Matthew Ryan Jonathan

Protein Secondary Structure Prediction

Computational Biology

16 December 2025

Introduction

Protein?

Fundamental workhorses of the cell. Their biological function is directly determined by their three-dimensional structure.

Vision

Our work focuses on Protein Secondary Structure Prediction (PSSP), which identifies local structural motifs including Alpha Helices (H), Beta Sheets (E), and Coils (C).

Dataset

PISCES culled PDB dataset

•Quality Filter: Resolution < 2.5 Å, R-factor < 0.25

•Non-Redundant: Sequence identity < 25% to prevent data leakage

•Chain Constraints: Length 40-10,000 residues

•Split Strategy: 3,000 training proteins / 600 held-out test proteins

•Labels: DSSP derived states collapsed to H (Helix), E (Sheet), C (Coil)

Feature Engineering

Sliding Window (w=17): Centered context for each residue.

BLOSUM62 matrix (evolutionary info) + Hydrophobicity + MW + Charge + Position + Type

One-hot encoding (20) + Normalized Position + Hydrophobicity + MW + Charge.

Model Architectures

Systematically compared five approaches, progressing from simple heuristics into advance deep learning to solve the secondary structure protein prediction.

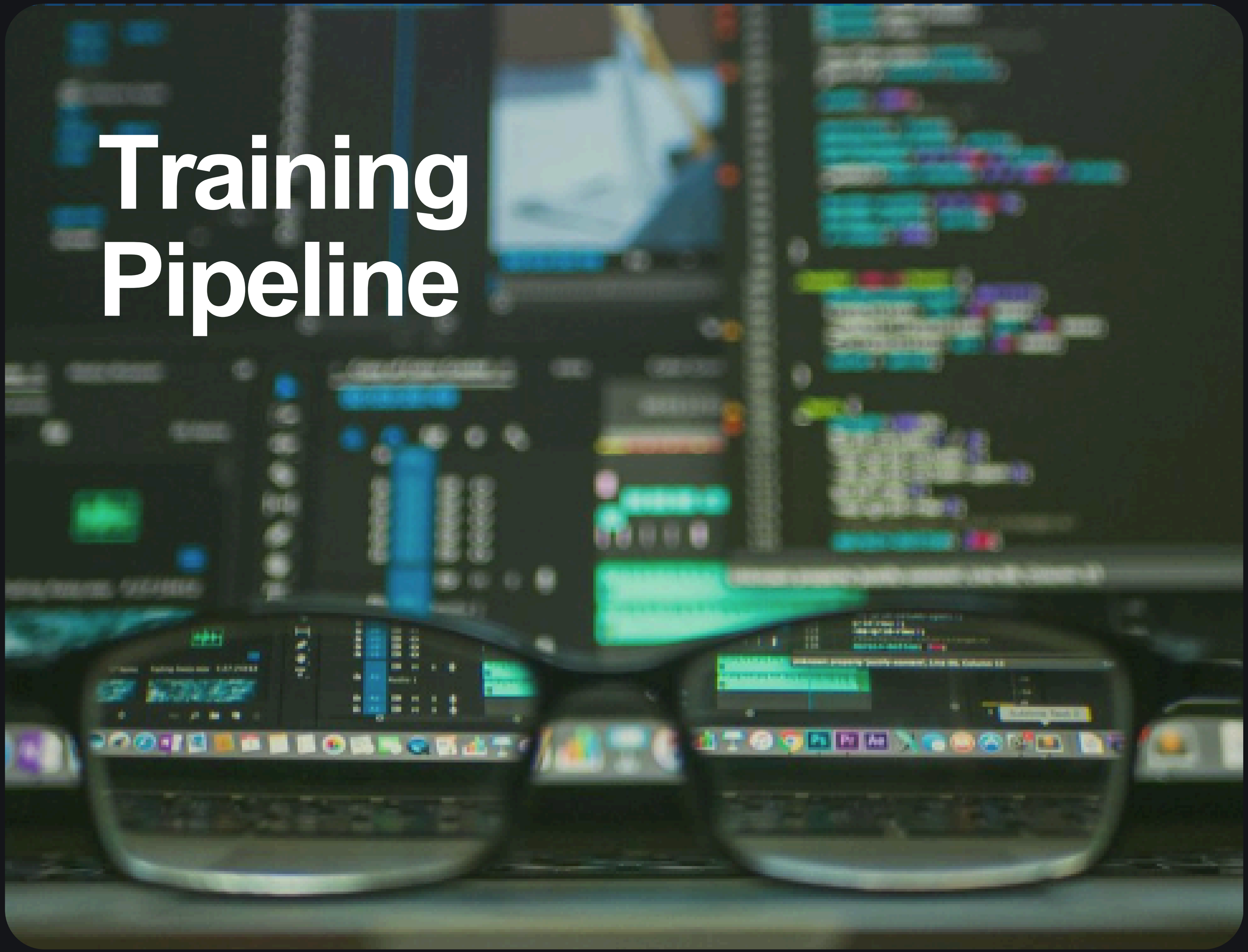
Rule-based

Decision Tree

Random Forest

Gradient Boosting

MLP



Training Pipeline

1. Data Ingestion
Parsing PISCES datasets, filtering protein chains by length (40-10k residues), and performing strict train-test splits (3000 train / 600 test).

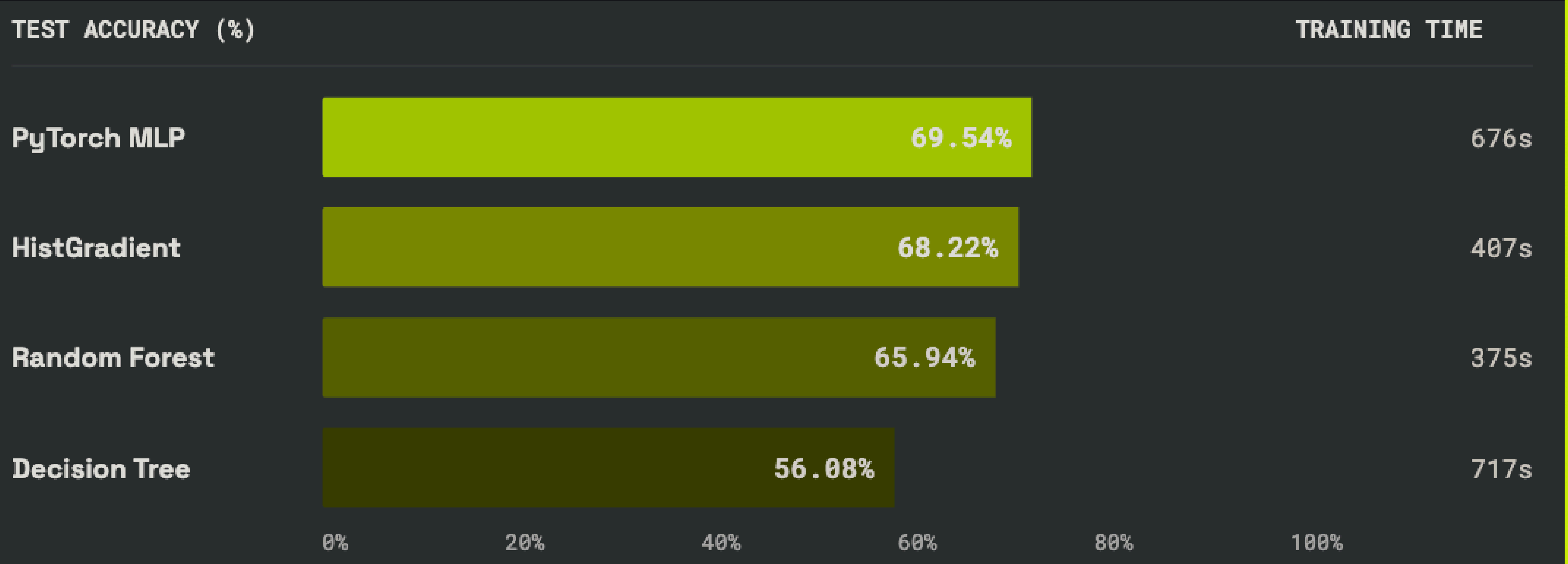
2. Feature Engineering
Generation of sliding window features (w=17) combining BLOSUM62 evolutionary matrices with physicochemical properties like hydrophobicity.

