UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Predicting Protein Function Using MLP Models

Biological Data

Kelahan , Cameron
ID: 2071947

Kralevska, Angela
ID: 2072071

Stefanovska, Elena
ID: 2085310

February 2024

Contents

Abstract

Protein function prediction, crucial in genome annotation, has advanced significantly with the rise of high-throughput sequencing. Gene Ontology (GO) plays a pivotal role in characterizing protein functions. Researchers employ diverse approaches, from traditional to cutting-edge deep learning, for efficient GO term prediction. This work reviews computational black-box GO annotation methods, exploring tools' potential in assigning GO terms. This work involves implementation of a software for protein function prediction. The objective is to build a predictive model surpassing CAFA baselines, specifically addressing Molecular Function (MF), Biological Process (BP), and Cellular Component (CC) aspects. Data including Protein embeddings from pre-trained language model, Protein Domains embeddings, Information Accretion (IA), has also been taken into consideration. Additionally, it discusses the challenges in the protein function prediction domain and outlines future directions.

Keywords: *Gene Ontology, Protein function prediction, Protein embeddings, Machine learning, CAFA5, Annotation, Information Accretion, Domain embeddings*

1 Introduction

The speed at which researchers can sequence genomes and identify protein sequences has increased somewhat exponentially in the last decade, doubling in size every few years (see figure 1). Experimental exploration of protein functions cannot keep up with with this pace, causing machine-learning based function predictions to become an important tool to assist in this area of bioinformatics. For example, researchers can start from the computationally predicted findings found in the UniProtKB TrEMBL database to expedite the experimental exploration process.

This paper demonstrates a machine-learning approach to computationally predict protein functions based on relevant information, such as their sequences, InterProT5 domains, and information accretion.

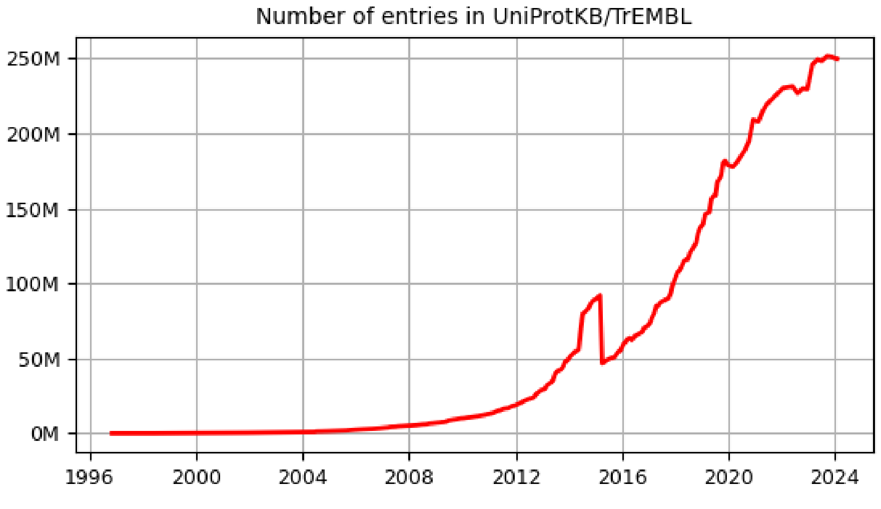

Figure 1: Growth of UniProtKB TrEMBL over the years [7].

Protein function prediction stands as a pivotal challenge in the realm of bioinformatics, where the expense and time associated with wet-lab experiments necessitate the development of computational methods for automated function determination. Determining the function of a protein is extremely useful for understanding diseases, developing new medicine, and growing our knowledge surrounding biological interactions in general. Common approaches for predicting protein functions include comparing sequences and structures to experimentally-annotated proteins, as well as considering the domains of the proteins. This work utilizes both the 'sequences' approach and the 'domain' approach, exploiting the observed correlation between similar sequences and similar functions or domains across proteins [5].

