THE UNIVERSITY of NEW ORLEANS

Bioinformatics & Machine Learning Lab

Intrinsically Disordered Protein Prediction using Pretrained Language Model

Md Wasi Ul Kabir (mkabir3@uno.edu), Marco Tabora (matabora@uno.edu), Md Tamjidul Hoque (thoque@uno.edu)

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

Introduction

- Traditionally, proteins have been characterized as having a fixed structure. However, many proteins contain disordered regions (also called Intrinsically Disordered Regions/Proteins) that lack a stable structure. They play significant biological roles despite their structural instability.
- Identifying these regions is an important challenge in modern bioinformatics, and machine learning techniques have proven an efficient way to tackle this problem via a computational approach.

Figure 1. Two states of a disordered protein

*Image By Lukasz Kozlowski , https://commons.wikimedia.org/w/index.php?curid=36298697

Motivation

- Disordered Proteins differ considerably from ordered proteins and can lead to a variety of severe illnesses.
- Identifying Disordered Regions/Proteins is a challenging and time-consuming process requiring specialized experimental analysis and identification tools.
- Disordered proteins have a wide range of applications in biology and medicine, such as developing new drugs and studying the mechanisms of diseases. This motivates us to develop a computational approach for disordered prediction.

Dataset

- We collected the training, test and validation set from the state-of-the-art (fIDPnn) method [1]. Table 1 shows the number of disordered and ordered residues in the training, test and validation set.
- We have removed sequences with unknown amino acid (X-tag) since they do not have specific physicochemical properties to get corresponding features in our methodology.

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

Disordered/Ordered	Train	Test	Validation
No. of Disordered residues	50387	17871	25004
No. of Ordered residues	169565	48675	4967
Total No. of Residues	219952	66546	29971

Feature Extraction

ESM Features

- We extracted features from Evolutionary Scale Modeling (ESM) [3], a protein language models developed by the Facebook Research team. ESM is a Transformer Model that is trained on 250 million protein sequences and 86 billion amino acids. The resulting model contains information about biological properties in its representations.

fIDPnn Features

- We collected 317 features from the fIDPnn tool. Some of these features include position specific scoring matrix(PSSM), with generation from PSI-BLAST, and disordered linker prediction via DFLpred.

Ankh Features

- Ankh [4], also a deep learning-based protein language model, and requires fewer parameters than other models. Ankh was developed using Google's TPU-V4 GPUs.

