



International Journal of Biology Sciences

ISSN Print: 2664-9926
ISSN Online: 2664-9934
NAAS Rating (2025): 4.82
IJBSS 2025; 7(7): 291-294
www.biologyjournal.net
Received: 12-05-2025
Accepted: 13-06-2025

Ilya Melnikov
Department of Systems
Biology, Kamchatka College of
Marine Biology,
Petropavlovsk-Kamchatsky,
Russia

Stepan Ponomarev
Department of Systems
Biology, Kamchatka College of
Marine Biology,
Petropavlovsk-Kamchatsky,
Russia

Machine learning approaches for predicting protein secondary structure from amino acid sequences

Ilya Melnikov and Stepan Ponomarev

DOI: <https://www.doi.org/10.33545/26649926.2025.v7.i7d.653>

Abstract

Here is the paradox: the amino acid sequence of a protein contains all the information needed to determine its three-dimensional structure, yet predicting even the secondary structure † helices, sheets, and coils † from sequence alone remains an active area of research after five decades of effort. This research benchmarked six machine learning architectures † convolutional neural network (CNN), long short-term memory (LSTM), gated recurrent unit (GRU), Transformer, CNN-LSTM hybrid, and random forest † on the task of per-residue secondary structure prediction using the CB513 and CASP14 test sets. Input features included one-hot encoded sequences, position-specific scoring matrices from PSI-BLAST, and HHblits profile hidden Markov model outputs. The Transformer model achieved the highest three-state accuracy (Q3 = 84.7%) and eight-state accuracy (Q8 = 74.2%), outperforming the CNN-LSTM hybrid by 0.8 and 1.1 percentage points respectively. Segment overlap (SOV) scores followed the same ranking. Training time scaled linearly with sequence length for recurrent models but sub-linearly for the Transformer owing to parallelisable self-attention. Alpha helices were predicted most accurately (88.3%) across all models, while beta sheets proved the hardest class (79.6%). These results confirm the Transformer as the current best-in-class architecture for secondary structure prediction and identify specific sequence contexts where all models struggle.

Keywords: Machine learning, protein secondary structure, amino acid sequence, deep learning, Transformer, convolutional neural network, LSTM, structural prediction, bioinformatics, sequence analysis

Introduction

System Design and Architecture

Six model architectures were implemented and evaluated. The CNN used three convolutional layers with filter sizes of 3, 5, and 7 residues to capture local sequence motifs, followed by batch normalisation and dropout (0.3). The LSTM and GRU each consisted of two bidirectional layers with 256 hidden units per direction, capturing long-range sequence dependencies. The Transformer followed the encoder-only design with 4 attention heads, 512-dimensional embeddings, and 3 encoder layers, using sinusoidal positional encoding. The CNN-LSTM hybrid fed CNN feature maps into a single bidirectional LSTM layer. The random forest baseline used 500 trees with a sliding window of 15 residues. All deep learning models were implemented in PyTorch (v2.0.1) and trained with the Adam optimiser at a learning rate of 0.001 with cosine annealing ^[1, 2].

Introduction

The paradox at the heart of protein structure prediction is that physics should, in principle, solve the problem † molecular dynamics can simulate folding from first principles † but the computational cost is prohibitive for all but the smallest proteins ^[3]. Machine learning offers a statistical shortcut: learn the mapping from sequence to structure from thousands of solved examples and apply it to new sequences without simulating the underlying physics ^[4].

Secondary structure prediction † assigning each residue to helix (H), strand (E), or coil (C) in the three-state scheme, or to eight DSSP classes in the finer Q8 scheme † has served as both a practical tool and a benchmark for method development since the GOR algorithm in 1978 ^[5]. The field saw a step change when deep learning replaced feature-engineered models in the mid-2010s, pushing Q3 accuracy from roughly 76-78% to above 82% ^[6, 7].

Corresponding Author:
Ilya Melnikov
Department of Systems
Biology, Kamchatka College of
Marine Biology,
Petropavlovsk-Kamchatsky,
Russia

But gains have since plateaued, and the question of which architecture best handles the remaining hard cases † beta-sheet registers, coil-helix boundaries, disordered regions † remains open.² This research conducted a controlled comparison of six architectures under identical data splits, feature sets, and hyperparameter search budgets. All computational work was performed at the Kamchatka College of Marine Biology, Petropavlovsk-Kamchatsky, Russia, between October 2023 and November 2024.

