Predicting the Subcellular Location of Proteins (COMP0082)

Rita Kurban

March 20, 2023

Abstract

Protein localisation is crucial in understanding cellular processes and developing innovative therapeutics. In this study, we compare the performance of two machine learning models, XGBoost and Meta ESM, in predicting subcellular protein localisation for five classes: cytosolic, mitochondrial, secreted, nuclear, and other. XGBoost is a well-established technique that utilises feature-based representations of protein sequences. Meta ESM is a state-of-the-art deep learning model that employs contextualised protein sequence embeddings. We evaluated both models on a diverse dataset of 11,000 protein sequences, observing that Meta ESM significantly outperformed XGBoost. Meta ESM achieved an F1 score of 0.88, while XGBoost reached 0.70. Furthermore, Meta ESM provides localisation probabilities for each class, offering valuable insights into prediction confidence. To make these predictions readily available to researchers, we developed an intuitive Flask web application that generates localisation probabilities for any input protein sequence. Our study underscores the significance of protein localisation prediction and showcases the potential of advanced deep learning techniques, such as Meta ESM, in this domain. By integrating multiple approaches and creating accessible tools, we can enhance protein localisation performance and reliability, thereby promoting further research in this area.

1 Introduction

Protein localisation determines the subcellular location of a protein within a cell. Understanding protein localisation is essential for elucidating protein function, interactions, and role in various cellular processes. Accurate protein localisation prediction can aid in developing innovative therapeutics and contribute to a deeper understanding of cellular mechanisms. In this study, we focus on predicting five subcellular localisations:

- **Cytosolic proteins** are located within the cell cytoplasm but outside the organelles. They play a crucial role in regulating various intracellular reactions.
- **Extracellular/Secreted proteins** are produced by cells and participate in cell signalling pathways. They are often involved in processes like breaking down proteins and complex sugars to enable their absorption into the bloodstream.
- **Nuclear proteins** reside within the cell's nucleoplasm and are essential for DNA replication, transcription, gene regulation, and epigenomic activation or silencing of genes.
- **Mitochondrial proteins** are found within the mitochondria, which contribute to processes that generate ATP in the cell, participating in essential functions like the citric acid cycle and mitochondrial gene transcription.
- This fifth category, **"other"**, includes prokaryotic proteins that sometimes contaminate samples during sequencing and sequences labelled as "none of the above" during the labelling process.

Traditionally, protein localisation has been determined through experimental techniques such as fluorescence microscopy, cell fractionation, and mass spectrometry [7]. While these methods provide valuable insights, they can be labour-intensive, time-consuming, and costly. As the volume of available protein sequence data has grown exponentially, there is an increasing need for efficient computational methods to predict protein localisation.

Over the years, various computational approaches have been developed to predict protein localisation accurately. Early methods primarily relied on amino acid composition-based techniques, which utilised global or local amino acid composition of protein sequences and statistical techniques, such as Chou's pseudo-amino acid composition approach [4]. Subsequently, signal-based approaches emerged, which identify distinct sequence patterns associated with various subcellular locations [2]. Domain-based methods were later developed, exploiting the presence of specific protein domains or motifs to predict subcellular localisation using protein domain databases and sequence alignment algorithms [5]. With the advent of machine learning, a new generation of methods was introduced, employing techniques such as Support Vector Machines (SVM), Artificial Neural Networks (ANN), and Random Forests (RF) to build predictive models based on various sequence features [8, 1, 11]. Hybrid methods have been proposed to improve prediction accuracy, integrating multiple features such as amino acid composition, sorting signals, and protein domains [9]. Ensemble methods have also gained popularity, combining multiple classifiers to achieve better predictive performance by employing techniques such as majority voting, bagging, and boosting [6].

The protein subcellular localisation prediction field has seen significant progress, and continued development in computational methods is vital to advance our understanding of cellular processes and facilitate drug discovery. Our approach builds upon recent advancements in deep learning and leverages a state-of-the-art model, Meta ESM, which employs contextualised protein sequence embeddings. This technique allows a more nuanced understanding of protein sequences, capturing complex patterns and relationships between amino acids. In this study, we compare the performance of Meta ESM with the well-established XGBoost model and demonstrate the deep learning approach's superiority in prediction accuracy and interpretability. By integrating advanced machine learning techniques and providing accessible tools for researchers, our work contributes to developing reliable and efficient protein localisation prediction methods.

