HOKKAIDO UNIVERSITY

Title	Accurate and Fast Prediction of Intrinsically Disordered Protein by Multiple Protein Language Models and Ensemble Learning
Author(s)	Xu, Shijie; Onoda, Akira
Citation	Journal of Chemical Information and Modeling, 64, 2901-2911 https://doi.org/10.1021/acs.jcim.3c01202
Issue Date	2023-10-26
Doc URL	https://hdl.handle.net/2115/100491
Rights	This document is the Accepted Manuscript version of a Published Article that appeared in final form in Journal of chemical information and modeling, copyright © 2023 American Chemical Society. To access the final published article, see ACS Articles on Request.
Type	journal article
File Information	MS_rev_unmarked_ci-2023-01202r_ShijieXu.pdf

Hokkaido University Collection of Scholarly and Academic Papers : HUSCAP

Accurate and Fast Prediction of Intrinsically Disordered Protein by Multiple Protein Language Models and Ensemble Learning

Shijie Xu[1] *and Akira Onoda*[1,2,*]

[1]Graduate School of Environmental Science, Hokkaido University, Sapporo, 060-0810, Japan

[2]Faculty of Environmental Earth Science, Hokkaido University, Sapporo, 060-0810, Japan

ABSTRACT

Intrinsically disordered proteins (IDPs) play a vital role in various biological processes and have attracted increasing attention over the past few decades. Predicting IDPs from primary structures of proteins offers a rapid and facile means of protein analysis without necessitating crystal structures. In particular, machine learning methods have demonstrated their potential in this field. Recently, protein language models (PLMs) are emerging as a promising approach to extracting essential information from protein sequences and have been employed in protein modeling to utilize its advantages of precision and efficiency. In this paper, we developed a novel IDP prediction method named IDP-ELM to predict the intrinsically disordered regions (IDRs) as well

as their functions including disorder flexible linkers and disordered protein-binding. This method utilizes the high-dimensional representations extracted from several state-of-the-art PLMs and predicts IDRs by ensemble learning based on bidirectional recurrent neural networks. The performance of the method was evaluated on two independent test datasets from CAID (Critical assessment of protein intrinsic disorder prediction) and CAID2, indicating notable improvements in terms of area under the receiver operating characteristic (AUC), Matthew's correlation coefficient (MCC), and F1 score. Moreover, IDP-ELM requires solely protein sequences as inputs and does not entail a time-consuming process of protein profile generation, which is a prerequisite for most existing state-of-the-art methods, enabling an accurate, fast, and convenient tool for proteome-level analysis. The corresponding reproducible source code and model weights are available at https://github.com/xu-shi-jie/idp-elm.

Introduction

Intrinsically disordered proteins (IDPs) and intrinsically disordered regions (IDRs) are proteins and regions that lack stable structures but play a crucial role in biological processes[1]. Recent research has demonstrated their widespread occurrence in domains of life[2]. For instance, over 6% of amino acid residues in annotated sequences exist as disordered in SwissProt[3]. This percentage increases to over 30% in eukaryotes[4]. Furthermore, some critical human diseases, such as Alzheimer's disease[5], Parkinson's disease[6], and cancers[7], have been linked to IDRs, which have attracted the interest of researchers in the application of drug design[8]. Despite the importance and potential, only thousands of IDPs have been experimentally identified and

recognized in DisProt database (V9.3, 2022 Dec) since it is significantly expensive and laborious to identify and annotate them[9].

Various computational methods have been thus proposed in the past few decades to predict IDRs directly from sequences, providing rapid and facile tools for protein analysis. With the assistance of deep learning, neural network-based methods[10-13] have been demonstrated to be useful in the field. These methods are built on different architectures such as recurrent neural networks (RNN)[14], convolutional neural networks[15], and the self-attention mechanism[16]. To achieve better performance, most existing methods depend directly or indirectly on protein profiles generated by multiple sequence alignments (MSAs), which are constructed by searching homologous sequences against the entire protein sequence database. However, the approach using MSA has two problems. First, MSAs are not always available for all proteins, such as those in the dark proteome[17, 18]. Second, generating MSAs involves a time-consuming search against the entire databases, such as UniProtKB, which contains over 227 million protein sequences[19]. Generating MSAs in a reasonable time becomes increasingly difficult since the size of the protein sequence database is rapidly increasing by modern sequencing techniques. The limitations hinder the applications of IDP predictors. In fact, some MSA-free predictors are therefore proposed[20, 21], enabling the prediction of IDRs in large-scale analysis such as human proteome IDR prediction with the use of protein sequences as inputs.

Recently, language models (LMs)[22] have emerged as a valuable method for sequence modeling in machine learning. These large-scale neural networks, often with hundreds of millions or billions of parameters, are trained on large-scale unlabeled language corpora, such as Wikipedia, to learn the latent patterns of sequences automatically through pre-training. The outputs of pre-trained LMs are usually represented as high-dimensional vectors that contain

plentiful contextual information of the sequence, which is referred to as "high-dimensional representations" and is represented in closer vectors for similar words and the farther vectors in dissimilar words. Methods based on pre-trained models have been shown to outperform conventional methods that do not use LMs in a wide range of machine-learning tasks, including machine translation and text generation[22, 23]. Protein language models (PLMs)[24-26] are LMs trained on protein sequences, which are capable of extracting essential information efficiently yet effectively. PLMs have shown competitive results compared to conventional methods in several protein-related tasks, including protein family classification[27], secondary structure prediction[26], residue-residue contact prediction[24], and three-dimensional structure prediction[25]. While initial work attempted to apply PLMs to IDP prediction tasks, it did not achieve state-of-the-art performance due to the limited model size (usually the number of parameters in the model) and disadvantages of the underlying architecture[28].

We hereby introduce a novel approach, IDP-ELM, aiming at the prediction of intrinsically disordered proteins and their functions by utilizing high-dimensional representations extracted from PLMs. The method employed state-of-the-art PLMs derived from previous studies[24-26]. We first implemented a module for secondary structure prediction based on a single PLM and bidirectional long short-term memory (BiLSTM) network, since predicted secondary structures have been found to be effective in improving the IDP predictions[13]. Subsequently, IDP is predicted by a bidirectional gated recurrent unit (BiGRU) network, using both the high-dimensional representations, with the predicted secondary structures as inputs. Furthermore, we employed ensemble learning to obtain better performance by averaging the output from each predictor based on different PLMs, since ensemble learning has been proven effective for IDP prediction[29]. Our results demonstrate that IDP-ELM outperforms previous predictors in terms of

AUC, MCC, and F1 metrics on both two challenging independent test datasets CAID[30] and CAID2[31].

In addition to accuracy, we show other two advantages of PLM-based methods. Our novel approach, based on PLM and predicted IDRs, predicts the functions of IDRs, including both disordered protein-binding (DP), disordered region that binds with other proteins, and disordered flexible linkers (DFL), disordered region that connects two domains, with state-of-the-art performance compared to the existing methods. PLM-based methods also significantly enhance the computational efficiency of IDP prediction. PLM-based methods also significantly enhance the computational efficiency of IDP prediction. We thus performed an accurate and fast prediction of human proteomes, including 34,877 sequences, to demonstrate the practical usefulness of our IDP-ELM in proteome-level analyses.

Methods

Datasets

For secondary structure prediction, we used the NetSurfP-3.0 dataset from the reported work[32]. The training dataset contains 10,337 non-redundant protein sequences, while the validation dataset has 500 such sequences. The performance of the secondary structure predictors was evaluated using three test datasets: TS115, comprising of 115 proteins; CB513, with 513 protein chains from 434 proteins; and CASP12, comprising of 21 proteins.

The same datasets in the reported work in fIDPnn[13] were used for training, validating, and testing our IDP predictors. fIDPnn was constructed from DisProt (V7.0) database and separated

into three datasets: training (445 proteins), validation (100 proteins), and testing (200 proteins). The combined training and test datasets were then clustered by cd-hit algorithm[33] with 25% sequence similarity. Sequences in the test dataset that shared <25% similarity to the training proteins were removed, resulting in 176 test proteins. We also employed the independent test dataset from CAID (Critical assessment of protein intrinsic disorder prediction)[30]. Furthermore, the sequences in CAID were clustered by cd-hit with 25% sequence similarity against fIDPnn training and validation datasets to ensure fairness in comparison with other predictors. The similarity within itself is also removed, resulting 81 non-redundant sequences, with 9,008 (15.35%) disordered residues and 49,663 (84.65%) ordered residues. It is referred as "non-redundant CAID" while original one is referred to as "redundant CAID" later.

Furthermore, we employed another test dataset from CAID2[31]. Similarly, we removed the redundant sequences against the training and validation datasets of fIDPnn, which is referred as "non-redundant CAID2" while the original one is referred to as "redundant CAID2". It is also noteworthy that the non-redundant CAID2 benchmark test dataset was considerably different from the non-redundant CAID benchmark test dataset, featuring a vast number of excluded residues (27,650, 60.03%) and disordered residues (4,825, 10.48%), as well as ordered residues (13,579, 29.48%) in 39 non-redundant sequences. In particular, these characteristics are quite different from fIDPnn training and validation datasets, thus being a greater challenge to predict.

For disordered flexible linker (DFL) prediction, we utilized benchmark datasets from the previously reported work[34] to train and evaluate our models, which include TR166 for training and validation, and TE82 for independent testing. To prevent the risk of overfitting, any two sequences in the training and validation dataset shared less than 25% sequence similarities. TR166 was divided into training and validation datasets which contain 133 and 33 sequences,

respectively. The independent test dataset TE82 comprised 82 sequences collected from DisProt (V8.0) database, as reported by the previous work[34]. The similarity between any two sequences in TR166 and TE82 was less than 25%. Detailed statistics are shown in Table 1.

Table 1. Statistics of Datasets for Disordered Flexible Linker Prediction

Dataset		# protein	# DFL	# non-DFL	# residue
TR144	Training	133	2,762 (6.1%)	13,900 (30.5%)	45,528
	Validation	33	908 (6.5%)	3,652 (26.0%)	14,031
TE82		82	2,223 (6.0%)	11,471 (31.1%)	36,938

Notes: mark "#" refers to the number.

