Elinvo (Electronics, Informatics, and Vocational Education) ISSN 2580-6424 (printed) | ISSN 2477-2399 (online)
Vol 9, No. 2, 2024, pp. 238-252

A Benchmark Study of Protein Embeddings in Sequence-Based Classification

Humasak Simanjuntak[1*], Lamsihar Siahaan[1], Patricia Dian Margaretha[1], Ruth Christine Manurung[1], Susi Purba[1], Rosni Lumbantoruan[1], Arlinta Barus[1], Helen Grace B. Gonzales[2]

[1]Institut Teknologi Del, Toba, Indonesia
[2]University of Science and Technology of Southern Philippines, Cagayan de Oro, Philippines

Article Info

Article history:

Received August 31, 2024
Revised September 23, 2024
Accepted October 07, 2024

Keywords:

Protein function prediction; protein embedding; protein sequence; protein function; sequence-based classification

Abstract

Proteins play a vital role in various tissue and organ activities and play a key role in cell structure and function. Humans can produce thousands of proteins, each consisting of tens or hundreds of interconnected amino acids. The sequence of amino acids determines the protein's 3D structure and conformational dynamics, which in turn affects its biological function. Understanding protein function is very important, especially for biological processes at the molecular level. However, extracting or studying features from protein sequences that can predict protein function is still challenging: it takes a long time, is an expensive process, and has yet to be maximized in accuracy, resulting in a large gap between protein sequence and function. Protein embedding is essential in function protein prediction using a deep learning model. Therefore, this study benchmarks three protein embedding models, ProtBert, T5, and ESM-2, as a part of function protein prediction using the LSTM Model. We delve into protein embedding performance and how to leverage it to find optimal embeddings for a given use case. We experimented with the CAFA-5 dataset to see the optimal embedding model in protein function prediction. Experiment results show that ESM-2 outperforms from ProtBert and T5. On training, the accuracy of ESM-2 is above 0.99, almost the same as T5, but still above ProtBert. Furthermore, testing on five samples of protein sequence shows that ESM2 has an average hit rate of 93.33% (100% for four samples and 66.67% for one sample).

Corresponding Author:

Email: humasak@gmail.com; humasak@del.ac.id

INTRODUCTION

Understanding protein function becomes crucial in the era of genetics and molecular biology that continues to develop [1]. Various genome sequencing projects have generated millions of protein sequences, especially with the advancement of sequencing technology and its application in the metagenomics area. This opens up opportunities for researching new drugs and helps overcome genetic diseases that are still a mystery in the medical world. Research on protein function also helps identify molecular pathways involved in biological processes, such as cell growth, immune response, and wound healing [2] [3].

Protein function prediction is one of the main tasks in bioinformatics that can help in various biological problems, such as understanding disease mechanisms or finding drug targets. Several approaches to predicting proteins have been developed, ranging from approaches based on protein sequences [1] [4] [5] [6] [7], gene expression [8], sequence similarity [9], network interaction [10] [11], and text mining [12]. However, determining protein function experimentally based on sequence still takes a long time, is an expensive process, and has a less-than-optimal level of accuracy, resulting in a large gap between protein sequence and function. In addition, previous studies have also had challenges in extracting or studying features from protein sequences that can predict function and ensuring that the prediction is consistent with the background biological knowledge of the function and its relationship.

In addition to sequence features, many other methods are available to predict protein function based on protein-protein interaction networks [10] [13] [14] and protein structure [15] [16] [17]. However, apart from the sequence, most features are challenging to obtain or unavailable for many proteins, limiting their scope. In previous studies, the performance of sequence-based function prediction methods was still lower than that of methods that combined many features.

In deep learning, especially in protein function prediction, a protein embedding model is crucial in representing amino acids' complex, high-dimensional sequences in a way that captures their biological properties and relationships. Furthermore, embedding models has an important role in dimensionality reduction, capturing relationships, improving prediction accuracy, handling sparse data, and transfer learning. Specifically for accuracy, the suitable embeddings capture the semantic relationships between data points, enabling the model to make more accurate predictions. These embeddings are then used as inputs for various predictive tasks, such as protein structure prediction, function annotation, and interaction analysis.

Protein embedding models have shown great promise in protein function prediction, but several challenges and limitations can affect their performance, including Data Sparsity and Imbalance, Contextual Information Loss, Transfer Learning Challenges, Sequence Length and Complexity, and Computational Limitations. Therefore, this study tries to conduct a study on three protein embeddings: ProtBert [18] [19], T5 [20], and ESM-2 [21], to see the performance of the prediction training results using the LSTM Model. Embedding models will be carried out on the CAFA-5 dataset to become input to the deep learning model. Furthermore, this paper contributes to benchmarking these three embedding models to learn the best model for capturing the semantic representation of protein sequence function.

METHODS

This study proposes using three embedding methods, Probert, T5, and ESM-2, to encode proteins' functional and structural properties based on their sequences into a machine-friendly format (vector representation). We compare which embedding method is the most effective in providing protein function prediction results based on the hit rate metric. The Function Prediction is built using the LSTM architecture. Figure 1 illustrates the process of predicting protein function using three different embedding models. The process begins by loading data from the CAFA-5 dataset, where data in TSV and FASTA formats are combined. Each embedding takes the protein's amino acid sequence and converts it into a numeric representation or vector. These vectors are then further processed using the LSTM model, which consists of two main layers with different unit sizes. This model is also equipped with a dropout layer to prevent overfitting. The final output of this model is used to predict protein function and is evaluated using relevant metrics. This approach combines methods from various embedding models to produce more accurate protein function predictions.

