Comparative Analysis of Embeddings and Neural Networks Models for Protein Functional Annotation Models

Boyu Zhang, Nitta Yusaku, Sahil Bishnoi, Baptiste Carbillet

1 Introduction

Protein function studies have gained attention in computational biology research community. Protein functions can be described by a bioinformatics framework called Gene ontology (GO). GO consists of three components: biological process (BP) to which proteins can contribute in biological reaction, molecular function (MF) that comprises a specific biological process, and cellular component (CC) that tells proteins' locations in a cell (Ashburner et al., 2000). These GO terms provide insightful knowledge that aids in understanding the biological profiles of target proteins. Despite the exponentially accumulated protein sequence data, a tiny fraction of those proteins are functionally annotated (You et al., 2018). Therefore, the elucidation of protein functions remains one of the greatest challenges in biology.

In response to the growing demand for an approach to fast and accurate protein functional annotation, scientists have leveraged and developed deep learning models to predict protein function annotations based on a wide range of information including amino acid sequences, protein structures, and protein-protein interactions (PPI). In light of recent advancements in natural language processing within the biological realm, our motivation for this project is to address the complexity of protein function annotation using multi-label classification methods enhanced by deep learning. We aim to develop neural networks that integrate multiple protein embeddings, incorporate additional biological data (e.g., InterPro annotations), and apply logical constraints to ensure biologically consistent predictions. By determining the optimal combination of embeddings, model architectures, and constraints, we strive to improve both accuracy and interpretability in predicting Gene Ontology (GO) terms.

In this project, we hope to contribute to improved learning of protein representations, which could aid people in designing more accurate functional annotation models. Such advancements would substantially benefit our knowledge base of protein functions, which contributes to a wide range of biological fields, including the identification of disease mechanisms, drug discovery, and evolutionary studies.

2 Related Work

Several studies have shown that computational problems that involve high dimensional data can be solved with deep learning-based techniques, and protein functional annotation problem is one of such examples. Data-driven representation approaches leveraging large language models and deep learning have outperformed traditional, rule-based methods in protein function prediction. Pretrained embeddings like ESM, ProtT5, and others capture meaningful sequence representations (Pan et al., 2023; Guan et al., 2024; Kulmanov et al., 2024). Studies incorporating additional biological features associated with proteins, such as hierarchical GO structures, evolutionary data, and protein–protein interactions, have demonstrated improved accuracy in functional annotation (Unsal et al., 2022; Pan et al., 2023; William L. Harrigan, 2024; Xiang et al., 2024). Advanced models, such as PFresGO (Pan et al., 2023), utilize attention mechanisms and hierarchical constraints. Transfer learning techniques leverage large unlabeled datasets to improve performance on downstream tasks. However, top-tier methods like DeepGO-SE (Kulmanov et al., 2024) achieve F1-scores near 0.74, indicating a gap between simpler approaches and state-of-the-art (SOTA) performance. Bridging this gap necessitates richer data integration, improved thresholding, and more computationally intensive strategies.

3 Proposed Method

3.1 Feature Engineering and Representation Learning

We employed four pre-trained embedding models (ESM2, ProtT5, TAPE, ProtBert) from Hugging Face, each offering unique representation strengths. For the tokenization, we used each model's tokenizer (e.g., T5 tokenizer for ProtT5, BERT tokenizer for ProtBert). The embedding were generated by using a mean-polling of the last layer hidden states to create fixed size vectors representation. The InterPro annotations were multi-hot encoded ; they provide informations about protein families, domains, and functional sites. Integrating these features alongside embeddings may help the model capture aspects of protein functionality not evident from sequence alone.

Below we details how the various embeddings were generated. The embeddings of the ESM2 model were already part of our dataset.

- ProtT5: the embeddings were generated using the Rostlab/prot_t5_xl_uniref50 (available on HuggingFace) model with efficient batch processing : the GPU utilization was optimized to manage effective embedding generation.
- TAPE : we used the ProteinBertModel from the tape library, and we applied a mean pooling to get fixed-size vectors.
- ProtBert : the embeddings were generated using the Rostlab/prot_bert model (available on HuggingFace). With this embeddings, we had to add a preprocessing step to get better tokenization : rare amino acids were replaced with 'X'.

All the embeddings were added as new column to the dataset and were converted to pytorch tensor to be passed to our models.

