Received: 18 March 2025 | Revised: 30 September 2025 | Accepted: 18 November 2025 | Published online: 23 December 2025

RESEARCH ARTICLE

Medinformatics
2025, Vol. 00(00) 1–9
DOI: 10.47852/bonviewMEDIN52025721

BON VIEW PUBLISHING

Proteinext: Protein Function Prediction with Sequence Embeddings and Natural Language Processing

Hailey Ledenko[1], Luke Coleman[1], G. Alvarado[1], Tyler Stratton[1], Boen Liu[2], Jie Hou[3], Dong Si[4], Lei Zhang[5], Rui Ding[5], Yang Wang[5] and Renzhi Cao[1,*]

[1] *Department of Computer Science, Pacific Lutheran University, USA*

[2] *Annie Wright Schools, USA*

[3] *Department of Computer Science, St. Louis University, USA*

[4] *School of Science, Technology, Engineering & Mathematics, University of Washington, USA*

[5] *Information Material and Intelligent Sensing Laboratory of Anhui Province, Anhui University, China*

Abstract: Proteins are fundamental to life, as they support vital processes in the body such as muscle development, cell growth, tissue repair, and immune defense. However, their complex structures and diverse functions make them challenging to fully understand. While recent advances enable efficient and accurate protein structure prediction, the challenge of predicting protein function remains. Although promising, current prediction methods suffer from slow performance, high computational demands, and struggle with handling highly specific proteins. Due to a rapid expansion of protein sequence databases, a computational method for predicting function directly from sequence is critical. Our solution to this ongoing challenge is Proteinext, an innovative method for protein function prediction that leverages advanced sequence representations and natural language processing (NLP) techniques. Proteinext leverages Meta's 15B-parameter evolutionary scale modeling to generate protein sequence embeddings, which are refined using a fine-tuned BigBird transformer-based NLP model. This combination results in a powerful model and method that significantly improves protein function prediction. The model was trained on 372,683 protein sequences from a combined dataset of Gene Ontology and Universal Protein Knowledgebase annotations. Proteinext represents a major step toward comprehensively understanding and predicting protein functions, achieving an impressive F_{max} score of 0.74 and S_{min} score of 0.39. This work underscores the potential of combining computational biology with NLP to address critical challenges in proteomics. Proteinext is available at https://github.com/Cao-Labs/AlphaAnalyzer.

Keywords: protein function prediction, machine learning, natural language processing

1. Introduction

Proteins, the building blocks of life, are the molecules that perform work in organic cells. Made up of chains of amino acids, proteins catalyze biochemical reactions, transport nutrients, recognize and transmit signals, and much more [1–6]. Protein characteristics are classified in three main ways: sequence, structure, and function. The sequence of a protein refers to the order in which amino acids are connected in the chain; this unique quality determines the structure, which is a three-dimensional representation of the protein's shape. Both the sequence and structure of a protein determine its function. Function can be defined in many ways, but in simple terms, it is the task(s) that the protein performs [7]. Gene Ontology (GO) notation is a standardized vocabulary developed to describe complex and multi-layered functions [8].

Understanding and accurately annotating protein function is crucial for expanding our knowledge of life at the molecular level. By identifying and characterizing protein functions, researchers can uncover the mechanisms underlying health, disease, and evolution, leading to advancements in drug discovery, genetic engineering, and synthetic biology. Furthermore, comprehensive protein annotation enhances our ability to interpret genomic data, paving the way for innovations in personalized medicine and biotechnological applications [9–15].

Traditionally, protein function is determined through experimental methods, which involve a combination of biochemical assays, genetic studies, high-throughput screening, and structural analysis techniques including X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, and cryo-electron microscopy [16–18]. However, this process requires tremendous amounts of time and resources, leading to an increased reliance on computational methods and Artificial Intelligence (AI) AI-driven predictions to complement and accelerate functional annotation

*Corresponding author: Renzhi Cao, Department of Computer Science, Pacific Lutheran University, USA. Email: caora@plu.edu

efforts. As protein sequencing technologies continue to decline in cost and become more reliableAs, advances in mass spectrometry-based proteomics, next-generation sequencing, and AI-driven annotation methods have been significantly enhanced, allowing our ability to identify and catalog proteins across diverse species and the size of protein databases, in ways like never before [2, 19–21]. Publicly available databases such as UniProt [22], PDB [23], AlphaFold DB [24–26], and Pfam [24, 27] are rapidly growing, incorporating newly sequenced proteins from previously unexplored organisms and environmental samples. This influx of data is further accelerated by machine learning and computational modeling, which aid in predicting protein structures, interactions, and potential functions with increasing accuracy. Given the rapid expansion of protein databases and the sheer volume of newly sequenced proteins, traditional experimental methods alone are insufficient to keep up with the growing demand for functional annotation. Therefore, a computational approach to extracting protein function directly from its sequence information is not just beneficial but essential in modern biological research [20, 28–30]. By leveraging machine learning, deep learning, and bioinformatics algorithms, computational methods can predict protein function based on sequence homology, evolutionary relationships, structural motifs, and biochemical properties. These approaches enable the rapid and scalable identification of protein roles, interactions, and potential applications in drug discovery, genetic engineering, and disease research [28, 31–33].

Many machine learning methods have achieved success in predicting protein function. Some notable methods are as follows: DeepGOPlus can make predictions based only on sequence data by utilizing a convolutional neural network (CNN) [34]. DeepAdd makes use of two CNNs and integrates the NLP Word2Vec to predict protein function [35]. Protein annotation with Z-score (PANNZER) uses a weighted k-nearest neighbor model to predict based on sequence data [15, 36]. As with all machine learning models, however, the accuracy of predictions varies depending on the specific model and data used for training. Protein sequences and structures are very complex, resulting in a nuanced "language" that models must understand and predict.

