Protein Language Model Ensembling for Gene Ontology Function Prediction under Positive–Unlabeled Supervision

Jacob Cobb, Tessa Everett, Lester Heredia Gopar, Tim Neumann

Abstract

The gap between discovered protein sequences and experimentally verified functions is growing rapidly. Traditional methods often fail to predict functions for new proteins that lack close evolutionary relatives. To address this, we combine the ProtT5 and ESM-2 protein language models to capture deep evolutionary and structural signals from sequence data. We train a neural network using a non-negative Positive-Unlabeled (nnPU) loss, which correctly treats missing annotations as ”unknown” rather than ”negative.” We also apply post-processing to ensure our predictions respect the biological hierarchy of the Gene Ontology. Finally, we evaluate our model using ”scaffold splits” to strictly separate protein families between training and testing. The results confirm our hypothesis: our model generalizes effectively to these completely new sequences, demonstrating that our approach learns meaningful biology rather than just memorizing training data.

Introduction

Proteins are fundamental to biological processes, yet the rapid explosion of genomic data has created a vast gap between known sequences and experimentally verified functions [1]. Traditional homology-based methods often struggle in this setting, especially for proteins that lack close evolutionary neighbors, leaving many sequences uncharacterized where functional inference matters most [2].

Recent advancements in natural language processing (NLP), particularly Transformer architectures, offer a promising alternative [3]. These models can learn contextual embeddings that capture evolutionary, structural, and functional signals directly from sequence, enabling functional inference even when no close homologs exist.

In this work, we investigate whether protein language model (PLM) embeddings, paired with lightweight neural classifiers and simple ontology-aware constraints, can improve function prediction in a CAFA-style evaluation framework [4]. We combine the complementary sequence-level representations of ProtT5 and ESM-2 into a unified feature vector [5, 6]. Our goal is to evaluate the extent to which contemporary sequence models can help close the widening gap between rapid sequence discovery and slow experimental annotation.

Background

Gene Ontology

The Gene Ontology (GO) is a structured, species-agnostic vocabulary designed to unify descriptions of gene product function across the tree of life [7, 8]. GO is organized as a *directed acyclic graph* (DAG) where each node corresponds to a biological concept (GO term) and edges capture relationships such as *is_a* or *part_of*.

GO is divided into three disjoint sub-ontologies, each representing a different dimension of biological knowledge:

- **Biological Process (BP)**: Coordinated biological programs.
- **Molecular Function (MF)**: Elementary biochemical activities, e.g.
- **Cellular Component (CC)**: Locations or complexes where functions occur.

In this work, we focus exclusively on predicting GO terms present in the fixed ontology release used by the CAFA6 challenge.

Protein Language Models

Protein language models (PLMs) apply large-scale self-supervised learning to unlabeled protein sequences, enabling them to learn contextual representations that capture evolutionary, structural, and functional directly from sequences [9].

In this project we use two state-of-the-art PLMs, ESM-2 and ProtT5, to generate fixed-length sequence embeddings for downstream GO prediction. using both models provides complementary information that can improve predictive performance.

Taxonomy

Taxonomic information provides an evolutionary context that complements sequence-based representations. Closely related organisms tend to share functional pathways, making taxonomy a coarse but biologically meaningful prior on GO term likelihoods.

We use organism identifiers from the NCBI Taxonomy database, a curated, hierarchical classification that links each taxon to its parent and reflects the current phylogenetic consensus [10]. This structure supports the idea that evolutionary relatedness can inform function prediction beyond what is captured by sequence similarity alone.

CAFA6 Dataset

The CAFA6 dataset [4] provides GO annotations for supervised training along with an evaluation framework that mimics how biological knowledge emerges over time. Unlike static benchmarks, CAFA evaluates predictions on proteins that acquire new experimental annotations only after model submission, meaning that models must predict both new functions for previously characterized proteins and functions for entirely new proteins appearing in the test superset. This forward-looking design offers a realistic test of a model's ability to generalize beyond the current annotation landscape.

Methods

Dataset Construction

The CAFA6 training dataset consists of 82,404 proteins with experimentally supported GO annotations. Each protein appears exactly once in the training set, with 81,183 unique amino acid sequences and 1,221 duplicated sequences. Sequence lengths range from 3 to 35,213 residues, with a median length of 409. The distribution is dominated by moderately sized proteins, and more than 98% of sequences fall below 2,048 amino acids, making them compatible with standard protein language model (PLM) embedding procedures.

