PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

DMFpred: Predicting protein disorder molecular functions based on protein cubic language model

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

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

* bliu@bliulab.net

OPEN ACCESS

Citation: Pang Y, Liu B (2022) DMFpred: Predicting protein disorder molecular functions based on protein cubic language model. PLoS Comput Biol 18(10): e1010668. https://doi.org/10.1371/journal.pcbi.1010668

Editor: Jeffrey Skolnick, Georgia Institute of Technology, UNITED STATES

Received: August 3, 2022

Accepted: October 19, 2022

Published: October 31, 2022

Data Availability Statement: The data used in this study are available at http://bliulab.net/DMFpred/.

Funding: This work was supported by the National Natural Science Foundation of China (No. 62271049 and U22A2039) and the Beijing Natural Science Foundation (No. JQ19019) to (BL). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Intrinsically disordered proteins and regions (IDP/IDRs) are widespread in living organisms and perform various essential molecular functions. These functions are summarized as six general categories, including entropic chain, assembler, scavenger, effector, display site, and chaperone. The alteration of IDP functions is responsible for many human diseases. Therefore, identifying the function of disordered proteins is helpful for the studies of drug target discovery and rational drug design. Experimental identification of the molecular functions of IDP in the wet lab is an expensive and laborious procedure that is not applicable on a large scale. Some computational methods have been proposed and mainly focus on predicting the entropic chain function of IDRs, while the computational predictive methods for the remaining five important categories of disordered molecular functions are desired. Motivated by the growing numbers of experimental annotated functional sequences and the need to expand the coverage of disordered protein function predictors, we proposed DMFpred for disordered molecular functions prediction, covering disordered assembler, scavenger, effector, display site and chaperone. DMFpred employs the Protein Cubic Language Model (PCLM), which incorporates three protein language models for characterizing sequences, structural and functional features of proteins, and attention-based alignment for understanding the relationship among three captured features and generating a joint representation of proteins. The PCLM was pre-trained with large-scaled IDR sequences and fine-tuned with functional annotation sequences for molecular function prediction. The predictive performance evaluation on five categories of functional and multi-functional residues suggested that DMFpred provides high-quality predictions. The web-server of DMFpred can be freely accessed from http://bliulab.net/DMFpred/.

Author summary

Intrinsically disordered proteins (IDPs) are proteins that are without stable three-dimensional (3D) structures in native physiologic conditions. The discovery of IDPs has disproved the idea that proteins must fold into 3D structures to accomplish their biological

functions. They are prevalent in eukaryotic organisms and carry out many critical functions in cellular regulation, signaling networks, and disease pathways. These functions of IDPs can be summarized into six general categories, including entropic chain, assembler, scavenger, effector, display site, and chaperone. Experimental identification of the molecular functions of IDP in the wet lab is an expensive and laborious procedure that is not applicable on a large scale. Some computational methods have been proposed to identify the entropic chain function of IDPs, while the predictive methods for the remaining but important functions of IDPs are desired. In this study, we proposed a disordered molecular function computational predictor of proteins, namely DMFpred. The DMFpred supports high-throughput sequences as input and computationally predicts five molecular functions of IDPs, including assembler, scavenger, effector, display site, and chaperone. The evaluation results suggested that DMFpred provides high-quality predictions, and the corresponding web server of DMFpred can be freely accessed from http://bliulab.net/DMFpred/.

This is a *PLOS Computational Biology* Methods paper.

Introduction

Proteins or regions that lack stable 3D-structures under the native physiologic conditions are known as intrinsically disordered proteins and regions (IDP/IDRs). Recent studies have suggested that IDP/IDRs are common in nature, with more than 30% of proteins in eukaryotes being disordered [1,2]. The widespread occurrence of IDP/IDRs alter the classical protein structure-function paradigm [3–5]. IDP/IDRs play essential roles in living organisms, the alteration of their functions are responsible for many human diseases such as cancer [6], Alzheimer's [7] and Parkinson's [8]. Exploring the molecular functional mechanism of IDP/IDRs will be helpful for a complete understanding of protein structures and functions, and will be also used to guide wet lab experiments and inform studies of rational drug design [9,10].

The functions of protein disordered regions arise from their native structural flexibility or from their ability to bind to partner molecules [4]. These disorder functions can be summarized as six categories: entropic chains, assembler, scavenger, effector, display site, and chaperone [4,11]. The disordered entropic chain benefits directly from its intrinsically disordered conformation without becoming structured, which serves as the connector between domains and structural elements making up domains [12]. Disordered assemblers bring together multiple binding partners, and promote the formation of large protein complexes [4,5,13]. Scavenger disordered regions in proteins store and neutralize small ligands, such as chromogranin, salivary glycoproteins and calcium-binding phosphoproteins [11,14,15]. Effectors interact with other partner proteins and modify their activity [16]. Some disordered regions serve as display sites, facilitating easy access and recognition of the post-translational modifications (PTMs) in proteins [17]. Disordered chaperone function makes the IDRs assisting RNA and protein molecules to reach their functionally folded states [18].

The intrinsically disordered is encoded in the protein sequence, motivating the development of computational sequence-based disorder predictors [19]. Currently, there are about 200 million disordered proteins have been identified experimentally and predictively [20]. In contrast, only thousands of disordered proteins have functional annotations [21,22]. This data

suggests that it is important to develop computational predictors for filling the deepening gap between annotated and unannotated disordered sequences. In this regard, several sequence-based computational predictors are proposed for predicting specific functions of disordered proteins. For example, the DFLpred [23] and APOD [24] are computational methods developed for predicting disordered linkers that fulfill entropic chain function in proteins. Besides, there are predictors for identifying disordered regions binding to specific types of molecular partners, including protein binding predictors [25–32], DNAs and RNAs binding predictors [33,34], and lipid binding predictors [35]. However, methods for predicting the other five classes (assembler, scavenger, effector, display site and chaperone) of molecular functions of IDRs are required.

Protein representation is critical for the construction of computational predictors. Protein sequence defines structure, which in turn dictates its function [4]. The intrinsically disordered proteins reassessed the classical sequence-structure-function paradigm [36], the complex sequence, structure, and functional properties of IDP/IDRs should be explored to fully represent the disordered proteins. By modelling the language's generative rules, the language model in natural language processing (NLP) comprehensively understands the language, and capture the semantic features of text, which is an indispensable technology in NLP. Protein sequences can be viewed as the language of genetics sharing high similarities with natural language sentences [37]. For example, the natural language sentences composed of words express their semantics, while proteins composed of residues perform various functions. Inspired by their similarities, the proteins can be represented and modelled by the language models.

In this paper, we proposed DMFpred predictor, which predicts five molecular functions of IDRs, including assembler, scavenger, effector, display site, and chaperone. DMFpred employs the Protein Cubic Language Model (PCLM) to learn protein representations, consisting of three types of protein language models and an attention-based language model alignment (ALAN) module. Three protein language models were used to capture protein sequences, structural, and functional features, respectively. The ALAN module extracts the relationship among three captured features and encodes the complementarity information. The key challenge in functional prediction is that the number of disordered sequences with functional annotations is relatively small. The transfer learning technology can transfer knowledge from tasks with plentiful training data to improve the performance of similar other tasks, which is especially useful for the task with limited training data [38]. Therefore, we first pre-trained PCLM with large IDRs sequences to capture the disordered features of proteins. Then the general disordered features were transferred separately to five different disorder functions prediction via model fine-tuning. Benefited from pre-training and function-specific fine-tuning of PCLM, DMFpred captures more relevant features of disorder molecular functions. The ablation experiment results demonstrated that each module of PCLM contributes to the predictive performance improvements. And further evaluation suggested that DMFpred provides high-quality predictions on all five categories of functional residues and multi-functional residues, whose residues carry more than one category of molecular functions. The corresponding web server of DMFpred was established and can be freely accessed from http://bliulab.net/DMFpred/.

Materials and methods

Benchmark datasets

The datasets used in this study were collected from DisProt [22], which is the major repository of manually curated functional annotations of intrinsically disordered proteins from literature. All sequences in the database are functionally annotated at the amino acid level. In this study,

Table 1. The number of functional residues in the DMF benchmark datasets.