Language processing techniques such as Natural Machine Translation are able to treat protein sequences and GO notation as a "language" of their own. For example, ProLanGO2 uses an encoder-decoder network of two recurrent neural networks to translate protein language into GO notation [1]. HiFun retrieves all reviewed protein sequences from the UniProt database to train the FastText sequence embedding model [37]. SPROF-GO first extracts sequence data using the protein language model ProtT5-XL-U50. Then, the sequence data is fed to two multilayer perceptron (MLP) , which are used to predict GO terms [38].

Beyond selecting optimal machine learning techniques, the way protein sequences are represented plays a crucial role in ensuring the reliability of results. While some approaches, like ProLanGO2, have developed custom systems for encoding protein sequences, the most effective methods leverage predictive protein structure models to enhance accuracy and functional insights. Evolutionary scale modeling (ESM) is Facebook's 15-billion-parameter protein language model used to predict structure, function, and other protein features. These predictions are in the form of an embedding, extracted from input sequence data. Although AlphaFold2 is the current state-of-the-art algorithm performing similar functions as ESM, it is 6× slower because it requires accessing a database to perform homology-based comparisons [39, 40]. Methods that utilize ESM tend to achieve better F_{max} and S_{min} scores than their baseline methods. For example, Transformers for high-performance language modeling (THPLM) uses an encoder-decoder transformer, which takes protein sequences and single-point variations generated by ESM-2 as input [41]. TransFun combines ESM's protein language model with AlphaFold2's predicted 3D structures to make predictions using graph neural networks [42]. Hierarchical embedding attention learning (HEAL) uses a hierarchical graph transformer, a graph convolutional network, and a multi-layer neural network along with sequence data from ESM-1b to predict function [25]. Each of these methods resulted in improved accuracy and a greater ability to extend to newly discovered proteins beyond those homologous with existing annotations.

Despite significant advancements in computational protein function prediction, existing methods still face several limitations in scope and applicability. THPLM, although very accurate, predicts only ΔΔG (changes in protein stability upon mutation) and not the function of a protein [41]. HEAL, another state-of-the-art method, relies on experimentally determined or computationally predicted protein structures as input, making it computationally expensive and time-intensive, thus limiting its scalability for large-scale functional annotation tasks [25]. Additionally, TransFun, though effective at predicting general functional categories, struggles with more specific GO terms, particularly those related to specialized biological pathways. For instance, it can annotate proteins under the broad category of "metabolic process," but it lacks precision in assigning deeper, more refined terms like "generation of precursor metabolites," which are critical for understanding context-specific protein functions [42]. To overcome these challenges and provide a more scalable, precise, and computationally efficient approach to protein function prediction, we introduce Proteinext—a novel framework designed to expand the capabilities of current methods while addressing their inherent limitations.

Proteinext advances the field of protein function prediction by addressing critical limitations of prior methods by uniquely combining large language and natural language processing models. The key contribution is the ability to accurately assign GO terms to proteins using their amino acid sequences. The framework is unique in its two-step process: first, it generates high-dimensional vector representations (embeddings) of protein sequences using the ESM-2 model, and second, it fine-tunes the BigBird natural language processing (NLP) model to classify these embeddings. This overcomes the input length limitations of previous models like bidirectional encoder representations from transformers (BERT), which cannot process the long sequences typical of protein embeddings without losing data. The dataset used to develop the model was a merging of the UniProtKB/Swiss-Prot and the GO knowledgebase, which resulted in 372,683 training entries. Proteinext addresses many of the limitations of current alternative models. Unlike THPLM, which is restricted to predicting protein stability changes (ΔΔG) rather than functional annotations, Proteinext directly predicts GO terms with fine-grained accuracy. It mitigates HEAL's dependence on 3D structural inputs, making it significantly more scalable and less computationally intensive, while surpassing TransFun in specificity, especially for deeper GO terms essential to understanding context-specific biological processes. Proteinext achieves high accuracy without requiring structural models or complex preprocessing steps by leveraging lightweight yet expressive ESM embeddings and a streamlined neural network architecture. This makes it uniquely suited for large-scale annotation of novel proteins, even those lacking close homologs, setting a new benchmark for efficient and precise protein function prediction. Ultimately, Proteinext provides a powerful, scalable, and

accessible solution poised to accelerate discovery across the biological sciences.

2. Materials and Methods

Proteinext utilizes ESM-2 and NLP to produce accurate protein function predictions given the sequence information. Below is an in-depth look at how Proteinext works and our workflow process.

2.1. Data preparation

GO provides a standardized framework to describe the functions, biological processes, and cellular locations of genes and their products across various species. It aims to unify the representation of gene and protein attributes, facilitating consistent annotations and data analysis in genomics and bioinformatics. The first dataset we utilized is the GO knowledgebase, which is available publicly at https://geneontology.org/docs/download-ontology/. The 2024-06-17 release of the GO database contains 42,093 entries. Annotations within each sub-ontology are structured into nodes of a directed acyclic graph, where the edges between nodes represent the relationships between protein functions [8]. Parent nodes are always broader than the children, meaning the deeper you go, the more specific and complex the node is.

The second dataset we used is from the Universal Protein Knowledgebase (UniProtKB). The public dataset can be downloaded here: https://www.uniprot.org/help/downloads/. It is split into two sections: Swiss-Prot and TrEMBL. Swiss-Prot contains manually annotated proteins, and TrEMBL contains proteins that are computationally found [22]. We only used UniProtKB/Swiss-Prot entries to train our model to avoid introducing other layers of computation into our data. We used the 2024_03 release, which contains 571,609 reviewed protein entries. To filter the data, we removed entries that contained sequences with length >= 30,000 and entries with duplicate sequences. After filtering, we were left with 483,428 unique entries.