Experimental Setup

Sliding window technique

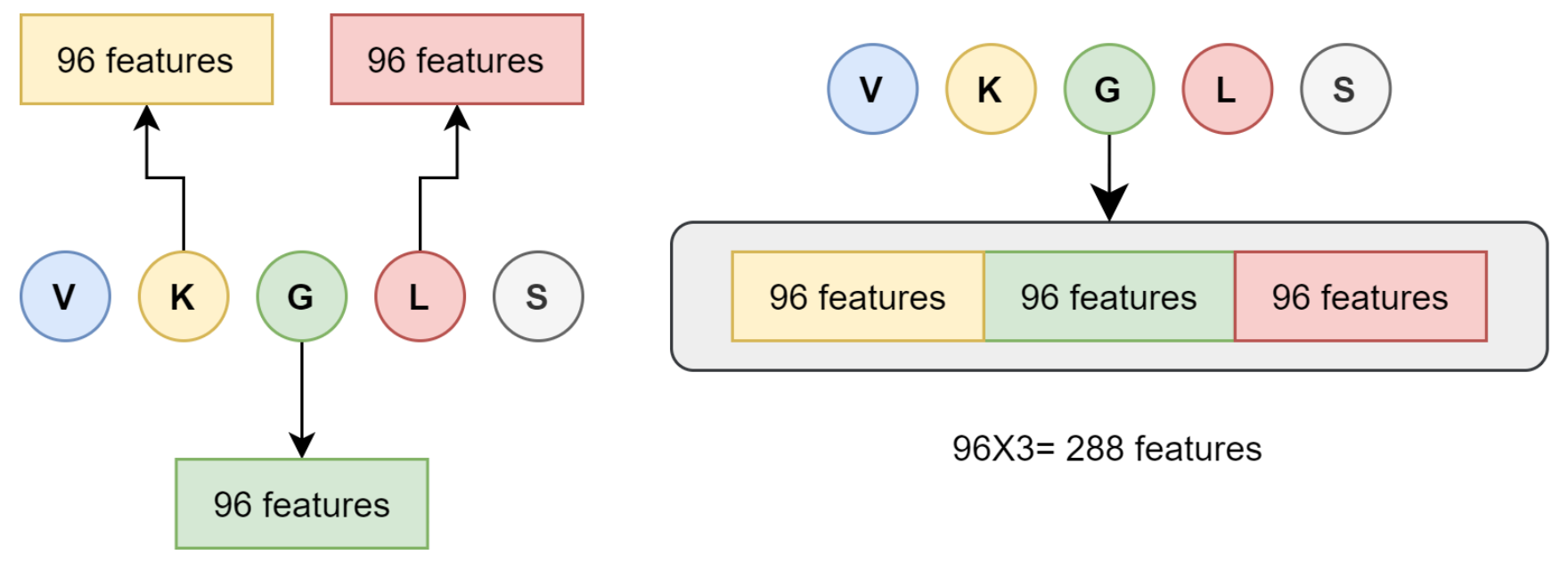


Figure 2. Illustration of sliding window technique to incorporate neighbor residues information. After feature selection, each residue is represented by 96 features. For sliding window size 3, the residues glycine(G) can be represented by concatenating the features from two of its neighbor's residues, lysine(K) and leucine(L), and the feature vector length is 96x3= 288 features.

Feature Extraction

We have extracted residue and protein level features from fIDPnn [1] tool and language model features from the protein language model (ESM) [3].

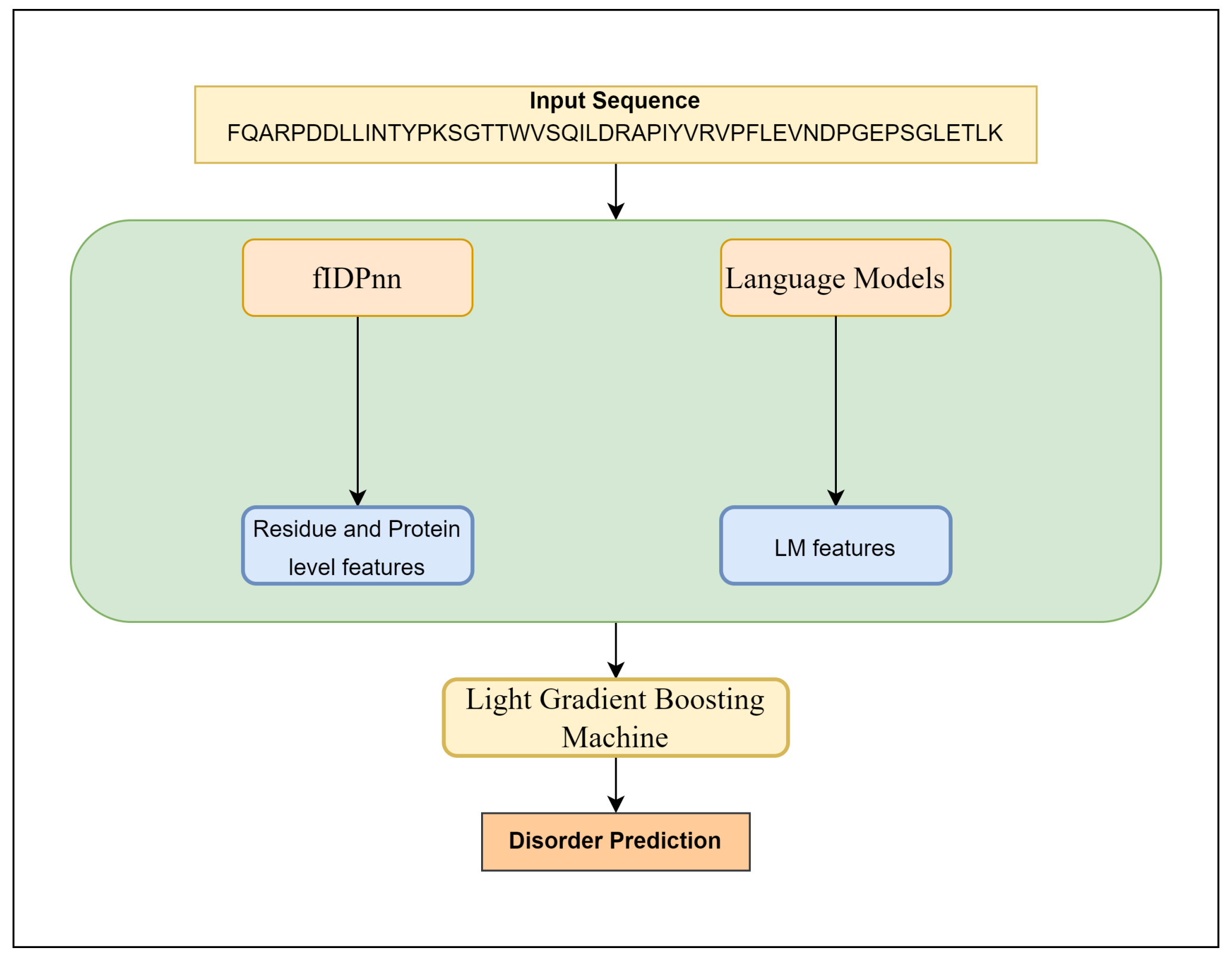
Table 2. The number of features extracted from different tools.