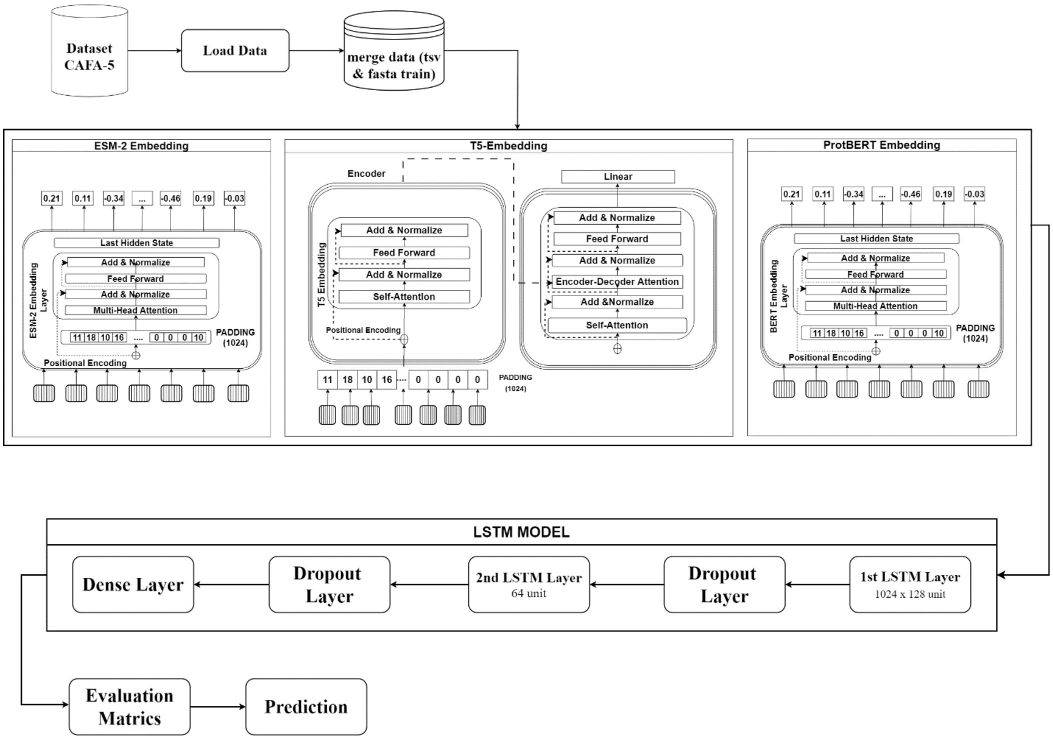


Figure 1. Proposed Research Design

Dataset

The dataset used in this study is the CAFA-5 (Critical Assessment of Functional Annotation) competition [22] [23]. CAFA-5 (Critical Assessment of Functional Annotation, 5th edition) is a community-driven challenge designed to evaluate the performance of computational methods for predicting protein function. The CAFA challenge focuses on assessing the ability of models to accurately annotate proteins with Gene Ontology (GO) terms, which describe biological processes, molecular functions, and cellular components. The goal of CAFA is to benchmark the progress of protein function prediction methods, providing a platform for comparing different approaches and pushing the field forward.

The dataset for CAFA-5 consists of a large collection of protein sequences that are either newly discovered or uncharacterized, meaning their functions are not yet known at the time of the challenge. Participants in the CAFA-5 challenge are tasked with predicting the functions of these proteins, and the predictions are later evaluated against experimentally verified annotations that become available over time. The dataset includes proteins from various species, providing a diverse range of sequences for testing the generalizability of prediction methods.

In addition to protein sequences, the CAFA-5 dataset typically includes associated metadata, such as organism information and related GO terms for proteins that have some existing annotations. This dataset serves as a valuable resource for researchers working on protein function prediction, allowing them to test and refine their models in a competitive setting with real-world relevance. The results of the challenge contribute to improving the accuracy and reliability of computational methods for protein annotation, which is critical for understanding protein roles in biology and disease.

This dataset consists of two main files, namely train and test files, with .tsv and .fasta formats. Each file in this dataset has attributes such as Protein_ID, Sequence, GO_Term_ID, and aspects of protein function that include Cellular Component (CC), Biological Process (BP), and Molecular Function (MF) categories. The training files are merged into a single dataset so all attributes from different files can be used, increasing the total data to obtain a more comprehensive dataset. It is essential because each protein can have many different protein functions, all of which are recorded in this dataset. The sample of the CAFA-5 dataset used can be seen in Table 1.

Table 1. Sample CAFA-5 dataset

Protein ID	Term ID	Ontology	Sequence
Q61824	GO:0071944 GO:0005575 GO:0005515 GO:0005488	CCO CCO MFO MFO	MAERPARRAPPARALLLALAGALLAPRAARGMSLWDQRGTYEVARALLSKDPGIPGQSIPAKDHPDVLTVLQLESRDLILSLERNEGLIANGFTETHYLQDGTDVSLTRNHTDHCYYH...
P29618	GO:0005515 GO:0005488 GO:0003674	MFO MFO MFO	MEQYEKEEKIGEGTYGVVYRARDKVTNETIALKKIRLEQEDEGVPSTAIREISLLKEMHHGNIVRLHDVIHSEKRIYLVFEYLDLDLKKFMDSCPEFAKNPTLIKSYLYQ...
A2RSY6	GO:0008344 GO:0032501 GO:0030534	BPO BPO BPO	MENMAEEELLPQEKEEAQVRVPTPAPDSAPVPAPAADTALDSAPTPDSDPAPALAPAPAPALSPSLASVPEEAESKRHISIQRRLADLEKLAFGTEGD...
Q9JIX8	GO:0005622 GO:0043229 GO:0031981	CCO CCO CCO	MWGRKRPNSSGETRGILSGNRGVDYGSGRGQSGPFEGRWRKLPKMPEAVGTDPSTSRKMAELEEVTLDGKPLQALRVTDLKAALEQRGLAKSGQKSALVKRLKGALMLENLQKHSTPHAA...
Q5F3W6	GO:0005515 GO:0005488 GO:0003674	MFO MFO MFO	MVDREQLVQKARLAEQAERYDDMAAAMKNVTELNEPLSNEERNLLSVAYKNVVGARRSSWRVISSIEQKTSADGNEKKIEMVRAYREKIEKELGAVCQDVLSLLDNYLIKNCSETQYESK...

Probert,

ProBERT embedding is a specialized adaptation of the BERT (Bidirectional Encoder Representations from Transformers) model, designed to handle legal language and documents [18] [19]. The foundational idea behind ProBERT is to fine-tune BERT on legal-specific corpora so that it captures the unique linguistic patterns and terminologies used in legal texts. BERT, initially introduced by Devlin et al., uses a transformer-based architecture that excels in understanding the context of words by processing the text in both directions (left-to-right and right-to-left) [24]. ProBERT builds upon this foundation, focusing on legal datasets to develop a model that is more adept at handling the intricacies of legal languages, such as statutory terms, legal jargon, and complex sentence structures. By training on legal documents, ProBERT aims to improve performance in tasks like legal document classification, contract analysis, and legal text retrieval.

The ProBERT model generates embeddings by passing the input text through multiple transformer layers, each of which refines the representation of the text. These embeddings are context-aware, meaning that they capture the meaning of a word or phrase in relation to the words around it. This is particularly important in legal texts, where context can dramatically alter the meaning of a term. For example, the word "consideration" in a legal context typically refers to something of value exchanged in a contract rather than the general meaning of careful thought. By producing embeddings that account for these legal-specific meanings, ProBERT can more accurately process legal documents for various applications, such as predicting legal outcomes or identifying relevant case law. One of the key advantages of ProBERT is its ability to handle polysemous terms, which are common in legal language. Traditional word embeddings might struggle with words that have multiple meanings, as they often assign a single vector to each word regardless of context. However, ProBERT's contextual embeddings enable it to differentiate between these meanings based on the surrounding text. This is particularly useful in legal applications, where precise interpretation of language is crucial. The development of ProBERT and its effectiveness in handling legal texts is detailed in related research works, such as the papers by Chalkidis et al. and Zheng et al., which explore how BERT-based models can be adapted for the legal domain through specialized training and fine-tuning on legal corpora [25] [26]. Figure 2 shows the BERT Embedding layer architecture that implemented in this research.

