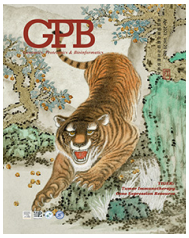
Genomics Proteomics Bioinformatics 21 (2023) 359–369

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com



METHOD

TransDFL: Identification of Disordered Flexible Linkers in Proteins by Transfer Learning

Yihe Pang [1], **Bin Liu** [1,2,*]

[1] *School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China*
[2] *Advanced Research Institute of Multidisciplinary Science, Beijing Institute of Technology, Beijing 100081, China*

Received 13 December 2021; revised 21 September 2022; accepted 14 October 2022
Available online 19 October 2022

Handled by Yu Xue

KEYWORDS

Intrinsically disordered protein;
Disordered flexible linker;
False positive rate;
Computational predictor;
Transfer learning

Abstract **Disordered flexible linkers** (DFLs) are the functional disordered regions in proteins, which are the sub-regions of intrinsically disordered regions (IDRs) and play important roles in connecting domains and maintaining inter-domain interactions. Trained with the limited available DFLs, the existing DFL predictors based on the machine learning techniques tend to predict the ordered residues as DFLs, leading to a high **false positive rate** (FPR) and low prediction accuracy. Previous studies have shown that DFLs are extremely flexible disordered regions, which are usually predicted as disordered residues with high confidence [$P(D) > 0.9$] by an IDR predictor. Therefore, transferring an IDR predictor to an accurate DFL predictor is of great significance for understanding the functions of IDRs. In this study, we proposed a new predictor called TransDFL for identifying DFLs by transferring the RFPR-IDP predictor for IDR identification to the DFL prediction. The RFPR-IDP was pre-trained with IDR sequences to learn the general features between IDRs and DFLs, which is helpful to reduce the false positives in the ordered regions. RFPR-IDP was fine-tuned with the DFL sequences to capture the specific features of DFLs so as to be transferred into the TransDFL. Experimental results of two application scenarios (prediction of DFLs only in IDRs or prediction of DFLs in entire proteins) showed that TransDFL consistently outperformed other existing DFL predictors with higher accuracy. The corresponding web server of TransDFL can be freely accessed at http://bliulab.net/TransDFL/.

Introduction

Intrinsically disordered regions (IDRs) are protein regions without stable three-dimensional (3D) structures, which are particularly common among eukaryotic organisms and viral proteomes [1]. Although the IDRs lack well-defined 3D structures, they carry out many critical functions, such as transcriptions, signal transmission, post-translational modifications, and multi-protein aggregation [2]. The functions of IDRs

* Corresponding author.
E-mail: bliu@bliulab.net (Liu B).

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation and Genetics Society of China.

https://doi.org/10.1016/j.gpb.2022.10.004

derive either from binding to molecular partners (such as DNA, RNA, and proteins) or directly from their native disordered states, where the former is called binding functions and the latter is called non-binding functions [3]. According to the DisProt database [4], about 75% of the non-binding functions are disordered flexible linkers (DFLs) [5,6]. DFLs serve as the linkers in multidomain proteins characterized by extremely structural flexibility, and can be located between inter- and intra-domain, which are different from generic linkers [5,7–10]. DFLs play essential roles for intramolecular allosteric regulation [2,11] and phase separation [12]. Identification of DFLs is crucial for comprehensively studying IDR functions. Experimental annotation of DFLs primarily relies on X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, and circular dichroism. In order to efficiently identify the DFLs, two computational methods have been developed only based on the protein sequences, including DFLpred [5] and APOD [6]. DFLpred identifies the DFLs via combining the logistic regression (LR) and four sequence-based features, including structure domain propensities, putative disordered regions, and two properties of spiral and turn formation. APOD incorporates various sequence profile features into support vector machines (SVMs) to further improve the predictive performance, such as evolutionary conservation and relative solvent accessible area.

These predictors successfully incorporate various sequence profile features for DFL prediction. However, DFLs are continuous regions in proteins, sharing global sequence patterns along the whole protein [5]. The global features of DFLs should be incorporated into the DFL predictors. Furthermore, DFLs are the sub-regions of IDRs, while sequences with unannotated disordered regions are common in nature [13–15]. As a result, DFL predictors tend to predict the ordered residues as DFLs, resulting in high false positive rate (FPR) and low prediction accuracy.

According to the recent Critical Assessment of protein Intrinsic Disorder prediction (CAID) experiment reports [16], great efforts have been made by researchers for the development of IDR predictors. **Figure 1** shows DFL prediction results on the DFL dataset TE82 [6] predicted by six state-of-the-art IDR predictors, including AUCpreD [17], SPINE-D [18], DISOPRED3 [19], SPOT-Disorder [20], IDP-Seq2Seq [15], and SPOT-Disorder2 [14]. We can see that DFLs can be predicted with high disordered probabilities [*i.e.*, extremely disordered state $P(D) > 0.9$] by different IDR predictors, providing an opportunity to predict the DFLs based on an IDR predictor (Figure 1A).

The information of disordered regions and functions are both encoded in their primary sequences, similar to the source language and target language sharing the same semantic [21] in the field of machine translation. For example, French and Portuguese are both from the Romance language family sharing similar grammatical structures, and the pre-trained French translation model can be transferred to Portuguese translation via transfer learning [22,23] (**Figure 2**A). Motivated by the similarities between protein sequences and natural languages, we treated the IDR prediction as the French translation and the DFL prediction as the Portuguese translation (Figure 2B) according to the discussion of predictive correlations between IDRs and DFLs in Figure 1. A new DFL predictor was proposed called TransDFL, which was transferred from an IDR predictor by the transfer learning technology. The IDR predictor RFPR-IDP [24] was pre-trained with the IDR data to learn the common features between DFLs and IDRs, and then it was transferred to DFL prediction by fine-tuning so as to capture the specific features of DFLs. The proposed TransDFL has the following advantages: (1) the predicted model employs the sequence labeling method by combining bi-directional long short-term memory (Bi-LSTM) neural network and convolutional neural network (CNN), which models the protein as a whole and captures the local and long-range interaction features among residues; (2) the disordered features learned from the pre-trained IDR predictor by transfer learning can reduce the incorrectly predicted DFL residues in the ordered regions, leading to a lower FPR.

We evaluated the performance of different DFL predictors in two scenarios: prediction of DFLs only in the IDRs (situation-I) and prediction of DFLs in the entire proteins (situation-II). Experimental results showed that TransDFL consistently outperformed existing predictors. Furthermore, the corresponding web server of TransDFL was established, which can be accessed at http://bliulab.net/TransDFL/.

Method

Datasets

In the pre-training phase, the IDR benchmark dataset was used to train the RFPR-IDP predictor [15]. To avoid the redundancy between the source and target domains, proteins sharing > 25% similarity with any protein in the DFL datasets (TR166, TE82, and TE64) were removed from the IDR benchmark dataset by using the BLASTClust search tool [25], leading to 2645 training IDR sequences and 1077 validation IDR sequences.