Methods	No. of Features
fIDPnn	317
ESM2_15B	5120
ESM2_3B	2560
ESM2_650M	1280
ESM2_150M	640
ESM2_35M	480
ESM2_8M	320
ESM-1b_650M	1280

Table 3. The number layers, parameters and embedding dimensions of ESM models.

Name	#layers	#params	Embedding Dim
ESM2	48	15B	5120
ESM2	36	3B	2560
ESM2	33	650M	1280
ESM2	30	150M	640
ESM2	12	35M	480
ESM2	6	8M	320
ESM-1b	33	650M	1280

Proposed Method



Input Sequence
FQARPDDLLINTYPKSGTTWVSQILDRAPIYVRVPFLEVNDPGEPSGLETLK

fIDPnn
Language Models
Residue and Protein level features
LM features
Light Gradient Boosting Machine
Disorder Prediction

Figure 3. The framework of the proposed method for disordered prediction. The proposed method collects features from fIDPnn, and Protein Language Models and trains a Light Gradient Boosting Machine for disorder prediction.

Performance Metrics

The following metrics are used to evaluate the predictive performance of machine learning methods along with the widely used *ROCAUC* metric.

Table 4. Performance Metrics

Name of Metric	Definition
True Positive (TP)	Correctly predicted positive samples
True Negative (TN)	Correctly predicted negative samples
False Positive (FP)	Incorrectly predicted positive samples
False Negative (FN)	Incorrectly predicted negative samples
F1-score (Harmonic mean of precision and recall)	$\frac{2TP}{2TP + FP + FN}$
Mathews Correlation Coefficient (MCC)	$\frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FN) \times (TP + FP) \times (TN + FP) \times (TN + FN}}$
Kappa	$\frac{2 \times (TP \times TN - FP \times FN)}{(TP + FN) \times (TP + FP) \times (TN + FP) \times (TN + FN)}$

Experimental Results

Table 5. Comparison of machine learning methods' prediction results in 10-fold cross-validation on the training dataset using ESM-1b model.

