A Hybrid Deep Neural Architecture For Residue Level Disorder Prediction

Md Wasi Ul Kabir, Ayon Dey, Farzeen Nafis, Md Tamjidul Hoque
Department of Computer Science, University of New Orleans
{mkabir3, adey, fnafis, thoque}@uno.edu

Introduction

Problem Statement:

- Intrinsically disordered proteins (IDPs) lack stable 3D structures yet play vital roles in biological processes.
- Structural flexibility contributes to diseases: cancer, neurodegenerative disorders.
- Experimental characterization remains challenging.

Motivation & Challenges :

- Critical for drug discovery, structural biology, and protein engineering.
- Computational prediction offers a scalable alternative to experiments.
- CAID assessments reveal limitations in existing predictors.
- Subtle or transient disorder characteristics are difficult to capture.
- Need for structure-aware prediction frameworks.

Proposed Method:

- Leverages evolutionary scale modeling (ESM2) protein language models with structural information from the protein data bank (PDB).
- Hybrid CNN–Transformer architecture for local and long-range dependencies.
- State-of-the-art performance: ROC-AUC 0.895, APS 0.778, F1 max 0.759.

Dataset

Data Source:

- DisProt database (prior to 2023_12 release).
- Initial: 2,845 proteins.
- Final training set: 2,020 proteins after quality filtering.
- Total amino acids: 1,043,829.

Data Processing:

- Removed sequences >2000 amino acids (reduces noise and class imbalance).
- Applied 25% sequence identity cutoff to avoid redundancy.
- Removed proteins with missing residues in PDB structures.

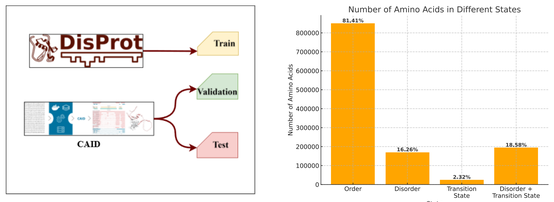

Figure 1: : Illustration of dataset splitting into training, validation, and test subsets derived from the DisProt database and the CAID benchmark dataset. Distribution of amino acids across different structural states (ordered, disordered, transition state) within the training dataset.

Protein Representation

ESM2 Language Model Embeddings:

- Large-scale Transformer trained on millions of protein sequences.
- Unsupervised learning captures evolutionary and structural information.
- Embeddings encode: secondary structure, disorder, binding affinity, functional domains.

Additional Features:

- DisPredict3.0 predictions - disorder-specific features from the previous SOTA model.
- Terminal residue encoding - captures N/C-terminal disorder enrichment.
- Statistical features - mean, variance, skewness, kurtosis from ESM2 embeddings.
- Sliding window analysis - local sequence context (optimal window size: 7).