For protein-binding prediction, the benchmark datasets from the reported work[35] was utilized to train and evaluate the models. These datasets were constructed from the DisProt database (V8.1), which includes IDRs that interact with other proteins. To avoid any overfitting, sequences within each dataset share less than 30% similarity with sequences in other datasets. The detailed statistics can be found in the literature[35].

Architecture of IDP-ELM

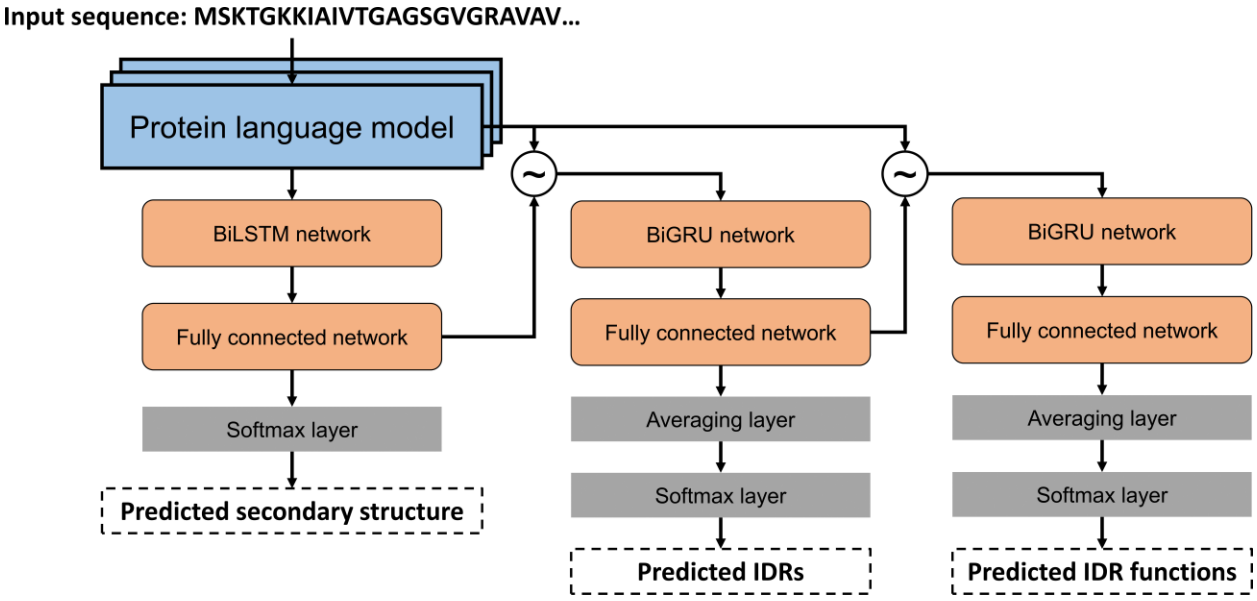


Figure 1. The architecture of IDP-ELM predictor.

Figure 1 illustrates the architecture of the IDP-ELM predictor. We employ nine distinct protein language models (PLMs) from the reported works[24-26] to extract high-dimensional representations for protein sequences. For each PLM, the representations are first fed into a bidirectional long short-term memory (BiLSTM) network, which is followed by a fully connected network and a softmax layer that predicts secondary structure. The intermediate outputs from a fully connected network, namely logits, are then concatenated with high-dimensional representations and fed into a bidirectional gated unit (BiGRU) network followed by a fully connected network to produce logits for IDRs. Similarly, the logits for IDR are also concatenated with high-dimensional representations and fed into another BiGRU network followed by a fully connected network to produce the logits of the function of IDRs. For secondary structure prediction, we employed the logits from the best-performed PLM (see Section **Results** for details) and converted them into predicted secondary structures by a softmax layer.

Ensemble learning. To predict IDRs as well as their functions, we utilize averaging layers and softmax layers. These layers have no parameters to train, and an averaging layer outputs the mean of the logits $x^i = (x_1^i, x_2^i)$ for $i = 1,2, \dots, 9$ from each individual predictor, represented by $x = \frac{1}{9}\sum_{i=1}^{9} x^i$. A softmax layer is then applied to convert the logits $x = (x_1, x_2)$ to a 2-dimensional vector (p_1, p_2), where p_1 is the probability of non-IDRs and p_2 is the probability of IDRs. This is achieved by the following equation:

$$(p_1, p_2) = \text{softmax}(x_1, x_2) = \frac{(e^{x_1}, e^{x_2})}{e^{x_1} + e^{x_2}}$$

Protein language models. We utilized various PLMs to extract high-dimensional representations for protein sequences, with different architectures, model sizes and pre-training datasets. ESM (evolutionary scale modeling) models from reported works[24, 25] are based on masked language models (MLM)[22] and its architecture is depicted in Figure 2. For each protein sequence x in the database X, a certain portion of 15% of its amino acid residues is masked, replaced, or kept while the remaining 85% are unmasked. The MLM predicts the probability of x_i given masked sequences $x_{/M}$, training to recover the original sequence before masking as accurately as possible. This is done to minimize the following loss function[24]:

$$L_{MLM} = E_{x \sim X} E_M \sum_{i \in M} -\log p(x_i | x_{/M})$$

where p is the probability predicted by MLM and E represents the expectation.

Through training on large-scale protein sequence datasets, the ESM models learns the latent relations between residues, which are called *attention maps* in MLM. The attention maps have been demonstrated to be related to the contact map of protein structures[36]. Additionally, ESM

models produce high-dimensional representations for each residue in the sequences, which can be utilized for downstream tasks.

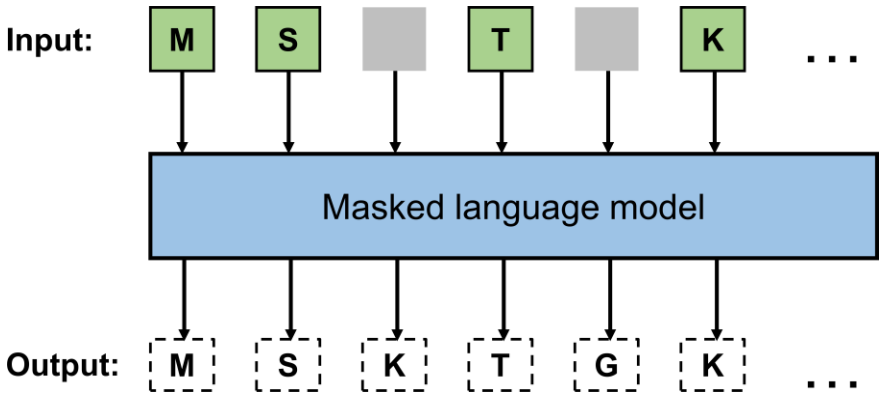


Figure 2. The architecture of masked language model.

The T5 (text-to-text transfer transformer) models were initially introduced in the reported work[37] to confirm the boundaries of transfer learning and attaining state-of-the-art proficiency in a variety of natural language processing tasks. These models have an encoder-decoder architecture, wherein the encoder is capable of learning from both directions of the sequence, while the decoder is limited to predicting tokens only from left to right (Figure 3). Typically, only the encoder is employed in practice, and T5 models for protein modeling[26], has been successfully utilized into various downstream tasks concerning protein modeling, exhibiting outstanding performance.

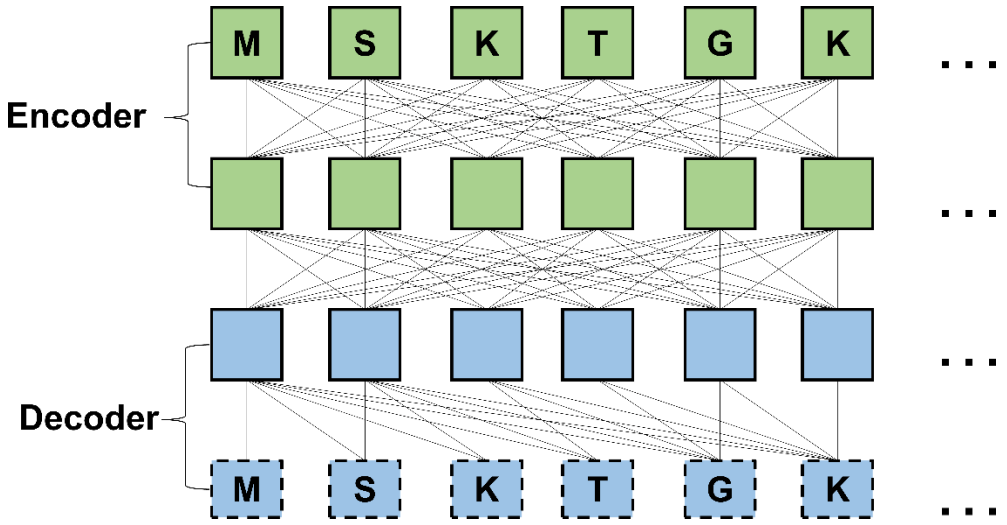


Figure 3. The architecture of T5 model.

The details of PLMs we employed are listed in Table 2 including hidden dimensions (Dim), number of layers (# layer), number of parameters (Size), and pre-training datasets (Dataset). For ESM models, UR50/S refers to the sparse Uniref50 datasets, and UR50/D refers to the dense Uniref50 datasets, which are constructed based on the UniParc dataset containing 250 million diverse protein sequences[38]. For T5 models, BFD refers to the Big Fantastic Database[39, 40] which contains 2.1 billion protein sequences and UR50 refers to Uniref50 dataset which contains 45 million protein sequences[41]. For more information of these pre-training datasets, see the original papers[24-26].