In the backdrop of Gene Ontology's hierarchical system, which categorizes and describes functions through a directed acyclic graph (DAG), our focus extends to building three distinct models for each sub-ontology. As we navigate through the subsequent sections, we will unveil the intricacies of our software implementation, the nuances of the training and validation data, and the benchmarks set by CAFA baselines, ultimately offering a comprehensive exploration of the advancements made in the realm of protein function prediction.

The paper is organized as follows. In section 2 we describe the data that we have at disposal. Section 3 presents data processing, model implementation, experiments and analysis on protein function prediction. Furthermore section 4 presents and discusses the resulting observations and section 5 concludes the paper.

2 Datasets

To both limit the amount of training data and consider only validated findings, the training dataset consists solely of proteins with functional annotations discovered with experimental evidence, listed via their associated gene ontology (GO) terms. On top of that, only protein sequences with a length < 2000 amino acids were kept and the GO terms being considered must have appeared at least 50, 250, or 50 times in the 'molecular_function,' 'biological_process,' and 'cellular_component' sub-ontologies respectively. This allows us to focus on a smaller training dataset for computational purposes and utilize the most common GO terms. Domain, family, and super-family information is also provided where known. We also calculate the information accretion (IA) [1] of the training set to improve the model's predictive abilities. In the end, the training set includes 123,969 unique proteins and 3,584 GO terms.

The test set consists of 1,000 proteins, which the models have not seen, with known functions from experimental research to act as the ground-truth.

2.1 The Gene Ontology (GO)

The Gene Ontology (GO) is a popular resource in bioinformatics that provides a standardized method of annotating functions of genes across different organisms. It is a controlled, structured vocabulary organized into three main categories: Biological Process (BP), Cellular Component (CC), and Molecular Function (MF) (definitions taken from [3]).

- **Biological Process (BP):** "The larger processes, or 'biological programs', accomplished by multiple molecular activities."

- **Cellular Component (CC):** "A location, relative to cellular compartments and structures, occupied by a macromolecular machine."
- **Molecular Function (MF):** "Molecular-level activities performed by gene products. Molecular function terms describe activities that occur at the molecular level, such as 'catalysis' or 'transport.' GO molecular function terms represent activities rather than the entities that perform the actions, and do not specify where, when, or in what context the action takes place."

Each term in GO is assigned a unique identifier (GO ID) and a definition. Terms are organized in a hierarchical, graph-like structure, where more specific terms (child terms) are linked to broader terms (parent terms) using specified relations such as "is a" or "part of".

GO is overseen by the Gene Ontology Consortium (GOC), a collaborative effort involving a team of international researchers who continuously update and expand the corpus based on expert curating and experimental evidence. The GO annotations provide insight into the functions of genes, facilitating data analysis, interpretation, and integration across different biological studies and databases.

As mentioned in the introduction, the GO terms being considered must have appeared at least 50, 250, or 50 times in the 'molecular_function,' 'biological_process,' and 'cellular_component' sub-ontologies respectively, representing the most common terms found in each hierarchy.

2.2 Protein Sequence Embeddings

Investigating the sequences of proteins offers important insight to their functions. In order to facilitate a computationally efficient approach, this research makes use of protein sequence embeddings created with the ProtT5 [4] protein sequence model. The purpose of using the pre-trained ProtT5 model is to learn patterns within given protein sequences that represent information about the evolution, structure, and function of proteins. To that end, the dataset used in this research supplies meaningful features and information from the chosen protein sequences extracted by the ProtT5 model.

In Fig. 2 we may observe the 2D mappings of the protein embeddings using two dimensionality reduction techniques, that is PCA and tSNE. Due to the high computational demand for calculating tSNE embeddings in 2D, the whole dataset was used just with the PCA method. In any case, it is a bit difficult to represent the three different clusters in 2D, nevertheless patterns of embeddings groupings may be observed.

