Protein Function Prediction Using Gene Ontology Annotations

Jakov Tushevski Emanuele Massoglia Gabriel Haas

January 17, 2026

1 Introduction

Protein function prediction is a fundamental problem in bioinformatics, driven by the rapid growth of available protein sequences and the limited throughput of experimental functional annotation. Computational approaches are therefore essential for large-scale functional characterization.

Gene Ontology (GO) provides a structured vocabulary for describing protein function and is organized as a directed acyclic graph (DAG), in which specific terms are linked to more general ones through relations such as *is_a* and *part_of*. According to the true-path rule, annotation of a protein with a GO term implies annotation with all of its ancestor terms.

GO annotations are divided into three independent sub-ontologies: Molecular Function (MF), Biological Process (BP), and Cellular Component (CC). These describe complementary aspects of protein function and define three distinct multi-label prediction tasks.

In this project, we perform protein function prediction under the experimental setup defined by the course assignment. Training and test datasets, along with precomputed protein representations and functional annotations, are provided and prefiltered according to fixed criteria. Separate predictive models are developed for MF, BP, and CC. Model evaluation follows CAFA-style principles, using K-fold cross-validation on the training set only, while test labels remain strictly hidden. Performance is summarized using the maximum F-score (F-max).

2 Data and Exploratory Analysis

2.1 Provided Dataset

All data used in this project are provided as part of the Biological Data course assignment. The training set consists of proteins annotated with GO terms from the MF, BP, and CC ontologies, with annotations already propagated according to the true-path rule.

To ensure tractability, the dataset is prefiltered according to assignment specifications. Only GO terms with at least 50, 250, and 50 instances in MF, BP, and CC respectively are retained. Proteins longer than 2000 amino acids are excluded, and only annotations supported by experimental evidence codes (EXP, TAS, and IC) are considered.

In addition to protein sequences, the dataset includes precomputed ProtT5 embeddings, InterPro domain annotations, and ontology structure files.

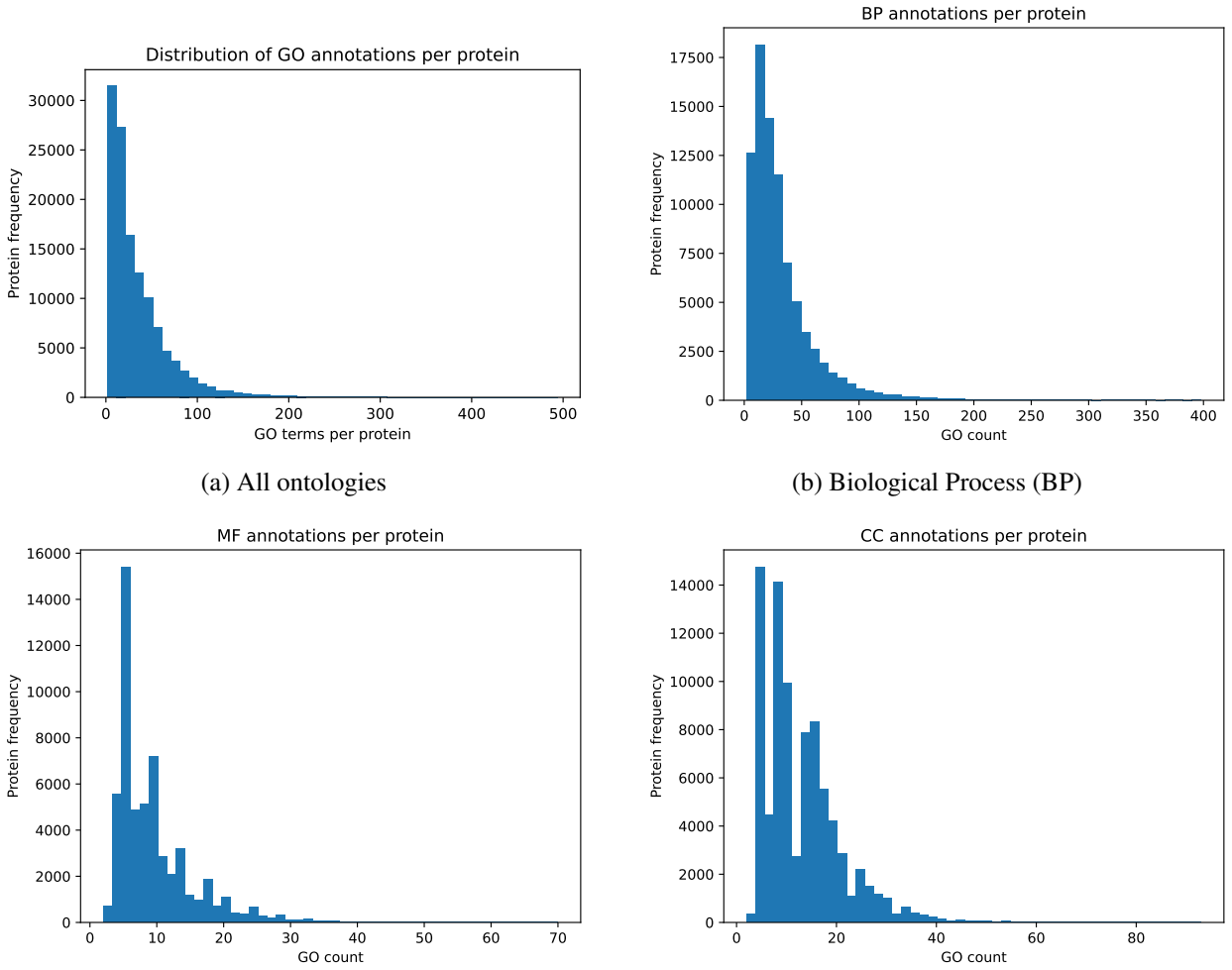
2.2 Exploratory Data Analysis

Exploratory data analysis was performed to characterize the distribution of GO annotations across proteins and sub-ontologies.

