Bioinformatics, 36(4), 2020, 1107–1113
doi: 10.1093/bioinformatics/btz691
Advance Access Publication Date: 5 September 2019
Original Paper

Structural bioinformatics

Identifying molecular recognition features in intrinsically disordered regions of proteins by transfer learning

Jack Hanson [1,*], Thomas Litfin[2], Kuldip Paliwal[1] and Yaoqi Zhou [2,*]

[1]Signal Processing Laboratory, Griffith University, Brisbane, QLD 4122, Australia and [2]Institute for Glycomics, School of Information and Communication Technology, Griffith University, Southport, QLD 4222, Australia

*To whom correspondence should be addressed.

Associate Editor: Jan Gorodkin

Received on March 5, 2019; revised on July 24, 2019; editorial decision on August 22, 2019; accepted on August 31, 2019

Abstract

Motivation: Protein intrinsic disorder describes the tendency of sequence residues to not fold into a rigid three-dimensional shape by themselves. However, some of these disordered regions can transition from disorder to order when interacting with another molecule in segments known as molecular recognition features (MoRFs). Previous analysis has shown that these MoRF regions are indirectly encoded within the prediction of residue disorder as low-confidence predictions [i.e. in a semi-disordered state P(D)≈0.5]. Thus, what has been learned for disorder prediction may be transferable to MoRF prediction. Transferring the internal characterization of protein disorder for the prediction of MoRF residues would allow us to take advantage of the large training set available for disorder prediction, enabling the training of larger analytical models than is currently feasible on the small number of currently available annotated MoRF proteins. In this paper, we propose a new method for MoRF prediction by transfer learning from the SPOT-Disorder2 ensemble models built for disorder prediction.

Results: We confirm that directly training on the MoRF set with a randomly initialized model yields substantially poorer performance on independent test sets than by using the transfer-learning-based method SPOT-MoRF, for both deep and simple networks. Its comparison to current state-of-the-art techniques reveals its superior performance in identifying MoRF binding regions in proteins across two independent testing sets, including our new dataset of >800 protein chains. These test chains share <30% sequence similarity to all training and validation proteins used in SPOT-Disorder2 and SPOT-MoRF, and provide a much-needed large-scale update on the performance of current MoRF predictors. The method is expected to be useful in locating functional disordered regions in proteins.

Availability and implementation: SPOT-MoRF and its data are available as a web server and as a standalone program at: http://sparks-lab.org/jack/server/SPOT-MoRF/index.php.

Contact: jack.s.hanson93@gmail.com or yaoqi.zhou@griffith.edu.au

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Intrinsic protein disorder describes the tendency of a protein to fold into a dynamic and flexible structure, rather than the previous widely held belief that all proteins should fold into a unique stable structure to provide their functionality. While this dogma still applies to many structured regions in proteins, the increasing evidence of intrinsic disorder across all domains of life has vindicated this change of perspective (Hu *et al.*, 2018). Functional and structural analysis of intrinsically disordered proteins and regions of proteins (IDRs) has revealed the evolutionary roles of these proteins and how they stem from their structural dynamics (Tompa, 2002; Uversky *et al.*, 2000; Wright and Dyson, 1999).

Proteins with intrinsic disorder have many functional benefits made available due to their transient structural states, and as such are able to fulfill specific roles in biology, particularly in signaling, regulatory and assembling functions (Dyson and Wright, 2005; Tompa, 2003). Moreover, so-called 'hub' proteins with a large repertoire of protein–protein interactions have been shown to have more disordered content than other proteins (Haynes *et al.*, 2006; Hu *et al.*, 2017a). This stems from a particular advantage of disordered regions to interact with numerous target molecules and undergo a transition from disorder to an ordered state in a process known as induced folding (Dyson and Wright, 2002; Receveur-Bréchot *et al.*, 2005; Uversky, 2002). These short, induced folding regions, known as molecular recognition features (MoRFs) (Mohan *et al.*, 2006), provide an interesting insight into the disorder continuum, blurring the point at which a residue can be categorically considered 'disordered' or not.

The computational prediction of these important functional sites has attracted increased attention in recent literature because of their potential roles as druggable disease targets (Kumar *et al.*, 2017; Metallo, 2010). So far, however, these MoRF prediction methods have had limited accuracies mostly due to the large imbalance of MoRF/non-MoRF residues and small MoRF databases for training and testing. In the sample of experimental data available, disorder is a minority class. For example, an analysis of >10 000 proteins from the MobiDB consensus database shows only a ≈6.6% coverage of disorder (Necci *et al.*, 2017). Trickling down from the available annotated disordered regions, recent analysis has shown that only up to 21% of IDRs contain a MoRF region (Yan *et al.*, 2016), inherently imposing a severe limitation on the positive samples available for training predictive models. In fact, the inadequate number of MoRF-annotated sequences has limited MoRF predictors to small-scale computational and machine learning algorithms, such as: Support Vector Machines (Vapnik, 1998) in OPAL (Sharma *et al.*, 2018c), OPAL+ (Sharma *et al.*, 2018b), MoRFPred (Disfani *et al.*, 2012), MoRFPred-Plus (Sharma *et al.*, 2018a), MoRFchibi (Malhis *et al.*, 2016), MFPSSMpred (Chun *et al.*, 2013) and in Fang *et al.* (2018); neural networks in MoRF-MLP (He *et al.*, 2019); and through the analysis of residue physicochemical interactions in ANCHOR (Mészáros *et al.*, 2009, 2018). This is in stark contrast to many other fields of structural bioinformatics, where data-hungry methods such as large recurrent and convolutional neural networks have become state-of-the-art (Hanson *et al.*, 2019b), such as backbone *cis*-isomer detection (Singh *et al.*, 2018), contact map prediction (Wang *et al.*, 2017), and, most relevantly to this work, intrinsic disorder prediction (Hanson *et al.*, 2017, 2019a; Klausen *et al.*, 2019).