2 Methods

2.1 Models

A brief outline of XGBoost and Meta ESM is provided subsequently. Specific implementational details are provided in the Preprocessing, Model Training, and Evaluation sections.

2.1.1 XGBoost

XGBoost is a popular gradient-boosting algorithm involving an ensemble of decision trees trained iteratively to minimise the loss function [3]. It is particularly effective in dealing with structured data and has been widely used in various machine-learning tasks, including classification and regression problems. XGBoost incorporates regularisation, sparsity-aware learning, and parallelisation to enhance performance, reduce overfitting, and speed up training time. The main hyperparameters of XGBoost include the number of estimators, maximum tree depth, and learning rate. These hyperparameters control the model's complexity, training speed, and the trade-off between model performance and overfitting.

2.1.2 Meta ESM

Meta ESM is a state-of-the-art general-purpose protein language model. It can be used to predict structure, function and other protein properties directly from individual sequences [10]. The model employs transformer-based architectures, proven highly effective in various natural language processing tasks, and adapts them to handle protein sequences. Meta ESM takes advantage of pre-trained embeddings and fine-tunes them on the target task, significantly reducing the required training data and improving generalisation. In the context of protein subcellular localisation prediction, Meta ESM learns complex patterns and dependencies within protein sequences, leading to more accurate predictions. Key hyperparameters in Meta ESM include the architecture size, learning rate, batch size, and fine-tuning strategy, which can be tuned to optimise the model's performance and prevent overfitting.

2.2 Data

The dataset used in this study consists of 11,224 protein sequences in FASTA format, representing five distinct subcellular localisation classes: cytosolic, mitochondrial, secreted, nuclear, and other. The data provides a diverse and comprehensive representation of protein sequences. It allows machine learning models to learn complex patterns and dependencies within the sequences to predict their subcellular localisation accurately. Additionally, a blind dataset of 20 sequences is provided to evaluate the performance of the models on previously unseen data, further validating their generalisation capabilities. Figure 1 illustrates the distribution of classes.

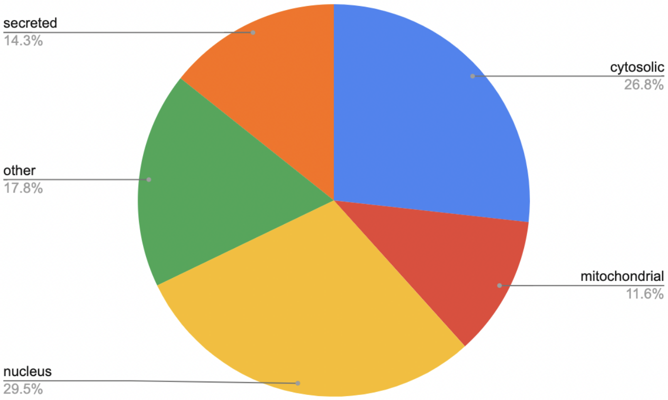

Figure 1: Distribution of protein localisation classes.

2.3 Feature Selection

In the XGBoost model, feature selection is a vital step in the modelling process. We extracted features using the PropParam module of the Biopython package, which offers various physicochemical properties of protein sequences. Furthermore, we included dipeptides, and local and global amino acid distribution, as they capture the local interactions between adjacent amino acids in protein sequences and the overall composition, providing valuable insights into their physicochemical properties and potential structural patterns. However, the high number of dipeptide features can increase the model's complexity. We addressed this issue by employing Principal Component Analysis (PCA), a dimensionality reduction technique that retains the most relevant information from the original dipeptide features while reducing the number of dimensions, simplifying the model, and mitigating the risk of overfitting. We selected two principal components that accounted for most of the variance and incorporated them into the feature set. To further refine the feature set, we eliminated features with a correlation exceeding 80% to avoid multicollinearity. The final set of features contained 70 features. It was organised into categories: basic physicochemical properties, local composition (protein sequence start and end), amino acid composition (percentage of each amino acid), and principal components for dipeptide distribution. Table 1 lists all features and their descriptions.

