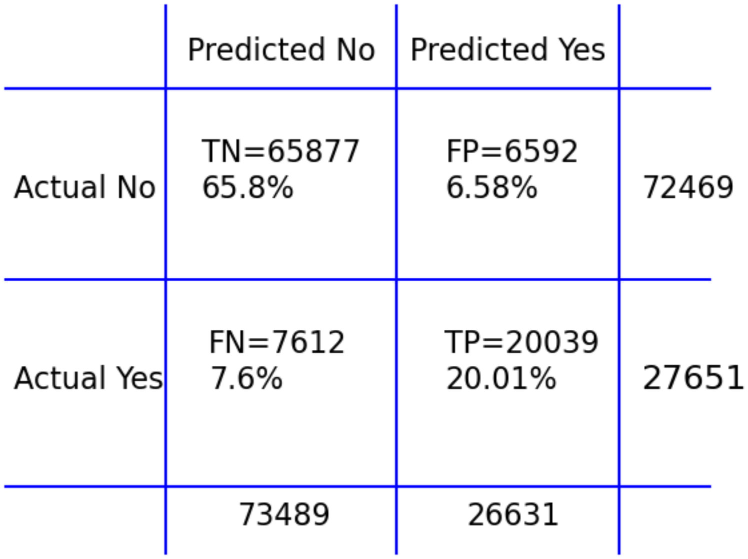
4 Results

For the testing of the model we used 2000 samples from the HMMDB as a training group, and 400 samples as test group. After training we used the model to predict the Hidden states of the test group, and measured the results using different parameters.



	Predicted No	Predicted Yes	
Actual No	TN=65877 65.8%	FP=6592 6.58%	72469
Actual Yes	FN=7612 7.6%	TP=20039 20.01%	27651
	73489	26631	

Confusion matrix, representing the tagging of each amino acid in each sequence in the test sequences. Where Positive/Yes represents an amino acid being inside the membrane (motif), and Negative/No means outside of the membrane. The percentile represents the respective value divided by the overall amino acids.

Testing methods

Avaraged Match Rate:

We wanted to find the correct match rate for different assignments $\left(\frac{\text{correct labeling}}{\text{overall labeling}}\right)$. since we saw that our overall success rate was 85% (fig matrix) we decided to test the match percentile of each sequence, and use them to determine if there are specific parameters that affect our sucess rate.

We decided to test the match rate **relative** to the length of a protein, and to the number of motifs, while most sequences were matched rather successfuly (mostly above 80%), there are inconsistencies by the different parameters, especially noticeable in the erratic changes when measuring by sequence length. We assume that the factor that causes it is either not among those we tested, or is to complex to predict using our selected model. We did notice however that shorter sequences (below 600 amino acids) can be more erratic than longer sequnces , though it might be caused by the fact that shorter sequences are more common, and suprisingly the long sequences and those with a large amount of motifs had similar of not better success rate.

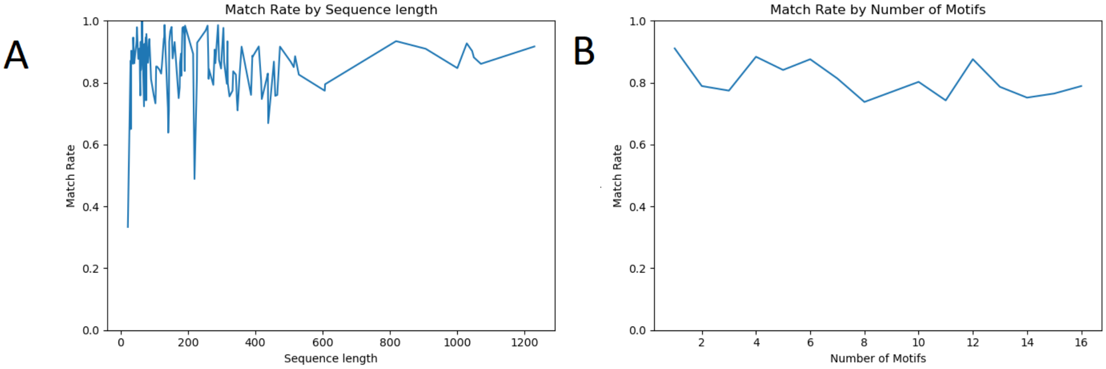


fig X, measuring the match success rate averaged for different parameters (meaning when multiple results have the same length/ number of motifs, they will be averaged to a single result). A: Measuring the match rate relative to the length of a sequence. B: Measuring the match rate relative to the number of transmembrane motifs.

