the presence or absence of conserved amino acid residues in a protein sequence provides important information on the structural and functional roles of the residues. However, the wide applicability of profile-based techniques is limited by time-consuming calculation of sequence profiles. Here we demonstrate that the performance gap between profile-based techniques and single-sequence methods can be reduced by using an ensemble of deep recurrent and convolutional neural networks that allow whole-sequence learning. In particular, the single-sequence method (called SPOT-Disorder-Single) is more accurate than SPOT-Disorder (a profile-based method) for proteins with few homologous sequences and comparable for proteins in predicting long-disordered regions. The method performance is robust across four independent test sets with different amounts of short- and long-disordered regions. SPOT-Disorder-Single is available as a Web server and as a standalone program at http://sparks-lab.org/jack/server/SPOT-Disorder-Single.



INTRODUCTION

Intrinsically disordered proteins (IDPs) or intrinsically disordered regions (IDRs) in proteins refer to proteins or protein regions that do not possess well-defined, three-dimensional structures at their corresponding physiological conditions. IDPs and IDRs are found abundantly in every domain of life[1,2] with a wide range of function[3,4] and are implicated in many diseases including cancer and neurodegenerative diseases.[5,6] The importance of intrinsic disorder in living organisms can be revealed by the fact that natural proteins are even more intrinsically disordered than proteins with random sequences,[7] likely due to the unique evolutionary advantages of flexibility and multistructural states that disordered proteins have over structured.[8,9]

IDPs and IDRs can be identified by a number of experimental techniques, such as missing regions in X-ray crystallography[10] and dynamics from nuclear magnetic resonance experiments.[11] This data has been collected and annotated in manually curated databases such as DisProt[12] and MobiDB.[13] However, these annotated proteins are only a tiny fraction of all known proteins. Given the high cost of identifying intrinsic disorder by experimental techniques, it is practically necessary to prioritize possible IDRs/IDPs by computational methods, prior to experimental studies.

More than 60 computational methods for ID prediction have been developed[14,15] since the first attempt was made in 1979 by Williams.[16] The first reliable, neural-network-based technique was developed almost 20 years later by Romero et al.[17] Many early methods utilized only single protein sequences and their derived information for prediction through window-based analysis of the physicochemical properties, amino acid propensities, and statistical potentials (e.g., IUPred,[18] Globplot,[19] and FoldIndex[20]). These methods have been shown to generally be outperformed by single-sequence machine learning methods,[21,22] such as the PONDR series,[23] DisEMBL,[24] CSpritz,[25] and Espritz.[26] However, these single-sequence-based techniques are often less accurate than machine-learning techniques using evolutionary sequence profiles generated from multiple sequence alignment.[22,27] This is because sequence profiles, generally created by programs such as PSI-Blast and HHBlits,[28,29] contain important information pertaining to the presence or the lack of evolutionarily conserved amino acid residues due to their respective structural and functional roles. Examples of some recent methods based on sequence profiles are SPOT-Disorder,[21] SPINE-D,[30] NetSurf,[31] MDFp2,[32] and AUCpred.[33]

However, generating the evolutionary profile for a given sequence is computationally intensive as the library size of known protein sequences has exponentially increased in recent years because of increasingly cheaper sequencing techniques. As a result, it is often too time-consuming to perform genome-

Received: September 17, 2018
Published: November 5, 2018

DOI: 10.1021/acs.jcim.8b00636
J. Chem. Inf. Model. 2018, 58, 2369−2376

Table 1. Architecture of Each of the Nine Ensemble Methods

model no.	RNN first	N_{LSTM}	$Blocks_{LSTM}$	N_{CNN}	K_{CNN}	$Blocks_{CNN}$	N_{FC}	$Blocks_{FC}$
0	N	200	2	-	-	-	500	1
1	Y	200	2	120	3	10	250	1
2	Y	200	2	60	3	10	250	2
3	Y	200	1	60	3	10	250	1
4	Y	200	2	60	3	10	125	1
5	Y	200	2	60	7	5	250	2
6	N	200	2	60	9	5	250	1
7	Y	200	2	60	15	5	250	1
8	Y	200	2	60	15	10	500	1

scale analysis of protein intrinsic disorder using profile-based techniques. Moreover, in real-world applications, the majority of proteins (>90%) do not belong to a large sequence cluster.[34] In other words, the quality of sequence profiles for the majority of proteins is poor due to limited evolutionary information. In this case, single-sequence-based techniques may well be more accurate than profile-based techniques as demonstrated for single-sequence prediction of solvent accessible surface area[35] and secondary structure.[36] Thus, it is highly desirable to have a highly accurate single-sequence-based method. Improving existing single-sequence-based methods also addresses the fundamental question of how far we can push the accuracy limit without relying on evolutionary information, knowing that protein intrinsic disorder is wholly determined by its sequence alone.

Improving over existing single-sequence methods is possible. This is because these single-sequence techniques are based on outdated machine learning architectures, such as a small SVM[37] or simple Neural Network,[38] except Espritz which employed a vanilla Recurrent-NN (RNN). On the other hand, several recent profile-based disorder predictors demonstrated the power of advanced machine-learning techniques in improving disorder prediction. Examples are deep convolutional neural fields,[33] deep bidirectional long short-term memory (LSTM) RNN,[21] and combined convolutional and LSTM networks.[31]

This work was inspired by our recent success in using an ensemble of coupled residual Convolutional Neural Networks (residual CNNs or ResNets)[39] and LSTM networks.[40] Such an ensemble allows the removal of prediction noise and improves robustness of performance in protein contact map prediction[41] and protein ω angle prediction.[42] Although LSTM methods have already provided high accuracies for protein disorder prediction in previous works,[21,31] their combinations with ResNets can increase the effectiveness in propagating information throughout the protein sequence. Here we showed that an ensemble of ResNets with LSTM networks leads to the most accurate single-sequence-based technique based on the comparison to existing state-of-the-art techniques.

■ METHODS AND DATA

Neural Network. SPOT-Disorder-Single utilizes an ensemble of 8 models each consisting of ResNets and/or LSTM BRNNs (Table 1). To simplify the description of their implementations, we discuss each of these architectures as a culmination of several functional blocks, as provided in Figure 1a-c, with each block representing the LSTM, ResNet, and Fully-Connected (FC) segments of the architecture, respectively. One example of the full model in the ensemble (Model 2 in Table 1) is illustrated as a flowchart of these blocks in Figure 1d.