To create our training and testing datasets, we first cross-referenced and merged the GO terms with the filtered Uniprot data. If a GO term did not have a matching ID in the Uniprot data, it was omitted. We then generated embeddings for each protein annotation using ESM-2 (650M), with a maximum embedding length of 1,024. Each embedding was rounded to three decimal places and appended to the input dataset. Our final processed dataset was partitioned into training and testing datasets using an 80:20 split, which resulted in 372,683 training and 93,172 testing entries, and all data can be found in Figure 1.

2.2. Model architecture and training

Proteinext can be divided into two steps, as described in Figure 2. First, protein sequences are transformed into numerical representations using pre-trained models. These embeddings capture complex biochemical and structural features of the sequences in a form that can be processed by machine learning algorithms. Second, these embeddings serve as input for fine-tuning advanced NLP models. By adapting these models specifically for protein data, Proteinext can improve its predictive accuracy for the protein function prediction problem.

We used ESM-2 to generate high-dimensional vector representations of protein sequences. This is beneficial to the accuracy of Proteinext because these embeddings capture the contextual relationships between amino acids that raw protein sequences can't. Using the 650M parameter model, we generated embeddings of length <= 1,024, extracting on the 33rd layer. Due to system restrictions, we randomly chose 350,000 entries from the training data. Then, we appended the embeddings to their respective rows in the training data, along with the protein ID and GO term(s).

We treat protein function annotation as a multi-label classification problem, allowing us to fine-tune pre-trained NLP models such as BERT, Longformer, and BigBird [43–45]. We experimented with these models throughout the process of designing Proteinext.

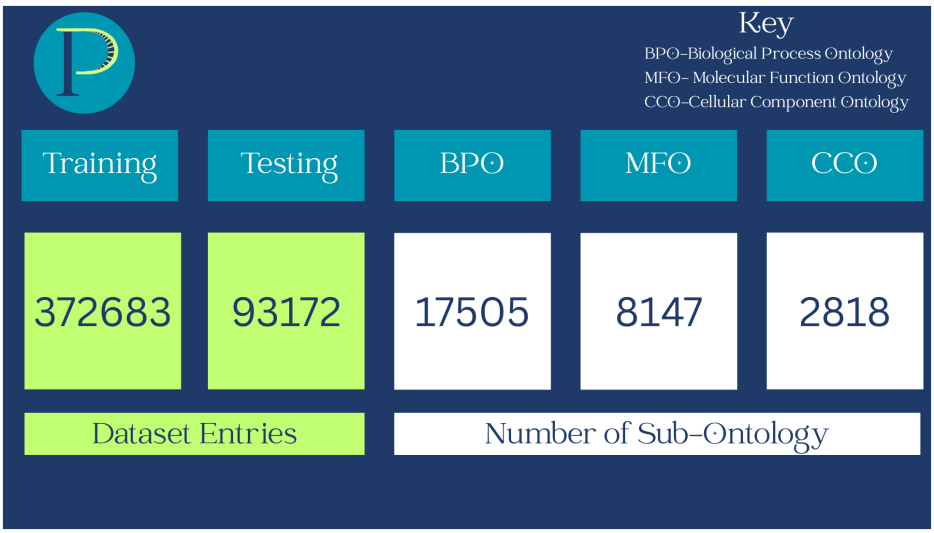

Figure 1. The number of entries in training and testing datasets after filtering and the number of unique GO terms in each sub-ontology

Note: BPO: biological process ontology; MFO: molecular function ontology; CCO: cellular component ontology

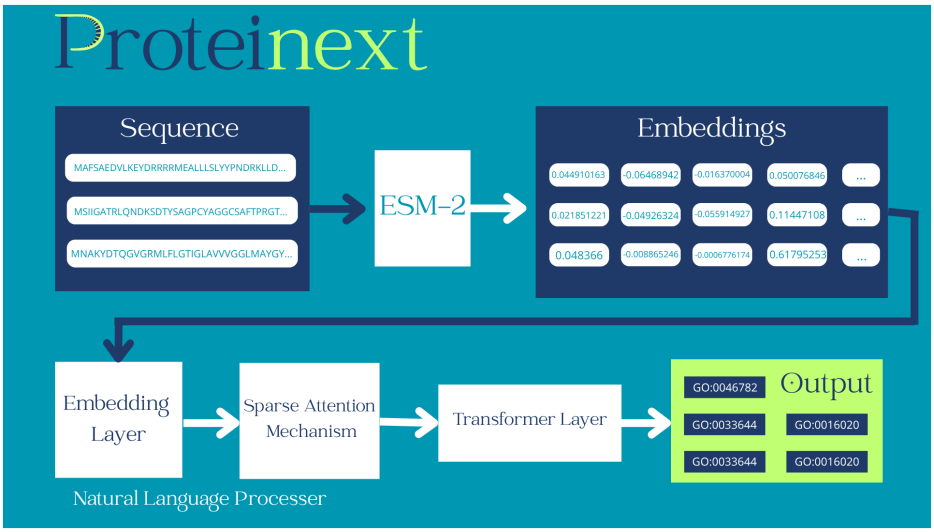

Figure 2. Flow chart of Proteinext

BERT is a popular, well-documented, transformer-based NLP model developed by Google. We were able to fine-tune BERT to our specific use case by utilizing the scikit-learn package's multi-label binarizer, which transformed a dictionary of every current GO term into a binarized format for classification. We then tuned it to make predicted classifications on ESM embedding representations of proteins. Although BERT is extremely robust, it is not without limitations. BERT's time and space complexities are both quadratic, which translates into a resource-heavy process to make predictions. Additionally, the model has a maximum token input length of 512, which is not suitable for long sequences. By default, ESM protein embeddings contain 1,024 tokens each, so in order to use BERT embeddings had to be truncated [43]. Multiple strategies of truncation have been developed, but all lead to data loss [46]. Each embedding is important to the overall representation of the protein sequence, so cutting out data could lead to inaccurate predictions. Because of this, we applied principal component analysis (PCA) to reduce the length of the embeddings and computational efficiency and to mitigate potential noise in the embedding space [47]. PCA is a technique that reduces the size of a dataset by selecting the most important elements. In this case, we reduced embeddings to length 510, leaving room for the start and end tokens of each sentence. We fine-tuned BERT with a learning rate of 1e-5 and 5 epochs. We tested on 500 entries with a threshold of 0.2. Utilizing BERT and ESM, we were able to achieve an F1 score of 0.83.