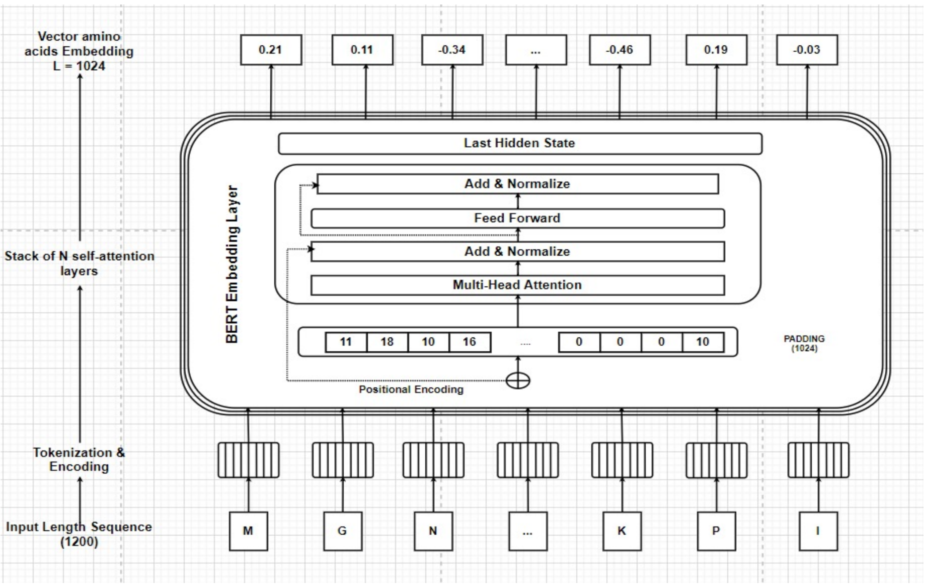


Figure 2. BERT Architecture for Protein Function Prediction

Text-To-Text Transfer Transformer (T5)

T5 is a versatile model initially designed for natural language processing (NLP) tasks, but it has also been adapted for protein sequence analysis, leading to the development of T5 protein embeddings. The T5 model, introduced by Raffel et al., is based on the transformer architecture and operates by converting all tasks into a text-to-text format *[20]*. When applied to protein sequences, T5 protein embeddings work by treating amino acid sequences as a form of "text" that can be encoded and processed in a similar way to human language. This approach enables the model to capture complex patterns and relationships within protein sequences, such as structural motifs, functional sites, and evolutionary conservation.

The T5 protein embeddings are generated by training the model on large datasets of protein sequences. By leveraging the transfer learning capabilities of T5, the model learns to represent protein sequences in a high-dimensional space where similar sequences or functional regions are positioned closely together. These embeddings can then be used for a variety of downstream tasks, such as protein function prediction, structure prediction, or identifying protein-protein interactions. One of the key strengths of T5 protein embeddings is their ability to generalize across different protein families and species, making them a powerful tool for analyzing diverse biological data.

The adaptation of T5 for protein sequences has been explored in several research papers, such as the work by Elnaggar et al., which demonstrated how transformer models like T5 can be fine-tuned for protein-related tasks *[18] [19]*. The use of T5 in protein embedding is part of a broader trend of applying advanced NLP models to biological sequences, which has shown significant promise in enhancing our understanding of proteins and their functions. By leveraging the capabilities of T5, researchers can develop more accurate and efficient methods for predicting protein properties, ultimately contributing to advancements in fields like bioinformatics, drug discovery, and molecular biology. Figure 3 shows the T5 layer architecture implemented in this research.

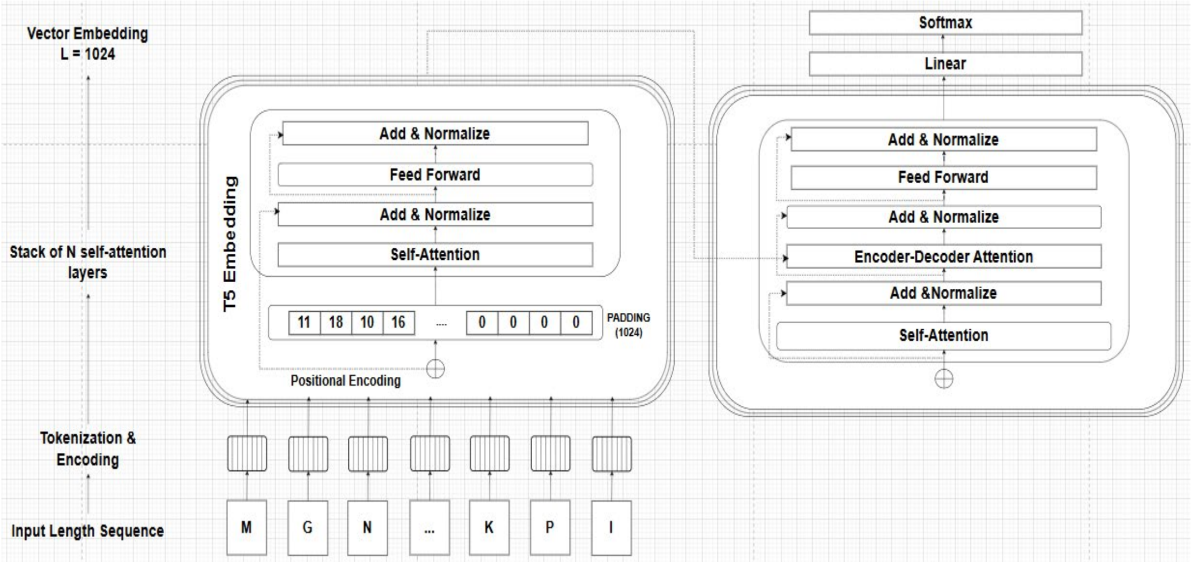


Figure 3. T5 Architecture for Protein Function Prediction

Evolutionary Scale Modeling 2 (ESM-2)

ESM-2 is a powerful protein language model that builds on the success of its predecessors, ESM-1b, by utilizing deep learning techniques to generate protein embeddings *[21]*. Developed by Meta AI, ESM-2 leverages the transformer architecture to analyze protein sequences and generate embeddings that capture intricate details about their structure and function. This model is trained on massive protein sequence datasets, allowing it to learn evolutionary patterns, structural motifs, and functional sites across a wide range of proteins. By embedding protein sequences into high-dimensional spaces, ESM-2 provides a powerful tool for understanding and predicting protein behavior, particularly in tasks such as protein structure prediction, functional annotation, and protein-protein interaction analysis.

ESM-2 embeddings work by transforming protein sequences into a series of vectors, where each vector represents a specific amino acid in the context of its surrounding sequence. The transformer architecture allows ESM-2 to consider long-range dependencies within the sequence, meaning that the model can capture complex interactions between amino acids that may be far apart in the linear sequence but close in the three-dimensional structure. This ability to model distant relationships is crucial for accurately predicting protein structures and functions, as proteins often rely on such interactions for their biological activity. ESM-2's embeddings have been shown to be effective in various downstream tasks, outperforming previous models in protein-related benchmarks.

The effectiveness of ESM-2 embeddings in protein sequence analysis has been highlighted in recent research, where the model's ability to generalize across diverse protein families was demonstrated and offered accurate predictions at the atomic level *[27]*. Additionally, ESM-2's scalability makes it suitable for large-scale protein analysis, enabling researchers to analyze millions of protein sequences efficiently and accurately *[28]*. This scalability, combined with the model's high accuracy, positions ESM-2 as a valuable resource in bioinformatics, particularly for tasks like de novo protein design and understanding the functional implications of genetic mutations. As protein language models continue to evolve, ESM-2 represents a significant step forward in the integration of deep learning and protein science. Figure 3 shows the ESM-2 layer architecture that implemented in this research.