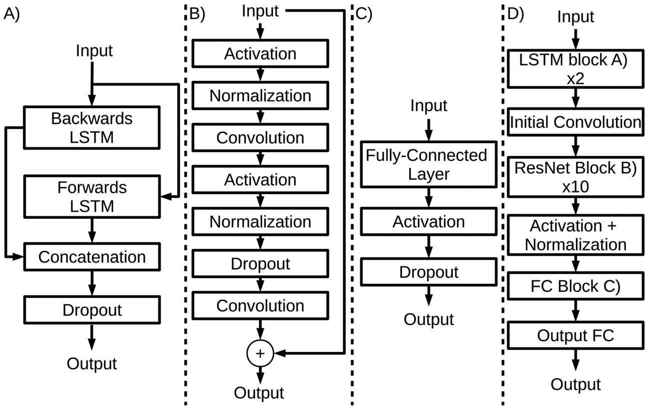


Figure 1. Each of the model blocks used in these experiments: A) the LSTM block; B) the preactivation ResNet block; C) the Fully-Connected (FC) block; and D) the architecture of Model 2 from Table 1, utilizing each of the three previous model blocks.

The LSTM blocks, shown in Figure 1a, consist of a bidirectional LSTM layer with N_{LSTM} one-cell memory blocks in each direction, concatenating together to provide an output of size $2 \times N_{LSTM}$.[40,43] Unlike our previous methods on contact map and ω angle prediction,[41,42] the LSTM blocks employed neither residual connections across the LSTM layer nor normalization of the LSTM layer because we found that they are not useful for improving performance in disorder prediction. A dropout of 50% is applied to the output of the LSTM layer.[44]

The ResNet blocks follow the preactivation architecture in He et al.,[39] in which the residual connection is applied over the unactivated outputs of the convolution layer, rather than the activation layer, as in He et al.[45] The layout used is shown in Figure 1b. The one-dimensional (1D) convolution layers applied in our models utilize a filter depth of N_{CNN} at a kernel size of K_{CNN}. Each convolution layer is normalized using the batch normalization technique[46] and activated by the Exponential Linear Unit (ELU).[47] Dropout is applied prior to the secondary convolution layer in the ResNet block, at a ratio of 25%. As the input needs to be prior to activation, we apply an unactivated convolution layer prior to the first ResNet block in each model. To compensate for this, the output of the last ResNet block is activated and normalized.

The FC blocks shown in Figure 1c are simply multilayer perceptron layers with a size of N_{FC}, an ELU activation, and dropout of 50%.[38] The output layer is a single neuron with a sigmoid activation and no dropout.

We trained numerous models, all formed from these three base blocks, and selected the top 9 models based on their performance on the validation set across several analysis metrics. These 9 models are then combined as an ensemble predictor, providing a more general output due to the suppression of spurious generalizations made after progressing through the unique learning path of each model.[48] While ensembles are generally a combination of several diverse machine learning techniques or other computational methods,[49] we have found that slightly altering the topologies of an already accurate model is sufficient to provide diverse and independent outputs, two fundamental properties of an accurate ensemble.[50] Thus, our final output is the mean of all 9 models (1 LSTM model, 1 ResNet-LSTM model, and 7 LSTM-ResNet models).

Each model was implemented in Tensorflow v1.4,[51] allowing us to accelerate training up to 20 times faster than CPU on an NVIDIA TitanX GPU.[52] The weights in the network were optimized using the Adam optimizer,[53] using the default hyperparameters.

Input Features. In this work, our input feature vector representing the protein chain is a binary one-hot matrix of size 20 × L with each row representing one amino acid type (1 for the amino acid type at that position and 0 otherwise), where L is the protein sequence length. As was noted by Heffernan et al.,[36] using a different matrix representation such as the BLOSUM62 matrix[54] or physicochemical properties of each amino acid[55] does not provide any substantial difference to performance as these are easily learnable by the network as a linear transformations on the one-hot vector. The input feature vector for each residue was standardized to have zero mean and unit variance according to the means and standard deviations of the training data.

Datasets. We used the same datasets as our previous protein disorder prediction work.[21,30] In brief, we obtained 4229 proteins which were randomly split into a train set of 3000 chains (DM3000) and a test set of 1229 chains (DM1229). These datasets were obtained from chains in the Disprot database:[56] 72 fully disordered chains from v5.0 of the Disprot database[12] and 4125 high-resolution structures derived from X-ray crystallography prior to the 5th of August, 2003. These sequences have a sequence similarity of <25% according to BlastClust.[28] A validation set was also randomly taken from the training set, leaving a final training and validation sets of 2700 and 300 proteins, respectively.

Additionally, we employed three independent test sets, SL329,[30,57] Mobi11924,[21,58,59] and Disprot267.[27] The SL329 dataset is a reduced set from the SL477 protein set released by Sirota et al.[57] By removing the overlap between DM4229 and SL477, we obtained 329 nonredundant protein sequences as an independent test set. The MobiDB dataset[58] consists of the entire DisProt database, indirectly inferred IDRs and IDPs from the Protein DataBank (PDB), and, most commonly, predictions from a large ensemble of various disorder predictors. The original set was reduced to 11924 proteins after filtering for 25% sequence identity against DM4229, removing sequences less than 30 residues or more than 20000 residues in length, and removing sequences with unknown amino acid residues. Finally, we also obtained the Disprot Complement set from Necci et al.,[27] which are newly annotated proteins in Disprot v7.0.[60] This dataset is dominated by long IDRs and IDPs. We removed one protein with unknown amino acids to obtain Disprot267.

Performance Evaluation. As our output is a singular node whose value has been compressed to be in the range of [0,1), we can treat our output as the probability of the residue at that position of being disordered. By using a cutoff threshold T for each model, we can linearly separate our outputs into the two classes and analyze the results based on the accuracies of these thresholded values. These thresholded values can thus be separated into True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN).

As protein disorder prediction has an innately skewed class distribution in DM4229, it would be trivial to gain ≈90% accuracy by simply designating every residue as ordered. As such, it is important to obtain skew-independent metrics with which we can analyze the performance of our model for both classes. The two simplest metrics are sensitivity ($\mathrm{Se} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$) and specificity ($\mathrm{Sp} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$), which can be seen as the accuracy of class 1 and class 0 prediction, respectively.