In fine-tuning phase, the TR166 [6] DFL benchmark dataset collected by Peng et al. [6] was used for model fine-tuning, and any two proteins in the dataset share sequence similarity < 25%. We randomly divided the DFL benchmark dataset into five subsets. Four of the subsets with 133 sequences were randomly selected as the training dataset for fine-tuning the model parameters, and the remaining subset with 33 sequences was employed as the validation dataset for model selection. This way ensures that there is no redundancy between the validation and training datasets.

In this study, TE82 and TE64 independent test sets were used for the performance evaluation of different DFL predictors. The TE82 test set has 82 sequences collected from the DisProt database (version 8.0) by Peng et al. [6], and the sequence similarity between the TE82 and TR166 datasets is < 25%. We constructed a new independent test set TE64 from the latest released DisProt database (version 9.0, September 2021) [4,26]. Following the previous annotation protocols [5,6], IDR proteins that have DFL functionally annotated regions in the database were collected as "DFL proteins". To reduce data redundancy and avoid overestimating the predictive performance, only the sequences sharing < 25% similarity with any protein in TE82, TR166, and IDR benchmark datasets were included in TE64. Finally, 64 sequences were collected as the TE64 independent test set.

All benchmark datasets used in this study can be downloaded online at http://bliulab.net/TransDFL/benchmark/.

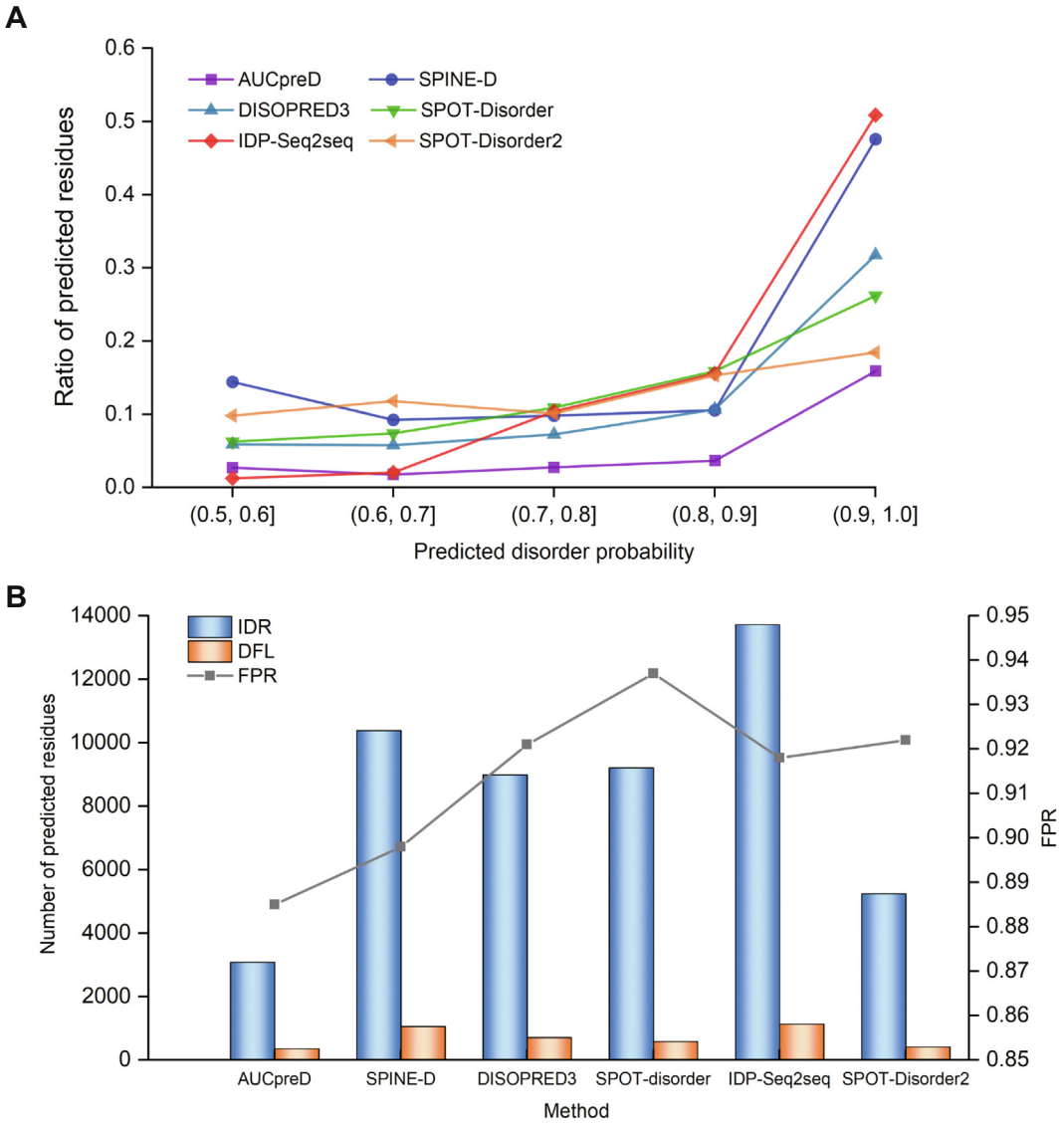


Figure 1 Applying the IDR predictors to DFLs

A. The relationships between the true DFLs and their probability scores predicted by IDR predictors. DFLs are preferred to be confidently predicted by IDR predictors with higher disordered probabilities, *i.e.*, extremely disordered state $P(D) > 0.9$. **B.** Histogram showing the number of predicted IDRs and DFLs by different IDR predictors with probabilities between 0.9 and 1.0. The line shows the corresponding FPR of each predictor, which equals the ratio of non-DFLs in predicted IDRs. DFL, disordered flexible linker; IDR, intrinsically disordered region; FPR, false positive rate.

The overview of TransDFL predictor

The overall flowchart of TransDFL is shown in **Figure 3**.

Sequence representation

In this study, the state-of-the-art IDR predictor RFPR-IDP [24] was employed to transfer into the DFL predictor. Two sequential features were used to represent the sequence in RFPR-IDP, including seven commonly used physicochemical properties [27] (steric parameter, polarizability, volume, hydrophobicity, isoelectric point, helix probability, and sheet probability), and position specific score matrix (PSSM) features generated by the PSI-BLAST tool [25] searching against the nrdb90 database (downloaded from https://ftp.ebi.ac.uk/pub/databases/nrdb90/). In this study, we also incorporated two additional features into RFPR-IDP so as to more comprehensively represent the DFL sequences, including the secondary structure (SS) features generated by the SPIDER tool [28,29], and the solvent accessibility (SA) features generated by the SABLE tool [30,31]. The linear combination of the 4-dimensional SS features, 1-dimensional SA feature, 7-dimensional physicochemical features (SEVEN), and 40-dimensional PSSM features, leads to a feature vector with 52 dimensions for representing a residue R_i as:

$$F_i = [f_1, f_2, \cdots, f_{52}]^T \quad (1)$$

The four features provide complementary information, and their combination leads to the best prediction performance (Table S1). Following previous studies [5,6], the local sliding window was applied to represent the residues.