Table 2. Details of Different Protein Language Models

Models	Dim	# layer	Size	Dataset
esm1b_t33_650M_UR50S	1280	33	650M	UR50/S
esm2_t6_8M_UR50D	320	6	8M	UR50/D
esm2_t12_35M_UR50D	480	12	35M	UR50/D
esm2_t30_150M_UR50D	640	30	150M	UR50/D

esm2_t33_650M_UR50D	1280	33	650M	UR50/D
esm2_t36_3B_UR50D	2560	36	3B	UR50/D
esm2_t48_15B_UR50D	5120	48	15B	UR50/D
prot_t5_xl_bfd	1024	24	~3B	BFD
prot_t5_xl_uniref50	1024	24	~3B	UR50

Evaluation metrics

To comprehensively assess the efficacy of our approach, we employed the following metrics:

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

$$\mathrm{SP} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$$

$$\mathrm{SN}\ (\mathrm{i.e.}, \mathrm{Rec}) = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

$$\mathrm{Pre} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

$$\mathrm{F1} = \frac{2 \times \mathrm{TP}}{2 \times \mathrm{TP} + \mathrm{FP} + \mathrm{FN}}$$

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}}$$

$$\mathrm{AUC} = \text{Area Under the Receiver Operating Characteristic Curve}$$

$\mathrm{TP}, \mathrm{FP}, \mathrm{FN}, \mathrm{TN}$ represent the true positive, false positive, false negative and true negative, respectively. Here, the term “true” signifies the prediction that align with the actual outcomes, while “positive” denotes disordered states, and vice versa. The Matthew’s correlation coefficient (MCC) ranges between –1 and 1. A value of –1 indicates complete disagreement between the predictor and the actual results, while a value of 1 signifies a perfect predictor. A value of 0 corresponds to the performance of a random prediction. The F1 score lies between 0 and 1. A values of 0 signifies complete disagreement between the predictor and the actual results, while a value of 1 represents a perfect prediction. A value of 0.5 corresponds to the performance of a

random prediction. Other measure also falls within the range of 0 to 1, with larger values indicating superior performance.

It is important to note that AUC measure differs from the others as it remains unbiased in the face of imbalanced classification and does not depend on the choice of threshold. For the remaining metrics, including precision (Pre), recall (Rec), accuracy (ACC), and F1 score, we have selected a threshold of 0.5 and do not explicitly mention it unless necessary.

Results

Hyper-parameter tuning, early stopping, and optimizers. We have employed a grid search approach to select the best hyper-parameters for IDP and function predictions, including batch size, hidden size, and the number of layers, while keeping the remaining hyper-parameters fixed. For IDP prediction, tuning ranges in [16,32,64] for batch size, [64,128,256] for hidden size, and [2,3] for the number of layers; for DFL prediction, tuning ranges in [8,16,32] for batch size, [16,32,64] for hidden size, and the number of layers is fixed with 1; for DP prediction, tuning ranges in [8,16,32] for batch size, [16,32,64,128] for hidden size, and the number of layers is fixed with 1. By performing the grid search, we chose the checkpoints that exhibit the maximum summation of validated AUC, F1, and MCC scores, resulting the hyper-parameters in Supplementary Table S1 for IDP prediction, Table S2 for DFL prediction, Table S3 for DP prediction. The hidden size and batch size for secondary structure prediction is fixed as 1024 and 16. Note that hidden sizes refer to the hidden size of BiLSTM or BiGRU networks.

To avoid the risk of overfitting, we have adopted an early stopping strategy. The validated AUC scores serve as the criterion for halting the training process when they stop to increasing over five epochs. The AdamW optimizer[42] has been employed to train the secondary structure

predictor, and Adam optimizer[43] has been employed for other training tasks. A learning rate of 10^{-3} and weight decays of 10^{-2} has been set for all computational experiments. Regarding the other optimizer options, we have utilized the default parameters as provided by the PyTorch library. For other detailed hyper-parameters, please refer to the Supplementary Table S4.

Sequence paddings and segmentation. While the PLMs can effectively extract high-dimensional representations for a batch of sequences, we instead encoded protein sequences one by one. Because user inputs for prediction are more likely to be sequential rather than concurrent. However, once the representations are extracted, they must be padded with zeroes within a mini batch before being fed into the BiGRU and subsequent networks. The previous report suggested the influence of sequence padding in protein prediction tasks[44]. We also observed a similar phenomenon in our situation, wherein accuracies slightly decrease from batchified prediction to single-sequence prediction. Some of the PLMs employed in our study impose limitations on the maximum sequences length. For instance, the esm1b model can only accept sequence with fewer than 1024 amino acids residues, while accounting for the inclusion of two special token that are the beginning and end of the sequence. Therefore, we divide protein sequences longer than 960 amino acid residues into multiple segments, with an additional 31 amino acids residues neighboring their left and right sides. Consequently, neighboring segments overlap by 31 amino acids, while the first and last segments exhibit a single overlapping region on one side.

Down-sampling strategy. Due to the highly imbalanced nature of the training datasets, where the number of positive labels is significantly smaller than the number of negative labels, we have implemented a down-sampling strategy. This strategy is designed to foster the model's focus on the positive labels and prevent it from excessively learning from the negative labels. Specifically, each min-batch training was performed by randomly masking a certain portion of residues that

are labeled as negative in the training datasets. For example, only 30% of the ordered residues are utilized for training the IDP predictor. It is worth noting that the down-sampling strategy employed in IDP-ELM has greatly enhanced its performance in terms of F1 score while keeping comparable AUC and MCC scores, as well as stabilizing the performance of a single predictor, and our ablation experiments support the conclusions (Supplementary Table S5).

PLM-based secondary structure predictions achieve state-of-the art performance

We implemented and evaluated the secondary structure predictors (SSPs) based on different PLMs, as depicted in Figure 4. Each SSP corresponds to a specific underlying PLM, and the evaluation encompasses tasks of predicting Q8 and Q3 states. Q8 entails predicting eight distinct states of secondary structure, 3_{10}-helix (G), α-helix (H), π-helix (I), β-strand I, bright (B), turn (T), bend (S) and others (C)[45]. On the other hand, Q3 involves predicting three general states: helix (including G, H, and I), strand (including B and E) and coil (including T, S, and C). In our experiments, we solely focus on predicting Q8 states, while the Q3 states are derived from Q8 through the aforementioned replacements.

We train multiple SSPs based on diverse PLMs and evaluate their performance on distinct test datasets. The results are displayed in Figure 4. Notably, the SSP utilizing a 3-billion-parameter PLM (esm2_t36_3B_UR50D) achieved the highest scores across all metrics, including Q8 and Q3, on CASP12, CB513, and TS115 test datasets. Consequently, we consider this model as our reported predictor. The Q3 scores on the CB513 and TS115 datasets reached 87.78% and 87.99%, respectively, which approach the theoretical limits of secondary structure prediction (88%–90%)[46]. Additionally, the Q3 score on the CASP12 test dataset reached 83.66%, albeit this dataset is small, containing only 21 sequences, and thus exhibiting a higher variance.

Furthermore, the performance of PLMs depends on factors such as model size, pre-trained dataset, and architecture. For instance, ESM2 PLMs (with prefix "esm2") pre-trained on the same UR50/D dataset and MLM architecture may yield different performance due to the variations in their model sizes. It is worth noting that larger PLMs do not necessarily guarantee improved accuracy. In certain cases, a SSP based on 15-billion-parameter ESM model (esm2_t48_15B_UR50D) may exhibit slight lower accuracy than a SSP based on 3-billion-parameter ESM model (esm2_t36_3B_UR50D) due to the limited size of the labeled training dataset, which is not sufficient to fully leverage the capabilities of the larger PLM. Moreover, two SSPs based on PLMs with similar model architectures and parameter sizes (esm1b_t33_650M_UR50S and esm2_t33_650M_50D) display different accuracy levels, attributed to the larger pre-training dataset and longer pre-training duration of the latter. Additionally, two SSPs based on T5 models (prot_t5_xl_bfd and prot_t5_xl_half_uniref50-enc) underperformed compared to the SSP based on the 3-billion-parameter ESM model (esm2_t36_3B_UR50D), despite having a similar model size of approximately 3 billion. This discrepancy arises because the T5 models, based on an encoder-decoder architecture, only utilize encoder components. In contrast, the 3-billion-parameter ESM model can harness the full parameters of the trained models.

Finally, we conducted a comparison with existing secondary structure prediction methods. As shown in Table 3, our predictor based on the 3-billion-parameter ESM PLM achieves state-of-the-art results, surpassing both the traditional MSA-based method NetSurP-2.0 and another predictor NetSurfP-3.0, which is also based on the PLM (esm1b_t33_650M_UR50S). It is also worth noting that by utilizing same PLM, our implementation demonstrated similar performance to NetSurfP-3.0.