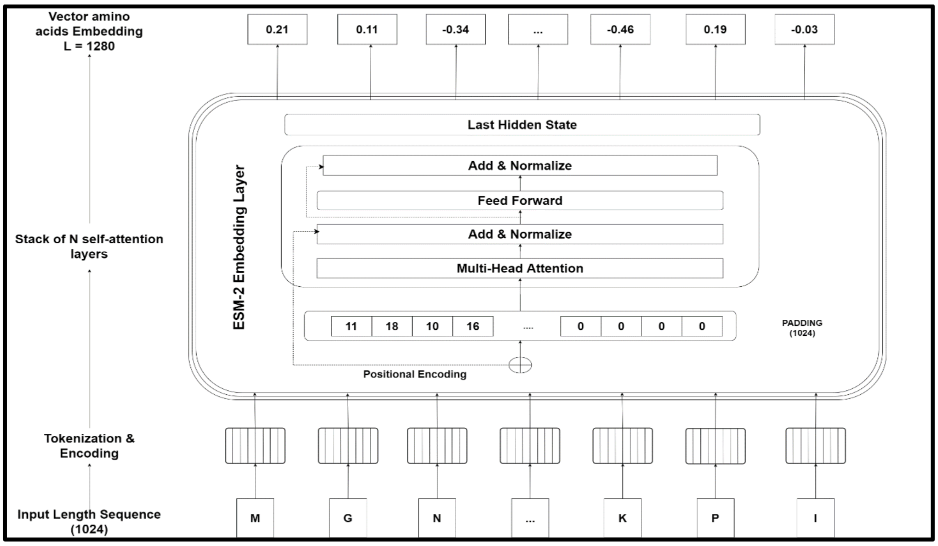


Figure 4. ESM-2 Architecture for Protein Function Prediction

RESULT AND DISCUSSION

The effectiveness of the embedding model is evaluated through the experiments to assess its capabilities. The following section presents and analyzes the performance evaluation results of the embedding model combined with LSTM model protein function prediction. The training model built using the LSTM architecture uses five hyperparameters, as shown in the following Table 2.

Table 2. LSTM Model Hyperparameter

No.	Parameter	Values
1	Optimizer	Adam
2	Batch Size	32
3	Learning Rate	0.001
4	Epoch	50
5	Jumlah layer	7

The metrics used to measure the results of protein function prediction are binary accuracy, AUC, Precision, Recall, and F-1 Score. The following figure compares the results of protein function prediction using the three protein embeddings.