In contrast, Meta ESM does not necessitate explicit feature selection. It learns relevant features from raw protein sequences using contextualised protein sequence embeddings, making it more efficient and effective at predicting subcellular localisation without manual feature selection compared to traditional machine learning techniques like XGBoost. The only preprocessing step that was required is the tokenisation of the sequences and their truncation, as the ESM model can only analyse 1024 tokens at a time. As a result, some longer sequences (around 10%) got truncated.

2.4 Preprocessing

The dataset used in this study did not require standardisation, as XGBoost is a tree-based model that is not sensitive to feature scaling. The classes were encoded using integer values 0, 1, 2, 3, 4 to represent the different subcellular locations. Some sequences containing amino acids other than the standard 20 amino acids were omitted from the analysis, removing a total of 64 sequences. This step ensured that the dataset consisted only of sequences with the standard amino acids, improving the model's ability to learn from the data effectively.

2.5 Model Training

The dataset was split into an 80:20 ratio, with 80% of the data used for training and 20% reserved for testing. The test set was only used to obtain the final model evaluation metrics, ensuring an unbiased assessment of the model's performance on unseen data. This approach helps to validate the model's generalisation capabilities and provides a more accurate representation of its performance in real-world applications.

2.5.1 XGBoost

The XGBoost model was evaluated using nested cross-validation, where the inner fold was utilised for hyperparameter tuning, and the outer fold was employed for model evaluation. This entire process was performed on the training set, ensuring that the test set remained untouched throughout the evaluation. The hyperparameters considered during the tuning process included the number of estimators, maximum tree depth, and the learning rate. Nested cross-validation provides a more robust estimate of model performance, accounting for variability in hyperparameter selection and reducing the risk of overfitting. The dataset did not contain any homologous sequences, which allowed for random cross-validation without the risk of overestimating the model's performance due to sequence similarity. The best hyperparameters for the XGBoost model were selected using the Optuna bayesian optimisation framework based on the F1 score, as this metric is particularly suitable for multi-class classification problems. By optimising the F1 score during hyperparameter tuning, we aimed to achieve a balanced model performance that accounts for precision and recall, ensuring that the model can effectively distinguish between different protein subcellular localisations. The parameters are summarised in Table 2.

2.5.2 Meta ESM

We tested the model with 35 million parameters. Due to the model's complexity and computational requirements, we did not perform hyperparameter tuning. However, we still used the same evaluation procedure as for the XGBoost model to ensure easy comparison. We let the model automatically identify the batch size that fits into memory. We evaluated the model using cross-validation on the training set and assessed its performance after one fine-tuning epoch. After obtaining satisfactory results, we evaluated the model on the test set to obtain the ultimate performance metrics.

2.6 Model Evaluation

Upon completing the tuning process, the models were trained on the entire training set and evaluated on the test set to obtain a final performance estimate. Four metrics were reported to comprehensively assess the results:

- **Precision:** Indicates the correctness of positive predictions, useful when false positives are costly.
- **Recall:** Measures the ability to identify all positive instances, important when false negatives are costly.
- **F1 score:** Harmonic mean of precision and recall, balancing their tradeoff, especially in imbalanced datasets.
- **Accuracy:** Proportion of correct predictions, widely used but can be misleading in imbalanced datasets.

Together, these metrics provide a detailed understanding of the model's performance, helping assess its suitability for a given task.
Additionally, a confusion matrix was generated to illustrate the models' performance further. The final ESM Meta model used to obtain the blind set predictions, leveraging its full potential for accurate subcellular localisation prediction. To estimate its confidence, class probabilities were extracted for each class. Probabilities above 80% were considered high confidence, those above 60% were deemed medium confidence, and probabilities below 60% were categorised as low confidence.

3 Results

The comparison of the results obtained from the XGBoost and Meta ESM models, though only partially equitable due to the inherent differences in their nature, showcases the potential of each approach. As Meta ESM is a pre-trained model, it is expected to have some advantages over traditional machine learning methods like XGBoost. Nevertheless, we attempted to employ identical procedures for both models to minimise potential bias in our evaluation process.
The cross-validation results for XGBoost and Meta ESM models are presented in Table 3, which displays the performance metrics for each model during the CV process. From the table, we can observe that the ESM model consistently outperforms the XGBoost model in terms of F1 scores (0.85 vs. 0.69), indicating that the ESM model is likely better suited for this specific classification task.
The final performance of each model on the test set using their optimal parameters is highlighted in Tables 4 and 5, which provide a comprehensive evaluation of their generalisation capabilities in terms of different evaluation metrics.