Dataset	Number of Assembler residue	Number of Chaperone residue	Number of Display-site residue	Number of Effector residue	Number of Scavenger residue
Training set	14177	2236	5041	12783	2202
Evaluation set	7764	904	932	4102	1022
TEST-1	4412	836	1001	3980	545

https://doi.org/10.1371/journal.pcbi.1010668.t001

we focused on five general categories of disordered molecular functions (DMFs), including assembler, scavenger, effector, display site and chaperone. Following the intrinsically Disordered Proteins Ontology (IDPO) schema in the DisProt, each of the five categories of function terms has one or two leaf terms (see S1 **Fig**). Here, we treat all the leaf terms as the same functional class as their root terms. The sequences in the database are functionally annotated with amino acids as the basic unit, and we collected a total of 590 sequences containing residues assigned at least one class of DMFs. For each class of function, we treat the residues annotated with the functional term class in the database as functional residues, and the others as non-functional residues. Then we assign all the functional residues in the sequences as label '1' and non-functional residues as label '0', leading to five lines of labels corresponding to five categories of DMFs annotations.

To avoid data redundancy, we performed the similarity clustering on the 590 sequences by using PSI-BLAST [39] by setting the threshold of 25%, and filtered sequences with pairwise sequence similarity >25%. This way ensured that the sequence similarity between any two sequences in the collections was lower than 25%. The remaining 541 proteins were randomly divided into training, evaluation, and test sets in a ratio of 6:2:2. Finally, 324 sequences were used as the training set for model training, 106 sequences were used as the valuation set for model selection, and 111 sequences were used as the independent test set (TEST-1) to evaluate predictive performance (**S1 Data**). The number of functional residues for the five categories of disordered molecular functions in the DMF benchmark datasets is given in **Table 1**.

Architecture of protein cubic language model

Sequence, structure and function language models. Sequence, structure, and function are three important aspects of proteins. Only one language model cannot fully characterize the three features. In this paper, we employed three types of language models for capturing the sequences, structural, and functional features of proteins.

Sequence language model. The amino acid sequence contains the evolutionary information of protein. Here, the bidirectional long short-term memory (Bi-LSTM) networks were employed as the sequence language model to capture the global correlation features of evolutionary information (see **Fig 1A**). By using the protein PSSM profile and HMM profile as the inputs of the sequence language model, the sequence features **Seq** can be calculated by [40]:

$$\mathbf{Seq} = [\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_L] \tag{1}$$

$$\mathbf{h}_i = Concat[LSTM_f(\mathbf{X}_{L\times 40}), LSTM_b(\mathbf{X}_{L\times 40})] \tag{2}$$

where $\mathbf{X}_{L\times 40}$ is the combination of PSSM and HMM matrix generated by PSI-BLAST [39] and HH-suits [41] respectively, and L is the length of the sequence. $LSTM_f$ and $LSTM_b$ indicate the forward and backward recurrent neural unit respectively. *Concat* represents the combination of vectors.

Structure language model

Protein structure reflects the results of local interaction among residues. The structure language model aims to capture structural features of the protein, and a convolutional-based model is used to capture structural local pattern features from the residue-residue contact map (CCM) (see Fig 1B). By taking CCMs as inputs, the structure features **Stc** can be calculated by [42]:

$$\mathbf{Stc} = relu[\mathrm{Conv}(\mathbf{Y}_{L\times L}, \mathbf{Filter}_{stc}) + \mathbf{b}_{stc}] \tag{3}$$

where $\mathbf{Y}_{L\times L}$ is the CCM profile generated by CCMpred [43,44], $\mathbf{Filter}_{stc}$ and $\mathbf{b}_{stc}$ are trainable variables, Conv represents convolution operator, and *relu* is the Rectified Linear Unit activation function [45].

Function language model

Functional conservative sequence segments also known as functional motifs hold particular functionality information of proteins. Previous researches [46–48] have shown that the motif-based convolution (MotifConv) by embedding particular motifs into the convolution kernel can learn the prior biological features. Inspired by MotifConv, the functional motif-based convolution was employed as the function language model to capture proteins' functional features (see Fig 1C). The 164 motifs used in this study were extracted from the Eukaryotic Linear Motif (ELM) database [49]. The letter-probability matrix of each motif is used to build the convolution kernel formulated as:

$$\mathbf{M}_1 = \begin{bmatrix} a_{1,1} & \cdots & a_{1,20} \\ \vdots & \ddots & \vdots \\ a_{l,1} & \cdots & a_{l,20} \end{bmatrix} \tag{4}$$

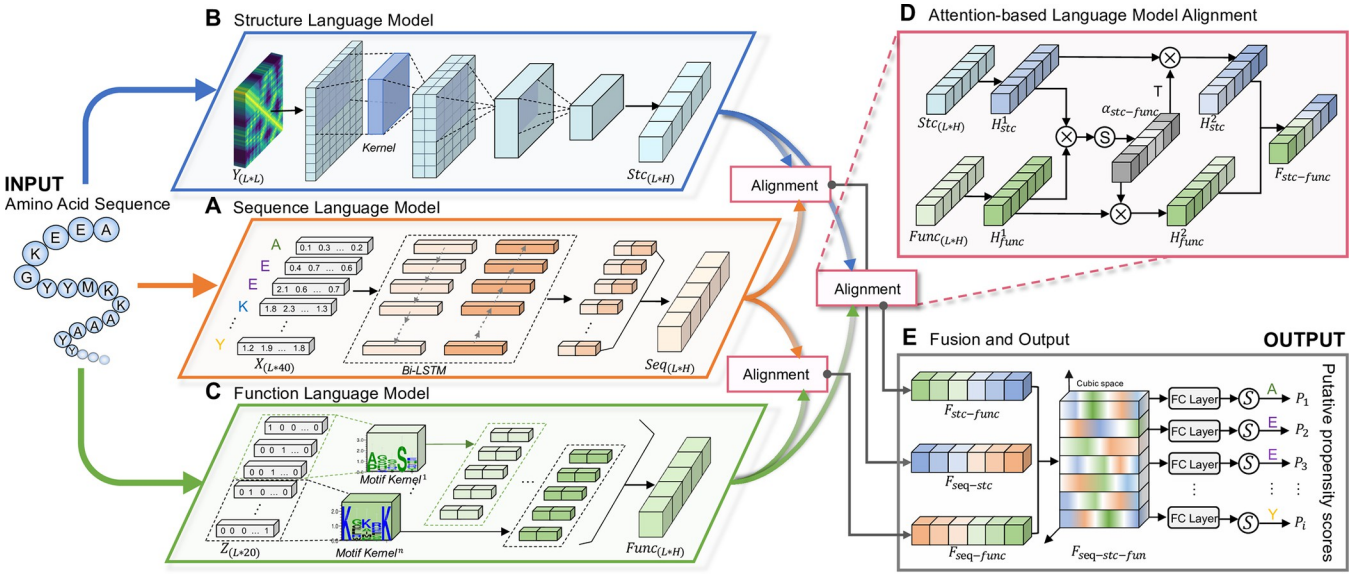

Fig 1. The architecture of protein cubic language model (PCLM). The PCLM contains five main modules: three protein language models (**A**. sequence, **B**. structure, and **C**. function language model), attention-based language model alignment module (**D.** ALAN), and the fusion and output layer (**E**). The input protein sequence is converted to sequence profile **X**, structure profile **Y**, and function profile **Z**, which are then fed into three protein language models to capture the sequence features **Seq**, the structure features **Stc**, and the function features **Func**. Next, three captured features are incorporated into the alignment features ($\mathbf{F}_{stc-func}$, $\mathbf{F}_{seq-stc}$ and $\mathbf{F}_{seq-func}$) by ALAN modules. Finally, the fusion and output layers merge the outputs of ALAN to calculate the propensity score P_i of disorder molecular function for each residue.

https://doi.org/10.1371/journal.pcbi.1010668.g001

where l is the length of motif, $a_{i,j}$ represents the frequency of standard amino acid. Then the function features **Func** can be calculated by:

$$\mathbf{Func} = relu[\mathrm{Conv}(\mathbf{Z}_{L\times 20}, \mathbf{M}) + \mathbf{b}_{func}] \tag{5}$$

where $\mathbf{Z}_{L\times 20}$ is the one-hot encoding matrix of protein sequence, **M** is the combination of 164 motif convolution kernel matrix, and $\mathbf{b}_{func}$ is trainable variable.