3.2 Model Architectures and Training

The baseline neural network we use is a feed forward MLP. The network begins with fixed-size input embeddings and passes them through fully connected hidden layers, progressively reducing dimensionality from 1024 to 512 units with ReLU activation functions. To mitigate overfitting, a 30% dropout rate is during training, enhancing the model's generalization capabilities. The architecture offers three distinct model variants: embedding-only models that leverage raw vector representations, concatenated models that merge embedding vectors with InterPro annotations into a unified input, and separate processing models that independently handle embeddings and InterPro annotations before feature fusion. The output is passed through a sigmoid to generate multi-label probability predictions. The network is trained with a binary cross entropy loss, to learn classification acroos all potential labels.

3.3 Logical Loss for Biological Consistency

To ensure predictions are biologically plausible and respect GO axioms, we introduced a logical loss function added to the BCE loss. This logical loss enforces constraints such as:

- A Implies B (NF1): If GO term A is predicted, B should also be predicted. We penalize cases where P(A) > P(B).
- Disjointness (NF2): Certain terms are mutually exclusive and cannot co-occur. We implement a penalization throught the loss function if $\sum P(\text{disjoint terms}) > 1$.
- A and B Imply C (NF3/NF4): If A and B are both predicted, C must also be predicted. We penalizes instances where min(P(A), P(B)) > P(C).

These constraints guide the model to produce biologically consistent outputs. Earlier attempts to incorporate network data or related GO terms directly led to data leakage and inflated metrics. The logical loss avoids such issues by penalizing biologically inconsistent predictions at training time without artificially augmenting predictions.

3.4 Handling Complexities

While per-residue embeddings might provide richer context, they were infeasible (50 hours/epoch on our GPU). We relied on mean-pooled embeddings. Adjusting decision thresholds for each label can improve the Precision-Recall trade-off and potentially increase F1-scores.

4 Dataset

We used the dataset available on the GitHub (`https://github.com/bio-ontology-research-group/deepgo2`) where the model from (Kulmanov et al., 2024) is implemented, with the EMS2 and ESM embeddings. The dataset is originally from UniProt (Universal Protein Resource) (Consortium, 2022), a comprehensive and widely used database of proteins' functional information. The dataset integrates three primary components (Figure 1):

- Protein Sequences: The amino acid sequences are the main input, they are transformed into embeddings with the pretrained models.

- InterPro Annotations: These annotations provide information about protein families, domains and functional sites. We treat them as multi-hot encoded features. Since each protein can belong to multiple InterPro categories, these annotations add complexity and biological richness.

- GO Labels (MF, BP, CC): Proteins have multiple GO terms assigned, creating a challenging multi-label setting. The relationships between proteins and GO terms are many-to-many, complicating the prediction of the exact set of terms for each protein.

For BP, the training dataset contains 52,584 proteins, 2,870 in the validation dataset, and 3,275 in the test set. There are in total 30,065 GO annotations that can be predicted. For the molecular function, the training dataset contains 38,533 proteins, 1,901 in the validation dataset, and 2,845 in the test set. There are in total 29,107 GO annotations that can be predicted. For the cellular component, the training dataset contains 52,072 proteins, 2,964 in the validation dataset, and 4,421 in the test set. There are in total 28,301 GO annotations that can be predicted.

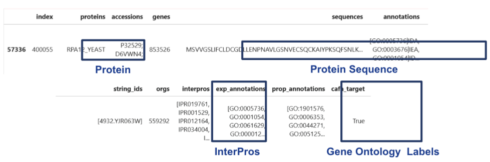

Figure 1: Format of the input dataset

Throught data analysis, we saw that most proteins in our dataset have less than 200 annotated functions (Figure 2). Nevertheless, more than half of the protein in the dataset have more than 40 GO terms (Figure 3).

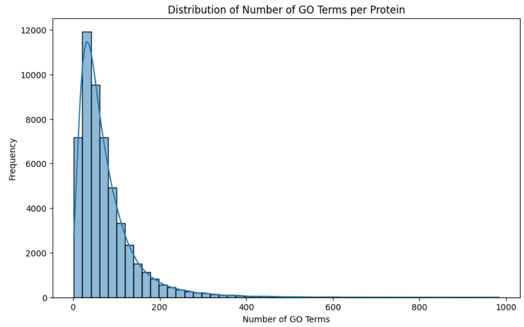

Figure 2: Distribution of the number of GO terms per protein (BP)

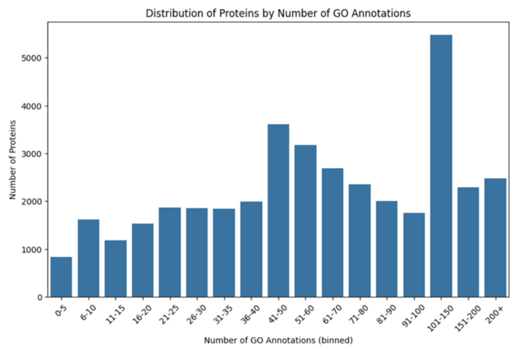

Figure 3: Distribution of Proteins by Number of GO Annotations (BP)

5 Experimental Results and Analysis

We tested embedding-only, concatenated, and separate-processing models. We used the following evaluation metrics for comparison:

- ROC AUC: Measures ranking ability. A high ROC AUC suggests good discrimination but can be misleading with class imbalance.

