Briefings in Bioinformatics, 2024, pp. 1–10

doi:
Paper

PAPER

IDP-EDL: Enhancing intrinsically disordered protein prediction by combining protein language model and ensemble deep learning

Junxi Xie [1], Xiaopeng Jin [1], Hang Wei [2] and Yumeng Liu [1,*]

[1] College of Big Data and Internet, Shenzhen Technology University, Shenzhen, 518118, Guangdong, China
[2] School of Computer Science and Technology, Xidian University, Xi'an, 710126, Shaanxi, China
[*] Corresponding author: Yumeng Liu, E-mail: liuyumeng@sztu.edu.cn

Abstract

Identification of intrinsically disordered regions (IDRs) in proteins is essential for understanding fundamental cellular processes. An accurate identification method of IDRs needs effective protein representations and appropriate algorithms. In previous studies, most feature construction was based on protein sequence profiles from multiple sequence alignment. However, the increase in the size of protein sequences causes computational complexity to increase sharply, posing a significant challenge in bioinformatics analysis. In this paper, we propose an accurate and stable method, IDP-EDL, based on a pretrained model and ensemble deep learning. IDP-EDL fuses three individual deep learning models, which can be divided into a Generic Features Extractor (GFE), a Specific Representations Extractor (SRE) , and a prediction layer. The GFE module utilizes a pretrained model to generate the generic features of proteins, thus bypassing time-consuming database searches. The SFE module further captures the specific representations of different types of disordered regions, including long disordered regions (LDRs) and short disordered regions (SDRs). The final prediction output is computed by a weighted voting of the results from these three models. In experiments, the feature method based on the pretrained protein language model ProteinBert achieved the best results. The ensemble deep learning model IDP-EDL can realize improvements in accuracy and stability compared to individual deep learning models. Compared with other methods, when evaluated on independent test sets, IDP-EDL showed equivalent or better performance. IDP-EDL is available at https://github.com/joestarXjx/IDP-EDL.

Key words: intrinsically disordered regions; pretrained protein language model; ensemble deep learning

Introduction

Intrinsically disordered proteins (IDPs) are protein regions that lack a stable three-dimensional structure under physiological conditions [1]. Intrinsically disordered regions (IDRs) are correlated to many important biological functions, and are widely involved in important physiological processes [1, 2, 3], such as regulation of transcription and translation, storage of small molecules, cellular signal transduction, and protein phosphorylation. Some diseases are also related to IDRs, such as cancer [2] and Alzheimer's disease [4]. Therefore, accurate identification of IDRs is essential in various biological processes.

Intrinsic disorder in proteins has been studied experimentally by methods including X-ray crystallography, nuclear magnetic resonance (NMR), and circular dichroism (CD) [5, 6]. However, these methods are not suitable for high-throughput data for reason of cost and time consumption. Many computational methods have been established to address these challenges.

Computational method typically include three essential components: sufficient training data, effective protein sequence features, and suitable models [7]. Several databases of experimentally determined IDRs have been established in recent years, such as Disprot [8] and MobiDB [9], offering an opportunity to train reliable computational models that accurately predict IDPs. Features of proteins should be extracted and encoded as numerical vectors for use in machine learning- and deep learning-based predictive modelling. Protein representation methods can be categorized as classical (i.e., model-driven) or data-driven [10], of which the first method employs predefined rules about properties that encapsulate the evolutionary relationships between proteins or the physicochemical properties of amino acids. Moreover, data-driven representations leverage statistical and machine learning algorithms trained for predefined tasks, such as the prediction of the next amino acid in a sequence [10]. Most previous studies have utilized classical representations as features [11], such as PSSM (Position-Specific Scoring Matrix) [12] and the seven physicochemical properties [9]. However, these methods, which are based on multiple sequence alignment, are time-consuming, and they require considerable computational resources. Recently, a few studies have pretrained deep neural

language models on protein sequences [13], such as ESM [14], TAPE-Transformer [15], ProtTrans [16], and ProteinBert [13]. These protein language models learn the implicit biochemical properties, secondary and tertiary structures, and inherent functional rules in protein sequences. Some studies have utilized pretrained models to extract features from protein sequences for downstream tasks, covering protein function, structure, post-translational modifications, and biophysical properties. These tasks have demonstrated promising results, inspiring us to use protein language models for disorder prediction. For example, LMDisorder [17] employed ProtTrans[16] to predict the intrinsic disorder regions of protein.

The exceptional flexibility and adaptability of ensemble methods and deep learning models have facilitated their extensive application in bioinformatics research. These two machine learning techniques have been commonly regarded as independent approaches in bioinformatics applications. AUCpreD [18] trains with a maximum-AUC algorithm combining Conditional random field and deep convolutional neural network, SPOT-Disorder [19] uses a model built with deep bidirectional LSTM recurrent neural networks in the problem of protein intrinsic disorder prediction, and RFPR-IDP [20] combines a convolutional neural network (CNN) and bidirectional long short-term memory. MFDp [21] integrates the DISOPRED2, DISOclust, and IUCpred models. SPOT-Disorder2 [22] combines five deep learning networks, fusing a residual convolutional network and long short-term memory. Some methods, such as SPINE-D [23], IDP-Seq2Seq [24] and IDP-FSP [25], further divide IDRs into long disordered regions (LDRs) and short disordered regions (SDRs), fusing length-dependent models trained separately on the corresponding dataset that includes disordered regions with specific length.

The emergence of ensemble deep learning, which combines the two machine learning techniques to achieve synergistic improvements in model accuracy and stability, has prompted a new wave of research and application. We propose an ensemble deep learning model, IDP-EDL, to predict disordered regions. This model integrates three deep learning models: IDP-EDL-G, IDP-EDL-L and IDL-EDL-S. The neural network of each model is trained on different types of datasets. The model includes a Generic Features (GFE), Specific Representations Extractor (SRE) and prediction layer. The GFE module utilizes the pretrained protein language model ProteinBert [13] to extract generic features of protein sequences. The SRE module further captures the features of different types of disordered regions, including both long and short disordered regions [23], with an attention mechanism to calculate the global associations among residues, and a CNN to capture local features. The final results are computed by the weighted voting of the predictive results of the three deep learning models.