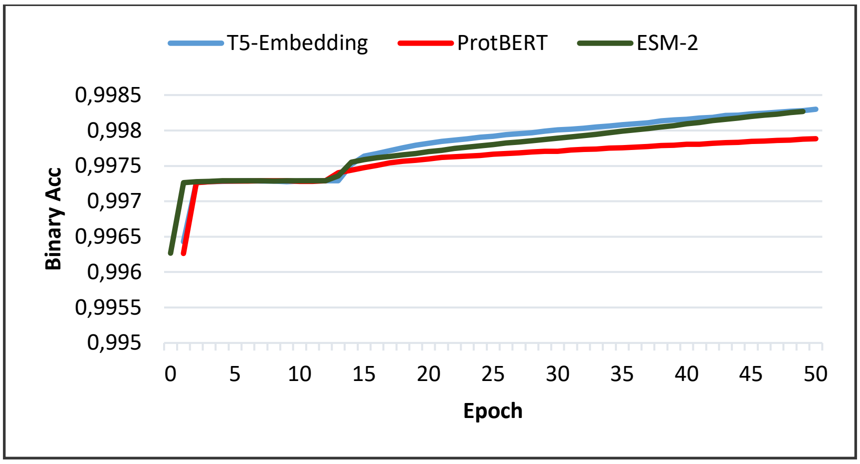


Figure 5. Comparison of Binary Accuracy Protein Function Prediction

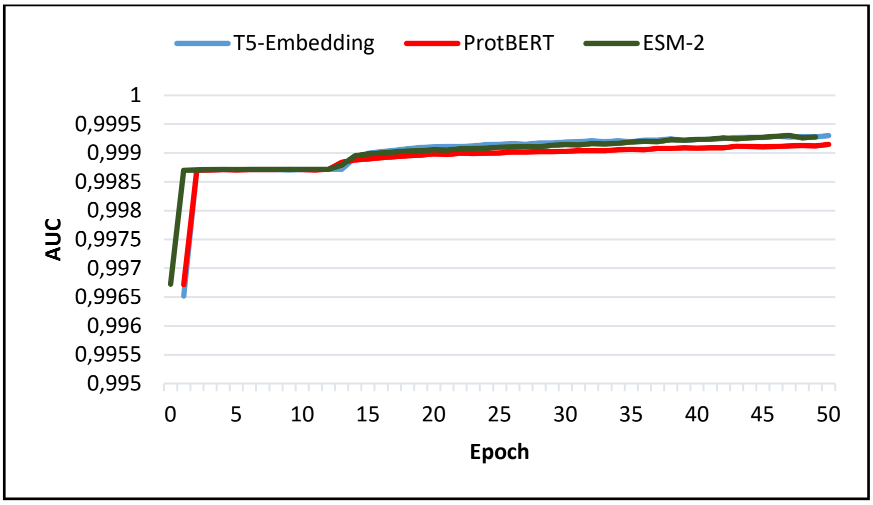


Figure 6. Comparison of AUC Protein Function Prediction

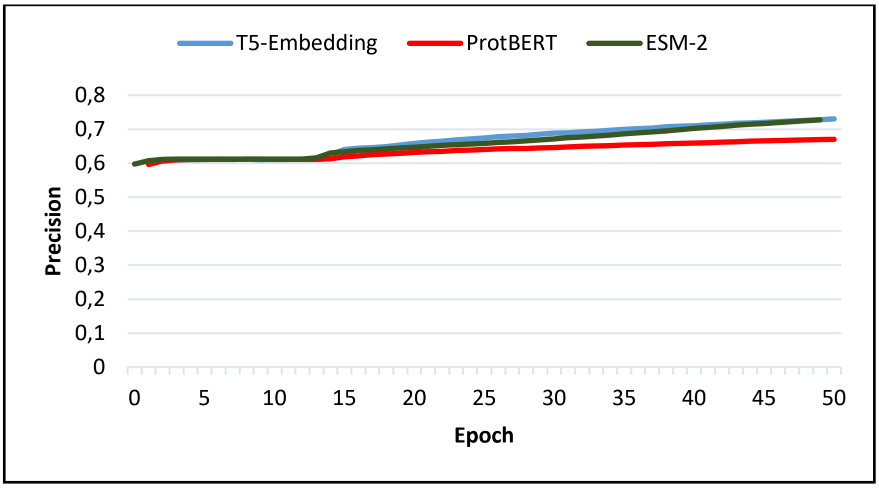


Figure 7. Comparison of Precision Protein Function Prediction

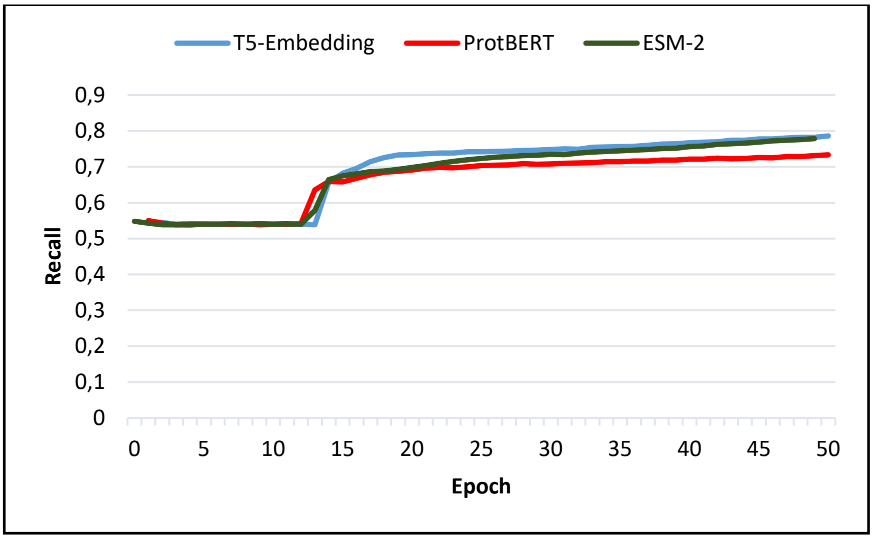


Figure 8. Comparison of Recall Protein Function Prediction

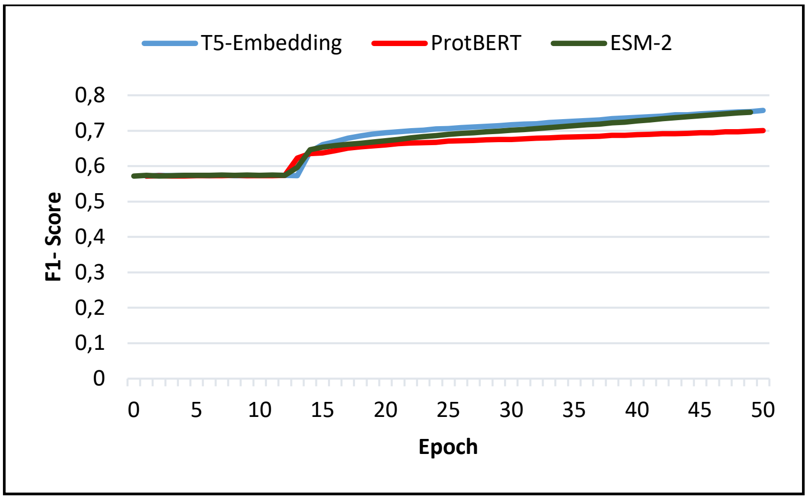


Figure 9. Comparison of F1-Score Protein Function Prediction

After training, we tested the protein function prediction using the model built with each different protein embedding. The model is tested using 5 data samples from CAFA-5: Q9CQV8, P62259, P68510, P61982, O70456. The testing result shows that the top 5 function predictions are GO:0005675, GO:0008150, GO:0110165, GO:0003674, GO:0005622. The following figures (Figure 10 and Figure 11) show the Comparison of Protein Function Prediction Results in the LSTM model with the T5, ProtBERT, and ESM-2 model embeddings.

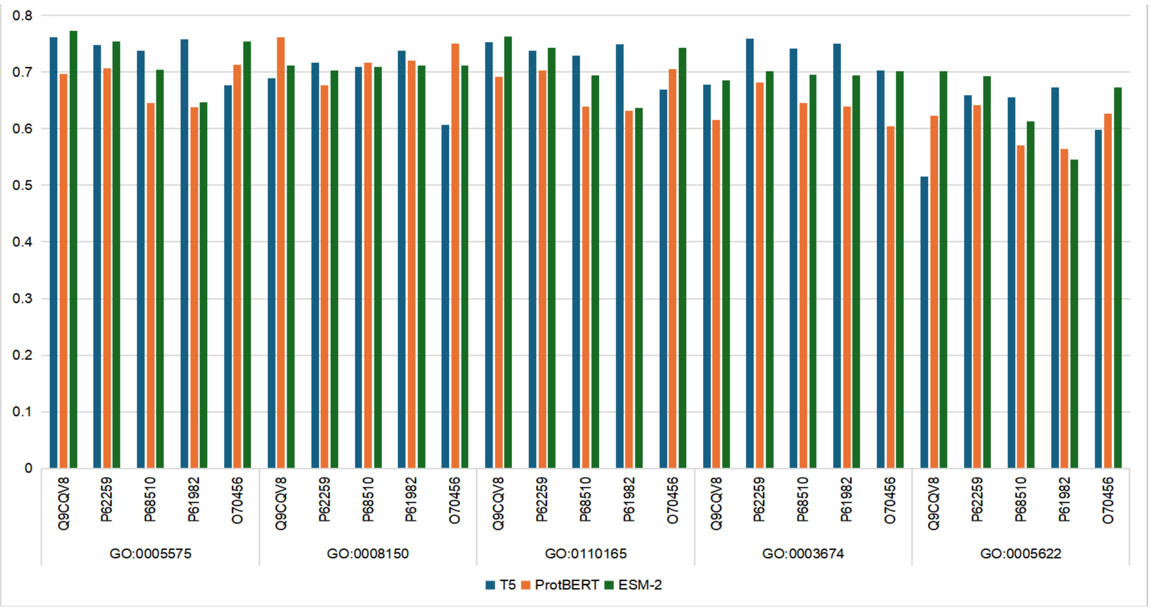
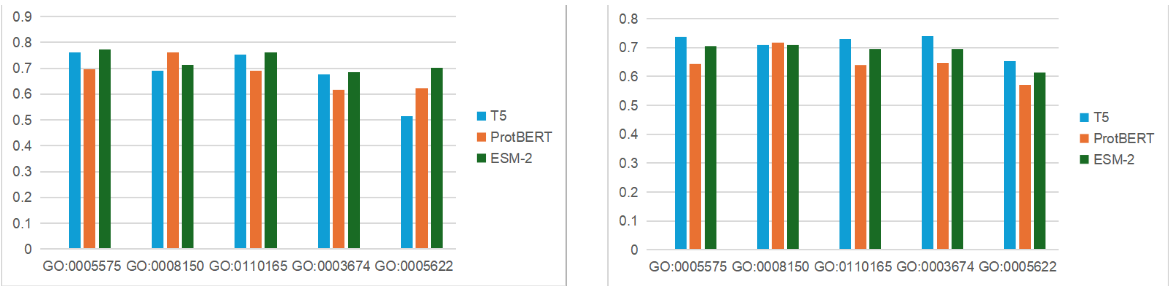
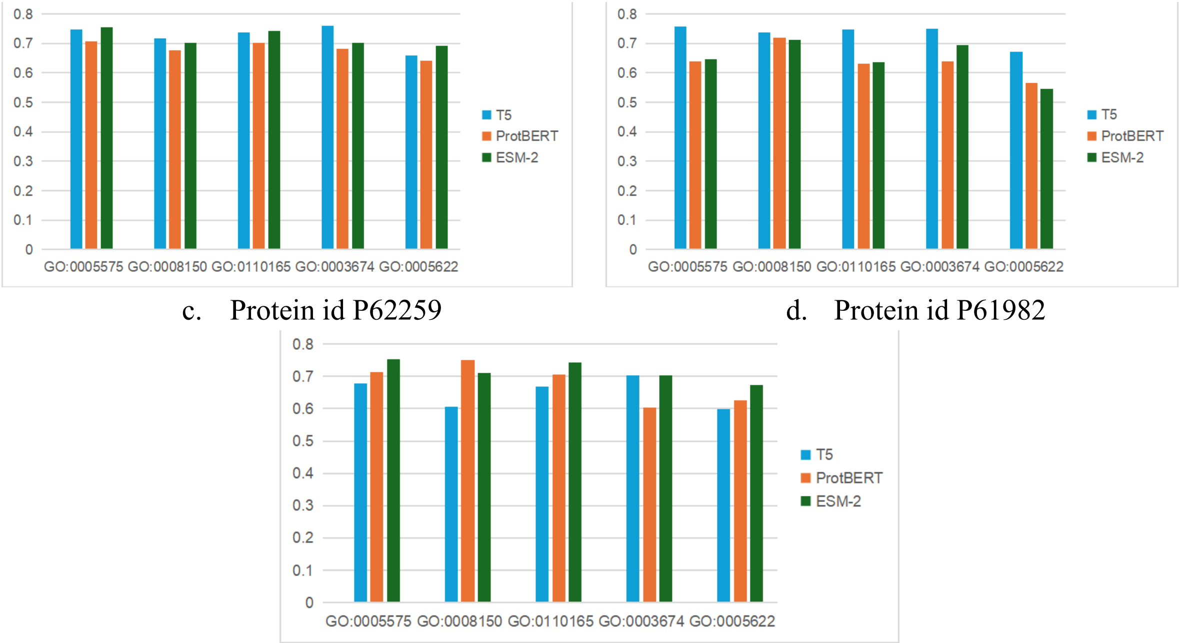