GO Term Selection. The full training set includes 26,125 distinct GO terms across the three ontologies. Because many terms are extremely rare, we restrict prediction to the 1,024 most frequent terms to prevent severe sparsity in the multi-label setting. This subset covers the overwhelming majority of annotated term–protein pairs in the training set while preserving representation of diverse biological functions.

Positive-Unlabeled (PU) Learning

GO annotation is an asymmetric problem: experimentally verified annotations are reliable positives (P), while unannotated entries are unlabeled (U) mixtures of true negatives and hidden positives. Treating U as negative introduces label noise. To address this, we minimize the non-negative PU (nnPU) risk [11].

Let x be a protein embedding and $y \in \{+1, -1\}$ the true label. We observe $s = 1$ if labeled, and $s = 0$ otherwise. The unbiased risk estimator decomposes the total risk $R(f)$ into positive and negative components using the class prior $\pi_p = P(y = +1)$:

$$R_{PU}(f) = \pi_p R_p^+(f) + R_u^-(f) - \pi_p R_p^-(f) \tag{1}$$

where R_u^- is the risk on unlabeled data treated as negative. To prevent overfitting on unlabeled data from driving the risk estimate negative, we employ the non-negative correction:

$$\widehat{R}_{nnPU}(f) = \pi_p \widehat{R}_p^+(f) + \max\Big(0, \widehat{R}_u^-(f) - \pi_p \widehat{R}_p^-(f)\Big) \tag{2}$$

This formulation ensures the model treats unlabeled data as ambiguous rather than definitively negative.

Taxonomy Integration. Each protein is associated with a species-level NCBI taxonomic identifier. The training set spans 1,381 unique taxa with a median of two proteins per taxon. To incorporate evolutionary context, we map each taxon ID to a learned 32-dimensional embedding vector that is concatenated with PLM-based sequence features.

Sequence Preprocessing. Protein sequences exceeding 8,000 residues were divided into contiguous chunks to fit within model limits for PLM embedding. The chunk embeddings were averaged to produce a single fixed-length representation per protein. This approach preserves global functional signal while accommodating rare extremely long sequences. All other sequences were embedded directly in full length.

Summary of Data Distributions. The final training set contains 537,027 protein–term annotation pairs and covers a wide range of eukaryotic and prokaryotic species. A large fraction of proteins originate from a small number of model organisms (e.g., *Homo sapiens*, *Mus musculus*, and *Arabidopsis thaliana*), while the long tail consists of hundreds of taxa with only one or two representatives. The test superset contains 224,309 proteins drawn from 8,453 unique taxa, many of which do not appear in the training set, requiring models to generalize to novel sequences and species.

Dataset and Preprocessing

We utilized the CAFA 6 training set derived from the Swiss-Prot database. The target labels are GO terms divided into three sub-ontologies: Molecular Function (MF), Biological Process (BP), and Cellular Component (CC).

- **Filtering:** Sequences longer than 1024 residues were excluded to fit memory constraints.
- **Label Encoding:** We selected the top N most frequent GO terms for each sub-ontology to create a computationally manageable target vector.

Feature Extraction: Protein Language Models

Instead of training from scratch, we utilized Transfer Learning. We extracted features using two distinct architectures:

1. **ESM-2 (Evolutionary Scale Modeling):** A transformer-based model trained on evolutionary patterns across millions of sequences.
2. **ProtT5:** A T5-based encoder-decoder model optimized for sequence-to-function translation.

We froze the weights of these pre-trained models and extracted the final hidden state representations (embeddings) for each sequence. These embeddings represent the semantic features of the amino acid chains.

Classification Model

We implemented a Multi-Layer Perceptron (MLP) as the classification head.

- **Input:** Concatenated embeddings from ESM-2 and ProtT5 (Dimension size: $D_{esm} + D_{t5}$).
- **Architecture:** Two hidden layers with ReLU activation, followed by Batch Normalization and Dropout ($p = 0.3$) to prevent overfitting.
- **Output:** A sigmoid activation layer producing probabilities for each GO term.