Attention based language model alignment

The primary sequences encode the disordered states of IDP/IDRs, which in turn determine functions. The potential correlations among sequence, structure and function are essential information for the protein representations. In this study, attention alignment models the correlations between protein features by calculating the attention alignment weights on two kinds of features (see Fig 1D). For example, given the sequence features **Seq**, structure features **Stc**, and function features **Func**, the attention-alignment weights $\boldsymbol{\alpha}_{seq-stc}$, $\boldsymbol{\alpha}_{seq-func}$ and $\boldsymbol{\alpha}_{stc-func}$ are calculated by:

$$\boldsymbol{\alpha}_{seq-stc} = softmax(\mathbf{H}^1_{seq}\mathbf{Seq} \times \mathbf{H}^1_{stc}\mathbf{Stc}) \tag{6}$$

$$\boldsymbol{\alpha}_{seq-func} = softmax(\mathbf{H}^1_{seq}\mathbf{Seq} \times \mathbf{H}^1_{func}\mathbf{Func}) \tag{7}$$

$$\boldsymbol{\alpha}_{stc-func} = softmax(\mathbf{H}^1_{stc}\mathbf{Stc} \times \mathbf{H}^1_{func}\mathbf{Func}) \tag{8}$$

where $\mathbf{H}^1_{seq}$, $\mathbf{H}^1_{stc}$ and $\mathbf{H}^1_{func}$ are the trainable weight variables. The attention-alignment weights between two kinds of features reflect matching patterns between different property aspects of the proteins. Weighted by the attention-alignment weights, the sequence features **Seq**, structure features **Stc** and function features **Func** captured by three language models can be enhanced and fused into the complementary features $\mathbf{F}_{seq-stc}$, $\mathbf{F}_{seq-func}$ and $\mathbf{F}_{stc-func}$:

$$\mathbf{F}_{seq-stc} = \mathrm{Concat}(\mathbf{H}^2_{seq}\boldsymbol{\alpha}^{\mathrm{T}}_{seq-stc}\mathbf{Seq}', \mathbf{H}^2_{stc}\boldsymbol{\alpha}_{seq-stc}\mathbf{Stc}') \tag{9}$$

$$\mathbf{F}_{seq-func} = \mathrm{Concat}(\mathbf{H}^2_{seq}\boldsymbol{\alpha}^{\mathrm{T}}_{seq-func}\mathbf{Seq}', \mathbf{H}^2_{func}\boldsymbol{\alpha}_{seq-func}\mathbf{Func}') \tag{10}$$

$$\mathbf{F}_{stc-func} = \mathrm{Concat}(\mathbf{H}^2_{func}\boldsymbol{\alpha}^{\mathrm{T}}_{stc-func}\mathbf{Stc}', \mathbf{H}^2_{stc}\boldsymbol{\alpha}_{stc-func}\mathbf{Func}') \tag{11}$$

where $\mathbf{H}^2_{seq}$ $\mathbf{H}^2_{stc}$ and $\mathbf{H}^2_{func}$ are the trainable variables, **Seq**′, **Stc**′ and **Func**′ indicate the transformed feature matrix of **Seq**, **Stc** and **Func**, respectively. The *softmax* is the activation function. The complementary features $\mathbf{F}_{seq-stc}$, $\mathbf{F}_{seq-func}$ and $\mathbf{F}_{stc-func}$ learn the correlations among sequence, structure, and functional properties of proteins, and these features are fed into the cubic fusion and output layers for calculating the predictive propensity score.

Cubic fusion and output layer

The cubic fusion module of PCLM merges the three alignment complementary features into latent cubic space, and obtains a joint representation matrix $\mathbf{F}_{seq-stc-func}$ of protein sequences:

$$\mathbf{F}_{seq-stc-func} = \mathbf{W}_x\mathbf{F}_{seq-stc} + \mathbf{W}_y\mathbf{F}_{seq-func} + \mathbf{W}_z\mathbf{F}_{stc-func} \tag{12}$$

$$\mathbf{F}_{seq-stc-func} = [\mathbf{F}_1, \cdots, \mathbf{F}_{L-1}, \mathbf{F}_L], \mathbf{F}_{seq-stc-func} \in R^{L\times n} \tag{13}$$

where L denotes the length of the input sequence, n denotes the dimension of features, $\mathbf{W}_x$,

$\mathbf{W}_y$, and $\mathbf{W}_z$ are the trainable weighted variables. Each vector $\mathbf{F}_i$ in the representation matrix represents the features of each residue in the sequence. The fully connected (FC) layer captures the global and local correlations between residues in the sequence so as to calculate the propensity score P_i for each residue:

$$\mathbf{P}[P_1, \ldots, P_i, \ldots, P_L] = Sigmoid(\mathbf{W}_f \mathbf{F}_{seq-stc-func} + \mathbf{b}_f) \quad (14)$$

where $\mathbf{W}_f$ *a*nd $\mathbf{b}_f$ represent the weighted and bias variables, respectively.

Pre-training of protein cubic language model

The transfer learning involves a model training strategy, which transfers the knowledge learned from the source domain to a new and different target domain. It is especially effective when the target domain has insufficient training data [38]. In this study, although we have relatively limited number of disorder functional annotation regions for PCLM model training, the number of intrinsically disordered regions (IDRs) is sufficient. The large number of IDRs will overcome the problem that model cannot be fully trained with insufficient disorder functional data, and the generic disordered features learned from IDR dataset can be transferred to facilitate the disorder molecular function prediction. Therefore, in this study, we employed the widely used IDP/IDR prediction benchmark dataset [40] as the pre-training dataset to pre-train PCLM model for predicting disordered regions in protein. To avoid data redundancy, we excluded sequences with >25% sequence similarity to the disordered functional benchmark datasets, and obtained 2639 sequences with 38134 IDRs and 1079 sequences with 16403 IDRs for model pre-training and validation, respectively (**S2 Data**). The binary cross-entropy loss function was used to calculate the loss score for model parameters optimizing [50]:

$$loss = -\sum_i [y_i log\ p_i + (1 - y_i) \log\ (1 - p_i)] \quad (15)$$

where p_i denotes the predictive score for residue R_i being disordered calculated by **Eq 14**, and y_i represents the actual label of disordered residue. The Adam optimizer [51] with a learning rate of 0.001 was employed to optimize the model parameters, and the model with the minimized loss score on the IDR validation set was saved as the pre-trained model.

Fine-tuning PCLM for predicting disordered molecular functions

In the fine-tuning stage, the pre-trained PCLM model was fine-tuned with functional specific data for predicting the disordered molecular functions in protein. Because of the differences between the five molecular functions, we separately fine-tuned PCLM with assembler, chaperone, display site, effector, and scavenger functional annotations in the DMF benchmark dataset, leading to five independent predicting PCLM models (see **Fig 2**). In the DMFpred

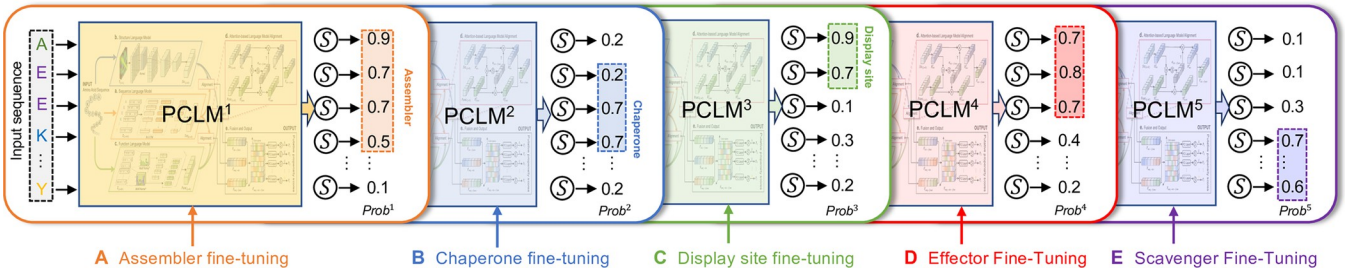

Fig 2. The functional specific fine-tuning of protein cubic language model (PCLM). The pretrained PCLM model was separately fine-tuned with five categories of disordered molecular functions into five corresponding PCLM$^{1\text{-}5}$ models for predicting assembler (**A**) chaperone (**B**) display site (**C**) effector (**D**) and scavenger (**E**) functional residues from the input sequence. $Prob_1$, $Prob_2$, $Prob_3$, $Prob_4$, and $Prob_5$ represent the predicted propensity scores for the five functions of input sequence, respectively.

https://doi.org/10.1371/journal.pcbi.1010668.g002

predictor, the five functional specific fine-tuned PCLM models work in parallel to produce five disordered molecular functional predictions for each residue in the input proteins. Here, we used the same loss function and optimizer as the ones used in the pre-training stage, but different learning rates to fine-tune the model parameters for each function. Parameters of all layers in PCLM were fine-tuned for achieving better performance, and this strategy has been adopted by many transfer learning based studies [52,53]. More detailed hyper-parameters for DMFpred are given in **S1 Table**.