Model	AUC	F1-score	Kappa	MCC
Light Gradient Boosting Machine	**0.982**	0.901	0.864	**0.867**
Decision Tree Classifier	0.933	**0.905**	**0.865**	0.865
Extra Trees Classifier	0.967	0.840	0.785	0.798
Gradient Boosting Classifier	0.936	0.798	0.727	0.736
Ridge Classifier	0.000	0.756	0.665	0.669
Linear Discriminant Analysis	0.917	0.760	0.668	0.670
Logistic Regression	0.917	0.756	0.660	0.661
Ada Boost Classifier	0.900	0.721	0.619	0.624
SVM - Linear Kernel	0.000	0.666	0.505	0.536
K Neighbors Classifier	0.801	0.619	0.455	0.455

Best score values are **bold-faced.**

Table 6. Analysis of the effect of feature sets on the test dataset.

Models	AUC	F1-score	Kappa	MCC
fIDPnn	0.837	0.558	0.445	0.469
fIDPnn+ESM2_650M	**0.843**	0.630	**0.500**	**0.501**
fIDPnn+AnkhLarge	0.841	0.630	**0.500**	**0.501**
fIDPnn+ESM2_3B	0.840	**0.639**	0.496	0.497
fIDPnn+AnkhBase	0.841	0.636	0.481	0.487