The existence of MoRFs in intrinsically disordered regions indicates that each protein can potentially contain up to three distinct structural regions in the disorder/order spectrum: structured regions, transitional intrinsically disordered MoRF regions and permanently intrinsically disordered regions. Similarly, Zhang *et al.* (2013) introduced the concept of 'semi-disorder' to designate these regions as neither fully disordered nor fully ordered. This semi-disordered state is a state separated from the structured state and the fully disordered state. Quantitative analysis indicates that semi-disordered residues are associated with protein aggregation and induced folding (Zhang *et al.*, 2013). The direct use of predicted disorder for binding site discrimination was also confirmed by DISOPRED3, which interpolated these binding site regions using an additional discriminative model on top of its disorder predictor (Jones and Cozzetto, 2015). In fact, we showed that the concept of the semi-disordered state can be used to predict MoRF regions with accuracy comparable to those methods dedicated for MoRF predictions when extracted linearly from the predicted likelihoods of disorder predictors (Hanson *et al.*, 2017, 2019a).

The above results suggest that disorder prediction and MoRF prediction are intrinsically connected with each other. In other words, the characterization of disordered regions learned from disorder prediction may be transferable to MoRF prediction. Transfer learning (Pan *et al.*, 2010) involves the repurposing of large models trained on one objective to a similar, secondary objective with insufficient data to train the large model from scratch (Goodfellow *et al.*, 2016). This is generally done by severing the culminating discriminatory layers [typically the final fully connected (FC) layer/s] of the initial model and replacing them with one or several untrained layers. By initializing the new model with an established, pre-learned internal representation of the original objective, training of the secondary objective can be greatly enhanced due to the model being placed in a much better location in the error plane than it would have been with random initialization.

Here, we have developed a method called SPOT-MoRF by applying transfer learning on the disorder prediction tool SPOT-Disorder2 (Hanson *et al.*, 2019a). This presents the first application of deep learning in MoRF prediction, hitherto limited to smaller-scale prediction tools due to the small training data pool. Building on the ensemble from SPOT-Disorder2, we show that the new method is substantially more accurate than direct training on MoRF prediction by randomly initialized neural networks and 14 previously developed techniques, highlighting the benefits of transfer learning for other smaller, niche fields in bioinformatics.

2 Methods and Data

2.1 Transfer learning

Transfer learning involves the extraction of a meaningful latent representation from a pretrained model to use for a new, similar objective. In this work, we use the learned internal representations of the current state-of-the-art protein intrinsic disorder predictor SPOT-Disorder2 (Hanson *et al.*, 2019a). SPOT-Disorder2 consists of an ensemble of both Inception-Residual-Squeeze and Excitation Networks (dubbed IncReSenets) (He *et al.*, 2016; Hu *et al.*, 2017b; Szegedy *et al.*, 2017) and Bidirectional Long Short-Term Memory (BLSTM) layers (Hochreiter and Schmidhuber, 1997; Schuster and Paliwal, 1997). We define layers consisting of convolutional or recurrent topologies as the 'contextual' layers of the model (i.e. the IncReSenet and BLSTM layers), and the concluding FC layers (Rumelhart *et al.*, 1986) as 'discriminative' layers.

In SPOT-Disorder2, it is the contextual layers which perform the bulk of the disorder characterization. BLSTM approaches have the ability to capture long-range dependencies through the enforcement of a constant error flow, which allows information to traverse through the sequence without being inhibited by affine weight transformations. On the other hand, deep convolutional neural network architectures, such as ResNets (He *et al.*, 2016) and Inception networks (Szegedy *et al.*, 2017), have found great success in many applications due to their relatively low parameter count despite their incredible network depth. This is largely due to the use of skip connections between layers which allow for unobstructed error propagation in network training, which act similarly to the constant error flow in LSTM cells. Squeeze and excitation networks are a slight modification applicable to these two convolutional neural network schema, in which another small set of weights is added which learns to intelligently control the contribution of each convolutional block to the skip connection (Hu *et al.*, 2017b). The simultaneous application of these architectures allows for both a wide and deep characterization of disorder, latently encoding MoRF regions. Additional architectural details are available in the accompanying manuscript (Hanson *et al.*, 2019a).

Generally speaking, SPOT-Disorder2 can be separated into a hierarchy of three general levels of data abstraction for the purpose of transfer learning: the output disorder likelihood, the FC layer directly preceding the final output layer and the final contextual layer (e.g. the final BLSTM or IncReSenet layer). Transferring the final output layer would be akin to using SPOT-Disorder2's predicted outputs as an input for a metapredictor, which does not transfer the model's latent characterization of disorder. Indeed, training with this transfer location led to poorer performance in initial trials. Therefore, we only considered the latter two of the transfer locations for these experiments, the layout of which is shown in Fig. 1. As this figure shows that the ensemble models of SPOT-Disorder2 are fed individually into a corresponding model for the use of ensemble MoRF prediction. These models are then averaged the same as SPOT-Disorder2 to provide our residue-wise ensemble MoRF prediction.

The models used in MoRF prediction all consist of one or several FC neural network layers placed at either the discriminative or contextual transfer point of SPOT-Disorder2. Explicitly, the MoRF models in Fig. 1 are independent from each other and are trained separately, but all utilize the same input data (SPOT-Disorder2's inputs) and similar architectures. The hyperparameters for each model are shown in Table 1. Dropout was used in most models at each layer (Srivastava *et al.*, 2014) during training (including the SPOT-Disorder2 layers). Each hidden layer is activated by the Rectified Linear Unit function (Nair and Hinton, 2010). The models were trained in Tensorflow v1.10 (Abadi *et al.*, 2016) using the in-built Adam optimizer (Kingma and Ba, 2014) with default parameters. The optimizer was switched to Stochastic Gradient Descent

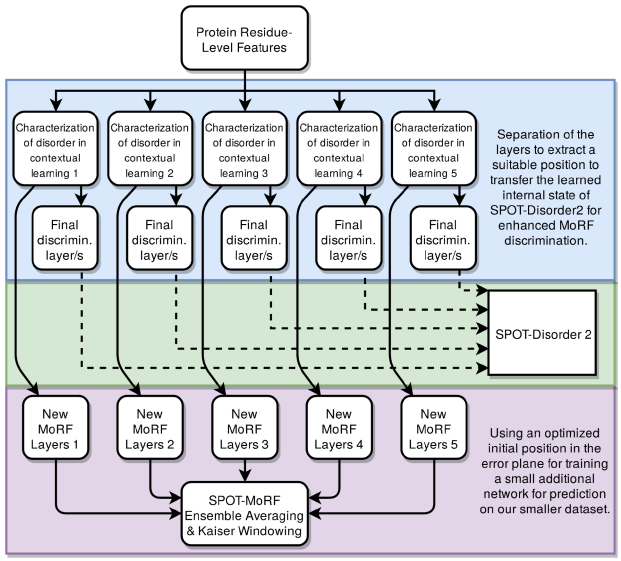

Fig. 1. The layout of our model, illustrating how the deep, latent disorder representations are transferred between SPOT-Disorder2 and our proposed method. Each trainable MoRF network consists of a FC network. In this figure, the characterization of disorder can be extracted from either the final contextual layer or the second-last discriminative (discrimin.) layer before the output of the disorder prediction model

