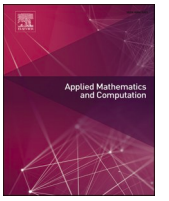
Applied Mathematics and Computation 472 (2024) 128630

Contents lists available at ScienceDirect

Applied Mathematics and Computation

journal homepage: www.elsevier.com/locate/amc



Full Length Article

DisPredict3.0: Prediction of intrinsically disordered regions/ proteins using protein language model

Md Wasi Ul Kabir, Md Tamjidul Hoque *

Department of Computer Science, University of New Orleans, New Orleans, LA, USA

ARTICLE INFO

Keywords:
Protein language models
Intrinsically disordered proteins
Predict disordered protein
Machine learning

ABSTRACT

Intrinsically disordered proteins (IDPs) or protein regions (IDRs) do not have a stable three-dimensional structure, even though they exhibit important biological functions. They are structurally and functionally very different from ordered proteins and can cause many critical diseases. Accurate identification of disordered proteins/regions significantly impacts fields such as drug design, protein engineering, protein design, and related research. However, experimental identification of IDRs is complex and time-consuming, necessitating the development of an accurate and efficient computational method. The recent development of deep learning methods for protein language models shows the ability to learn evolutionary information from billions of protein sequences. This motivates us to develop a computational method, named DisPredict3.0, to predict proteins' disordered regions (IDRs) using evolutionary information from a protein language model. Compared to the state-of-the-art method in the CAID (2018) assessment, DisPredict3.0 has an improvement of 2.51 %, 16.13 %, 17.98 %, and 11.94 % in terms of AUC, F1-score, MCC, and kappa, respectively. In addition, in the CAID-2 assessment (2022), DisPredict3.0 shows promising results and is ranked first for disorder residue prediction on the Disorder-NOX dataset. The DisPredict3.0 webserver is available at https://bmll.cs.uno.edu.

1. Introduction

Intrinsically disordered proteins and protein regions are important to understanding protein functions. Despite lacking a well-defined secondary and tertiary structure, they still play various crucial roles. Disordered proteins are the main reason related to various diseases [1], including cancer, cardiovascular disease, Alzheimer's disease, Parkinson's disease, etc. There are abundant disordered protein sequences available in organisms, i.e., eukaryotes, in which disordered proteins are found in over 30 % of proteins [2]. Scientists have known about the disordered region of proteins for 20 years [3], yet the accuracy of the existing computational methods to predict the disordered regions/proteins is low [1,2,4-7].

Community-driven assessments have been conducted to evaluate disorder predictors. The first such assessment was held in 2005 during the Critical Assessment of Protein Structure Prediction (CASP) [8], where disorder predictors were evaluated. Recently, a new assessment initiative, named the Critical Assessment of Protein Intrinsic Disorder (CAID), was initiated. CAID is specifically designed to evaluate disordered protein prediction. The inaugural CAID event took place in 2018 [9] and evaluated several recent methods, i.e., flDPnn [2], AUCpreD [10], ESpritz [11], RawMSA [12], SPOT-Disorder2 [6], and SPOT-Disorder-Single [13].

The recent advancements in neural networks have inspired researchers to apply deep learning techniques to predict disordered

* Corresponding author.
E-mail address: thoque@uno.edu (M.T. Hoque).

https://doi.org/10.1016/j.amc.2024.128630
Received 2 August 2023; Received in revised form 20 January 2024; Accepted 21 February 2024

proteins. In CAID (2022) assessment [5], the majority of the recent methods were developed using deep learning approaches. In AUCpreD [10], the authors implemented Deep Convolutional Neural Fields (DeepCNF), which combines deep convolutional neural networks (DCNN) and conditional random fields (CRF) to understand intricate sequence-structure relationships in a hierarchical way and correlations among neighboring residues. The model was trained to maximize the Area under the ROC curve (AUC) to tackle the problem of imbalanced classes. In SPOT-Disorder2 [6], the authors employed an ensemble of deep Squeeze-and-Excitation residual inception and long short-term memory (LSTM) networks for predicting disordered proteins, with evolutionary information and predicted structural properties as input. Klausen et al. [14] introduced a method, NetSurfP-2.0, which utilizes an architecture comprising convolutional and long short-term memory neural networks. This single integrated model can predict disordered protein regions, secondary structure, solvent accessibility, structural disorder, and backbone dihedral angles for each residue [14]. In NetSurfP-3.0 [15], the authors use a pre-trained protein language model to enhance the runtime efficiency of NetSurfP-2.0 by two orders of magnitude without compromising its prediction accuracy. Dosztányi et al. [16] presented a method, IUPred, which is based on an energy estimation approach. The authors introduced new smoothing functions to prediction, which reduce noise and enhance the performance of predictions. IUPred utilizes an energy estimation approach to predict disordered proteins, emphasizing the contrasting biophysical characteristics exhibited by ordered and disordered regions [16]. In the latest version of IUPred, IUPred2A [17], the authors combined energy estimation-based predictions for ordered and disordered residues provided by IUPred2.

Walsh et al. [11] developed ESpritz, an ensemble of predictors for protein disorder, utilizing a bidirectional recursive neural network as its foundation. Similarly, Emenecker et al. [5] introduced a consensus score to address individual methods' potential biases or drawbacks. This score integrates the results from multiple independent disorder predictors to forecast whether a residue will be ordered or disordered [5]. They also used a bidirectional recurrent neural network to train the consensus disorder scores from 12 proteomes. Meanwhile, Hu et al. designed flDPnn [2], a deep neural network method for predicting disordered proteins, ranked as the top-performing method in the CAID 2018 evaluation. flDPnn extracts and encodes both structural and functional data from various sources, integrating the profile information at the residue, window, and protein levels. After combining all these features, they trained their neural network to predict disordered proteins. Additionally, flDPnn can predict protein, DNA, RNA binding, and linkers in disordered regions or proteins.

One of the most remarkable characteristics of proteins is that their structure and function are encoded in their amino acid sequence. Protein sequences bear several resemblances to human language. For example, just as human languages are represented, proteins are composed of 20 standard amino acids [18]. Biological databases are exponentially expanding in size, especially in terms of protein sequences. By leveraging these extensive databases, protein research can employ Natural Language Processing (NLP) techniques, such as training language models like BERT, to capture proteins' properties [18]. Language models have recently gained significant attention in the bioinformatics research community. Recent literature shows language models, such as Evolutionary Scale Modeling (ESM) [19], Evolutionary Scale Modeling 2 (ESM2) [20], ProtTrans [21], and Tranception [22], that encode the representation of proteins. These language models have demonstrated the ability to replace multiple sequence alignment (MSA) input requirements from AlphaFold2 [23] and accurately predict protein structures [24,25]. One of these recent protein language models, Evolutionary Scale Modeling (ESM), is trained on 86 billion amino acids across 250 million protein sequences spanning evolutionary diversity [19]. ESM is a Transformer-based deep neural network model with 34 layers that was pre-trained on the UniParc database to produce vector representations that encode protein sequences [18]. These representations are learned solely from the sequence data, and the authors have found evidence suggesting that the learned representation space contains evolutionary information about the structure of proteins, from the level of biochemical properties of amino acids to the remote homology of proteins [19]. Previous research studies have highlighted the significance of evolutionary features in addressing bioinformatics problems [2,6,26,27]. In addition to evolutionary information, the model also learned about proteins' secondary and tertiary structures [19]. This recent development in large language model motivates us to investigate the effects of the protein language model on the prediction of disordered regions/proteins.

Continuing our previous effort in the DisPredict [7] and DisPredict2 [28] methods, we have created a new disorder predictor known as DisPredict3.0. Built upon the foundation of a protein language model, DisPredict3.0 is trained on the representation from large protein language model in combination with features extracted from the flDPnn method. We further reduce the dimension of the protein representation using Principal Component Analysis (PCA) and apply the optimized Light Gradient Boosting Machine algorithm for model training. Additionally, we optimize the threshold to address the issue of imbalanced classes. Our study demonstrates that the representation of the protein's language model enhances protein disorder prediction. In the following sections, we first outline the dataset used for training and testing the DisPredict3.0 method. We then delve into the architecture of DisPredict3.0. Finally, we compare the performance of DisPredict3.0 with existing methods as described in the literature.