Best score values are **bold-faced.**

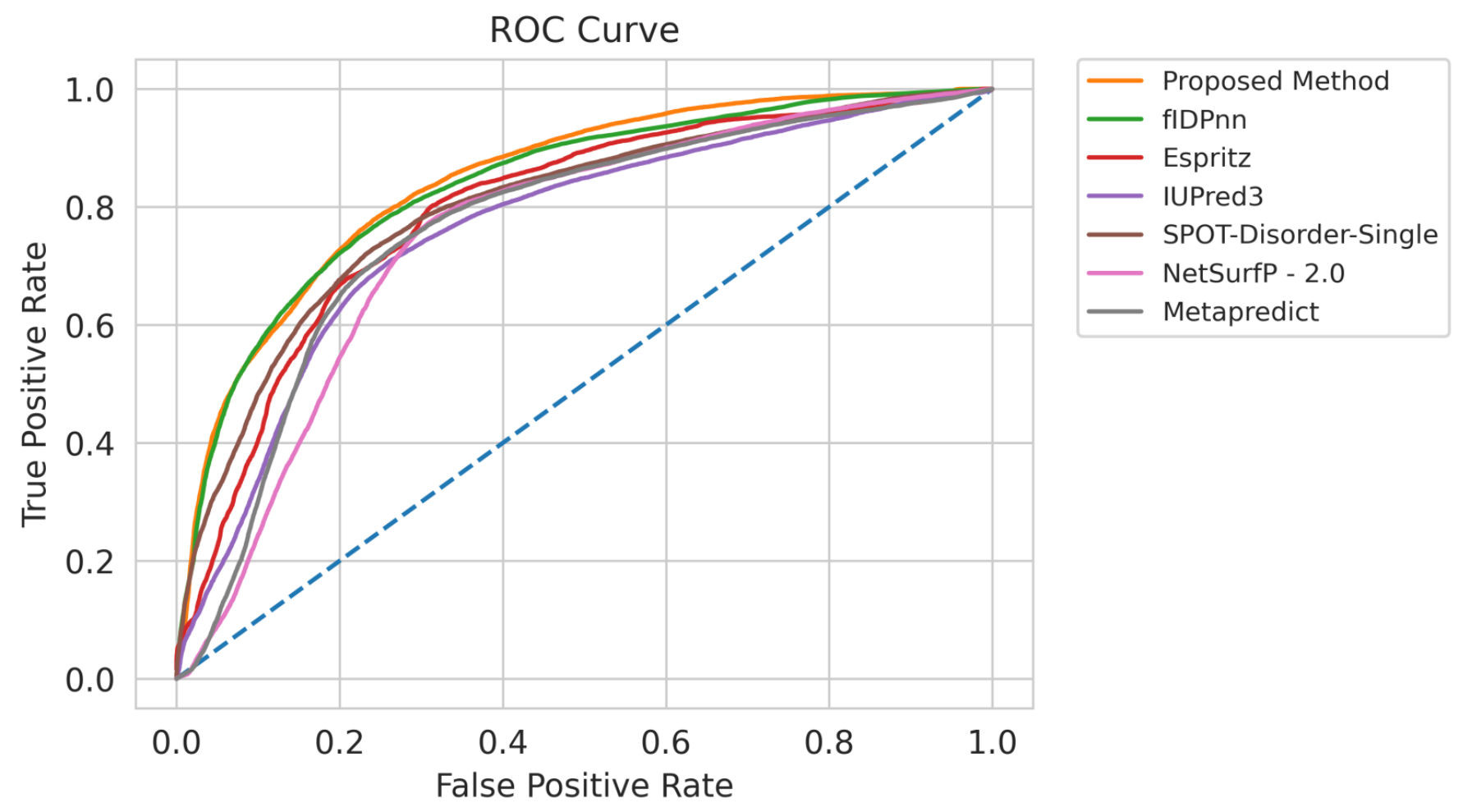


Figure 4. Comparison of proposed method with the six disordered predictors on the test dataset in terms of ROC curve.