Longformer is an NLP that is designed to address BERT's limitations, allowing input sequences up to length 4,096 [44]. It uses a sparse attention mechanism instead of the traditional dense attention mechanism used in BERT, creating a model that is computationally cheaper than other NLPs. While the sliding window method reduces space and time complexity to linear, the model is still receiving sequence inputs that are double the length of BERT. This leads to the downfall of our use of this model; resources for Longformer are extremely limited, and it requires double the time to train compared to BERT. Due to this, we were unable to train and test Longformer for protein function prediction.

BigBird is Google's attempt at extending BERT for longer sequences. Similar to Longformer, BigBird supports sequences up to length 4,096. BigBird was fine-tuned similarly to BERT, using a multi-label binarizer for classification. Its sparse attention mechanism combines random attention, global attention, and a sliding window to create a versatile model that is highly effective with long sequences [45]. Consequently, BigBird's time complexity is sub-quadratic, which makes it slower than Longformer. However, its effectiveness in predicting protein function allowed us to look past this downfall.

To fine-tune BigBird, we first loaded the pre-trained model from Hugging Face's Transformers library. For label preparation, we used scikit-learn's MultiLabelBinarizer to transform the GO knowledgebase into a binary matrix, where each valid GO term is represented as a separate column. The binarizer was initialized on the complete GO knowledgebase and then fitted on the subset of GO terms present in the training dataset to ensure consistent encoding across train/validation splits. For optimization, we began with the AdamW optimizer using a learning rate of 1e-5 and fine-tuned for 5 epochs. Each epoch processed approximately 43,750 mini-batches of sequence embeddings paired with their corresponding GO term labels. To stabilize training and reduce overfitting, we employed a learning rate scheduler with dynamic adjustment: whenever the validation loss plateaued, the learning rate was reduced by a factor of 0.5, with a minimum floor of 1e-7. During each of the five epochs, the model was trained on 43,750 batches of sequence embeddings and their corresponding GO term labels. Model checkpoints were saved at the end of each epoch, with the best checkpoint selected based on validation loss.

In summary, our new method integrates state-of-the-art protein language models with advanced NLP architectures in a way that adapts their strengths to the challenges of protein function prediction. In the first stage, high-dimensional embeddings generated by ESM-2 capture biochemical and structural relationships between amino acids, providing a rich numerical foundation beyond raw sequence input. These embeddings are then coupled with fine-tuned transformer models—BERT, Longformer, and BigBird—to handle the classification of protein functions framed as a multi-label problem. To address sequence length and computational bottlenecks, our method introduces practical adaptations, such as PCA-based dimensionality reduction to preserve information while fitting within BERT's token limits and leveraging the sparse attention mechanisms of Longformer and BigBird to extend context handling up to 4,096 tokens. This integration of pre-trained protein embeddings with adapted NLP models is novel

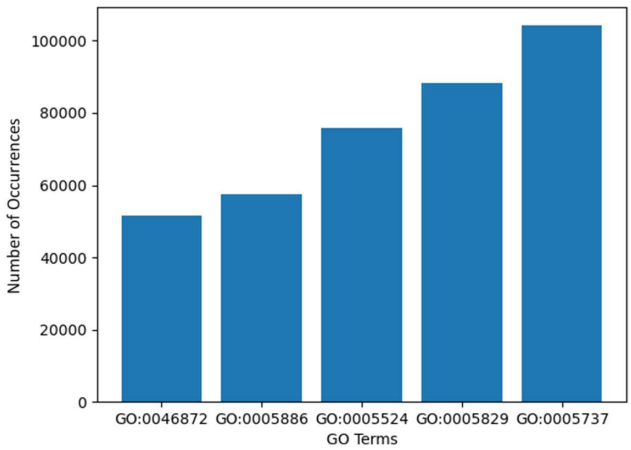

Figure 3. Most common GO terms from our training dataset

in its ability to bridge biological sequence data and transformer architectures, yielding improved predictive accuracy while balancing efficiency, scalability, and robustness across different model constraints.

3. Results and Discussion

3.1. Common GO terms and relationship between GO term and sequence length

First of all, we removed all protein sequences with length longer than 1,024, and because of that, 246,603 protein sequences were removed out of 465,853 sequences.

Next, we graphed the five most common GO terms as described in Figure 3. GO:0005737 was the most common. It refers to cytoplasm, which is the liquid that fills the inside of a cell. Next was GO:0005829, which refers to cytosol—a component of cytoplasm that contains particulate matter, such as protein complexes. The third most common GO term, GO:0005524, refers to adenosine 5'-triphosphate (ATP) binding, which is when a protein binds to ATP. With this, the proteins are given energy to work with. The fourth and fifth most common terms were GO:0005886 and GO:0046872, which corresponded to plasma membrane and metal ion binding. All five of these components are crucial for most cells to function and therefore are incredibly common.

We specifically monitored predictions for the most common GO terms to ensure that Proteinext is not simply biased toward predicting frequent labels, which could artificially inflate performance metrics. By checking the model's outputs on these high-frequency terms, we can verify that Proteinext is learning meaningful patterns in the data rather than defaulting to naive predictions based on label prevalence. This step serves as a safeguard to confirm that our model captures true functional signals rather than relying on label imbalance.