2. Material and method

In this section, we delve into various aspects of our research methodology. Firstly, we present an overview of the dataset used for training our model. This encompasses the sources from which the data was gathered, the selection criteria for data inclusion, and the preprocessing steps. Additionally, we outline how the dataset was divided into training, validation, and test sets, while highlighting the considerations made to ensure the diversity and representativeness of the dataset. Next, we describe the architecture of our proposed model, DisPredict3.0. We present in detail the various components of the model and how they interact with each other. Then, we describe the metrics used to evaluate the model's performance.

2.1. Dataset

DisPredict3.0 is trained and validated using the same dataset as the flDPnn method [2]. However, our test set differs from the flDPnn method, as we exclude a protein sequence (DP01026) that contains an unidentified amino acid (X). The authors [2] of the flDPnn method curated the 745 proteins dataset from the DisProt 7.0 database [29]. The ordered and disordered annotations are collected from the DisProt database. The regions not defined as disordered in the DisProt database are considered ordered. The DisProt database is the gold standard for disordered regions/proteins annotation and has been consistently updated. The recent release is from June 2022 at the time of writing this manuscript. The training, test, and validation set contains 445, 175, and 100 proteins, respectively. The training and test set share less than 25 % pairwise sequence identity to avoid overlap between the train and test set. Table 1 shows the statistics of the dataset. The dataset is imbalanced and significantly skewed towards ordered residues. In the training dataset, 77.1 % of the residues are ordered, while 22.9 % are disordered. Likewise, the test and validation datasets contain 73.1 % and 83.4 % ordered residues, respectively, and 26.9 % and 16.6 % disordered residues.

2.2. Architecture of dispredict3.0

DisPredict3.0 leverages the flDPnn method to gather residue-level, window-level, and protein-level sequence information. These features are extracted from various other methods used by the flDPnn method, such as PSIPRED [30], IUPred [17], and PSI-BLAST [31]. Inspired by the impressive performance of the flDPnn method in the CAID (2018) competition [9], we decided to use all pertinent features to train our model. In addition to this, we extract evolutionary data using the Transformer-based protein language model known as Evolutionary Scale Modeling (ESM) from Facebook AI Research [19]. ESM is trained on large protein sequences and can represent residues as vector representations.

DisPredict3.0 extracts representations from all 34 layers of the ESM-1b (esm1b_t33_650M_UR50S) pre-trained model [19], representing each residue with a matrix of size (34×1281). While extracting features from protein sequences, we encountered two challenges: high dimensional data and ESM's limitation of processing only sequences with a maximum of 1024 tokens. To manage the high dimensionality of the ESM features, we employed the Principal Component Analysis (PCA) technique [26], a widely used method for reducing dimensions in high-dimensional data. Specifically, we utilized incremental PCA to transform the dimensions (34×1281) into a vector size of 5124, preserving 96.02 % of the original data variance. To address the limited tokens, we split sequences exceeding 1024 amino acids into smaller segments and independently extracted features from each. These features are then combined to form a comprehensive representation of the entire protein sequence. Table 2 lists the number of features extracted using the flDPnn and ESM methods.

After collecting all the features from different methods, we trained an optimized Light Gradient Boosting Machine (LightGBM) algorithm [32]. LightGBM is a popular gradient-boosting framework based on tree-based learning algorithms. The algorithm develops trees leaf by leaf and selects the leaf that it thinks would result in the highest reduction in loss [32]. Furthermore, LightGBM uses a highly optimized histogram-based decision tree learning method, which offers significant efficiency and memory usage benefits [32]. The algorithm can train the model fast, even with a large dataset [32]. The architecture of DisPredict3.0 is presented in Fig. 1.

2.3. Performance evaluation metrics

The performance of DisPredict3.0 is evaluated using a suite of metrics that are commonly adopted for handling imbalanced datasets. These include the Area Under the Receiver Operating Characteristic curve (AUC), F1-score, Mathews Correlation Coefficient (MCC), and Kappa score. The AUC metric is a graphical representation of the performance of a classification model at all classification thresholds. AUC is particularly valuable because it is independent of any particular threshold, offering a comprehensive measure of model performance across all possible decision boundaries. In contrast, the F1-score, MCC, and Kappa scores are threshold-dependent metrics. This means their calculations are tied to a specific decision threshold, making them sensitive to changes in threshold. Due to its robustness and threshold independence, we have chosen the AUC as our primary evaluation metric to compare the performance of DisPredict3.0 with other methods in the field. The definitions and formulas of all the selected metrics are provided in Table 3 for further clarity.

3. Results

We provided a comprehensive evaluation of DisPredict3.0 method. We start by comparing the effectiveness of various machine-learning methodologies. Then, we explore the effectiveness of each feature using ablation analysis. Subsequently, we measure how DisPredict3.0 performs against current state-of-the-art methods in the field. We further test the proposed method specifically for fully

Table 1
Statistics of ordered and disordered residues in the training, test, and validation dataset.

Disordered/Ordered	Train	Test	Validation
No. of Disordered residues	50,387 (22.9 %)	17,871 (26.9 %)	4967 (16.6 %)
No. of Ordered residues	169,565 (77.1 %)	48,675 (73.1 %)	25,004 (83.4 %)
Total No. of Residues	219,952	66,546	29,971

Table 2
The number of features extracted from different methods.

Methods	No. of Features
flDPnn	317
ESM	5124
Total Features	5441

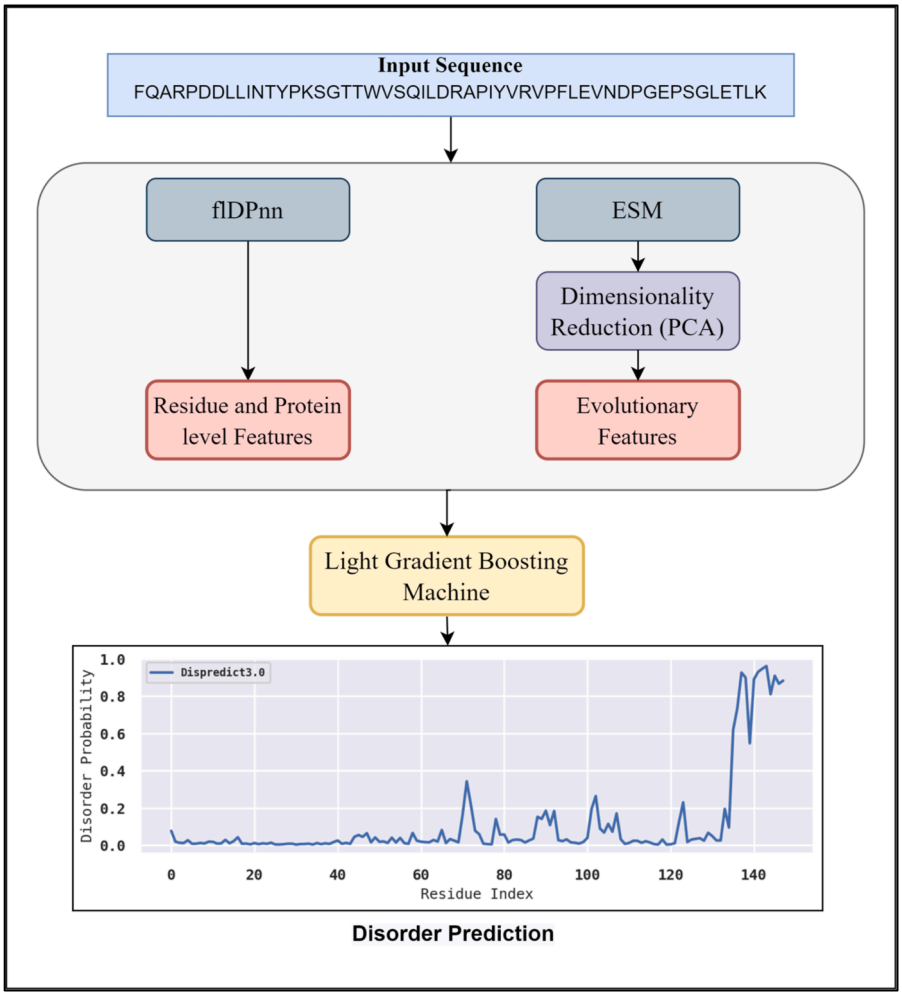


Fig. 1. The framework of the DisPredict3.0 method for disordered prediction. DisPredict3.0 collects features from flDPnn, and ESM Model. The principal component analysis is used to reduce the dimension of the ESM features, and an optimized Light Gradient Boosting Machine is trained for disorder prediction.

disordered protein predictions to measure its performance. In addition, we demonstrate case studies for three proteins with three-dimensional structures to evaluate the performance. Finally, we examine the runtime complexity of our proposed method, which is a critical aspect of real-world applications.