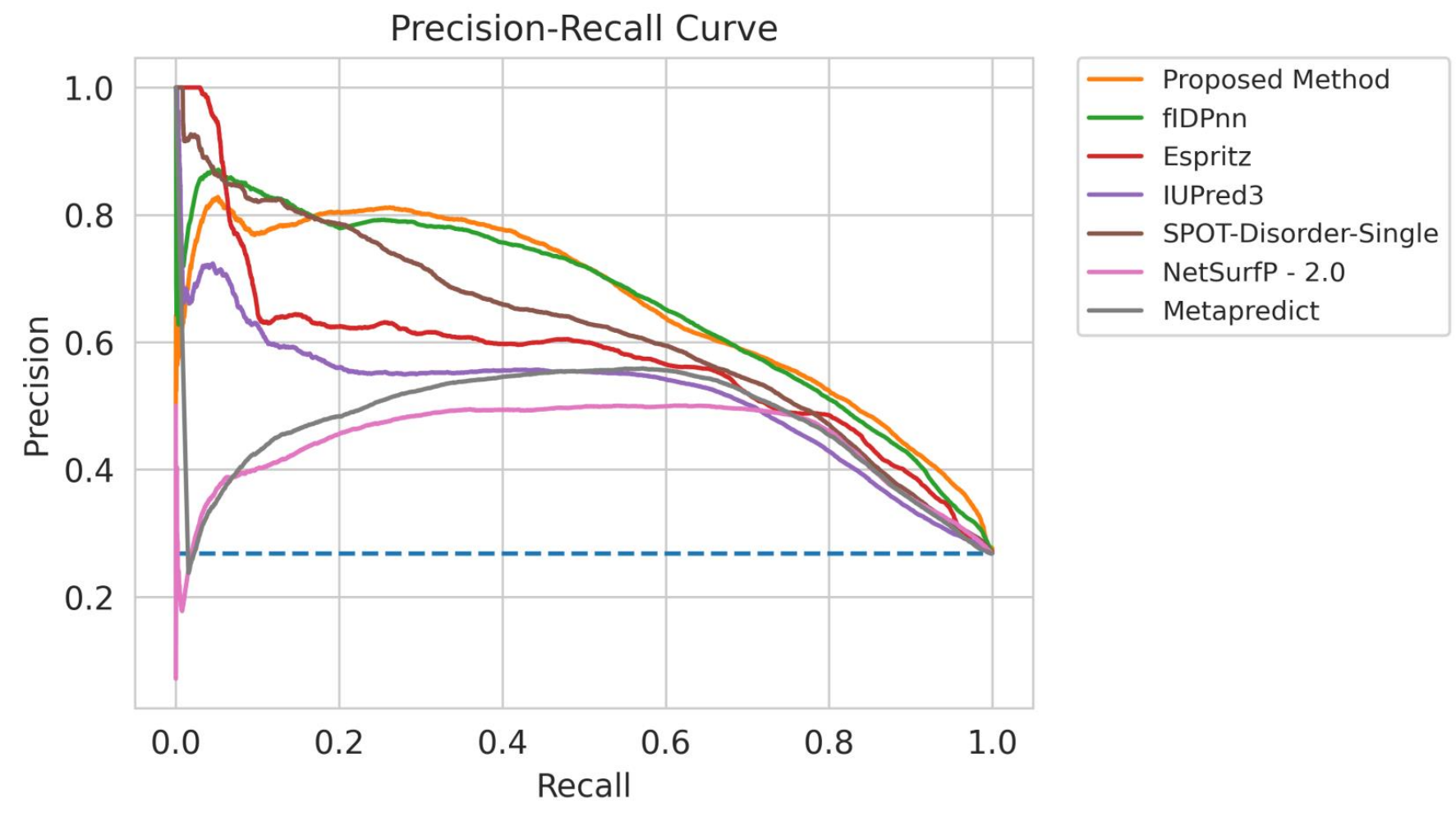
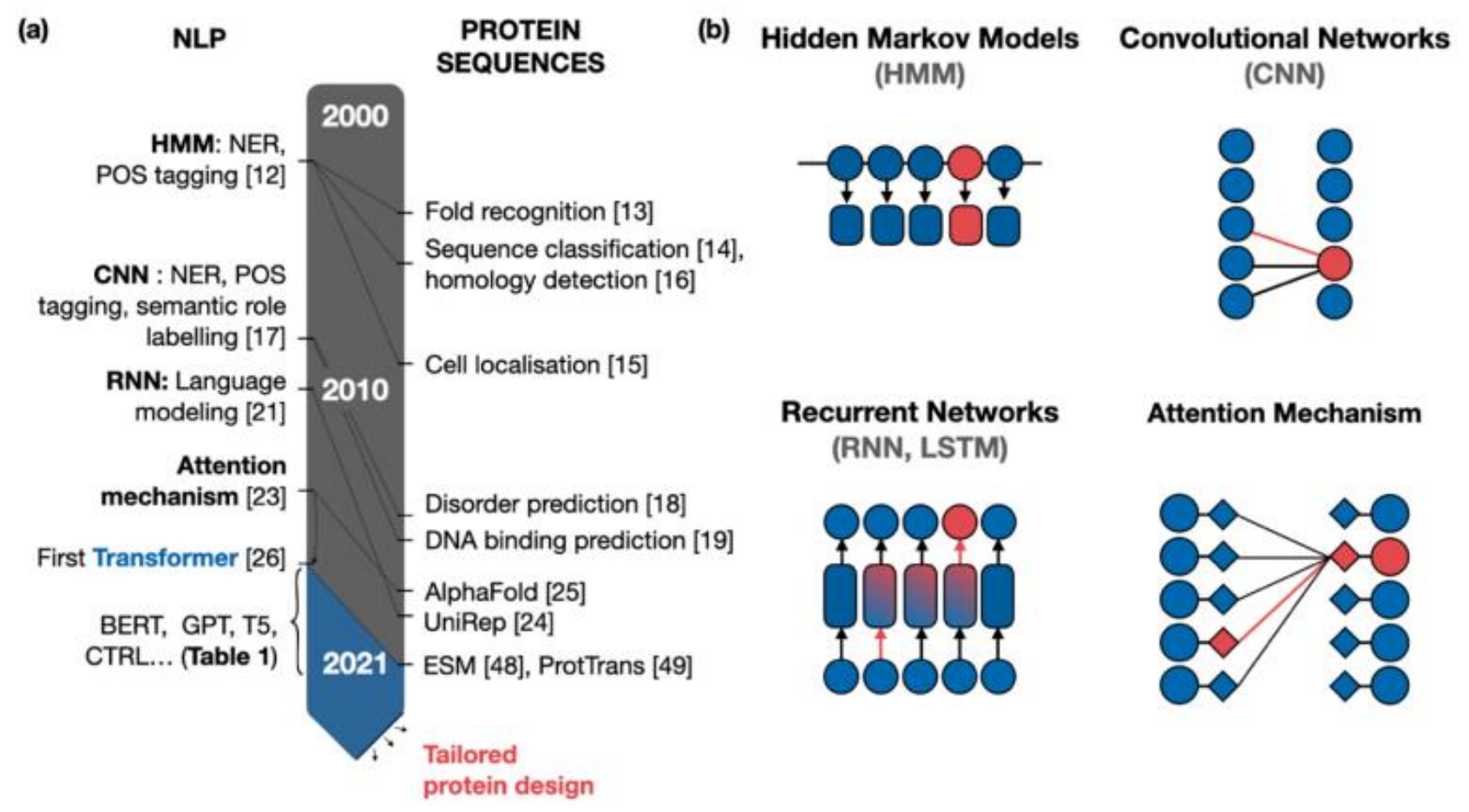