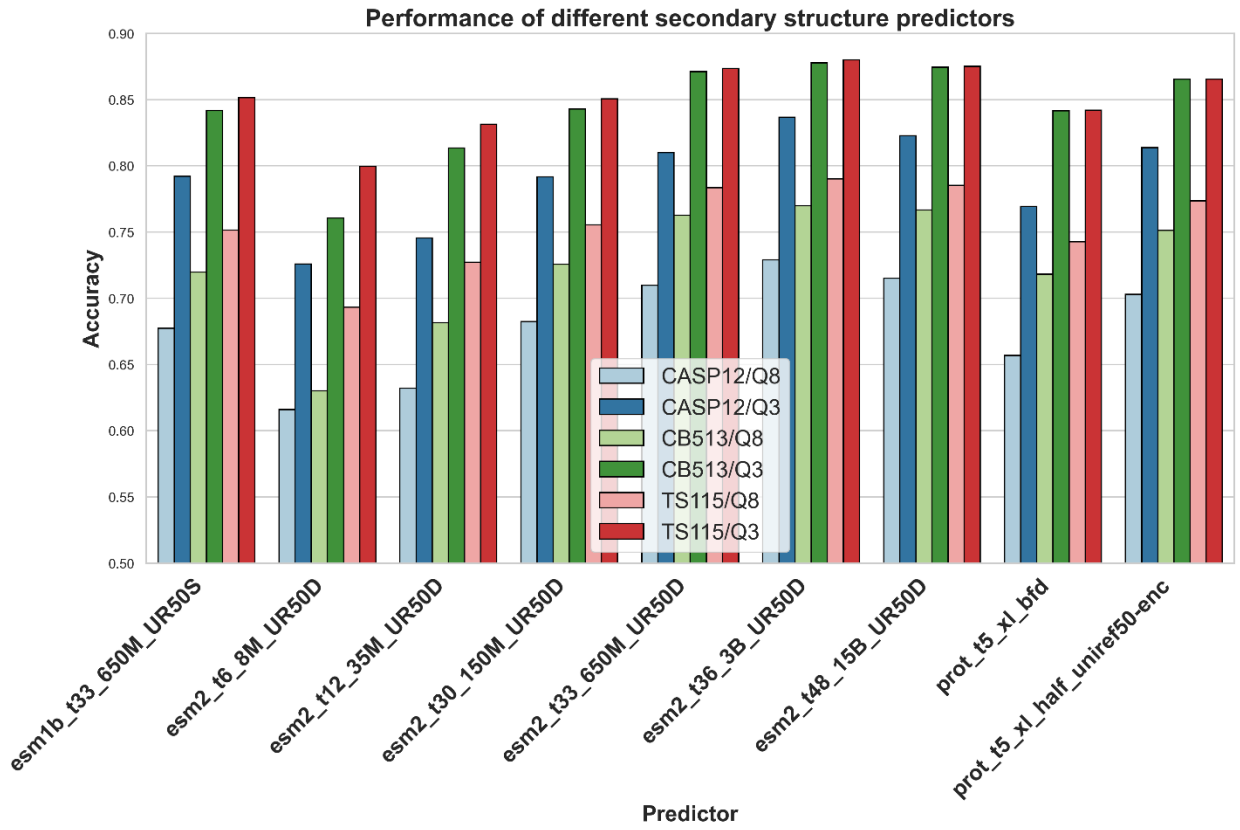


Figure 4. Accuracy of secondary structure predictors based on different PLMs.

Table 3. Performance of Various SSPs

Name	CASP12		CB513		TS115	
	Q8	Q3	Q8	Q3	Q8	Q3
IDP-ELM, esm2 3B (This work)	**0.7291**	**0.8368**	**0.7698**	**0.8778**	**0.7902**	**0.8799**
IDP-ELM, esm1b (This work)	0.6774	0.7922	0.7197	0.8419	0.7515	0.8515
NetSurfP-2.0	0.699	0.810	0.713	0.845	0.740	0.849
NetSurfP-3.0	0.669	0.791	0.711	0.846	0.749	0.856

Ensemble learning improves the performance of IDP prediction

We conduct a comparison of our IDP predictors based on different PLMs, evaluating various metrics on the test dataset (Table 4). Different PLM-based predictors exhibit notable variations in performance. The smallest PLM (esm2_t6_8M_UR50D) performs unfavorably in all metrics due to its limited model size. However, larger PLMs do not consistently outperform smaller ones, e.g., when comparing 150M ESM model (esm2_t30_150M_UR50D) to 650M ESM model (esm2_t33_650M_UR50D). The latter has worse scores except for the Rec score. The performance of each predictor also displays variability across different metrics, as seen in the comparison between 3B ESM model (esm2_t36_3B_UR50D) and 15B ESM model (esm2_t48_15B_UR50D). We conclude that this variability arises due to several factors: 1) IDRs exhibit an imbalanced distribution within proteins, leading to greater variance in performance. 2) The labeled datasets, encompassing training, validation, and test data, are relatively small, limiting the predictive capabilities of these PLM-based IDP predictors. 3) PLMs are primarily designed to capture self-interactions among residues, and larger PLMs possess greater capacity for learning and storing such interaction patterns across billions of data points. Since IDRs, being unstructured regions in proteins, may lack strong correlations with other residues in the same protein, the advantages conferred by large model sizes offer limited benefits in the context of IDP prediction.

Despite the disparities among different PLM-based IDP predictors, we employ ensemble learning to augment their performance, named *ensemble*. As depicted in the final line of Table 4, we observe improvements in AUC, F1, and MCC scores, as well as favorable results in Rec and Pre scores for the ensemble predictor, thereby underscoring the necessity of employing such technique. It is important to note that the Rec and Pre are biases metrics: although the 15B ESM

model achieves the highest Rec score, it attains the second-worse Pre score, while the 150M ESM model achieves the highest Pre score but has a worse Rec score.

Furthermore, we compared the DFL predictors and DP predictors utilizing different PLMs (Supplementary Tables S10 and Table S11). The results indicate that ensemble learning indeed enhances the performance compared to the single predictor, as evidenced by improved scores in AUC, F1, and MCC scores.

Table 4. Performance of IDP Prediction Based on Different PLMs on Test Dataset

Models	AUC	F1	MCC	Rec	Pre
esm1b_t33_650M_UR50S	0.8345	0.6310	0.4785	0.7265	0.5577
esm2_t6_8M_UR50D	0.8108	0.5881	0.4150	0.8515	0.4491
esm2_t12_35M_UR50D	0.8271	0.6217	0.4626	0.7849	0.5147
esm2_t30_150M_UR50D	0.8334	0.6321	0.4812	0.7156	**0.5660**
esm2_t33_650M_UR50D	0.8321	0.6284	0.4746	0.8271	0.5066
esm2_t36_3B_UR50D	0.8278	0.6217	0.4625	0.7640	0.5241
esm2_t48_15B_UR50D	0.8332	0.5957	0.4301	**0.8744**	0.4518
prot_t5_xl_bfd	0.8351	0.6157	0.4589	0.8610	0.4792
prot_t5_xl_uniref50	0.8395	0.6031	0.4400	0.8639	0.4633
ensemble	**0.8469**	**0.6325**	**0.4814**	0.8344	0.5092

IDP-ELM based IDP predictions achieved the state-of-the-art performance

We compared the performance of various IDP predictors by evaluating on the non-redundant CAID test dataset. The predictions of the compared predictors were obtained from the CAID competition, and the results on the redundant CAID are also provided in Supplementary Table S6. It is important to note that we utilized the default threshold of 0.5 for evaluating all predictors, while the metrics reported in CAID paper[30] are based on the threshold selection that maximizes the F1 score.

Our IDP-ELM exhibits a remarkable advantage in terms of AUC (0.8554), F1 (0.5031), and MCC (0.4332), and it achieves top performance on Rec (0.8724), as shown in Supplementary Table S7. IDP-ELM outperforms other state-of-the-art predictors except the value of Pre (0.3535), which is relatively lower than the best achieved values of 0.5042 by fIDPnn. Figure 5 presents the top 10 IDP predictors plotted with AUC, F1, and MCC metrics. Additionally, the ROC (receiver operating characteristic) curve, displayed in Figure 6A, clearly illustrates the significant improvement of IDP-ELM compared to all other predictors. The precision-recall curve demonstrates that our IDP-ELM attains the highest F1 score when the optimal thresholds are chosen (Figure 6B), where the gray dashed curves represent the points with same F1 scores. Furthermore, we conducted a Student's *t*-test to compare the predictions of IDP-ELM with those of other IDP predictors, and the results indicate that IDP-ELM significantly outperforms all existing methods (P-value<0.05).

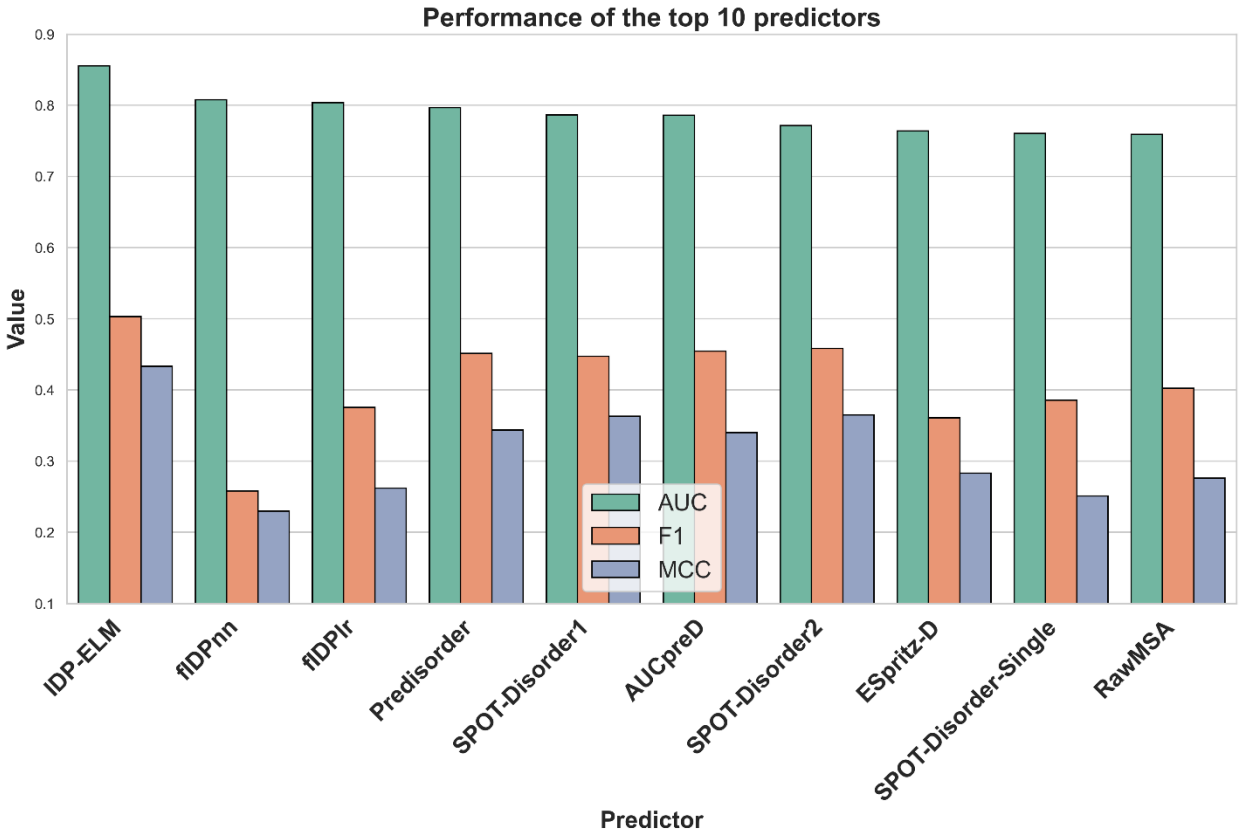


Figure 5. Performance of top 10 predictors on non-redundant CAID test dataset.