Materials and methods

Datasets

The datasets used in this study were obtained from previous studies [26, 24], and contain 5589 protein sequences, where the sequence similarity between any two proteins is less than 25%. These proteins were randomly divided into training and validation datasets containing 4360 and 1229 proteins, respectively. Generally, long disordered regions (LDRs) are defined as disordered regions with more than or equal to 30 residues, while short disordered regions (SDRs) have fewer than 30 residues [23]. Based on the length of disordered regions, we further divided the training dataset into LDR and SDR sub-training datasets, the first containing proteins with at least one LDR, and the second containing proteins with at least one SDR but without LDRs. The validation set was treated in the same manner. The datasets are described in Table 1, and can be formatted as

$$\begin{cases} S_{all}^{train} = S_{long}^{train} \cup S_{short}^{train} \\ S_{all}^{valid} = S_{long}^{valid} \cup S_{short}^{valid} \end{cases} . \quad (1)$$

Three independent test datasets with different ratios of LDRs and SDRs were used to comprehensively compare the performance of different methods: MXD494 [27], SL329 [28], and Disprot504 [18]. Table 2 lists the number of proteins, ordered proteins, and disordered residues in each datasets, as well as the percentage of each type.

Residue representation

The extraction of numerical features from protein sequences is necessary before they can be utilized by machine learning or deep learning algorithms. The application of large language models has revolutionized the way protein sequences are

Table 1. Statistical information of training and validation datasets

Dataset	Residue level		Protein level	
	Disordered residue (percent)	Ordered residue (percent)	LDR protein (percent)	SDR protein (percent)
S_{all}^{train}	149183 (11.6%)	1135478 (88.4%)	872 (20.0%)	3488 (80.0%)
S_{long}^{train}	100586 (26.3%)	281964 (73.7%)	872 (100.0%)	0 (0.0%)
S_{short}^{train}	48597 (5.4%)	853514 (94.6%)	0 (0.0%)	3488 (100.0%)
S_{all}^{valid}	29082 (9.5%)	276748 (90.5%)	144 (11.7%)	1085 (88.3%)
S_{long}^{valid}	12504 (26.8%)	34159 (73.2%)	144 (100.0%)	0 (0.0%)
S_{short}^{valid}	16578 (6.4%)	242589 (93.6%)	0 (0.0%)	1085 (100.0%)

Table 2. Statistical information of independent test datasets

Dataset	Residue level		Protein level	
	Disordered residue (percent)	Ordered residue (percent)	LDR protein (percent)	SDR protein (percent)
MXD494	44087 (22.4%)	152414 (77.6%)	248 (50.2%)	246 (49.8%)
SL329	39544 (42.4%)	51292 (57.6%)	234 (71.1%)	95 (28.9%)
Disprot504	74454 (24.7%)	226992 (75.3%)	504 (100.0%)	0 (0.0%)

analyzed and interpreted. These models, inspired by natural language processing models like BERT [29, 30] and GPT, have been adapted and fine-tuned to process and extract meaningful information from protein sequences. Protein language models are trained on massive amounts of sequence data to encode protein sequences into dense, continuous vector representations, capturing complex structural and functional information.

We leveraged the pretrained protein language model ProteinBert to extract the sequence embedding. ProteinBert improves upon the classic Transformer/BERT architecture, and takes advantage of the unique characteristics of proteins [13]. We obtain residue-level features derived from the hidden states from the last layer of ProteinBERT, which is a fixed-size matrix. The resulting embedding serves as residue representations for protein sequences.

IDP-EDL architecture

Figure 1 shows the IDP-EDL architecture. The framework fuses three deep learning models, each trained separately on the corresponding dataset [31]. Each model includes a Generic Features Extractor (GFE), Specific Representations Extractor (SFE) and prediction layer. The protein sequence is input to the GFE module to generate the generic features of protein sequences. The SFE module further captures specific characteristics of different types of disordered regions, including LDRs and SDRs. The prediction layer predicts the propensity of each residue to be disordered. The final prediction results are obtained by the weighted voting of the output results from the three deep learning models [32].

Generic Features Extractor

The GFE module utilizes the pretrained ProteinBert model to yield the generic features of protein sequences [13]. Formally, a protein sequence can be represented as:

$$S = a_1, a_2, a_3, ..., a_L, \tag{2}$$

where a_i is the residue at position i and L is the length of the sequence. The protein sequence is encoded into a fixed-length vector composed of 26 integer tokens,

$$T = t_1, t_2, t_3, ..., t_n, \tag{3}$$

where n is the fixed sequence length chosen for the batch, and t_i represents the token of the ith residue. These tokens represent the 20 standard amino acids, selenocysteine (U), an undefined amino acid (X), another amino acid ($other$), and three additional tokens ($start$, end and pad) [13]. The fixed-length vectors are input to the pretrained layers, yielding the residue-level feature matrix,

$$X = x_1, x_2, x_3, ..., x_n, \tag{4}$$

where x_i is the feature vector of the ith residue. The corresponding mask vector, $M = m_1, m_2, m_3, ..., m_n$, is generated simultaneously, where the value at position i is

$$m_i = \begin{cases} 0, & \text{if } t_i \text{ is } start, end \text{ or } pad \\ 1, & \text{otherwise.} \end{cases} \tag{5}$$

Specific Representations Extractor

The SRE module further captures specific characteristics of different types of disordered regions (LDRs and SDRs) from the features extracted by the GFE module. The SRE modules are composed of different neural networks, including an attention-based SRE module and a CNN-based SRE module, which are utilized by the three independent deep learning models. The attention-based SRE module is employed by IDP-EDL-G and IDP-EDL-L, and is composed of an additive self-attention layer [33] and a bidirectional bidirectional Gate Recurrent Unit (Bi-GRU) layer [34]. The attention layer is used to calculate the global associations among residues. As shown in Figure 2, the protein sequence features extracted by the GFE block are input to the additive self-attention layer. The attention scores are computed through a feedforward neural network with a single hidden layer,

$$e_{ij} = V \tanh(W_k x_i + W_q x_j), \tag{6}$$

where W_k and W_q are trainable weight matrices, and V is a trainable weight vector. The padded parts of the sequence are disregarded via the mask vector M, and the attention weights between the ith residue and jth residue $alpha_{ij}$ are calculated using the softmax function. Then the attention vector $attn_j$ can be calculated as the weighted sum of the attention weights and the input vector,

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{j=1} \exp(e_{ij})} \tag{7}$$

$$attn_j = \sum_{i=1} \alpha_{ij} h_i. \tag{8}$$

The feature vector x_j and attention vector $attn_j$ are concatenated and fed into the Bi-GRU layer to further capture long-term dependency information.