Evaluation criteria

DMFpred generates two forms of outputs: the real-valued propensity score (the likelihood of residue with the given function) and binary results (residue with or without the given function). Binary predictions were converted from the propensities: one residue is predicted as functional residue if its propensity score is greater than a given threshold. Otherwise, it is predicted as the non-functional residue. The receiver operating characteristic curve (ROC) and AUC value (area under ROC curve) were utilized to evaluate the predictive performance of the real-valued propensity prediction. Sensitivity (Sn), specificity (Sp) and accuracy (ACC) were used for the evaluation of the binary results. Since the dataset is imbalanced, *i.e.* there are many more non-functional residues than the functional residues. Therefore, two metrics, balanced accuracy (BACC) and the Matthews Correlation Coefficient (MCC) were used to measure the predictive performance.

Disordered residues interact with multiple partners with more than one functions are called the multi-functional residues. The residue-level functional prediction of these multi-functional residues can be treated as a multi-label learning task, and five example-based metrics were utilized to evaluate the performance of DMFpred on multi-functional residues [54]:

$$\begin{cases} Hammingloss = \frac{1}{p}\sum_{i=1}^{p}\frac{1}{q}|h(x_i)\Delta Y_i| \\ Accuracy_{exam} = \frac{1}{p}\sum_{i=1}^{p}\frac{|h(x_i)\cap Y_i|}{|h(x_i)\cup Y_i|} \\ Precision_{exam} = \frac{1}{p}\sum_{i=1}^{p}\frac{|h(x_i)\cap Y_i|}{|h(x_i)|} \\ Recall_{exam} = \frac{1}{p}\sum_{i=1}^{p}\frac{|h(x_i)\cap Y_i|}{|Y_i|} \\ F1_{exam} = \frac{2\times Precision_{exam}\times Recall_{exam}}{Precision_{exam}+Recall_{exam}} \end{cases} \quad (16)$$

where p indicates the total number of samples, q indicates the number of labels, $h(x_i)$ is the predicted label set and Y_i is the true label set. Δ represents the symmetric difference between two sets.

Results and discussion

Functional specific fine-tuning achieves better performance

In order to investigate the differences among five categories of molecular functions, we performed the cross-functional validation on the benchmark datasets. To avoid the overestimation caused by the multi-functional residues, sequences that only belonging to one class function in the training and validation sets are used to fine-tune and test the PCLM model. The AUC evaluation results are shown in **Fig 3**. From **Fig 3**, we can see that model fine-tuned and tested on the same function achieves the best performance, while cross-functional

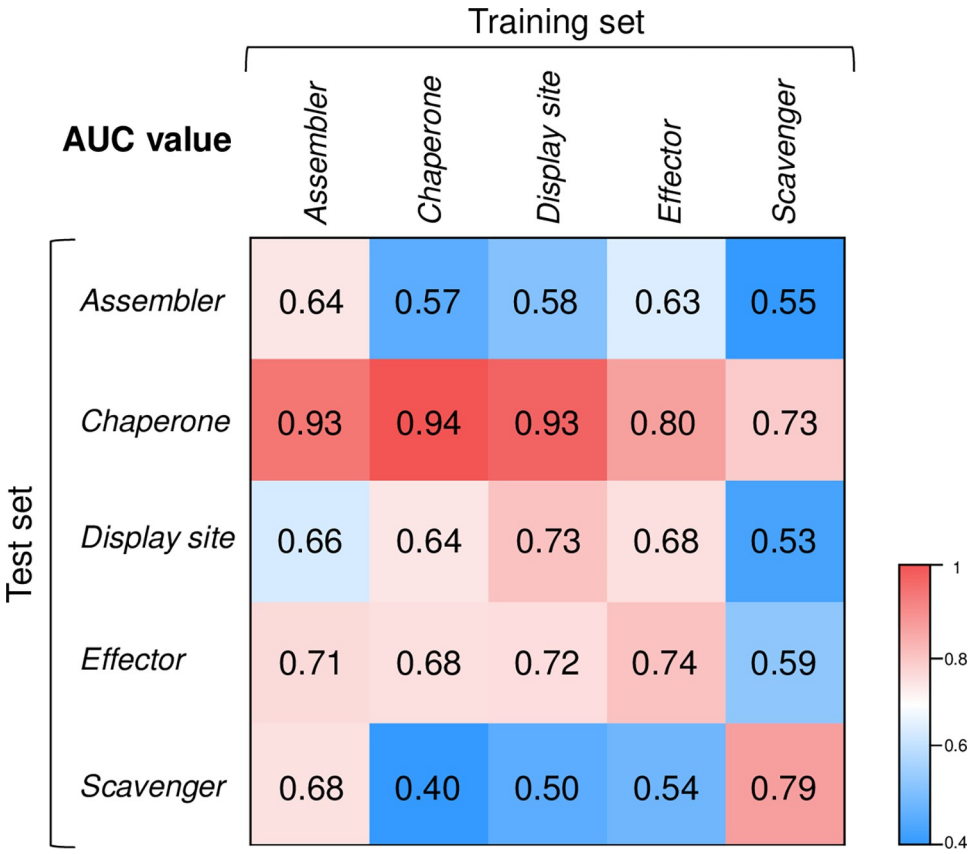

Fig 3. Cross-functional validation results. Functions on the x-axis were used to fine-tune PCLM model, and functions on the y-axis were used for model validation.

https://doi.org/10.1371/journal.pcbi.1010668.g003

predictors achieve lower performances. These predictive results suggest that specialized predictors are required for each functional category, and function-specific fine-tuning is the key to achieve better predictive performance of each disordered molecular function.

Ablation analysis of protein cubic language models

To verify the contribution of three language models to DMFpred, we performed an ablation analysis. The PCLM models with different combinations of three language models were individually fine-tuned on five molecular function training data, and the corresponding AUC values for each function evaluated on validation dataset are shown in Fig 4. We can see that (i) predictors with the combination of three language models consistently achieve the best performances for all five functions; (ii) the prediction performance of predictor decreased by dropping the structural language model. Predictors with only sequence language model performed the worst. These results are not surprising because three language models capture the sequence, structural, and functional features of proteins, and these three features are complementary, and contribute to the functional prediction. As a result, predictors incorporating the three protein language models achieve the best performance.

Attention based language model alignment learns the correlation patterns

In order to investigate the performance improvement of attention-based language model alignment (ALAN) to the proposed predictor. We compared the performance of predictors for

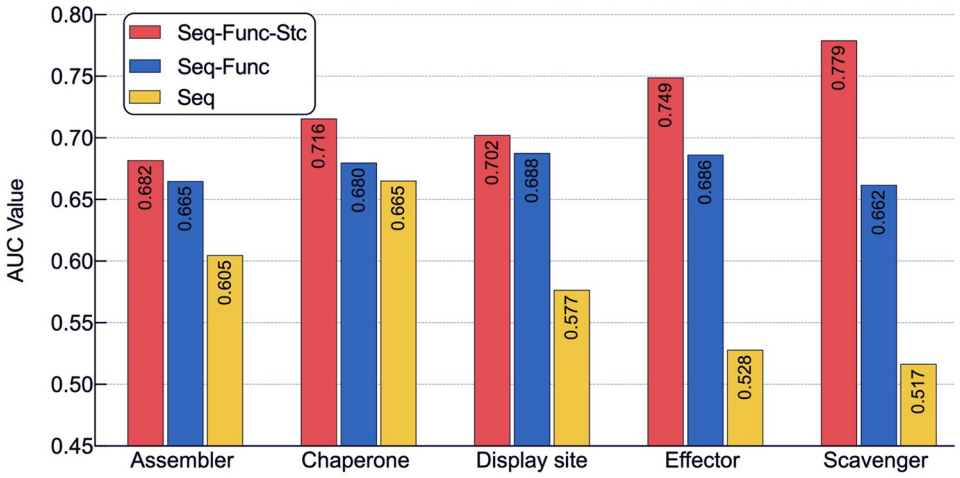

Fig 4. The predictive results of PCLM model in DMFpred with different language models. *Seq* represents the PCLM with only sequence language model, *Seq-Func* denotes the PCLM with the combination of sequence language model and function language model, and *Seq-Func-Stc* stands for the PCLM model, which is the combination of sequence language model, function language model and structure language model. The AUC values were calculated on the validation data set.

https://doi.org/10.1371/journal.pcbi.1010668.g004

predicting five disordered molecular functions by using PCLM model with and without the ALAN module. The PCLM model without ALAN directly feed the features captured by the three language models to the fusion and output layers (see Fig 1) to calculate prediction results. The two types of models were independently fine-tuned with five different functions, and the results evaluated on the validation dataset are shown in Fig 5. From this figure, we can see that predictors with ALAN consistently outperform the predictors without ALAN on five classes of functions, demonstrating the effectiveness of the ALAN module. Furthermore, we note that the predictor for Scavenger function with an ALAN achieves better performance in terms of AUC value. These results may be caused by the fact that the complementary features captured by the ALAN module supplemented the inadequate sequence, structure and functional features learned from limited annotated sequences. This improvement is especially manifested in the Scavenger function with a relatively small number of annotated sequences. Benefitted from the features captured by ALAN, predictor can make more accurate prediction leading to better performance.