Figure 10. Comparison of Testing Results on 5 Protein Samples



a. Protein id Q9CQV8

b. Protein id P68510



c. Protein id P62259

d. Protein id P61982

e. Protein id O70456

Figure 11. The detail of testing result (a,b,c,d,e) for function prediction for each protein ID.

In this study, we analyzed the performance of three embedding models: T5, ProtBERT, and ESM-2, in predicting protein function based on several Gene Ontology (GO) Term IDs and Protein IDs. The analysis shows that each model performs differently depending on the specific GO Term ID and Protein ID. For GO:0005575 (Extracellular Region), the ESM-2 model generally provided better prediction results than the other models. For example, in the Q9CQV8 protein, ESM-2 produces the highest score of 0.772703, followed by T5, with a score of 0.761126 and ProtBERT, with a score of 0.696704. A similar pattern can be seen in the P62259 and O70456 proteins, where ESM-2 excelled with scores of 0.753928 and 0.754111, respectively. However, T5 delivered the best results for protein P68510 with a score of 0.738088, while for protein P61982, T5 also excelled with a score of 0.757815. ProtBERT has shown quite competitive but not dominant performance during this GO Term. On GO:0008150 (Metabolic Process), the model performance could be more consistent, with no model being dominantly superior. ProtBERT presents the best results on proteins Q9CQV8 and O70456, with scores of 0.760884 and 0.750726, respectively. On the other hand, T5 excelled on proteins P62259 and P61982, with scores of 0.716674 and 0.737991, respectively. ESM-2 shows good but not dominant performance on this GO Term, with its best score on protein P68510 of 0.709408, but still below ProtBERT and T5.

GO:0110165 (Cellular Anatomical Entity) analysis reveals that ESM-2 provided the best overall prediction results. For proteins Q9CQV8 and P62259, ESM-2 gave the highest scores of 0.762396 and 0.742384. However, T5 excelled on proteins P68510 and P61982 with scores of 0.729143 and 0.748615, respectively. On protein O70456, ESM-2 again excelled with a score of 0.743279, indicating that this model can understand cellular anatomical entities well.

In GO:0003674 (Molecular Function), T5 consistently achieves better than other models. T5 provides the highest scores for all proteins except O70456, where ESM-2 slightly outperformed T5 with scores of 0.702123 and 0.702747. T5 shows the highest scores of 0.759577 for protein P62259 and 0.750126 for protein P61982, indicating that T5 has an advantage in understanding the molecular function of proteins.

For GO:0005622 (Intracellular), ESM-2 tends to be superior overall, giving the highest scores to four of the five proteins tested. For example, ESM-2 gives the highest scores to proteins Q9CQV8, P62259, P68510, and O70456 with scores of 0.700982, 0.692235, 0.612979, and 0.672331, respectively. However, in protein P61982, T5 gives the highest score of 0.673006. ProtBERT shows promising results but is not dominant on this GO Term, with its best score on protein Q9CQV8 of 0.623393. Overall, the ESM-2 model provides the best prediction results for most GO Term IDs tested, especially for GO:0005575, GO:0110165, and GO:0005622. This suggests that ESM-2 can better capture the complex relationships in amino acid sequences when optimised explicitly for protein data. However, T5 performs very well and even outperforms some GO Term IDs, especially for GO:0003674, demonstrating its ability to understand proteins' sequence context and molecular function. ProtBERT, although not consistently superior, performs competitively on some proteins and specific GO Term IDs, especially for GO:0008150, suggesting that this model still has the potential to predict protein function depending on its specific context.

In T5 embedding, the T5-Large model configuration is used, which has 24 encoder layers and 24 decoder layers. The encoder generates an internal representation of the input sequence, while the decoder transforms this representation into the desired output sequence. Each layer in the encoder and decoder is equipped with self-attention and feed-forward neural networks, with the decoder also utilizing cross-attention to interact with the encoder. With a hidden size of 1024 and 16 attention heads, this T5 model can capture various essential aspects of the protein sequence in detail. There are several variants of T5: T5-Small has 60 million parameters, T5-Base has 220 million parameters, T5-Large has 770 million parameters, and T5-3B has 3 billion parameters. Using T5-Large, which has 770 million parameters, this model can capture more information and complex relationships in protein sequences, resulting in a more accurate and detailed representation for tasks such as protein function prediction.

ProtBERT is a variant of the BERT model specifically adapted for protein data. This embedding model uses a bidirectional transformer architecture, allowing for a bidirectional understanding of the amino acid sequence context, both forward and backwards. This is important because information in an amino acid sequence can be contextual and dependent on previous and subsequent amino acid sequences. ProtBERT is specifically designed to capture patterns of relationships in amino acid sequences better than models that do not use a bidirectional approach. ProtBERT has 110 million parameters with 12 encoder layers, each with 12 attention heads. Each attention head allows the model to focus on different parts of the amino acid sequence simultaneously, thus capturing different types of relationships and interactions between amino acids. The hidden size in ProtBERT is 1024, meaning each token (amino acid) is represented as a vector with 1024 dimensions. This representation is rich in contextual information that is valuable for various bioinformatics tasks.

The bidirectional transformer architecture ProtBERT uses consists of several key components, including a self-attention mechanism and a feed-forward neural network. The self-attention mechanism allows the model to consider each amino acid in the sequence relative to all other amino acids, providing a rich and comprehensive context. The feed-forward neural network processes this information to produce a more robust and informative final representation. In addition, ProtBERT is trained on an extensive protein sequence data set, making it able to understand and recognize various patterns and motifs in amino acid sequences. This training process allows ProtBERT to learn effective representations of protein sequences, which can be used for various applications such as protein function prediction, secondary structure determination, and protein-protein interaction analysis.

ESM-2 is a transformer model specifically optimized for protein data. This model is designed to capture complex relationships in amino acid sequences by leveraging evolutionary information contained in protein sequences. ESM-2's ability to leverage evolutionary information allows the model to understand a broader biological context, which is essential for predicting complex and specific protein functions. ESM-2 has a variable number of parameters, ranging from 8 million to 15 billion, with 33 layers. With a large enough number of parameters and well-tuned, ESM-2 can learn and understand complex patterns in protein data, making it very effective for tasks such as protein structure prediction and protein functional classification. In understanding patterns in protein data, having more parameters in a model like ESM-2 helps the model capture more complex and detailed patterns. With a larger capacity, the model can learn a more detailed and accurate representation of the protein data, resulting in better and more accurate predictions.