Hierarchical Post-Processing

Raw neural network outputs often violate GO constraints (e.g., predicting a child term with higher probability than its parent). We implemented a consistency enforcement step. Let $P(t)$ be the predicted probability of term t. We iterated through the GO graph structure provided in the `go-basic.obo` file. For every parent term p and child term c, we enforced:

$$P(p)_{final} = \max(P(p)_{raw}, P(c)_{raw})$$

This ensures that the score of a general term is never lower than the score of its specific descendants.

Results

Kaggle Evaluation Procedure.

Model performance is assessed using the CAFA-style, time-delayed evaluation framework. At submission time, t_0, the training set is composed of all known Gene Ontology (GO) annotations. The Kaggle competition then commences a waiting period. After this period, all proteins that accumulated new experimentally validated GO annotations between t_0 and the evaluation time t_1 were selected as benchmarks. This yields three independent benchmark sets—one each for Molecular Function (MF), Biological Process (BP), and Cellular Component (CC)—where a protein is included in an ontology's benchmark if and only if it gained new experimental annotations in that ontology during $(t_0, t_1]$. A protein may therefore appear in multiple benchmark sets. Any annotation that appears in the training data is given weight 0, so the test values reflect the model's ability to predict unseen annotations.

Information-Accretion Weights. Each GO term f is assigned an information-accretion weight $\mathrm{ia}(f)$, equivalent to the information content $\mathrm{ic}(f)$. These weights, provided by the organizers, reflect the rarity and specificity of GO terms: root-level terms receive weight 0, while deep, rarely observed terms receive larger weights. [12]

Weighted Precision and Recall. For a score threshold $\tau \in [0.01, 1.00]$, let $P_i(\tau)$ be the set of terms predicted for protein i with score $\geq \tau$, and let T_i be its experimentally verified GO terms (expanded by the ontology closure). Weighted precision and weighted recall are defined as:

$$\mathrm{pr}(\tau) = \frac{1}{m(\tau)} \sum_{i=1}^{m(\tau)} \frac{\sum_{f \in P_i(\tau) \cap T_i} \mathrm{ia}(f)}{\sum_{f \in P_i(\tau)} \mathrm{ia}(f)}, \tag{3}$$

$$\mathrm{rc}(\tau) = \frac{1}{n_e} \sum_{i=1}^{n_e} \frac{\sum_{f \in P_i(\tau) \cap T_i} \mathrm{ia}(f)}{\sum_{f \in T_i} \mathrm{ia}(f)}. \tag{4}$$

Here, $m(\tau)$ is the number of benchmark proteins for which the method makes at least one prediction at threshold τ, and n_e is the total number of proteins used for evaluation.

Maximum Weighted F-measure. Each ontology is summarized using the maximum information-weighted F-measure, over all thresholds:

$$F_{\max} = \max_{\tau} \left(\frac{2\,\mathrm{pr}(\tau)\,\mathrm{rc}(\tau)}{\mathrm{pr}(\tau) + \mathrm{rc}(\tau)} \right). \tag{5}$$

Final Score. Each submission yields three $F_{\max}$ values (MF, BP, and CC). The competition's overall performance score is the arithmetic mean:

$$\mathrm{S_K} = \frac{1}{3} \left(F_{\max}^{\mathrm{MF}} + F_{\max}^{\mathrm{BP}} + F_{\max}^{\mathrm{CC}} \right). \tag{6}$$

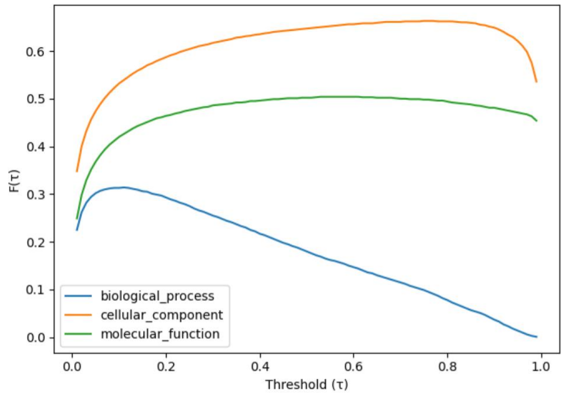
This is the procedure used to test Kaggle submissions. [12]