Lastly, we created a scatter plot to show a very interesting and vital relationship, as shown in Figure 4. Despite average sequence length growing as the number of GO terms increases, the sequences with fewer GO terms tend to have far more cases of long amino acid sequences. This was surprising, as we initially theorized that having more GO terms would result in a greater sequence length. Upon further reflection, this observation may be explained by the functional complexity and domain structure of proteins. Proteins with very long sequences but few GO annotations may contain repetitive regions, large unstructured domains, or nonfunctional extensions (e.g., signal peptides, disordered regions, or large low-complexity regions) that do not contribute additional functional annotations. In contrast, proteins with multiple GO terms often have more modular domain architectures where distinct functional domains correspond to distinct GO annotations, but these individual domains may not necessarily require very long sequences. Another possible explanation is annotation bias: some very long proteins may be insufficiently annotated due to limited experimental characterization, resulting in fewer assigned GO terms despite their potential functional complexity. Conversely, well-studied multifunctional proteins may accumulate more GO terms even if their overall sequence length is moderate. This finding highlights the importance of considering both sequence content and annotation completeness in protein function prediction tasks. Simply relying on sequence length as a proxy for functional complexity can be misleading. It also emphasizes that models like Proteinext need to capture nuanced sequence features beyond simple length-based patterns to accurately predict GO annotations.

3.2. Results and discussion

We tested Proteinext on 500 unseen sets of sequence embeddings from the merged Uniprot and GO knowledgebase testing dataset. To calculate precision, a method to check the similarity of GO terms was necessary. We first had to propagate the GO tree, so we were able to visualize the relationships between GO terms. Then, we checked the similarity between predicted and actual terms for each sequence by comparing the distance of terms to the root node. We applied depth-first search to find the common ancestors and ultimately calculate the precision. Proteinext achieved an F_{max} score of 0.74 and an S_{min} score of 0.39, where F_{max} refers to the highest harmonic mean of precision and recall across all thresholds, indicating how well the model balances identifying correct functions while avoiding false positives. S_{min} refers to how far the model's predictions deviate from the true functional annotations; a low S_{min} demonstrates fewer and less severe errors in predicting the protein functions.

Our approach was rigorously evaluated using UniProt data, where it achieved noteworthy performance metrics: precision of 0.83 and recall of 0.59. These results demonstrate Proteinext's ability to balance accuracy and comprehensiveness, ensuring both high-confidence predictions and meaningful functional coverage. By combining the strengths of NLP models and ESM-based representations, Proteinext bridges the gap between sequence data and biological function, offering a scalable and efficient solution for large-scale protein annotation [32, 40, 48–50].

Proteinext is a powerful tool for predicting protein functions, but it has certain limitations that are actively being addressed through ongoing improvements. Although it couldn't use the entire dataset of 372,683 entries due to system limits, it still trained effectively on 94% of the data, with only a small impact on its performance. However, its precision (0.83) and recall (0.59)—measures of how accurate and thorough it is—could be better, especially in identifying more protein functions correctly. Right now, Proteinext relies only on protein sequences and doesn't use structural information, which is key to understanding how proteins work. Adding structural data from tools like AlphaFold could make its predictions much sharper by highlighting important

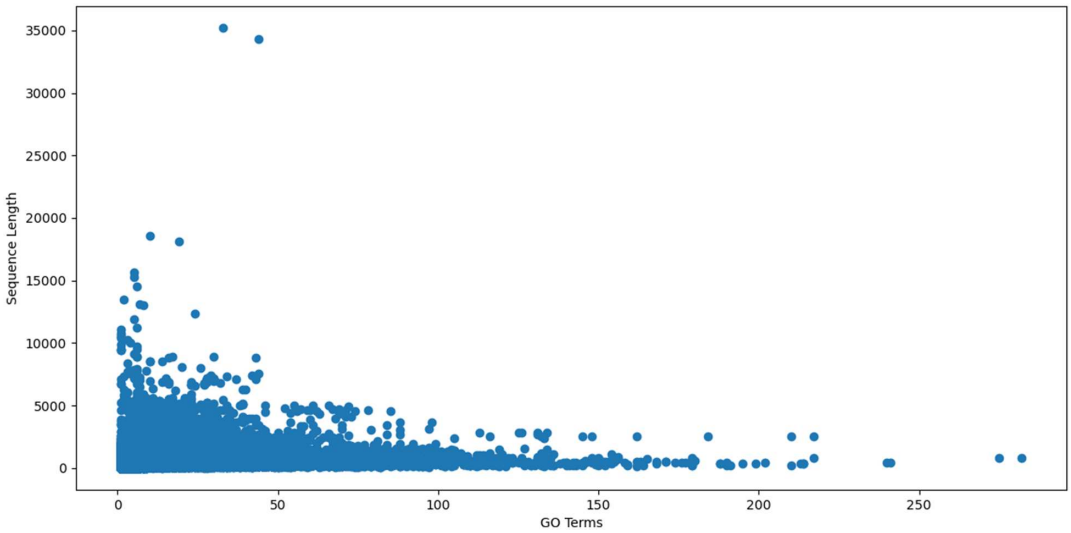

Figure 4. Sequence length based on the number of GO terms

features like binding sites. Moving forward, we plan to boost computing power to handle all the data, tweak the model to catch more protein functions, and blend in structural insights for richer, more precise results. These steps aim to make Proteinext an even stronger bridge between protein sequences and their roles in biology [51–54]. For future research, we plan to run comparisons of benchmarks between the Proteinext model and alternative state-of-the-art methods including DeepGO, THPLM, HEAL, and TransFun, which will more clearly demonstrate the advantages and weaknesses of the methods. While the current study followed a standard 80:20 split for the training and testing datasets, additional cross-validation would provide stronger evidence of reproducibility in the future.