The specificity and sensitivity of a model can be combined to form the Receiver Operating Characteristic (ROC) curve. The Area Under the ROC Curve (AUC) is an unbiased metric which provides the probability that a randomly sampled positive sample will obtain a higher ranking than a randomly sampled negative sample.[61] For comparison between two ROC curves, it is often necessary to calculate the significance between the AUC values using the P-value in order to reject the null hypothesis.[62]

Finally, we use the Matthew's Correlation Coefficient (MCC) to gauge the correlation of the predicted labels and true labels.[63] The formula for the MCC value is

$$\mathrm{MCC} = \frac{\mathrm{TP}\cdot\mathrm{TN} + \mathrm{FN}\cdot\mathrm{FP}}{\sqrt{(\mathrm{TP} + \mathrm{FN})(\mathrm{TP} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})}} \tag{1}$$

Supplementary to these analysis metrics (which are based on a residue-level analysis), we can also analyze the predicted disorder content on a per-protein basis. We take the Pearson Correlation Coefficient (PCC)[64] and calculate the Root Mean Squared Error (RMSE) of the predicted and actual disordered content for each protein in the test set. The RMSE is provided by

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{n=0}^{N-1}(y_n - x_n)^2}{N}} \tag{2}$$

where N is the number of proteins in the dataset, and y_n and x_n are the actual and predicted disordered content for protein n, respectively. The content is normalized for each protein by the number of annotated residues as to not bias longer proteins in the dataset.

The maximum value of all of our per-residue analysis metrics is 1, indicating that the predictor which obtains the highest value for AUC of the ROC and/or MCC can be considered to be the most accurate for disorder prediction. The predictor which minimizes the RMSE can be considered to perform the best for that analysis.

Method Comparison. Because this method is a single-sequence based technique, we will mostly compare to other single-sequence-based methods. To this end, we downloaded the standalone versions of MobiDB-lite[65] (available at http://protein.bio.unipd.it/mobidblite/) and DisEMBL (available at http://dis.embl.de/html/download.html) and accessed the online server of IUPred2A[66] (Server URL: https://iupred2a.

Table 2. Performance of the Proposed Methods on All Test Sets

Dataset	AUC_{ROC}	MCC	Pr	Se	Sp	no. Ord	no. Dis
Validation	0.888	0.575	0.712	0.517	0.979	61231	6083
DM1229	0.868	0.518	0.707	0.432	0.981	276748	29082
SL329	0.887	0.604	0.939	0.563	0.972	51292	39544
Mobi11249	0.858	0.438	0.561	0.389	0.980	2917685	194753
Disprot267	0.788	0.425	0.528	0.668	0.787	82989	29579

elte.hu/), sequence-only Espritz (X-ray-, NMR-, and Disprot-trained) (Server URL: http://protein.bio.unipd.it/espritz/), and PONDR-VLXT (Server URL: http://www.pondr.com/). The MobiDB-lite version downloaded utilizes the short and long IUPred, DisEMBL 465, and hot-loops, sequence-only Espritz D/N/X, and Globplot predictors in its consensus modeling. In addition, we compare to the profile-based technique SPOT-Disorder as a baseline for profile-based methods (available athttp://sparks-lab.org/jack).

Table 3. Performance of the Ensemble Methods on the DM1229 Set

Model	AUC	MCC	Pr	Se	Sp
Model 0	0.862	0.490	0.683	0.404	0.980
Model 1	0.862	0.500	0.714	0.400	0.983
Model 2	0.861	0.496	0.744	0.375	0.986
Model 3	0.863	0.498	0.728	0.387	0.985
Model 4	0.849	0.489	0.761	0.355	0.988
Model 5	0.854	0.496	0.678	0.418	0.979
Model 6	0.852	0.498	0.679	0.421	0.979
Model 7	0.841	0.482	0.644	0.421	0.976
Model 8	0.858	0.504	0.686	0.425	0.980
SPOT-Disorder-Single	0.868	0.518	0.707	0.432	0.981

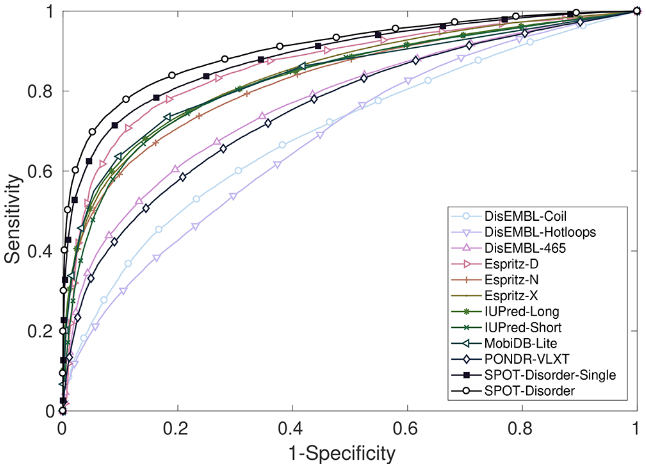


Figure 2. Receiver Operating Characteristic curve for all predictors on the SL329 dataset.

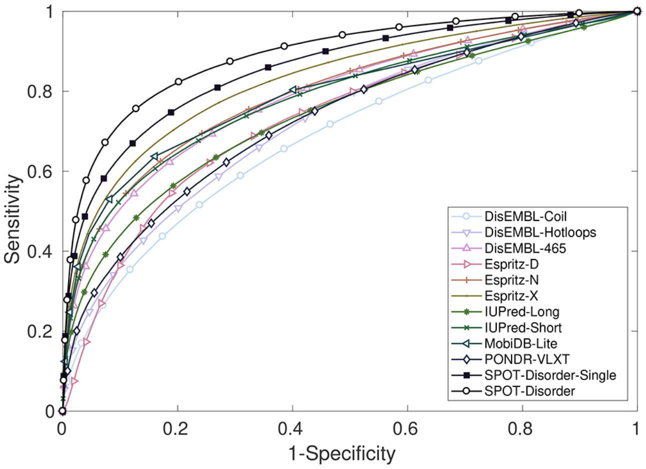


Figure 3. Receiver Operating Characteristic curve for all predictors on the Mobi11924 dataset.