2.3 Inter-Pro Domains (IPR)

In order to further enrich our understanding of protein function, we leverage the comprehensive resources offered by InterPro [8]. We utilized InterPro as a resource that identifies protein domains, which are conserved and independently folding units within a protein that often

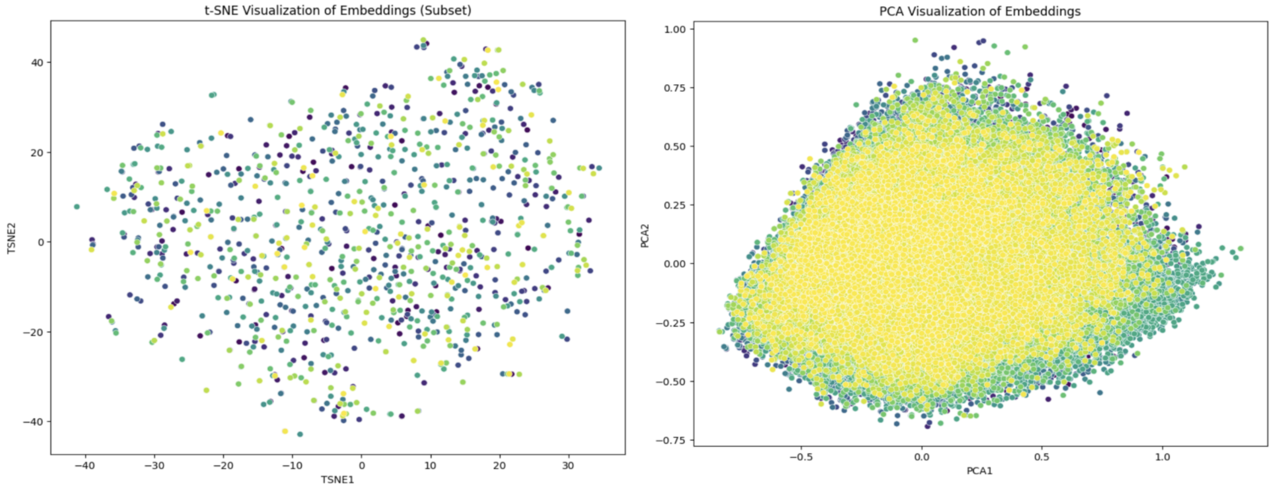

Figure 2: tSNE - over subset of data (left) and PCA 2D visualizations of the T5 protein sequence embeddings (right)

correspond to specific functions[6]. Our methodology relies on two key datasets: protein2ipr and ipr2go.

The protein2ipr dataset acts as a vital intermediary, connecting proteins to their associated InterPro domains. Leveraging the ipr2go dataset, we establish links between InterPro domains and Gene Ontology (GO) terms, facilitating the connection between protein domains and functional annotations. In our dataset construction, we utilize these datasets to generate features, mapping proteins to their domains and connecting these domains to their respective GO terms.

Furthermore, to enhance the representational efficiency of InterPro (IPR) domain data, Principal Component Analysis (PCA) is applied to generate meaningful representations of the domain embeddings, in order to avoid the high-dimensional sparse one-hot encoding for each protein. Since this representation displays embeddings clustering as observed from Fig. 3 as well as dense feature representation for each protein, we decided to further take it into consideration and employ it together with the T5 embeddings as additional feature vectors in our datasets. This approach in turn has improved our models' performance.