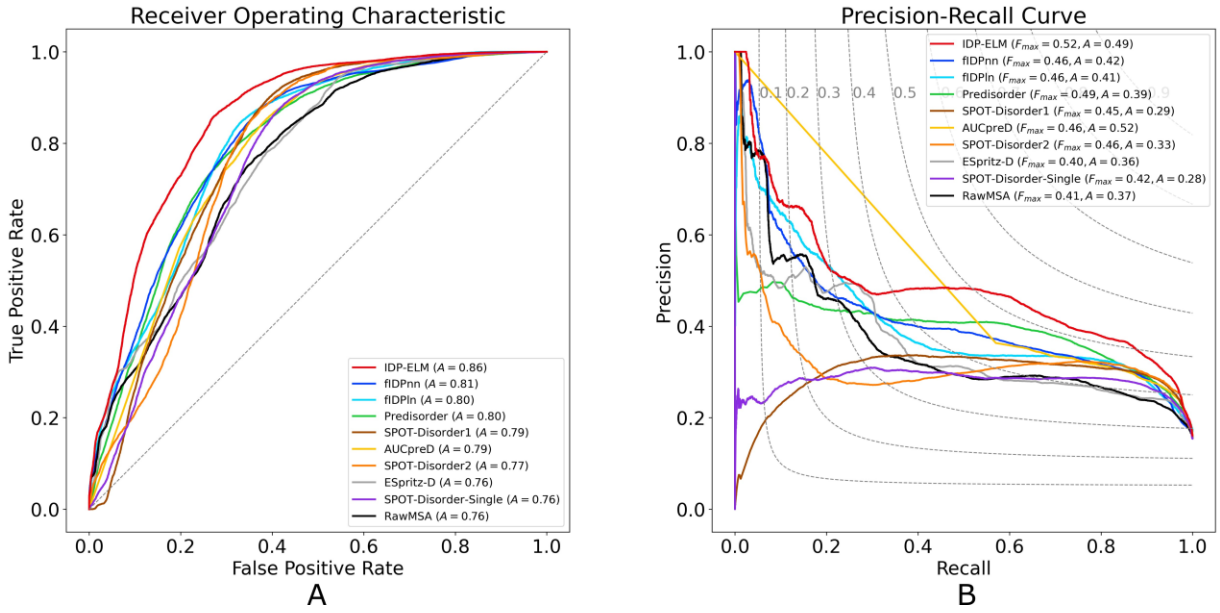


Figure 6. A. Receiver operating characteristic curves and B. Precision-recall curves of top 10 IDP predictors on non-redundant CAID test dataset.

To comprehensively assess the performance of various IDP predictors, we constructed a new test dataset based on the CAID2 competition[31] test dataset. The predictions of the compared predictors were obtained from the CAID2 competition, and the results on redundant CAID2 can also be found in Supplementary Table S8. It is worth noting that some predictors, e.g., SPOT-Disorder2 failed on a large part of sequences (64 out of 284) since it relies on many input features from other complicated predictors that may fail to produce features for some sequences, while our IDP-ELM always produces reliable predictions since it requires only the sequence.

Our IDP-ELM exhibits a remarkable advantage in terms of AUC (0.9788), F1 (0.9186), MCC (0.8717), and Rec (0.9457), while it achieves top performance on Pre (0.8930), as shown in Supplementary Table S9. Figure 7 presents the top 10 IDP predictors plotted with AUC, F1, and

MCC metrics. In addition, the ROC (receiver operating characteristic) curve in Figure 8A clearly illustrates the significant improvement of IDP-ELM compared to all other predictors. The precision-recall curve, depicted in Figure 8B, demonstrates that our IDP-ELM attains the highest F1 score when the optimal thresholds are chosen, where the gray dashed curve represent points with same F1 scores. Furthermore, we conducted a Student's *t*-test to compare the predictions of IDP-ELM with those of other IDP predictors, and the results indicate IDP-ELM outperforms all existing methods significantly (P-value < 0.05) except for RawMSA (P-value = 0.2361). However, RawMSA obviously underperforms our IDP-ELM on non-redundant CAID test dataset.

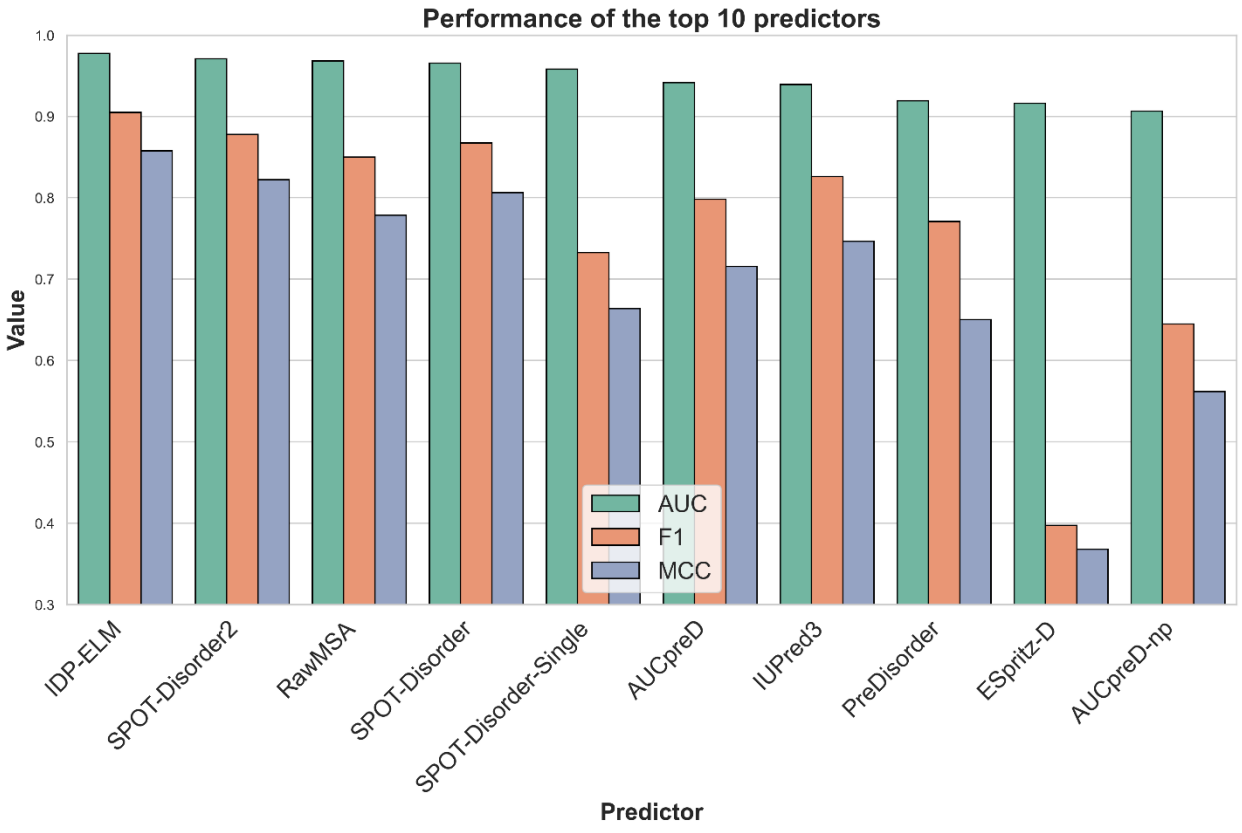


Figure 7. Performance of top 10 IDP predictors on non-redundant CAID2 test dataset.

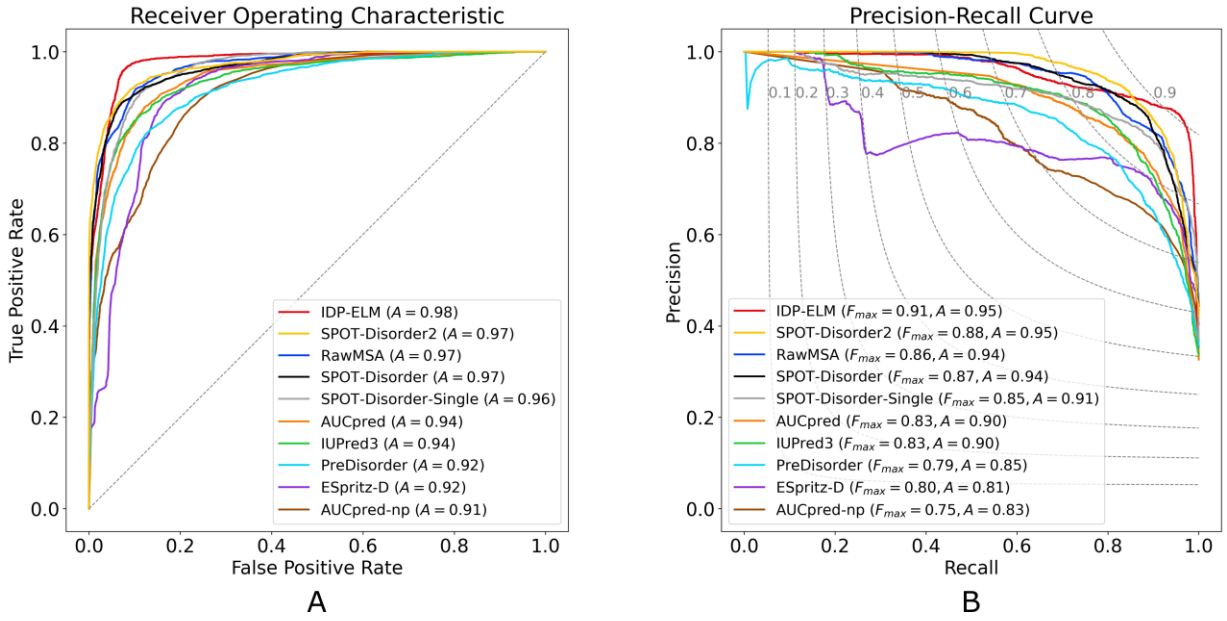


Figure 8. A. Receiver operating characteristic curves and B. Precision-recall curves of top 10 IDP predictors on non-redundant CAID2 test dataset.