■ RESULTS

The results of SPOT-Disorder-Single on the validation and four independent test sets are presented in Table 2. As Table 2 shows, these datasets have substantially different ratios of ordered to disordered residues ranging from 15:1 in Mobil1249, 10:1 in validation, 9.5:1 in DM1229, 3.8:1 in Disprot267, and 1.3:1 in SL329. This is largely due to different numbers of fully disordered proteins included in each set. Thus, it is not surprising that the performance varies across different datasets. Nevertheless, the AUC values for validation, DM1229, SL329, and Mobil1249 varied slightly from 0.86 to 0.89, suggesting an overall robustness of the method. The DisProt267 set, the latest annotations from DisProt v7.0, turns out to be the hardest dataset with an AUC of 0.788.

Using the DM1229 test set as an example, Table 3 compared the performance of individual models with those of the final consensus. Although the AUC value only improved by 0.005 over the best single model, the MCC value increases by 3%. The improvement is statistically significant with a P-value of 1 × 10^{-3}.

Figures 2, 3, and 4 show the ROC curves produced by various methods for three independent test sets SL329, Mobi11924, and DisProt267, respectively. Performance measures of these methods are also tabulated in Table 4. SPOT-Disorder-single has the highest AUC and MCC values for SL329 (5% higher in AUC and 9% higher in MCC than the next best) and Mobi11924 (3% higher in AUC and 16% higher in MCC than the next best) among all single-sequence methods compared. The AUC by SPOT-Disorder-single is only 3−4% lower than that by the profile-based method SPOT-Disorder for SL329 and Mobi11924. For DisProt267, SPOT-Disorder-Single is only slightly worse than Espritz trained on the DisProt dataset in AUC (0.781 versus 0.796) and slightly better in MCC (0.416 versus 0.414). However, Espritz-DisProt performs significantly worse than SPOT-Disorder-Single for Mobil11924 (0.732 versus 0.858 in AUC, 0.209 versus 0.442 in MCC), suggesting Espritz-DisProt is not as general as SPOT-Disorder-Single. This is possibly due to the fact that Espritz-Disprot was trained on long-disordered chains from the Disprot database, which forms the bulk of the Disprot267 dataset.

Long-disordered regions have different preference for amino acid residues.[15] Thus, it is of interest to examine if the method

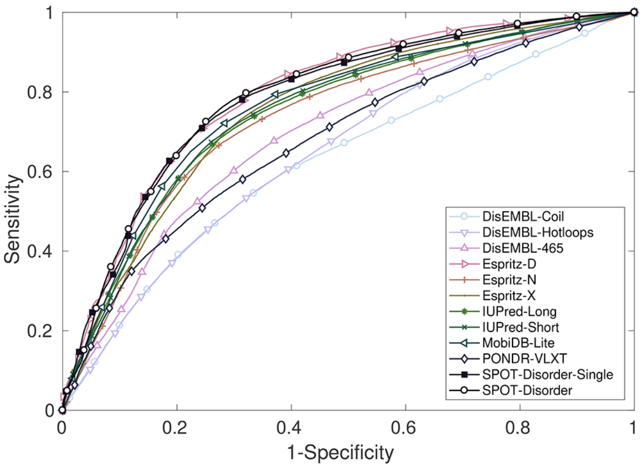


Figure 4. Receiver Operating Characteristic curve for all predictors on the Disprot267 dataset.

developed has a particular bias toward sizes of intrinsically disordered regions. Figure 5 plots MCC values as disordered and ordered regions are binned according to their sizes. The largest Mobi11924 set was used so that we have sufficient statistics in each bins. Except for a few methods, the majority has a similar trend: the medium size around 30 residues has the best discrimination between unstructured and structured regions. SPOT-Disorder-Single is comparable to a few methods in short IDRs but are more accurate than all single-sequence methods in long-disordered regions. In fact, it is even closer to the MCC values of the profile-based SPOT-Disorder in long-disordered regions.

The above analysis is based on performance per residues. Another way to evaluate method performance is to examine the overall performance at the protein level: the fraction of the residues in disordered regions (disordered content). The performance can be measured by RMSE and PCC values. SPOT-Disorder-Single was not specifically trained for protein-level disordered contents. Thresholds used for optimizing MCC values are not necessarily optimized to measure disorder contents. Table 4 compared RMSE and PCC values given by SPOT-Disorder-Single to other single-sequence-based techniques. If the threshold for optimizing MCC values is employed, the performance of our method is comparable to other methods.

To understand the performance difference between a profile-based and single-sequence methods, we plot the AUC value as a function of the Number of effective homologous sequences (Neff) in Figure 6. Neff is a parameter used in HHblits to measure the effective size of its homologous sequence cluster. The figure confirms that SPOT-Disorder-Single performs significantly better than the profile-based SPOT-Disorder when there is a lack of homologous sequences (very low Neff), similar to single-sequence-based method for solvent accessibility[35] and secondary structure prediction.[36]

For illustrative purposes, we present the predictions on cellular tumor antigen p53 (UniProt ID: P04637), a tumor suppressing protein found in multicellular organisms containing several IDR[67] in Figure 7. We selected the models DisEMBL-465, Espritz-X, and IUPred-Long for comparison due to their high performance across each test set. Espritz-D is potentially biased as this protein also exists in the DisProt database (DisProt ID: DP00086) and was thus not chosen. As shown in Figure 7, SPOT-Disorder-Single is able to accurately map all ordered and disordered regions, achieving an MCC of 0.826 for this protein, higher than the next highest of 0.759 achieved by IUPred-Long. While it achieves the third-lowest sensitivity (94.8% vs 100% for Espritz-X and 96.0% for IUPred-Long), SPOT-Disorder-Single achieves the highest precision of 85.6% due to its lower occurrence of false positives, compared to 65.9% and 78.0% for Espritz-X and IUPred-Long, respectively.

