Predicting protein subcellular localisation using Deep Learning

Introduction

Proteins makeup the building blocks inside the cell, responsible for various functions including transportation, catalysing reactions, providing structural support etc. Each protein is targeted to a particular subcellular location where it performs this function. This is achieved through specific amino acid sequences that act as signalling sequences. Mislocalisation of proteins often result in disease phenotypes including cancer, neurodegenerative diseases etc. Accurate prediction of protein localisation from its amino acid sequence can be vital in understanding cellular mechanisms and developing tools/drugs that targets specific subcellular locations.

While experimental techniques like fluorescence microscopy and mass spectrometry-based methods are used to study subcellular localisation, they are not scalable to the entire proteome of the organism. Thus, using computational methods to decipher the complex patterns that determine the localisation becomes the better alternative. Earlier iterations relied on manually extracted features such as know sequence motifs and known signal peptides. However, complex patterns in the sequences are lost in this method. Recent advancements in protein language models have enabled the automatic learning of high-dimensional, context-aware sequence representations.

In this project, we used three different embedders, ESM-2 (Evolutionary scale modelling), ProtBERT and ProtT5. All three are protein language models with varying levels of complexity. ESM-2, developed by Meta AI, consists of 6 transformer layers and 8 milion trainable parameters, which offers speed while compromising on accuracy. ProtBERT is also a transformer-based model with 30 layers and 420 million parameters. ProtT5 is the most complex of the three with 48 layers (encoder and decoder with 24 layers each) and ~3 billion parameters. We checked the embedder performance and speed on a sample dataset to identify the embedder that works well with the computational resources at our disposal. With the selected model, we created and trained a simple feed-forward neural network to assess it's performance in predicting subcellular localisation of the sequences.