To further analyse the information learned by the ALAN module, we visualized the attention-alignment weights between sequence and structure features. Two protein examples (DisProt ID: DP02925 and DP00284) selected from the independent test set (TEST-1) were visualized in Fig 6, from which we can see that the specific segments in the sequences map with the highest attention weights, and these sequence segments corresponding to the experimentally determined functional motifs searched from the ELM database [49] by FIMO tools (https://meme-suite.org/meme/tools/fimo). These results indicate that the ALAN can capture critical correlation patterns by modelling the relationship between different protein features. This prior biological knowledge captured by ALAN complements the original sequence, structure and functional attributes of proteins, providing a powerful protein representation.

Model pre-training facilitates feature correlation

In order to explore the contribution of model pre-trained with disordered proteins, we compare the predictive power of features extracted between models directly trained with molecular functional sequences (DT in Fig 7) and the fine-tuned model based on pre-training with IDRs (PT in Fig 7). Following previous studies [23, 24], the absolute point-biserial correlation

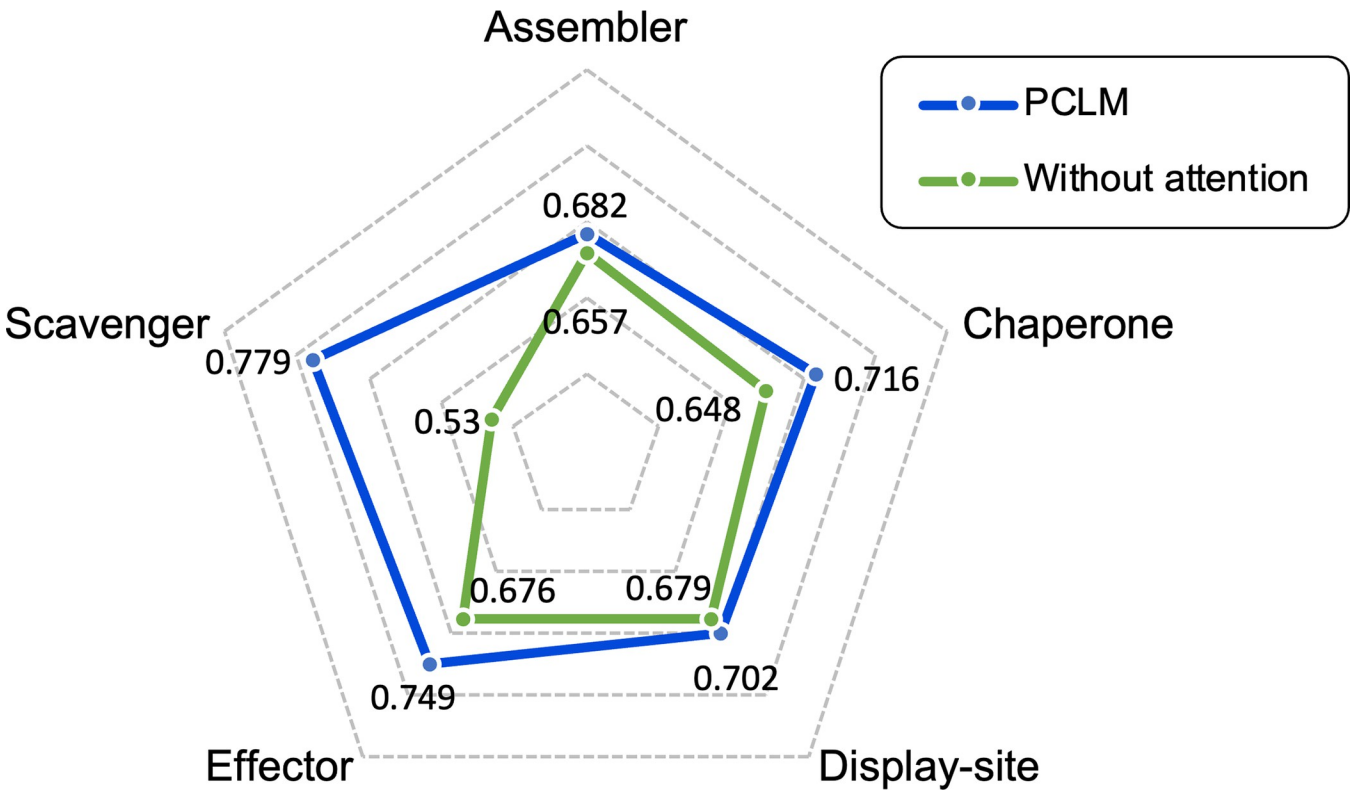

Fig 5. The predictive results (AUC values) of DMFpred with and without attention-based language model alignment. The *PCLM* represents the entire PCLM predictive model, while the *Without attention* denotes the PCLM model without the ALAN module. Both two models were independently fine-tuned and evaluated for the five molecular functions on the training dataset and validation dataset, respectively.

https://doi.org/10.1371/journal.pcbi.1010668.g005

(PBC) score is used to quantify the feature predictive qualities, which reflects the correlation between numeric and binary variables:

$$PBC = \frac{m_1 - m_0}{s_n}\sqrt{\frac{n_0 \times n_1}{n \times n}} \quad (17)$$

where n_0 and n_1 indicate the number of functional and non-functional residues, m_0 and m_1 indicate the average values of features of functional and non-functional residues, s_n is the standard deviation of all values of features, and n is the total number of residues. The PBC score results for five functions on the TEST-1 independent test set are shown in Fig 7. From this figure, we observe that the features captured by the pre-trained model are consistently outperformed that directly trained model on all five functions. This is because model pre-trained with IDR sequences captures more disordered features than directly trained on limited functional sequences. As the functional residues are the sub-set of disordered regions, the common disordered features captured by pre-trained model facilitate to distinguish disordered functional residues from ordered residues, leading to a robust predictive quality.

Overall results

To our best knowledge, DMFpred is currently the only predictor for predicting the five general molecular functions of disordered proteins. There are two forms of outputs of DMFpred: real-valued propensity results and binary results. We used the ROC curve and AUC value for evaluating the real-valued predictive results. Sn, Sp, ACC and two metrics for imbalanced datasets