3.1. Prediction of intrinsic disorder

We experimented with the eight machine learning methods with default parameter settings to find the best method for disorder prediction. Supplementary Table S1 shows the 10-fold cross-validation results on the training dataset in terms of the metrics AUC, F1-score, kappa, and MCC. We chose AUC as a primary metric for selecting the best model as AUC is not dependent on the threshold. We found that, on average, the Light Gradient Boosting Machine performs the best in terms of AUC.

Table 3
Performance Metrics to evaluate the performance of disordered protein.

Metric	Definition
TP	True Positive: Correctly predicted positive samples
TN	True Negative: Correctly predicted negative samples
FP	False Positive: Incorrectly predicted positive samples
FN	False Negative: Incorrectly predicted negative samples
F1-score	$\frac{2TP}{2TP+FP+FN}$
MCC	$\frac{(TP\times TN)-(FP\times FN)}{\sqrt{(TP+FN)\times(TP+FP)\times(TN+FP)\times(TN+FN)}}$
Kappa	$\frac{2\ \times(TP\times TN-FP\times FN)}{(TP+FN)\times(TP+FP)\times(TN+FP)\times\ (TN+FN)}$

3.2. Ablation analysis

We employ an ablation analysis to determine the importance of each feature set in prediction. The outcomes presented in Supplementary Table S2 and Table 4 exhibit the results obtained by gradually increasing the feature set on the validation and test dataset, respectively. In their work on ESM [19], the authors recommended exploring the representation of each layer to identify the most suitable layer representation for a specific problem. Therefore, we initiated experiments using the representations from the first (0), last (33), and middle (15) layers of ESM. Our findings revealed that combining representations from different ESM layers enhances prediction accuracy. This discovery inspired us to incorporate all representations (dimension 34×1281) as input for DisPredict3.0 and utilize Principal Component Analysis (PCA) to reduce the dimensionality. Experimental results indicate that the reduced dimension obtained through PCA improves the AUC and F1 scores compared to utilizing a single layer both validation and test set.

3.3. Threshold optimization and hyperparameter selection

Threshold optimization plays a crucial role in the machine learning pipeline, particularly when the training dataset exhibits a significant class imbalance. The imbalanced nature of the dataset causes the trained model to display a bias towards one class. This situation motivates us to seek the ideal threshold value, considering the highly imbalanced nature of our training dataset (as shown in Table 1). To address the imbalanced order/disorder ratio, we optimize the threshold based on the ROC-AUC curve. By calculating the average of the optimal thresholds obtained from both the training (0.585) and validation datasets (0.178), we determine the best threshold to be 0.382.

Additionally, we conduct parameter optimization to identify the optimal number of trees in the Light Gradient Boosting Machine (LightGBM). We employ the grid search technique to explore different parameter values, ultimately selecting the best parameter value of n_estimators = 1000. For the remaining parameters, we utilize the default values provided by scikit-learn.

3.4. Comparison with existing methods

We conducted a thorough comparison between DisPredict3.0 and seven recently published disorder predictors. All evaluation results were obtained by executing these predictors on a Linux server. When compared to the state-of-the-art methods (flDPnn) in CAID (2018), DisPredict3.0 exhibited substantial improvements across multiple performance metrics. Specifically, our proposed method demonstrated enhancements of 2.51 % in AUC, 16.13 % in F1-score, 17.98 % in MCC, and 11.94 % in kappa. These superior results are depicted in Table 5. It is important to note that the results in Table 5 differ from those in Table 4, as they have been improved through threshold optimization. Furthermore, Fig. 2 displays the AUC, F1-score, MCC, and kappa scores achieved by the seven selected disorder

Table 4
Analysis of the effect of different feature sets created from ESM and flDPnn on theTest set.

Name	AUC (Imp%)	F1-score (Imp%)	Kappa (Imp%)	MCC (Imp%)	Number of features
Base Scores (flDPnn)	0.837	0.558	0.445	0.469	317
flDPnn features	0.834 (−0.36 %)	0.575 (3.05 %)	0.447 (0.45 %)	0.455 (−2.99 %)	317
flDPnn + Layer 0	0.831 (−0.72 %)	0.580 (3.94 %)	0.454 (2.02 %)	0.461 (−1.71 %)	1598
flDPnn + Layer 0 and 33	0.851 (1.67 %)	0.610 (9.32 %)	0.487 (9.44 %)	0.491 (4.69 %)	2879
flDPnn + Layer 33	0.852 (1.79 %)	0.615 (10.22 %)	0.495 (11.24 %)	0.501 (6.82 %)	1598
flDPnn + Layer 15	0.851 (1.67 %)	0.620 (11.11 %)	0.502 (12.81 %)	0.508 (8.32 %)	1598
flDPnn + Layer 0 and 15	0.849 (1.43 %)	0.624 (11.83 %)	0.506 (13.71 %)	0.512 (9.17 %)	2879
flDPnn + Layer 15 and 33	0.854 (2.03 %)	0.632 (13.26 %)	0.517 (16.18 %)	0.523 (11.51 %)	2879
flDPnn + Layer 0, 15 and 33	0.857 (2.39 %)	0.631 (13.08 %)	**0.518 (16.40 %)**	**0.525 (11.94 %)**	4160
flDPnn + PCA	**0.858 (2.51 %)**	**0.633 (13.53 %)**	0.517 (16.25 %)	0.522 (11.34 %)	5441

The best score values are bold-faced. Here, 'Imp' stands for improvement. The 'Imp' indicates the improvement in percentage achieved by DisPredict3.0 over the Base Scores (flDPnn) for the corresponding evaluation metric.

Table 5
Performance comparison of DisPredict3.0 with the seven disorder predictors.

Method	AUC	F1-score	Kappa	MCC
NetSurfP-2.0	0.753	0.536	0.348	0.350
IUPred3	0.762	0.584	0.407	0.412
Metapredict	0.766	0.595	0.411	0.423
Espritz	0.795	0.544	0.398	0.401
SPOT-Disorder-Single	0.802	0.596	0.448	0.448
NetSurfP-3.0	0.799	0.601	0.424	0.434
flDPnn (Base Scores)	0.837	0.558	0.445	0.469
DisPredict3.0	**0.858**	**0.648**	**0.525**	**0.525**
(Imp%)	2.51 %	16.13 %	17.98 %	11.94 %

The best score values are bold-faced. Here, 'Imp' stands for improvement. The 'Imp' indicates the improvement in percentage achieved by DisPredict3.0 over the flDPnn method for the corresponding evaluation metric.

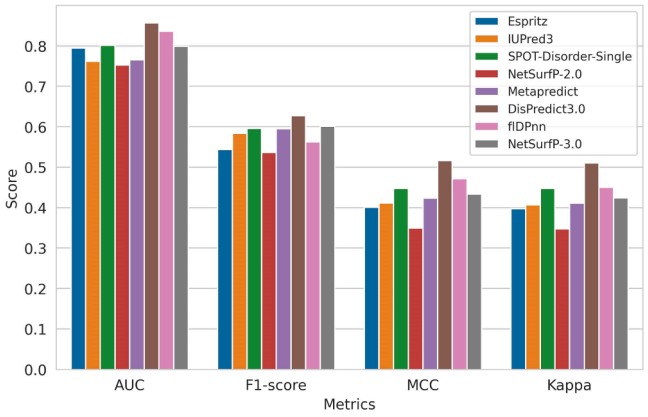


Fig. 2. Comparison of the performance of DisPredict3.0 with the seven disorder predictors on the test dataset. DisPredict3.0 achieves an AUC of 0.858 and shows better results than other methods in terms of AUC, F1-score, MCC, and kappa metric.

predictors, reaffirming the exceptional performance of DisPredict3.0.

In addition, we plotted the ROC and precision-recall curves to compare DisPredict3.0 with existing methods, as illustrated in Figs. 3 and 4. These curves demonstrate that our proposed method outperforms other methods in terms of both the ROC and precision-recall curves, further substantiating its superior performance.