3.1 XGBoost

Table 4 demonstrates that XGBoost exhibits varying performance across different classes. It performs best on the "other" class, followed by the "secreted" class. The "mito" class shows moderate performance, while the "nucleus" and "cyto" classes exhibit the lowest performance. The model particularly struggles to distinguish between the "cyto" and "nucleus" classes, as evidenced by the high number of misclassifications between them.
The confusion matrix 2 provides a further visualisation of these results, highlighting the confusion occurring between "cyto" and "nucleus" classes.

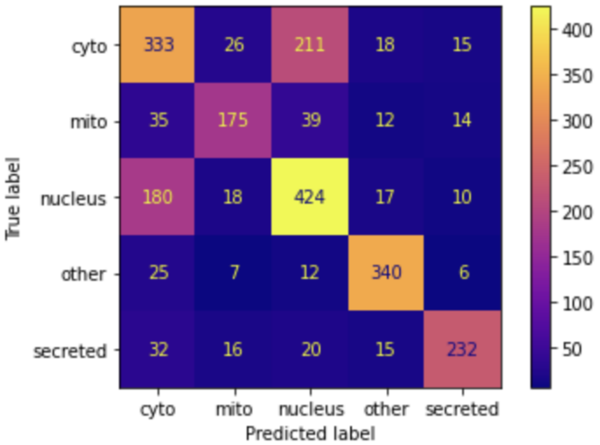

Figure 2: XGBoost Confusion Matrix.

3.2 Meta ESM

Meta ESM demonstrates improved performance compared to the XGBoost as demonstrated by Table 5 and Figure 3. The model seems to have a better understanding of the ”other” and ”secreted” classes, exhibiting higher precision and recall values. Additionally, the performance on the ”mito” class has also improved, indicating the model’s enhanced ability to distinguish between different subcellular localisations. Overall, the updated table and confusion matrix reflect a more accurate and reliable model for predicting subcellular localisations.

3.3 Final Predictions

Table 6 details the final predictions of the Meta ESM on the blind set and their associated confidence levels. The best performing model across five fine-tuning epochs was used. The model was trained on the training set and evaluated on the test set. Training on all data without a validation set is not advisable in deep learning as it is possible to run into issues with overfitting or underfitting, as there will be no reliable way to monitor the model performance during training. The table presents the predictions and associated confidence levels for a set of 20 sequences. It shows that most of the predictions have high confi-

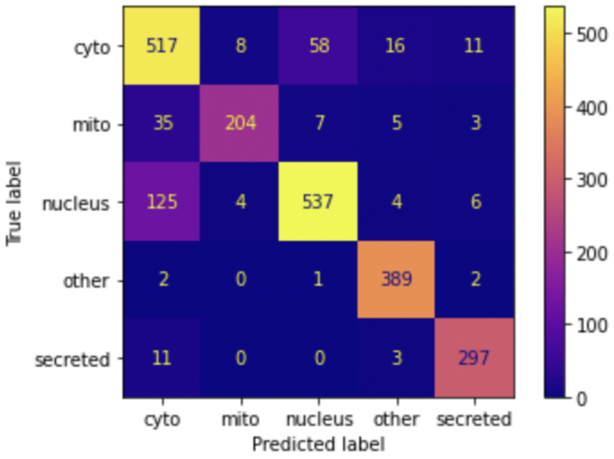

Figure 3: Meta ESM Confusion Matrix.

dence, indicating that the model is quite certain about its predictions for those sequences. However, there are two sequences (SEQ02 and SEQ11) with low confidence (< 0.6) and one sequence (SEQ10) with medium confidence (< 0.8), that suggest that the predictions for these sequences might be less reliable. The predictions cover a diverse range of classes, including ”cyto”, ”mito”, ”nucleus”, ”other”, and ”sectered” (same as ”extr”), demonstrating the model’s capability to classify sequences into different categories.