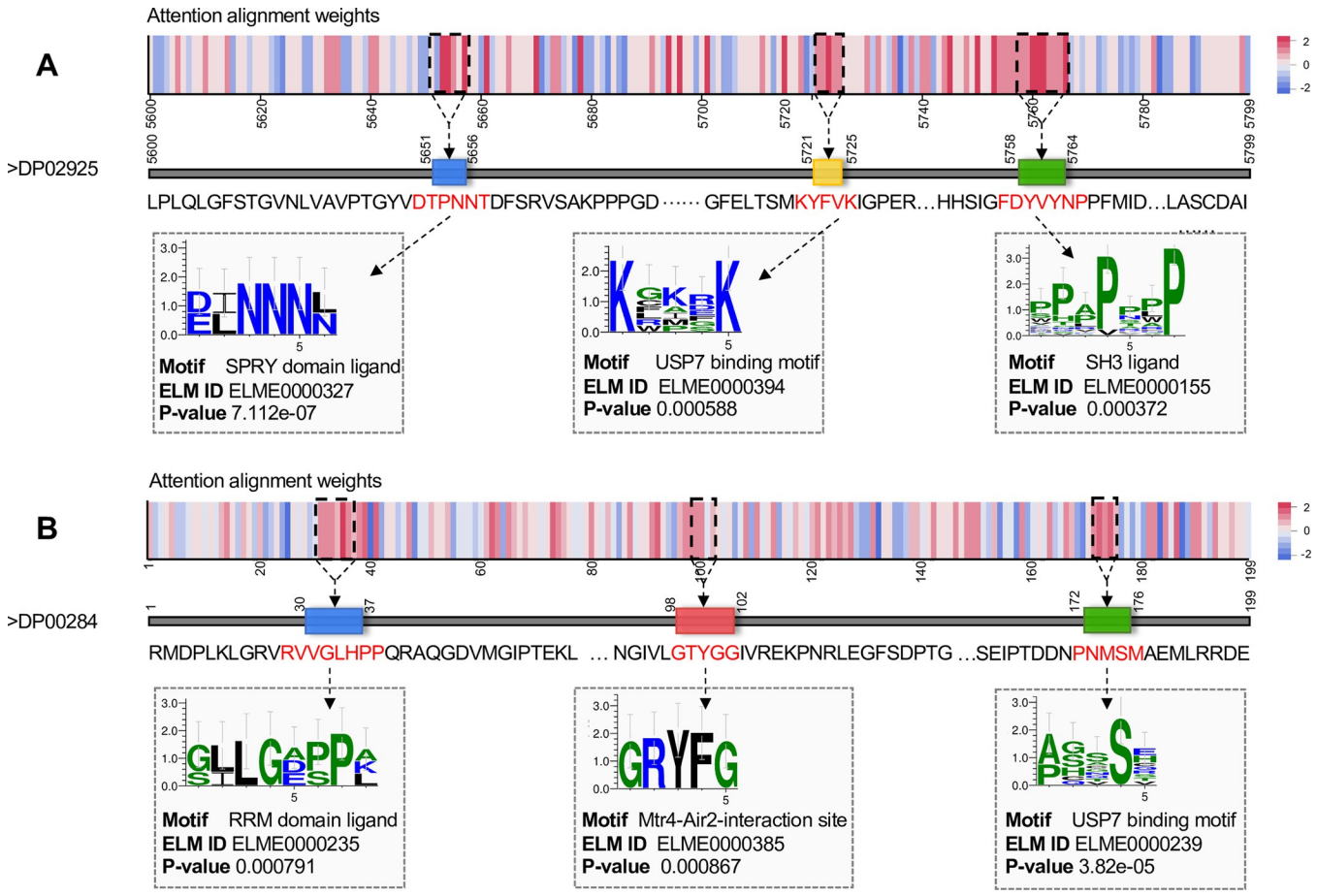

Fig 6. The attention alignment weight visualizations.

https://doi.org/10.1371/journal.pcbi.1010668.g006

(BACC and MCC) were used to assess binary results. The evaluation results on the TEST-1 independent test set are shown in Table 2 (the ROC curve and thresholds settings see S2 Fig, S2 **Table**). From Table 2, we can see that DMFpred provides accurate predictive performance for all five functional categories in terms of AUC values. The Sn, Sp and ACC results show the ability of DMFpred to correctly predict functional and non-functional residues, demonstrating the predictive performance.

In order to further evaluate the predictive performance of the predictor, we constructed a new independent test set (TEST-2) with the sequences newly added into the DisProt database during July 2021 to June 2022 by following the same dataset collection protocols. TEST-2 contains 47 proteins with 5780 functional residues, including 3753 assemblers, 218 chaperones, 855 display sites, 682 effectors and 272 scavengers. The prediction results of DMFpred on TEST-2 are shown in S3 Table. From these results, we can see that the predictive results achieved by DMFpred on the new independent test set TEST-2 are highly comparable with those on the independent test dataset TEST-1, indicating that the performance of DMFpred predictor is stable.

Predictive results on the multi-functional residues

The disordered residues interacting with multiple partners with more than one functions are called multi-functional residues. In order to investigate the performance of DMFpred

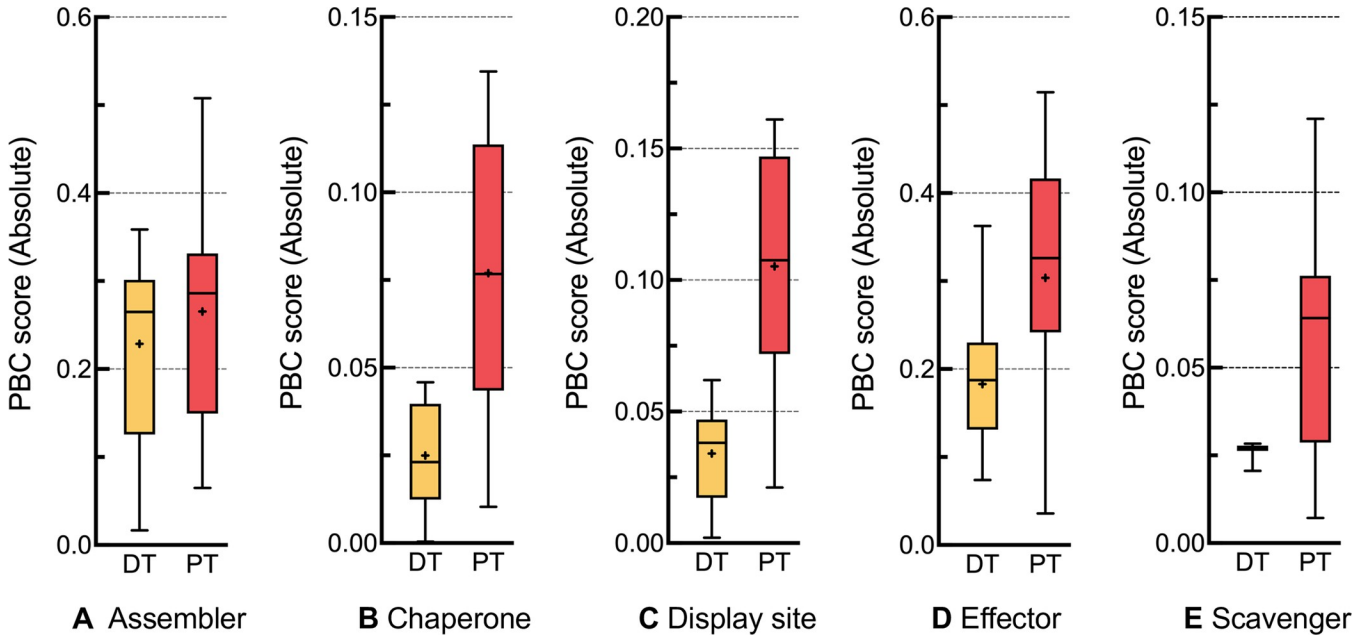

Fig 7. Absolute PBC score distributions on five functions. DT represents the PCLM models directly trained with molecular functional data, and PT represents the PCLM models pre-trained with disordered proteins.

https://doi.org/10.1371/journal.pcbi.1010668.g007

Table 2. The predictive performance of DMFpred for five categories of molecular functions on TEST-1 independent test set.

Function	AUC	Sn	Sp	ACC	BACC	MCC
Assembler	0.682	0.428	0.804	0.778	0.616	0.143
Chaperone	0.716	0.379	0.919	0.912	0.650	0.120
Display-site	0.702	0.291	0.962	0.952	0.627	0.155
Effector	0.749	0.663	0.741	0.736	0.703	0.215
Scavenger	0.779	0.999	0.520	0.524	0.761	0.095

https://doi.org/10.1371/journal.pcbi.1010668.t002

Table 3. The predictive results for multi-functional residues on TEST-1 independent test set.

Predictor	Hamming loss	$Accuracy_{exam}$	$Precision_{exam}$	$Recall_{exam}$	$F1_{exam}$
DMFpred	0.413	0.404	0.504	0.671	0.576
Baseline	0.508	0.285	0.409	0.485	0.444

https://doi.org/10.1371/journal.pcbi.1010668.t003

predictor for predicting these multi-functional residues, we collected all the residues with at least two functional annotations from TEST-1 dataset, and obtained a total number of 1352 multi-functional residues for performance evaluation. We compare DMFpred with a random baseline predictor generating the multi-functional labels for each residue with a probability of 0.5, and the evaluation results are shown in Table 3. From this table, we can see the followings: (i) compared with the baseline predictor, DMFpred achieves lower Hamming loss, but higher accuracy, which indicates DMFpred can accurately predict more multi-functional residues than the baseline predictor. (ii) DMFpred achieves higher performance than the baseline method in terms of precision, recall rate and F1 value. These results are not surprising because DMFpred was fine-tuned with function-specific labels on the benchmark dataset so as to learn the discriminative features of each function. Benefitting from the accurate prediction for five functions, DMFpred achieves better performance for predicting multi-functional residues.

Conclusion

Intrinsically disordered proteins/regions perform various molecular functions in living organisms. These functions of IDP/IDRs can be summarized as six general categories, including entropic chains, assembler, scavenger, effector, display site and chaperone. Motivated by the growing numbers of the annotated disordered sequences and the need to expand the coverage of disordered protein function predictors, we introduce the disordered molecular functional predictor called DMFpred, covering five important categories: disordered assembler, scavenger, effector, display site and chaperone. It has the following advantages: 1) DMFpred employed the protein cubic language model (PCLM) that incorporates three protein language models for characterizing sequence, structure, and functional attributes of proteins. PCLM employed attention-based language model alignment to capture the sequence-structure-function correlation and learn a joint representation of proteins. 2) Benefited from the pre-training and function-specific fine-tuning of PCLM, DMFpred captures discriminative features for five functional categories prediction. 3) The evaluation results on five categories of functional and multi-functional residues suggest that DMFpred provides high quality predictions. 4) The web-server of DMFpred is established and can be freely accessed from http://bliulab.net/DMFpred/, which will be helpful to researchers working on the related fields.