False negative rate:

Next we decided to determine how likely we are to produce False Positive results relative to the same parameters. For each labeled sequence we counted the amount of times we labled a background/outside acid as a motif/inside acid. Because a longer sequence has more labels that could be wrong, we tested the relationship between the length of the sqeunece and the number of motifs, to the number of FP assignments, and the rate of FP assignments normalized by the length of the sequence.

The results were quite similar to those of the match percentile, while longer sequences expectedly had more overall FP lables, when the number of FP was normalized by the length of the sequence, the rate of FP was less erratic for smaller proteins, and similar with proteins with different number of motifs.

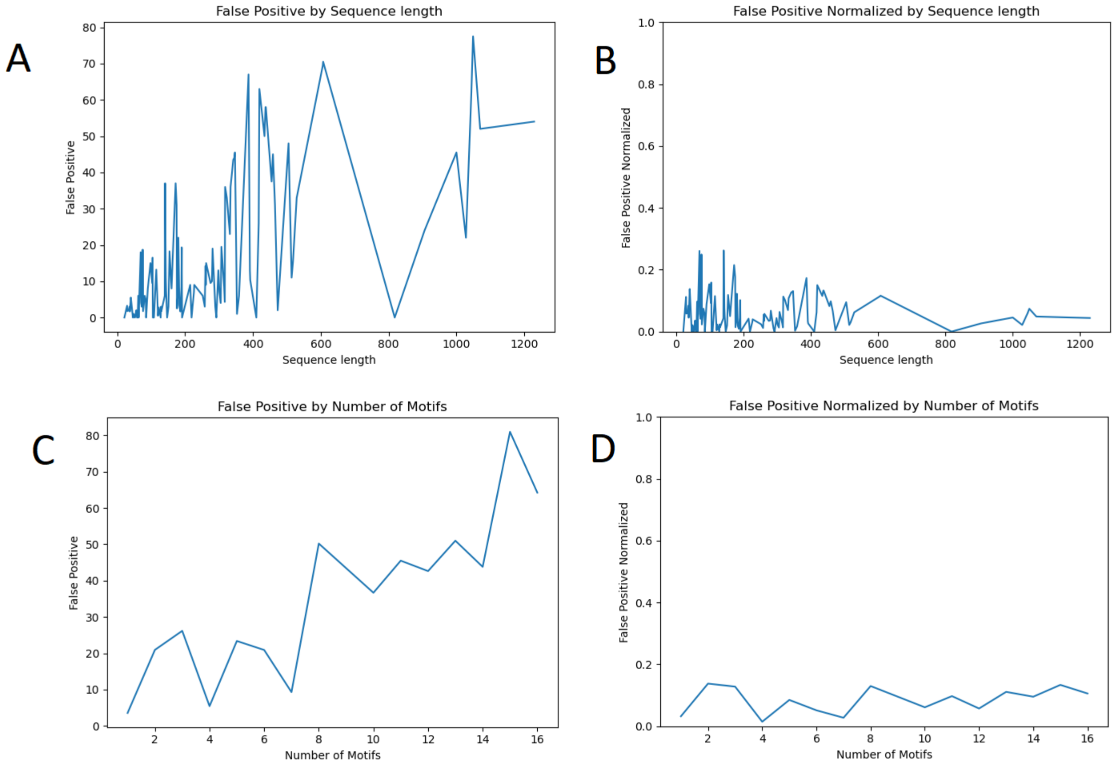


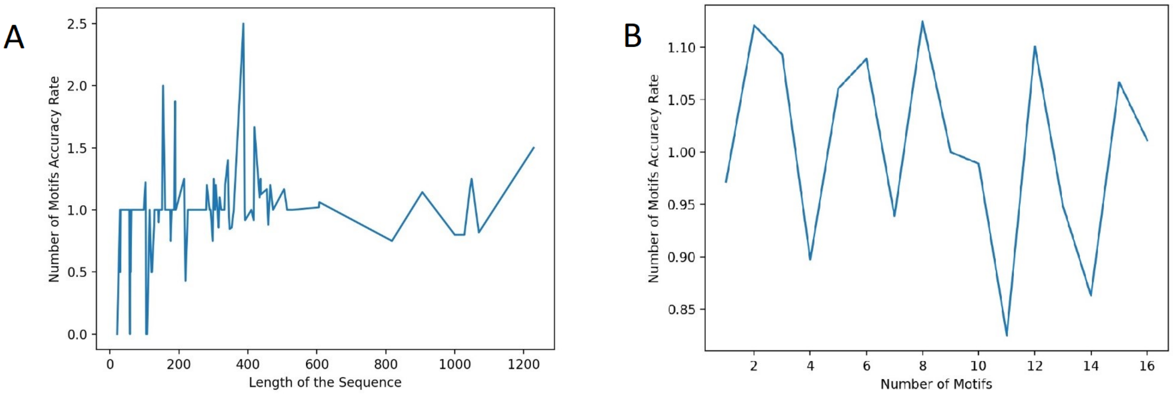
fig X+1, measuring the overall FP labels, and the FP lable rate averaged for different parameters (meaning when multiple results have the same length/ number of motifs, they will be averaged to a single result). A: Measuring the overall FP relative to sequence length. B: Measuring the FP rate relative to sequence length. C: Measuring the overall FP relative to number of motifs. D: Measuring the FP rate relative to the number of motifs

