Predicting Protein Secondary Structure Elements Using Dense Neural Networks

Introduction

Understanding the relationship between structure and functions of proteins has been a long time an endeavor of structural biology. Early advancements in resolving the structure of proteins lied within experimental approaches. However, their resource inefficiency together with the increasing availability of protein sequences led to the rise of computational approaches for protein structure prediction [1]. One crucial part of the protein structure are secondary structure elements, which are formed by hydrogen bonds between backbone molecules. One broad classification of those elements is in the so called-three-state model, which distinguishes α-helix, β-sheets, and coils (Fig. 1).

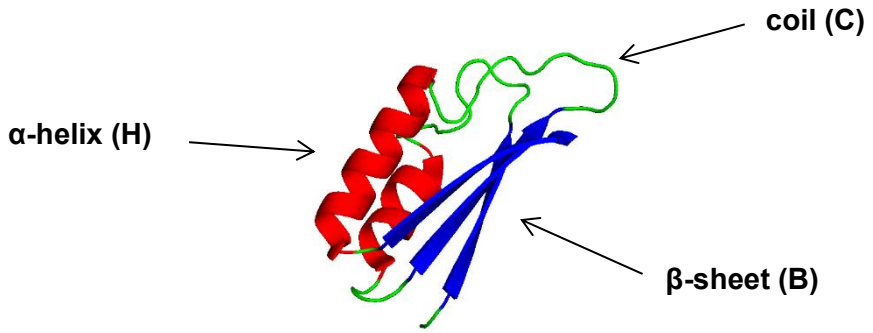


Figure 1: Secondary structure elements can be divided in alpha-helix (red), β-sheets (blue), and coil (green). (Adapted from PDB ID: 3GC6 [2])

While secondary structure predictors originally focused solely on the analyzing amino acid propensities around a "window" of central residue, the inclusion of evolutionary information in the models significantly increased their performance [3]. Especially applying neural networks, including recurrent (RNN), convolutional (CNN) and fully connected fully dense networks, increased prediction accuracies [4]. However, selecting the correct model architecture and hyperparameters is crucial for their efficiency.
That is why in this paper we sought to construct a secondary structure predictor based on fully connected network and test the influence of hyperparameters on its performance.

Methods

Data

For training and evaluating the model, training, cross validation (cv) test and blind test data were provided as PSSM files and files with assigned secondary structure (dssp). The whole data was utilized as described by Elez[5]. The cv split divided the data into 5 cv models that were each trained and evaluated for each model.

Model

In total five models (model_default, model_reg, model_tanh, model_extra_layers (Fig. 2), model_less_layers) were compared for window sizes 11, 13, 15 and 17 and the predictive performance was assessed. An outline of all models can be found in Fig. 2 and the Supplementary Material.