3. Model Training
Parallel execution of Sklearn ensembles and PyTorch MLP training (30 epochs) with real-time validation loss monitoring and checkpointing.

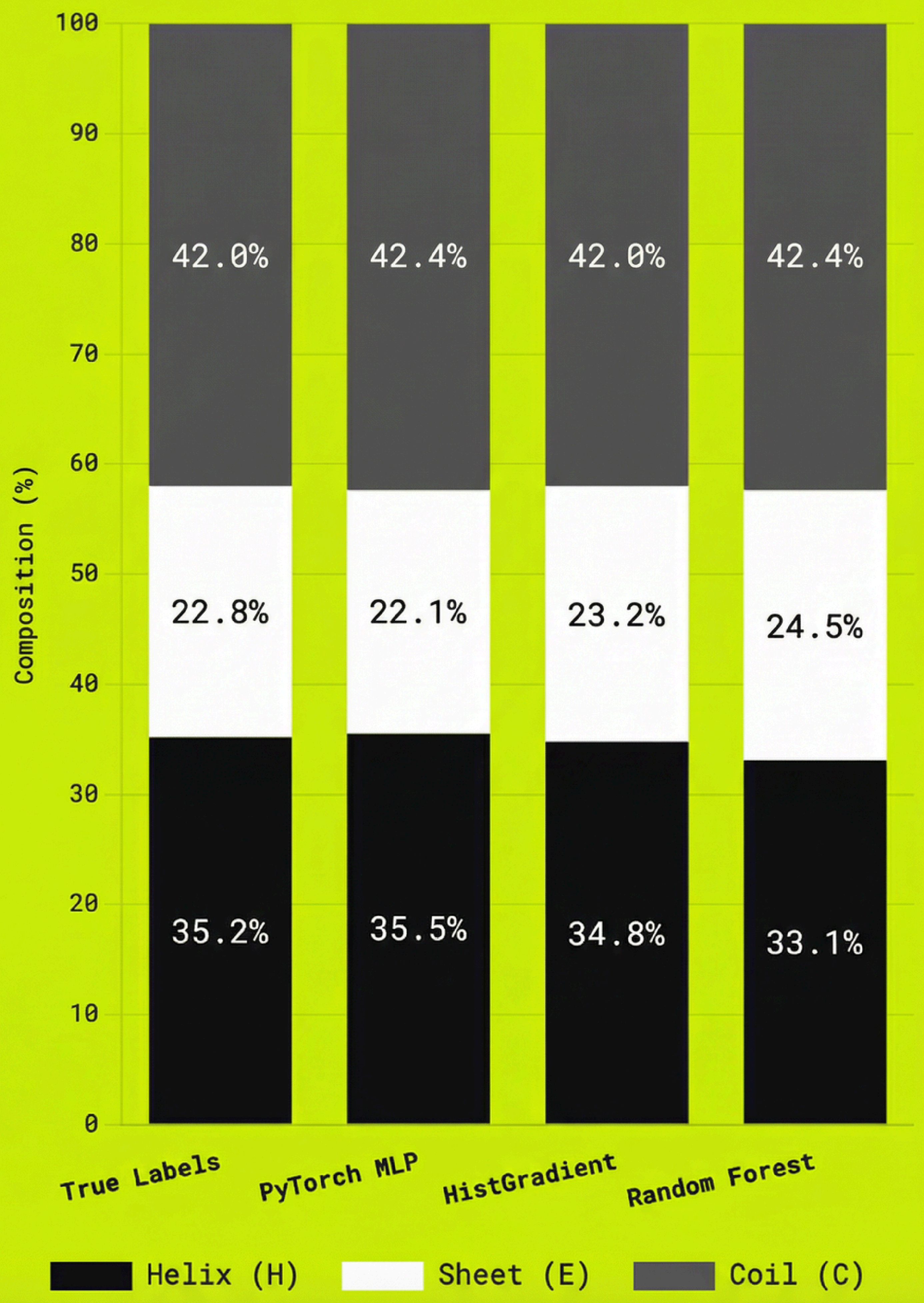
4. Evaluation
Post-processing (smoothing), comprehensive metric calculation, model serialization (.pkl), and integration into the interactive Streamlit dashboard.