4 Discussion

In this study, we aimed to predict the subcellular localisation of proteins using two different models: XGBoost and Meta ESM. We used a dataset with 11,224 protein sequences divided into five classes. The dataset was split into an 80:20 ratio, with the test set reserved for final model evaluation. Both models were evaluated using nested cross-validation on the training set, where the inner fold was used for hyperparameter tuning (only for XGBoost) and the outer fold for evaluation. The best hyperparameters were picked using the F1 score, which is a good metric for multi-class classification. The XGBoost model required feature selection and extraction using the PropParam module and principal component analysis (PCA) for dipeptide distribution. The Meta ESM model, however, did not require explicit feature selection as it learns relevant features from the raw protein sequences using contextualised protein sequence embeddings. Both models were evaluated on the test set to obtain final performance estimates, with XGBoost achieving an F1 score of 0.70 and Meta ESM reaching 0.88. Although the comparison between XGBoost and Meta ESM is only partially fair because Meta ESM is a pre-trained model, we attempted to reduce bias by using identical procedures for both models. Reporting precision, recall, F1 score, and accuracy allowed us to understand the model performances comprehensively. The results indicate that the Meta ESM model significantly outperforms the XGBoost model in predicting subcellular protein localisation, demonstrating the effectiveness of using pre-trained models for such tasks.

In addition to open-sourcing the code[1], we created an interactive GUI to predict the subcellular localisation of a given protein sequence. The application takes any protein sequence as input and uses the best-preforming ESM model to predict the most likely subcellular localisation. The output is displayed on the web page in a clear and easy-to-understand format. The application can be easily accessed and used by anyone with an internet connection.

5 Conclusion

In conclusion, our results suggest that pre-trained deep learning models, such as Meta ESM, can outperform traditional machine learning techniques, such as XGBoost, for protein subcellular localisation prediction. Future work could involve optimising the models for speed and accuracy, possibly by fine-tuning hyperparameters or exploring alternative architectures. Additionally, training the models on larger datasets with a wider range of protein sequences and localisations could help improve their generalisation and robustness. Another potential avenue for future work is exploring ensemble methods that combine the strengths of multiple models for improved performance.

[1] https://github.com/ritakurban/protein-localizer

References

[1] José Juan Almagro Armenteros et al. "DeepLoc: prediction of protein subcellular localization using deep learning". In: *Bioinformatics* 33.21 (2017), pp. 3387–3395.

[2] José Juan Almagro Armenteros et al. "SignalP 5.0 improves signal peptide predictions using deep neural networks". In: *Nature biotechnology* 37.4 (2019), pp. 420–423.

[3] Tianqi Chen et al. "Xgboost: extreme gradient boosting". In: *R package version 0.4-2* 1.4 (2015), pp. 1–4.

[4] Kuo-Chen Chou. "Prediction of protein cellular attributes using pseudo-amino acid composition". In: *Proteins: Structure, Function, and Bioinformatics* 43.3 (2001), pp. 246–255.

[5] Chittibabu Guda and Shankar Subramaniam. "TARGET: a new method for predicting protein subcellular localization in eukaryotes". In: *Bioinformatics* 21.21 (2005), pp. 3963–3969.

[6] Manish Kumar and Gajendra PS Raghava. "Prediction of nuclear proteins using SVM and HMM models". In: *BMC bioinformatics* 10.1 (2009), pp. 1–10.

[7] Ravindra Kumar and Sandeep Kumar Dhanda. "Bird eye view of protein subcellular localization prediction". In: *Life* 10.12 (2020), p. 347.

[8] Ravindra Kumar et al. "Protein sub-nuclear localization prediction using SVM and Pfam domain information". In: *PloS one* 9.6 (2014), e98345.

[9] Bo Li et al. "Prediction of protein subcellular localization based on fusion of multi-view features". In: *Molecules* 24.5 (2019), p. 919.

[10] Zeming Lin et al. "Evolutionary-scale prediction of atomic level protein structure with a language model". In: *bioRxiv* (2022), pp. 2022–07.

[11] Lingling Zhao et al. "Deep forest-based prediction of protein subcellular localization". In: *Current Gene Therapy* 18.5 (2018), pp. 268–274.