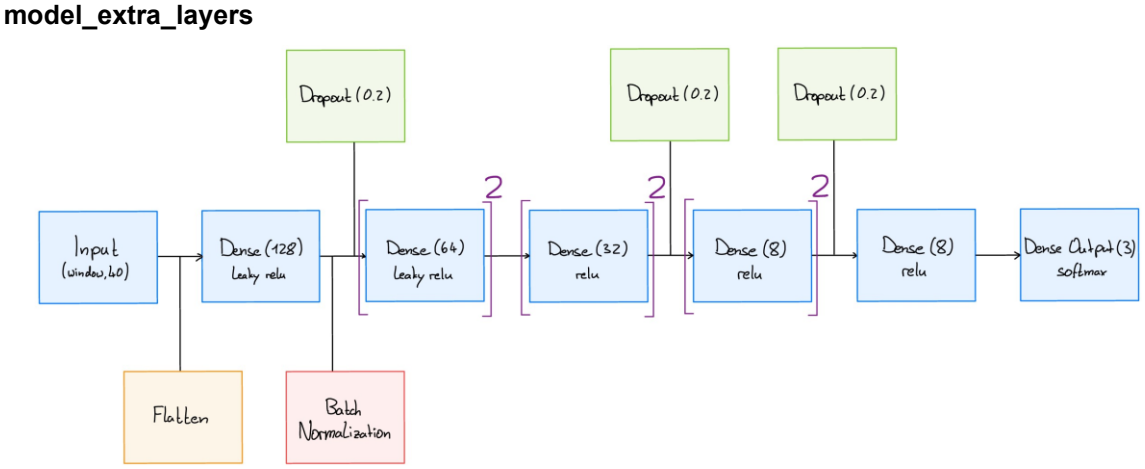


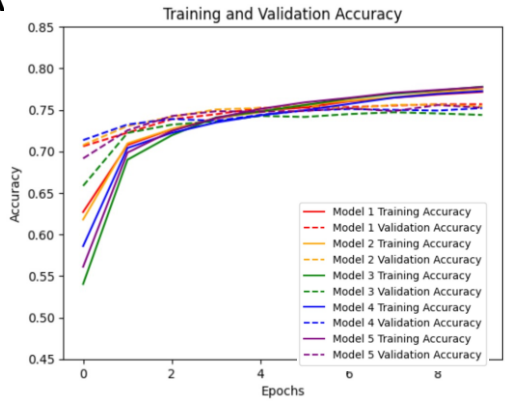
Figure 2: The neural network architecture for the model_extra_layers can be used for prediction. The superscript two indicates that the neural block is present twice in succession.

Results

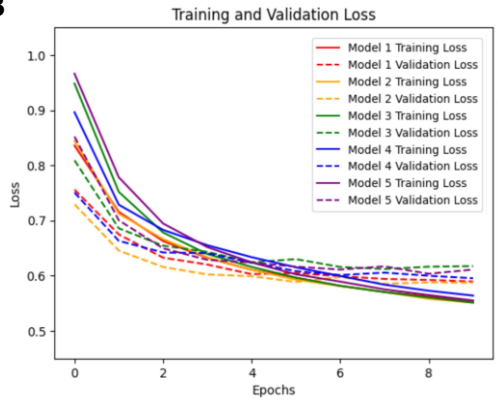
a) The best performing model is the model_extra_layers with window size 17

A range of models with different hyperparameters were tested for their predictive performance. The model model_extra_layers (Fig. 2) with window size 17 achieved the highest accuracy on the blind test data with 74.4% for the cross validation (cv) model 3. The prediction accuracies on the training and validation set (Fig. 3A) ranged from 74 to 78% (Fig. 3D). However, for epochs higher than 6, the validation accuracy did not increase, hinting to potential overfitting for the total epoch number of 10.

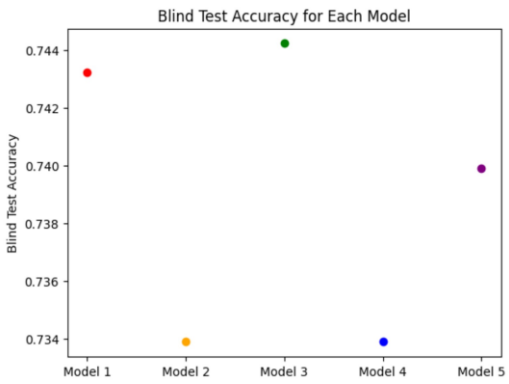
A



B



C



D

	Accuracy		
	Training	CV test	Blind test
CV model 1	0.772	0.757	0.743
CV model 2	0.775	0.754	0.734
CV model 3	**0.778**	**0.744**	**0.744**
CV model 4	0.7773	0.752	0.734
CV model 5	0.777	0.752	0.740

Figure 3: Model_extra_layers showed highest performance. A) Training and validation accuracy and (B) loss over 10 epochs. C) Accuracy on the blind test for all cv models. D) Summary of the accuracies of all cv models. The highest performing model is marked in green.