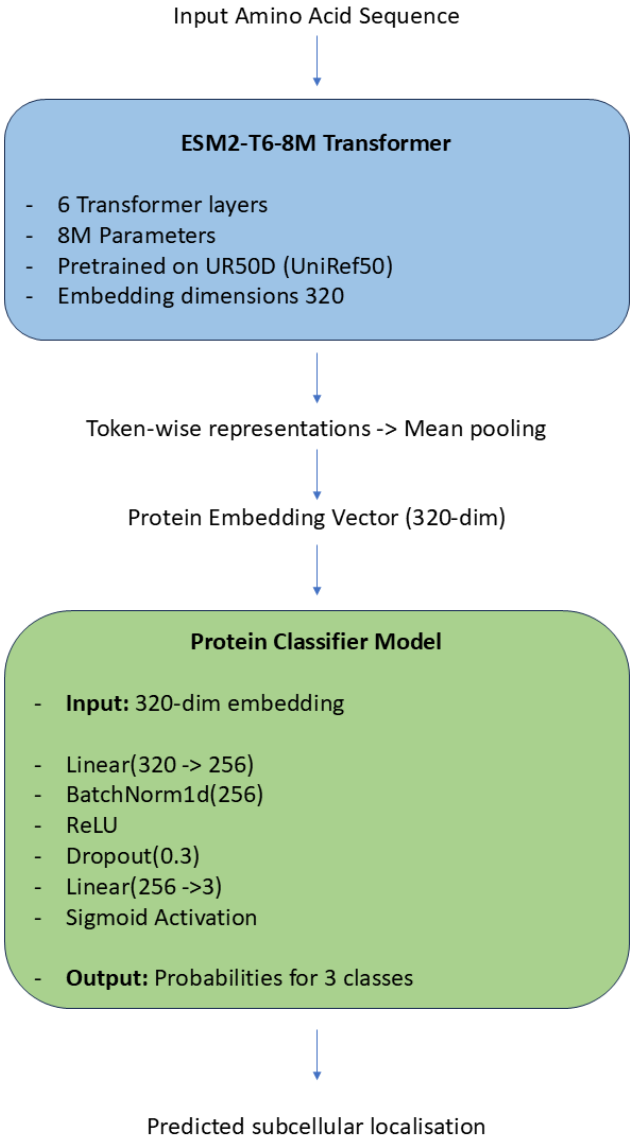

Figure 1 Model architecture for the Protein classifier

Results

Preparing the data

We used the Uniprot database to procure all the reviewed human proteins with their sequences and subcellular localisation. However, the subcellular localisation column was not ready-to-use for the models. So, we filtered the dataset such that binary indicator columns (1/0) were generated for each location header (Nucleus, Cytoplasm, Secreted) indicating their presence/absence respectively. We only selected Nucleus, Cytoplasm and secreted as we believe these three compartments offer less overlaps as opposed to other proteins. We also only considered sequences less than 500 amino acids for this preliminary model. This was done as

larger datasets couldn't be processed with the resources provided by the free version of Colab. This modified data was used to test the embedders and then create the primary model.

Testing different embedders

To test the effectiveness of different protein language models, we sample a stratified subset of sequences using multilabel stratification to preserve the distribution of localisation labels. We used 0.6% of the total dataset (43 sequences) to account for the computational power of our device. Three pre-trained protein embedding models- ESM-T6-8M, ProtBERT and ProtT5-XL were employed to generate fixed-size representations for each sequence. ESM2 embeddings were obtained using mean pooling over token representations, while ProtBERT and ProtT5 embeddings were generated with their respective tokenizers, followed by attention-masked mean pooling. As illustrated here, the ESM2 embedder far surpassed the rest, completing in dramatically less than a minute. ProtBERT, the middle-sized model, took a few minutes to work on the same data. ProtT5-XL took longest by far, taking more than a quarter of an hour to accomplish what took ESM2 less than a minute. Such vast disparity in speed underscores how model complexity and size directly affects throughput. Practically, employing ProtT5-XL would translate to extremely longer prediction waiting times, while ESM2 provides nearly instantaneous generation of embeddings for small batches. How much could be batch-processed also varied: The small size of ESM2 permitted us to embed multiple sequences simultaneously, whereas the larger models essentially processed one sequence at a time (their size limiting batching), again adding to ProtT5's long runtime.