- Hamming Loss: Fraction of incorrectly predicted labels. Lower values indicate fewer label-wise errors.

- Subset Accuracy: Requires predicting all labels for an instance correctly. Often near zero for complex multi-label tasks, highlighting difficulty.

- Precision and Recall: Precision checks correctness of predicted positives, Recall measures how many of the true positives are identified.

- F1-Score: Harmonic mean of Precision and Recall, balancing both metrics.

After training multiple model architectures across different embedding types and integrating

GO axioms, our models obtained outstanding prediction performance across different embeddings (Table 1).

5.1 Key Insights

5.1.1 Hamming Loss Over Epochs

Looking at Figure 4, The training and validation Hamming Loss plot show steady improvements with training, indicating the model's ability to reduce incorrect predictions. In terms of training loss, it reduces steadily, indicating learning stability. In terms of validation loss, it flattens early, reflecting conservativeness in avoiding false positives.

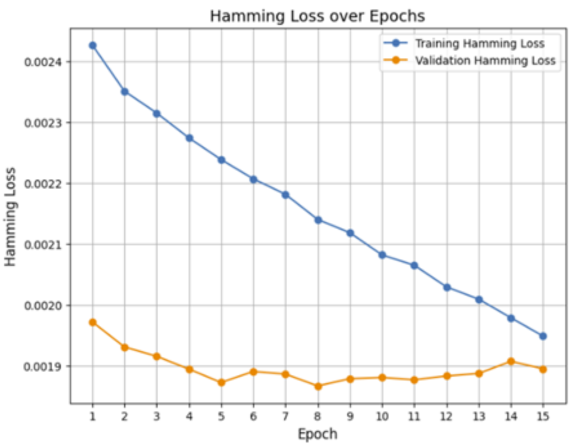

Figure 4: Hamming Loss

5.1.2 ROC Curves for GO Terms

Looking at Figure 5, for the most common labels, AUC ranges from 0.70 to 0.80, demonstrating good discrimination for frequent labels. For the rarest labels, AUC varies widely (0.22 to 0.97), highlighting challenges in predicting rare terms.

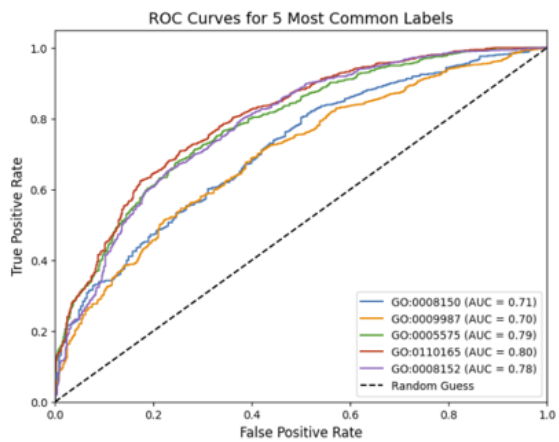

Figure 5: ROC curves

5.1.3 Distribution of GO Term Predictions

Looking at Figure 6, the histogram compares predicted and true counts of GO terms per protein. True GO terms span a wide range, with some proteins annotated with up to 700 terms. However, model predicted fewer GO terms as predicting all GO terms remain challenging.

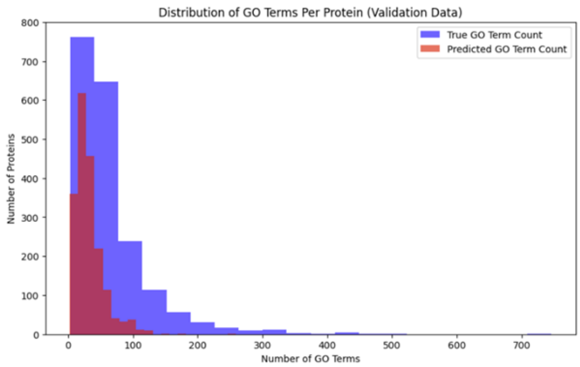

Figure 6: Distribution of GO terms

5.1.4 Evaluation on Test Data

In the end, evaluating our best-performing model (e.g., ESM2 embedding-only) on the test dataset produced similar results to validation. High ROC AUC, low Hamming Loss, and moderate F1-scores were observed, suggesting that the model's strengths and weaknesses generalize beyond the training environment (Figure 7).