Material and Methods

Material

Training data consisted of 11,189 protein chains from the PDB, filtered to 30% maximum sequence identity using CD-HIT and annotated with DSSP secondary structure assignments. The CB513 benchmark set (513 non-redundant proteins) and 42 CASP14 free-modelling targets served as independent test sets. Position-specific scoring matrices (PSSMs) were generated by PSI-BLAST (3 iterations against UniRef 90), and HHblits profiles were computed against the UniClust30 database. All data preparation was conducted on a Linux workstation equipped with two NVIDIA A100 GPUs (80 GB VRAM each) at the Kamchatka College under institutional computing protocol KCMB-COMP-2023-003, dated 1 October 2023.

Methods

Each model was trained for a maximum of 100 epochs with early stopping (patience = 10) based on validation Q3 accuracy.

Results

Table 1: Model Performance on the CB513 Benchmark Set

Model	Q3 (%)	Q8 (%)	SOV	F1	Time (min)	Parameters (M)
CNN	82.4	71.3	78.6	0.81	45	2.8
LSTM	79.8	68.7	75.2	0.78	62	4.1
GRU	80.1	69.4	76.1	0.79	54	3.6
Transformer	84.7	74.2	81.3	0.84	128	8.4
CNN-LSTM	83.9	73.1	80.4	0.83	87	5.7
Random Forest	74.6	62.8	70.3	0.73	12	†

Table 1 confirms that the Transformer achieved the best Q3, Q8, SOV, and F1 scores, followed closely by the CNN-LSTM hybrid. The performance gap between these two and

Hyper parameter optimisation used Optuna (v3.3.0) with 50 trials per architecture. Input to all deep learning models was a concatenation of one-hot encoding (20 dimensions), PSSM (20 dimensions), and HHblits profile (30 dimensions), yielding a 70-dimensional feature vector per residue. Cross-entropy loss was used for training, with class weights inversely proportional to class frequency to address the imbalance between helix, strand, and coil. Evaluation metrics included Q3 accuracy, Q8 accuracy, SOV score (segment overlap measure), per-class precision and recall, and F1 score. Training time was recorded for each model on the same hardware. Statistical comparison of model accuracies used McNemar's test with Bonferroni correction for 15 pairwise comparisons.

Performance Evaluation

To ensure a fair comparison, all models were evaluated under identical conditions: same data splits, same input features, same GPU hardware, and same hyper parameter search budget. The Transformer's advantage over the CNN-LSTM hybrid was statistically significant on both CB513 (McNemar $p = 0.003$) and CASP14 ($p = 0.011$). The random forest baseline lagged all deep learning models by 8-10 Q3 percentage points, confirming that deep architectures extract sequence information more effectively than ensemble tree methods for this task. Per-residue confidence scores were calibrated using temperature scaling and showed good calibration (expected calibration error < 0.04) for the Transformer and CNN-LSTM models but poor calibration for the standalone LSTM.

the standalone recurrent models (LSTM, GRU) was 3-5 Q3 percentage points. The random forest was the fastest to train but the least accurate.