Exemplary visualizations

To test our IDP-ELM, two protein sequences were analyzed by different top predictors, and the results of their predictions are visualized in Figure 9 and Figure 10, demonstrating the accuracy of our IDP-ELM. Note that some top predictors, e.g., SPOT-Disorder2 failed to predict these sequences, and we do not compare them here.

The first example is SNARE (Vesicle transport through interaction with t-SNAREs homolog 1A, UniProt: Q9JI51, DisProt: DP01499), which was initially discovered to be unstructured (residue 115-192) in solution but assumed to be a helical conformation when engaged in interactions with other molecules[47]. The presence of diverse conformations within its structure makes it challenging to predict disorder accurately. However, our IDP-ELM successfully

captures the disordered regions except a few false positives in the terminal, while other predictors fail, with predicting nothing in the IDR or too many false positives (Figure 9).

The second example showcases the HspB6 (Heat shock protein beta-6, UniProt: O14558, DisProt: DP01131), which has a long disordered region spanning from residue 1 to 56. Since the full-length human HSPB6 have not been crystallized by extensive attempts, this protein was crystallized by removing poorly ordered region by limited proteolysis techniques[48]. Notably, only the prediction from our IDP-ELM matches with the sequence of a long-disordered region with a short false positive region in the tail of the sequence, while the others fail with predicting nothing in the IDR or too many false positives in ordered regions (Figure 10).

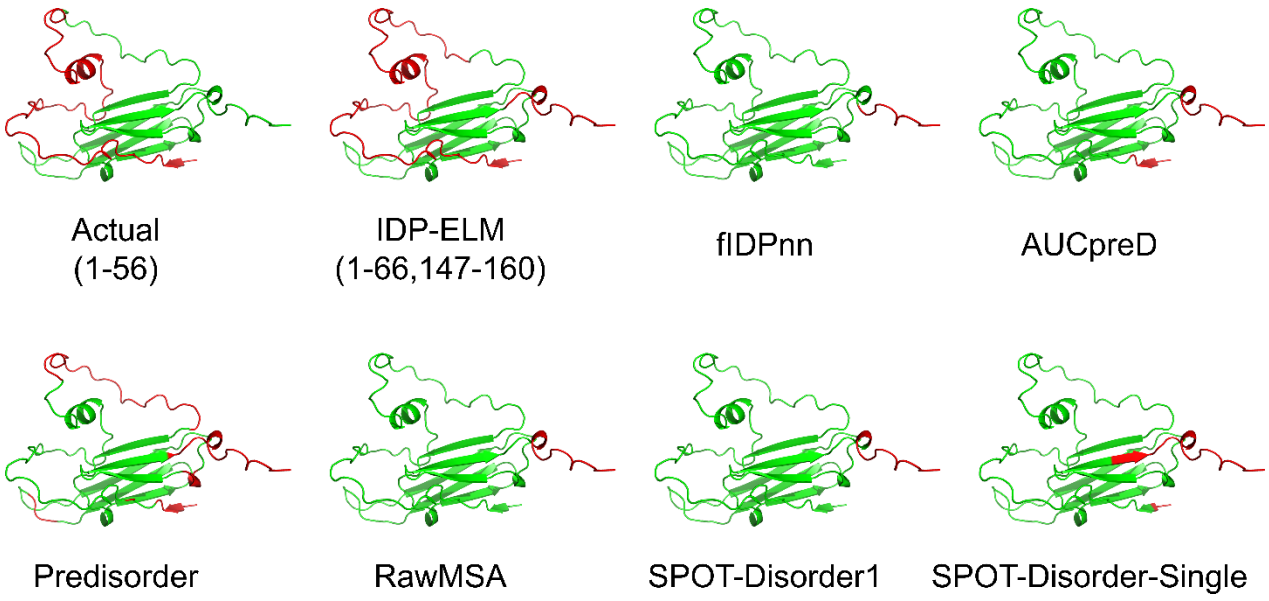


Figure 9. Visualization of selected example (DisProt ID: DP01499)
3D structures are predicted by AlphaFold2, and cartoons are visualized by PyMOL. The residues are colored in red (disordered) and green (ordered).

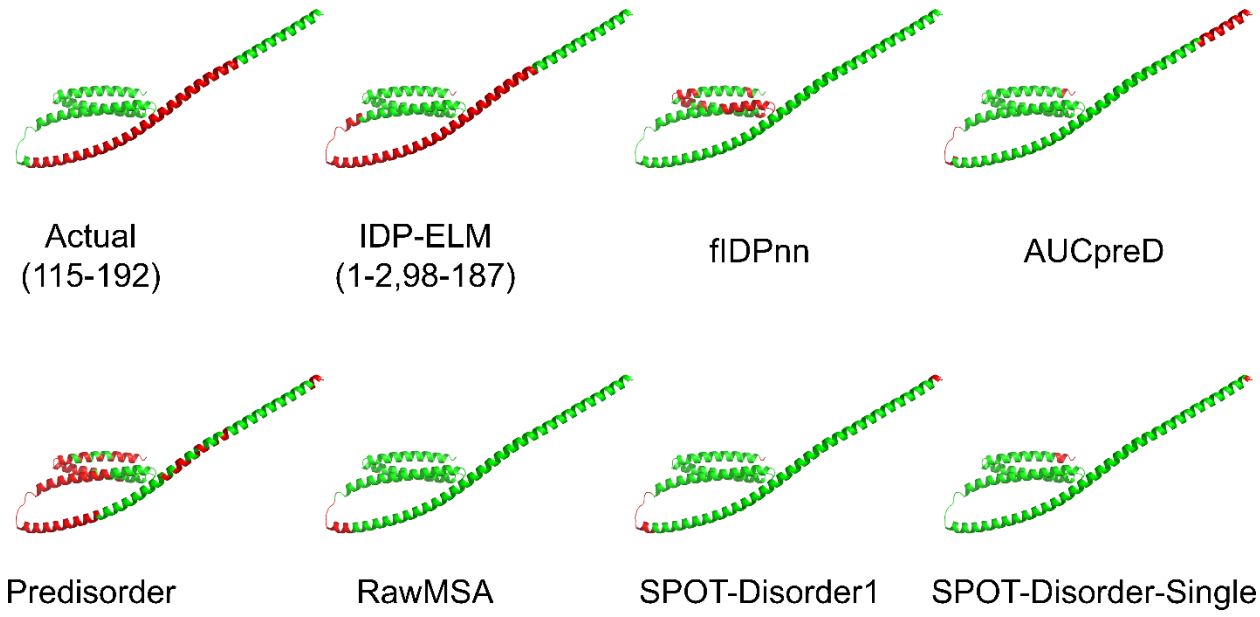


Figure 10. Visualization of selected example (DisProt ID: DP01131)
3D structures are predicted by AlphaFold2, and cartoons are visualized by PyMOL. The residues are colored in red (disordered) and green (ordered).

PLM-based IDR function predictions achieve state-of-the-art performance

We thoroughly evaluate the performance of our disordered flexible linker (DFL) predictor on an independent test dataset. Our implementation utilizing PLM attained the highest AUC, MCC, and Rec scores compared to other state-of-the-art methods, and the second-best Pre score (Table 5). The other three methods including TransDFL[49], APOD[34], and DFLpred[50] employed various input features while our methods employed only high-dimensional representations as well as predicted IDRs. Considering the performance of DFL prediction are limited by the extremely limited labelled data size, our method shows the superior advantage of high-dimensional representations in such a scenario.

Table 5. Comparison of Different DFL Predictors on the Test Dataset

Predictor	AUC	MCC	Rec	Pre
IDP-ELM (This work)	**0.8670**	**0.5221**	**0.6482**	0.5647
TransDFL	0.835	0.435	0.452	**0.586**
APOD	0.816	0.418	0.512	0.512
DFLpred	0.637	0.145	0.179	0.337

Furthermore, we evaluate our disordered protein-binding (DP) predictor on the independent test dataset. Our PLM-based implementation outperforms other state-of-the-art methods in terms of AUC, F1, Rec, and Pre scores (Table 6). Note that here we employed thresholds that fit specificity as 0.8, which is the same as the settings of other compared methods in Table 6. The other four DP prediction methods including DeepDISOBind[35], MoRFChibiLight[51], DisoRDPbind[52], and ANCHOR2[53] employed various input features while our methods employed only high-dimensional representations as well as predicted IDRs. Our method shows the superior advantage of high-dimensional representations in such a scenario.

Table 6. Comparison of Different DP Predictors on the Test Dataset

Predictor	AUC	F1	Rec	Pre
IDP-ELM (This work)	**0.8748**	**0.4202**	**0.6711**	**0.3058**
DeepDISOBind	0.771	0.313	0.595	0.212
MoRFChibiLight	0.735	0.270	0.502	0.185
DisoRDPbind	0.727	0.249	0.456	0.171
ANCHOR2	0.719	0.270	0.501	0.185

Fast IDP prediction for human proteome

We executed our IDP-ELM on the human proteome to showcase the advantages of our methodology in terms of efficiency. The human proteome dataset, encompassing 34,877 proteins, was obtained from MobiDB dataset[54]. Since human proteome have not been annotated experimentally, we employed the Mobi-lite annotations (V5.0 released in July 2022), which were derived from the consensus of eight distinct IDP predictors, demonstrating a 62.5% agreement rate and undergoing subsequent processing[29, 55]. It is worth noting that the Mobi-lite ranked 12th in CAID2 and 20th in CAID, implying that they serve as mere references and cannot be considered absolute truths.

It is important to note the computational configurations which we adopted in human proteome prediction. Throughout the prediction pipeline, we utilized three NVIDIA 3090Ti 24 GB GPUs to accommodate the substantial demands of the ever largest PLM esm2_t48_15B_UR50D, which necessitates approximately 70 GB of GPU memory. The distribution of all ESM models across multiple GPUs was facilitated using the Python package accelerate 0.18.0[56] with Transformers 4.26.0. Apart from the native automatic mixed precision in PyTorch, no further optimizations were implemented.