The CNN-based SRE module within IDP-EDL-S is composed of two Bi-GRU layers and a CNN [35]. As shown in Figure 3, the protein sequence features are fed into the first Bi-GRU layer, which returns the entire sequence of hidden state vectors, $H = h_1, h_2, h_3, ..., h_n$, which can be calculated as

$$\begin{aligned} h_i &= \text{Bi-GRU}(x_i) \\ &= \overrightarrow{\text{GRU}}(x_i) \oplus \overleftarrow{\text{GRU}}(x_i) \\ &= (\overrightarrow{\mathbf{h}}_i \oplus \overleftarrow{\mathbf{h}}_i), \end{aligned} \tag{9}$$

where i is the time step of predicting the ith residue. The CNN is used to better extract the features of SDRs [35]; its structure is shown in the Figure 4.

Prediction layer

The vectors $H = h_1, h_2, h_3, ..., h_n$, as obtained from the SRE module, are input to a fully connected layer to obtain the final output, $Y = y_1, y_2, y_3, ..., y_n$, i.e.,

$$Y = sigmoid(HW + B), \tag{10}$$

where H is a learnable weight matrix, and b is a bias term. For a target residue, the final output probability predicted by IDP-PLM is computed as the weighted sum of the predictive results of the three models,

$$Output = 0.5a + 0.25b + 0.25c, \tag{11}$$

where a, b, and c are the propensities of the target residue to be disordered, as predicted by IDP-PLM-G, IDP-PLM-L,and IDP-PLM-S, respectively. When the output probability

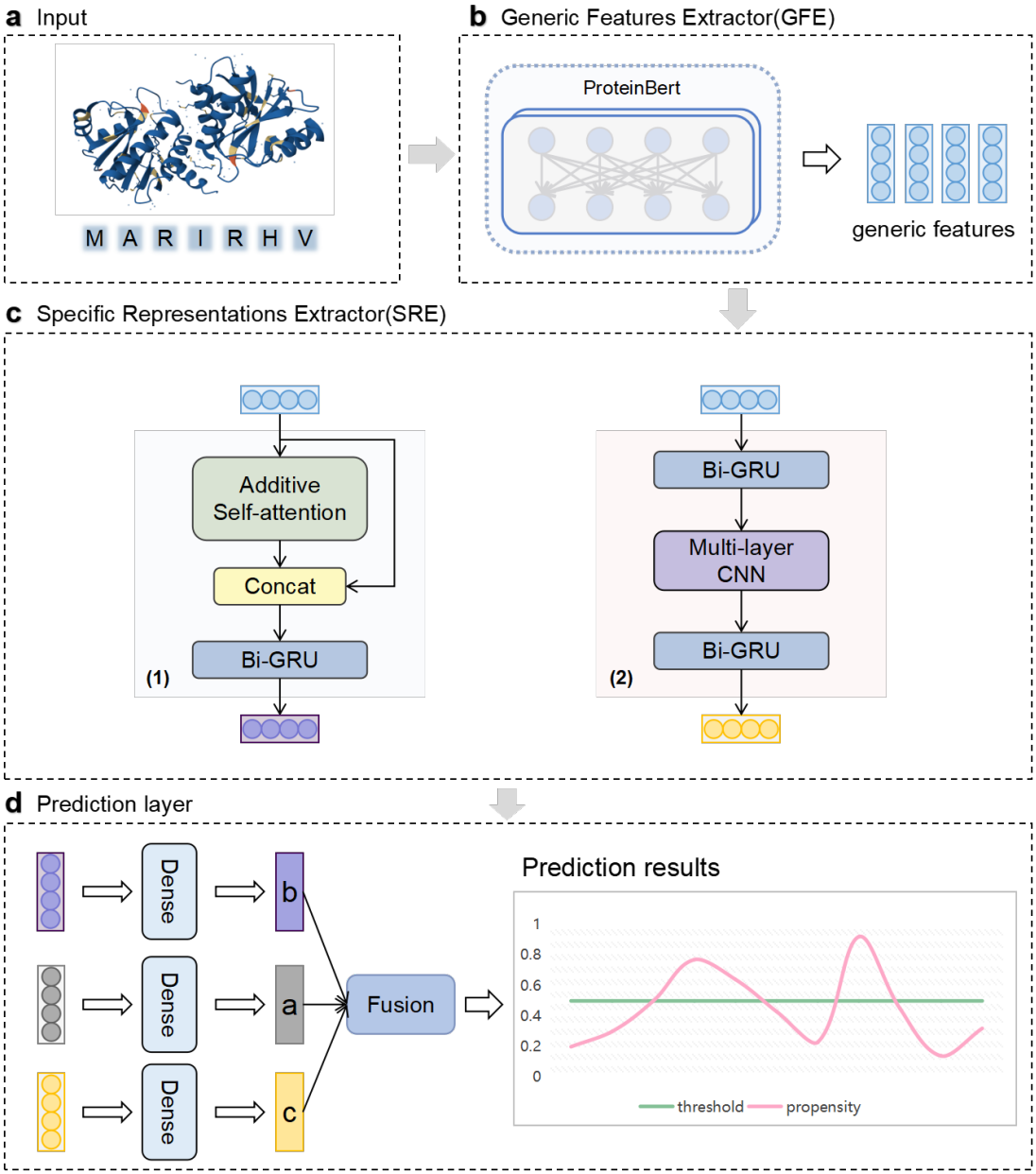

Fig. 1. IDP-EDL architecture. (a) Protein sequence to be predicted is input; (b) GFE module is used to extract generic features; (c) SFE module futher capture features of different types of disordered regions, using different neural networks. Neural network (1) is constructed from IDP-EDL-G and IDP-EDL-L, and neural network (2) is constructed from IDP-EDL-S; (d) Output of SRE module is sent to prediction layer to calculate final prediction results.

is greater than or equal to 0.5, the target residue is predicted as disordered residue, and otherwise as ordered residue. The fusion method is described as follows: from a total of four votes, IDP-EDL-L and IDP-EDL-S each has one vote, and IDP-EDL-G has two votes, as it is responsible for predicting both LDRs and SDRs. The prediction results of IDP-EDL can be obtained according to Equation 11.