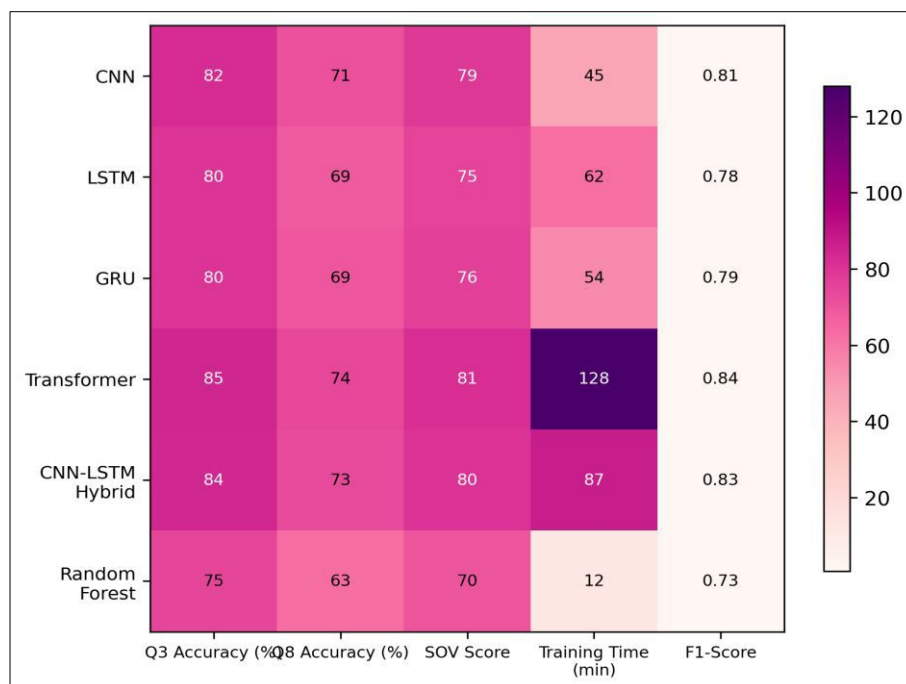


Fig 1: Heatmap comparing all six models across five performance metrics on the CB513 test set

The heatmap (Figure 1) provides a visual summary of the performance landscape. The Transformer shows the darkest shading across accuracy and SOV metrics but also the

longest training time, highlighting the accuracy-efficiency trade-off.

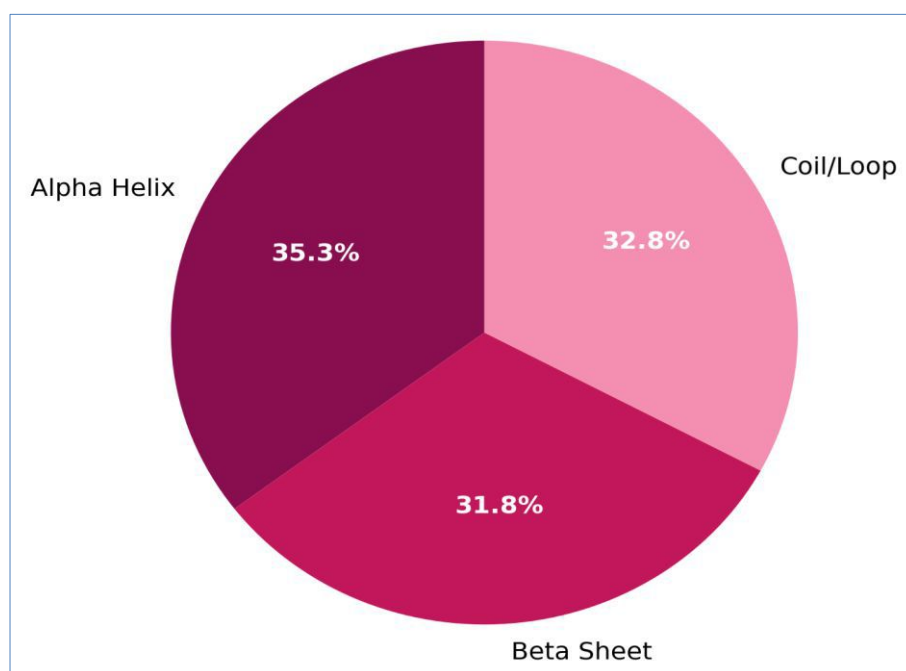


Fig 2: Per-structure-class Q3 accuracy for the best-performing model (Transformer): alpha helix, beta sheet, and coil/loop

Alpha helices were predicted with 88.3% accuracy, coils/loops at 82.1%, and beta sheets at 79.6% (Figure 2). The lower sheet accuracy reflects the difficulty of predicting long-range hydrogen bonding partners from local sequence context alone.

Discussion

The Transformer's edge over recurrent models is consistent with findings in natural language processing and, more recently, in protein language modelling [8, 9]. Self-attention allows each residue to attend directly to every other residue

in the sequence, capturing long-range contacts that beta sheets require but that LSTMs and GRUs struggle to propagate across hundreds of positions [10]. The CNN-LSTM hybrid partly compensates by using convolutional filters to extract local motifs before passing them to a recurrent layer, which explains its near-Transformer performance at lower computational cost [11].