4. Conclusion

Proteinext presents a promising advancement in protein function prediction by leveraging state-of-the-art NLP models and ESM-2 Sequence embeddings. The model achieved promising results: an F_{max} score of 0.74, an S_{min} of 0.39, a precision of 0.83, and a recall of 0.59. This highlights its effectiveness in making high-confidence and broad protein function predictions and suggests that even for proteins lacking close homologs in current databases, contemporary NLP tools can successfully interpret the intricate link between protein sequences and their biological functions. The model's impressive performance demonstrates how protein language models can directly derive significant functional signals from sequence data.

Proteinext's ability to use BigBird's sparse attention mechanism to analyze full-length ESM-2 embeddings (1,024 tokens) addresses an important limitation of conventional NLP models like BERT, which necessitate dimensionality reduction or truncation. Two significant advantages of Proteinext are scalability for large-scale annotation and enhanced specificity for fine-grained GO word prediction. Limitations still exist, though, including reliance on sequence data alone and a moderate recall (0.59). These outcomes are consistent with our findings, which are displayed in Figures 3 and 4 and indicate that rare terms had poorer accuracy, while longer sequences (>300 aa) and specific common GO terms (cytoplasm, ATP binding, etc.) were better predicted.

Proteinext effectively predicts protein function, but integrating additional features beyond sequence data could further improve the model. Future developments will incorporate structural information from AlphaFold predictions and primary/secondary structural embeddings, while upgrading to advanced language models such as ESM-3. The research team plans to enhance training datasets with underrepresented GO terms and rigorously benchmark performance against THPLM and TransFun. These enhancements will boost annotation accuracy for novel proteins while preserving the model's existing capability to predict fine-grained functional terms.

Proteinext accurately predicts protein functions from UniProtKB/Swiss-Prot sequences. The publicly available model bridges sequence data and biological function, enabling efficient large-scale protein annotation.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The data that support the findings of this study are openly available in GitHub at https://github.com/Cao-Labs/AlphaAnalyzer.

Author Contribution Statement

Hailey Ledenko: Conceptualization, Methodology, Software, Writing – original draft, Writing – review & editing, Visualization. **Luke Coleman:** Conceptualization, Methodology, Software, Writing – original draft, Writing – review & editing. **G. Alvarado:** Writing – original draft, Writing – review & editing, Visualization. **Tyler Stratton:** Writing – original draft, Writing – review & editing. **Boen Liu:** Writing – original draft, Writing – review

& editing. **Jie Hou:** Writing – original draft, Writing – review & editing. **Dong Si:** Writing – original draft, Writing – review & editing. **Lei Zhang:** Writing – original draft, Writing – review & editing. **Rui Ding:** Writing – original draft, Writing – review & editing. **Yang Wang:** Writing – original draft, Writing – review & editing. **Renzhi Cao:** Conceptualization, Writing – original draft, Writing – review & editing, Supervision, Project administration, Funding acquisition.

References

[1] Hippe, K., Gbenro, S., & Cao, R. (2020). ProLanGO2: Protein function prediction with ensemble of Encoder-Decoder networks. In *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, 1–6. https://doi.org/10.1145/3388440.3414701

[2] Jiang, Y., Oron, T. R., Clark, W. T., Bankapur, A. R., D'Andrea, D., Lepore, R., . . . , & Radivojac, P. (2016). An expanded evaluation of protein function prediction methods shows an improvement in accuracy. *Genome Biology*, *17*(1), 184. https://doi.org/10.1186/s13059-016-1037-6

[3] Zhou, N., Jiang, Y., Bergquist, T. R., Lee, A. J., Kacsoh, B. Z., Crocker, A. W., . . . , & Friedberg, I. (2019). The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *Genome Biology*, *20*(1), 244. https://doi.org/10.1186/s13059-019-1835-8

[4] Tran, T. K., Tran, M. C., Joseph, A., Phan, P. A., Grau, V., & Farmery, A. D. (2024). A systematic review of machine learning models for management, prediction and classification of ARDS. *Respiratory Research*, *25*(1), 232. https://doi.org/10.1186/s12931-024-02834-x

[5] Rost, B., Liu, J., Nair, R., Wrzeszczynski, K. O., & Ofran, Y. (2003). Automatic prediction of protein function. *Cellular and Molecular Life Sciences CMLS*, *60*(12), 2637–2650. https://doi.org/10.1007/s00018-003-3114-8

[6] Lawson, C. L., Kryshtafovych, A., Pintilie, G. D., Burley, S. K., Černý, J., Chen, V. B., . . . , & Chiu, W. (2024). Outcomes of the EMDataResource cryo-EM ligand modeling challenge. *Nature Methods*, *21*(7), 1340–1348. https://doi.org/10.1038/s41592-024-02321-7

[7] Eisenberg, D., Marcotte, E. M., Xenarios, I., & Yeates, T. O. (2000). Protein function in the post-genomic era. *Nature*, *405*(6788), 823–826. https://doi.org/10.1038/35015694

[8] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., . . . , & Sherlock, G. (2000). Gene ontology: Tool for the unification of biology. *Nature Genetics*, *25*(1), 25–29. https://doi.org/10.1038/75556

[9] Radivojac, P., Clark, W. T., Oron, T. R., Schnoes, A. M., Wittkop, T., Sokolov, A., . . . , & Friedberg, I. (2013). A large-scale evaluation of computational protein function prediction. *Nature Methods*, *10*(3), 221–227. https://doi.org/10.1038/nmeth.2340

[10] Vu, T. T. D., Kim, J., & Jung, J. (2024). An experimental analysis of graph representation learning for Gene Ontology based protein function prediction. *PeerJ*, *12*, e18509. https://doi.org/10.7717/peerj.18509

[11] Rigden, D. J. (2009). *From protein structure to function with bioinformatics*. Netherlands: Springer. https://doi.org/10.1007/978-1-4020-9058-5

[12] Zheng, L., Shi, S., Lu, M., Fang, P., Pan, Z., Zhang, H., . . . , & Zhu, F. (2024). AnnoPRO: A strategy for protein function annotation based on multi-scale protein representation and a hybrid deep learning of dual-path encoding. *Genome Biology*, *25*(1), 41. https://doi.org/10.1186/s13059-024-03166-1

[13] Ilgisonis, E. V., Pogodin, P. V., Kiseleva, O. I., Tarbeeva, S. N., & Ponomarenko, E. A. (2022). Evolution of protein functional annotation: Text mining study. *Journal of Personalized Medicine*, *12*(3), 479. https://doi.org/10.3390/jpm12030479

[14] Amiri-Dashatan, N., Koushki, M., Abbaszadeh, H.-A., Rostami-Nejad, M., & Rezaei-Tavirani, M. (2018). Proteomics applications in health: Biomarker and drug discovery and food industry. *Iranian Journal of Pharmaceutical Research*, *17*(4), 1523–1536.