A Technical Appendix

The following technical approaches, libraries and techniques were used in this work:

- Python programming language.
- PyTorch and Transformers libraries were used to fine-tune the ESM protein localisation model.
- Scikit-learn library was used for XGBoost model training and evaluation.
- Hugging Face's Transformers library was used for text encoding and decoding.
- Pandas and NumPy libraries were used for data manipulation and analysis.
- Biopython package for XGBoost feature extraction.
- Matplotlib library was used for data visualisation.
- Principal Component Analysis (PCA) was used for dimensionality reduction.
- Nested cross-validation and Optuna were used for hyperparameter tuning and model evaluation.
- Wandb was used for tracking and visualising experiment results.
- Flask was used to create a web application for protein localisation prediction.

In summary, we used a combination of programming languages, libraries, and frameworks to preprocess, model, and evaluate our data. We employed various techniques such as hyperparameter tuning and feature reduction to optimise the performance of our models. Finally, we deployed the ESM model as a web application using the Flask web framework. For further details and implementation specifics, please refer to the Github repository.

Features	**Description**
Molecular Weight	The molecular weight of the protein.
Aromaticity	The relative frequency of aromatic amino acids in the protein sequence.
Instability Index	A measure of the protein stability.
Isoelectric Point	The pH at which the protein carries no net electrical charge.
Helix Fraction	The fraction of amino acids participating in alpha-helical structures.
Sheet Fraction	The fraction of amino acids participating in beta-sheet structures.
GRAVY	The grand average of hydropathy (hydrophobicity or hydrophilicity) of the protein.
Charge at pH 7	The net charge of the protein at pH 7.
Local Composition Start (20)	Composition of amino acids at the start of the protein sequence, first 50 amino acids.
Local Composition End (20)	Composition of amino acids at the end of the protein sequence, last 50 amino acids.
Amino Acid Composition (20)	Percentage of each amino acid in the protein sequence.
Principal Components (2)	Principal components for dipeptide distribution derived from the entire dipeptide distribution.

Table 1: Features and their descriptions.

Hyperparameter	**Value**
n_estimators	159
max_depth	8
learning_rate	0.3

Table 2: Optimal hyperparameters for the final XGBoost model.

Fold	XGBoost	ESM
1	0.7004	0.8396
2	0.6876	0.8490
3	0.7048	0.8411
4	0.6765	0.8430
5	0.6892	0.8588
Mean ± STD	0.6917 ± 0.0099	0.8463 ± 0.0068

Table 3: Cross-validation F1 scores for XGBoost and ESM models across five folds.

Class	Precision	Recall	F1-score
Cyto	0.5504	0.5523	0.5513
Mito	0.7231	0.6364	0.6769
Nucleus	0.6006	0.6533	0.6260
Other	0.8586	0.8718	0.8652
Secreted	0.8375	0.7365	0.7830
Mean ± STD	0.7140 ± 0.1294	0.6901 ± 0.1224	0.7005 ± 0.1203

Table 4: XGBoost Test Set Model Evaluation.

Class	Precision	Recall	F1-score
Cyto	0.76	0.85	0.80
Mito	0.94	0.79	0.86
Nucleus	0.89	0.77	0.83
Other	0.92	0.99	0.95
Secreted	0.95	0.96	0.95
Mean ± STD	0.8920 ± 0.0891	0.8720 ± 0.0832	0.8780 ± 0.0827

Table 5: Meta ESM Test Set Model Evaluation.

Sequence	Prediction	Confidence
SEQ01	OTHR	HIGH
SEQ02	CYTO	LOW
SEQ03	MITO	HIGH
SEQ04	CYTO	HIGH
SEQ05	CYTO	HIGH
SEQ06	MITO	HIGH
SEQ07	EXTR	HIGH
SEQ08	MITO	HIGH
SEQ09	EXTR	HIGH
SEQ10	EXTR	MEDIUM
SEQ11	NUCL	LOW
SEQ12	CYTO	HIGH
SEQ13	NUCL	HIGH
SEQ14	CYTO	HIGH
SEQ15	OTHR	HIGH
SEQ16	NUCL	HIGH
SEQ17	OTHR	HIGH
SEQ18	NUCL	HIGH
SEQ19	OTHR	HIGH
SEQ20	CYTO	HIGH

Table 6: Sequence Predictions and Confidence.