Evaluation metrics

To predict IDRs in proteins is a binary classification task. We access the performance of binary classification models [22] using the metrics of Sn (Sensitivity), Sp (Specificity), MCC (Matthew's correlation coefficient), BACC (Balanced Accuracy), F1 score, and AUC (Area under the ROC curve). These metrics are calculated as

$$\begin{cases} S_n = \frac{TP}{TP+FN} \\ S_P = \frac{TN}{TN+FP} \\ BACC = \frac{1}{2}(Sn + Sp) \\ F1 = 2 \times \frac{TP}{2\times\ TP+FP+FN} \\ MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}, \end{cases} \tag{12}$$

where TP, TN, FP, and FN are the respective numbers of true positives, true negatives, false positives, and false negatives.

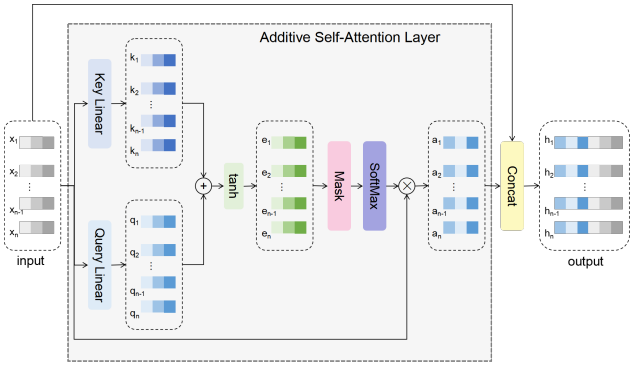

Fig. 2. Structure of additive self-attention layer.

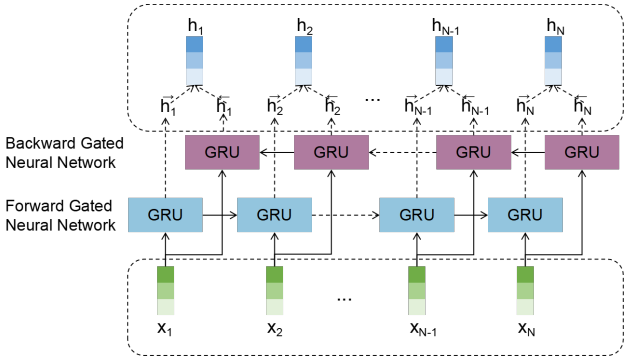

Fig. 3. Structure of bidirectional Gated Recurrent Unit layer.

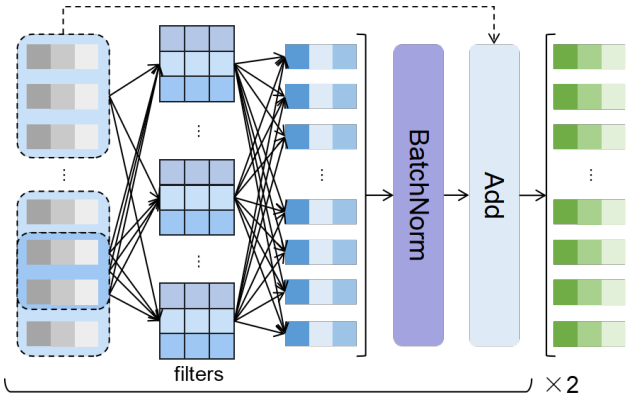

Fig. 4. Structure of multi-layer Convolutional Neural Network.

Results and discussion

Feature comparison

In the realm of bioinformatics, it is essential to extract the features of protein sequences for the prediction of IDRs. Some state-of-the-art methods generate the protein features based on sequence profiles from multiple sequence alignment, from evolution-related, residue-level, and structure-level perspectives. Pretrained language models are seeing increased use as encoders to generate features of proteins. We employ a lightweight pretrained model, ProteinBert, to extract generic features of proteins for the prediction of IDRs.

We compare three commonly used features to those generated by ProteinBert. Evolution-related features are constructed by the Position-Specific Scoring Matrix (PSSM), which captures the evolutionary information of protein sequences. Residue-related features are constituted by a matrix from the AAindex database, which can represent the physicochemical properties of individual amino acids and their dyadic combinations. Structure-level features are constituted by the matrix of potential energy arising from contact between amino acid pairs within protein structures. For a fair test, each experimental model uses the same hidden layer (a single Bi-GRU layer), and is independently trained on the same training dataset. For the test dataset, we selected proteins with sequence lengths less than 512 in the validation set.

Table 3 shows the comparison results on the test set of each method. Among them, the combination of PSSM, AAindex, and Energy features can improve model performance, as compared to the individual PSSM, AAindex, or Energy features, indicating that these three features are complementary, and together, can more comprehensively express the properties of proteins. Our feature method performs best, with MCC, BACC, and AUC of 0.550, 0.717, and 0.885, respectively, which are 15.3% (MCC), 8.6% (BACC), and 3.1% (AUC) higher than the combination method (PSSM+AAindex+Energy). This indicates that the features generated by ProteinBert are effective for the prediction of IDRs, and ProteinBert can capture the evolution information and structure information of proteins and the properties of each amino acid.

Table 3. Performance comparison of features on test dataset.

Features	Sn	Sp	BACC	MCC	AUC
PSSM	0.323	0.989	0.656	0.467	0.856
AAindex	0.286	0.992	0.639	0.449	0.853
Energy	0.311	0.990	0.651	0.460	0.857
PSSM+AAindex	0.318	0.991	0.654	0.471	0.857
PSSM+AAindex+Energy	0.331	0.990	0.660	0.477	0.858
ProteinBert	0.450	0.984	0.717	0.550	0.885

Training methods for deep learning models

There are two approaches to the employment of pretrained protein language models in bioinformatics. The first method engenders the pretrained model's role as a feature extractor, converting the raw sequences into a fixed-size feature matrix, which is subsequently input to the target model. The second approach incorporates the pretrained model as part of the target model. We explore the suitability of these methods for our deep learning models. As shown in Figure 5, the pretrained layers are frozen, and only the newly added layers are allowed to train in the first method. In the second approach, the entire model, including pretrained and newly added layers, is trained based on transfer learning. We propose a method that combines these two approaches, initially freezing the pretrained layers while training the other layers, and then unfreezing all layers while training the entire model.