From the analysis above, it can be seen that ESM-2 often excels in many categories of protein function, mainly because its architecture is optimized for protein data. T5 also shows strong performance in many cases, although only sometimes superior. ProtBERT shows variable performance and, in some cases, is inferior to ESM-2 and T5. The best model selection depends on the data type and task specifications, but ESM-2 seems to be a superior choice for complex and specific protein function prediction tasks.

Table 3. The Detail Comparison Function Prediction Result for Each Embedding Model

Protein_ID	Predict Go_Term & Ontology Protbert	Prob.score Eksperimen Protbert	Hit Rate Protbert	Predict Go_Term & Ontology T5-embedding	Prob.score Eksperimen T5-embedding	Hit Rate T5-embedding	Predict Go_Term & Ontology ESM-2	Prob.score Eksperimen ESM-2	Hit Rate ESM-2
Q9CQV8	GO:0005575 GO:0008150 GO:0110165 GO:0003674 GO:0005622 GO:0043226 GO:0043229 GO:0005488 GO:0043227	0.7196 0.6960 0.7162 0.5905 0.6624 0.5760 0.5627 0.5284 0.5169	66,6%	GO:0071944 GO:0005575 GO:0110165 GO:0016020 GO:0005886 GO:0005515 GO:0005488 GO:0003674 GO:0008150 GO:0005622	0.3736 0.6243 0.6210 0.3702 0.3674 0.3671 0.4419 0.5684 0.7690 0.5732	60%	GO:0005575 GO:0008150 GO:0110165 GO:0003674 GO:0005622 GO:0043226 GO:0043229 GO:0005488 GO:0043227	0.7727 0.7118 0.7623 0.6851 0.7009 0.5258 0.6129 0.5843 0.5917	66.67%
P62259	GO:0005575 GO:0008150 GO:0110165 GO:0003674 GO:0005622 GO:0043226 GO:0043229 GO:0005488 GO:0043227	0.7141 0.6769 0.7111 0.6027 0.6569 0.5804 0.5650 0.5583 0.5217	100%	GO:0005622 GO:0043229 GO:0031981 GO:0043226 GO:0110165 GO:0070013 GO:0005730 GO:0043233 GO:0043232 GO:0005575 GO:0043227 GO:0031974 GO:0005634 GO:0043231 GO:0043228 GO:0005488 GO:0008150 GO:0003674	0.6888 0.5989 0.3430 0.6172 0.7481 0.3322 0.3242 0.3354 0.3160 0.7501 0.5587 0.3274 0.3202 0.3380 0.3281 0.5943 0.5786 0.6206	66.67%	GO:0005575 GO:0008150 GO:0110165 GO:0003674 GO:0005622 GO:0043226 GO:0043229 GO:0005488 GO:0043227	0.7539 0.7023 0.7423 0.7019 0.6922 0.6168 0.5937 0.6357 0.5305	100%
P68510	GO:0005575 GO:0008150 GO:0110165 GO:0003674 GO:0005622 GO:0043226 GO:0043229 GO:0005488 GO:0043227	0.7366755 0.6632989 0.7336384 0.5833655 0.6826415 0.6067184 0.5915128 0.53948826 0.54647386	66,6%	GO:0005575 GO:0008150 GO:0110165 GO:0003674 GO:0005622 GO:0009987 GO:0005488	0.6500436 0.77116716 0.6332418 0.661271 0.53523976 0.5620894 0.54087865	100%	GO:0005575 GO:0008150 GO:0110165 GO:0003674 GO:0005622 GO:0009987 GO:0043229 GO:0005488	0.7041 0.7094 0.6935 0.6951 0.6129 0.5159 0.5070 0.5979	100%
P61982	GO:0005575 GO:0008150 GO:0110165 GO:0003674 GO:0005622 GO:0043226 GO:0043229 GO:0005488 GO:0043227	0.7456174 0.6551619 0.7427267 0.58277696 0.68954253 0.6168528 0.60087526 0.53723305 0.55723524	66,6%	GO:0005575 GO:0008150 GO:0110165 GO:0003674 GO:0005622 GO:0009987 GO:0005488	0.67450005 0.7457478 0.65944827 0.62957084 0.55256057 0.5431961 0.52205485	100%	GO:0005575 GO:0008150 GO:0110165 GO:0003674 GO:0005622 GO:0009987 GO:0005488	0.6468 0.7118 0.6367 0.6941 0.5448 0.5059 0.5970	100%
O70456	GO:0005575 GO:0008150 GO:0110165 GO:0003674 GO:0005622 GO:0043226 GO:0005488	0.65836674 0.7150305 0.6545256 0.5959167 0.59024465 0.50187826 0.5188333	100%	GO:0005575 GO:0008150 GO:0110165 GO:0003674 GO:0005622 GO:0009987 GO:0043226 GO:0043229 GO:0005488 GO:0043227	0.82507706 0.6729239 0.8145351 0.63139975 0.7351699 0.53219 0.6408058 0.6083025 0.55399746 0.54590976	100%	GO:0005575 GO:0008150 GO:0110165 GO:0003674 GO:0005622 GO:0009987 GO:0043226 GO:0043229 GO:0005488 GO:0043227	0.7541 0.7109 0.7432 0.7021 0.6723 0.5125 0.5896 0.5622 0.6243 0.5022	100%

Table 4. Summary of Hit Rate Comparison Function Prediction Result for Each Embedding Model

Protein_ID	Hit Rate Eksperimen ProtBERT	Hit Rate Eksperimen T5-Embedding	Hit Rate Eksperimen ESM-2
Q9CQV8	66.6%	60%	66.67%
P62259	100%	66.67%	100%
P68510	66.6%	100%	100%
P61982	66.6%	100%	100%
O70456	100%	100%	100%

Tables 3 and Table 4 compare the performance of protein function prediction based on sequence, measured using the Hit Rate on five protein IDs, namely Q9CQV8, P62259, P68510, P61982, and O70456. The results show that the performance of the LSTM model using the ESM-2 embedding model gives better results, where 54 samples get a hit rate of 100% and 1 sample gets above 60%. This shows consistent and convincing protein predictions on ESM-2

CONCLUSION

The most acceptable embedding for protein function prediction based on sequence effectively captures the intricate relationships within the protein's amino acid sequence, including evolutionary, structural, and functional patterns. Based on the study and experiment, three embedding models produce good training results (accuracy > 99%, precision > 60%, and recall > 70%) and prediction results (hit rate > 79%). The ESM-2 embedding model, trained explicitly on large-scale protein sequence datasets, stands out for its ability to model long-range dependencies and complex interactions within the sequence. ESM-2 embedding model generates high-dimensional embeddings that accurately reflect the biological nuances of proteins, making them particularly effective for function prediction tasks (the training accuracy above 99% and average prediction hit rate 93.33%). By utilizing embeddings tailored to protein sequences and trained on vast evolutionary data, researchers can achieve more accurate and reliable predictions, ultimately advancing our understanding of protein functions and their implications in biology and medicine.