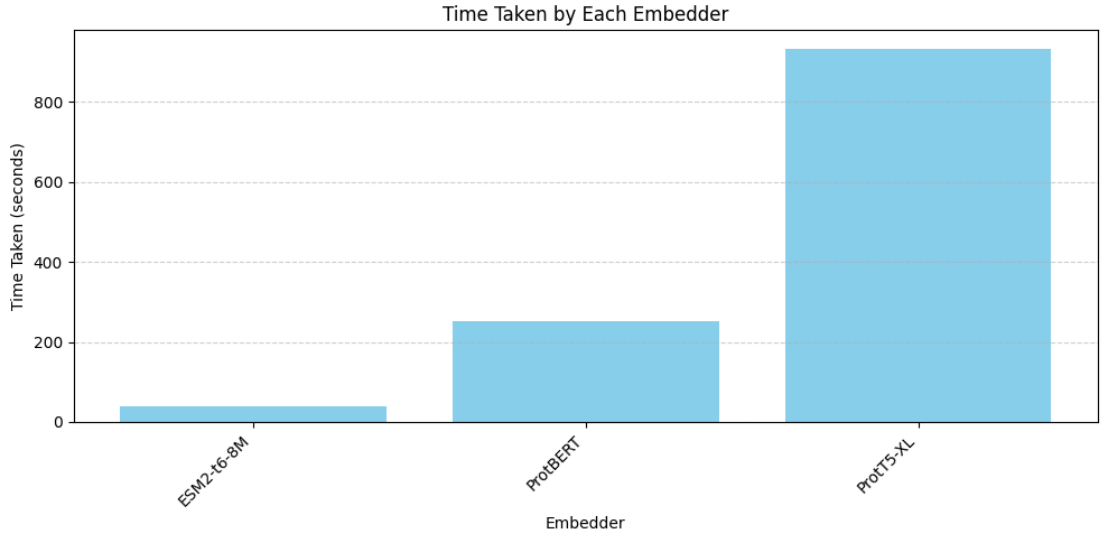

Figure 2 Time taken by each embedder

These embeddings were then used to train a feedforward neural network using weighted cross-entropy loss to account for weight imbalance. The model consisted of a linear input layer, ReLU activation, dropout for regularisation and a final linear output layer. The dataset was split with stratification and model performance was evaluated using accuracy, weighted precision, recallm F1-score and ROC-AUC curve. ROC curves were plotted for each embedder to visualise their capability.

Model	Accuracy	Weighted Precision	Weighted Recall	Weighted F1-score	Weighted AUC
ESM-T6-8M	0.44	0.4	0.44	0.42	0.67
ProtBERT	0.66	0.74	0.66	0.64	0.79
ProtT5-XL	0.77	0.81	0.77	0.77	0.87

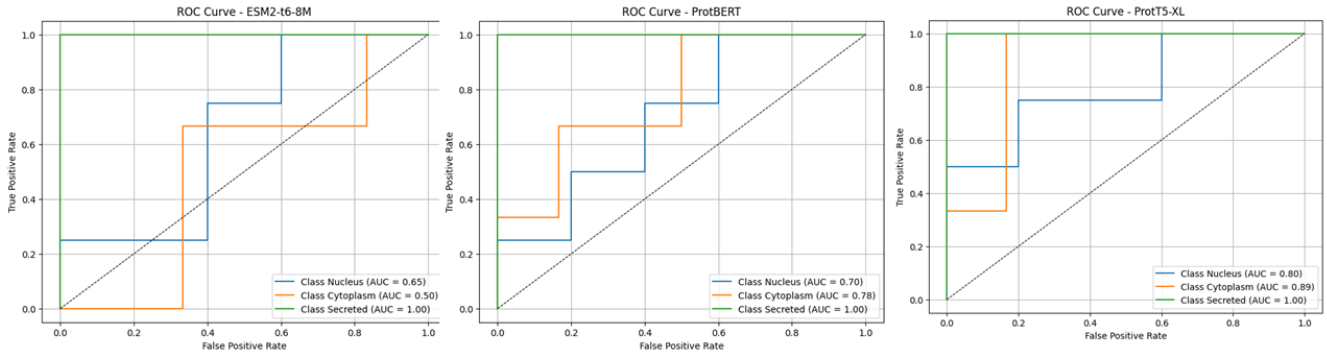

Figure 3 ROC curves for the three models

Aside from speed and memory, we looked at how each embedder affects model accuracy at the task of localization prediction. Once embeddings were taken from each model, a simple classifier was trained to predict one of three localization labels (Nucleus, Cytoplasm, Secreted) for the test sample. In spite of the tiny test set, the trend remained evident: bigger embedders provided better predictive performance. With the use of ESM2 embeddings, the classifier was only 44.4% accurate on the held-out test (getting 4 out of 9 examples right). Accuracy rose to 66.7% with ProtBERT embeddings and to 77.8% with ProtT5-XL embeddings. That is, the ProtT5-XL derived features resulted in nearly twice the accuracy of ESM2's in this sample. Precision and recall went in the same direction: weighted precision, for instance, went from 40.0% with ESM2 to 74.8% with ProtBERT, to as high as 81.5% with ProtT5-XL, suggesting that the bigger models not just made more correct predictions, but also were more stable and certain in making these predictions. That makes intuitive sense – enormous ProtT5-XL probably preserves more biologically relevant signal (protein motifs, evolutionary context, etc.), providing the downstream classifier with a better foundation to separate localization signals, while ESM2's less detailed embeddings lose some of that. ProtBERT's results fell in between, far improved over ESM2 but not quite up to ProtT5's mark. These findings emphasize that accuracy improves in proportion to embedder model size for this assignment, though tested at a small level.