In the ablation study, all methods utilized the same network architecture, including the GFE module and a prediction layer. We initialized weights from the pretrained ProteinBert model. In the first and second sets of experiments, the model was trained on the LDRs (S_{long}^{train}) and SDRs (S_{short}^{train}) training dataset, respectively. In the third set of experiments, the model was trained on the mixed training dataset, S_{all}^{train}. All methods were tested on the validation dataset with the

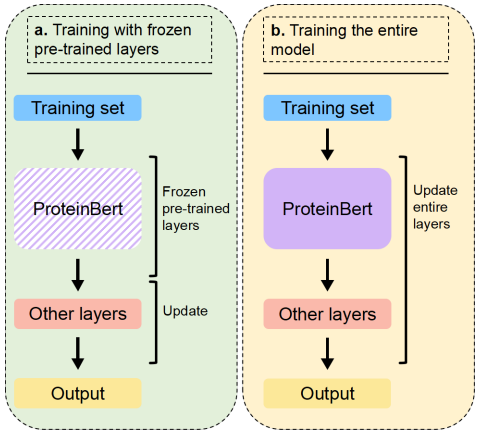

Fig. 5. Implementation of training method

same disordered region type. Table 4 compares the results of these experiments. We can see that the method combining two approaches consistently outperforms the other two methods, because it retains the generic features learned by the pretrained model, and adjusts its weights and biases according to the IDR prediction data, thereby better adapting to identify disordered regions in proteins. We implemented all training using this method.

Table 4. Performance of implementation methods on different datasets.

Test set	Method	Sn	Sp	BACC	MCC	AUC
S_{long}^{valid}	Freezing	0.533	0.960	0.747	0.579	0.857
	Fine-tuning	0.567	0.959	0.763	0.604	0.869
	Combination	0.580	0.949	0.765	0.594	0.871
S_{short}^{valid}	Freezing	0.374	0.992	0.683	0.515	0.897
	Fine-tuning	0.406	0.990	0.698	0.529	0.900
	Combination	0.404	0.991	0.697	0.532	0.903
S_{all}^{valid}	Freezing	0.405	0.988	0.697	0.532	0.874
	Fine-tuning	0.441	0.986	0.713	0.549	0.878
	Combination	0.436	0.985	0.711	0.546	0.880

SRE module can capture features of different types of IDRs

An intrinsically disordered protein may contain both LDRs and SDRs. Among these, LDRs tend to occur at the N-terminal and C-terminal, and are relatively easier to identify. SDRs are short and important regions that are discretely distributed, and computational methods may easily filter them out. Due to the different amino acid compositions and properties of SDRs and LDRs, the SRE module is employed to further capture the characteristics of different types of IDRs (SDRs and LDRs). The SRE module consists of two deep learning networks, including attention- and CNN-based networks, where the former captures the patterns of LDRs, and the latter extracts features from SDRs.

In the ablation study, the training sets were divided into two sub-training sets (S_{long}^{train} and S_{short}^{train}), and the validation set was partitioned, using the same method, into S_{long}^{valid} and S_{short}^{valid}. To validate the effectiveness of the SRE block in identifying different types of IDRs, the baseline model was composed solely of the pretrained model layers and a prediction layer, without the SRE block. As shown in Table 5, in the first set of experiments, both IDP-EDL-L and the baseline model were independently trained on S_{long}^{train} and tested on S_{long}^{valid}. Similarly, in the second set of experiments, IDP-EDL-S and the baseline model were trained on S_{short}^{train} and evaluated on the S_{short}^{valid}. Both IDP-EDL-L and IDP-EDL-S outperformed the baseline model, with Sn at 0.624 (16.4%), 0.520 (29.7%), BACC at 0.780 (4%), 0.746 (7.3%), and F1 at 0.694 (5.8%), 0.538 (2.9%), indicating that the incorporation of the SRE block enhanced the model's ablity to capture the characteristics of different types of IDRs, and to identify SDRs in proteins. In the third set of experiments, IDP-EDL-G, initializing the network architecture to be the same as IDP-EDL-L, was trained on S_{all}^{train} and tested on S_{all}^{valid}. We can see that IDP-EDL-G outperformed the baseline model.

Table 5. Performance of deep learning and baseline models.

Test set	Model	Sn	Sp	BACC	MCC	F1
S_{long}^{valid}	IDP-EDL-L	0.609	0.939	0.774	0.598	0.686
	Base model	0.532	0.965	0.748	0.589	0.654
S_{short}^{valid}	IDP-EDL-S	0.554	0.968	0.761	0.515	0.546
	Base model	0.401	0.991	0.696	0.529	0.524
S_{all}^{valid}	IDP-EDL-G	0.433	0.986	0.710	0.544	0.553
	Base model	0.407	0.988	0.697	0.535	0.536

Ensemble deep learning method can improve predictive performance

Traditionally, deep learning and ensemble learning, as machine learning techniques, have been regarded as independent research methods in bioinformatics applications. Ensemble deep learning, which combines them, has shown improved accuracy and stability. In our study, IDP-EDL is an ensemble of three deep learning models–IDP-EDL-L, IDP-EDL-S, and IDP-EDL-G–whose neural networks are trained separately on their respective types of training datasets. Table 6 shows the performance of these models on different types of test datasets, from which we can see the following:

- IDP-EDL-S and IDP-EDL-L performed best on the respective SDRs and LDRs validation datasets, which indicates that a specific deep learning model can improve the predictive performance for disordered regions with specific lengths;

Table 6. Performance of IDP-EDL-G, IDP-EDL-L, IDP-EDL-S, IDP-EDL on validation datasets.

Test set	Predictor	Sn	Sp	BACC	MCC	AUC
S_{all}^{valid}	IDP-EDL-G	0.463	0.982	0.722	0.546	0.880
	IDP-EDL-L	0.567	0.938	0.752	0.474	0.852
	IDP-EDL-S	0.488	0.973	0.730	0.526	0.863
	IDP-EDL	0.485	0.979	0.732	0.552	0.881
S_{long}^{valid}	IDP-EDL-G	0.500	0.971	0.735	0.576	0.853
	IDP-EDL-L	0.617	0.938	0.778	0.602	0.876
	IDP-EDL-S	0.430	0.979	0.704	0.537	0.831
	IDP-EDL	0.507	0.973	0.740	0.587	0.872
S_{short}^{valid}	IDP-EDL-G	0.436	0.983	0.710	0.501	0.880
	IDP-EDL-L	0.526	0.938	0.733	0.395	0.843
	IDP-EDL-S	0.533	0.972	0.752	0.517	0.897
	IDP-EDL	0.468	0.980	0.724	0.509	0.884

Table 7. Performance of various methods on MXD494.