At the time of writing this manuscript, we participated in the CAID-2 assessment [33] conducted in 2022. Figs. 5–7 and

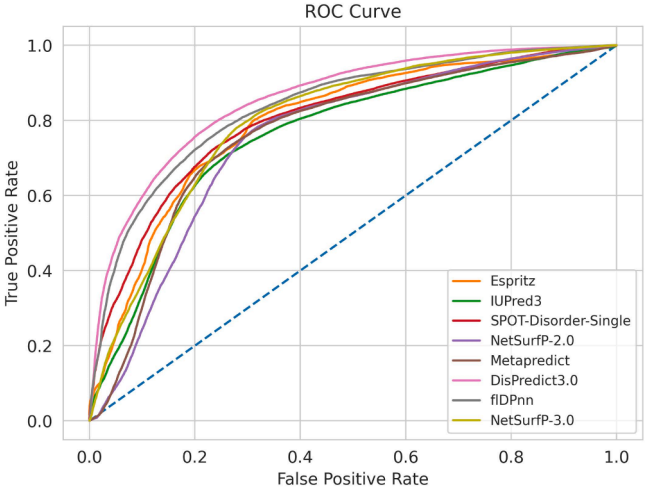


Fig. 3. The performance comparison of existing methods on the test dataset using ROC curves. These curves plot the TPR, which is the proportion of true positive cases correctly identified by the model, against the FPR, which is the proportion of false positive cases incorrectly identified as positive by the model at different thresholds.

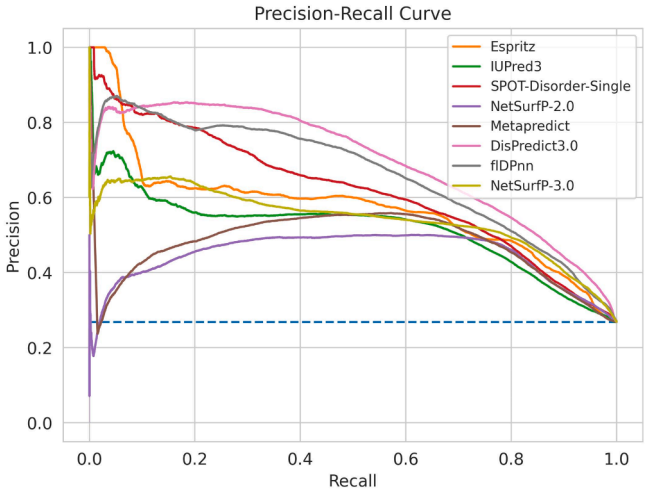


Fig. 4. The performance comparison of existing methods on the test dataset using the Precision-Recall curve. The figure plots the precision, the proportion of true positive predictions among all positive predictions, against the recall, which is the proportion of true positive predictions among all true positive instances at different thresholds.

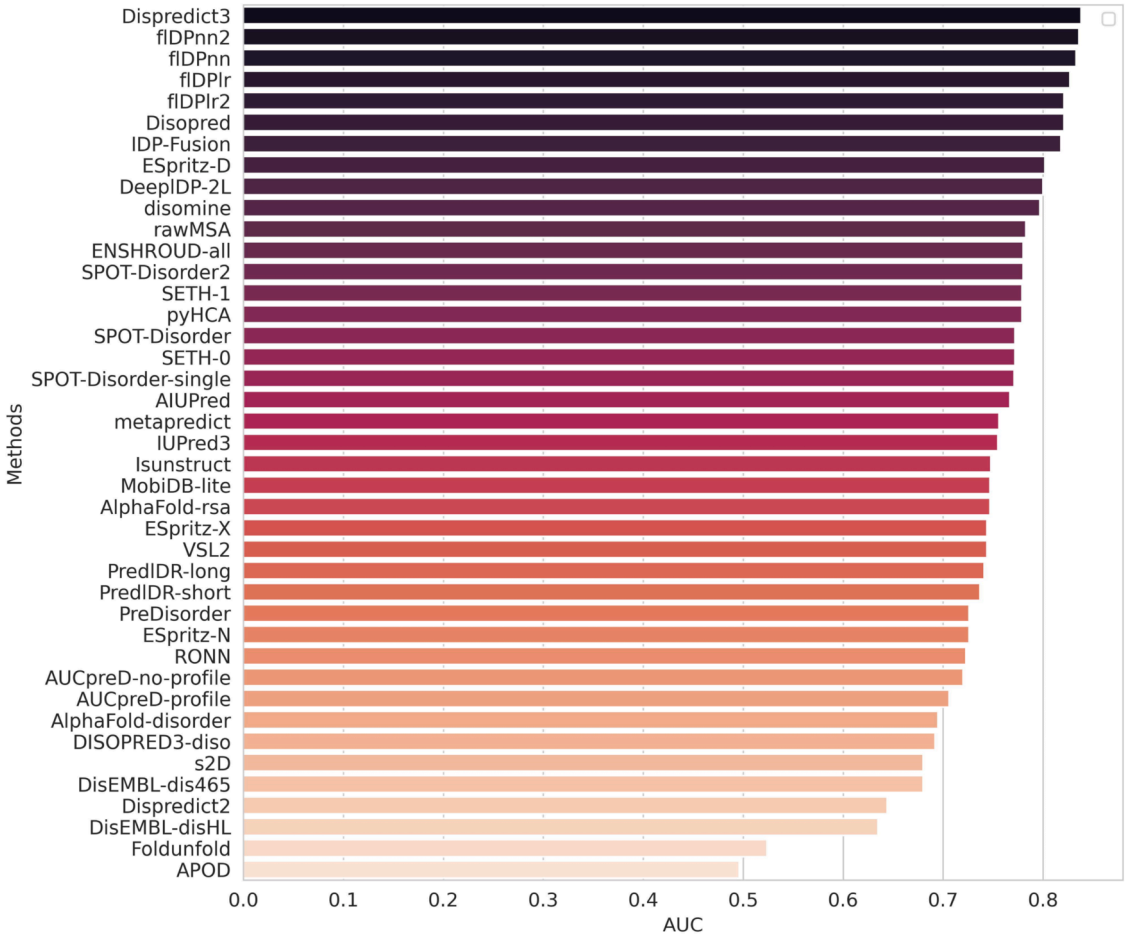


Fig. 5. Performance comparison of DisPredict3.0 on CAID2 on Disorder NOX dataset. DisPredict3.0 ranked first among 41 methods in terms of AUC.