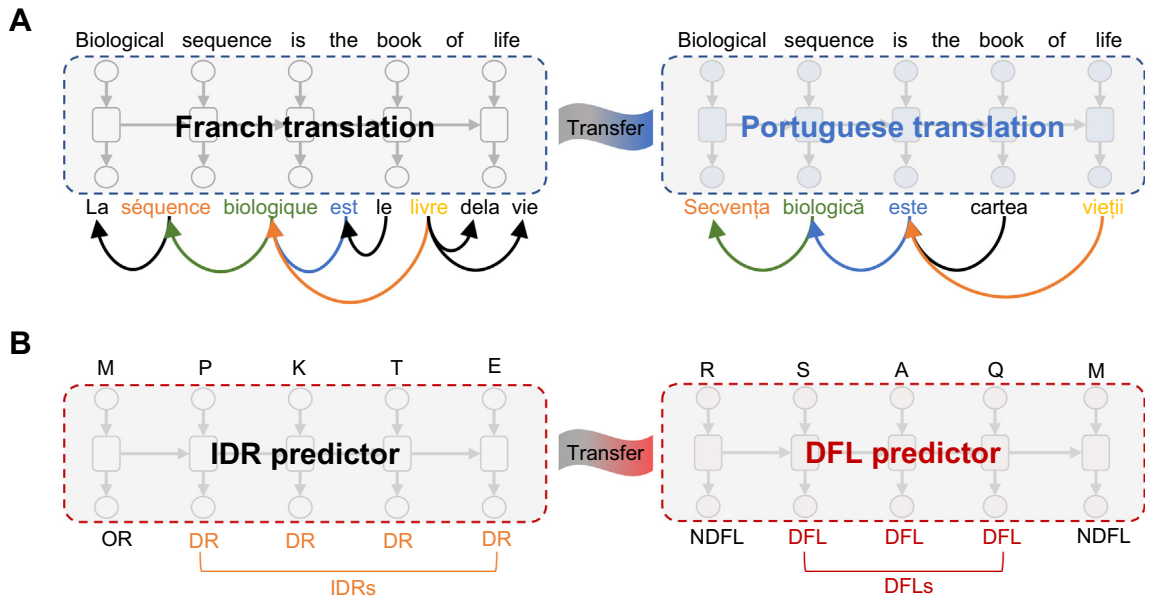


Figure 2 Comparison between the transfer learning frameworks for machine translation and DFL prediction

A. Linguistic commonalities learned from the language translation pairs, *i.e.*, the English–French parallel corpus can be adapted to the English–Portuguese translation by transfer learning. **B.** In DFL prediction, the IDR predictor RFPR-IDP trained with the IDR datasets was transferred to predict the DFLs by transfer learning. OR, ordered region; DR, disordered region; NDFL, non-DFL.

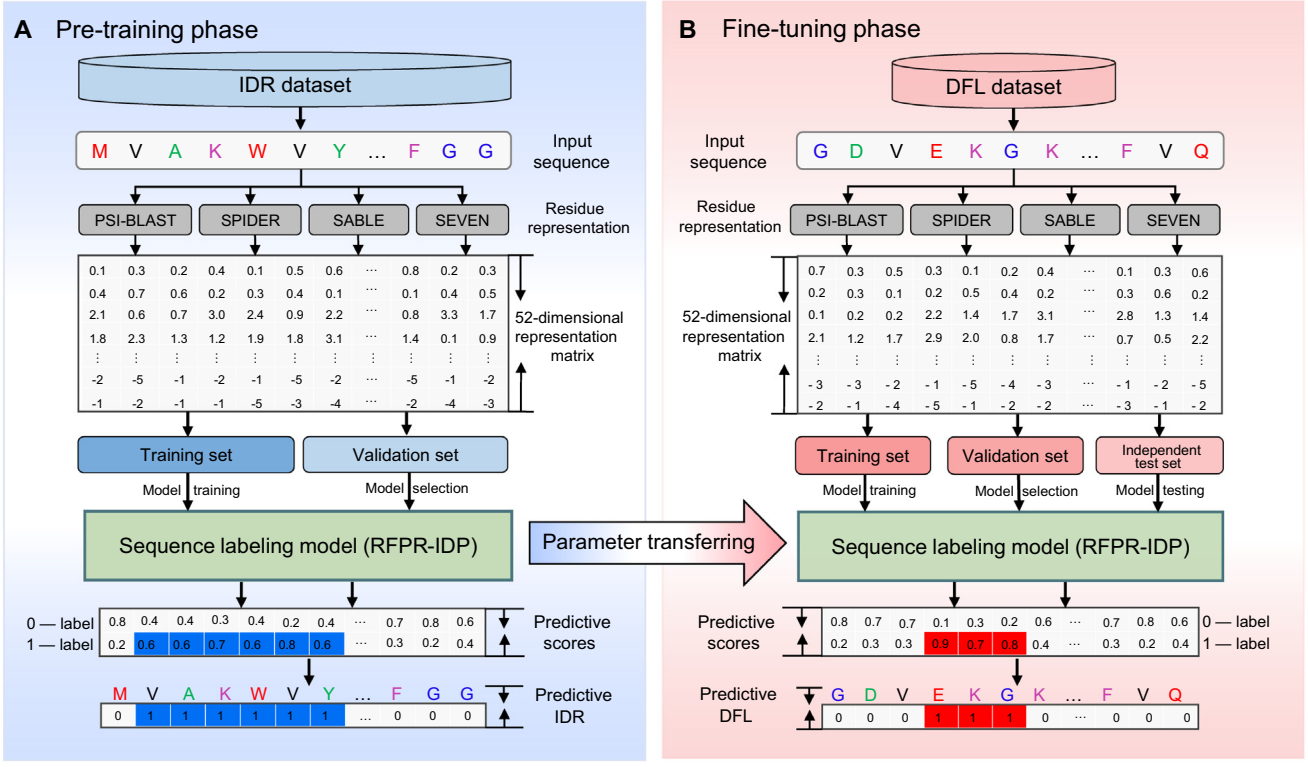


Figure 3 The flowchart of the TransDFL predictor

A. In the pre-training phase, the IDR dataset was used for pre-training the sequence labeling model of the IDR predictor RFPR-IDP. **B.** In the fine-tuning phase, the DFL dataset was used to fine-tune the sequence labeling model for DFL prediction through transfer learning.

The sequence labeling model transferred to TransDFL

The sequence labeling model is able to incorporate the correlation among adjacent residues and capture the interaction features of residues along the whole proteins. Two IDR predictors IDP-Seq2Seq [15] and RFPR-IDP [24] based on the sequence labeling model were used to be transferred to TransDFL. However, due to the insufficient number of DFL training sequences, the IDP-Seq2Seq using a more complex network structure is not suitable. Therefore, the RFPR-ID predictor by a combination of Bi-LSTM and CNN is more suitable for transferring to TransDFL. The model architecture is shown in **Figure 4**. The Bi-LSTM layer with a forward and a backward LSTM network layer was adopted to capture the global correlation features. For each residue R_i, the correlation feature vector H_i is calculated by [24]:

$$H_i = \left[\overrightarrow{h_i}; \overleftarrow{h_i}\right] \tag{2}$$

$$\overrightarrow{h_i}, \overrightarrow{c_i} = LSTM_f(win_k F_i, \overrightarrow{h_{i-1}}; \overrightarrow{c_{i-1}}) \tag{3}$$

$$\overleftarrow{h_i}, \overleftarrow{c_i} = LSTM_b(win_k F_i, \overleftarrow{h_{i-1}}; \overleftarrow{c_{i-1}}) \tag{4}$$

where $\overrightarrow{h_i}$ and $\overleftarrow{h_i}$ represent the forward and backward output feature vectors of R_i, respectively. The $win_k F_i$ is the feature representation vector of R_i, which is the combination of the corresponding feature vectors of target residue R_i and its $k-1$ neighboring residues.