We went with ESM2, despite its shortcomings in prediction accuracy owing to the computational resource constraints we had.

Protein Classifier

Protein sequences were annotated with subcellular localisation labels – Nucleus, Cytoplasm and Secreted – were embedded using the ESM2-T6-8M transformer model. Fixed-length sequence embeddings were obtained by applying mean-pooling over the per-residue

representations generated by the model. To asses the intrinsic structure of the embedded data, we applied Uniform Manifold Approximation and Projection (UMAP) for dimensionality reduction. From figure 4, we see that considerable overlap is observed between the classes, reflecting the complex multi-class nature of the dataset. The decision to choose the labels was driven by the relatively lower degree between them. A silhouette score of 0.34 was obtained, indicating modest but acceptable separation in the embedding space.

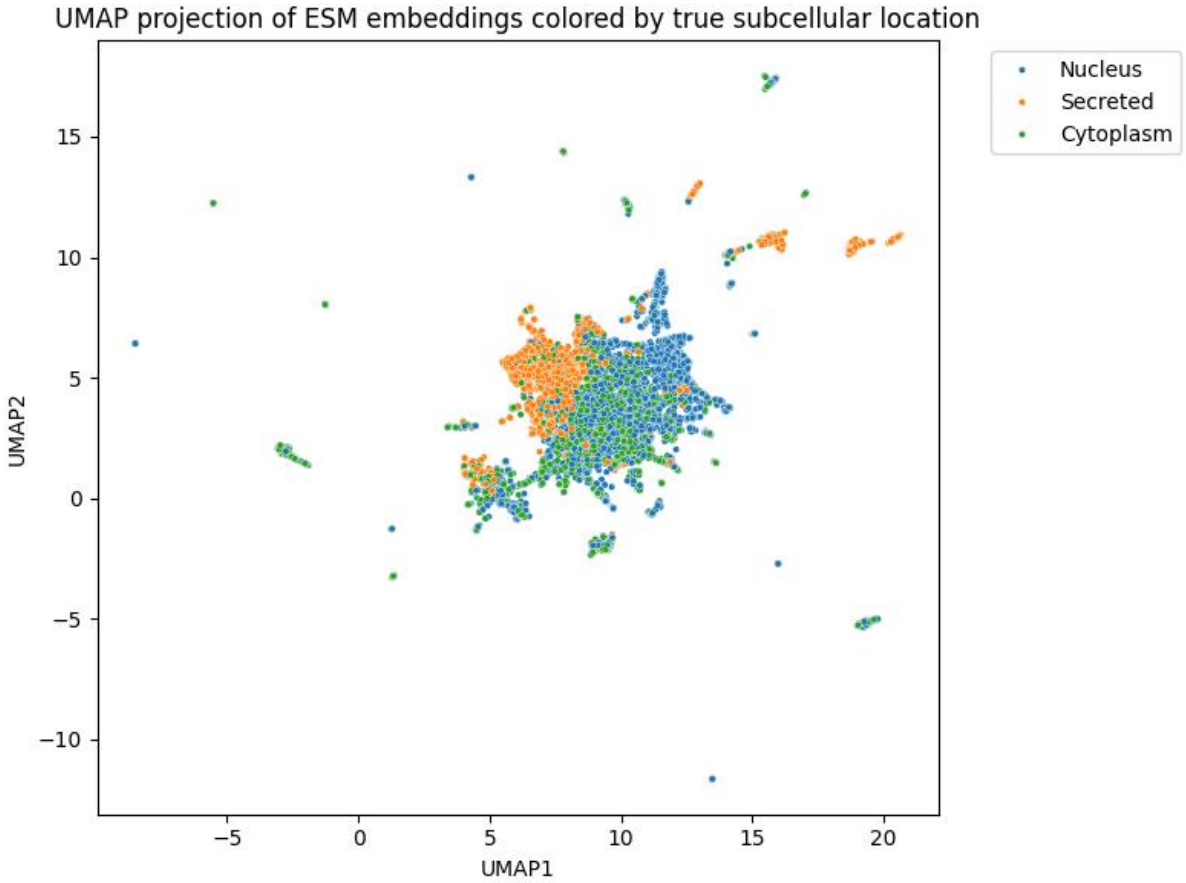

Figure 4 UMAP projection of the ESM embeddings

These embeddings were subsequently fed into a simple feed-forward neural network with one hidden layer and dropout for regularisation for classification. The model was trained using binary cross-entropy loss with class-specific positive weights to account for class imbalance. The model was then evaluated using metrics like F1-score, accuracy and ROC curve. The best performing model, using validation loss, achieved reasonable classification performances. However, the performance is dampened by the nature of dataset, as we are using just a binary indicator for each location. The precision-recall curves and ROC curves indicate this ambiguity presented by the overlapping target classes for each protein.