Predictor	Sn	Sp	BACC	MCC	AUC	Rank		
						AUC	BACC	MCC
IDP-EDL	0.631	0.865	0.749	0.480	**0.842**	1	6	1
DeepIDP-2L [26]	0.737	0.776	0.757	0.452	**0.825**	2	2	5
IDP-Seq2Seq [24]	0.743	0.791	0.767	0.475	**0.825**	2	1	2
MFDp [21]	0.746	0.768	0.757	0.451	**0.821**	4	2	6
MD [36]	0.673	0.813	0.743	0.444	**0.821**	4	7	7
RFPR-IDP [20]	0.749	0.758	0.754	0.442	**0.821**	4	4	8
SPOT-Disorder [19]	0.626	0.851	0.739	0.457	**0.813**	7	9	4
SPINE-D [23]	0.787	0.698	0.742	0.411	**0.803**	8	8	10
AUCpreD [18]	0.521	0.881	0.701	0.411	**0.800**	9	15	10
DISOPRED3 [37]	0.622	0.820	0.721	0.410	**0.800**	9	12	12
IDP-FSP [25]	0.670	0.831	0.751	0.465	**0.794**	11	5	3
PONDER-FIT [38]	0.631	0.821	0.726	0.419	**0.790**	12	10	9
IUPred-long [39]	0.581	0.841	0.711	0.405	**0.784**	13	13	14
DISOPRED2 [40]	0.647	0.800	0.724	0.406	**0.781**	14	11	13
IUPred-short [39]	0.522	0.866	0.694	0.389	**0.781**	14	16	15
DISpro [41]	0.303	0.940	0.622	0.318	**0.775**	16	19	18
RONN [42]	0.664	0.754	0.709	0.368	**0.764**	17	14	16
Ucon [43, 44]	0.554	0.787	0.671	0.313	**0.741**	18	18	19
NORSnet [43, 44]	0.532	0.829	0.681	0.347	**0.738**	19	17	17
PROFbval [45]	0.835	0.387	0.611	0.196	**0.697**	20	20	20

- IDP-EDL performed best on general datasets containing both LDRs and SDRs, illustrating that the fusion of three deep learning models can achieve improved model accuracy.

The models have demonstrated a significant performance improvement on the corresponding disordered regions datasets. However, a noticeable decrease in performance is observed on different types of disordered region datasets. This is attributed to the extraction of the feature information from the protein sequence. IDP-EDL can achieve stable performance in predicting different IDRs, for two reasons:

- Different types of disordered regions have different characteristics. IDP-EDL-S can capture the characteristics of SDRs, while IDP-EDL-L can capture the characteristics of LDRs;
- IDP-EDL is an ensemble of three deep learning models, which are complementary.

Furthermore, the lengths of disordered regions in newly sequenced proteins are typically unknown. The stable performance of IDP-EDL in predicting disordered regions of varying lengths is especially useful for practical applications.

Comparison with other methods on independent test sets. To comprehensively evaluate the performance and generalization capability of IDP-EDL, we compared it with other existing methods on three distinct independent test sets, each containing varying proportions of LDRs and SDRs, as shown in Table 2. The prediction results of the commonly used MXD494 and SL329, and updated Disprot504 are shown in Tables 7, 8, and 9, respectively.

On the MXD494 and SL329 datasets, IDP-EDL achieved comparable or better performance with other methods, with the highest values of both AUC and MCC. The AUC and MCC of IDP-EDL are 0.842 and 0.480, respectively, on MXD494, which are 2% and 1% higher than those of the second-best method, IDP-Seq2Seq. Further evaluation of IDP-EDL against the top-performing methods was conducted on Disprot504, and the results of various methods are listed in Tables 9, from which we can see that IDP-EDL outperformed the other methods. Among all the methods compared, stable performance on these three independent test sets demonstrates that IDP-EDL is insensitive to different proportions of LDRs and SDRs.

Conclusion

In this paper, we proposed an ensemble deep learning model to predict the intrinsically disordered regions in proteins. IDP-EDL is an ensemble of three deep learning models–IDP-EDL-G, IDP-EDL-L, and IDP-EDL-S–each composed of a Generic Features Extractor (GFE) module, Specific Representations Extractor (SRE) module, and prediction layer. Inspired by the use of pretrained large language models in natural language processing to address downstream tasks, we employed a pretrained protein language model, ProteinBert, to extract the generic features of protein sequences to identify the intrinsically disordered regions. Based on the BERT/Transformer architecture and unsupervised learning methods, ProteinBert learns the intrinsic patterns of amino acid properties, structural information and evolutionary information implied in protein sequences from a large volume of data. This is why the pretrained model is capable of identifying disordered regions within proteins. After feature construction, to further capture the specific characteristics of different types of disordered regions, we introduced a Specific Representations Extractor (SRE) module, which employs an attention-based neural network and CNN to capture the respective features of LDRs and SDRs. Ensemble deep learning, combining the machine learning techniques of ensemble learning and deep learning, improves the performance of deep learning models. Based on ensemble deep learning, IDP-EDL achieves accuracy, stability, and robustness compared to individual deep learning models.

Table 8. Performance of various methods on SL329.