DISCUSSION

In this paper, we have developed a method called SPOT-Disorder-Single to predict intrinsic disorder of proteins by using the sequence of amino acids only, with an ensemble of coupled deep recurrent and convolutional neural networks. We showed that employing the ensemble of deep neural networks allows the method to improve over other sequence-only techniques in terms of the ability to separate intrinsically disordered from structured regions of proteins. The robust performance is demonstrated by using several independent test sets (DM1229, SL329, Mobi11249, and Disprot267) with varying amounts of disordered content in short- and long-

Table 4. Performance of Several Predictors on the Independent Test Sets[d]

	SL329				Mobi11924				Disprot267			
Predictor	AUC	MCC	RMSE	PCC	AUC	MCC	RMSE	PCC	AUC	MCC	RMSE	PCC
SPOT-Disorder[a]	0.905	0.672	0.358	0.646	0.891	0.498	0.160	0.436	0.792	0.431	0.298	0.515
PONDR	0.755	0.384	0.330	0.569	0.73	0.191	0.256	0.276	0.680	0.234	0.28	0.439
IUP-Short	0.830	0.506	0.389	0.615	0.784	0.343	0.138	0.369	0.720	0.273	0.258	0.143
IUP-Long	0.838	0.552	0.386	0.587	0.741	0.273	0.159	0.359	0.706	0.292	0.305	0.201
DisEMBL-Coils	0.694	0.226	0.376	0.335	0.687	0.105	0.554	0.062	0.623	0.152	0.378	0.161
DisEMBL-Hotloops	0.682	0.243	0.366	0.408	0.724	0.169	0.312	0.196	0.614	0.128	0.246	0.310
DisEMBL-465	0.770	0.404	0.440	0.565	0.788	0.325	**0.117**	0.333	0.657	0.179	**0.244**	0.216
Espritz-Xray	0.842	0.543	0.324	**0.660**	0.829	0.354	0.187	0.388	0.754	0.372	0.269	0.591
Espritz-NMR	0.826	0.473	**0.302**	0.610	0.796	0.239	0.303	0.342	0.735	0.303	0.320	0.447
Espritz-Disprot	0.863[b]	0.608[b]	0.358	0.594	0.732	0.209	0.404	0.208	**0.796**	0.414	0.439	0.500
MobiDB-lite	0.835	0.554	0.369	0.643	0.791	0.376	0.140	**0.394**	0.762	0.397	0.253	**0.607**
Spot-Disorder-Single[c]	**0.887**	**0.604**	0.403	0.600	**0.857**	**0.438**	0.139	0.393	0.788	**0.425**	0.269	0.600

[a] Profile-based method. [b] The training set for Espritz-Disprot had significant overlap with the SL329 set (23 proteins left after filtering for 25% sequence similarity). The results are listed here for completeness. [c] Using the threshold from the validation set in Table 2 for MCC calculation. [d] The best sequence-based method is in boldface, while the second best is underlined.

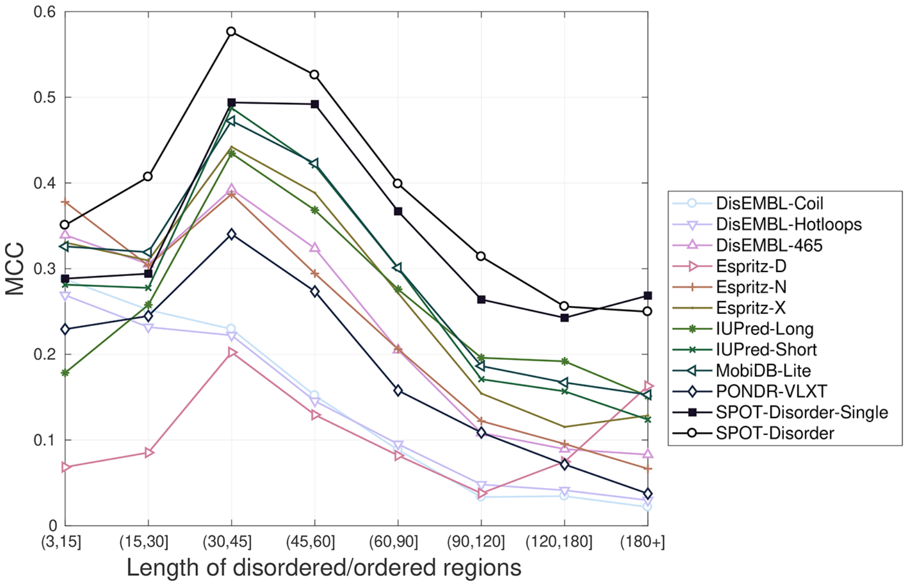


Figure 5. MCC of each predictor on the Mobi11924 dataset when binned by the length of both ordered and disordered regions.

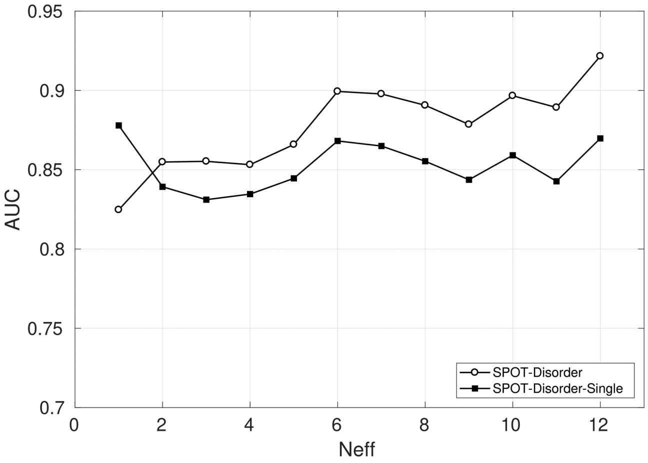


Figure 6. AUC of the ROC curve when binned by the cumulative Neff value of the proteins from the Mobi9981 dataset.

disordered regions. More importantly, we showed that the single-sequence method is more accurate than the profile-based technique for proteins with few homologous sequences.

One advantage of SPOT-Disorder-Single is its speed. We obtained the computing time required for processing the large Mobi11924 dataset. The timing was measured on CPU (Intel Xeon CPU E5-1650 v2 @ 3.50 GHz) and GPU (Nvidia GTX Titan X). This ended up making no difference during testing, as the CPU and GPU versions both took 85 min for a full ensemble pass over the data without parallelization. By comparison, the profile-based SPOT-Disorder took more than 3 weeks for the same set, even with parallel processing of the profiles over multiple machines. A single-machine implementation would have taken over a month. Meanwhile, the Espritz server, which splits incoming jobs into 8 parallel batches for processing on a high-performance computing cluster, processed the Mobi11924 set in approximately 20 min. Thus, SPOT-Disorder-Single is competitive in terms of computing speed against other single-sequence servers as well. This method allows a fast genome-scale prediction that is often too time-consuming for profile-based techniques.

