Protein Structure Prediction using Attention-ProteinMeNet

Oxford Brookes University in collaboration with Chengdu University of Technology

Author: Jia Xin Yue

Supervisor: Dr Grace Ugochi Nneji

Abstract

1

Protein secondary structure prediction is key to understanding function and supporting drug design. Traditional experimental methods are costly and slow, while conventional algorithms lack accuracy on complex sequences.

Attention-ProteinMeNet combines convolutional layers, bidirectional LSTM, and attention mechanisms to improve prediction performance. Trained on RCSB-PDB and CB513 datasets, the model achieved validation accuracies of 96.49% and 94.15%, with high ROC-AUC and F1-scores.

SHAP analysis enhances interpretability, and a graphical interface enables both single and batch predictions. The model provides a reliable and efficient tool for protein structure analysis.

Dataset

2

Dataset 1: RCSB-PDB：Contains high-quality, experimentally annotated protein sequences. Used for model training and validation.

Dataset 2: CB513: A benchmark dataset with 513 proteins for testing model generalization.

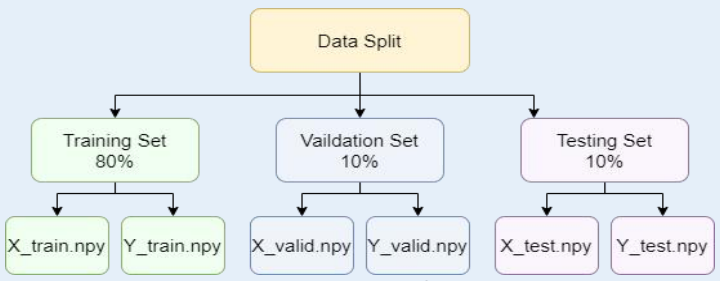

Figure 1.Dataset Split

Preprocessing includes:
SST8 → SST3 label mapping
One-hot encoding
Sequence padding & normalization
80/10/10 train/val/test split

Model Explainability--SHAP

5

To enhance model transparency and build trust. uses methods such as SHAP.

Confirms biological relevance of focused sequence regions.

Improves model trust and interpretability.

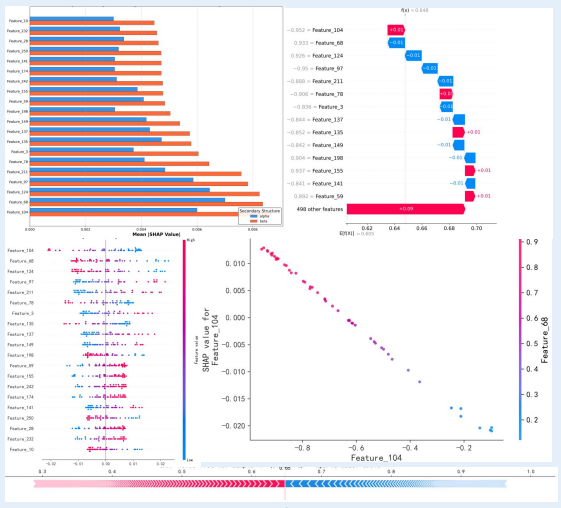

Figure 7. SHAP

Proposed Model

3

Attention-ProteinMeNet

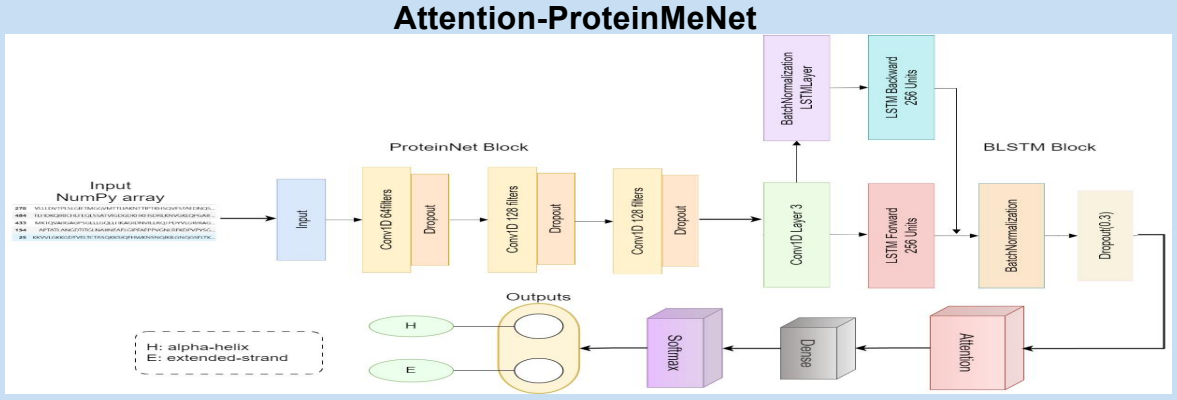

Figure 2. Architectural Overview of Attention ProteinMeNet Model

Figure2 shows Attention-ProteinMeNet is a deep learning model that combines three key components:ProteinNet (Conv1D layers) to extract local sequence features. Bidirectional LSTM to capture long-range dependencies. Attention mechanism to focus on key residues influencing structure.
The final output layer uses softmax for residue classification.This architecture improves accuracy, generalization, and interpretability in secondary structure prediction.