Table 1. The hyperparameters of each model used in SPOT-MoRF

ID	L	N	Transfer Loc.	Dropout
0	3	100	Contextual	0.0
1	3	100	Discrim.	0.2
2	3	100	Contextual	0.2
3	3	100	Contextual	0.5
4	3	200	Contextual	0.5

Note: The ID also corresponds to the SPOT-Disorder2 model used in transfer learning. L, N and 'Dropout' refer to the number of FC layers appended to the severed SPOT-Disorder2 model, the number of neurons in these layers and the level of dropout applied, respectively.

with momentum (0.9) and a small learning rate (0.001) when the model stopped improving (after 10 epochs on the validation set) to reduce generalization errors between training and testing (Keskar and Socher, 2017). Early stopping was again used in this stage based on the validation set performance of the past five epochs to minimize the risk of overtraining on the small training dataset, with the highest-performing iteration being used for analysis. Note that in transfer learning the extracted weights from the transferred models can be frozen (i.e. not trained further), but in this experiment it was found empirically useful to adapt the trained networks for our purposes in training. Using this methodology, transferring from the contextual transfer point of model 0 from SPOT-Disorder2 (e.g.) provides our MoRF objective with 2.82×10^6 pretrained parameters. The new untrained parameter count of SPOT-MoRF model 0 is much lower, at 2.6×10^4.

The use of a Nvidia GTX 1080 Graphics Processing Unit allowed us to accelerate training over standard CPU training; each model taking only 45 s/epoch to train for each of our trained parameters in a grid search (SPOT-Disorder2 models 1–5, 1–3 layers, between 50 and 1500 nodes/layer, dropout 0–0.5 and 3 transfer locations). The final ensemble was selected based on the set which provided the highest ensemble performance on the validation set, not based on individual performance alone. This was done so that the selection method would prioritize orthogonal approaches over their baseline accuracy, providing the final model with a richer variety of learned MoRF representations.

Similarly to the MoRF prediction task in SPOT-Disorder2 (Hanson *et al.*, 2019a), we employ a post-processing smoothing window of size w_L on the output ensemble likelihoods y. However, rather than a rectangular window, we apply a Kaiser window ($\beta = 0.5$) to reduce the weight of longer-range samples (Kaiser, 1966). The following equation describes this process to obtain our final discrete output $\hat{y}$:

$$\hat{y}(i) = \begin{cases} 1 & \text{if } \left[\sum_{j=i-w_L}^{i+w_L} w(j-i) \times y(j)\right] \geq T \\ 0 & \text{elsewhere} \end{cases}, \quad (1)$$

where T is our operating threshold and w is our w_L-sized Kaiser window. The parameters w_L and T are chosen in validation. When $j < 0$ or $j \geq L$, where L is the protein length, values of 0 are substituted for $y(j)$.

2.2 Datasets

We utilize the training and testing dataset (Train, Test464) created by Disfani *et al.* (2012). Since its release, these datasets have been used to train and benchmark almost all subsequent machine-learning-based MoRF predictors (Malhis *et al.*, 2016; Sharma *et al.*, 2018a, c; Yan *et al.*, 2016). The proteins in Train and Test464 are extracted from proteins deposited in the PDB prior to 2008. Furthermore, we utilize the Exp53 (Exptest) set from Malhis *et al.* (2016), which consists of a set of 53 experimentally validated MoRF sequences. The distinction between experimental validation and the sets from Disfani *et al.* (2012) is that the Exptest MoRF regions are guaranteed to be *bona fide* MoRFs disordered in their natural state.

In this work, we updated the Test2012 set from Disfani *et al.* (2012) to include a much larger set of recently identified MoRFs, by following a similar protocol to Test2012's creation. We collected all protein–peptide complex structures from the BioLiP database (Yang *et al.*, 2012), accessed June 21, 2019 and mapped peptide sequences to UniProt IDs based on SIFTS annotations (Velankar *et al.*, 2012). We used a simple heuristic to filter out putative MoRFs which were likely to also be structured in the free state. All PDBs associated with the given UniProt IDs were collected and the structured residues were mapped to the full-length sequence by a sequential alignment with an affine gap penalty. Structured proteins which were not identical (95%) to the aligned region of the full-length sequence (e.g. chimeric proteins) were ignored. Putative MoRFs which overlapped a structured region >5 residues longer than all MoRFs associated with that protein were assumed to be structured in the free state and excluded from the dataset. We found that the labels generated by our protocol were mostly consistent with Test2012 except for cases where the peptide fragment was mapped to different full-length sequences (e.g. 3unnB is mapped to Q14676 by our protocol which does not match the Test2012 sequence). We elected to use the SIFTS PDB mapping as it is an independent protocol which avoids the problem of ambiguous full-length sequences. Six of the original Test2012 peptides were also removed from the dataset as they were judged to be synthetic peptides based on the SIFTS annotations (e.g. 3avcD). Finally, we removed proteins of length <31 residues and >5000 residues due to the length constraints of other predictors.

In this work, we utilize Exptest and Test2019 as independent tests and Test464 as a validation set (henceforth dubbed 'Validation' to avoid confusion). While Validation has already been clustered to have <30% sequence similarity to the training set, it contains a large number of homologous sequences. In fact, more than a third of the proteins in Validation share >90% sequence similarity to another sequence within the set (Malhis *et al.*, 2016). Thus, removing this intra-dataset homology will limit the biasing of our results towards certain sequence clusters prominent in the data. Furthermore, to ensure our testing is blind and unbiased, we also cluster our Validation and testing sets against the training and validation sets used in both SPOT-Disorder2 and SPOT-MoRF. This clustering is performed at 30% sequence identity using Blastclust (Altschul *et al.*, 1997). We do not cluster the Train set to preserve the number of training samples.