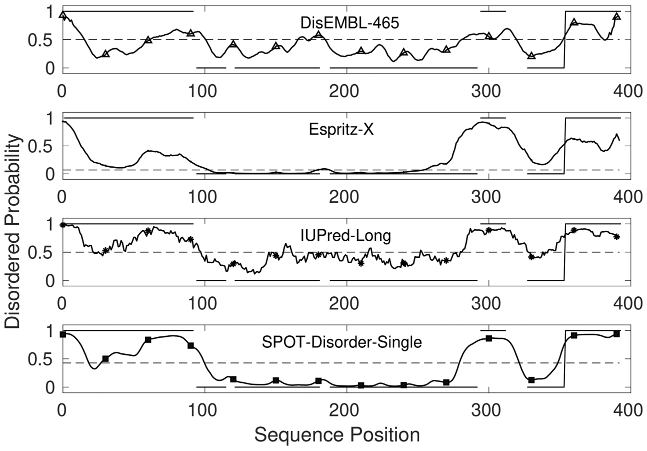


Figure 7. Predictions of several different predictors on the p53 protein (UniProt ID: P04637). The dashed horizontal line in each graph represents the threshold of that predictor (e.g., 0.426 for SPOT-Disorder-Single), and the remaining disjointed lines are the labels for the target protein. There are several discontinuities (residues 92, 115−119, 181−186, 292, 312−325) where the curated and indirectly inferred labels are conflicting in the MobiDB database.

Another advantage of SPOT-Disorder-Single is that its performance is more accurate than the modern profile-based technique SPOT-Disorder for proteins with few homologous sequences. This is important as most proteins have few homologous sequences.[34] Moreover, the method has a comparable performance to SPOT-Disorder for long-disordered regions as shown in Figure 5. This significantly enhances the quality of the results for genome-scale prediction where locating a stretch of disordered regions separating structured domains is often needed.

■ AUTHOR INFORMATION

Corresponding Authors

*E-mail: k.paliwal@griffith.edu.au.

*E-mail: yaoqi.zhou@griffith.edu.au.

ORCID

Jack Hanson: 0000-0001-6956-6748

Notes

The authors declare no competing financial interest.

To further facilitate the usage, SPOT-Disorder-Single is freely available as a Web-based server, and a downloadable version for local implementation and all datasets employed in this study for this article may be accessed at http://sparks-lab.org.

■ ACKNOWLEDGMENTS

We would like to thank M. Necci and S. C. Tosatto for providing us the dataset of DisProt267 and the results of the methods they tested. This work was supported by Australia Research Council DP180102060 to Y.Z. and K.P. and in part by the National Health and Medical Research Council (1121629) of Australia to Y.Z. We also gratefully acknowledge the use of the High Performance Computing Cluster "Gowonda" to complete this research and the aid of the research cloud resources provided by the Queensland Cyber Infrastructure Foundation (QCIF).

■ REFERENCES

(1) Xue, B.; Dunker, A. K.; Uversky, V. N. 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*, 137−149.

(2) Peng, Z.; Yan, J.; Fan, X.; Mizianty, M. J.; Xue, B.; Wang, K.; Hu, G.; Uversky, V. N.; Kurgan, L. Exceptionally Abundant Exceptions: Comprehensive Characterization of Intrinsic Disorder in All Domains of Life. *Cell. Mol. Life Sci.* **2015**, *72*, 137−151.

(3) Dyson, H. J.; Wright, P. E. Intrinsically Unstructured Proteins and Their Functions. *Nat. Rev. Mol. Cell Biol.* **2005**, *6*, 197−208.

(4) Dunker, A. K.; Silman, I.; Uversky, V. N.; Sussman, J. L. Function and Structure of Inherently Disordered Proteins. *Curr. Opin. Struct. Biol.* **2008**, *18*, 756−764.

(5) Uversky, V. N.; Oldfield, C. J.; Dunker, A. K. Intrinsically Disordered Proteins in Human Diseases: Introducing the D2 Concept. *Annu. Rev. Biophys.* **2008**, *37*, 215−246.

(6) Shigemitsu, Y.; Hiroaki, H. Common Molecular Pathogenesis of Disease-related Intrinsically Disordered Proteins Revealed by NMR Analysis. *J. Biochem.* **2018**, *163*, 11−18.

(7) Yu, J.-F.; Cao, Z.; Yang, Y.; Wang, C.-L.; Su, Z.-D.; Zhao, Y.-W.; Wang, J.-H.; Zhou, Y. Natural Protein Sequences Are More Intrinsically Disordered Than Random Sequences. *Cell. Mol. Life Sci.* **2016**, *73*, 2949−2957.

(8) Receveur-Bréchot, V.; Bourhis, J.-M.; Uversky, V. N.; Canard, B.; Longhi, S. Assessing Protein Disorder and Induced Folding. *Proteins: Struct., Funct., Genet.* **2006**, *62*, 24−45.

(9) Uversky, V. N. Functions of Short Lifetime Biological Structures at Large: the Case of Intrinsically Disordered Proteins. *Briefings in Funct. Gen.* **2018**, ely023.

(10) DeForte, S.; Uversky, V. N. Resolving the Ambiguity: Making Sense of Intrinsic Disorder When PDB Structures Disagree. *Protein Sci.* **2016**, *25*, 676−688.

(11) Konrat, R. NMR Contributions to Structural Dynamics Studies of Intrinsically Disordered Proteins. *J. Magn. Reson.* **2014**, *241*, 74−85.

(12) Sickmeier, M.; Hamilton, J. A.; LeGall, T.; Vacic, V.; Cortese, M. S.; Tantos, A.; Szabo, B.; Tompa, P.; Chen, J.; Uversky, V. N. DisProt: the Database of Disordered Proteins. *Nucleic Acids Res.* **2007**, 35, D786−D793.

(13) Piovesan, D.; Tabaro, F.; Paladin, L.; Necci, M.; Mičetić, I.; Camilloni, C.; Davey, N.; Dosztányi, Z.; Mészáros, B.; Monzon, A. M. MobiDB 3.0: More Annotations for Intrinsic Disorder, Conformational Diversity and Interactions in Proteins. *Nucleic Acids Res.* **2018**, *46*, D471−D476.

(14) Meng, F.; Uversky, V. N.; Kurgan, L. Comprehensive Review of Methods for Prediction of Intrinsic Disorder and Its Molecular Functions. *Cell. Mol. Life Sci.* **2017**, *74*, 3069−3090.

(15) He, B.; Wang, K.; Liu, Y.; Xue, B.; Uversky, V. N.; Dunker, A. K. Predicting Intrinsic Disorder in Proteins: an Overview. *Cell Res.* **2009**, *19*, 929−949.