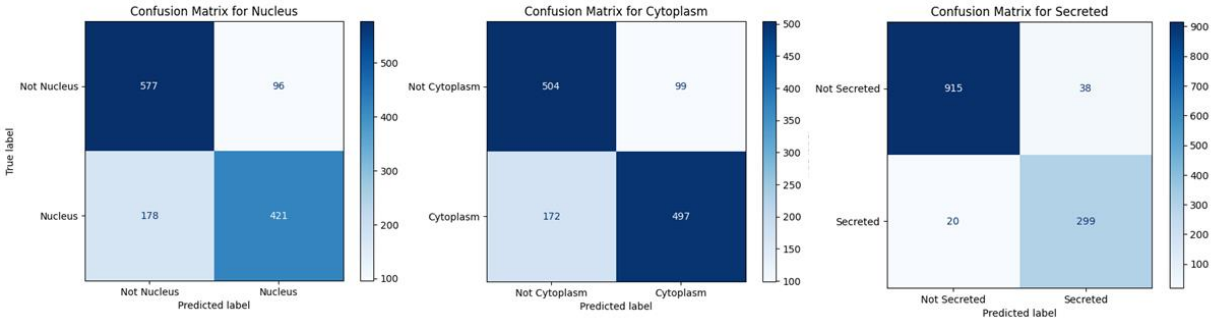

Figure 5 Confusion Matrices for the three locations

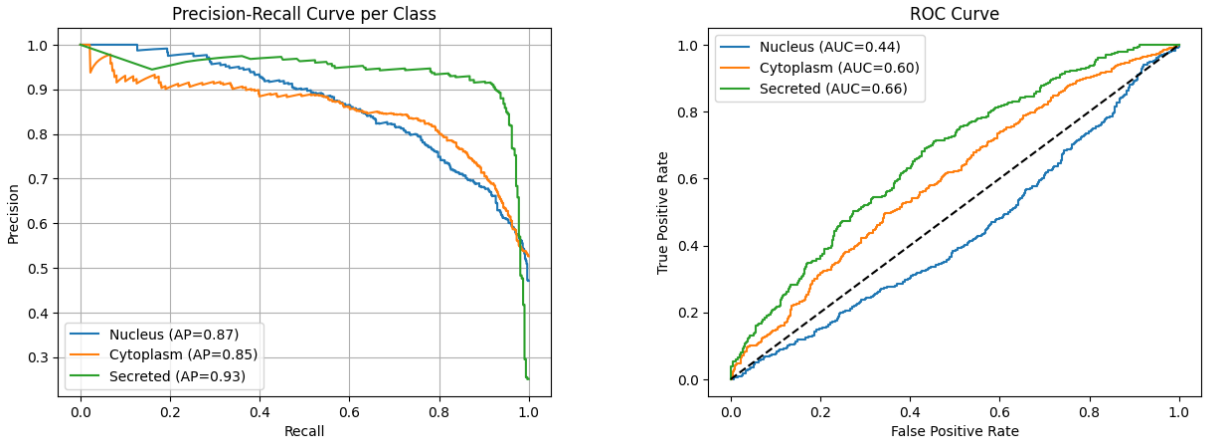

Figure 6 Precision-Recall and ROC curves for each location

Secreted proteins show the highest classification performance with an average precision (AP) of 0.93. Nucleus and Cytoplasm performed reasonably good, albeit with slightly lower AP values of 0.87 and 0.85 respectively. The secreted proteins perform best with an area under the curve (AUC) of 0.66 while the AUC scores of Cytoplasm and Nucleus are 0.6 and 0.44. This suggests, that despite high precision-recall performance, the classifier struggles to clearly separate true and false positives. This discrepancy can be attribute to class imbalance and partial label overlap where the binary nature of the labels might not fully capture the multi-localisation tendencies of certain proteins.

Conclusions and Future Directions

Although the model captured some structural information, the model's performance was limited due to the multi-class nature of the problem. This introduced ambiguity, as the classifier struggled to distinguish mutually exclusive classes when trained with overlapping binary targets. Furthermore, the use of a relatively small and shallow model might have restricted the model's capacity to learn complex patterns. In the future iterations, we plan to do the following:

- Reframe the task as a true multiclass classification problem, where each protein is assigned to exactly one class. We could try to use probabilities for each protein and their target location. This would allow the model to learn more distinct class boundaries with a softmax activation and cross-entropy loss.
- Experiment with larger or more expressive protein language models like ESM-1b, ProtT5 or ProtBERT which may yield richer results.
- Can incorporate sequence-based attention models like Long Short-term models (LSTMs) directly into the pipeline to allow the model to dynamically focus on functionally relevant residues.