Number of Motifs Accuracy:

Since the number of times the protein crosses the membrane can have a massive effect on the protein structure, we decided to test how accurately we are able to predict the number of transmembrane motifs for a specific sequence (meaning, how many times a specific protein will cross the memrane).

As before, the results showed greater variation for shorter sequneces when measured by length, and relatively lesser variation when measured by number of motifs.

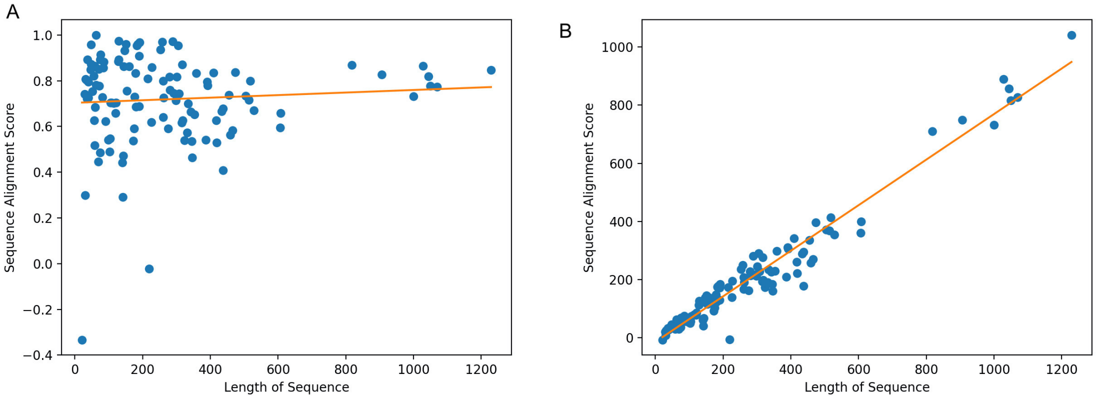
we



A: The ratio between predicted number of motifs and the true number of motifs, for each sequence length. B: The ratio between predicted number of motifs and the true number of motifs, for each number of motifs.

Sequence Alignment:

To get another measure of how similar our predicted sequences were to the actual ones, we preformed a sequence alignment test on our predicted sequences. This test gives us a better indication of the predicted general structure of the proteins and not just the prediction of the state of each amino acid. We constructed a score matrix as such: 1 point for match, -1 for mismatch, and -2 for gap. Because a longer sequence has more instances that could be wrong, we examined the relationship between the average sequence alignment score (and the same score normalized by the length of the sequence) and the sequence length. The model was able to predict well more than half of the sequence for most lengths. For a decent number of lengths, the model did very well and predicted accurately most of the sequence, with a several nearly perfect scores. This indicates that the model was able to identify and assign quite well the appropriate states.



A : The average sequence alignment score normalized by the length of the sequence, for each sequence length. B: The average sequence alignment score for each sequence length.

5 Conclusions

In this project we wanted to create a model that predicts alpha helix transmembranal domains for specific protein sequence. while most of our predictions were reletively accurate, there are places where the model is a bit lacking. In all tested parameters we found that shorter sequences are more likely to show erratic behavior, while it could be caused by lack of data, it is also possible that shorter sequences show a less predictable behavior.

It should also be noted that to make the model less complex, we elected to only predict alpha-helix transmembranal regions. While simpler, it might also reduce our ability for prediction, as the existance of multiple distinct **models** tagged as background might hinder our ability to determine what is a background.

If given more time we would have expended our model to **include different motifs, and maybe create hidden states that are more specilized to specific motifs.**