Table 2. An overview of the datasets used in this work and their residue-wise MoRF/non-MoRF propensities

Dataset	Sequences	MoRFs	Non-MoRFs	MoRF ratio (%)
Train	394	4977	169 604	2.93
Validation	265	3264	116 041	2.74
Exptest	39	1812	12 482	12.68
Test2019	850	14 643	437 798	3.24

Finally, due to limitations introduced by the large computational pipeline for SPOT-Disorder2 (namely the two-dimensional prediction of a contact map in SPOT-Contact), we remove all proteins of length >1500 (Hanson *et al.*, 2018a). Despite this limitation, statistics from the latest UniProtKB version (June 2019) indicate that >99% of proteins are eligible for prediction by our method (UniProt, 2014). After clustering and thresholding sequence length, our Train, Validation, Exptest and Test2019 sets contain 394, 265, 39 and 850 sequences, respectively. The MoRF content of each dataset is shown in Table 2.

2.3 Performance evaluation

Providing a balanced analysis of MoRF predictions is key to obtaining a true representation of the prediction performance. This is particularly important in MoRF prediction as MoRF residues account for only 3% of residues across our Train, Validation and Test2019 datasets. The first skew-independent metrics can be obtained through the use of binary classification analysis to obtain the True Positives (TP), False Positives (FP), True Negatives (TN) and False Negatives (FN), as determined by a classification threshold on the model's predicted likelihoods and their corresponding labels. These metrics can be combined to find the sensitivity ($\text{Sens} = \frac{\text{TP}}{\text{TP+FN}}$) and specificity ($\text{Spec} = \frac{\text{TN}}{\text{TN+FP}}$), the class-dependent accuracies of class 1 (MoRF) and class 0 (non-MoRF), respectively. We can also find the accuracy of the positive predictions of a model through the precision metric ($\text{Prec} = \frac{\text{TP}}{\text{TP+FP}}$). This metric is particularly important as minor decreases in specificity can flood the performance of a model with FP's due to the large class disparity, which would otherwise be unnoticed without Precision analysis. To select a threshold for our predictor, we use the value which maximizes the Matthew's Correlation Coefficient (MCC) on the validation set (Matthews, 1975).

To get a singular metric for overall performance analysis, we can use the Area Under the Curve of the Receiver Operating Characteristic (AUC_{ROC}) (Hanley and McNeil, 1982) or Precision–Recall (AUC_{PR}) curves (Davis and Goadrich, 2006).

2.4 Method comparison

We compare to a set of previously released MoRF prediction tools. We downloaded the standalone version of MoRFPred-Plus (Sharma *et al.*, 2018a) (Available: https://github.com/roneshsharma/MoRFpred-plus), OPAL, OPAL+ and PROMIS (Sharma *et al.*, 2018b, c) (Available: https://github.com/roneshsharma? tab=repositories) and DISOPRED3 (Jones and Cozzetto, 2015) (Available: http://bioinf.cs.ucl.ac.uk/psipred/). We used the web servers of MoRFPred (Disfani *et al.*, 2012), fMoRFPred (Yan *et al.*, 2016), DisoRDPbind (Peng and Kurgan, 2015) (Server URL: http://biomine.cs.vcu.edu/#webservers), ANCHOR2 (Mészáros *et al.*, 2018) (Server URL: https://iupred2a.elte.hu/) and through the RESTful interface of MoRFchibi (Malhis *et al.*, 2016) (Server URL: https://gsponerlab.msl.ubc.ca/software/morf_chibi/). We further separate the predictions from MoRFchibi into the three provided flavors: MoRFchibi, MoRFchibi-Web and MoRFchibi-Lite. We also use the same dual-thresholding method for binding and MoRF regions from SPOT-Disorder and SPOT-Disorder2, respectively (Hanson *et al.*, 2017, 2019a). For completeness, we would like to mention that a new method called MoRF-MLP was recently published (He *et al.*, 2019). We were unable to make a direct comparison between our method and this new method because it does not have a standalone package or an online server. It, however, performs worse than OPAL and PROMIS which are compared here.

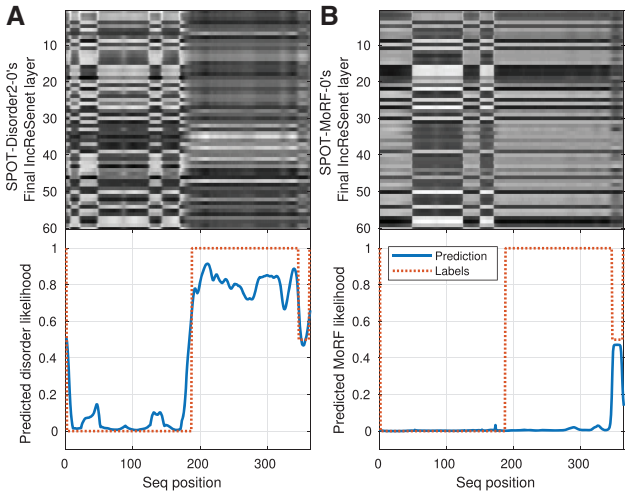

Fig. 2. The activations of the transfer location (top) and model outputs (bottom) from (**A**) SPOT-Disorder2 model 0 and (**B**) SPOT-MoRF model 0 for protein P53990 from Test2019. Labels of 0, 0.5 and 1 correspond to order, MoRF and disordered regions, respectively

3 Results

As a proof of concept for the use of transfer learning in this work, we can analyze the activations and outputs of SPOT-Disorder2 to see if there is any correlation between the output disorder prediction and any MoRF component for an input protein. In Fig. 2A, we present the activations of the transfer point and outputs of SPOT-Disorder2 model 0 for the protein P53990 from Test2019, which has annotated ordered, disordered and MoRF regions (represented by labels of 0, 1 and 0.5 in the bottom graph, respectively). As this graph shows, there is a direct correlation between certain patterns in the activation of the contextual layer and the high- and low-ranked predictions of disorder. Pertinently to this work, however, there is a direct match of uncertainty in the disorder predictions and the MoRF region at residues 347–362. This is identified by SPOT-MoRF model 0, as shown in Fig. 2B, where the output spikes at this predicted semi-disorder region, correctly identifying the MoRF and predicting both the ordered and disordered regions as non-MoRFs. This despite the somewhat similar activations of the contextual layer in SPOT-MoRF and SPOT-Disorder2.