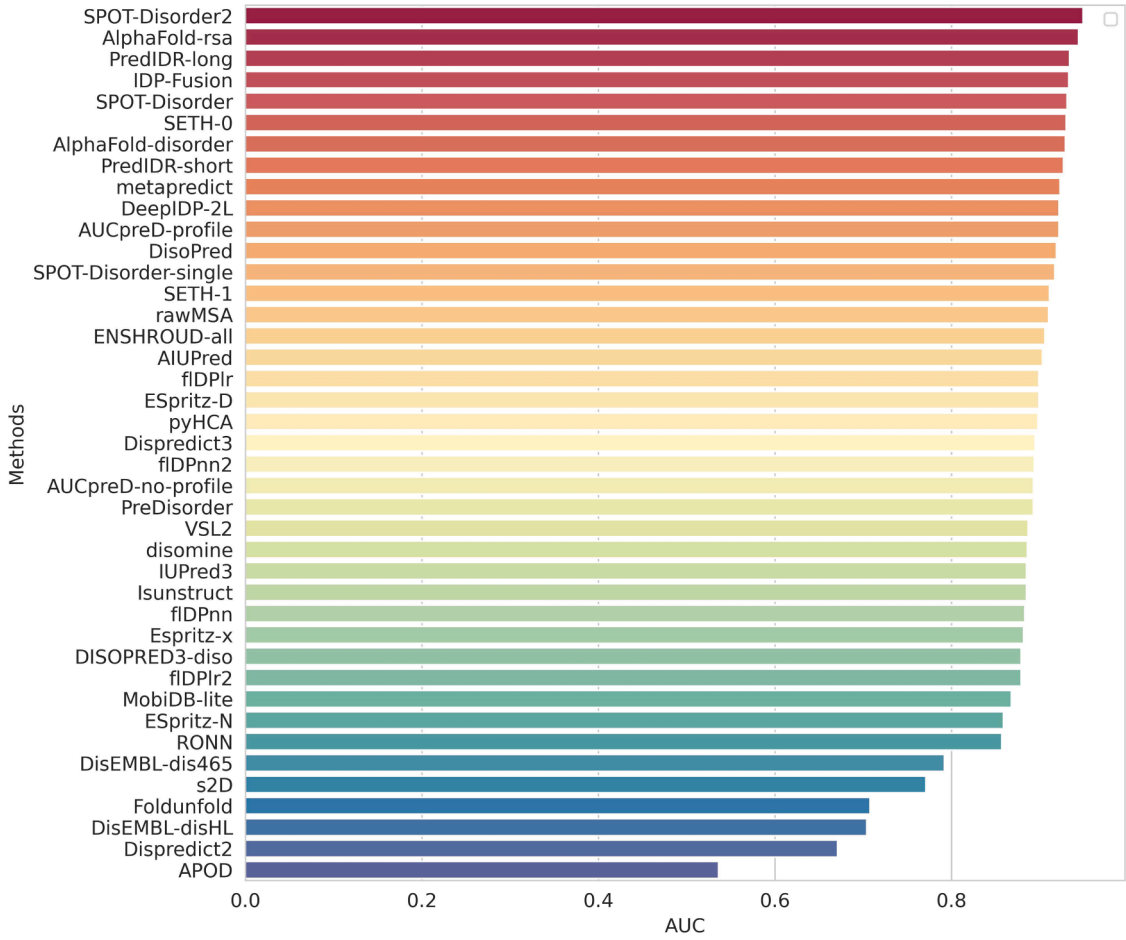


Fig. 6. Performance comparison of DisPredict3.0 on CAID2 on Disorder PDB dataset. This test dataset was created in CAID2 from experimentally derived PDB structures.

Supplementary Tables S3, S4, and S5 present the results of DisPredict3.0 in comparison to over forty existing disorder predictors [2,6,11,28,34-38]. In CAID assessment, DisPredict3.0 achieved first place on the Disorder NOX dataset (Disorder-no X-ray dataset) and sixth place for linker prediction out of 41 predictors, as indicated in Figs. 5 and 7 (Supplementary Tables S3 and S5) [33].

However, our methods did not perform as well on the Disorder-PDB dataset, as shown in Fig. 6 (Supplementary Table S4). This outcome was expected since our model was trained using annotations from the DisProt dataset rather than the experimentally derived PDB dataset. These results were gathered from the official website of CAID2 (https://caid.idpcentral.org/challenge#Benchmarking) [33]. The findings demonstrate that DisPredict3.0 possesses the capability to predict disordered regions accurately.

3.5. *Prediction of fully disordered proteins*

We also explore the prediction capabilities of DisPredict3.0 specifically for fully disordered proteins, where the disordered region encompasses at least 95 % of their sequences. Fig. 8 illustrates the performance of the proposed method in comparison to existing methods. The MCC, F1-score, and Kappa metric were calculated with an optimized threshold. While SPOT-Disorder-Single [6] and IUPred3 [1] demonstrate better performance in terms of AUC (Area Under the Curve) and MCC (Matthews Correlation Coefficient), DisPredict3.0 exhibits favorable results in terms of F1-score and Kappa.

3.6. *Case study*

We further explore how well DisPredict3.0 performs on both disordered and fully disordered proteins. To understand which regions are correctly predicted, we visualize the three-dimensional protein structures. Since most of the disordered proteins do not have an experimentally derived structure, we evaluate our method using predicted structures from AlphaFold2 [23]. Fig. 9(a) depicts the true annotated region of the Midkine b protein (DP00885) on its predicted structure. While DisPredict3.0 shows good performance, it fails to predict a small alpha helix in the middle of the protein, as illustrated in Fig. 9(b). Fig. 9(c) displays the predicted regions using the

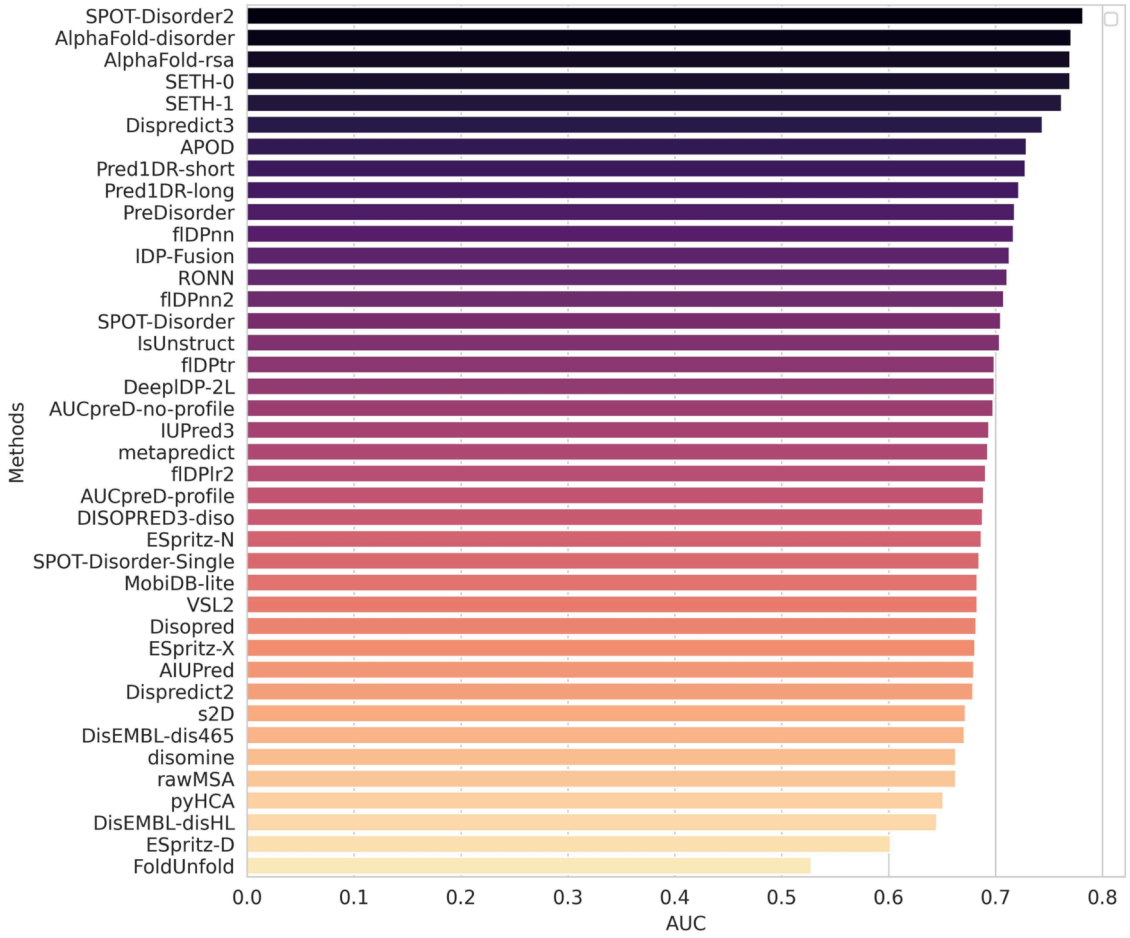


Fig. 7. Performance comparison of DisPredict3.0 on CAID2 on Linker prediction. DisPredict3.0 ranked sixth among 41 methods in terms of AUC.

