Binding Residue Prediction in Disordered Proteins Using Protein Language Models

Farzeen Nafees (fnafees@uno.edu), Md Wasi Ul Kabir (mkabir3@uno.edu), Md Tamjidul Hoque (thoque@uno.edu)
Department of Computer Science, University of New Orleans, New Orleans, LA, USA

Introduction

Background

- **Intrinsically Disordered Regions (IDRs)** refer to regions of protein sequences that do not maintain a consistent 3D structure under varying conditions.
- A **binding site** in a protein sequence refers to a specific pocket of a protein that interacts with other molecules.
- **ESM-2** is a transformer-based **Protein Language Model** that generates rich embeddings for protein sequences.
- **ESMDisPred:** Our lab's top ranked predictor for protein disorder.

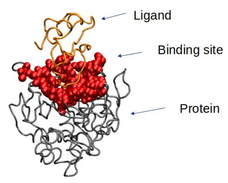

Figure 1. Schematic illustration of a protein binding site showing the specific pocket region where molecular interactions occur between the protein and potential ligand molecules.

Motivation and Objective

- Understanding where binding occurs is crucial for understanding signaling proteins, transcription factors, etc.
- Traditional computational methods for locating binding sites focus on **rigid binding** rather than on binding within disordered proteins, which are difficult to analyze.
- Data analysis revealed that in our dataset of disordered proteins, over 90% of amino acids in binding regions were also classified as disordered.

Proposed Solution

- We developed a novel disorder-augmented approach that explicitly incorporates disorder probabilities into binding site prediction.
- We leverage our top-ranked ESMDisPred model to generate disorder probability features for each amino acid residue.
- We integrate these disorder probabilities with ESM2 protein language model embeddings, creating a comprehensive feature set that captures both sequence context and structural flexibility (Figure 3).
- By directly encoding disorder information, our models can effectively predict binding sites in the challenging disordered regions where over 90% of binding residues are located.

Dataset

- We obtained protein sequence data for 2,020 proteins from DisProt and filtered out proteins with incomplete data, including those already in the test set.
- We identified and removed amino acid residues marked as "missing" in PDB from our sequences.
- Existing methods such as fldpnn and Evolutionary Scale Modeling (ESM2) were then used to extract additional features from amino acids.
- Our test set was collected from Critical Assessment of Intrinsic Disorder (CAID).

Dataset

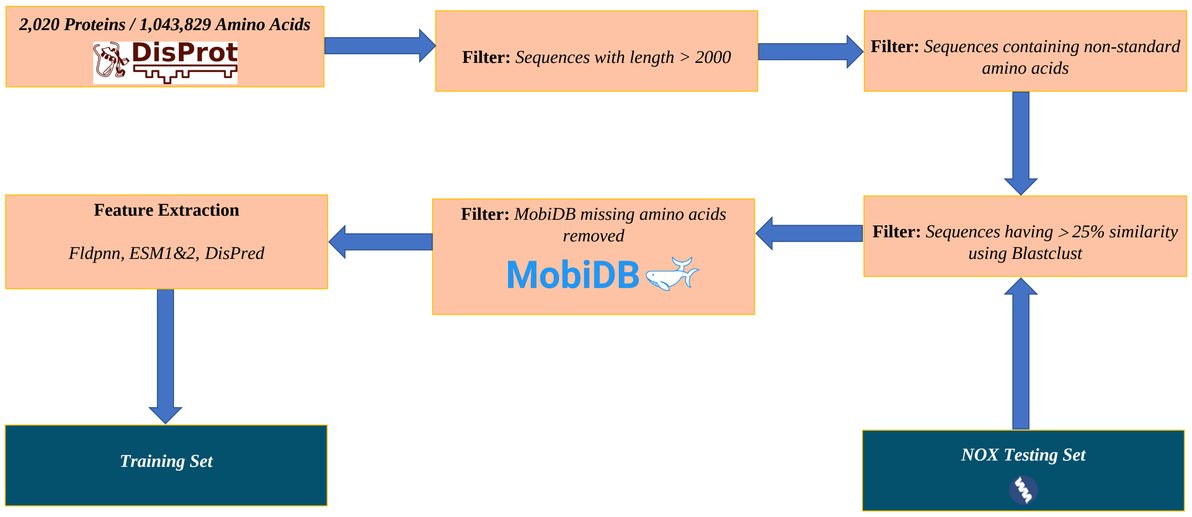

Figure 2. Multi-stage data processing workflow illustrating the curation of the training dataset from DisProt and the independent test set from CAID, with intermediate filtering steps to ensure data quality.