To illustrate the usefulness of our transfer learning approach, we compare several neural-network-based approaches to MoRF prediction. In this analysis, we compare the proposed approach to a directly trained model utilizing the same architecture as model 0 from SPOT-Disorder2 and ensembles of FC (MoRF-FC) and BLSTM (MoRF-LSTM) networks utilizing the inputs and outputs of each corresponding SPOT-Disorder2 component model. Early stopping was used to avoid excessive overtraining due to training such large networks on small datasets, particularly for the directly trained method. All of these methods use the same window as SPOT-MoRF, shown in Eq. (1). A baseline performance is provided by the linear extraction of MoRF residues from the outputs of SPOT-Disorder2 using the dual-thresholding method as per Hanson *et al.* (2019a). The results of these methods on the Validation, Exptest and Test2019 sets are shown alongside SPOT-MoRF in Table 3. The directly trained model is shown to be substantially inferior in all analyzed metrics (surprisingly bar AUC_{ROC} in Validation) to its corresponding singular SPOT-MoRF model utilizing transfer learning, illustrating that deep methods cannot effectively be used in this problem without suitable weight initialization (i.e. weight transfer). This is particularly the case in the Test2019 set, where transfer

Table 3. A comparison of the proposed SPOT-MoRF approach to other machine learning methods for the Validation, Test and Exptest sets

Method	Validation			Exptest			Test2019		
	AUC_{ROC}	AUC_{PR}	MCC	AUC_{ROC}	AUC_{PR}	MCC	AUC_{ROC}	AUC_{PR}	MCC
Direct training	0.756	0.157	0.214	0.731	0.299	0.225	0.745	0.108	0.118
SPOT-MoRF (model 0)	0.755	0.184	0.249	0.789	0.387	0.316	0.786	0.160	0.211
SPOT-Disorder2	—	—	0.123	—	—	0.260	—	—	0.127
MoRF-FC	0.719	0.087	0.144	0.681	0.211	0.173	0.676	0.073	0.086
MoRF-LSTM	0.766	0.165	0.232	0.750	0.323	0.217	0.819	0.174	0.224
SPOT-MoRF	0.787	0.227	0.282	0.813	0.421	0.325	0.835	0.208	0.253

Note: The results are separated into singular methods based on model 0 (top) and ensemble (bottom) methods.

learning outperforms the direct model by 78% in MCC and 47% in AUC_{PR}. Moreover, the proposed SPOT-MoRF method is shown to substantially outperform its FC and LSTM ensemble counterparts, vindicating this technique as a valid method for predicting niche structural properties in proteins from small datasets. The LSTM-based method MoRF-LSTM performs better than MoRF-FC due to its use of contextual information, but is itself outperformed by SPOT-MoRF by 13% in MCC and 19% in AUC_{PR} on the Exptest set. Interestingly, SPOT-Disorder2 considerably outperforms MoRF-FC and MoRF-LSTM on the Exptest set, indicating that linear discrimination of MoRF residues performs better than a randomly initialized complex neural network for this data.

It is interest to observe how substantially the performance of a singular model can be improved through the use of an ensemble. As such, Supplementary Table S1 shows the individual performance of each ensemble component model. All models have a consistent improvement from the Validation set to the Exptest and Test2019 test. One strength of the ensemble approach is the ability of the model to inherently adapt to different distributions of MoRF residues, exemplified in the component models' varying performances in the Test2019 (lower MoRF propensity) and Exptest (higher MoRF propensity) sets. Overall, the benefit of the ensemble is particularly noticeable in the AUC_{PR} of each dataset, with the ensemble outperforming the next-best model by 23%, 7% and 17% in the Validation, Exptest and Test2019 sets, respectively, indicating that the rare MoRF class can become more distinct when individual models' generalization errors are removed in the ensemble averaging.

3.1 Comparison of SPOT-MoRF to other predictors

We compare the results of SPOT-MoRF to the 14 other current predictors described in Section 2.4 on the Exptest set in Table 4. This set contains proteins with experimentally validated MoRFs, with some proteins containing multiple MoRF regions. Thus, the precision of the models in this dataset will be generally higher than previous results from Test464 and Test2012 due to the removal of potentially false-negative labels. SPOT-MoRF achieves the highest AUC_{ROC} and AUC_{PR} by 2% and 11%, respectively, and attains the joint-highest MCC with MoRFchibi-Web of 0.33 (a 3% improvement on the next-best). Even on this small dataset, the improvement in AUC_{ROC} over all methods except OPAL+ (P-value <0.023) is statistically significant with a P-value of $\leq 3 \times 10^{-4}$. Moreover, it must be noted that OPAL and PROMIS were validated (fine-tuned) on this set (Sharma *et al.*, 2018c). Supplementary Fig. S1 shows the ROC and PR curves of all compared single-threshold models on the Exptest set. SPOT-MoRF provides a more precise MoRF prediction for both low and high sensitivity values, and is only lower than any other methods between sensitivities of 0.15–0.35. The operating threshold of SPOT-MoRF reflects this, obtaining the second-highest operating precision of 64%, a 36% relative increase on MoRFchibi-Web despite its comparable MCC. Furthermore, the ROC curve in Supplementary Fig. S1 shows that SPOT-MoRF obtains a higher sensitivity for almost all specificities. The high performance in Exptest by SPOT-MoRF is significant because the fractions of MoRFs in Train and Validation (2–3%) are so different to the MoRF content in Exptest (12.7%, Table 2), indicating robust training.

Table 5 shows the performance of the compared predictors on the Test2019 set. This set provides an updated performance analysis for the compared predictors, as most have only been benchmarked with proteins deposited in the PDB prior to 2013. SPOT-MoRF obtains the highest performance across all analyzed metrics, soundly beating the next-best OPAL+ in AUC_{ROC}, AUC_{PR} and MCC by 0.835 to 0.805, 0.208 to 0.157 and 0.253 to 0.192, respectively. The difference in AUC_{ROC} is statistically significant, obtaining a P-value of $\leq 1 \times 10^{-7}$. The ROC and PR curves for this dataset are shown in Fig. 3. In both graphs, SPOT-MoRF shows a clear superiority to the other methods at all sensitivities (except for the region of <5% sensitivity to PROMIS in the PR curve). The gap between methods in the PR curve is reflected in SPOT-MoRF obtaining a much higher operating precision at a similar sensitivity to other methods (a 59% improvement in precision over the next-best OPAL+).