2.4 Information Accretion (IA)

Information Accretion (IA) [2], is a measure that quantifies the information added to an ontology annotation by a given node v when its parents $Pa(v)$ are already annotated. The formula for IA is given by:

$$IA(v) = \log_2 \left(\frac{Pr(v|Pa(v))}{Pr(v)} \right)$$

IA reflects how the annotation of a specific term v implies the annotation of its parent terms $Pa(v)$ to ensure a consistent subgraph. To calculate IA, empirical distributions are used based on observed annotations in a dataset. This involves computing $Pr(v)$ as the proportion of proteins annotated with term v and $Pr(Pa(v))$ as the proportion of proteins annotated with

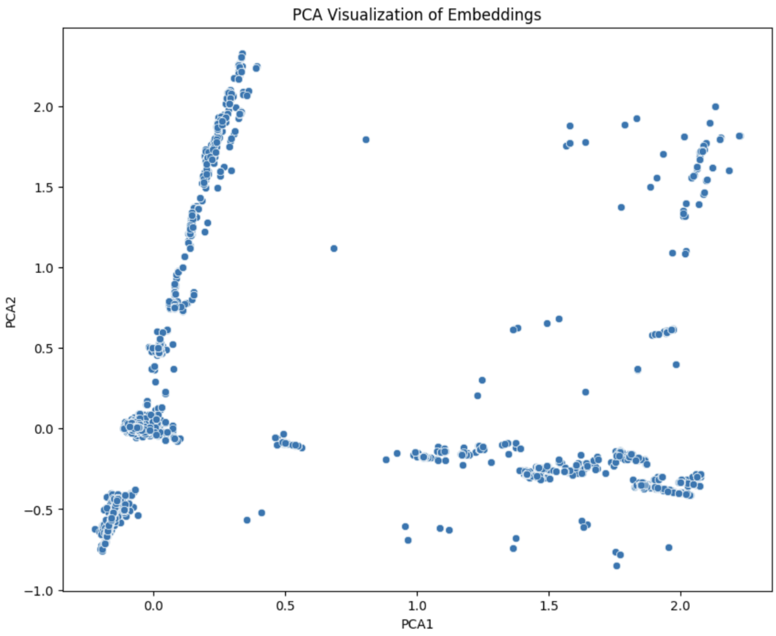

Figure 3: PCA 2D visualizations of the protein domain embeddings.

terms $Pa(v)$. Regularization is applied to handle terms with $Pr(v) = 0$ by introducing an artificial record for all terms in the counts.

For generating the corresponding IA.txt file for our proteins, we used the details and code provided in the GitHub repository[1].

2.4.1 Intuition Checks

1. **Unseen Term with Large IA:** An unseen term X having a large IA indicates that many proteins in the dataset are annotated with the parent term(s) of X, demonstrating significant information accrual.
2. **Term with IA=0:** IA=0 occurs when $Pr(v) = Pr(Pa(v))$, suggesting that every protein annotated with term v is also annotated with all parent terms $Pa(v)$, resulting in no additional information.
3. **Deeper Term with Lower IA:** Deeper terms usually have lower IA, indicating more specialized definitions that are less frequently annotated. Exceptions may occur when term v is frequently annotated in proteins with annotations $Pa(v)$.

3 Experiments and Implementation

3.1 Data Processing

The data processing stage involved the preparation and organization of the dataset. The original dataset was split into training and validation sets using a stratified approach based on protein

[1] https://github.com/claradepaolis/InformationAccretion

IDs. Specifically, 80% of the protein IDs were allocated to the training set, and the remaining 20% were used for validation. The data separated in this way was saved for subsequent model training. At the same time, this ensured a balanced distribution of proteins in both sets.

Furthermore the data processing pipeline involves:

- **Loading and Splitting:** Training dataset is split into train/validation sets.
- **T5 Sequence Embeddings:** Converted to numpy arrays for both training and test sets.
- **IPR Domains:** Embeddings generated, and PCA applied for dimensionality reduction.
- **Information Accretion (IA):** Calculated for training set using external tool, that has been previously referenced. GO terms mapped, and results saved for the purposes of CAFA evaluation analysis.

Moreover, the distribution of the data labeled for each GO Term is given in Fig.4. Biological process is the most abundant, followed by significantly smaller quantity of molecular function and cellular component annotations.