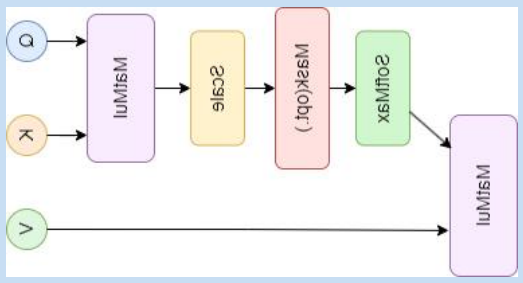

Figure 3. Attention Block

Figure 3 shows The attention module assigns dynamic weights to each residue in the sequence, allowing the model to selectively focus on structurally important positions.
By evaluating the contribution of each amino acid to the final prediction, attention enhances both the accuracy and interpretability of the model. It is especially effective in identifying long-range interactions that traditional models often miss, making it a key component in improving structural prediction performance.

Result Analysis

4

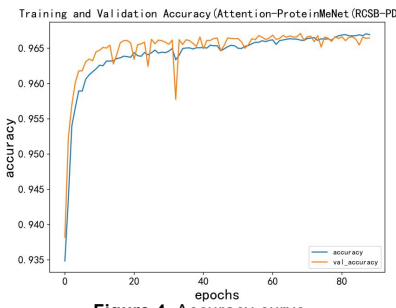

Figure 4. Accuracy curve

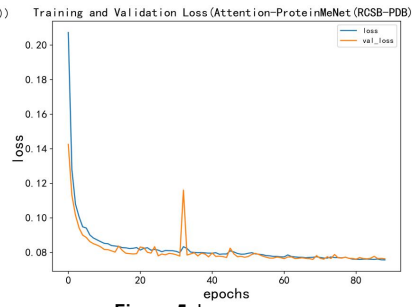

Figure 5. Loss curve

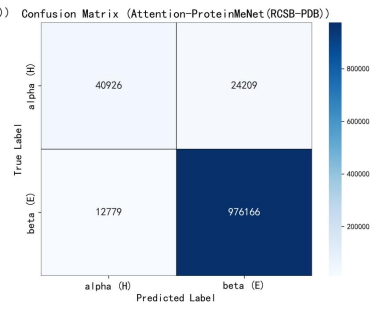

Figure 6. Confusion Matrix

The model shows strong performance on the RCSB-PDB dataset. As shown in the accuracy and loss curves (Figures 5 & 6), training converges quickly and remains stable, reaching a validation accuracy of 96.49% with minimal overfitting. The confusion matrix (Figure 7) indicates high classification accuracy for both helix (H) and strand (E) structures, demonstrating balanced and reliable predictions.

GUI Deployment

6

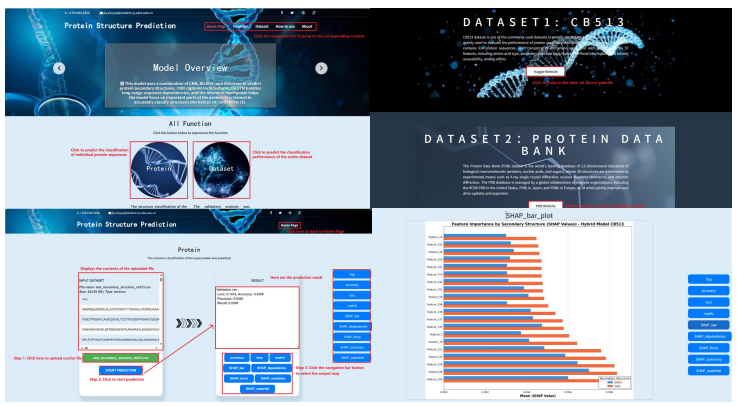

Figure 8 shows an interactive web-based GUI allows users to upload protein sequences, run predictions, and view results with SHAP-based interpretation. It supports both single and batch input, and provides easy access to dataset info and model outputs, making the tool accessible to non-technical users.

Conclusions

7

• This project proposed a novel deep learning model, Attention-ProteinMeNet, for protein secondary structure prediction.

• ProteinNet for local feature extraction

• BLSTM for sequence dependency modeling

• Attention mechanism to focus on key residues

• the model achieves both high accuracy and strong generalization

•The model achieved 96.49% validation accuracy on the RCSB-PDB dataset and 94.15% on CB513.