Table 4. Performance comparison on the Exptest set

Predictor	AUC_{ROC}	AUC_{PR}	MCC	Se	Sp	Pr
ANCHOR2	0.548	0.130	0.039	44.48	61.21	14.27
DISOPRED3	0.545	0.159	0.069	16.28	90.12	19.31
MoRFPred	0.615	0.186	0.106	15.51	93.13	24.69
DisoRDPbind	0.628	0.205	0.140	38.85	78.97	21.15
fMoRFPred	0.656	0.219	0.097	7.89	97.35	30.17
MoRFPred-Plus	0.675	0.234	0.189	53.20	72.96	22.22
MoRFchibi	0.704	0.306	0.206	9.99	98.98	58.77
PROMIS	0.757	0.316	0.287	48.34	85.37	32.42
OPAL+	0.795	0.368	0.288	29.03	94.78	44.65
OPAL	0.782	0.376	0.281	68.60	71.37	25.81
MoRFchibi-Lite	0.768	0.379	0.316	37.80	92.41	41.97
MoRFchibi-Web	0.754	0.379	0.328	40.95	91.65	41.59
SPOT-MoRF	0.813	0.421	0.325	25.33	97.21	56.88
SPOT-Disorder	—	—	0.185	35.82	85.27	26.10
SPOT-Disorder2	—	—	0.260	77.15	61.58	22.57

Table 5. Performance comparison on the Test2019 set

Predictor	AUC_{ROC}	AUC_{PR}	MCC	Se	Sp	Pr
DisoRDPbind	0.606	0.045	0.033	15.22	90.36	5.02
DISOPRED3	0.543	0.046	0.037	11.96	93.36	5.68
ANCHOR2	0.620	0.049	0.071	45.84	72.25	5.24
MoRFPred[a]	0.647	0.069	0.092	23.93	91.14	8.29
MoRFchibi	0.682	0.074	0.062	4.15	99.18	14.52
fMoRFPred	0.660	0.080	0.092	9.20	98.17	14.40
MoRFPred-Plus	0.762	0.114	0.156	64.71	74.46	7.81
MoRFchibi-Lite	0.773	0.126	0.155	24.65	95.19	14.63
MoRFchibi-Web	0.793	0.142	0.190	36.08	92.96	14.63
OPAL	0.788	0.152	0.190	49.77	87.37	11.64
PROMIS	0.784	0.154	0.183	42.15	90.10	12.47
OPAL+	0.805	0.157	0.192	25.86	96.33	19.05
SPOT-MoRF	0.835	0.208	0.253	24.94	98.09	30.35
SPOT-Disorder	—	—	0.089	30.76	86.69	7.18
SPOT-Disorder2	—	—	0.127	71.71	63.17	6.11

[a]Reported on 849 proteins due to server error (missing Q2HR73).

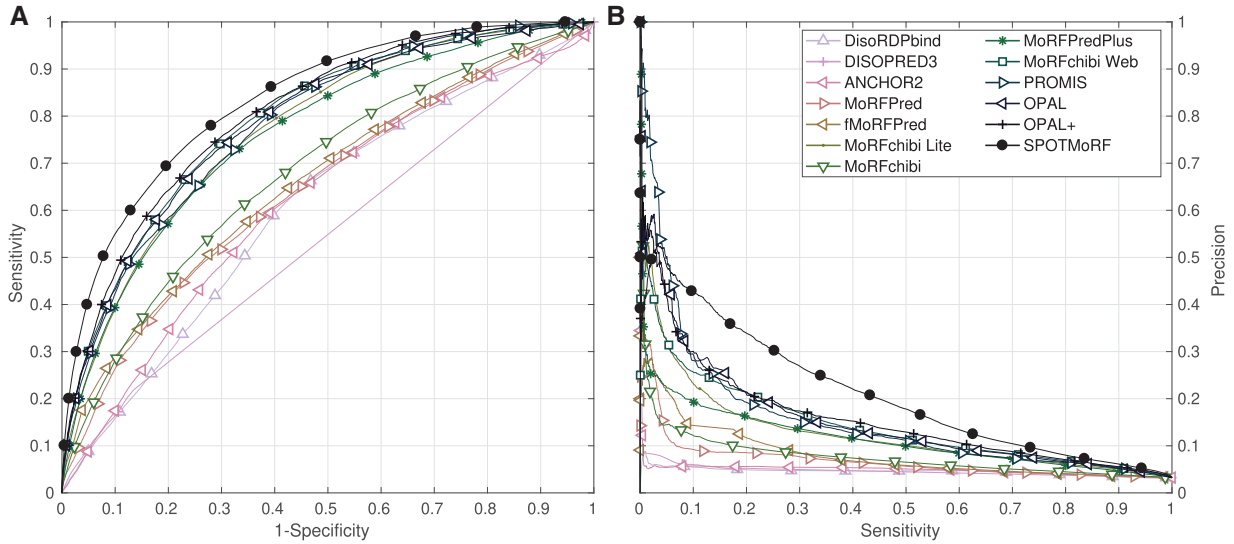

Fig. 3. The Receiver-Operating Characteristic and Precision–Recall curves in (**A**) and (**B**), respectively, for the Test2019 set by SPOT-MoRF and other methods as labeled

As we use the original Test464 set as a validation set in this work, it is important to gauge whether or not the improvement of our method over others is due to our transfer learning approach rather than the effectively larger seen data pool. To analyze this, we extracted a secondary validation set of 40 proteins from our Train set containing <30% sequence similarity to each other and the Train set according to Blastclust, reducing our Train set to 314 proteins. This train set is comparable in size to the training set used during 5-fold cross validation on 421 proteins in previous work, and smaller than those works which use the full set (Disfani *et al.*, 2012; Malhis *et al.*, 2016; Sharma *et al.*, 2018c). We then retrained our ensemble on the new training and validation data, utilizing the same training procedure and hyperparameters. The performance of the model across both of our test sets is only lower when compared to SPOT-MoRF, obtaining an AUC_{ROC} and AUC_{PR} score of 0.810 and 0.406 on the Exptest set and 0.826 and 0.181 on the Test2019 set, respectively. These scores are lower than the results obtained by SPOT-MoRF in Tables 4 and 5, but still higher than the competing methods (e.g. 16% improvement in AUC_{PR} over the second-best OPAL+ on the Test2019 set), suggesting that the bulk of the improvement of our method over others stems from the transfer learning approach rather than the use of an extended validation set.