[15] Yan, T.-C., Yue, Z.-X., Xu, H.-Q., Liu, Y.-H., Hong, Y.-F., Chen, G.-X., . . . , & Xie, T. (2023). A systematic review of state-of-the-art strategies for machine learning-based protein function prediction. *Computers in Biology and Medicine*, *154*, 106446. https://doi.org/10.1016/j.compbiomed.2022.106446

[16] Wang, S., Bian, J., Huang, X., Zhou, H., & Zhu, S. (2025). PubLabeler: Enhancing automatic classification of publications in UniProtKB using protein textual description and Pub MedBERT. *IEEE Journal of Biomedical and Health Informatics*, *29*(5), 3782–3791. https://doi.org/10.1109/JBHI.2024.3520579

[17] Liu, W., Wang, Z., You, R., Xie, C., Wei, H., Xiong, Y., . . . , & Zhu, S. (2024). PLMSearch: Protein language model powers accurate and fast sequence search for remote homology. *Nature Communications*, *15*(1), 2775. https://doi.org/10.1038/s41467-024-46808-5

[18] Yan, H., Wang, S., Liu, H., Mamitsuka, H., & Zhu, S. (2024). GORetriever: Reranking protein-description-based GO candidates by literature-driven deep information retrieval for protein function annotation. *Bioinformatics*, *40*(Supplement_2), ii53–ii61. https://doi.org/10.1093/bioinformatics/btae401

[19] Friedberg, I. (2006). Automated protein function prediction—The genomic challenge. *Briefings in Bioinformatics*, *7*(3), 225–242. https://doi.org/10.1093/bib/bbl004

[20] Chen, J., Gu, Z., Lai, L., & Pei, J. (2023). In silico protein function prediction: The rise of machine learning-based approaches. *Medical Review*, *3*(6), 487–510. https://doi.org/10.1515/mr-2023-0038

[21] Stephenson, N., Shane, E., Chase, J., Rowland, J., Ries, D., Justice, N., . . . , & Cao, R. (2019). Survey of machine learning techniques in drug discovery. *Current Drug Metabolism*, *20*(3), 185–193. https://doi.org/10.2174/1389200219666180820112457

[22] The UniProt Consortium. (2018). UniProt: The universal protein knowledgebase. *Nucleic Acids Research*, *45*(D1), D158-D169. https://doi.org/10.1093/nar/gkw1099

[23] Burley, S. K., Bhatt, R., Bhikadiya, C., Bi, C., Biester, A., Biswas, P., . . . , & Zardecki, C. (2025). Updated resources for exploring experimentally-determined PDB structures and Computed Structure Models at the RCSB Protein Data Bank. *Nucleic Acids Research*, *53*(D1), D564–D574. https://doi.org/10.1093/nar/gkae1091

[24] Procházka, D., Slaninákova, T., Olha, J., Rošinec, A., Grešová, K., Jánošová, M., . . . , & Antol, M. (2024). AlphaFind: Discover structure similarity across the proteome in AlphaFold DB. *Nucleic Acids Research*, *52*(W1), W182–W186. https://doi.org/10.1093/nar/gkae397

[25] Gu, Z., Luo, X., Chen, J., Deng, M., & Lai, L. (2023). Hierarchical graph transformer with contrastive learning for protein

function prediction. *Bioinformatics*, *39*(7), btad410. https://doi.org/10.1093/bioinformatics/btad410

[26] Baker, K., Hughes, N., & Bhattacharya, S. (2024). An interactive visualization tool for educational outreach in protein contact map overlap analysis. *Frontiers in Bioinformatics*, *4*, 1358550. https://doi.org/10.3389/fbinf.2024.1358550

[27] Paysan-Lafosse, T., Andreeva, A., Blum, M., Chuguransky, S. R., Grego, T., Pinto, B. L., . . . , & Bateman, A. (2025). The Pfam protein families database: Embracing AI/ML. *Nucleic Acids Research*, *53*(D1), D523–D534. https://doi.org/10.1093/nar/gkae997

[28] Kulmanov, M., Guzmán-Vega, F. J., Duek Roggli, P., Lane, L., Arold, S. T., & Hoehndorf, R. (2024). Protein function prediction as approximate semantic entailment. *Nature Machine Intelligence*, *6*(2), 220–228. https://doi.org/10.1038/s42256-024-00795-w

[29] Idhaya, T., Suruliandi, A., & Raja, S. P. (2024). A comprehensive review on machine learning techniques for protein family prediction. *The Protein Journal*, *43*(2), 171–186. https://doi.org/10.1007/s10930-024-10181-5

[30] Wang, B., & Li, W. (2024). Advances in the application of protein language modeling for nucleic acid protein binding site prediction. *Genes*, *15*(8), 1090. https://doi.org/10.3390/genes15081090

[31] Wang, W., Shuai, Y., Zeng, M., Fan, W., & Li, M. (2025). DPFunc: Accurately predicting protein function via deep learning with domain-guided structure information. *Nature communications*, *16*(1), 70. https://doi.org/10.1038/s41467-024-54816-8