The same increasing divergence between the sets was found when analyzing the loss over the epochs in Fig. 3B. Furthermore, Fig. 3C illustrates that there was a high prediction accuracy on the blind test data of the cv model 1, 3 and 5 (around 74.2%), while it was slightly lower for the cv models 2 and 4 (around 73.4%). Fig. 3D summarizes all the performances and shows that the accuracies decreased about 1-2% when comparing the training to the blind test accuracy.

b) Ablation study shows influence of hyperparameter choice

The identification of the highest scoring model was part of a wider ablation study which compared five different model architectures (Fig. 2, Fig. S1) for four different window sizes. The model architectures were constructed in a way to allow for the analysis of the impact of added regularization, applied tanh activation functions and changes in the model complexity through variations in the layer number on a default model.
Table 1 shows the cv model that scored the highest on the blind test set for all parameter variations. It was notable that when increasing the window size from 11 to 17 for the input shape, the accuracies increased about 0.3 – 0.9%, independently from the chosen model architecture. Moreover, one could observe that additional L2 regularization to the default model reduced the accuracy about 1.8 – 2 %. Similarly, applying a tanh activation function instead of relu functions decreased the accuracy, however, to a smaller extent (0.4 – 1%).
Interestingly, changing the number of layers did not have significant impact on the maximum blind test accuracy.

Table 1: Maximum prediction accuracy on the blind test set for different model architectures and window sizes depends on the chosen hyperparameters. The brackets indicate the cv model that the maximum accuracy originated from. The highest performing model is marked in green.

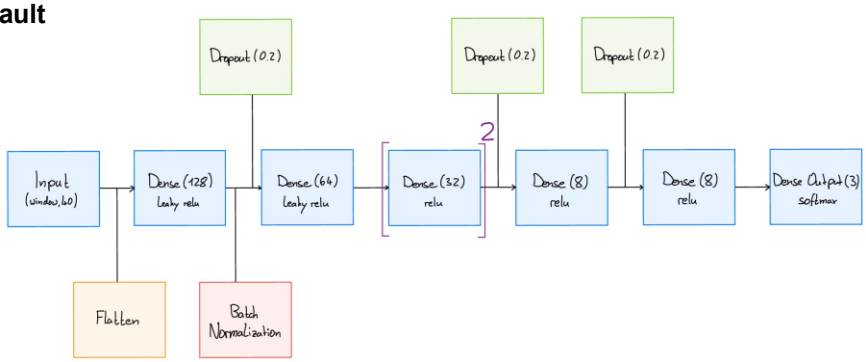
	Maximum Blind Test Accuracy for different window sizes			
	Window 11	**Window 13**	**Window 15**	**Window 17**
model_default	0.739 (CV mod 3)	0.738 (CV mod 4)	0.743 (CV mod 3)	0.742 (CV mod 2)
model_reg	0.713 (CV mod 2)	0.720 (CV mod 5)	0.723 (CV mod 1)	0.724 (CV mod 1)
model_tanh	0.729 (CV mod 4)	0.734 (CV mod 3)	0.736 (CV mod 3)	0.734 (CV mod 3)
model_extra_layers	0.737 (CV mod 3)	0.738 (CV mod 1)	0.742 (CV mod 1)	**0.744** (CV mod 3)
model_less_layers	0.736 (CV mod 3)	0.738 (CV mod 5)	0.741 (CV mod 3)	0.741 (CV mod 3)

Discussion

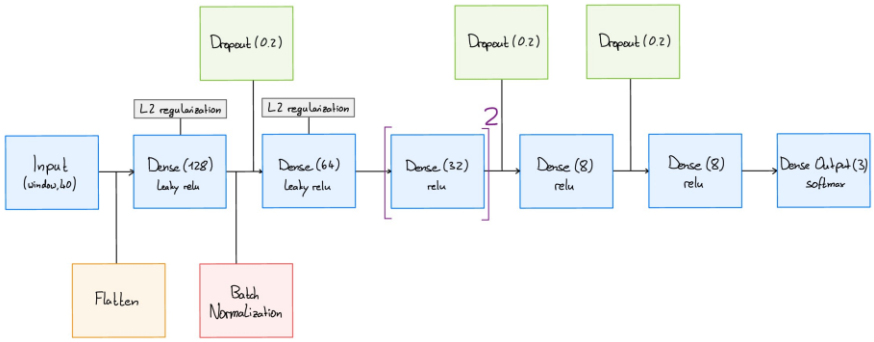
We successfully constructed several model architectures with accuracies more than 70%, with the highest reaching 74%. The accuracy is lower than in more advanced approaches with around 80% [6], but is still high enough to infer that the windows can capture some contextual dependencies. Though, RNNs and CNNs potentially could account for wider dependencies [4].
Nevertheless, it needs to be noted that the highest scoring model was slightly overfitted (Fig. 3B), which could be attributed to its higher complexity through the additional layers, which calls for additional regularization techniques and drop out. Interestingly, most parameter changes did not severely impact the models' performance (maximal 2% difference in accuracy), which could be due to the simple architecture. The reduction in accuracy when applying the tanh activation function indicates potential saturation. However, multiple runs of the networks including statistical tests are needed to assess the significancy of the results. Additionally, increasing the epochs could improve the accuracy on the blind test set for models where no overfitting occurred.
All in all, our investigations and ablation study highlight the sensibility of neural networks to overfitting and emphasize that more extensive hyperparameters searches are required for potential model improvement.