Our Evaluation Procedure

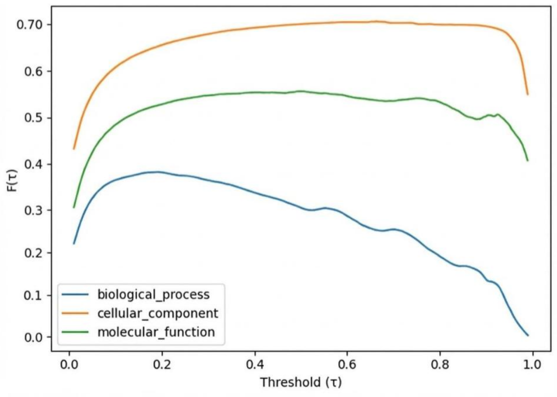
To investigate how data partitioning affects model generalization, we trained two models under different splitting strategies: a *random split* model m_r and a *scaffold split* model m_s. Our primary objective was to assess the extent to which each model relied on sequence memorization versus learned, biologically meaningful representations.

We evaluated both models on two categories of test sets. The first category consisted of proteins held out entirely from training—both their sequences and GO annotations were unseen. Because the held-out sets differed between the two partitioning strategies, we denote them t_r (for the random split) and t_s (for the scaffold split). As seen in table 1, m_r performed better on its own held-out set t_r, since random partitioning often places homologous or near-identical proteins across the train and test sets. Such homology leakage enables m_r to succeed through implicit memorization of family-level similarity rather than through learning generalizable sequence–function relationships.

In contrast, m_s faced a more challenging held-out set t_s: the scaffold split enforces family-level separation, reducing trivial nearest-neighbor matching and providing a more realistic estimate of functional generalization.



(a) Weighted F-measure vs. τ for m_s evaluated on held out set t_s



(b) Weighted F-measure vs. τ for m_r evaluated on held out set t_r

Table 1: F_{max} scores evaluated on t_r and t_s

	Random	**Scaffold**
Overall	0.521	0.494
BP	0.336	0.314
CP	0.685	0.663
MF	0.543	0.504

The second evaluation set, denoted t_k, was the Kaggle challenge set. Unlike the first category, t_k includes not only entirely unseen proteins but also proteins previously observed during training that have since been annotated with newly acquired GO terms. This setting therefore evaluates a model's ability to generalize *beyond its training label space*, predicting novel functions for familiar sequences. Under this regime, we expect that the scaffold-trained model m_s would outperform m_r, as the scaffold split forces the model to learn higher-level sequence–function relationships rather than exploiting memorized family-level similarity.

The results support this expectation: on t_k, m_s exceeds m_r by 0.031 weighted $F_{\max}$, corresponding to an improvement of over 20% relative to m_r. This gap indicates that the inductive biases introduced by scaffold splitting—namely reduced homology leakage and stronger pressure toward

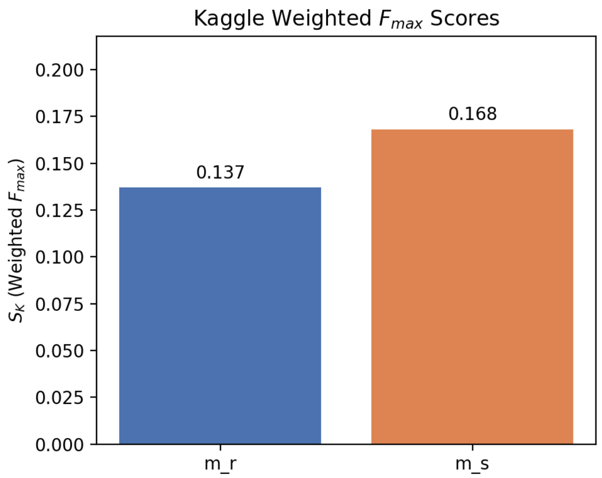


Figure 2: Kaggle weighted F_{max} scores (S_K) for random split model m_r and scaffold split model m_s.

learning transferable representations—lead to improved generalization when models must predict functions that extend beyond the exact annotation patterns seen during training.

Discussion

A key challenge revealed by our experiments is the difficulty of achieving robust generalization across both sequence and taxonomic diversity. While scaffold splitting reduces direct homology leakage, many test proteins come from taxa largely absent in the training set, making taxon-aware modeling essential for realistic functional prediction. Our simple learned taxonomy embedding captures broad lineage structure but does not fully represent deep evolutionary relationships, limiting performance on unseen species. Future work should explore hierarchical or graph-based taxonomy encodings and evaluate taxon-conditioned models to better handle out-of-distribution species.