We generated a length-time curve, with the x-axis representing the sequence length of protein sequences and the y-axis denoting the inference time for each length (Figure 11). The point on the graph suggests a linear increase in inference time as the sequence length expands. While a naïve self-attention mechanism would yield a time complexity of $O(N^2)$ where N is the length of the sequence, our approach involves truncating protein sequences to a constant length, resulting in a linear time complexity. For most protein sequences comprising hundreds of amino acid residues, IDP-ELM can deliver predictions in just one second. We also employed the Mobi-lite

annotations to evaluate the predicted results, which yielded scores of 0.8823 for AUC, 0.6520 for F1, 0.5458 for MCC, 0.9030 for Rec, and 0.5101 for Pre.

Furthermore, we compared the speed on the redundant CAID dataset, consisting of 646 sequences and 336,595 residues. The average prediction time per sequence of our method was 0.8 seconds, outperforming several state-of-the-art methods mentioned in the cited work[57]: flDPnn (0.5-1 min), SPOT-Disorder2 (>10 min), SPOT-Disorder-Single (0.8-1 min), RawMSA (>10 min), SPOT-Disorder (5 min) and AUCpreD-np (0.5 min).

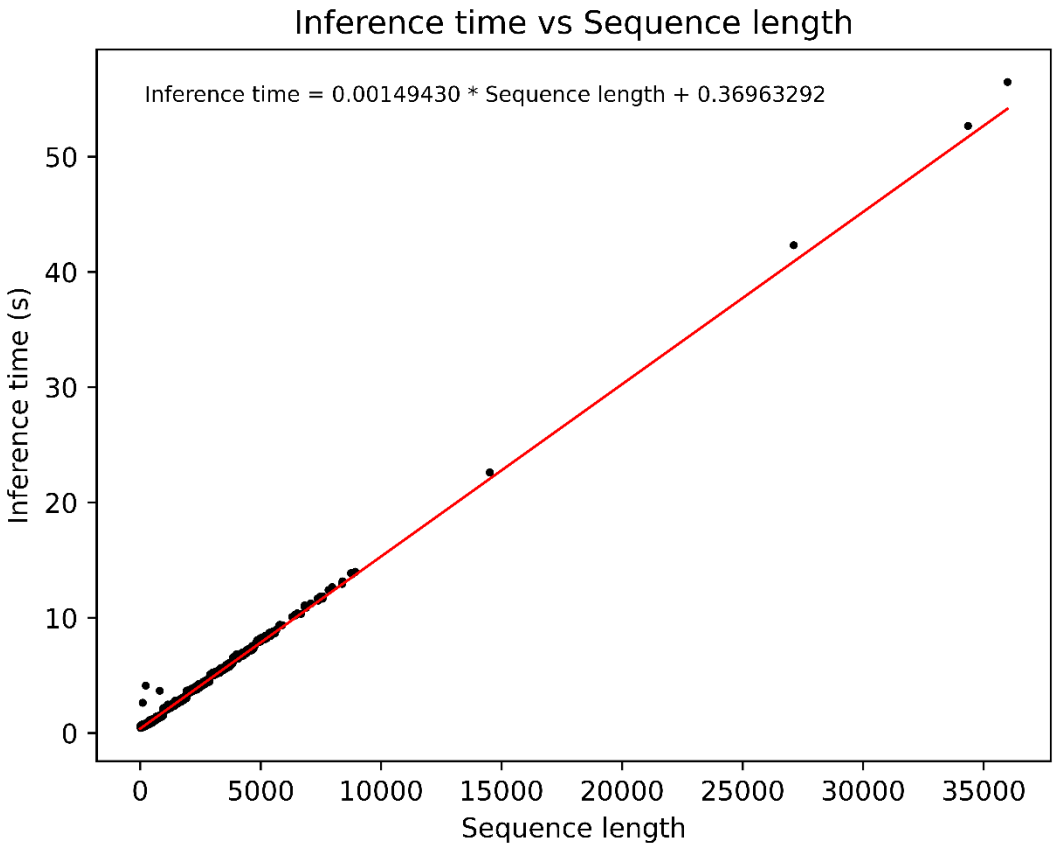


Figure 11. Inference time of IDP-ELM for human proteome prediction.

Conclusion

In conclusion, our study presents a novel IDP prediction method named IDP-ELM, which leverages the power of protein language models (PLMs) and ensemble learning to accurately predict intrinsically disordered regions (IDRs) and their associated functions, including disordered flexible linkers and disordered protein-binding. By utilizing high-dimensional representations from multiple PLMs and employing averaging strategy from ensemble learning, IDP-ELM achieves the state-of-the-art performance of IDR prediction on two challenging independent test datasets CAID and CAID2. Notably, IDP-ELM always offers a rapid and accurate prediction by solely utilizing protein sequences as inputs, while traditional approaches rely on time-consuming MSA generation and may fail to predict some sequences. This paradigm enables efficient proteome-level analysis without compromising accuracy. Furthermore, our predictions of IDR functions showcase advancements over existing IDR function prediction methods on several test datasets, underscoring the robustness and effectiveness of the IDP-ELM approach.

Overall, our research contributes to the growing field of IDP prediction and highlights the potential of PLMs in extracting vital information from protein sequences. The IDP-ELM method not only enhances the accuracies of IDP prediction but also opens avenues for further exploration and advancements in understanding the role of intrinsically disordered proteins in biological processes.

Authors' contributions

All authors conceived and designed the study. **SX** developed method, performed data analysis, wrote the paper. **AO** supervised the research project, and critically reviewed and revised the manuscript. All authors approved the final manuscript.

Competing interests

The authors have declared no competing interests.

Data and Software Availability

The corresponding reproducible source code, model weights, datasets are publicly available at https://github.com/xu-shi-jie/idp-elm.

Supporting Information Available

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/XXX.

Hyper-parameters for training IDP, DP, and DFL predictors, the performance of the single predictor based on different protein language models, and the comparison of various IDP prediction methods on CAID and CAID2 test datasets (PDF).

Acknowledgements

This work was supported by Hokkaido University DX Doctoral Fellowship (JST SPRING, Grant Number JPMJSP2119) to S.X. and JST SATREPS Grant Number JPMJSA2206. to A.O.

References

(1) Habchi, J.; Tompa, P.; Longhi, S.; Uversky, V. N. Introducing Protein Intrinsic Disorder. *Chem. Rev.* **2014**, *114*, 6561-6588.
(2) Peng, Z. L.; Yan, J.; Fan, X.; Mizianty, M. J.; Xue, B.; Wang, K.; Hu, G.; Uversky, V. N.; Kurgan, L. Exceptionally Abundant Exceptions: Comprehensive Characterization of Intrinsic Disorder in All Domains of Life. *Cell. Mol. Life Sci.* **2015**, *72*, 137-151.
(3) Boeckmann, B.; Bairoch, A.; Apweiler, R.; Blatter, M. C.; Estreicher, A.; Gasteiger, E.; Martin, M. J.; Michoud, K.; O'Donovan, C.; Phan, I. The SWISS-PROT Protein Knowledgebase and Its Supplement TrEMBL in 2003. *Nucleic Acids Res.* **2003**, *31*, 365-370.
(4) Xue, B.; Dunker, A. K.; Uversky, V. N. Orderly Order in Protein Intrinsic Disorder Distribution: Disorder in 3500 Proteomes from Viruses and the Three Domains of Life. *J. Biomol. Struct. Dyn.* **2012**, *30*, 137-149.
(5) Carballo-Pacheco, M.; Strodel, B. Comparison of Force Fields for Alzheimer's A β42: A Case Study for Intrinsically Disordered Proteins. *Protein Sci.* **2017**, *26*, 174-185.
(6) Coskuner, O.; Uversky, V. N. Intrinsically Disordered Proteins in Various Hypotheses on the Pathogenesis of Alzheimer's and Parkinson's Diseases. *Prog. Mol. Biol. Transl. Sci.* **2019**, *166*, 145-223.
(7) Rajagopalan, K.; Mooney, S. M.; Parekh, N.; Getzenberg, R. H.; Kulkarni, P. A Majority of the Cancer/Testis Antigens are Intrinsically Disordered Proteins. *J. Cell. Biochem.* **2011**, *112*, 3256-3267.
(8) Ambadipudi, S.; Zweckstetter, M. Targeting Intrinsically Disordered Proteins in Rational Drug Discovery. *Expert Opin. Drug Discovery* **2016**, *11*, 65-77.
(9) Quaglia, F.; Meszaros, B.; Salladini, E.; Hatos, A.; Pancsa, R.; Chemes, L. B.; Pajkos, M.; Lazar, T.; Pena-Diaz, S.; Santos, J. DisProt in 2022: Improved Quality and Accessibility of Protein Intrinsic Disorder Annotation. *Nucleic Acids Res.* **2022**, *50*, D480-D487.
(10) Liu, Y. M.; Wang, X. L.; Liu, B. IDP-CRF: Intrinsically Disordered Protein/Region Identification Based on Conditional Random Fields. *Int. J. Mol. Sci.* **2018**, *19*, 2483.
(11) Tang, Y. J.; Pang, Y. H.; Liu, B. IDP-Seq2Seq: Identification of Intrinsically Disordered Regions Based on Sequence to Sequence Learning. *Bioinformatics* **2020**, *36*, 5177-5186.
(12) Liu, Y. M.; Wang, X. L.; Liu, B. RFPR-IDP: Reduce the False Positive Rates for Intrinsically Disordered Protein and Region Prediction by Incorporating Both Fully Ordered Proteins and Disordered Proteins. *Briefings Bioinf.* **2021**, *22*, 2000-2011.
(13) Hu, G.; Katuwawala, A.; Wang, K.; Wu, Z. H.; Ghadermarzi, S.; Gao, J. Z.; Kurgan, L. flDPnn: Accurate Intrinsic Disorder Prediction with Putative Propensities of Disorder Functions. *Nat. Commun.* **2021**, *12*, 4438.