Proposed Method

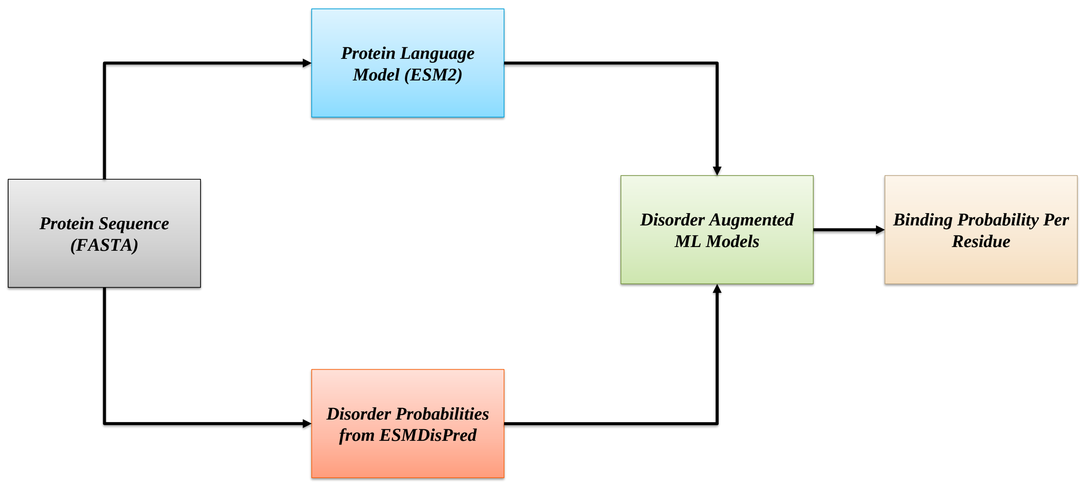

Figure 3. Workflow illustrating the incorporation of disorder information into binding site prediction, where disorder probabilities from ESMDisPred are used as additional features alongside ESM2 embeddings to improve binding residue classification

Machine Learning Models

- We evaluated diverse model architectures to identify the most effective approach for binding site prediction:
- XGBoost, LightGBM, and CatBoost were employed for their superior handling of imbalanced datasets and ability to capture complex feature interactions.
- Multi-layer Perceptron (MLP) architecture was implemented to learn non-linear representations from high-dimensional protein embeddings.
- Random Forest and Extra Trees were used to aggregate predictions and reduce overfitting through bootstrap aggregation.
- Logistic Regression, Linear Discriminant Analysis (LDA), and K-Nearest Neighbors (KNN) served as performance benchmarks for comparison.

Gradient Boosting: XGBoost, LightGBM, CatBoost

Ensemble Learners: Random Forest, Extra Trees

Deep Learner: Multi-layer Perceptron

Baseline Models: Logistic Regression, Linear Discriminant Analysis, K-Nearest Neighbors

Results

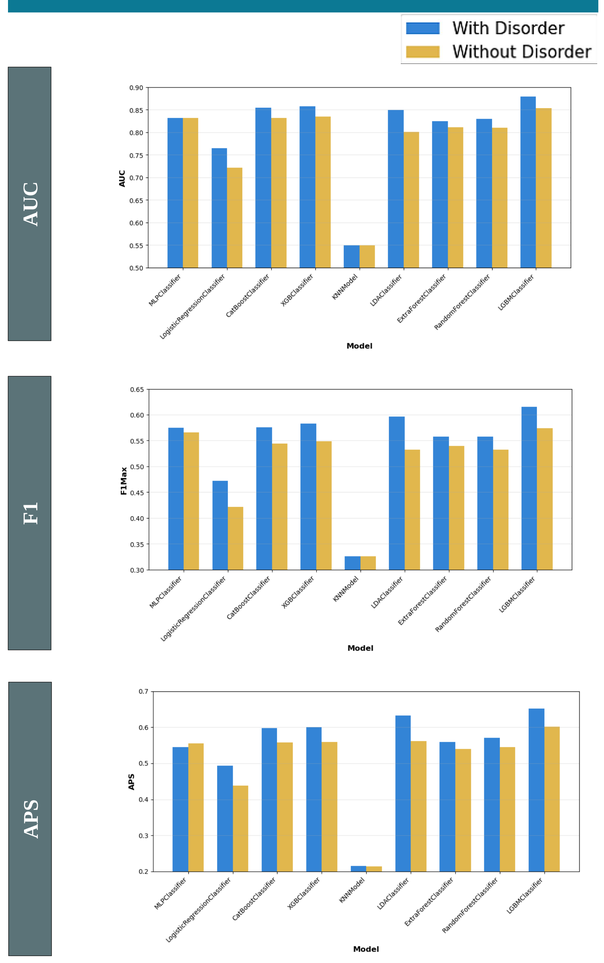

Figure 4. Performance benchmarking results displaying F1, AUC, and APS scores for multiple model architectures, enabling direct comparison between gradient boosting, deep learning, ensemble, and baseline classification methods

Conclusion

- This study demonstrates that **explicitly** including disorder probabilities improves binding prediction.
- Over 90% of binding residues in our dataset occur in disordered regions, supporting our approach.
- Our disorder-augmented models outperform traditional approaches across multiple evaluation metrics.

Future Plans

- In the future, we will optimize model performance by exploring ensemble strategies and deep learning architectures.
- We will further improve prediction capabilities through data techniques to alleviate the severe **class imbalance** between binding and non-binding residues.
- We plan to expand these techniques to investigate the impact of including disorder probabilities to predict other protein functional states, such as **linker** and **transition** states.

Acknowledgements

Funding: This research was supported by an Institutional Development Award (IDeA) from the National Institute of General Medical Sciences of the National Institutes of Health under Grant No. P20GM103424-21. Additional support was provided by LA EPSCoR through the NSF EPSCoR SURE Program, funded by the National Science Foundation under Cooperative Agreement No. OIA-2437963.