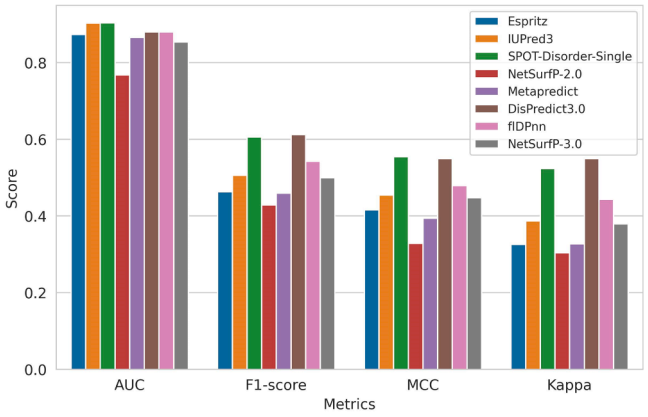
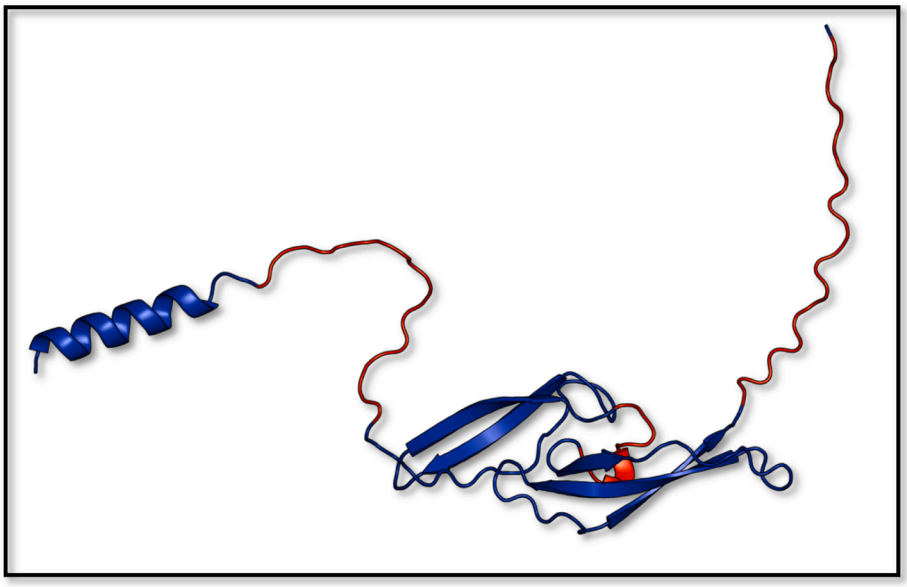


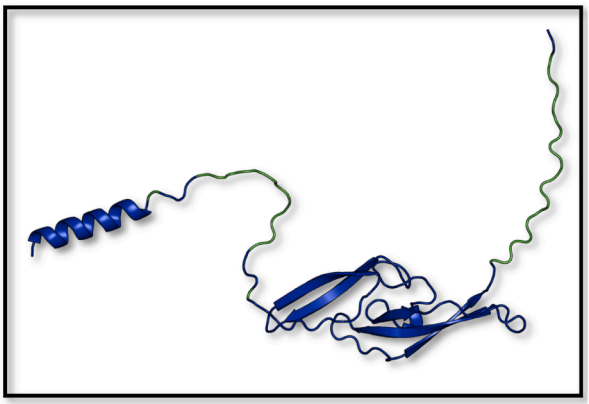
Fig. 8. Comparative Analysis of the performance of DisPredict3.0 on fully disordered proteins. The result shows that DisPredict3.0 shows promising results compared to the existing methods.

fldpnn method, which is also not able to predict the small alpha helix region. In contrast, Fig. 9(d) shows the predicted regions in terms of amino acid sequences, where DisPredict3.0 demonstrates better performance compared to the fldpnn method.

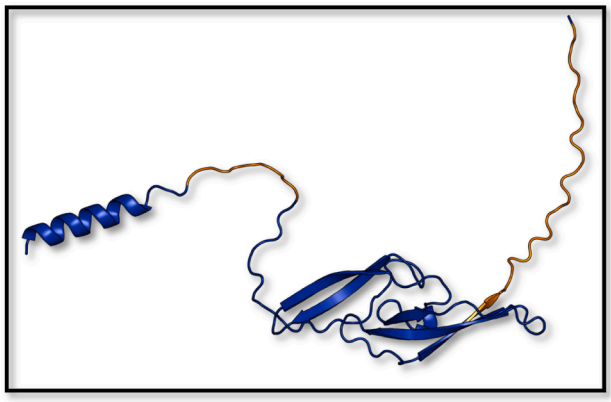
We also explore the prediction of disordered regions for Hypoxanthine phosphoribosyltransferase (DP01027) protein, as illustrated in Fig. 10. Fig. 10(a) displays the true regions annotated in DisProt. Fig. 10(b) and 10(c) present the predicted regions by the DisPredict3.0 and fldpnn methods, respectively. In this case, DisPredict3.0 successfully predicted the alpha helix region, although a few



a) True regions



b) DisPredict3.0 predicted regions



c) fldpnn predicted regions

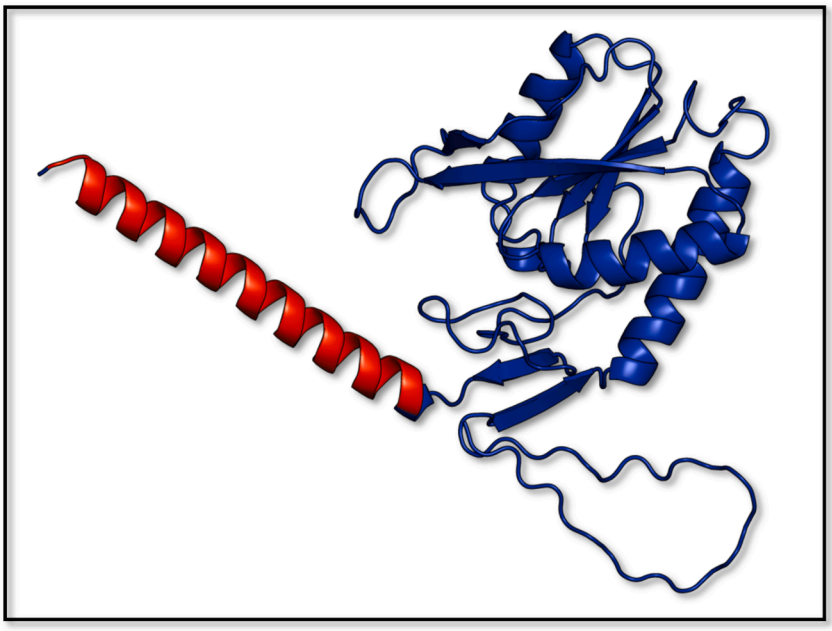
```
Protein ID : DP00885
Sequence Length : 147
True Regions :
MRSLFSIALVLVVVSMITVEASMKKKEKGKEPKADAECS...PCNWKKDFGADCKYK...NAECQTTIKVSKPCTPKTPKPKGGEKKKGKGKEN
Dispredict3.0 Predicted Regions :
MRSLFSIALVLVVVSMITVEASMKKKEKGKEPKADAECS...PCNWKKDFGADCKYK...NAECQTTIKVSKPCTPKTPKPKGGEKKKGKGKEN
fldpnn Predicted Regions :
MRSLFSIALVLVVVSMITVEASMKKKEKGKEPKADAECS...PCNWKKDFGADCKYK...NAECQTTIKVSKPCTPKTPKPKGGEKKKGKGKEN
```

(d) True and predicted regions in terms of amino acids

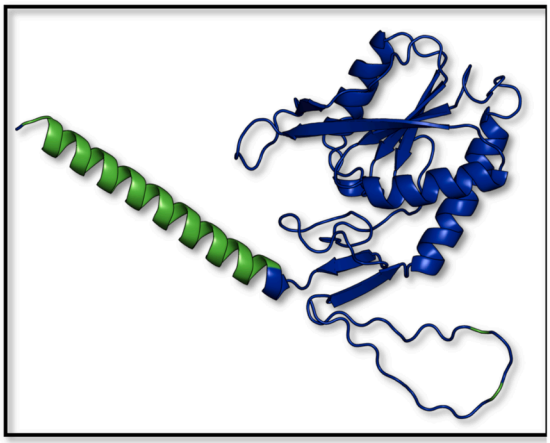
Fig. 9. The performance of DisPredict3.0 on the Midkine b (DP00885) protein. a) Displays the disordered regions of Midkine b (DP00885), as annotated in DisProt, highlighted in red. b) Illustrates the regions predicted by the DisPredict3.0 method in green. c) Shows the regions predicted using the fldpnn method in orange. d) Demonstrates the protein sequence, featuring both the true and predicted regions. The true region is highlighted in red, while the regions predicted by DisPredict3.0 are marked in green, and those predicted by fldpnn are marked in yellow.

residues were not accurately labeled, as shown in Fig. 10(b). DisPredict3.0 demonstrated superior performance compared to the fldpnn method in accurately predicting the disordered regions. The figures were generated using the open-source version of PyMOL [39].