Supporting information

S1 Fig. The five disordered molecular functions and their sub-level annotations collected from the DisProt database.
(TIF)

S2 Fig. The ROC curves of DMFpred for predicting five disordered molecular functions.
(TIF)

S1 Table. The hyper-parameters of PCLM in DMFpred.
(DOCX)

S2 Table. The thresholds used for binary results of DMFpred. The thresholds were selected according to the highest MCC values in the validation set.
(DOCX)

S3 Table. The predictive performance of DMFpred on the TEST-2 independent test set.
(DOCX)

S1 Data. The molecular function benchmark dataset.
(DOCX)

S2 Data. The IDRs pre-training dataset.
(DOCX)

Author Contributions

Conceptualization: Yihe Pang, Bin Liu.

Data curation: Yihe Pang.

Formal analysis: Yihe Pang.

Funding acquisition: Bin Liu.

Investigation: Yihe Pang.

Methodology: Yihe Pang.

Project administration: Yihe Pang, Bin Liu.

Resources: Yihe Pang, Bin Liu.

Software: Yihe Pang.

Supervision: Bin Liu.

Validation: Yihe Pang, Bin Liu.

Visualization: Yihe Pang.

Writing – original draft: Yihe Pang.

Writing – review & editing: Yihe Pang, Bin Liu.

References

1. Xue B, Dunker AK, Uversky VN. Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life. J Biomol Struct Dyn. 2012; 30(2):137–49. Epub 2012/06/19. https://doi.org/10.1080/07391102.2012.675145 PMID: 22702725.
2. Peng Z, Yan J, Fan X, Mizianty MJ, Xue B, Wang K, et al. Exceptionally abundant exceptions: comprehensive characterization of intrinsic disorder in all domains of life. Cell Mol Life Sci. 2015; 72(1):137–51. Epub 2014/06/19. https://doi.org/10.1007/s00018-014-1661-9 PMID: 24939692.
3. Dunker AK, Brown CJ, Lawson JD, Iakoucheva LM, Obradovic Z. Intrinsic disorder and protein function. Biochemistry. 2002; 41(21):6573–82. Epub 2002/05/23. https://doi.org/10.1021/bi012159+ PMID: 12022860.
4. 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(13):6589–631. Epub 2014/04/30. https://doi.org/10.1021/cr400525m PMID: 24773235; PubMed Central PMCID: PMC4095912.
5. Tompa P. The interplay between structure and function in intrinsically unstructured proteins. FEBS Lett. 2005; 579(15):3346–54. Epub 2005/06/10. https://doi.org/10.1016/j.febslet.2005.03.072 PMID: 15943980.
6. Iakoucheva LM, Brown CJ, Lawson JD, Obradovic Z, Dunker AK. Intrinsic disorder in cell-signaling and cancer-associated proteins. J Mol Biol. 2002; 323(3):573–84. Epub 2002/10/17. https://doi.org/10.1016/s0022-2836(02)00969-5 PMID: 12381310.
7. Melo AM, Coraor J, Alpha-Cobb G, Elbaum-Garfinkle S, Nath A, Rhoades E. A functional role for intrinsic disorder in the tau-tubulin complex. Proc Natl Acad Sci U S A. 2016; 113(50):14336–41. Epub 2016/12/03. https://doi.org/10.1073/pnas.1610137113 PMID: 27911791; PubMed Central PMCID: PMC5167143.
8. Dev KK, Hofele K, Barbieri S, Buchman VL, van der Putten H. Part II: alpha-synuclein and its molecular pathophysiological role in neurodegenerative disease. Neuropharmacology. 2003; 45(1):14–44. Epub 2003/06/20. https://doi.org/10.1016/s0028-3908(03)00140-0 PMID: 12814657.
9. Cheng Y, LeGall T, Oldfield CJ, Mueller JP, Van YY, Romero P, et al. Rational drug design via intrinsically disordered protein. Trends Biotechnol. 2006; 24(10):435–42. Epub 2006/08/01. https://doi.org/10.1016/j.tibtech.2006.07.005 PMID: 16876893.
10. Uversky VN. Intrinsically disordered proteins and novel strategies for drug discovery. Expert Opin Drug Discov. 2012; 7(6):475–88. Epub 2012/05/09. https://doi.org/10.1517/17460441.2012.686489 PMID: 22559227.
11. Tompa P. Intrinsically unstructured proteins. Trends Biochem Sci. 2002; 27(10):527–33. Epub 2002/10/09. https://doi.org/10.1016/s0968-0004(02)02169-2 PMID: 12368089.
12. Daughdrill GW, Narayanaswami P, Gilmore SH, Belczyk A, Brown CJ. Dynamic behavior of an intrinsically unstructured linker domain is conserved in the face of negligible amino acid sequence conservation. J Mol Evol. 2007; 65(3):277–88. Epub 2007/08/28. https://doi.org/10.1007/s00239-007-9011-2 PMID: 17721672.
13. Uversky VN. Disorder in the lifetime of a protein. Intrinsically Disord Proteins. 2013; 1(1):e26782. Epub 2013/11/07. https://doi.org/10.4161/idp.26782 PMID: 28516024; PubMed Central PMCID: PMC5424783.

14. Daniels AJ, Williams RJ, Wright PE. The character of the stored molecules in chromaffin granules of the adrenal medulla: a nuclear magnetic resonance study. Neuroscience. 1978; 3(6):573–85. Epub 1978/01/01. https://doi.org/10.1016/0306-4522(78)90022-2 PMID: 692872

15. Holt C. Unfolded phosphopolypeptides enable soft and hard tissues to coexist in the same organism with relative ease. Curr Opin Struct Biol. 2013; 23(3):420–5. Epub 2013/04/30. https://doi.org/10.1016/j.sbi.2013.02.010 PMID: 23622834.

16. Galea CA, Wang Y, Sivakolundu SG, Kriwacki RW. Regulation of cell division by intrinsically unstructured proteins: intrinsic flexibility, modularity, and signaling conduits. Biochemistry. 2008; 47(29):7598–609. Epub 2008/07/17. https://doi.org/10.1021/bi8006803 PMID: 18627125; PubMed Central PMCID: PMC2580775.

17. Diella F, Haslam N, Chica C, Budd A, Michael S, Brown NP, et al. Understanding eukaryotic linear motifs and their role in cell signaling and regulation. Front Biosci. 2008; 13:6580–603. Epub 2008/05/30. https://doi.org/10.2741/3175 PMID: 18508681.

18. Young JC, Agashe VR, Siegers K, Hartl FU. Pathways of chaperone-mediated protein folding in the cytosol. Nat Rev Mol Cell Biol. 2004; 5(10):781–91. Epub 2004/10/02. https://doi.org/10.1038/nrm1492 PMID: 15459659.

19. Necci M, Piovesan D, Predictors C, DisProt C, Tosatto SCE. Critical assessment of protein intrinsic disorder prediction. Nat Methods. 2021; 18(5):472–81. Epub 2021/04/21. https://doi.org/10.1038/s41592-021-01117-3 PMID: 33875885; PubMed Central PMCID: PMC8105172.

20. 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(D1):D361–D7. Epub 2020/11/26. https://doi.org/10.1093/nar/gkaa1058 PMID: 33237329; PubMed Central PMCID: PMC7779018.

21. Piovesan D, Tabaro F, Micetic I, Necci M, Quaglia F, Oldfield CJ, et al. DisProt 7.0: a major update of the database of disordered proteins. Nucleic Acids Res. 2017; 45(D1):D219–D27. Epub 2016/12/03. https://doi.org/10.1093/nar/gkw1056 PMID: 27899601; PubMed Central PMCID: PMC5210544.

22. 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(D1):D269–D76. Epub 2019/11/13. https://doi.org/10.1093/nar/gkz975 PMID: 31713636; PubMed Central PMCID: PMC7145575.

23. Meng F, Kurgan L. DFLpred: High-throughput prediction of disordered flexible linker regions in protein sequences. Bioinformatics. 2016; 32(12):i341–i50. Epub 2016/06/17. https://doi.org/10.1093/bioinformatics/btw280 PMID: 27307636; PubMed Central PMCID: PMC4908364.