(16) Williams, R. J. P. The Conformation Properties of Proteins in Solution. *Biological Reviews* **1979**, *54*, 389−437.

(17) Romero, P.; Obradovic, Z.; Kissinger, C.; Villafranca, J.; Dunker, A. Identifying Disordered Regions in Proteins From Amino Acid Sequence. *Int. Conf. Neural Networks* **1997**, *1997*, 90−95.

(18) Dosztányi, Z.; Csizmok, V.; Tompa, P.; Simon, I. IUPred: Web Server for the Prediction of Intrinsically Unstructured Regions of Proteins Based on Estimated Energy Content. *Bioinformatics* **2005**, *21*, 3433−3434.

(19) Linding, R.; Russell, R. B.; Neduva, V.; Gibson, T. J. GlobPlot: Exploring Protein Sequences for Globularity and Disorder. *Nucleic Acids Res.* **2003**, *31*, 3701−3708.

(20) Prilusky, J.; Felder, C. E.; Zeev-Ben-Mordehai, T.; Rydberg, E. H.; Man, O.; Beckmann, J. S.; Silman, I.; Sussman, J. L. FoldIndex© A Simple Tool to Predict Whether a Given Protein Sequence Is Intrinsically Unfolded. *Bioinformatics* **2005**, *21*, 3435−3438.

(21) Hanson, J.; Yang, Y.; Paliwal, K.; Zhou, Y. Improving Protein Disorder Prediction by Deep Bidirectional Long Short-Term Memory Recurrent Neural Networks. *Bioinformatics* **2017**, *33*, 685−694.

(22) Liu, Y.; Wang, X.; Liu, B. A Comprehensive Review and Comparison of Existing Computational Methods for Intrinsically Disordered Protein and Region Prediction. *Briefings Bioinf.* **2017**, DOI: 10.1093/bib/bbx126.

(23) Romero, P.; Obradovic, Z.; Li, X.; Garner, E. C.; Brown, C. J.; Dunker, A. K. Sequence Complexity of Disordered Protein. *Proteins: Struct., Funct., Genet.* **2001**, *42*, 38−48.

(24) Linding, R.; Jensen, L. J.; Diella, F.; Bork, P.; Gibson, T. J.; Russell, R. B. Protein Disorder Prediction: Implications for Structural Proteomics. *Structure* **2003**, *11*, 1453−1459.

(25) Walsh, I.; Martin, A. J.; Di Domenico, T.; Vullo, A.; Pollastri, G.; Tosatto, S. C. CSpritz: Accurate Prediction of Protein Disorder Segments With Annotation for Homology, Secondary Structure and Linear Motifs. *Nucleic Acids Res.* **2011**, *39*, W190−W196.

(26) Walsh, I.; Martin, A. J.; Di Domenico, T.; Tosatto, S. C. ESpritz: Accurate and Fast Prediction of Protein Disorder. *Bioinformatics* **2012**, *28*, 503−509.

(27) Necci, M.; Piovesan, D.; Dosztányi, Z.; Tompa, P.; Tosatto, S. C. A Comprehensive Assessment of Long Intrinsic Protein Disorder From the DisProt Database. *Bioinformatics* **2018**, *34*, 445−452.

(28) Altschul, S. F.; Madden, T. L.; Schäffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs. *Nucleic Acids Res.* **1997**, *25*, 3389−3402.

(29) Remmert, M.; Biegert, A.; Hauser, A.; Söding, J. HHblits: Lightning-Fast Iterative Protein Sequence Searching by HMM-HMM Alignment. *Nat. Methods* **2012**, *9*, 173−175.

(30) Zhang, T.; Faraggi, E.; Xue, B.; Dunker, A. K.; Uversky, V. N.; 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.

(31) Klausen, M. S.; Jespersen, M. C.; Nielsen, H.; Jensen, K. K.; Jurtz, V. I.; Soenderby, C. K.; Sommer, M. O. A.; Winther, O.; Nielsen, M.; Petersen, B. NetSurfP-2.0: Improved Prediction of Protein Structural Features by Integrated Deep Learning. *bioRxiv* **2018**, 1−14.

(32) Mizianty, M. J.; Stach, W.; Chen, K.; Kedarisetti, K. D.; Disfani, F. M.; Kurgan, L. Improved Sequence-based Prediction of Disordered Regions With Multilayer Fusion of Multiple Information Sources. *Bioinformatics* **2010**, *26*, i489−i496.

(33) Wang, S.; Ma, J.; Xu, J. AUCpreD: Proteome-Level Protein Disorder Prediction by AUC-Maximized Deep Convolutional Neural Fields. *Bioinformatics* **2016**, *32*, i672−i679.

(34) Ovchinnikov, S.; Park, H.; Varghese, N.; Huang, P.-S.; Pavlopoulos, G. A.; Kim, D. E.; Kamisetty, H.; Kyrpides, N. C.; Baker, D. Protein Structure Determination Using Metagenome Sequence Data. *Science* **2017**, *355*, 294−298.

(35) Faraggi, E.; Zhou, Y.; Kloczkowski, A. Accurate Single-sequence Prediction of Solvent Accessible Surface Area Using Local and Global Features. *Proteins: Struct., Funct., Genet.* **2014**, *82*, 3170−3176.

(36) Heffernan, R.; Paliwal, K.; Lyons, J.; Singh, J.; Yang, Y.; Zhou, Y. Single-Sequence-Based Prediction of Protein Secondary Structure, Backbone Angles, Solvent Accessibility, Half-Sphere Exposure, and Contact Number by Long Short-Term Memory Bidirectional Recurrent Neural Networks. *J. Comput. Chem.* **2018**, *39*, 2210−2216.

(37) Vapnik, V. N. *Statistical Learning Theory*; Wiley, New York, 1998; Vol. *1*.

(38) Rumelhart, D. E.; Hinton, G. E.; Williams, R. J. *Learning Internal Representations By Error Propagation*; 1985.

(39) He, K.; Zhang, X.; Ren, S.; Sun, J. Identity Mappings in Deep Residual Networks. *Euro. Conf. Comp. Vis.* **2016**, *9908*, 630−645.

(40) Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comp.* **1997**, *9*, 1735−1780.