Model	GO	Test ROC AUC	Test Precision	Test Recall	Test F-1 score
ESM2+ Embedding Only	MF	0.980889	0.681853	0.355364	0.467223
	BP	0.981382	0.663407	0.31744	0.429974
	CC	0.977754	0.703793	0.276432	0.398629
PROTT5 + Embedding Only	MF	0.981338	0.648102	0.371073	0.471937
	BP	0.981372	0.606122	0.351105	0.444783
	CC	0.975997	0.643371	0.30635	0.413088

Figure 7: Test Results

5.2 Comparison with Other Baseline

Our best F1-scores (~ 0.47) are significantly lower than those of DeepGO-SE (~ 0.73) (Kulmanov et al., 2024). DeepGO-SE's advantage likely comes from more extensive pre-training, richer biological data (e.g., protein–protein interactions), and possibly more sophisticated threshold optimization. While logical constraints and InterPro data improved interpretability and certain metrics, bridging the gap to SOTA performance will require integrating additional biological signals, exploring per-residue embeddings if feasible, refining thresholds, and potentially leveraging more advanced architectures.

Table 1: Performance comparison of models across different ontologies

Model	Ontology	ROC AUC	Hamming Loss	Subset Accuracy	Precision	Recall	F1-score
Baseline	BP	0.9828	0.0019	0.0000	0.6413	0.2984	0.4073
	MF	0.9797	0.0019	0.0000	0.6223	0.3317	0.4327
	CC	0.9791	0.0018	0.0165	0.6633	0.2604	0.3740
ESM2 + Concatenated	BP	0.977431	0.001944	0.001045	0.587627	0.299012	0.396345
	MF	0.974165	0.002054	0.000000	0.563993	0.364484	0.442804
	CC	0.970639	0.001876	0.012146	0.583501	0.279332	0.377803
ESM2 + Embedding Only	BP	**0.981682**	0.001896	0.000000	0.608950	0.312121	**0.412706**
	MF	**0.981189**	0.001925	0.000526	0.625882	0.349409	**0.448459**
	CC	**0.978053**	0.001788	0.019231	0.646021	0.271800	**0.382620**
ESM2 + Separate Processing	BP	0.968781	0.001982	0.001742	0.571605	0.283828	0.379311
	MF	0.965748	0.002060	0.000000	0.571113	0.321793	0.411645
	CC	0.962261	0.001932	0.010796	0.550706	0.284571	0.375241
ProtBERT + Concatenated	BP	0.965645	0.002039	0.000697	0.539161	0.307429	0.391580
	MF	0.964469	0.002053	0.001052	0.577853	0.309413	0.403025
	CC	0.953937	0.001905	0.011808	0.569883	0.266987	0.363620
ProtBERT + Embedding Only	BP	0.979728	0.001917	0.000697	0.637310	0.236405	0.344880
	MF	0.979214	0.001967	0.000526	0.622771	0.308872	0.412941
	CC	0.976463	0.001810	0.017544	0.653400	0.238665	0.349624
ProtBERT + Separate Processing	BP	0.963647	0.002037	0.001394	0.542610	0.289818	0.377830
	MF	0.961134	0.002042	0.000000	0.584215	0.305257	0.400992
	CC	0.956806	0.001885	0.012146	0.586596	0.255308	0.355772
ProtT5 + Concatenated	BP	0.969801	0.001990	0.000348	0.561381	0.307999	0.397766
	MF	0.968845	0.002086	0.000000	0.554528	0.347569	0.427309
	CC	0.962816	0.001955	0.007422	0.536917	0.296238	**0.381814**
ProtT5 + Embedding Only	BP	**0.981373**	0.001929	0.000697	0.593436	0.305783	**0.403601**
	MF	**0.981339**	0.001933	0.000000	0.634537	0.323173	**0.428241**
	CC	**0.975998**	0.001814	0.016532	0.629905	0.266805	**0.374841**
ProtT5 + Separate Processing	BP	0.968860	0.002004	0.001045	0.551236	0.327874	0.411179
	MF	0.968403	0.002028	0.000000	0.581215	0.338531	0427855
	CC	0.959671	0.001892	0.013158	0.575033	0.274706	0.371796
TAPE + Concatenated	BP	0.973319	0.001931	0.000000	0.697833	0.167874	0.270641
	MF	0.969968	0.001968	0.000526	0.673571	0.235151	0.348601
	CC	0.967131	0.001817	0.006073	0.694298	0.194601	0.303997
TAPE + Embedding Only	BP	0.970104	0.001951	0.000000	0.652370	0.184105	0.287169
	MF	0.969953	0.001999	0.000000	0.644737	0.239242	0.348986
	CC	0.964226	0.001843	0.004723	0.676111	0.183929	0.289187
TAPE + Separate Processing	BP	0.962678	0.002021	0.001394	0.565209	0.229036	0.325978
	MF	0.955930	0.002042	0.000000	0.593779	0.278996	0.379622
	CC	0.951920	0.001900	0.008097	0.596590	0.209765	0.310393