Predictor	Sn	Sp	BACC	MCC	AUC	Rank		
						AUC	BACC	MCC
IDP-EDL	0.68	0.97	0.820	0.68	**0.904**	1	4	1
DeepIDP-2L [26]	0.74	0.92	0.830	0.68	**0.904**	1	1	1
SPOT-Disorder [19]	0.65	0.96	0.805	0.65	**0.901**	3	7	4
IDP-Seq2Seq [24]	0.71	0.92	0.822	0.67	**0.899**	4	2	3
AUCpreD [18]	0.63	0.96	0.795	0.64	**0.887**	5	9	5
SPINE-D [23]	0.82	0.80	0.815	0.61	**0.886**	6	5	9
DISOPRED3 [37]	0.67	0.92	0.796	0.62	**0.880**	7	8	7
RFPR-IDP [20]	0.78	0.84	0.809	0.62	**0.879**	8	6	7
MFDp [21]	0.88	0.62	0.750	0.51	**0.873**	9	15	14
MD [36]	0.66	0.89	0.775	0.58	**0.864**	10	11	11
IDP-FSP [25]	0.75	0.89	0.821	0.65	**0.864**	10	3	6
DISOPRED2 [40]	0.69	0.90	0.795	0.59	**0.858**	12	9	10
DISOClust [46]	0.81	0.70	0.755	0.51	**0.846**	13	14	14
PONDR-FIT [38]	0.61	0.91	0.760	0.55	**0.843**	14	12	12
IUpred-long [39]	0.60	0.92	0.760	0.55	**0.839**	15	12	12
IUpred-short [39]	0.50	0.94	0.720	0.50	**0.829**	16	17	17
NORSnet [43, 44]	0.54	0.92	0.730	0.51	**0.815**	17	16	14
Ucon [43, 44]	0.59	0.81	0.700	0.42	**0.779**	18	18	18
PONDERVL-XT [25]	0.59	0.78	0.685	0.38	**0.755**	19	19	19

Table 9. Performance of various methods on Disprot504.

Predictor	Sn	Sp	BACC	MCC	AUC	Rank		
						AUC	BACC	MCC
IDP-EDL	0.632	0.775	0.703	0.374	**0.762**	1	1	1
DeepIDP-2L [26]	0.688	0.718	0.703	0.361	**0.758**	2	1	2
IDP-Seq2Seq [24]	0.646	0.727	0.686	0.334	**0.741**	3	3	3
SPINE-D [23]	0.752	0.613	0.683	0.315	**0.738**	4	4	6
SPOT-Disorder [19]	0.594	0.761	0.677	0.326	**0.732**	5	6	4
AUCpreD [18]	0.497	0.822	0.660	0.315	**0.729**	6	8	6
IUCpred-Long [39]	0.575	0.772	0.674	0.323	**0.725**	7	7	5
RFPR-IDP [20]	0.723	0.634	0.681	0.314	**0.720**	8	5	8
IUCpred-Short [39]	0.482	0.817	0.649	0.295	**0.718**	9	9	9
DISOPRED3 [37]	0.510	0.773	0.641	0.267	**0.697**	10	10	10

Key Points

- The feature construction method of IDP-EDL involves generating features from raw protein sequences using protein pretrained language model, thus no need for database searches.
- IDP-EDL employs specific representation extractor to capture the characteristics of different types of disordered regions.
- IDP-EDL is an ensemble of three deep learning medels-IDP-EDL-L, IDP-EDL-S, and IDP-EDL-G. Each model is trained separately on their respective types of training datasets.
- Experiments demonstrate that IDP-EDL is an accurate and robust predictor, which has better performance than state-of-the art predictors.

Funding

This work was supported by the National Natural Science Foundation of China (Grant No.62302316, 62302317, 62302359 and 62472152), Shenzhen Science and Technology Program (Grant No.RCBS20221008093227027), Shenzhen Colleges and Universities Stable Support Program (Grant No.2022071518360 2001), Natural Science Foundation of Top Talent of SZTU (Grant No.GDRC202319), Natural Science Basic Research Program of Shaanxi(No.2023-JC-QN-0636).

References

1. H Jane Dyson and Peter E Wright. Intrinsically unstructured proteins and their functions. *Nature reviews Molecular cell biology*, 6(3):197–208, 2005.
2. Lilia M Iakoucheva, Celeste J Brown, J David Lawson, Zoran Obradović, and A Keith Dunker. Intrinsic disorder in cell-signaling and cancer-associated proteins. *Journal of molecular biology*, 323(3):573–584, 2002.
3. Damiano Piovesan, Francesco Tabaro, Ivan Mičetić, Marco Necci, Federica Quaglia, Christopher J Oldfield, Maria Cristina Aspromonte, Norman E Davey, Radoslav Davidović, Zsuzsanna Dosztányi, et al. Disprot 7.0: a major update of the database of disordered proteins. *Nucleic acids research*, 45(D1):D219–D227, 2017.
4. Vladimir N Uversky, Christopher J Oldfield, and A Keith Dunker. Intrinsically disordered proteins in human diseases: introducing the d2 concept. *Annu. Rev. Biophys.*,

37(1):215–246, 2008.