Another limitation concerns label selection. Restricting training to the top 1024 GO terms improves tractability but may obscure meaningful functions in the long tail. More principled strategies—such as leveraging information-accretion weights or dynamically selecting terms based on ontology structure—could lead to denser, better-calibrated supervision. Relatedly, temporal holdout evaluation highlights that incompleteness of GO remains a central issue; refining label selection to emphasize well-curated branches may reduce noise and improve downstream robustness.

Methodologically, our work evaluated only one architecture (a shallow MLP over concatenated PLM embeddings). Testing multiple architectures would provide a clearer view of what inductive biases matter for GO prediction. Promising directions include models that integrate structure (e.g., AlphaFold features), graph neural networks that operate directly over the GO DAG, and deeper multimodal ensembles that fuse sequence, taxonomy, and predicted structure.

Overall, our findings highlight both the promise and the limitations of using multimodal protein language model embeddings for large-scale GO prediction. While our baseline model demonstrates meaningful generalization under more realistic scaffold-based evaluation, substantial room remains for improvement in handling taxonomic diversity, capturing rare functional labels, and designing architectures that more effectively integrate biological structure.

References

[1] Serbulent Unsal et al. "Learning functional properties of proteins with language models". In: *Nature Machine Intelligence* 4.3 (2022), pp. 227–245. DOI: 10.1038/s42256-022-00457-9. URL: https://doi.org/10.1038/s42256-022-00457-9.

[2] J.-Y. Chen et al. "Evaluating the advancements in protein language models for encoding strategies in protein function prediction: a comprehensive review". In: *Frontiers in Bioengineering and Biotechnology* 13 (2025), p. 1506508. DOI: 10.3389/fbioe.2025.1506508.

[3] Ashish Vaswani et al. *Attention Is All You Need.* 2023. arXiv: 1706.03762 [cs.CL]. URL: https://arxiv.org/abs/1706.03762.

[4] Iddo Friedberg et al. *CAFA 6 Protein Function Prediction.* https://kaggle.com/competitions/cafa-6-protein-function-prediction. Kaggle. 2025.

[5] Ahmed Elnaggar et al. "ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44.10 (2022), pp. 7112–7127. DOI: 10.1109/TPAMI.2021.3095381.

[6] Zeming Lin et al. "Evolutionary-scale prediction of atomic-level protein structure with a language model". In: *Science* 379.6637 (2023), pp. 1123–1130. DOI: 10.1126/science.ade2574. eprint: https://www.science.org/doi/pdf/10.1126/science.ade2574. URL: https://www.science.org/doi/abs/10.1126/science.ade2574.

[7] M. Ashburner and et al. "Gene Ontology: tool for the unification of biology". In: *Nature Genetics* 25.1 (2000), pp. 25–29. DOI: 10.1038/75556.

[8] The Gene Ontology Consortium. "The Gene Ontology knowledgebase in 2023". In: *Genetics* 224.1 (2023), iyad031. DOI: 10.1093/genetics/iyad031.

[9] Lei Wang et al. *A Comprehensive Review of Protein Language Models*. 2025. arXiv: 2502.06881 [q-bio.BM]. URL: https://arxiv.org/abs/2502.06881.

[10] Scott Federhen. "The NCBI Taxonomy database". In: *Nucleic Acids Research* 40.D1 (Dec. 2011), pp. D136–D143. ISSN: 0305-1048. DOI: 10.1093/nar/gkr1178. eprint: https://academic.oup.com/nar/article-pdf/40/D1/D136/9480848/gkr1178.pdf. URL: https://doi.org/10.1093/nar/gkr1178.

[11] Ryuichi Kiryo et al. "Positive-Unlabeled Learning with Non-Negative Risk Estimator". In: *Advances in Neural Information Processing Systems*. Vol. 30. DOI: 10.48550/arXiv.1703.00593. Curran Associates, Inc., 2017, pp. 1675–1685. URL: https://proceedings.neurips.cc/paper/2017/hash/7cce53cf90577442771720a370c3c723-Abstract.html.

[12] Yuxiang Jiang et al. "An expanded evaluation of protein function prediction methods shows an improvement in accuracy". In: *Genome Biology* 17.1 (2016), p. 184. DOI: 10.1186/s13059-016-1037-6.