•Explainable AI (SHAP analysis) enhanced the interpretability of model predictions.

•A GUI tool was developed to support real-time and batch structure prediction for practical use in biological research.

References

9

- J. Hong, Z.-H. Zhan, L. He, Z. Xu, and J. Zhang, 'Protein Structure Prediction Using A New Optimization-Based Evolutionary and Explainable Artificial Intelligence Approach', IEEE Trans. Evol. Comput., pp. 1–1, 2024, doi: 10.1109/TEVC.2024.3366814.
- S. Prasad, N. Nandhini, R. Singh, A. Anuradha, L. Varshitha Averineni, and S. Debnath, 'Perspectives of machine learning on protein structure prediction and function', in 2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), May 2023, pp. 385–390. doi: 10.1109/ICACITE57410.2023.10183157.
- R. K. Deepak, and M. K. Praveen, 'A Review of Machine Learning Techniques and Applications for Health Care', IEEE Access, pp. 4-8, 2021, doi: 10.1109/ICATME50232.2021.9732761.
- X. Qiu, H. Li, G. Ver Steeg, and A. Godzik, 'Advances in AI for Protein Structure Prediction: Implications for Cancer Drug Discovery and Development', Biomolecules, vol. 14, no. 3, Art. no. 3, Mar. 2024, doi: 10.3390/biom14030339.
- A. Paiardini, 'Protein Structure Prediction in Drug Discovery', Biomolecules, vol. 13, no. 8, Art. no. 8, Aug. 2023, doi: 10.3390/biom13081258.
- A. Shehu and L. E. Kavraki, 'Modeling Structures and Motions of Loops in Protein Molecules', Entropy, vol. 14, no. 2, pp. 252–290, Feb. 2012, doi: 10.3390/e14020252.
- T. Selwate, M. A. Kamble, P. M. Sabale, D. Dhabarde, K. Dongarwar, and J. Baheti, 'Protein Structure Prediction: A Computational Approach to Unraveling Molecular Mysteries', in Deep Learning and Computer Vision: Models and Biomedical Applications: Volume 1, U. N. Dulhare and E. H. Houssein, Eds., Singapore: Springer Nature, 2025, pp. 63–87. doi: 10.1007/978-981-96-1285-7_4.
- 'DNACoder: a CNN-LSTM attention-based network for genomic sequence data compression | Neural Computing and Applications'. Accessed: Dec. 18, 2024.. Available: https://link.springer.com/article/10.1007/s00521-024-10130-4
- X. Ma and E. Hovy, 'End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF', May 29, 2016, arXiv: arXiv:1603.01354. doi: 10.48550/arXiv.1603.01354.
- N. P, K. M. Sudar, V. S. Sri, Nikitha. V, V. S. S. Reddy, and V. M, 'Enhancing Protein Structure Generation Through Deep Learning Techniques', in 2024 Third International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS), Mar. 2024, pp. 1–6. doi: 10.1109/INCOS59338.2024.10527559.
- J. L. Filgueiras, D. Varela, and J. Santos, 'Protein structure prediction with energy minimization and deep learning approaches', Nat. Comput., vol. 22, no. 4, pp. 659–670, Dec. 2023, doi: 10.1007/s11047-023-09943-4.
- Z. Shi and B. Li, 'Graph neural networks and attention-based CNN-LSTM for protein classification', Feb. 22, 2023, arXiv: arXiv:2204.09486. doi: 10.48550/arXiv.2204.09486.
- M. M. Mohamed Mufassirin, M. A. H. Newton, J. Rahman, and A. Sattar, 'Multi-S3P: Protein Secondary Structure Prediction With Specialized Multi-Network and Self-Attention-Based Deep Learning Model', IEEE Access, vol. 11, pp. 57083–57096, 2023, doi: 10.1109/ACCESS.2023.3282702.
- A. Golwelkar and A. Kothari, 'A Review of Protein Sequences of COVID-19 Using Machine Learning and Deep Learning Approaches', in 2023 IEEE International Conference on ICT in Business Industry & Government (ICTBIG), Dec. 2023, pp. 1–9. doi: 10.1109/ICTBIG59752.2023.10456322.
- 'A New Approach Of Applying Deep Learning To Protein Model Quality Assessment | IEEE Conference Publication | IEEE Xplore'. Accessed: Oct. 29, 2024. Available: https://ieeexplore.ieee.org/document/8983005
- J. S. A, K. Merriliance, and N. Soundiraraj, 'Integrating Deep Learning with Structural Bioinformatics using Next-Generation Protein Stability Prediction', in 2024 International Conference on Inventive Computation Technologies (ICICT), Apr. 2024, pp. 1252–1257. doi: 10.1109/ICICT60155.2024.10544908.