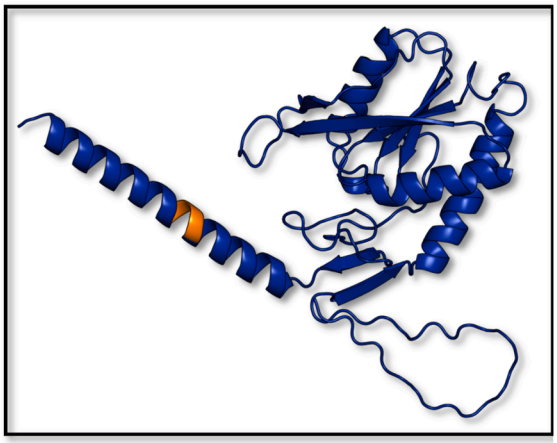
To investigate fully disordered regions, we analyzed the Seed maturation PM28 protein (DP01088), which comprises 89 residues.



a) True regions



b) DisPredict3.0 predicted regions



c) fldpnn predicted regions

```
Protein ID : DP01027
Sequence Length : 241
True Regions :
MFIRLFAREEEGEKEGKICH...SADYVVANIPNAFVIGYGLDYDDTYRELRDIVVLRPEVYAEREAARQKKQRAIGSADTDRDAKREFHSKY
Predicted Regions :
MFIRLFAREEEGEKEGKICH...SADYVVANIPNAFVIGYGLDYDDTYRELRDIVVLRPEVYAEREAARQKKQRAIGSADTDRDAKREFHSKY
fldpnn Predicted Regions :
MFIRLFAREEEGEKEGKICH...SADYVVANIPNAFVIGYGLDYDDTYRELRDIVVLRPEVYAEREAARQKKQRAIGSADTDRDAKREFHSKY
```

(d) Comparison based on the protein sequence

Fig. 10. The performance of DisPredict3.0 on disordered proteins was evaluated by comparing the observed true regions with the predicted regions in 3D protein structures. a) Displays the disordered regions of Hypoxanthine phosphoribosyltransferase (DP01027) based on DisProt annotations, highlighted in red. b) Illustrates the predicted regions from the DisPredict3.0 method in green. c) Shows the predicted regions using the fldpnn method in orange. d) Presents the protein sequence, indicating both the true and predicted regions. The true region is highlighted in red, while the predicted regions are indicated in green for DisPredict3.0 and in yellow for the fldpnn method, respectively.

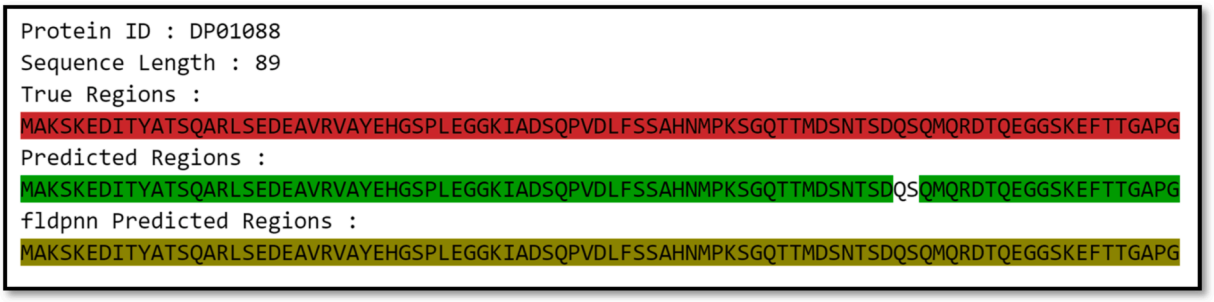


Fig. 11. The effectiveness of DisPredict3.0 on fully disordered proteins is illustrated in the figure, which showcases the true and predicted regions in terms of the amino acid sequence. The true region is highlighted in red, while the predicted regions are denoted in green for DisPredict3.0 and in yellow for fldpnn.

Fig. 11 shows that DisPredict3.0 accurately predicted the entire region, except for a small segment highlighted in white. Both DisPredict3.0 and the fldpnn method were able to predict the protein as fully disordered.

3.7. Computational time

DisPredict3.0 was trained on a Linux Machine equipped with 64 cores (32 physical cores and 32 virtual cores) and 768 GB of RAM. The training time was approximately 6 min after the features were loaded into memory, demonstrating the speed of the Light Gradient Boosting Machine algorithm. The primary time-consuming task is loading the extensive language model into memory, which takes up 7.29 GBs. However, once the model is loaded, DisPredict3.0 efficiently predicts disordered regions/proteins for the test set. During the CAID-2 evaluation, the average execution time per protein for the DisPredict3.0 method was approximately 3.3 min (estimated from the plot) when run within a Docker container [40]. Again, the predominant portion of time is spent loading the sizable language model into memory.

To further improve computational efficiency, we have implemented a parallel version of DisPredict3.0 that utilizes multiple CPUs. We wrote a script to run the DisPredict3.0 method in parallel to leverage the power of additional CPUs. Figs. 12 and 13 and Supplementary Table S6 present the runtime analysis for the parallel DisPredict3.0 method. Utilizing ten cores, DisPredict3.0 successfully predicted 176 protein sequences from the test set in less than 23 min. Fig. 13 illustrates that, on average, the parallel DisPredict3.0 can process a protein sequence in approximately 7 s by using 10 CPUs. We observed a minimal performance improvement when experimenting with more than ten cores, as indicated in the figure. This significant speedup demonstrates the efficiency of the DisPredict3.0 method in the proteome scale. Additionally, we have compared the runtime of our proposed method with existing methods in Supplementary Table S7, with the runtimes of these existing methods collected from [41]. The table demonstrates, on average (based on our test set), that the parallel version of DisPredict3.0 can make predictions in less than 0.2 min.

4. Conclusions

This study introduces DisPredict3.0, a novel disorder predictor that leverages a protein language model representation to enhance

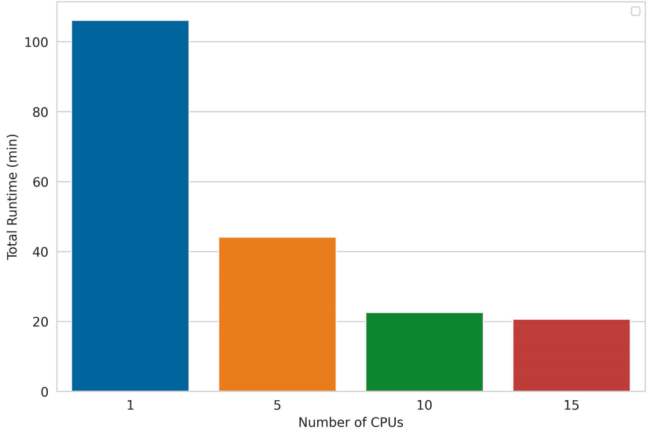


Fig. 12. Runtime analysis of parallel DisPredict3.0 with multiple CPUs. The x-axis represents the total time in minutes required to predict the protein sequences on the test dataset. The y-axis indicates the number of CPUs utilized for prediction.

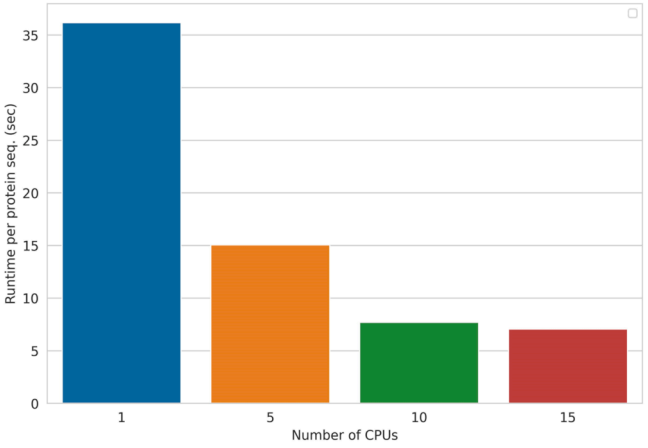


Fig. 13. Average performance of parallel runs of the DisPredict3.0 method. The x-axis shows the number of CPUs, and the y-axis shows the average time per protein sequence in seconds.

disorder prediction performance. DisPredict3.0 utilizes Principal Component Analysis (PCA) to reduce the dimensionality of the language model representation and train an optimized Light Gradient Boosting Machine for disorder prediction. Experimental results demonstrate that DisPredict3.0 outperforms existing methods for disorder prediction, with comparable performance for fully disordered regions. Despite its reliance on a large protein language model, the parallel version of DisPredict3.0 is computationally fast and efficient. Utilizing recent advancements in protein structure prediction can improve the performance of disorder prediction by leveraging predicted protein structures to aid in identifying disordered regions. Additionally, fine-tuning the language model can further enhance disorder prediction performance. The CAID-2 evaluation demonstrates that DisPredict3.0 performs very well compared to other recent methods. The DisPredict3.0 method is publicly accessible on GitHub and can be run using Docker/Singularity for reproducible results. To further facilitate user accessibility, we have also developed a web server. We believe that this method will assist researchers in identifying disordered regions/proteins, comprehending the role of disordered regions/proteins in living cells, and developing drugs for critical diseases.