Results

Our PyTorch MLP model achieves a competitive 69.54% accuracy without relying on computationally expensive Multiple Sequence Alignments (MSA). While current SOTA methods reach 80-85%, our approach offers sub-second inference and high feature interpretability, making it ideal for rapid screening.



Structure Distribution Comparison



Architecture

Integration of BiLSTM/GRU networks and Transformer-based attention mechanisms to better capture long-range dependencies in protein sequences.

Training

Implementation of advanced data augmentation strategies, class weighting for imbalance handling, and multi-task learning objectives.

Features

Incorporation of Position-Specific Scoring Matrices (PSSMs), predicted solvent accessibility, and contact maps to enrich input data representations.

Deployment

Development of a scalable RESTful API with Docker containerization and GPU acceleration for high-throughput batch processing.

Conclusion

This project successfully establishes a production-ready machine learning pipeline for Protein Secondary Structure Prediction (PSSP). By systematically comparing five distinct modeling approaches, we demonstrated that deep learning architectures can achieve competitive results without relying on computationally expensive sequence alignments.

- Performance: PyTorch MLP achieved 69.54% accuracy, effectively capturing local structural dependencies.
- Robustness: Validated the efficacy of combining BLOSUM62 evolutionary data with physicochemical properties.
- Impact: Delivered an open-source interactive dashboard, bridging the gap between theoretical ML and practical biology.

References

[1] J. Jumper et al., ”Highly accurate protein structure prediction with AlphaFold,” Nature, vol. 596, pp. 583-589, 2021.

[2] P. Y. Chou and G. D. Fasman, ”Prediction of protein conformation,” Biochemistry, vol. 13, no. 2, pp. 222-245, 1974.

[3] J. Garnier, D. J. Osguthorpe, and B. Robson, ”Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins,” J. Mol. Biol., vol. 120, pp. 97-120, 1978.

[4] B. Rost and C. Sander, ”Prediction of protein secondary structure at better than 70% accuracy,” J. Mol. Biol., vol. 232, pp. 584-599, 1993.

[5] H. Kim and H. Park, ”Protein secondary structure prediction based on an improved support vector machines approach,” Protein Eng., vol. 16, no. 8, pp. 553-560, 2003.

[6] S. Wang, J. Peng, J. Ma, and J. Xu, ”Protein secondary structure prediction using deep convolutional neural fields,” Scientific Reports, vol. 6, 18962, 2016.

[7] R. Heffernan et al., ”Capturing non-local interactions by long short-term memory bidirectional recurrent neural networks for improving prediction of protein secondary structure,” Bioinformatics, vol. 33, no. 18, pp. 2842- 2849, 2017.

[8] M. S. Klausen et al., ”NetSurfP-2.0: Improved prediction of protein structural features by integrated deep learning,” Proteins, vol. 87, no. 6, pp. 520-527, 2019.

[9] G. Wang and R. L. Dunbrack Jr., ”PISCES: a protein sequence culling server,” Bioinformatics, vol. 19, no. 12, pp. 1589-1591, 2003.

[10] W. Kabsch and C. Sander, ”Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features,” Biopolymers, vol. 22, no. 12, pp. 2577-2637, 1983.