Future work in protein function prediction based on the sequence will likely focus on enhancing the accuracy and scalability of protein embeddings by integrating multimodal data and more advanced deep learning architectures. One promising direction is incorporating structural and experimental data alongside sequence information to create richer embeddings that capture the linear sequence and the three-dimensional conformation of proteins. Additionally, exploring more sophisticated transformer models or hybrid approaches that combine evolutionary information with functional annotations could improve predictive power. Training models on even larger and more diverse protein datasets, possibly incorporating unsupervised or self-supervised learning methods, will be crucial as computational resources expand. Furthermore, fine-tuning these models for specific protein families or functions and making them more interpretable will be essential for applying them in practical biological and clinical settings, such as drug discovery and personalized medicine.

ACKNOWLEDGMENT

This work is supported by the Ministry of Education, Culture, Research and Technology of the Republic of Indonesia through Funding for Research and Community Service Programs for Fiscal Year 2024.

REFERENCES

[1] M. Kulmanov e R. Hoehndorf, “DeepGOPlus: improved protein function prediction from sequence,” *Bioinformatics,* vol. 36, nº 2, p. 422–429, 2020.

[2] M. Lee, “Recent Advances in Deep Learning for Protein-Protein Interaction Analysis: A Comprehensive Review,” *Molecules,* vol. 28, nº 13, p. 5169, 2023.

[3] F. Soleymani , E. Paquet, H. Viktor, W. Michalowski e D. Spinello, “Protein–protein interaction prediction with deep learning: A comprehensive review,” *Computational and Structural Biotechnology Journal,* vol. 20, pp. 5316-5341, 2022.

[4] R. You, Z. Zhang, Y. Xiong, F. Sun, H. Mamitsuka e S. Zhu, “GOLabeler: improving sequence-based large-scale protein function prediction by learning to rank,” *Bioinformatics,* vol. 34, nº 14, p. 2465–2473, 2018.

[5] M. Kulmanov, M. A. Khan e R. Hoehndorf, “DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier,” *Bioinformatics,* vol. 34, nº 4, p. 660–668, 2018.

[6] M. Kulmanov e R. Hoehndorf, “DeepGOZero: improving protein function prediction from sequence and zero-shot learning based on ontology axioms,” *Bioinformatics,* vol. 38, p. i238–i245, June 2022.

[7] S. Yao, R. You, S. Wang, Y. Xiong, X. Huang e S. Zhu, “NetGO 2.0: improving large-scale protein function prediction with massive sequence, text, domain, family and network information,” *Nucleic Acids Research,* vol. 49, nº W1, p. W469–W475, 2021.

[8] S. Makrodimitris, M. J. T. Reinders e R. C. H. J. van Ham, “Metric learning on expression data for gene function prediction,” *Bioinformatics,* vol. 36, nº 4, p. 1182–1190, 2020.

[9] E. Lavezzo, M. Falda, P. Fontana, L. Bianco e S. Toppo, “Enhancing protein function prediction with taxonomic constraints – The Argot2.5 web server,” *Methods,* vol. 93, pp. 15-23, 2016.

[10] R. You, S. Yao, Y. Xiong, X. Huang, F. Sun, H. Mamitsuka e S. Zhu, “NetGO: improving large-scale protein function prediction with massive network information,” *Nucleic Acids Research,* vol. 47, nº W1, p. W379–W387, 2019.

[11] B. Dunham e M. K. Ganapathiraju, “Benchmark Evaluation of Protein–Protein Interaction Prediction Algorithms,” *Molecules,* vol. 27, nº 1, pp. 1-21, 22 December 2021.

[12] K. M. Verspoor , “Roles for Text Mining in Protein Function Prediction,” *Biomedical Literature Mining,* vol. 1159, p. 95–108, 2014.

[13] Z. Gao, J. Chenran, J. Zhang, X. Jiang, L. Li, P. Zhao, H. Yang, Y. Huan e J. Li, “Hierarchical graph learning for protein–protein interaction,” *Nature Communications,* vol. 14, p. 1093, 25 February 2023.

[14] K. Jha, K. Sourav e S. Saha , “Graph-BERT and language model-based framework for protein–protein interaction identification,” *Scientific Reports,* vol. 13, p. 5663, 06 April 2023.

[15] J. Pereira, A. J. Simpkin, M. D. Hartmann, D. J. Rigden, R. M. Keegan e A. N. Lupas, “High-accuracy protein structure prediction in CASP14,” *Proteins: Structure, Function, and Bioinformatics,* vol. 89, p. 1687–1699, 2021.

[16] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov e O. Ronneberger, “Highly accurate protein structure prediction with AlphaFold,” *nature ,* vol. 596, p. 583–589, 2021.

[17] V. Gligorijević, P. D. Renfrew, T. Kosciolek, J. K. Leman e D. Berenberg, “Structure-based protein function prediction using graph convolutional networks,” *nature,* vol. 12, p. 3168, 2021.

[18] A. Elnaggar, M. Heinzinger, C. Dallago, G. Rehawi, Y. Wang, L. Jones, T. Gibbs, T. B. Fehér, C. Angerer, M. Steinegger, D. Bhowmik e B. Rost , “ProtTrans: Towards Cracking the Language of Life’s Code Through Self-Supervised Deep Learning and High Performance Computing,” *bioRxiv,* 12 July 2020.

[19] A. Elnaggar, M. Heinzinger, C. Dallago, G. Rehawi, Y. Wang, L. Jones, T. Gibbs, T. Feher, C. Angerer, M. Steinegger, D. Bhowmik e B. Rost, “ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning,” *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 44, nº 10, pp. 7112-7127, October 2022.

[20] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li e P. J. Liu, “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer,” *Journal of Machine Learning Research,* vol. 21, pp. 1-67, 2020.

[21] Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, A. d. S. Costa, M. Fazel-Zarandi, T. Sercu, S. Candido e A. Rives, “Evolutionary-scale prediction of atomic level protein structure with a language model,” *bioRxiv,* 21 December 2022.

[22] N. Zhou, Y. Jiang, T. R. Bergquist e A. J. Lee, “The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens,” *bioRXiv,* pp. 1-48, 29 May 2019.

[23] CAFA, “https://biofunctionprediction.org/cafa/,” CAFA, 2024. [Online]. Available: https://biofunctionprediction.org/cafa/. [Acesso em 31 8 2024].

[24] J. Devlin, M.-W. Chang, K. Lee e K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” em *The 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019.

[25] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras e I. Androutsopoulos, “LEGAL-BERT: The Muppets straight out of Law School,” em *Findings of the Association for Computational Linguistics: EMNLP 2020*, Association for Computational Linguistics, 2020.

[26] L. Zheng, N. Guha, R. B. Anderson, P. Henderson e D. E. Ho, “When Does Pretraining Help? Assessing Self-Supervised Learning for Law and the CaseHOLD Dataset of 53,000+ Legal Holdings,” em *Proceeding Eighteenth International Conference for Artificial Intelligence and Law (ICAIL’21)*, New York, 2021.

[27] Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin , R. Verkuil, O. Kabeli, Y. Shmueli, A. D. S. Costa, M. Fazel-Zarandi, T. Sercu, S. Candido e A. Rives, “Evolutionary-scale prediction of atomic-level protein structure with a language model,” *Science,* vol. 379, nº 6637, pp. 1123-1130, 17 March 2023.

[28] J. Gong, L. Jiang, Y. Chen, Y. Zhang, X. Li, Z. Ma, Z. Fu, F. He, P. Sun, Z. Ren e M. Tian, “THPLM: a sequence-based deep learning framework for protein stability changes prediction upon point variations using pretrained protein language model,” *Bioinformatics,* vol. 39, nº 11, November 2023.