Predicting Protein Secondary Structure using a Machine Learning model.

Hidalgo-Moreno, Javier

Link to the SPP [Google collab notebook](#).

Introduction

Proteins are fundamental biological macromolecules whose functions are intrinsically linked to their three-dimensional structures. Therefore, predicting their structure through computational methods has been a major challenge in bioinformatics. Advancements in machine learning (ML) have revolutionized this field, enabling more accurate predictions by leveraging deep learning architectures and ensemble models trained on evolutionary features (Srivastava et al., 2025).

This report details the development of an ML model designed to predict secondary structures from PSSM or FASTA inputs and output results in DSSP format. The model combines convolutional neural networks (CNNs) with bidirectional neural networks, achieving approximately 70% accuracy on the test set.

Material and methods

Data

The dataset used for the model development included PSSM files, FASTA files and DSSP files from 1,200 proteins. PSSM matrices are obtained by retrieving evolutionary information through multiple sequence alignment, FASTA files contain just the amino acid sequences and DSSP files contain the output labels H (for α-helix), E (for β-sheet) and C (for coil), representing secondary structure elements.

The FASTA and DSSP were one-hot encoded using 20 and 4 classes respectively. Then, all data was padded to obtain the same dimension, which is need when using CNN layers. FASTA and PSSM files were padded with zeroes. However, DSSP was padded to a fourth dummy class, X, as padding with zeroes was problematic for the model. Nonetheless, this class was excluded from the model evaluation.

Model architecture

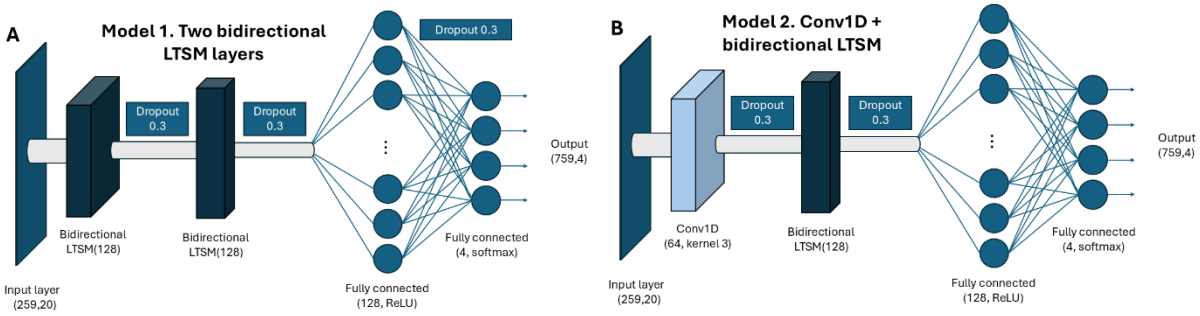


Fig. 1. Architecture of the two biredictional LTSM layers model (A) and the model that combines a 1D-convolutional layer with a bidirectional LTSM layer.

The model architecture was taken from (Singh et al., 2021). It consists of two bidirectional LTSM layers with 0.3 dropout regularization, and a fully connected dense layer (Fig. 1A). Here, we also test an alternative model that combines a 1D-convolutional layer with a bidirectional LTSM layer. It has

batch normalization and dropout regularization, and also a fully connected layer (Fig. 1B) Both models' output layer is a four-neurons fully connected dense layer with softmax activation that will predict the 4 classes in DSSP (coil, helix, strand or padded). For each architecture, two models were trained separately: one for PSSM files and one for FASTA files.

The models were trained using Adam optimizer and cross-entropy loss function for 15 epochs.

Model evaluation

The models were compared in terms of accuracy and loss for the validation and the training set. To not consider the results from the dummy class padded, a new metric was introduced: masked and loss accuracy, which only consider coil, strand and helix classes. The test set was evaluated through confusion matrixes, from which the accuracy, precision and the Matthew's confusion coefficient (MCC) was calculated.

To understand the contribution of each layer in the model ablation studies were performed on the model that combines CNN and LTSM layers. In each case, one layer was ablated: 1D-Convolutional, LTSM bidirectional or dense layer. Additionally, we tested the contribution of dropout to the model performance. We used the same metrics for comparison with the unablated model.