24. Peng Z, Xing Q, Kurgan L. APOD: accurate sequence-based predictor of disordered flexible linkers. Bioinformatics. 2020; 36(Suppl_2):i754–i61. Epub 2021/01/01. https://doi.org/10.1093/bioinformatics/btaa808 PMID: 33381830; PubMed Central PMCID: PMC7773485.

25. Sharma R, Bayarjargal M, Tsunoda T, Patil A, Sharma A. MoRFPred-plus: Computational Identification of MoRFs in Protein Sequences using Physicochemical Properties and HMM profiles. J Theor Biol. 2018; 437:9–16. Epub 2017/10/19. https://doi.org/10.1016/j.jtbi.2017.10.015 PMID: 29042212.

26. Disfani FM, Hsu WL, Mizianty MJ, Oldfield CJ, Xue B, Dunker AK, et al. MoRFpred, a computational tool for sequence-based prediction and characterization of short disorder-to-order transitioning binding regions in proteins. Bioinformatics. 2012; 28(12):i75–83. Epub 2012/06/13. https://doi.org/10.1093/bioinformatics/bts209 PMID: 22689782; PubMed Central PMCID: PMC3371841.

27. Hanson J, Litfin T, Paliwal K, Zhou Y. Identifying molecular recognition features in intrinsically disordered regions of proteins by transfer learning. Bioinformatics. 2020; 36(4):1107–13. Epub 2019/09/11. https://doi.org/10.1093/bioinformatics/btz691 PMID: 31504193.

28. Meszaros B, Simon I, Dosztanyi Z. Prediction of protein binding regions in disordered proteins. PLoS Comput Biol. 2009; 5(5):e1000376. Epub 2009/05/05. https://doi.org/10.1371/journal.pcbi.1000376 PMID: 19412530; PubMed Central PMCID: PMC2671142.

29. Meszaros B, Erdos G, Dosztanyi Z. IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. Nucleic Acids Res. 2018; 46(W1):W329–W37. Epub 2018/06/04. https://doi.org/10.1093/nar/gky384 PMID: 29860432; PubMed Central PMCID: PMC6030935.

30. Malhis N, Jacobson M, Gsponer J. MoRFchibi SYSTEM: software tools for the identification of MoRFs in protein sequences. Nucleic Acids Res. 2016; 44(W1):W488–93. Epub 2016/05/14. https://doi.org/10.1093/nar/gkw409 PMID: 27174932; PubMed Central PMCID: PMC4987941.

31. Sharma R, Sharma A, Raicar G, Tsunoda T, Patil A. OPAL+: Length-Specific MoRF Prediction in Intrinsically Disordered Protein Sequences. Proteomics. 2019; 19(6):e1800058. Epub 2018/10/17. https://doi.org/10.1002/pmic.201800058 PMID: 30324701.

32. Sharma R, Raicar G, Tsunoda T, Patil A, Sharma A. OPAL: prediction of MoRF regions in intrinsically disordered protein sequences. Bioinformatics. 2018; 34(11):1850–8. Epub 2018/01/24. https://doi.org/10.1093/bioinformatics/bty032 PMID: 29360926.

33. Peng Z, Kurgan L. High-throughput prediction of RNA, DNA and protein binding regions mediated by intrinsic disorder. Nucleic Acids Res. 2015; 43(18):e121. Epub 2015/06/26. https://doi.org/10.1093/nar/gkv585 PMID: 26109352; PubMed Central PMCID: PMC4605291.

34. Zhang F, Zhao B, Shi W, Li M, Kurgan L. DeepDISOBind: accurate prediction of RNA-, DNA- and protein-binding intrinsically disordered residues with deep multi-task learning. Brief Bioinform. 2022; 23(1). Epub 2021/12/15. https://doi.org/10.1093/bib/bbab521 PMID: 34905768.

35. Katuwawala A, Zhao B, Kurgan L. DisoLipPred: Accurate prediction of disordered lipid binding residues in protein sequences with deep recurrent networks and transfer learning. Bioinformatics. 2021. Epub 2021/09/07. https://doi.org/10.1093/bioinformatics/btab640 PMID: 34487138.

36. Wright PE, Dyson HJ. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. J Mol Biol. 1999; 293(2):321–31. Epub 1999/11/05. https://doi.org/10.1006/jmbi.1999.3110 PMID: 10550212.

37. Searls DB. The language of genes. Nature. 2002; 420(6912):211–7. Epub 2002/11/15. https://doi.org/10.1038/nature01255 PMID: 12432405.

38. Pan SJ, Yang Q. A survey on transfer learning. IEEE Transactions on knowledge and data engineering. 2009; 22(10):1345–59.

39. 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 (17):3389–402. Epub 1997/09/01. https://doi.org/10.1093/nar/25.17.3389 PMID: 9254694; PubMed Central PMCID: PMC146917.

40. Tang YJ, Pang YH, Liu B. IDP-Seq2Seq: identification of intrinsically disordered regions based on sequence to sequence learning. Bioinformatics. 2021; 36(21):5177–86. Epub 2020/07/24. https://doi.org/10.1093/bioinformatics/btaa667 PMID: 32702119.

41. Steinegger M, Meier M, Mirdita M, Vohringer H, Haunsberger SJ, Soding J. HH-suite3 for fast remote homology detection and deep protein annotation. BMC Bioinformatics. 2019; 20(1):473. Epub 2019/09/16. https://doi.org/10.1186/s12859-019-3019-7 PMID: 31521110; PubMed Central PMCID: PMC6744700.

42. Liu B, Li CC, Yan K. DeepSVM-fold: protein fold recognition by combining support vector machines and pairwise sequence similarity scores generated by deep learning networks. Brief Bioinform. 2020; 21 (5):1733–41. Epub 2019/10/31. https://doi.org/10.1093/bib/bbz098 PMID: 31665221.

43. Balakrishnan S, Kamisetty H, Carbonell JG, Lee SI, Langmead CJ. Learning generative models for protein fold families. Proteins. 2011; 79(4):1061–78. Epub 2011/01/27. https://doi.org/10.1002/prot.22934 PMID: 21268112.

44. Ekeberg M, Lovkvist C, Lan Y, Weigt M, Aurell E. Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. Phys Rev E Stat Nonlin Soft Matter Phys. 2013; 87(1):012707. Epub 2013/02/16. https://doi.org/10.1103/PhysRevE.87.012707 PMID: 23410359.

45. Nair V, Hinton GE, editors. Rectified linear units improve restricted boltzmann machines. Proceedings of the 27th International Conference on International Conference on Machine Learning; 2010.

46. Li CC, Liu B. MotifCNN-fold: protein fold recognition based on fold-specific features extracted by motif-based convolutional neural networks. Brief Bioinform. 2020; 21(6):2133–41. Epub 2019/11/28. https://doi.org/10.1093/bib/bbz133 PMID: 31774907.

47. Zhang J, Chen Q, Liu B. iDRBP_MMC: Identifying DNA-Binding Proteins and RNA-Binding Proteins Based on Multi-Label Learning Model and Motif-Based Convolutional Neural Network. J Mol Biol. 2020; 432(22):5860–75. Epub 2020/09/14. https://doi.org/10.1016/j.jmb.2020.09.008 PMID: 32920048.

48. 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. 2020;PP. Epub 2020/10/23. https://doi.org/10.1109/TCBB.2020.3031888 PMID: 33090951.

49. Kumar M, Gouw M, Michael S, Samano-Sanchez H, Pancsa R, Glavina J, et al. ELM-the eukaryotic linear motif resource in 2020. Nucleic Acids Res. 2020; 48(D1):D296–D306. Epub 2019/11/05. https://doi.org/10.1093/nar/gkz1030 PMID: 31680160; PubMed Central PMCID: PMC7145657.

50. Christoffersen P, Jacobs K. The Importance of the Loss Function in Option Valuation. CIRANO. 2003; 72(2):291–318.

51. Kingma D, Ba J. Adam: A Method for Stochastic Optimization. International Conference on Learning Representations2015. p. 1–11.

52. Singh J, Hanson J, Paliwal K, Zhou Y. RNA secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning. Nat Commun. 2019; 10(1):5407. Epub 2019/11/30. https://doi.org/10.1038/s41467-019-13395-9 PMID: 31776342; PubMed Central PMCID: PMC6881452.

53. Zhang J, Yan K, Chen Q, Liu B. PreRBP-TL: Prediction of Species-Specific RNA-Binding Proteins Based on Transfer Learning. Bioinformatics. 2022. Epub 2022/02/18. https://doi.org/10.1093/bioinformatics/btac106 PMID: 35176130.

54. Zhang M-L, Zhou Z-H. A review on multi-label learning algorithms. IEEE transactions on knowledge and data engineering. 2013; 26(8):1819–37.