4 Discussion

In this work, we proposed SPOT-MoRF, a prediction tool built by repurposing an accurate protein intrinsic disorder predictor to the prediction of MoRFs in proteins. This presents the first MoRF predictor built utilizing deep learning, achieving state-of-the-art results on two independent datasets while being trained using the same training data as almost all other predictors compared. The use of an ensemble s*et al*lows the method to make more confident predictions due to the removal of spurious false generalizations in singular component models, reflected in its increased AUC_{PR} over other methods. Following from this, our model also achieves high MCCs across all datasets, meaning that the model predicts with a high level of MoRF/non-MoRF separation for each model's prescribed threshold. This is consistent over three independent testing datasets with varying levels of MoRF content, indicating a robust performance afforded by its deep learning foundation.

One challenge in MoRF prediction is the lack of large annotated datasets for training and testing. The established test sets in the literature, Test2012 and Exp53, contain only 45 and 53 sequences, respectively (Malhis *et al.*, 2016). To address this gap in the literature, we have created a much larger test set Test2019 which consists of 930 (850 with sequence length $\leq$ 1500) new MoRF-annotated proteins. Test2019 provides a much more thorough and current analysis than the currently used datasets in the literature, and is formed using a similar protocol to its predecessor, Test2012. The higher performance of SPOT-MoRF on this dataset points to a more consistent prediction quality of the proposed method for unseen data than other predictors. This is after clustering each of our testing sets against the validation and training sets of both SPOT-MoRF and SPOT-Disorder2 at 30% similarity according to Blastclust.

Despite its strengths, this method also presents several drawbacks potentially limiting usage. First of all, the large pipeline for SPOT-MoRF, which contains two evolutionary profile generation tools (HHBlits and PSI-Blast) (Altschul *et al.*, 1997; Remmert *et al.*, 2012) and large computational prediction tools for contact map and 1-D structural property prediction (Hanson *et al.*, 2018a, 2019c), can be quite slow, especially compared to much faster prediction methods such as fMoRFpred and MoRFchibi. Furthermore, the use of SPOT-Contact in the SPOT-MoRF pipeline limits the length of input proteins due to the two-dimensional prediction space, depending on the computational capabilities of the user's workstation. These problems can be alleviated by utilizing a smaller secondary structure predictor in place of SPOT-1D at the inputs of SPOT-MoRF. To address this shortcoming, we can calculate the results on the 83 long proteins ($\geq$1500 residues) omitted from all of our test datasets (3 from Exptest and 80 from Test2019) by using the modified outputs of the secondary structure prediction method SPIDER3 (Heffernan *et al.*, 2017). The proteins in this long set have an average sequence length of 2196 and MoRF coverage of 0.95%. Supplementary Table S2 shows that without the use of SPOT-1D, SPOT-MoRF can still perform accurately, obtaining the highest AUC_{PR} for this dataset and obtaining the second-highest AUC_{ROC} (0.781 versus 0.822 of MoRFchibi-Web) and second-highest MCC values (0.092 versus 0.109 of MoRFchibi-Web), even without being specifically trained for detecting such sparse MoRF regions. We have implemented this feature as an option in our software package for local usage. Another option to potentially make this process more efficient is to apply transfer learning to our single-sequence intrinsic disorder prediction tool SPOT-Disorder-Single (Hanson *et al.*, 2018b). This work is in progress.

Funding

This work was supported by the Australian 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).

Conflict of Interest: none declared.

References

Abadi,M. *et al*. (2016) Tensorflow: large-scale machine learning on heterogeneous distributed systems. *CoRR*, **abs/1603.04467.**

Altschul,S.F. *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3389–3402.

Chun,F. *et al*. (2013) Sequence-based prediction of molecular recognition features in disordered proteins. *J. Med. Bioeng*., **2**, 110–114.

Davis,J. and Goadrich,M. (2006). The relationship between precision-recall and roc curves. In *Proceedings of the 23rd International Conference on Machine Learning*, pp. 233–240. ACM.

Disfani,F.M. *et al*. (2012) MoRFpred, a computational tool for sequence-based prediction and characterization of short disorder-to-order transitioning binding regions in proteins. *Bioinformatics*, **28**, i75–i83.

Dyson,H.J. and Wright,P.E. (2002) Coupling of folding and binding for unstructured proteins. *Curr. Opin. Struct. Biol*., **12**, 54–60.

Dyson,H.J. and Wright,P.E. (2005) Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol*., **6**, 197–208.

Fang,C. *et al*. (2018) Identifying MoRFs in disordered proteins using enlarged conserved features. In: *Proceedings of the 2018 6th International Conference on Bioinformatics and Computational Biology*, pp. 50–54. ACM.

Goodfellow,I. *et al*. (2016) *Deep Learning*, Vol. **1**. MIT Press, Cambridge, MA, USA.

Hanley,J.A. and McNeil,B.J. (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, **143**, 29–36.

Hanson,J. *et al*. (2017) Improving protein disorder prediction by deep bidirectional long short-term memory recurrent neural networks. *Bioinformatics*, **33**, 685–694.

Hanson,J. *et al*. (2018a) Accurate prediction of protein contact maps by coupling residual two-dimensional bidirectional long short-term memory with convolutional neural networks. *Bioinformatics*, **34**, 4039–4045.

Hanson,J. *et al*. (2018b) Accurate single-sequence prediction of protein intrinsic disorder by an ensemble of deep recurrent and convolutional architectures. *J. Chem. Inf. Model*., **58**, 2369–2376.

Hanson,J. *et al*. (2019a) Enhancing protein intrinsic disorder prediction by utilizing deep squeeze and excitation residual inception and long short-term memory networks. *Genom. Proteom. Bioinf*., in press.

Hanson,J. *et al*. (2019b) Getting to know your neighbor: protein structure prediction comes of age with contextual machine learning. *J. Comput. Biol*., in press. doi: 10.1089/cmb.2019.0193.

Hanson,J. *et al*. (2019c) Improving prediction of protein secondary structure, backbone angles, solvent accessibility, and contact numbers by using predicted contact maps and an ensemble of recurrent and residual convolutional neural networks. *Bioinformatics*, **35**, 2403–2410.

Haynes,C. *et al*. (2006) Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes. *PLoS Comput. Biol*., **2**, e100.

He,K. *et al*. (2016) Deep residual learning for image recognition. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778.

He,H. *et al*. (2019) Prediction of MoRFs in protein sequences with MLPs based on sequence properties and evolution information. *Entropy*, **21**, 635.