The convolutional layer was used to capture the local correlation features H'_i of R_i:

$$H'_i = conv(W, H) + b \tag{5}$$

where W is the convolutional kernel and b is the bias parameter matrix.

Then, a fully-connected layer was used to predict the label of each residue, mapping the output feature vector H'_i from the CNN layer to a probability score p_i of R_i being a positive residue, which is calculated by [32]:

$$O_i^1 = tanh(W_1^1 H'_i + b_1^1) \tag{6}$$

$$O_i^2 = tanh(W_1^2 O_i^1 + b_1^2) \tag{7}$$

$$p_i = softmax(W_2 O_i^2 + b_2) \tag{8}$$

where O_i^1 and O_i^2 represent the output vectors of the first and second fully-connected layers, respectively; W_1^1, W_1^2, and W_2 are the trainable weight parameter vectors; b_1^1, b_1^2, and b_2 are the trainable bias parameter vectors; *tanh* is the hyperbolic tangent activation function [33]; and *softmax* is the soft argmax activation function [34].

The pre-training phase

In the pre-training phase, the RFPR-IDP predictor was pre-trained with the IDR dataset (Figure 3A). The pre-trained model was optimized based on the binary cross entropy loss function calculated by [35]:

$$\text{loss} = -\sum_{i=1}^{L}(y_i \log(p_i) + (1-y_i)\log(1-p_i)) \tag{9}$$

where L is the length of a sequence, p_i is the predictive probability score of residue R_i being an IDR residue [Equation (8)], and y_i is the corresponding real label (0 or 1). All the model parameters were optimized by minimizing the loss function value on the IDR validation set. The model pre-trained with the source domain IDR dataset learns the common characteristics shared with IDRs and DFLs, which can be used for DFL prediction by transfer learning. The hyperparameters of RFPR-IDP in the pre-training phase are given in Table S3.

The fine-tuning phase

Because DFLs are the extremely flexible disordered regions predicted by IDR predictors with high probabilities, this pattern can be transferred to identify the DFLs via transfer learning (Figure 3B). Different from the model directly trained with the target dataset with a limited number of samples, the model fine-tuned based on the pre-trained parameters avoids overfitting, and improves the prediction performance in the target domain [36].

The weighted binary cross entropy loss function was employed in the fine-tuning phase:

$$\text{loss} = -\sum_{i=1}^{L}[w \times y_i \log(p_i) + (1-w)(1-y_i)\log(1-p_i)] \tag{10}$$

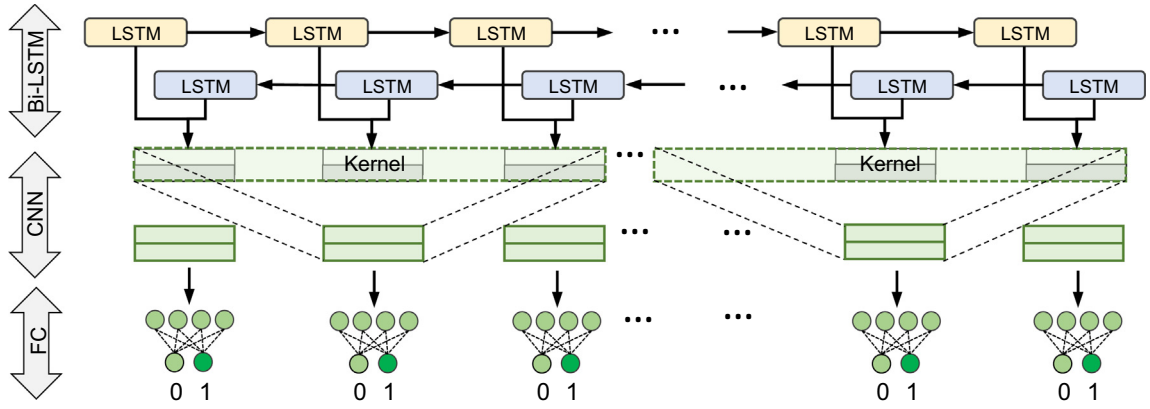


Figure 4 The sequence labeling model architecture transferred to TransDFL

FC, fully-connected; CNN, convolutional neural network; LSTM, long short-term memory; Bi-LSTM, bi-directional long short-term memory.

where the w is the weight coefficient of DFL residue, optimized according to the best area under the receiver operating characteristic curve (AUROC) on the DFL validation dataset (Table S2). The model was implemented by the Tensorflow framework [37]. Adam algorithm [38] with a learning rate of 0.0008 was used for parameter optimization. All the parameters of the pre-trained RFPR-IDP model were fine-tuned on the validation set of the DFL benchmark dataset according to the minimum loss. The hyperparameters in the fine-tuning phase are given in Table S4.

Performance evaluation strategy

In this study, the AUROC was used to evaluate the overall performance of different methods [39–41]. Besides, following previous studies [6,42], the Matthews correlation coefficient (MCC) [29,43], precision (Pre), and recall (Rec) [44] were used to evaluate the predictive quality of a predictor:

$$\begin{cases} MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \\ Pre = \frac{TP}{TP+FP} \\ Rec = \frac{TP}{TP+FN} \end{cases} \quad (11)$$

where true positive (TP) is the number of DFL residues correctly predicted as DFLs, false positive (FP) is the number of non-DFL residues incorrectly predicted as DFLs, true negative (TN) is the number of non-DFL residues correctly predicted as non-DFLs, false negative (FN) is the number of DFL residues incorrectly predicted as non-DFLs. Given a threshold thd, a residue R_i is classified as a DFL residue, if its predictive probability score $p_i \geq thd$. Otherwise, it is predicted as a non-DFL residue.

Results and discussion

Predicting the DFL residues in disordered regions

We compared the performance of TransDFL to the other two DFL predictors (DFLpred [5] and APOD [6]) on two independent test sets (TE82 and TE64). Following previous studies [5,6], disordered regions without functional annotations and ordered regions were not evaluated (situation-I). The results of different predictors on the TE82 and TE64 datasets are shown in **Tables 1** and **2**, respectively. From these results, we can see that TransDFL outperforms DFLpred by 0.198 in terms of AUROC on the TE82 dataset, and achieves highly comparable performance with APOD. Particularly, the precision-recall (PR) curves (Figure S2) showed that the prediction results of TransDFL and APOD were complementary and their differences were significant ($P = 0.001$; Tables 1 and 2).