Supplementary material

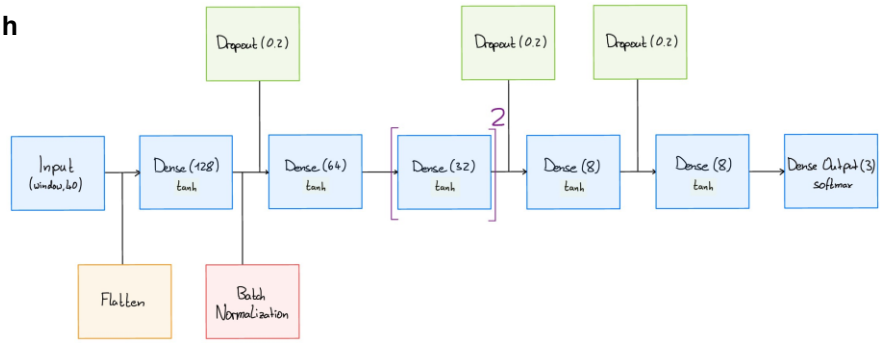
model_default



model_reg



model_tanh



model_less_layers

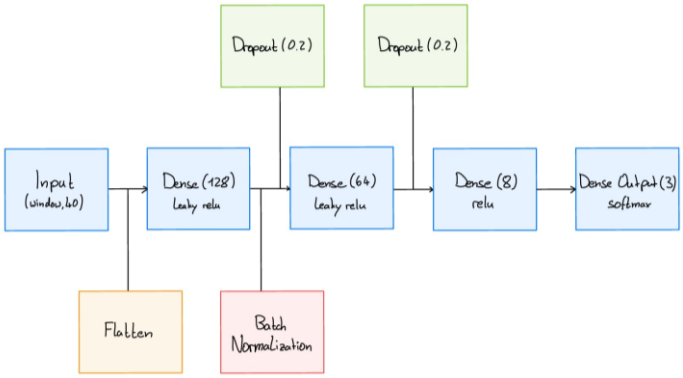


Figure S1: Different model architectures were used for the ablation study.

References

1 Li, D., Li, T., Cong, P., Xiong, W. & Sun, J. A novel structural position-specific scoring matrix for the prediction of protein secondary structures. *Bioinformatics* **28**, 32-39 (2012).

2 Egea, P. F. *et al.* Insights into the mechanism of bovine CD38/NAD+ glycohydrolase from the X-ray structures of its Michaelis complex and covalently-trapped intermediates. *PLoS One* **7**, e34918 (2012).

3 Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol* **292**, 195-202 (1999). https://doi.org:10.1006/jmbi.1999.3091

4 Ismi, D. P., Pulungan, R. & Afiahayati. Deep learning for protein secondary structure prediction: Pre and post-AlphaFold. *Comput Struct Biotechnol J* **20**, 6271-6286 (2022). https://doi.org:10.1016/j.csbj.2022.11.012

5 Elez, K. (2018). Predicting secondary structure of proteins: a comparison between GOR method and Support Vector Machines (Project Report No. 2). Laboratory of Bioinformatics, University of Bologna. Retrieved from https://github.com/katarinaelez/protein-ss-pred

6 Kulikova, A. V. *et al.* Two sequence-and two structure-based ML models have learned different aspects of protein biochemistry. *Scientific Reports* **13**, 13280 (2023).