Results

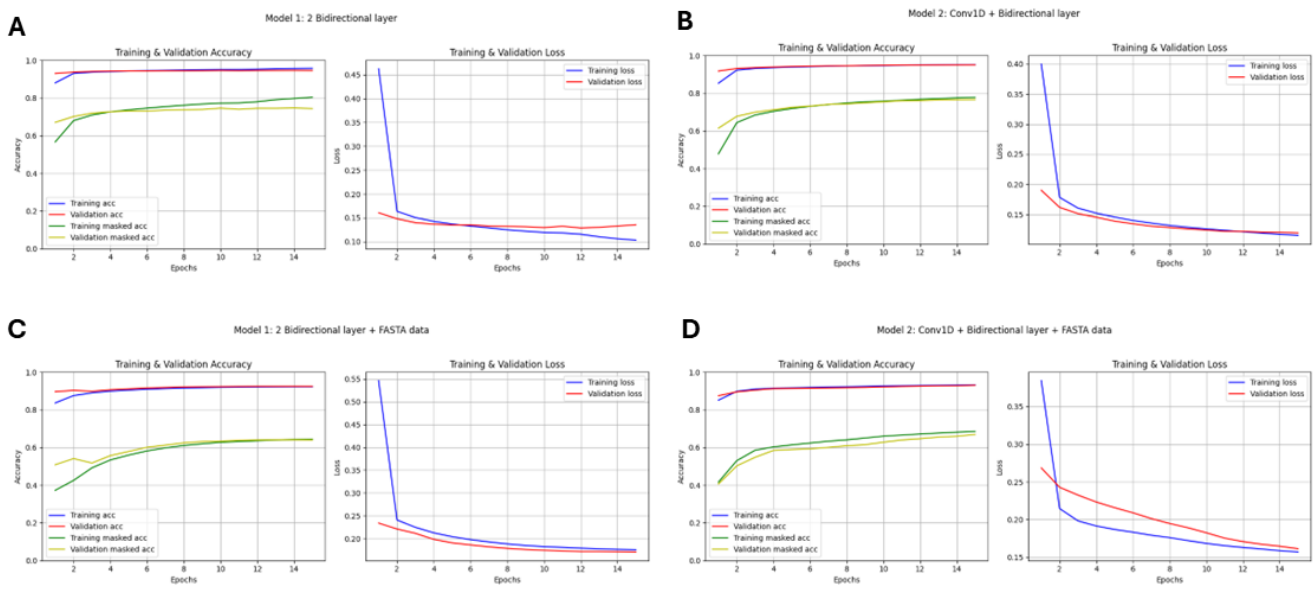


Fig. 2. Accuracy and loss curves for model 1 (A and C) and model 2 (B and D) trained with PSSM data (A-B) and FASTA data (C-D)

The results from the accuracy and the loss curves show similar patterns (Fig 2). The accuracy stabilizes at early epochs around 95% for the accuracy and 80% for the mask accuracy in PSSM data (Fig. 2A-B) and 60% for the FASTA files (Fig. 2C-D). The curves plateau and are similar for the validation and training set, indicating a decent learning procedure. However, there is still room for improvement of the masked accuracy.