Predicting the DFL residues in the entire proteins

DFLs are flexible linkers in disordered regions. The existing two predictors (APOD and DFLpred) focus on predicting DFLs only in the functionally annotated disordered regions. However, the information on the disordered regions is not always available [15,45–47]. In order to more comprehensively and fairly evaluate the performance of different methods, they were evaluated by identifying DFLs in the entire protein sequences in TE82 and TE64 datasets (situation-II). Their AUROC values are shown in **Figure 5**, and the PR curves and the area under the PR curve (AUPRC) are shown in Figure S3, from which we can see the followings: (1) compared with the results in situation-I, the performance of these three predictors decreased, indicating that predicting DFLs in the entire sequence is more challenging; (2) TransDFL obviously outperforms DFLpred and APOD on both two datasets in terms of AUROC.

Performance comparison between TransDFL and IDR predictors

In order to investigate the performance of IDR predictors for predicting DFLs, we employed six state-of-the-art IDR predictors to identify DFLs. Because the average length of DFLs is 47 amino acids (aas) in the TR166 dataset (Figure S1), a target residue is considered as a DFL residue if its 46 neighboring residues are disordered residues (the target is in the middle).

Table 1 Performance comparison of TransDFL and other DFL predictors on the TE82 independent test set evaluated in situation-I

Method	Pre	Rec	MCC	AUROC	*P* value
DFLpred	0.337	0.179	0.145	0.637	1.360E−147
APOD	0.512	0.512	0.418	0.816	0.001
TransDFL	0.586	0.452	0.435	0.835	/

Note: The results of APOD and DFLpred were obtained from a previous study [6] evaluated on the same TE82 dataset. The *thd* of TransDFL was set as 0.16, which is equal to the ratio of DFL residues in the dataset. The *P* value was calculated by *t*-test based on the probabilities predicted by different methods. Pre, precision; Rec, recall; MCC, Matthews correlation coefficient; AUROC, area under receiver operating characteristic curve; DFL, disordered flexible linker.

Table 2 Performance comparison of TransDFL and other DFL predictors on the TE64 independent test set evaluated in situation-I

Method	Pre	Rec	MCC	AUROC	*P* value
DFLpred	0.189	0.193	0.108	0.675	5.330E−15
APOD	0.195	0.600	0.223	0.751	0.001
TransDFL	0.207	0.722	0.273	0.784	/

Note: The results of APOD and DFLpred were calculated according to the predicted results obtained by running corresponding web servers. The *thd* of TransDFL was set as same as in Table 1. The *P* value was calculated by *t*-test based on the probabilities predicted by different methods.

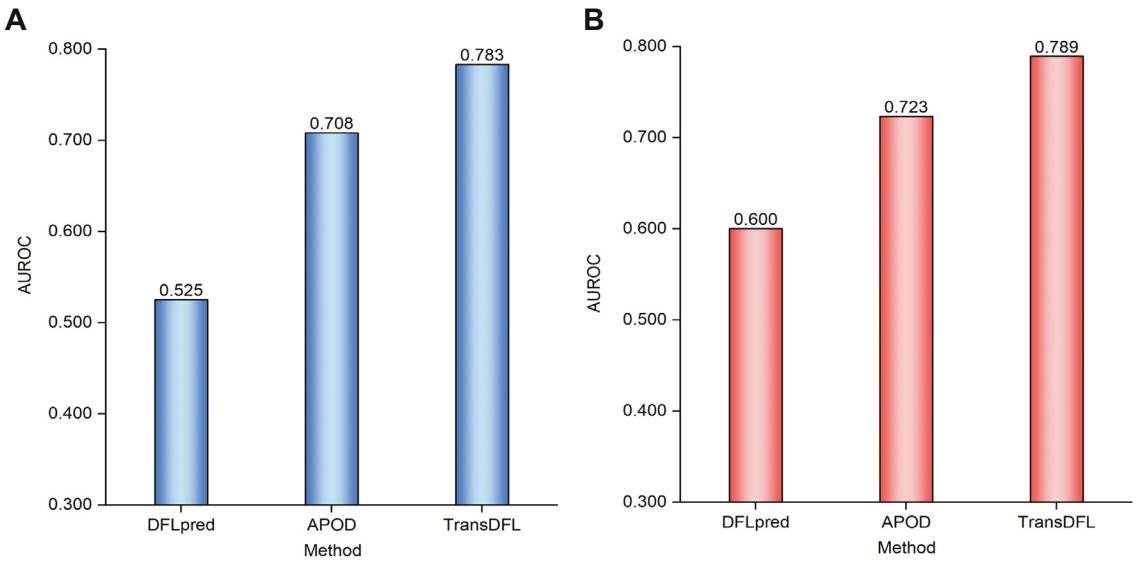


Figure 5 Comparison of different DFL predictors in situation-II

A. AUROC values on the TE82 independent test set. **B.** AUROC values on the TE64 independent test set. AUROC, area under the receiver operating characteristic curve.

The results of different methods evaluated in two situations on the TE82 independent test set are shown in Tables S5 and S6. From these results, we can see that the six IDR predictors are not effective enough for identifying DFLs compared with the specific DFL predictor TransDFL in both two evaluation situations.

Transfer learning obviously reduces the false positives

In order to explore the contribution of transfer learning to the performance improvement of TransDFL, we compared the FPR in all predictions (FPR^{ALL}) and the FPR in the ordered region (FPR^{OR}) of three different predictors. The FPR^{OR} is calculated as the ratio of the number of falsely predicted DFL residues in ordered regions to the number of all the positively predicted DFL residues, where the ordered residues are annotated according to the DisProt database (version 8.2). For fair evaluation, the FPRs of different predictors were compared under the same number of positively predicted residues. As shown in **Figure 6**, TransDFL achieves the lowest FPR^{ALL} and FPR^{OR}. These results are not surprising because TransDFL employs the transfer learning framework pre-trained with the IDR dataset to capture the common characteristics between IDRs and DFLs. Therefore, compared with the other predictors only trained with DFLs and disordered residues, TransDFL can obviously reduce the number of incorrectly predicted DFL residues in the ordered regions so as to reduce the overall false positive predictions.

The prediction results of a protein (DisProt ID: DP01080; PDB ID: 1OCB) from the TE82 independent test set obtained by different predictors were visualized by the PyMOL software (https://pymol.org/2/). As shown in **Figure 7**, although most of the DFLs can be correctly predicted by TransDFL, DFLpred, and APOD, the false positives predicted by TransDFL are obviously fewer than those predicted by DFLpred and APOD evaluated in situation-II. The false positives predicted by TransDFL are in the disordered regions near the true DFLs, while most of the false positives predicted by APOD and DFLpred are located in the ordered regions far away from the true DFLs. These results are fully consistent with the observations in Figure 6.

In order to explore the contribution of the model pre-trained with disordered proteins, we compared the predictive performance between the TransDFL model directly trained with DFLs (TransDFL-DT) and the fine-tuned model based on pre-training with IDRs (TransDFL). The evaluation results showed that TransDFL consistently outperformed TransDFL-DT on two independent test sets in both two situations (**Table 3**), indicating that transfer learning contributes to the predictive performance improvement of TransDFL.

The sequence labeling model facilitates the stable performance on different lengths of DFL regions

In order to investigate the performance of TransDFL for predicting DFL regions with different lengths, we divided the protein sequences in the TE82 independent test set into five groups according to their DFL lengths. As shown in **Figure 8**, compared with APOD and DFLpred, TransDFL is insensitive to the lengths of DFL regions, and achieves better and more stable performance. There are two reasons. First, TransDFL employs the sequence labeling model based on deep learning technology, which is able to capture the local and global interactions among the residues and the sequence patterns of the DFLs. In contrast, all the other two classifiers are classification-based methods predicting each residue in a separate manner. Second, benefitting from the deep neural networks, the sequence labeling model in TransDFL captures the general disordered characteristics of DFLs from the large IDR dataset, which facilitates the DFL prediction.