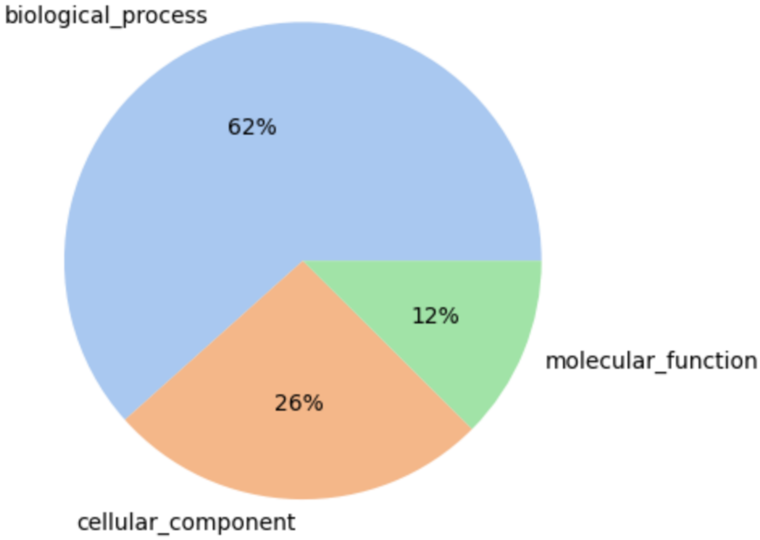

Figure 4: Proportion of data per aspect in the training set

In addition, in Fig. 5 we may observe the distribution of the 50 most frequent terms used for labelling the proteins for the train and validation sets. It is worth noting that we exhibit the same distribution of terms in both sets.

3.2 Model Selection

In this section, we elaborate on the process of model selection for predicting the assigned terms, such that the same pipeline is applied separately to each aspect. The entire pipeline is organized into several key steps, encompassing data loading, feature engineering, model architecture, training, evaluation, and CAFA evaluation.

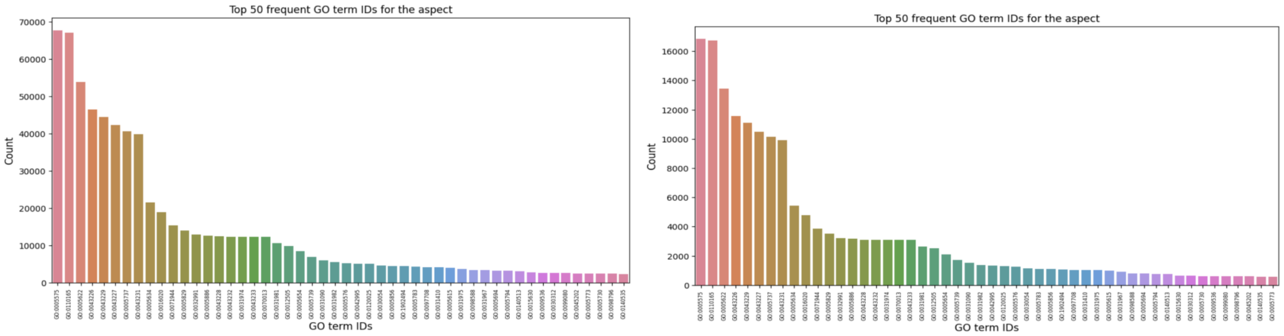

Figure 5: Distribution of the top 50 most frequent terms in the training set within train (left) and validation (right) dataset

3.2.1 Data Loading and Label Encoding

Following data loading, T5 embeddings and domain embeddings are loaded and processed separately. These embeddings capture essential information about protein sequences and domains. The embeddings are combined into a unified dataframe containing all features.

To enable model training, the Gene Ontology (GO) terms corresponding to the aspect are one-hot encoded for both the training and validation sets. The label encoding is based on a predefined order of labels, ensuring consistency across the training and validation datasets.