Figure 5. Comparison of proposed method with the six disordered predictors on the test dataset in terms of Precision-Recall curve.

Protein Language Models

- Protein sequences are very similar to natural languages. Protein representation learning methods are called protein language models.
  - Evolutionary Scale Modeling (ESM) [3]
  - Ankh [4]
  - ProteinBERT [5]
  - TAPE-Transformer [6]
  - ProtTrans [7]
- Recently, Transformer-based models such as ESM and Ankh have performed well for protein prediction features, so we have selected them as the main features to use within the experiment.
- One reason for the success of transformer-based models is their ability to capture long-range dependencies in the amino acid sequences. These models can represent these interactions by attending to distant parts of the sequence, allowing them to better capture the underlying patterns in the data.



(a)
NLP
PROTEIN SEQUENCES
2000
2010
2021
HMM: NER, POS tagging [12]
CNN : NER, POS tagging, semantic role labelling [17]
RNN: Language modeling [21]
Attention mechanism [23]
First Transformer [26]
BERT, GPT, T5, CTRL... (Table 1)
Fold recognition [13]
Sequence classification [14], homology detection [16]
Cell localisation [15]
Disorder prediction [18]
DNA binding prediction [19]
AlphaFold [25]
UniRep [24]
ESM [48], ProtTrans [49]
Tailored protein design
(b)
Hidden Markov Models (HMM)
Convolutional Networks (CNN)
Recurrent Networks (RNN, LSTM)
Attention Mechanism

Figure 6. NLP methods and their application in protein research [2].

Conclusion and Future Plans

- In this study, we presented a disordered protein predictor that uses the representation from a protein language model and initial results show that it helps improve disordered protein prediction performance.
- We are experimenting with the latest Protein Language model, which has 8 million to 15 billion parameters.
- So far, we could only experiment with four pre-trained models. In the future, we plan to see the performance of other pre-trained language models.
- We also plan to evaluate the performance of other machine learning methods, including deep neural networks (LSTM, Transformers).

References

1. G. Hu *et al.*, "flDPnn: Accurate intrinsic disorder prediction with putative propensities of disorder functions," *Nat Commun*, vol. 12, no. 1, p. 4438, Dec. 2021, doi: 10.1038/s41467-021-24773-7
2. Ferruz, Noelia, and Birte Höcker. "Controllable Protein Design with Language Models." *Nature Machine Intelligence*, vol. 4, no. 6, June 2022, pp. 521–32. *arXiv.org*
3. A. Rives et al., "Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences," Proc Natl Acad Sci USA, vol. 118, no. 15, Art. no. 15, Apr. 2021, doi: 10.1073/pnas.2016239118.
4. Elnaggar, Ahmed, et al. *Ankh: Optimized Protein Language Model Unlocks General-Purpose Modelling*. arXiv, 16 Jan. 2023. *arXiv.org*, https://doi.org/10.48550/arXiv.2301.06568.
5. Brandes, Nadav, et al. "ProteinBERT: A Universal Deep-Learning Model of Protein Sequence and Function." Bioinformatics, edited by Pier Luigi Martelli, vol. 38, no. 8, Apr. 2022, pp. 2102–10.
6. Rao, Roshan, et al. Evaluating Protein Transfer Learning with TAPE. arXiv, 19 June 2019. arXiv.org
7. Elnaggar, Ahmed, et al. ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Deep Learning and High Performance Computing. arXiv, 4 May 2021

test