ESMDisPred Framework & Model Architectures

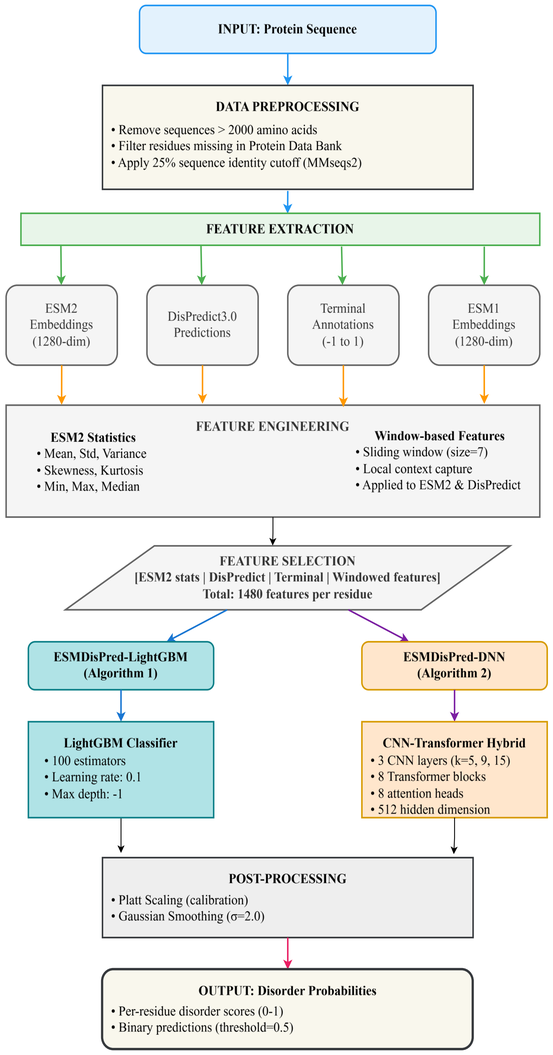

Figure 2: ESMDisPred pipeline for protein disorder prediction.

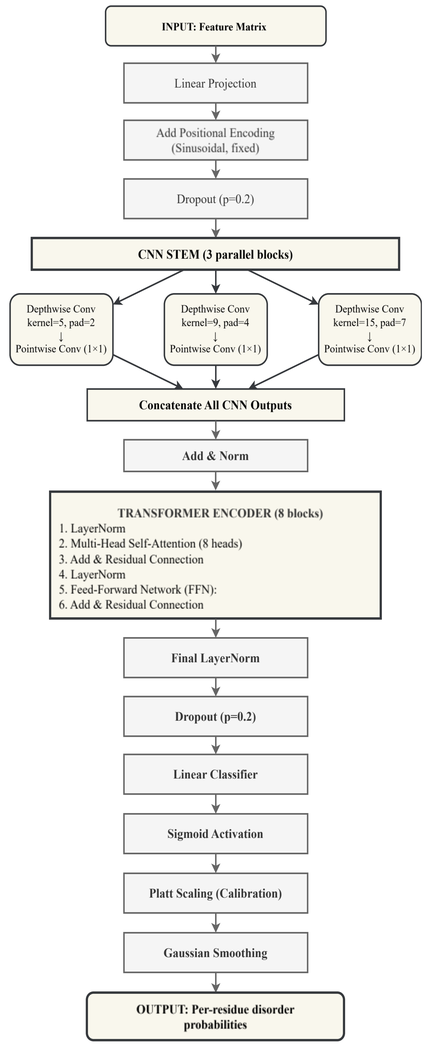

Figure 3: Neural network architecture of ESMDisPred-DNN for disorder prediction.

Results

- ESMDisPred-DNN significantly outperforms ALL other methods ($p < 0.05$).
- Second tier: ESMDisPred-2PDB and DisorderUnetLM (no significant difference).
- Within ESMDisPred family: ESMDisPred-2 > ESMDisPred-1

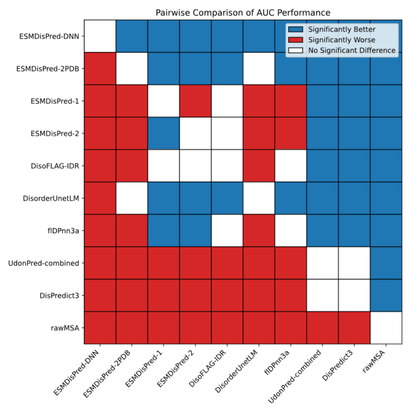

Figure 4: DeLong confidence interval comparison for the top 10 disorder prediction methods from the CAID (Critical Assessment of Intrinsic Disorder) results.