Overall, while the current pipeline demonstrates the potential of lightweight protein embeddings for protein localisation prediction, future iterations with the above-mentioned additions will better align the model structure and label format with the biological nature of subcellular localisation to achieve robust and biologically meaningful predictions.

References

1. Zhang, X., Tseo, Y., Bai, Y. *et al.* Prediction of protein subcellular localization in single cells. *Nat Methods* (2025). https://doi.org/10.1038/s41592-025-02696-1
2. Huang WL, Tung CW, Ho SW, Hwang SF, Ho SY. ProLoc-GO: utilizing informative Gene Ontology terms for sequence-based prediction of protein subcellular localization. BMC Bioinformatics. 2008 Feb 1; 9:80. doi: 10.1186/1471-2105-9-80. PMID: 18241343; PMCID: PMC2262056
3. Almagro Armenteros, J.J., et al. (2019). "Detecting sequence signals in targeting peptides using deep learning." *PMC*. https://doi.org/10.1371/journal.pcbi.1007562
4. Rives, A., et al. (2021). "Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences." *PNAS*. https://doi.org/10.1073/pnas.2016239118
5. Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, Smetanin N, Verkuil R, Kabeli O, Shmueli Y, Dos Santos Costa A, Fazel-Zarandi M, Sercu T, Candido S, Rives A. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science. 2023 Mar 17;379(6637):1123-1130. doi: 10.1126/science.ade2574. Epub 2023 Mar 16. PMID: 36927031.
6. Weissenow K, Heinzinger M, Rost B. Protein language-model embeddings for fast, accurate, and alignment-free protein structure prediction. Structure. 2022 Aug 4;30(8):1169-1177.e4. doi: 10.1016/j.str.2022.05.001. Epub 2022 May 23. PMID: 35609601.
7. Brandes N, Ofer D, Peleg Y, Rappoport N, Linial M. ProteinBERT: a universal deep-learning model of protein sequence and function. Bioinformatics. 2022 Apr 12;38(8):2102-2110. doi: 10.1093/bioinformatics/btac020. PMID: 35020807; PMCID: PMC9386727.