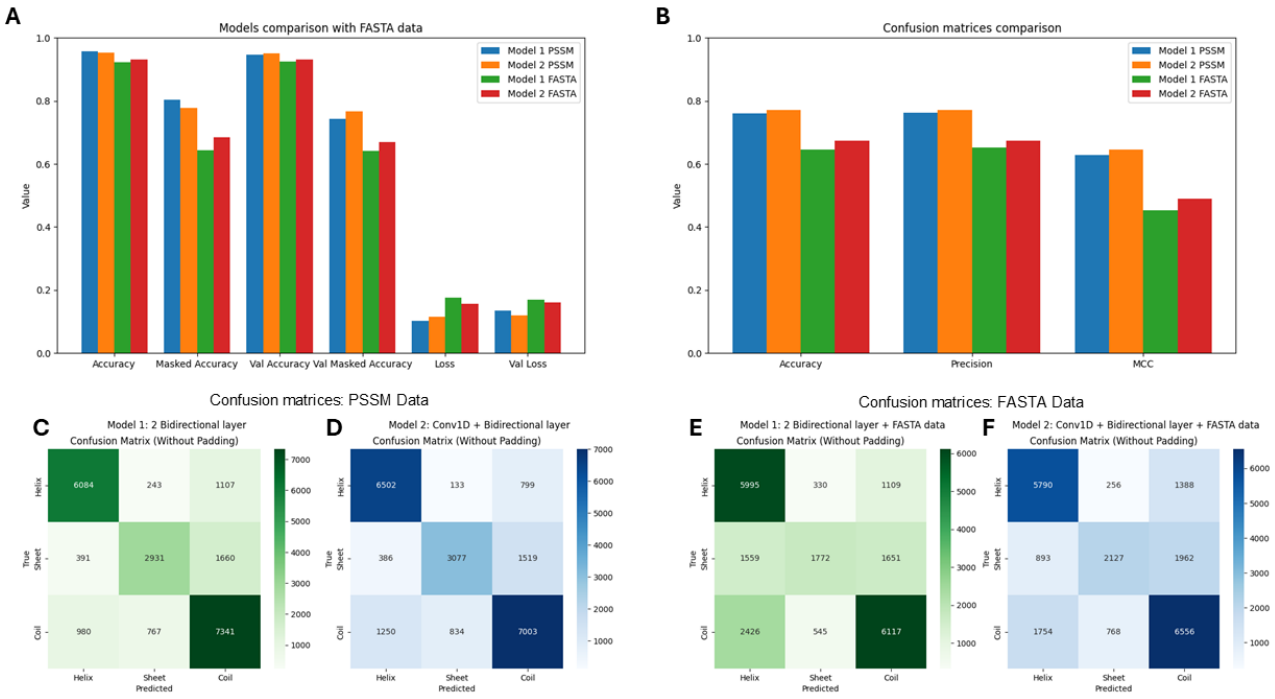


Fig. 3. Metrics comparison of model 1 and 2. A. Comparison of accuracy, masked accuracy and loss in the training and validation set. B. Accuracy, precision and MCC metrics extracted from the test set. B-F. Confusion matrices of model 1 (C and E) and model 2 (D and F) in both PSSM (C-D) and FASTA data (E-F).

Both models showed better performance with PSSM data than with FASTA, which was expected provided that PSSM also includes evolutionary data. Metric given by model 1 and model suggests that both models have a similar performance, yielding a ~70% masked accuracy (Fig. 3A). In the same way, the confusion matrices (Fig. 3C-F) and their metrics (Fig. 3B) looked very similar for both models, being coil and helix the most predicted secondary structures. Nonetheless, provided that the metrics from model 2 are slightly better and its lower computational cost lower, it was used to perform ablation studies.

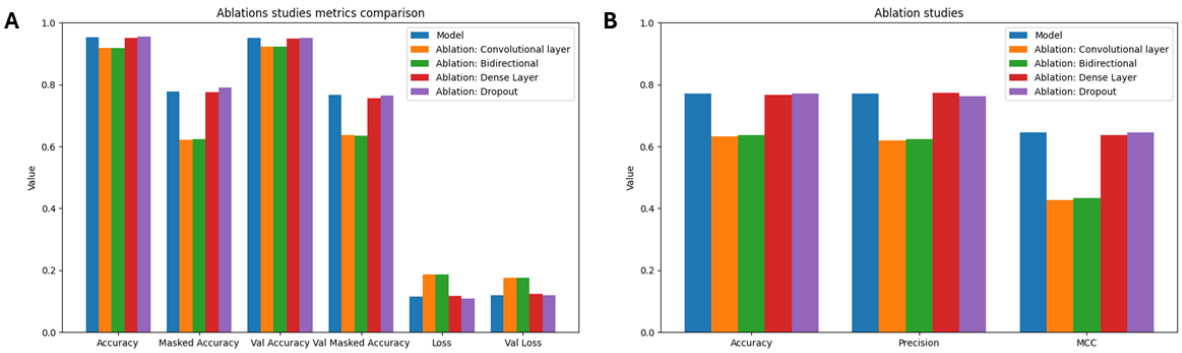


Fig. 4. Comparison of the metrics obtained in ablation studies compared to the original model (in blue). A. Metrics of the validation and training sets. B. Metrics of the test set based on confusion matrix.

The results from the ablation studies show a decrease in accuracy, precision and MCC, and an increase in loss in the models lacking the bidirectional and the convolutional layer (Fig 4A-B), marking its importance in the model performance. On the other hand, the ablation of the dense layer and the dropout regularization did not undermine the model performance significantly.

Discussion

The model that combines convolutional and LTSM bidirectional layers achieved a better accuracy and reduced the computational cost (from 333,956 trainable parameters to 87,108). The first convolutional network searches for short patterns of amino acids (kernel= 3), which can help to the identification of α-helix motifs as they interact with amino acids 3.6 positions forward. Then, the bidirectional LTSM layer allows the identification of long-range dependencies thanks to its memory cell. These patterns are also importance because amino acids in β-sheets can interact with other amino acids that are further away in the sequence or because α-helices can extend for many residues. The importance of these layers was demonstrated through ablation studies (Fig. 4), as their removal negatively impacted the model's performance

The addition of the dummy class for padding the label data created problems when analysing the accuracy and the confusion matrix data. The high number of outputs with the padded label diluted the data, leading to inflated metrics that did not reflect the actual model performance (Supl. Fig. 1). This was corrected creating a personalised metric and excluding the fourth class from the confusion matrix. A future implementation would be trying to mask the data inside the model, so it does not learn to classify padded data.