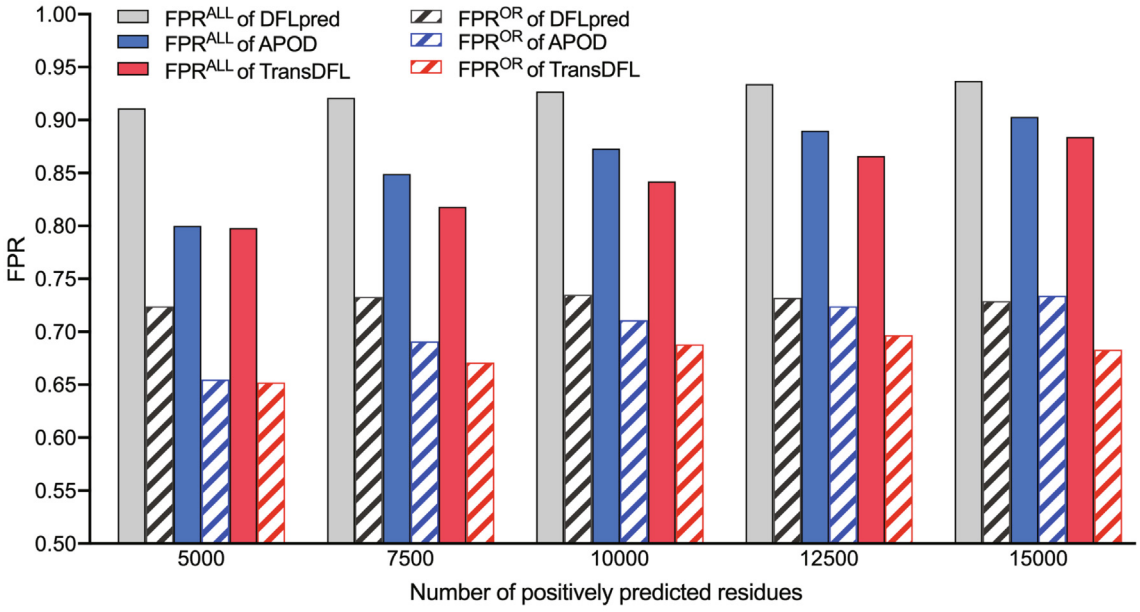


Figure 6 Comparison of FPRs of different predictors on the TE82 test set
FPR^{ALL}, false positive rate in all predictions; FPR^{OR}, false positive rate in the ordered region.

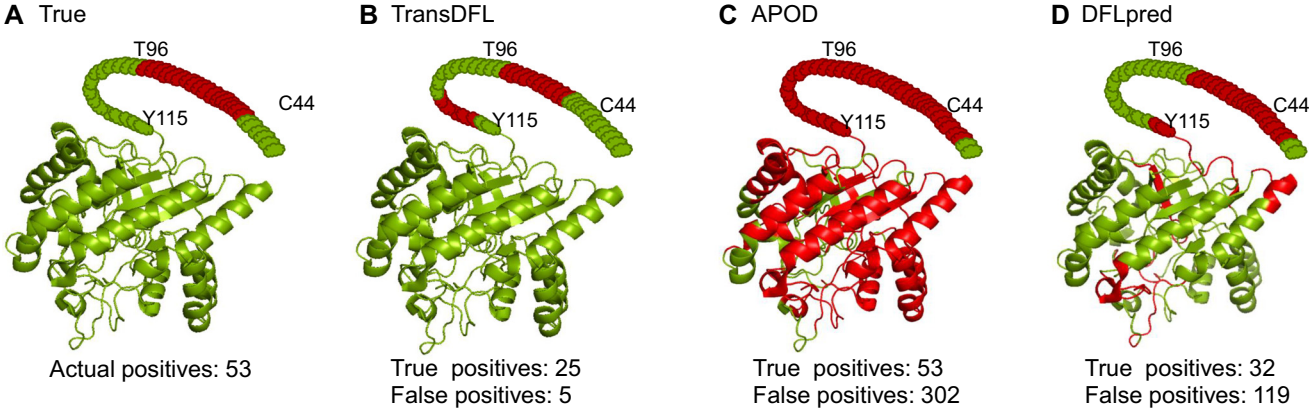


Figure 7 Visualization of predictive results
The predictive results of a protein (DisProt ID: DP01080; PDB ID: 1OCB) were visualized by the PyMOL software (https://pymol.org/2/). The true and predicted DFL residues are shown in red. **A.** The true DFLs. **B.** DFLs predicted by TransDFL. **C.** DFLs predicted by APOD. **D.** DFLs predicted by DFLpred.

Table 3 Performance of TransDFL predictors based on different models

Dataset	Model	Situation-I				Situation-II			
		Pre	Rec	MCC	AUROC	Pre	Rec	MCC	AUROC
TE82	TransDFL-DT	0.552	0.250	0.298	0.746	0.010	0.581	0.120	0.705
	TransDFL	0.207	0.722	0.273	0.789	0.149	0.727	0.241	0.783
TE64	TransDFL-DT	0.254	0.481	0.166	0.697	0.080	0.680	0.113	0.642
	TransDFL	0.275	0.518	0.289	0.784	0.207	0.722	0.273	0.789

Note: TransDFL-DT refers to the model directly trained with the DFL training set. TransDFL refers to the transferred model with pre-training on the IDR dataset. IDR, intrinsically disordered region.

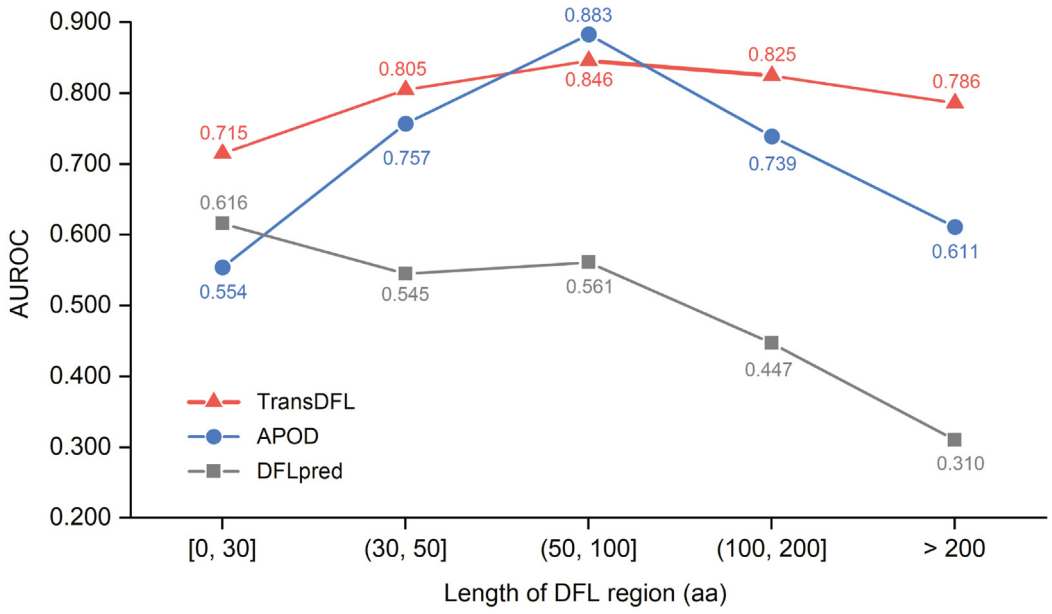


Figure 8 **Performance comparison of TransDFL, APOD, and DFLpred for predicting proteins with different lengths of DFL regions**
aa, amino acid.