Heffernan,R. *et al*. (2017) Capturing non-local interactions by long short term memory bidirectional recurrent neural networks for improving prediction of protein secondary structure. *Bioinformatics*, **33**, 2842–2849.

Hochreiter,S. and Schmidhuber,J. (1997) Long short-term memory. *Neural Comput*., **9**, 1735–1780.

Hu,G. *et al*. (2017a) Functional analysis of human hub proteins and their interactors involved in the intrinsic disorder-enriched interactions. *Int. J. Mol. Sci*., **18**, 2761.

Hu,J. *et al*. (2017b) Squeeze-and-excitation networks. arXiv:1709.01507.

Hu,G. *et al*. (2018) Taxonomic landscape of the dark proteomes: whole-proteome scale interplay between structural darkness, intrinsic disorder, and crystallization propensity. *Proteomics*, **18**, 1800243.

Jones,D.T. and Cozzetto,D. (2015) DISOPRED3: precise disordered region predictions with annotated protein-binding activity. *Bioinformatics*, **31**, 857–863.

Kaiser,J. (1966) Digital filters. In *System Analysis by Digital Computer*, Chapter 7. Wiley, NY, USA.

Keskar,N.S. and Socher,R. (2017) Improving generalization performance by switching from Adam to SGD. arXiv:1712.07628.

Kingma,D.P. and Ba,J. (2014) Adam: a method for stochastic optimization. *CoRR*, abs/1412.6980.

Klausen,M.S. *et al*. (2019) NetSurfP-2.0: improved prediction of protein structural features by integrated deep learning. *Proteins*, **87**, 520.

Kumar,D. *et al*. (2017) Therapeutic interventions of cancers using intrinsically disordered proteins as drug targets: c-myc as model system. *Cancer Inf*., **16**, 1176935117699408.

Malhis,N. *et al*. (2016) MoRFchibi SYSTEM: software tools for the identification of MoRFs in protein sequences. *Nucleic Acids Res*., **44**, W488–W493.

Matthews,B.W. (1975) Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochim. Biophys. Acta*, **405**, 442–451.

Mészáros,B. *et al*. (2009) Prediction of protein binding regions in disordered proteins. *PLoS Comput. Biol*., **5**, e1000376.

Mészáros,B. *et al*. (2018) IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Res*., **46**, W329–W337.[10.1093/nar/gky384].

Metallo,S.J. (2010) Intrinsically disordered proteins are potential drug targets. *Curr. Opin. Chem. Biol*., **14**, 481–488.

Mohan,A. *et al*. (2006) Analysis of molecular recognition features (MoRFS). *J. Mol. Biol*., **362**, 1043–1059.

Nair,V. and Hinton,G.E. (2010) Rectified linear units improve restricted Boltzmann machines. In: *Proceedings of the 27th International Conference on Machine Learning*, pp. 807–814.

Necci,M. *et al*. (2017) MobiDB-lite: fast and highly specific consensus prediction of intrinsic disorder in proteins. *Bioinformatics*, **33**, 1402–1404.

Pan,S.J. *et al*. (2010) A survey on transfer learning. *IEEE Trans. Knowl. Data Eng*., **22**, 1345–1359.

Peng,Z. and Kurgan,L. (2015) High-throughput prediction of RNA, DNA and protein binding regions mediated by intrinsic disorder. *Nucleic Acids Res*., **43**, e121.

Receveur-Bréchot,V. *et al*. (2005) Assessing protein disorder and induced folding. *Proteins*, **62**, 24–45.

Remmert,M. *et al*. (2012) HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods*, **9**, 173–175.

Rumelhart,D.E. *et al*. (1986) Learning representations by back-propagating errors. *Nature*, **323**, 533.

Schuster,M. and Paliwal,K.K. (1997) Bidirectional recurrent neural networks. *IEEE Trans. Signal Process*., **45**, 2673–2681.

Sharma,R. *et al*. (2018a) Morfpred-plus: computational identification of MoRFS in protein sequences using physicochemical properties and HMM profiles. *J. Theor. Biol*., **437**, 9–16.

Sharma,R. *et al*. (2018b) OPAL+: length-specific MoRF prediction in intrinsically disordered protein sequences. *Proteomics*, **19**, e1800058.

Sharma,R. *et al*. (2018c) OPAL: prediction of MoRF regions in intrinsically disordered protein sequences. *Bioinformatics*, **34**, 1850–1858.

Singh,J. *et al*. (2018) Detecting proline and non-proline cis isomers in protein structures from sequences using deep residual ensemble learning. *J. Chem. Inf. Model*., **58**, 2033–2042.

Srivastava,N. *et al*. (2014) Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res*., **15**, 1929–1958.

Szegedy,C. *et al*. (2017) Inception-v4, inception-ReSnet and the impact of residual connections on learning. In *AAAI*, Vol. **4**, AAAI Press, San Francisco, California, p. 12.

Tompa,P. (2002) Intrinsically unstructured proteins. *Trends Biochem. Sci*., **27**, 527–533.

Tompa,P. (2003) The functional benefits of protein disorder. *J. Mol. Struct. Theochem*, **666**, 361–371.

UniProt (2014) UniProt: a hub for protein information. *Nucleic Acids Res*., **43**, D204–D212.

Uversky,V.N. (2002) What does it mean to be natively unfolded? *Eur. J. Biochem*., **269**, 2–12.

Uversky,V.N. *et al*. (2000) Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins*, **41**, 415–427.

Vapnik,V.N. (1998) *Statistical Learning Theory*, Vol. **1**. Wiley, NY, USA.

Velankar,S. *et al*. (2012) Sifts: structure integration with function, taxonomy and sequences resource. *Nucleic Acids Res*., **41**, D483–D489.

Wang,S. *et al*. (2017) Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLoS Comput. Biol*., **13**, 1–34.

Wright,P.E. and Dyson,H.J. (1999) Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J. Mol. Biol*., **293**, 321–331.

Yan,J. *et al*. (2016) Molecular recognition features (MoRFs) in three domains of life. *Mol. Biosyst*., **12**, 697–710.

Yang,J. *et al*. (2012) BioLiP: a semi-manually curated database for biologically relevant ligand–protein interactions. *Nucleic Acids Res*., **41**, D1096–D1103.

Zhang,T. *et al*. (2013) Intrinsically semi-disordered state and its role in induced folding and protein aggregation. *Cell. Biochem. Biophys*., **67**, 1193–1205.