Data availability

The DisPredict3.0 webserver is available at https://bmll.cs.uno.edu. The code and data related to the development of Dispredict3.0 can be found here https://github.com/wasicse/Dispredict3.0.

Acknowledgments

The authors would like to thank Austin Schmidt for thoroughly reviewing the manuscript. The research reported in the paper was partially supported by an Institutional Development Award (IDeA) from the National Institute of General Medical Sciences of the National Institutes of Health under grant number P2O GM103424-21.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.amc.2024.128630.

References

[1] G. Erdős, M. Pajkos, Z. Dosztányi, IUPred3: prediction of protein disorder enhanced with unambiguous experimental annotation and visualization of evolutionary conservation, Nucleic Acids Res. 49 (W1) (2021) W297–W303.
[2] G. Hu, et al., flDPnn: accurate intrinsic disorder prediction with putative propensities of disorder functions, Nat. Commun. 12 (1) (2021) 4438.
[3] F. Quaglia, et al., DisProt in 2022: improved quality and accessibility of protein intrinsic disorder annotation, Nucleic Acids Res. 50 (D1) (2022) D480–D487.
[4] A. Del Conte, et al., CAID prediction portal: a comprehensive service for predicting intrinsic disorder and binding regions in proteins, Nucleic. Acids. Res. 51 (W1) (2023) W62–W69.
[5] R.J. Emenecker, D. Griffith, A.S. Holehouse, Metapredict: a fast, accurate, and easy-to-use predictor of consensus disorder and structure, Biophys. J. 120 (20) (2021) 4312–4319.
[6] J. Hanson, et al., SPOT-Disorder2: improved protein intrinsic disorder prediction by ensembled deep learning, Genomics. Proteomics. Bioinformatics. 17 (6) (2019) 645–656.
[7] S. Iqbal, M.T. Hoque, DisPredict: a predictor of disordered protein using optimized RBF Kernel, PLoS. One 10 (10) (2015) e0141551.
[8] E. Melamud, J. Moult, Evaluation of disorder predictions in CASP5, Proteins 53 (S6) (2003) 561–565.
[9] C. Predictors, et al., Critical assessment of protein intrinsic disorder prediction, Nat. Methods 18 (5) (2021) 472–481.

[10] S. Wang, J. Ma, J. Xu, AUCpreD: proteome-level protein disorder prediction by AUC-maximized deep convolutional neural fields, Bioinformatics. 32 (17) (2016) i672–i679.
[11] I. Walsh, et al., ESpritz: accurate and fast prediction of protein disorder, Bioinformatics. 28 (4) (2012) 503–509.
[12] C. Mirabello, B. Wallner, rawMSA: end-to-end deep learning using raw multiple sequence alignments, PLoS. One 14 (8) (2019) e0220182.
[13] J. Hanson, K. Paliwal, Y. Zhou, Accurate single-sequence prediction of protein intrinsic disorder by an ensemble of deep recurrent and convolutional architectures, J. Chem. Inf. Model. 58 (11) (2018) 2369–2376.
[14] M.S. Klausen, et al., NetSurfP-2.0: improved prediction of protein structural features by integrated deep learning, Proteins 87 (6) (2019) 520–527.
[15] M.H. Høie, et al., NetSurfP-3.0: accurate and fast prediction of protein structural features by protein language models and deep learning, Nucleic Acids Res. 50 (W1) (2022) W510–W515.
[16] Z. Dosztányi, Prediction of protein disorder based on IUPred: prediction of Protein disorder based on IUPred, Protein Sci. 27 (1) (2018) 331–340.
[17] B. Mészáros, G. Erdős, Z. Dosztányi, IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding, Nucleic Acids Res. 46 (W1) (2018) W329–W337.
[18] N. Ferruz, B. Höcker, Controllable protein design with language models, Nat. Mach. Intell. 4 (6) (2022) 521–532.
[19] A. Rives, et al., Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences, Proceed. Nat. Acad. Sci. 118 (15) (2021) e2016239118.
[20] Z. Lin, et al., Evolutionary-scale prediction of atomic-level protein structure with a language model, Science (1979) 379 (6637) (2023) 1123–1130.
[21] A. Elnaggar, et al., ProtTrans: Towards cracking the language of life's code through self-supervised learning, Cold Spring Harbor Laboratory, 2020.
[22] P. Notin, et al., Tranception: protein fitness prediction with autoregressive transformers and inference-time retrieval, ArXiv. (2022).
[23] J. Jumper, et al., Highly accurate protein structure prediction with AlphaFold, Nature 596 (7873) (2021) 583–589.
[24] Z. Lin, et al., Language Models of Protein Sequences At the Scale of Evolution Enable Accurate Structure Prediction, Cold Spring Harbor Laboratory, 2022.
[25] R. Wu, et al., High-resolution de novostructure prediction from primary sequence, bioRxiv. 07 (21) (2022) 500999, 2022.
[26] A. Mishra, M.W.U. Kabir, M.T. Hoque, diSBPred: a machine learning based approach for disulfide bond prediction, Comput. Biol. Chem. 91 (2021) 107436.
[27] A. Mishra, et al., AIRBP: accurate identification of RNA-binding proteins using machine learning techniques, Artif. Intell. Med. 113 (2021) 102034.
[28] S. Iqbal, M.T. Hoque, Estimation of position specific energy as a feature of protein residues from sequence alone for structural classification, PLoS. One 11 (9) (2016) e0161452.
[29] D. Piovesan, et al., DisProt 7.0: a major update of the database of disordered proteins, Nucleic Acids Res. 45 (D1) (2017) D219–D227.
[30] D.W.A. Buchan, D.T. Jones, The PSIPRED protein analysis workbench: 20 years on, Nucleic Acids Res. 47 (W1) (2019) W402–W407.
[31] S.F. Altschul, et al., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25 (17) (1997) 3389–3402.
[32] G. Ke, et al., LightGBM: a highly efficient gradient boosting decision tree, in: in Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, California, USA, Curran Associates Inc., 2017, pp. 3149–3157.
[33] A. Del Conte, et al., CAID prediction portal: a comprehensive service for predicting intrinsic disorder and binding regions in proteins, Nucleic Acids Res. (2023).
[34] D. Piovesan, A.M. Monzon, S.C.E. Tosatto, Intrinsic protein disorder and conditional folding in AlphaFoldDB, Protein Sci. 31 (11) (2022) e4466.
[35] O.V. Galzitskaya, S.O. Garbuzynskiy, M.Y. Lobanov, FoldUnfold: web server for the prediction of disordered regions in protein chain, Bioinformatics. 22 (23) (2006) 2948–2949.
[36] D. Ilzhöfer, M. Heinzinger, B. Rost, SETH predicts nuances of residue disorder from protein embeddings, Front. Bioinform. 2 (2022).
[37] M.Y. Lobanov, I.V. Sokolovskiy, O.V. Galzitskaya, IsUnstruct: prediction of the residue status to be ordered or disordered in the protein chain by a method based on the Ising model, J. Biomol. Struct. Dyn. 31 (10) (2013) 1034–1043.
[38] Z. Peng, Q. Xing, L. Kurgan, APOD: accurate sequence-based predictor of disordered flexible linkers, Bioinformatics. 36 (Suppl_2) (2020) i754–i761.
[39] Schrodinger, L., The PyMOL molecular graphics system. 2010.
[40] Disorder in CAID-2, in CASP15. 2022: Antalya, Turkey.
[41] B. Zhao, L. Kurgan, Deep learning in prediction of intrinsic disorder in proteins, Comput. Struct. Biotechnol. J. 20 (2022) 1286–1294.