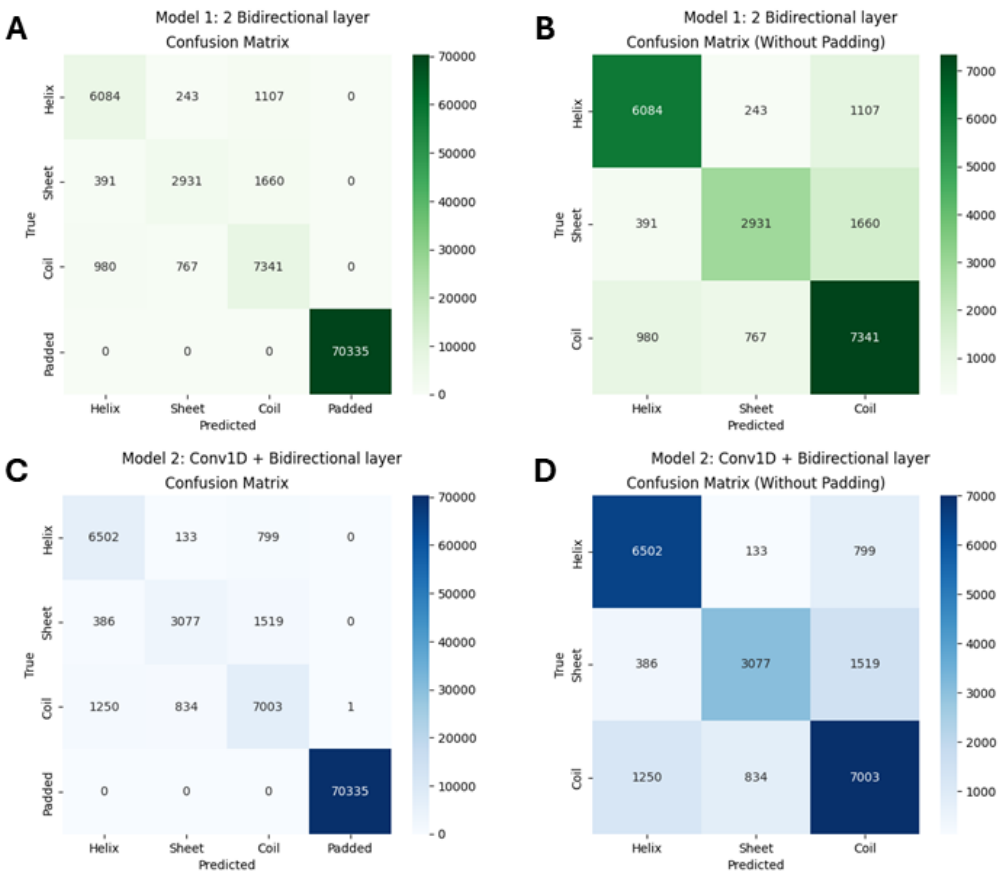
Acknowledgements

I acknowledge the use of ChatGPT [https://chatgpt.com/] to help me write and correct the code. I have revised it and I take full responsibility for it.

References

Singh, J., Litfin, T., Paliwal, K., Singh, J., Hanumanthappa, A. K., & Zhou, Y. (2021). SPOT-1D-Single: improving the single-sequence-based prediction of protein secondary structure, backbone angles, solvent accessibility and half-sphere exposures using a large training set and ensembled deep learning. *Bioinformatics*, *37*(20), 3464–3472. https://doi.org/10.1093/bioinformatics/btab316

Srivastava, G., Liu, M., Ni, X., Pu, L., Brylinski, M. (2025). Machine Learning Techniques to Infer Protein Structure and Function from Sequences: A Comprehensive Review. In: Kloczkowski, A., Kurgan, L., Faraggi, E. (eds) Prediction of Protein Secondary Structure. Methods in Molecular Biology, vol 2867. Humana, New York, NY. https://doi.org/10.1007/978-1-0716-4196-5_5



Suplementary figure 1. The addition of a 'padded' class dilluted the data giving unreliable metrics. Confusion matrices without the elimination of 'padded class' (A and C) vs. with a 3-class representation (B and D)