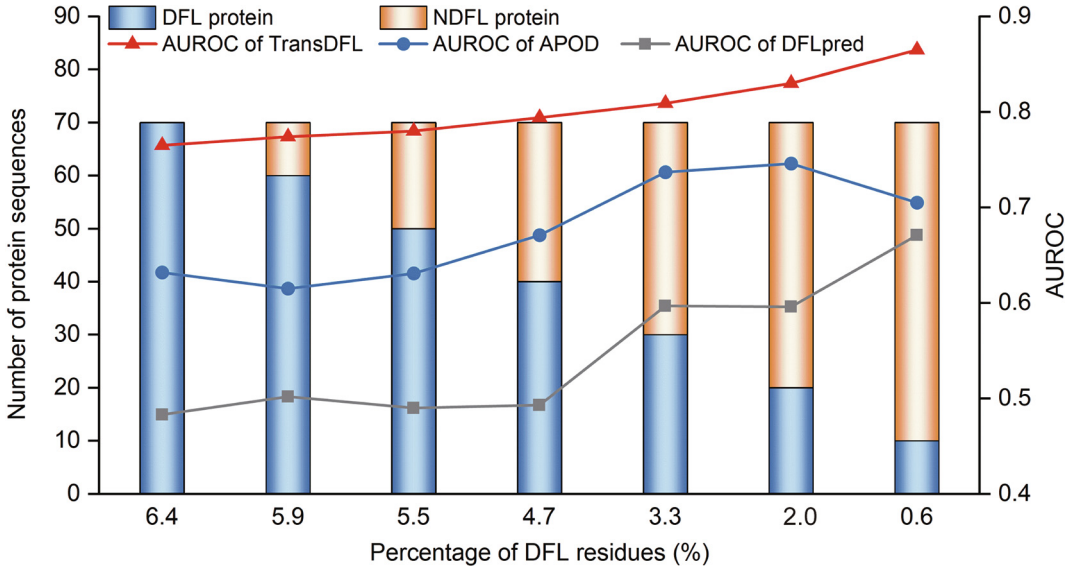


Figure 9 **Predictive results of TransDFL, APOD, and DFLpred on real-word datasets with different percentages of DFL residues**

Lower FPR leads to better performance in the real-world application

According to the latest DisProt database, only 6.3% of the 4438 annotated disordered regions are DFLs [26]. There are even many more disordered proteins without DFLs in MobiDB [48]. As a result, the percentage of DFL residues is much lower than 6.3% in nature. Therefore, for real-world applications, it is important for a DFL predictor to deal with the extremely imbalanced problem (*i.e.*, the number of non-DFL residues is much higher than the number of DFL residues). In this regard, seven datasets were constructed based on TE82 and TE64 with different percentages of DFL residues. The performance of different predictors on the seven datasets is shown in Figure 9. We observed that TransDFL consistently outperformed both APOD and DFLpred, especially for the datasets with fewer DFL residues. These results indicate that TransDFL is able to solve the imbalanced problem, and therefore, it is more suitable for real-world applications.

Conclusion

Inspired by the similarity between protein sequences and natural language sentences, we applied the transfer learning derived from the machine translation to the DFL identification, and a new predictor TransDFL was proposed.

TransDFL was constructed by transferring the state-of-the-art IDR predictor RFPR-IDP into the current DFL predictor. It has the following advantages: (1) TransDFL employs the sequence labeling model to capture the global sequence patterns of DFLs; (2) benefitting from transfer learning, TransDFL is the first deep learning predictor for DFL prediction and achieves state-of-the-art performance with the lowest FPR. The web server of TransDFL was established, which can be freely accessed at http://bliulab.net/TransDFL.

Code availability

The source code of TransDFL is available at https://ngdc.cncb.ac.cn/biocode/tools/BT007312.

Data availability

The web server of TransDFL can be freely accessed at http://bliulab.net/TransDFL/.

Competing interests

Both authors have declared no competing interests.

CRediT authorship contribution statement

Yihe Pang: Methodology, Software, Validation, Formal analysis, Data curation, Writing – original draft, Writing – review & editing. **Bin Liu:** Conceptualization, Resources, Writing – review & editing, Supervision, Project administration, Funding acquisition. Both authors have read and approved the final manuscript.

Acknowledgments

This work was supported by the National Key R&D Program of China (Grant No. 2018AAA0100100), and the Beijing Natural Science Foundation, China (Grant No. JQ19019). We are very much indebted to the three anonymous reviewers, whose constructive comments are very helpful for strengthening the presentation of this paper.

Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gpb.2022.10.004.

ORCID

ORCID 0000-0001-7338-5739 (Yihe Pang)
ORCID 0000-0002-8520-8374 (Bin Liu)