(41) Hanson, J.; Paliwal, K.; Litfin, T.; Yang, Y.; Zhou, Y. Accurate Prediction of Protein Contact Maps by Coupling Residual Two-Dimensional Bidirectional Long Short-Term Memory with Convolutional Neural Networks. *Bioinformatics* **2018**, DOI: 10.1093/bioinformatics/bty481.

(42) Singh, J.; Hanson, J.; Heffernan, R.; Paliwal, K.; Yang, Y.; Zhou, Y. Detecting Proline and Non-Proline Cis Isomers in Protein Structures from Sequences Using Deep Residual Ensemble Learning. *J. Chem. Inf. Model.* **2018**, *58*, 2033−2042.

(43) Schuster, M.; Paliwal, K. K. Bidirectional Recurrent Neural Networks. *IEEE Trans. Sig. Proc.* **1997**, *45*, 2673−2681.

(44) Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks From Overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929−1958.

(45) He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning For Image Recognition. *Proc. IEEE Conf. Comp. Vis. and PR.* **2016**, 770−778.

(46) Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *Int. Conf. ML.* **2015**, 448−456.

(47) Clevert, D.-A.; Unterthiner, T.; Hochreiter, S. Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs). *arXiv preprint arXiv:1511.07289* 2015.

(48) Hansen, L. K.; Salamon, P. Neural Network Ensembles. *IEEE Trans. Patt. Anal. Mach. Learn.* **1990**, *12*, 993−1001.

(49) Dehzangi, A.; Paliwal, K.; Sharma, A.; Dehzangi, O.; Sattar, A. A Combination of Feature Extraction Methods With an Ensemble of Different Classifiers for Protein Structural Class Prediction Problem. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **2013**, *10*, 564−575.

(50) Dietterich, T. G. Ensemble Methods in Machine Learning. *Int. Workshop on Multiple Classifier Systems.* **2000**, *1857*, 1−15.

(51) Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G. S.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Goodfellow, I. J.; Harp, A.; Irving, G.; Isard, M.; Jia, Y.; Józefowicz, R.; Kaiser, L.; Kudlur, M.; Levenberg, J.; Mané, D.; Monga, R.; Moore, S.; Murray, D. G.; Olah, C.; Schuster, M.; Shlens, J.; Steiner, B.; Sutskever, I.; Talwar, K.; Tucker, P. A.; Vanhoucke, V.; Vasudevan, V.; Viégas, F. B.; Vinyals, O.; Warden, P.; Wattenberg, M.; Wicke, M.; Yu, Y.; Zheng, X. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *CoRR*. 2016, abs/1603.04467. https://arxiv.org/abs/1603.04467 (accessed Nov 9, 2018).

(52) Oh, K.-S.; Jung, K. GPU Implementation of Neural Networks. *Pattern Recognition* **2004**, *37*, 1311−1314.

(53) Kingma, D. P.; Ba, J. Adam: A Method for Stochastic Optimization. *CoRR*. 2014, abs/1412.6980.https://arxiv.org/abs/1412.6980 (accessed Nov 9, 2018).

(54) Henikoff, S.; Henikoff, J. G. Amino Acid Substitution Matrices From Protein Blocks. *Proc. Natl. Acad. Sci. U. S. A.* **1992**, *89*, 10915−10919.

(55) Meiler, J.; Müller, M.; Zeidler, A.; Schmäschke, F. Generation and Evaluation of Dimension-reduced Amino Acid Parameter Representations by Artificial Neural Networks. *J. Mol. Model.* **2001**, *7*, 360−369.

(56) Vucetic, S.; Obradovic, Z.; Vacic, V.; Radivojac, P.; Peng, K.; Iakoucheva, L. M.; Cortese, M. S.; Lawson, J. D.; Brown, C. J.; Sikes, J. G. DisProt: A Database of Protein Disorder. *Bioinformatics* **2005**, *21*, 137−140.

(57) Sirota, F. L.; Ooi, H.-S.; Gattermayer, T.; Schneider, G.; Eisenhaber, F.; Maurer-Stroh, S. Parameterization of Disorder Predictors for Large-scale Applications Requiring High Specificity by Using an Extended Benchmark Dataset. *BMC Genomics* **2010**, *11*, S15.

(58) Potenza, E.; Di Domenico, T.; Walsh, I.; Tosatto, S. C. MobiDB 2.0: An Improved Database of Intrinsically Disordered and Mobile Proteins. *Nucleic Acids Res.* **2015**, *43*, D315−D320.

(59) Walsh, I.; Giollo, M.; Di Domenico, T.; Ferrari, C.; Zimmermann, O.; Tosatto, S. C. Comprehensive Large-Scale Assessment of Intrinsic Protein Disorder. *Bioinformatics* **2015**, *31*, 201−208.

(60) Piovesan, D.; Tabaro, F.; Mičetić, I.; Necci, M.; Quaglia, F.; Oldfield, C. J.; Aspromonte, M. C.; Davey, N. E.; Davidović, R.; Dosztányi, Z. DisProt 7.0: A Major Update of the Database of Disordered Proteins. *Nucleic Acids Res.* **2017**, *45*, D219−D227.

(61) Fawcett, T. An Introduction To ROC Analysis. *Patt. Rec. Lett.* **2006**, *27*, 861−874.

(62) Hanley, J. A.; McNeil, B. J. The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve. *Radiology* **1982**, *143*, 29−36.

(63) Matthews, B. W. Comparison of the Predicted and Observed Secondary Structure of T4 Phage Lysozyme. *Biochim. Biophys. Acta, Protein Struct.* **1975**, *405*, 442−451.

(64) Pearson, K. Note on Regression and Inheritance in the Case of Two Parents. *Proc. R. Soc. London* **1895**, *58*, 240−242.

(65) Necci, M.; Piovesan, D.; Dosztányi, Z.; Tosatto, S. C. MobiDB-lite: Fast and Highly Specific Consensus Prediction of Intrinsic Disorder in Proteins. *Bioinformatics* **2017**, *33*, 1402−1404.

(66) Mészáros, B.; Erdös, G.; Dosztányi, Z. IUPred2A: Context-Dependent Prediction of Protein Disorder as a Function of Redox State and Protein Binding. *Nucleic Acids Res.* **2018**, *46*, W329−W337.

(67) Kastan, M. B.; Onyekwere, O.; Sidransky, D.; Vogelstein, B.; Craig, R. W. Participation of p53 Protein in the Cellular Response to DNA Damage. *Cancer Res.* **1991**, *51*, 6304−6311.