[32] Chen, J.-Y., Wang, J.-F., Hu, Y., Li, X.-H., Qian, Y.-R., & Song, C.-L. (2025). Evaluating the advancements in protein language models for encoding strategies in protein function prediction: A comprehensive review. *Frontiers in Bioengineering and Biotechnology*, *13*, 1506508. https://doi.org/10.3389/fbioe.2025.1506508

[33] Meng, L., & Wang, X. (2024). TAWFN: A deep learning framework for protein function prediction. *Bioinformatics*, *40*(10), btae571. https://doi.org/10.1093/bioinformatics/btae571

[34] Kulmanov, M., & Hoehndorf, R. (2020). DeepGOPlus: Improved protein function prediction from sequence. *Bioinformatics*, *36*(2), 422–429. https://doi.org/10.1093/bioinformatics/btz595

[35] Du, Z., He, Y., Li, J., & Uversky, V. N. (2020). Deepadd: Protein function prediction from k-mer embedding and additional features. *Computational Biology and Chemistry*, *89*, 107379. https://doi.org/10.1016/j.compbiolchem.2020.107379

[36] Törönen, P., & Holm, L. (2022). PANNZER—A practical tool for protein function prediction. *Protein Science*, *31*, 118–128. https://doi.org/10.1002/pro.4193

[37] Wu, J., Qing, H., Ouyang, J., Zhou, J., Gao, Z., Mason, C. E., ..., & Shi, T. (2023). HiFun: Homology independent protein function prediction by a novel protein-language self-attention model. *Briefings in Bioinformatics*, *24*(5), bbad311. https://doi.org/10.1093/bib/bbad311

[38] Yuan, Q., Xie, J., Xie, J., Zhao, H., & Yang, Y. (2023). Fast and accurate protein function prediction from sequence through pretrained language model and homology-based label diffusion. *Briefings in Bioinformatics*, *24*(3), bbad117. https://doi.org/10.1093/bib/bbad117

[39] Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., . . . , & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

[40] Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., . . . , & Rives, A. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, *379*(6637), 1123–1130. https://doi.org/10.1126/science.ade2574

[41] Gong, J., Jiang, L., Chen, Y., Zhang, Y., Li, X., Ma, Z., . . . , & Tian, M. (2023). THPLM: A sequence-based deep learning framework for protein stability changes prediction upon point variations using pretrained protein language model. *Bioinformatics*, *39*(11), btad646. https://doi.org/10.1093/bioinformatics/btad646

[42] Boadu, F., Cao, H., & Cheng, J. (2023). Combining protein sequences and structures with transformers and equivariant graph neural networks to predict protein function. *Bioinformatics*, *39*(Supplement_1), i318–i325. https://doi.org/10.1093/bioinformatics/btad208

[43] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[44] Beltagy, I., Peters, M. E., & Cohan, A. (2020). *Longformer: The long-document transformer*. arXiv. https://doi.org/10.48550/arXiv.2004.05150

[45] Zaheer, M., Guruganesh, G., Dubey, K. A., Ainslie, J., Alberti, C., Ontanon, S., . . . , & Ahmed, A. (2020). Big bird: Transformers for longer sequences. *Advances in Neural Information Processing Systems*, *33*, 17283–17297. https://dl.acm.org/doi/abs/10.5555/3495724.3497174

[46] Sun, C., Qiu, X., Xu, Y., & Huang, X., Sun, M., Huang, x., ji, H., Liu, Z., & Liu, Y. (2019). How to fine-tune BERT for text classification? In *China National Conference on Chinese Computational Linguistics*, 194–206. https://doi.org/10.1007/978-3-030-32381-3_16

[47] Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202

[48] Ferruz, N., Schmidt, S., & Höcker, B. (2022). ProtGPT2 is a deep unsupervised language model for protein design. *Nature Communications*, *13*(1), 4348. https://doi.org/10.1038/s41467-022-32007-7

[49] Zhang, C., Wang, Q., Li, Y., Teng, A., Hu, G., Wuyun, Q., & Zheng, W. (2024). The historical evolution and significance of multiple sequence alignment in molecular structure and function prediction. *Biomolecules*, *14*(12), 1531. https://doi.org/10.3390/biom14121531

[50] Chandra, A., Tünnermann, L., Löfstedt, T., & Gratz, R. (2023). Transformer-based deep learning for predicting protein properties in the life sciences. *eLife*, *12*, e82819. https://doi.org/10.7554/eLife.82819

[51] Zhao, C., Liu, T., & Wang, Z. (2024). PANDA-3D: Protein function prediction based on AlphaFold models. *NAR Genomics and Bioinformatics*, *6*(3), lqae094. https://doi.org/10.1093/nargab/lqae094

[52] Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., . . . , & Jumper, J. M. (2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3.

Nature, *630*(8016), 493–500. https://doi.org/10.1038/s41586-024-07487-w

[53] Pak, M. A., Markhieva, K. A., Novikova, M. S., Petrov, D. S., Vorobyev, I. S., Maksimova, E. S., . . . , & Ivankov, D. N. (2023). Using AlphaFold to predict the impact of single mutations on protein stability and function. *PLoS ONE*, *18*(3), e0282689. https://doi.org/10.1371/journal.pone.0282689

[54] Ma, W., Zhang, S., Li, Z., Jiang, M., Wang, S., Lu, W., . . . , & Wei, Z. (2022). Enhancing protein function prediction performance by utilizing AlphaFold-predicted protein structures. *Journal of Chemical Information and Modeling*, *62*(17), 4008–4017. https://doi.org/10.1021/acs.jcim.2c00885

How to Cite: Ledenko, H., Coleman, L., Alvarado, G., Stratton, T., Liu, B., Hou, J., …, & Cao, R. (2025). Proteinext: Protein Function Prediction with Sequence Embeddings and Natural Language Processing. *Medinformatics*. https://doi.org/10.47852/bonviewMEDIN52025721