References

[1] Basile W, Salvatore M, Bassot C, Elofsson A. Why do eukaryotic proteins contain more intrinsically disordered regions? PLoS Comput Biol 2019;15:e1007186.
[2] Dyson HJ, Wright PE. Intrinsically unstructured proteins and their functions. Nat Rev Mol Cell Biol 2005;6:197–208.
[3] van der Lee R, Buljan M, Lang B, Weatheritt RJ, Daughdrill GW, Dunker AK, et al. Classification of intrinsically disordered regions and proteins. Chem Rev 2014;114:6589–631.
[4] Quaglia F, Meszaros B, Salladini E, Hatos A, Pancsa R, Chemes LB, et al. DisProt in 2022: improved quality and accessibility of protein intrinsic disorder annotation. Nucleic Acids Res 2022;50:D480–7.
[5] Meng F, Kurgan L. DFLpred: high-throughput prediction of disordered flexible linker regions in protein sequences. Bioinformatics 2016;32:i341–50.
[6] Peng Z, Xing Q, Kurgan L. APOD: accurate sequence-based predictor of disordered flexible linkers. Bioinformatics 2020;36:i754–61.
[7] Dunker AK, Brown CJ, Lawson JD, Iakoucheva LM, Obradovic Z. Intrinsic disorder and protein function. Biochemistry 2002;41:6573–82.
[8] Chen X, Zaro JL, Shen WC. Fusion protein linkers: property, design and functionality. Adv Drug Deliv Rev 2013;65:1357–69.
[9] Szabo B, Horvath T, Schad E, Murvai N, Tantos A, Kalmar L, et al. Intrinsically disordered linkers impart processivity on enzymes by spatial confinement of binding domains. Int J Mol Sci 2019;20:2119.
[10] George RA, Heringa J. An analysis of protein domain linkers: their classification and role in protein folding. Protein Eng 2002;15:871–9.
[11] Sorensen CS, Kjaergaard M. Effective concentrations enforced by intrinsically disordered linkers are governed by polymer physics. Proc Natl Acad Sci U S A 2019;116:23124–31.
[12] Harmon TS, Holehouse AS, Rosen MK, Pappu RV. Intrinsically disordered linkers determine the interplay between phase separation and gelation in multivalent proteins. Elife 2017;6:e30294.
[13] Liu Y, Wang X, Liu B. A comprehensive review and comparison of existing computational methods for intrinsically disordered protein and region prediction. Brief Bioinform 2019;20:330–46.
[14] Hanson J, Paliwal KK, Litfin T, Zhou Y. SPOT-Disorder2: improved protein intrinsic disorder prediction by ensembled deep learning. Genomics Proteomics Bioinformatics 2019;17:645–56.
[15] Tang YJ, Pang YH, Liu B. IDP-Seq2Seq: identification of intrinsically disordered regions based on sequence to sequence learning. Bioinformatics 2021;36:5177–86.
[16] Necci M, Piovesan D, Predictors C, DisProt C, Tosatto SCE. Critical assessment of protein intrinsic disorder prediction. Nat Methods 2021;18:472–81.
[17] Wang S, Ma J, Xu J. AUCpreD: proteome-level protein disorder prediction by AUC-maximized deep convolutional neural fields. Bioinformatics 2016;32:i672–9.
[18] Zhang T, Faraggi E, Xue B, Dunker AK, Uversky VN, Zhou Y. SPINE-D: accurate prediction of short and long disordered regions by a single neural-network based method. J Biomol Struct Dyn 2012;29:799–813.
[19] Jones DT, Cozzetto D. DISOPRED3: precise disordered region predictions with annotated protein-binding activity. Bioinformatics 2015;31:857–63.
[20] Hanson J, Yang Y, Paliwal K, Zhou Y. Improving protein disorder prediction by deep bidirectional long short-term memory recurrent neural networks. Bioinformatics 2017;33:685–92.
[21] Brown PF, Cocke J, Della Pietra SA, Della Pietra VJ, Jelinek F, Lafferty J, et al. A statistical approach to machine translation. Comput Linguist 1990;16:79–85.
[22] Zoph B, Yuret D, May J, Knight K. Transfer learning for low-resource neural machine translation. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016:1568–75.
[23] Gu J, Wang Y, Chen Y, Li VOK, Cho K. Meta-learning for low-resource neural machine translation. Proceedings of the 2018

Conference on Empirical Methods in Natural Language Processing :3622–31.
[24] Liu Y, Wang X, Liu B. RFPR-IDP: reduce the false positive rates for intrinsically disordered protein and region prediction by incorporating both fully ordered proteins and disordered proteins. Brief Bioinform 2021;22:2000–11.
[25] Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–402.
[26] Hatos A, Hajdu-Soltesz B, Monzon AM, Palopoli N, Alvarez L, Aykac-Fas B, et al. DisProt: intrinsic protein disorder annotation in 2020. Nucleic Acids Res 2020;48:D269–76.
[27] Meiler J, Muller M, Zeidler A, Schmaschke F. Generation and evaluation of dimension-reduced amino acid parameter representations by artificial neural networks. J Mol Model 2001;7:360–9.
[28] Yang Y, Heffernan R, Paliwal K, Lyons J, Dehzangi A, Sharma A, et al. SPIDER2: a package to predict secondary structure, accessible surface area, and main-chain torsional angles by deep neural networks. Methods Mol Biol 2017;1484:55–63.
[29] Guo L, Jiang Q, Jin X, Liu L, Zhou W, Yao S, et al. A deep convolutional neural network to improve the prediction of protein secondary structure. Curr Bioinform 2020;15:767–77.
[30] Adamczak R, Porollo A, Meller J. Accurate prediction of solvent accessibility using neural networks-based regression. Proteins 2004;56:753–67.
[31] Smolarczyk T, Roterman-Konieczna I, Stapor K. Protein secondary structure prediction: a review of progress and directions. Curr Bioinform 2020;15:90–107.
[32] Pang Y, Liu B. SelfAT-Fold: protein fold recognition based on residue-based and motif-based self-attention networks. IEEE/ACM Trans Comput Biol Bioinform 2022;19:1861–9.
[33] Karlik B, Olgac AV. Performance analysis of various activation functions in generalized MLP architectures of neural networks. Int J Artif Intell Exp Syst 2011;1:111–22.
[34] Jiang M, Liang Y, Feng X, Fan X, Pei Z, Xue Y, et al. Text classification based on deep belief network and softmax regression. Neural Comput Appl 2018;29:61–70.
[35] Zhang Z, Sabuncu M. Generalized cross entropy loss for training deep neural networks with noisy labels. Adv Neural Inf Process Syst 2018;31:8778–88.
[36] Pan SJ, Yang Q. A survey on transfer learning. IEEE Trans Knowl Data Eng 2009;22:1345–59.
[37] Abadi M, Barham P, Chen JM, Chen ZF, Davis A, Dean J, et al. TensorFlow: a system for large-scale machine learning. Proceedings of the 12th USENIX conference on Operating Systems Design and Implementation 2016:265–83.
[38] Kingma DP, Ba J. Adam: a method for stochastic optimization. 3rd International Conference for Learning Representations 2015:1–11.
[39] Davis J, Goadrich M. The relationship between precision-recall and ROC curves. Proceedings of the 23rd International Conference on Machine Learning 2006:233–40.
[40] An JY, Zhou Y, Zhang L, Niu Q, Wang DF. Improving self-interacting proteins prediction accuracy using protein evolutionary information and weighed-extreme learning machine. Curr Bioinform 2019;14:115–22.
[41] Yang H, Luo Y, Ren X, Wu M, He X, Peng B, et al. Risk prediction of diabetes: big data mining with fusion of multifarious physical examination indicators. Inf Fusion 2021;75:140–9.
[42] Jing X, Dong Q, Lu R, Dong Q. Protein inter-residue contacts prediction: methods, performances and applications. Curr Bioinform 2019;14:178–89.
[43] Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta 1975;405:442–51.
[44] Ikram N, Qadir MA, Afzal MT. SimExact – an efficient method to compute function similarity between proteins using Gene Ontology. Curr Bioinform 2020;15:318–27.
[45] Katuwawala A, Ghadermarzi S, Kurgan L. Computational prediction of functions of intrinsically disordered regions. Prog Mol Biol Transl Sci 2019;166:341–69.
[46] Habchi J, Tompa P, Longhi S, Uversky VN. Introducing protein intrinsic disorder. Chem Rev 2014;114:6561–88.
[47] Qi YJ, Lu HN, Zhao YM, Wang Z, Ji YJ, Jin NZ, et al. Screening and analysis of hypolipidemic components from Shuangdan capsule based on pancreatic lipase. Curr Bioinform 2020;15:478–92.
[48] Piovesan D, Necci M, Escobedo N, Monzon AM, Hatos A, Micetic I, et al. MobiDB: intrinsically disordered proteins in 2021. Nucleic Acids Res 2021;49:D361–7.