2.2.1 Distribution of GO Annotations per Protein

The number of GO annotations per protein exhibits a strongly right-skewed distribution, with most proteins associated with a limited number of terms and a small subset annotated with many terms. This long-tailed behavior is especially pronounced when considering all ontologies jointly.

Biological Process annotations show the highest density and variability, with proteins often annotated to many BP terms. In contrast, MF and CC annotations display more compact distributions, reflecting more localized and constrained functional descriptions.



(a) All ontologies

(b) Biological Process (BP)

(c) Molecular Function (MF)

(d) Cellular Component (CC)

Figure 1: Distribution of GO annotations per protein across all ontologies and per sub-ontology. Biological Process exhibits higher annotation density and variability than Molecular Function and Cellular Component.

2.2.2 Annotation Coverage Across Ontologies

Annotation coverage differs substantially across sub-ontologies. Out of 123,969 proteins in the training set, approximately 84,638 proteins have at least one CC annotation, 83,064 have BP annotations, and only 55,698 have MF annotations. These differences reflect the availability of experimental evidence rather than true absence of function.

The substantial fraction of proteins lacking annotations in each ontology motivates treating MF, BP, and

CC as independent prediction tasks and excluding proteins without positive labels during ontology-specific training.

3 Methods

3.1 Feature Construction

All models operate on numerical protein representations provided with the dataset; no raw sequence preprocessing is performed.

3.1.1 InterPro Domain Features

InterPro domain annotations are used to represent conserved functional and structural regions. Only the 1000 most frequent domains observed in the training set are retained to control dimensionality. Each protein is encoded as a binary multi-hot vector indicating the presence or absence of these domains.

3.1.2 ProtT5 Embeddings

ProtT5 embeddings provide dense sequence-level representations derived from a pretrained protein language model. For residue-level embeddings, mean pooling is applied to obtain fixed-length vectors.

3.1.3 Combined Representation

The final input for each protein is obtained by concatenating its InterPro domain vector with its ProtT5 embedding. ProtT5 features are standardized to zero mean and unit variance, while InterPro features remain binary.

3.2 Label Construction

Training labels are derived from the provided GO annotation file and are already propagated according to the true-path rule. Labels are separated by ontology, and only GO terms meeting the predefined frequency thresholds are retained.

Binary label matrices are constructed independently for MF, BP, and CC. Proteins without positive labels in the target ontology are excluded during training to avoid treating missing annotations as negative evidence.

3.3 Neural Network Architecture

Separate feed-forward neural networks are trained for each ontology. All models are multi-layer perceptrons with ReLU activations in hidden layers and sigmoid outputs, optimized using the Adam optimizer and binary cross-entropy loss.

Architectures differ by ontology:

- **MF:** two hidden layers (512, 256 units), dropout 0.5 and 0.3.
- **BP:** two hidden layers (1024, 1024 units), dropout 0.15.
- **CC:** three hidden layers (2024, 1024, 512 units), dropout 0.4, 0.3, 0.3.

Early stopping based on validation loss is applied to all models.

3.4 Evaluation Protocol

Model selection and performance estimation are conducted using K-fold cross-validation on the training set only, following CAFA-style evaluation principles. Performance is measured using the maximum F-score (F-max), computed by sweeping the decision threshold over predicted probabilities.

4 Results

4.1 Cross-Validation Performance

Table 1 reports the mean and standard deviation of F-max across cross-validation folds.

Table 1: Cross-validation performance measured by F-max (mean $\pm$ standard deviation).

Ontology	F-max (mean)	F-max (std)
Cellular Component (CC)	0.6900	$\pm$ 0.0018
Molecular Function (MF)	0.6757	$\pm$ 0.0043
Biological Process (BP)	0.4480	$\pm$ 0.0047

Performance is highest for CC and MF and substantially lower for BP, consistent with the greater complexity and annotation density observed for BP.

5 Discussion

Differences in predictive performance across ontologies reflect underlying biological organization. Cellular Component and Molecular Function annotations are often closely tied to conserved sequence features and domains, making them more amenable to sequence-based prediction.

In contrast, Biological Process annotations describe higher-level phenomena involving multiple proteins and pathways. Many such processes cannot be inferred from sequence-derived features alone, which likely limits achievable performance.

Despite these challenges, the low variance across cross-validation folds indicates stable training and reliable performance estimation.

6 Conclusion

We developed a protein function prediction pipeline based on supervised learning and evaluated it under CAFA-style assessment principles. Using InterPro domain features and ProtT5 embeddings, separate multi-label classifiers were trained for the MF, BP, and CC ontologies.

The approach achieves strong and stable performance for Molecular Function and Cellular Component prediction, while highlighting the inherent difficulty of Biological Process prediction. Overall, this work demonstrates both the effectiveness and limitations of sequence-based machine learning approaches for automated protein function annotation.