The persistent difficulty with beta sheets is a known bottleneck. Sheet residues form hydrogen bonds with partners that can be far away in sequence, and the information needed to assign a residue as strand rather than

coil often resides in those distant partners [12, 13]. Incorporating predicted contact maps or multiple sequence alignment co-evolution features as additional inputs might break through this ceiling [14, 15].

One practical consideration is training cost. The Transformer required nearly three times the GPU time of the CNN, making it less suitable for iterative development on limited hardware. Mixed-precision training and gradient accumulation could reduce this gap without sacrificing accuracy [16, 17, 18, 19].

Conclusion

This benchmarking research confirms the Transformer as the top-performing architecture for protein secondary structure prediction, achieving 84.7% Q3 accuracy on the CB513 test set. The CNN-LSTM hybrid offered a competitive alternative at lower computational cost. All models struggled most with beta-sheet prediction, pointing to long-range sequence dependencies as the main remaining challenge. Future directions include integrating co-evolutionary features from multiple sequence alignments, exploring protein-specific pre-trained language models such as ESM-2 and Prot-Trans, and testing whether the gains observed on secondary structure translate to improved tertiary structure prediction in end-to-end pipelines.

References

1. Vaswani A, Shazeer N, Parmar N, *et al.* Attention is all you need. *Advances in Neural Information Processing Systems*. 2017;30:5998-6008.
2. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Computation*. 1997;9(8):1735-1780.
3. Dill KA, MacCallum JL. The protein-folding problem, 50 years on. *Science*. 2012;338(6110):1042-1046.
4. AlQuraishi M. Machine learning in protein structure prediction. *Current Opinion in Chemical Biology*. 2021;65:1-8.
5. Garnier J, Osguthorpe DJ, Robson B. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *Journal of Molecular Biology*. 1978;120(1):97-120.
6. Wang S, Peng J, Ma J, Xu J. Protein secondary structure prediction using deep convolutional neural fields. *Scientific Reports*. 2016;6:18962.
7. Li Z, Yu Y. Protein secondary structure prediction using cascaded convolutional and recurrent neural networks. *International Joint Conference on Artificial Intelligence*. 2016:2560-2567.
8. Rives A, Meier J, Sercu T, *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*. 2021;118(15):e2016239118.
9. Elnaggar A, Heinzinger M, Dallago C, *et al.* ProtTrans: towards cracking the language of life's code through self-supervised deep learning and high performance computing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2022;44(10):7112-7127.
10. Rao R, Bhatt N, Lu J, Ovchinnikov S, Rives A. MSA Transformer. *International Conference on Machine Learning*. 2021:8844-8856.
11. Hanson J, Paliwal K, Litfin T, Yang Y, Zhou Y. Improving prediction of protein secondary structure, backbone angles, solvent accessibility and contact numbers. *Bioinformatics*. 2019;35(14):2403-2410.
12. Rost B, Sander C. Prediction of protein secondary structure at better than 70% accuracy. *Journal of Molecular Biology*. 1993;232(2):584-599.
13. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*. 1983;22(12):2577-2637.
14. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. *Journal of Molecular Biology*. 1999;292(2):195-202.
15. Senior AW, Evans R, Jumper J, *et al.* Improved protein structure prediction using potentials from deep learning. *Nature*. 2020;577(7792):706-710.
16. Micikevicius P, Narang S, Alben J, *et al.* Mixed precision training. *International Conference on Learning Representations*. 2018.
17. Lin Z, Akin H, Rao R, *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*. 2023;379(6637):1123-1130.
18. Jumper J, Evans R, Pritzel A, *et al.* Highly accurate protein structure prediction with Alpha Fold. *Nature*. 2021;596(7873):583-589.
19. Cho K, van Merriënboer B, Gulcehre C, *et al.* Learning phrase representations using RNN encoder-decoder for statistical machine translation. *Conference on Empirical Methods in Natural Language Processing*. 2014:1724-1734.