Methods	AUC	Average Precision	F1 max
ESMDisPred-DNN	0.895	0 .778	0.759
ESMDisPred-2PDB	0.885	0.754	0.749
ESMDisPred-1	0.876	0.745	0.720
ESMDisPred-2	0.872	0.743	0.724
DisoFLAG-IDR	0.868	0.714	0.671
DisorderUnetLM	0.862	0.702	0.647
flDPnn3a	0.860	0.700	0.668
UdonPred-combined	0.854	0.668	0.671
DisPredict3.0	0.850	0.666	0.658
rawMSA	0.850	0.671	0.641

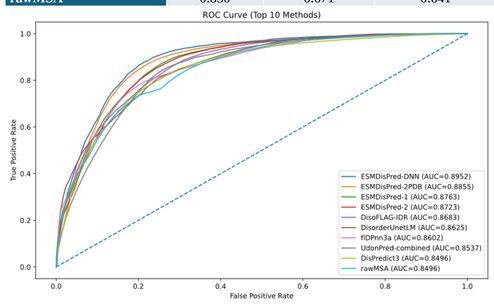

Figure 6: Comparison of ESMDisPred models against other top-performing predictors on the Disorder NOX dataset (CAID2).

Computational Efficiency

Infrastructure:

- LONI HPC: 32-core Intel Xeon Platinum 8358, 512GB RAM.
- NVIDIA A100 GPU (40GB memory).
- Containerized (Docker/Apptainer) for reproducibility.

Runtime Performance:

- Average: ~95 seconds per protein on CPU (CAID3 dataset, 232 sequences).
- Substantially faster on GPU hardware.
- Excellent balance between accuracy and speed.
- rawMSA, flDPnn3b: Much longer runtimes.
- AlphaFold-disorder: 35,000s per protein (highest cost).
- ESMDisPred: High accuracy with low computational cost.

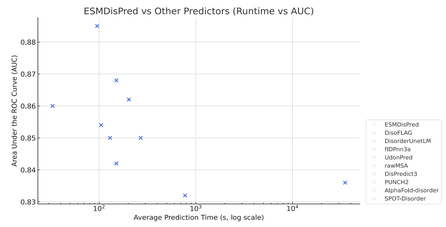

Figure 7: The average prediction time (log scale, in seconds) and area under the ROC curve (AUC) for ten leading protein disorder predictors evaluated in the CAID3 challenges.

Conclusions & Future Work

State-of-the-Art Performance:

- Highest ROC-AUC (0.895), APS (0.778), F1 max (0.759) on CAID3.
- Statistically significant improvements over all competing methods.
- First CNN–Transformer hybrid for disorder prediction.
- Effectively captures local motifs AND long-range dependencies.
- Structure-aware through ESM2 + PDB integration.

Robust Feature Engineering:

- Terminal residue encoding addresses biological insight (N/C-terminal enrichment).
- Statistical + windowing strategies enhance representation.
- Post-processing improves prediction stability.

Computational Efficiency:

- 95s/protein enables proteome-scale applications.
- Favorable accuracy-speed trade-off.

Biological Impact

- Accurate IDP prediction advances drug discovery.
- Identifies flexible regions prone to disease-related modifications.
- Enables protein engineering with disorder considerations.

Web Server: https://bmll.cs.uno.edu.
Source Code: https://github.com/wasicse/ESMDisPred.

References

- Aspromonte et al. (2024). DisProt in 2024. *Nucleic Acids Res*, 52(D1):D434-D441.
- Rives et al. (2021). Biological structure and function from 250M protein sequences. *PNAS*, 118(15).
- Lin et al. (2023). ESMFold: Evolutionary-scale structure prediction. *Science*, 379(6637):1123-1130.
- Kabir & Hoque (2024). DisPredict3.0: Protein language model for disorder. *Appl Math Comput*, 472:128630.
- Mehdiabadi et al. (2025). CAID3: Predicting disorder with language models. *Proteins*, 93(8):e70045.

Acknowledgements

Funding: Research supported by an Institutional Development Award (IDeA) from the National Institute of General Medical Sciences of the National Institutes of Health under grant number **P20GM103424-21.**

Computing Resources: Louisiana Optical Network Infrastructure (LONI) for high-performance computing access.