(14) Rumelhart, D. E.; Williams, R. J. *Learning Internal Representations by Error Propagation*; MIT Press, **1987**.
(15) Fukushima, K. Neocognitron - A Self-Organizing Neural Network Model for a Mechanism of Pattern-Recognition Unaffected by Shift in Position. *Biol. Cybern.* **1980**, *36*, 193-202.
(16) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; Polosukhin, I. Attention is All You Need. *arxiv* **2017**.
(17) Perdigao, N.; Rosa, A. Dark Proteome Database: Studies on Dark Proteins. *High-Throughput* **2019**, *8*, 8.
(18) Zarin, T.; Strome, B.; Nguyen Ba, A. N.; Alberti, S.; Forman-Kay, J. D.; Moses, A. M. Proteome-Wide Signatures of Function in Highly Diverged Intrinsically Disordered Regions. *Elife* **2019**, *8*.
(19) Bateman, A.; Martin, M. J.; Orchard, S.; Magrane, M.; Ahmad, S.; Alpi, E.; Bowler-Barnett, E. H.; Britto, R.; Cukura, A.; Denny, P. UniProt: The Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.* **2022**, *32*, D115-D119.
(20) Wang, S.; Ma, J. Z.; Xu, J. B. AUCpreD: Proteome-Level Protein Disorder Prediction by AUC-Maximized Deep Convolutional Neural Fields. *Bioinformatics* **2016**, *32*, 672-679.
(21) Hanson, J.; Paliwal, K.; Zhou, Y. Accurate Single-Sequence Prediction of Protein Intrinsic Disorder by an Ensemble of Deep Recurrent and Convolutional Architectures. *J. Chem. Inf. Model.* **2018**, *58*, 2369-2376.
(22) Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *arxiv* **2019**.
(23) Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. Improving Language Understanding by Generative Pre-Training. *arxiv* **2018**.
(24) Rives, A.; Meier, J.; Sercu, T.; Goyal, S.; Lin, Z.; Liu, J.; Guo, D.; Ott, M.; Zitnick, C. L.; Ma, J. Biological Structure and Function Emerge from Scaling Unsupervised Learning to 250 Million Protein Sequences. *Proc. Natl. Acad. Sci. U. S. A.* **2021**, *118*, e2016239118.
(25) Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; Verkuil, R.; Kabeli, O.; Shmueli, Y. Evolutionary-Scale Prediction of Atomic-Level Protein Structure with a Language Model. *Science* **2023**, *379*, 1123-1130.
(26) Elnaggar, A.; Heinzinger, M.; Dallago, C.; Rihawi, G.; Wang, Y.; Jones, L.; Gibbs, T.; Feher, T.; Angerer, C.; Steinegger, M. ProtTrans: Towards Cracking the Language of Life's Code Through Self-supervised Deep Learning and High Performance Computing. *arxiv* **2020**.
(27) Jin, C.; Cukier, R. I. Machine Learning Can be Used to Distinguish Protein Families and Generate New Proteins Belonging to Those Families. *J. Chem. Phys.* **2019**, *151*.
(28) Heinzinger, M.; Elnaggar, A.; Wang, Y.; Dallago, C.; Nechaev, D.; Matthes, F.; Rost, B. Modeling Aspects of the Language of Life Through Transfer-Learning Protein Sequences. *BMC Bioinf.* **2019**, *20*, 1-17.
(29) Necci, M.; Piovesan, D.; Dosztanyi, Z.; Tosatto, S. C. E. MobiDB-lite: Fast and Highly Specific Consensus Prediction of Intrinsic Disorder in Proteins. *Bioinformatics* **2017**, *33*, 1402-1404.
(30) Necci, M.; Piovesan, D.; Tosatto, S. C. E.; Predictors, C.; Curators, D. Critical Assessment of Protein Intrinsic Disorder Prediction. *Nat. Methods* **2021**, *18*, 472-481.
(31) Portal, C. P. *CAID-2 references*. **2022**. https://caid.idpcentral.org/challenge#Data (accessed 2023 Sep 13th).

(32) Høie, M. H.; Kiehl, E. N.; Petersen, B.; Nielsen, M.; Winther, O.; Nielsen, H.; Hallgren, J.; Marcatili, P. NetSurfP-3.0: Accurate and Fast Prediction of Protein Structural Features by Protein Language Models and Deep Learning. *Nucleic Acids Res.* **2022**, *50*, W510-W515.
(33) Huang, Y.; Niu, B. F.; Gao, Y.; Fu, L. M.; Li, W. Z. CD-HIT Suite: A Web Server for Clustering and Comparing Biological Sequences. *Bioinformatics* **2010**, *26*, 680-682.
(34) Peng, Z.; Xing, Q.; Kurgan, L. APOD: Accurate Sequence-Based Predictor of Disordered Flexible Linkers. *Bioinformatics* **2020**, *36*, i754–i761.
(35) Zhang, F. H.; Zhao, B.; Shi, W. B.; Li, M.; Kurgan, L. DeepDISOBind: Accurate Prediction of RNA-, DNA- and Protein-Binding Intrinsically Disordered Residues with Deep Multi-Task Learning. *Briefings Bioinf.* **2022**, *23*, bbab521.
(36) Rao, R.; Meier, J.; Sercu, T.; Ovchinnikov, S.; Rives, A. Transformer Protein Language Models are Unsupervised Structure Learners. *Biorxiv* **2020**.
(37) Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y. Q.; Li, W.; Liu, P. J. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.* **2020**, *21*, 5485-5551.
(38) Suzek, B. E.; Wang, Y.; Huang, H.; McGarvey, P. B.; Wu, C. H.; Consortium, U. UniRef Clusters: A Comprehensive and Scalable Alternative for Improving Sequence Similarity Searches. *Bioinformatics* **2015**, *31*, 926-932.
(39) Steinegger, M.; Söding, J. Clustering Huge Protein Sequence Sets in Linear Time. *Nat. Commun.* **2018**, *9*, 2542.
(40) Steinegger, M.; Mirdita, M.; Soding, J. Protein-Level Assembly Increases Protein Sequence Recovery from Metagenomic Samples Manyfold. *Nat. Methods* **2019**, *16*, 603-606.
(41) Bairoch, A.; Bougueleret, L.; Altairac, S.; Amendolia, V.; Auchincloss, A.; Puy, G. A.; Axelsen, K.; Baratin, D.; Blatter, M. C.; Boeckmann, B. The Universal Protein Resource (UniProt). *Nucleic Acids Res.* **2008**, *36*, D190-D195.
(42) Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. *arXiv* **2017**.
(43) Kingma, D. P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2014**.
(44) Lopez-del Rio, A.; Martin, M.; Perera-Lluna, A.; Saidi, R. Effect of Sequence Padding on the Performance of Deep Learning Models in Archaeal Protein Functional Prediction. *Sci. Rep.* **2020**, *10*, 14634.
(45) Kabsch, W.; Sander, C. Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen‐Bonded and Geometrical Features. *Biopolymers* **1983**, *22*, 2577-2637.
(46) Yang, Y.; Gao, J.; Wang, J.; Heffernan, R.; Hanson, J.; Paliwal, K.; Zhou, Y. Sixty-Five Years of the Long March in Protein Secondary Structure Prediction: The Final Stretch? *Briefings Bioinf.* **2018**, *19*, 482-494.
(47) Zwilling, D.; Cypionka, A.; Pohl, W. H.; Fasshauer, D.; Walla, P. J.; Wahl, M. C.; Jahn, R. Early Endosomal SNAREs Form a Structurally Conserved SNARE Complex and Fuse Liposomes with Multiple Topologies. *EMBO J.* **2007**, *26*, 9-18.
(48) Weeks, S. D.; Baranova, E. V.; Heirbaut, M.; Beelen, S.; Shkumatov, A. V.; Gusev, N. B.; Strelkov, S. V. Molecular Structure and Dynamics of the Dimeric Human Small Heat Shock Protein HSPB6. *J. Struct. Biol.* **2014**, *185*, 342-354.
(49) Pang, Y.; Liu, B. TransDFL: Identification of Disordered Flexible Linkers in Proteins by Transfer Learning. *Genomics, Proteomics Bioinf.* **2022**.
(50) Meng, F. C.; Kurgan, L. DFLpred: High-Throughput Prediction of Disordered Flexible Linker Regions in Protein Sequences. *Bioinformatics* **2016**, *32*, 341-350.

(51) Malhis, N.; Jacobson, M.; Gsponer, J. MoRFchibi SYSTEM: Software Tools for the Identification of MoRFs in Protein Sequences. *Nucleic Acids Res.* **2016**, *44*, W488-W493.
(52) Oldfield, C. J.; Peng, Z.; Kurgan, L. Disordered RNA-Binding Region Prediction with DisoRDPbind. *RNA Chaperones: Methods and Protocols* **2020**, 225-239.
(53) 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*, W329-W337.
(54) Piovesan, D.; Del Conte, A.; Clementel, D.; Monzon, A. M.; Bevilacqua, M.; Aspromonte, M. C.; Iserte, J. A.; Orti, F. E.; Marino-Buslje, C.; Tosatto, S. C. E. MobiDB: 10 Years of Intrinsically Disordered Proteins. *Nucleic Acids Res.* **2023**, *51*, D438-D444.
(55) Necci, M.; Piovesan, D.; Clementel, D.; Dosztanyi, Z.; Tosatto, S. C. E. MobiDB-lite 3.0: Fast Consensus Annotation of Intrinsic Disorder Flavors in Proteins. *Bioinformatics* **2020**, *36*, 5533-5534.
(56) Sylvain Gugger, L. D., Thomas Wolf, Philipp Schmid, Zachary Mueller, Sourab Mangrulkar. *Accelerate: Training and Inference at Scale Made Simple, Efficient and Adaptable.* **2022**. https://github.com/huggingface/accelerate (accessed 2023 Sep 13th).
(57) Zhao, B.; Kurgan, L. Deep Learning in Prediction of Intrinsic Disorder in Proteins. *Comput. Struct. Biotechnol. J.* **2022**, *20*, 1286-1294.

Accurate and Fast Prediction of Intrinsically Disordered Protein by Multiple Protein Language Models and Ensemble Learning

Shijie Xu and Akira Onoda*