6 Conclusion

Our work demonstrates the complexity and potential of deep learning approaches for multi-label protein function annotation. Integrating multiple embeddings and InterPro annotations would provide the model with diverse feature sets. Moreover, logical loss ensured that predictions adhered to GO axioms, enhancing biological consistency and interpretability. Despite achieving high ROC AUC and low Hamming Loss, the model's conservatism led to lower Recall and moderate F1-scores compared to SOTA model (DeepGO-SE). Among all the method options, ESM2-based embedding-only models performed significantly well in function prediction, underscoring the importance of high-quality embeddings. However, predicting the exact combination of GO terms remains difficult, reflected in low subset accuracy and a significant gap from SOTA performance. Future work may involve incorporating protein–protein interaction data, exploring more granular embeddings (if computationally feasible), refining thresholding strategies, and employing more complex architectures to improve Recall and approach SOTA performance levels.

7 Limitations

The main limitation was computational feasibility. While per-residue embeddings might offer more detailed insight, generating them was impractical (50 hours/epoch). We also did not integrate protein–protein interaction data, which could provide essential contextual clues. Addressing these limitations in future work may significantly improve performance and narrow the gap to top-tier models.

8 Source Code

For more details and code, please visit: GitHub Repository

References

Michael Ashburner, Carol A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, James T. Dwight, Janan T. Eppig, Mary A. Harris, David P. Hill, Laurie Issel-Tarver, Amy Kasarskis, Suzanna Lewis, Joel C. Matese, Jill E. Richardson, Mark Ringwald, Gerald M. Rubin, and Gavin Sherlock. 2000. Gene ontology: Tool for the unification of biology. *Nature Genetics*, 25(1):25–29.

The UniProt Consortium. 2022. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research*, 51(D1):D523–D531.

Jiaojiao Guan, Yongxin Ji, Cheng Peng, Wei Zou, Xubo Tang, Jiayu Shang, and Yanni Sun. 2024. Phago: Protein function annotation for bacteriophages by integrating the genomic context.

Maxat Kulmanov, Francisco J. Guzmán-Vega, Patricia Duek Roggli, Ludovic Lane, Stephan T. Arold, and Robert Hoehndorf. 2024. Protein function prediction as approximate semantic entailment. *Nature Machine Intelligence*, 6(2):220–228.

Tianyi Pan, Cong Li, Yan Bi, Zongyue Wang, Robin B. Gasser, Anthony W. Purcell, Tatsuya Akutsu, Geoffrey I. Webb, Seiya Imoto, and Jingyu Song. 2023. PFresGO: An attention mechanism-based deep-learning approach for protein annotation by integrating gene ontology inter-relationships. *Bioinformatics (Oxford, England)*, 39(3):btad094.

Sinan Unsal, Hakan Atas, Murat Albayrak, Kutay Turhan, A Cihan Acar, and Tunca Doğan. 2022. Learning functional properties of proteins with language models. *Nature Machine Intelligence*, 4(3):227–245.

Barbra D. Ferrell et Al William L. Harrigan. 2024. Improvements in viral gene annotation using large language models and soft alignments. *BMC Bioinformatics*, 25(165).

Wenxiang Xiang, Zhiyu Xiong, Haoming Chen, Jiawei Xiong, Wenqing Zhang, Zhiwei Fu, Ming Zheng, Bin Liu, and Qian Shi. 2024. FAPM: Functional annotation of proteins using multi-modal models beyond structural modeling. BioRxiv preprint.

Ren You, Xiaohua Huang, and Shengyu Zhu. 2018. Deeptext2go: Improving large-scale protein function prediction with deep semantic text representation. *Methods*, 145:82–90.