5. Véronique Receveur-Bréchot, Jean-Marie Bourhis, Vladimir N Uversky, Bruno Canard, and Sonia Longhi. Assessing protein disorder and induced folding. *Proteins: Structure, Function, and Bioinformatics*, 62(1):24–45, 2006.
6. Robert Konrat. Nmr contributions to structural dynamics studies of intrinsically disordered proteins. *Journal of Magnetic Resonance*, 241:74–85, 2014.
7. Zexi Yang, Yan Wang, Xinye Ni, and Sen Yang. Deepdrp: Prediction of intrinsically disordered regions based on integrated view deep learning architecture from transformer-enhanced and protein information. *International Journal of Biological Macromolecules*, 253:127390, 2023.
8. András Hatos, Borbála Hajdu-Soltész, Alexander M Monzon, Nicolas Palopoli, Lucía Álvarez, Burcu Aykac-Fas, Claudio Bassot, Guillermo I Benítez, Martina Bevilacqua, Anastasia Chasapi, et al. Disprot: intrinsic protein disorder annotation in 2020. *Nucleic acids research*, 48(D1):D269–D276, 2020.
9. Emilio Potenza, Tomás Di Domenico, Ian Walsh, and Silvio CE Tosatto. Mobidb 2.0: an improved database of intrinsically disordered and mobile proteins. *Nucleic acids research*, 43(D1):D315–D320, 2015.
10. Serbulent Unsal, Heval Atas, Muammer Albayrak, Kemal Turhan, Aybar C Acar, and Tunca Doğan. Learning functional properties of proteins with language models. *Nature Machine Intelligence*, 4(3):227–245, 2022.
11. Yumeng Liu, Xiaolong Wang, and Bin Liu. A comprehensive review and comparison of existing computational methods for intrinsically disordered protein and region prediction. *Briefings in bioinformatics*, 20(1):330–346, 2019.
12. Stephen F Altschul, Thomas L Madden, Alejandro A Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–3402, 1997.
13. Nadav Brandes, Dan Ofer, Yam Peleg, Nadav Rappoport, and Michal Linial. Proteinbert: a universal deep-learning model of protein sequence and function. *Bioinformatics*, 38(8):2102–2110, 2022.
14. Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021.
15. Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Peter Chen, John Canny, Pieter Abbeel, and Yun Song. Evaluating protein transfer learning with tape. *Advances in neural information processing systems*, 32, 2019.
16. Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, et al. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(10):7112–7127, 2021.
17. Yihe Pang and Bin Liu. Idp-lm: Prediction of protein intrinsic disorder and disorder functions based on language models. *PLOS Computational Biology*, 19(11):e1011657, 2023.
18. Sheng Wang, Jianzhu Ma, and Jinbo Xu. Aucpred: proteome-level protein disorder prediction by auc-maximized deep convolutional neural fields. *Bioinformatics*, 32(17):i672–i679, 2016.
19. Jack Hanson, Yuedong Yang, Kuldip Paliwal, and Yaoqi Zhou. Improving protein disorder prediction by deep bidirectional long short-term memory recurrent neural networks. *Bioinformatics*, 33(5):685–692, 2017.
20. Yumeng Liu, Xiaolong Wang, and Bin Liu. Rfpr-idp: reduce the false positive rates for intrinsically disordered protein and region prediction by incorporating both fully ordered proteins and disordered proteins. *Briefings in bioinformatics*, 22(2):2000–2011, 2021.
21. Marcin J Mizianty, Wojciech Stach, Ke Chen, Kanaka Durga Kedarisetti, Fatemeh Miri Disfani, and Lukasz Kurgan. Improved sequence-based prediction of disordered regions with multilayer fusion of multiple information sources. *Bioinformatics*, 26(18):i489–i496, 2010.
22. Jack Hanson, Kuldip K Paliwal, Thomas Litfin, and Yaoqi Zhou. Spot-disorder2: improved protein intrinsic disorder prediction by ensembled deep learning. *Genomics, Proteomics and Bioinformatics*, 17(6):645–656, 2019.
23. Tuo Zhang, Eshel Faraggi, Bin Xue, A Keith Dunker, Vladimir N Uversky, and Yaoqi Zhou. Spine-d: accurate prediction of short and long disordered regions by a single neural-network based method. *Journal of Biomolecular Structure and Dynamics*, 29(4):799–813, 2012.
24. Yi-Jun Tang, Yi-He Pang, and Bin Liu. Idp-seq2seq: identification of intrinsically disordered regions based on sequence to sequence learning. *Bioinformatics*, 36(21):5177–5186, 2020.
25. Yumeng Liu, Shengyu Chen, Xiaolong Wang, and Bin Liu. Identification of intrinsically disordered proteins and regions by length-dependent predictors based on conditional random fields. *Molecular Therapy-Nucleic Acids*, 17:396–404, 2019.
26. Yi-Jun Tang, Yi-He Pang, and Bin Liu. Deepidp-2l: protein intrinsically disordered region prediction by combining convolutional attention network and hierarchical attention network. *Bioinformatics*, 38(5):1252–1260, 2022.
27. Zhen-Ling Peng and Lukasz Kurgan. Comprehensive comparative assessment of in-silico predictors of disordered regions. *Current Protein and Peptide Science*, 13(1):6–18, 2012.
28. Fernanda L Sirota, Hong-Sain Ooi, Tobias Gattermayer, Georg Schneider, Frank Eisenhaber, and Sebastian Maurer-Stroh. Parameterization of disorder predictors for large-scale applications requiring high specificity by using an extended benchmark dataset. *BMC genomics*, 11:1–17, 2010.
29. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
30. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
31. Yue Cao, Thomas Andrew Geddes, Jean Yee Hwa Yang, and Pengyi Yang. Ensemble deep learning in bioinformatics. *Nature Machine Intelligence*, 2(9):500–508, 2020.

32. Ammar Mohammed and Rania Kora. A comprehensive review on ensemble deep learning: Opportunities and challenges. *Journal of King Saud University-Computer and Information Sciences*, 35(2):757–774, 2023.
33. Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.
34. Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.
35. Xiao-Juan Zhu, Chao-Qin Feng, Hong-Yan Lai, Wei Chen, and Lin Hao. Predicting protein structural classes for low-similarity sequences by evaluating different features. *Knowledge-Based Systems*, 163:787–793, 2019.
36. Avner Schlessinger, Marco Punta, Guy Yachdav, Laszlo Kajan, and Burkhard Rost. Improved disorder prediction by combination of orthogonal approaches. *PloS one*, 4(2):e4433, 2009.
37. David T Jones and Domenico Cozzetto. Disopred3: precise disordered region predictions with annotated protein-binding activity. *Bioinformatics*, 31(6):857–863, 2015.
38. Bin Xue, Roland L Dunbrack, Robert W Williams, A Keith Dunker, and Vladimir N Uversky. Pondr-fit: a meta-predictor of intrinsically disordered amino acids. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, 1804(4):996–1010, 2010.
39. Zsuzsanna Dosztányi, Veronika Csizmok, Peter Tompa, and István Simon. Iupred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics*, 21(16):3433–3434, 2005.
40. Jonathan J Ward, Jaspreet S Sodhi, Liam J McGuffin, Bernard F Buxton, and David T Jones. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *Journal of molecular biology*, 337(3):635–645, 2004.
41. Jianlin Cheng, Michael J Sweredoski, and Pierre Baldi. Accurate prediction of protein disordered regions by mining protein structure data. *Data mining and knowledge discovery*, 11:213–222, 2005.
42. Zheng Rong Yang, Rebecca Thomson, Philip McNeil, and Robert M Esnouf. Ronn: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. *Bioinformatics*, 21(16):3369–3376, 2005.
43. Avner Schlessinger, Jinfeng Liu, and Burkhard Rost. Natively unstructured loops differ from other loops. *PLoS computational biology*, 3(7):e140, 2007.
44. Avner Schlessinger, Marco Punta, and Burkhard Rost. Natively unstructured regions in proteins identified from contact predictions. *Bioinformatics*, 23(18):2376–2384, 2007.
45. Avner Schlessinger, Guy Yachdav, and Burkhard Rost. Profbval: predict flexible and rigid residues in proteins. *Bioinformatics*, 22(7):891–893, 2006.
46. Liam J McGuffin. Intrinsic disorder prediction from the analysis of multiple protein fold recognition models. *Bioinformatics*, 24(16):1798–1804, 2008.