3.2.2 Model Architecture

Different types of models were explored and compared using the validation set's F1 score, including Linear Regression, Convolutional Neural Network, and Multi-Layer Perceptron (MLP) models. The model architecture that yielded the highest performance was a shallow multi-layer perceptron, created with a regularized linear model. It comprises a dense hidden layer with a ReLU activation function, followed by dropout regularization to prevent overfitting. The final layer utilizes the sigmoid activation function, as this is a multi-label classification task (figure 6 shows the MLP diagram).

The compiled model is trained on the training set using the Adam optimizer with a learning rate of 0.01. Training occurs over 50 epochs with a batch size of 256.

3.2.3 10-fold Cross-Validation

For further model evaluation purposes, 10-fold Cross-Validation of the MLP model is also performed. It uses all the features (T5 protein embeddings, IPR domain embeddings), and investigates variable hidden layer sizes for the different aspects as well. The model is trained on 10 separate folds of the training set using the Adam optimizer with a learning rate of 0.01. Training occurs over 50 epochs per fold with a batch size of 500. The results following this section demonstrate highly similar performance within each fold, which are at the same time near to our models trained and evaluated on the whole training set we have at disposal. Therefore we will further investigate the evaluation plots and metrics on the whole dataset per

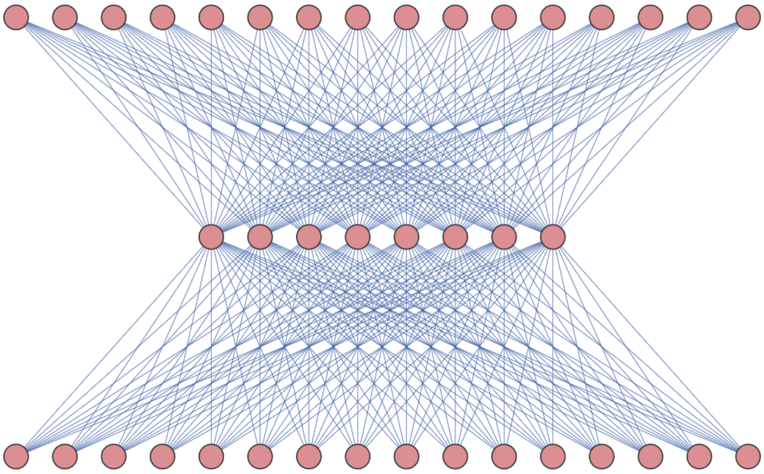

Figure 6: Visualization of the MLP; each node on the input (top) and hidden (middle) layers represent 64 nodes. The final layer consists of the number of possible labels for each specific sub-ontology. There was also a dropout layer after the hidden layer, with a dropout rate of 0.2.

each aspect only.

3.3 Statistical Analysis

During training, the validation set was used to monitor the model's performance and prevent overfitting. The trained model was saved for later use, and the training history was stored for analysis. Various visualizations are generated to facilitate a detailed analysis of the model's performance. Confusion matrices provide insights into specific label predictions, while ROC curves showcase the model's discriminatory power (see section 4). The results of CAFA evaluation are further scrutinized to understand the model's relevance in a broader biological context.

3.3.1 Proteins vs Assigned Labels Analysis

This section outlines the analysis of proteins versus assigned labels for different gene ontology aspects, namely Cellular Component (CC), Molecular Function (MF), and Biological Process (BP), and within their respective train and validation sets. What may be observed from analysing the plots for example for Cellular Component in Fig.7 is that there is similar labels distribution within the train and test set, and this situation is the same for each aspect. In general, the conclusion is there are no drastic differences between the ratio of the labels distribution within aspect. In other words, in this way we confirm the similar distribution of the train and test data, and between aspects. Therefore in spite of this there are no signs of threats to model performance.

Nevertheless, what may also be concluded is that as expected, most proteins are assigned zero terms. In addition, the largest number of labels that a protein gets within the cellular component dataset is 82 in train set and 93 in test set respectively.

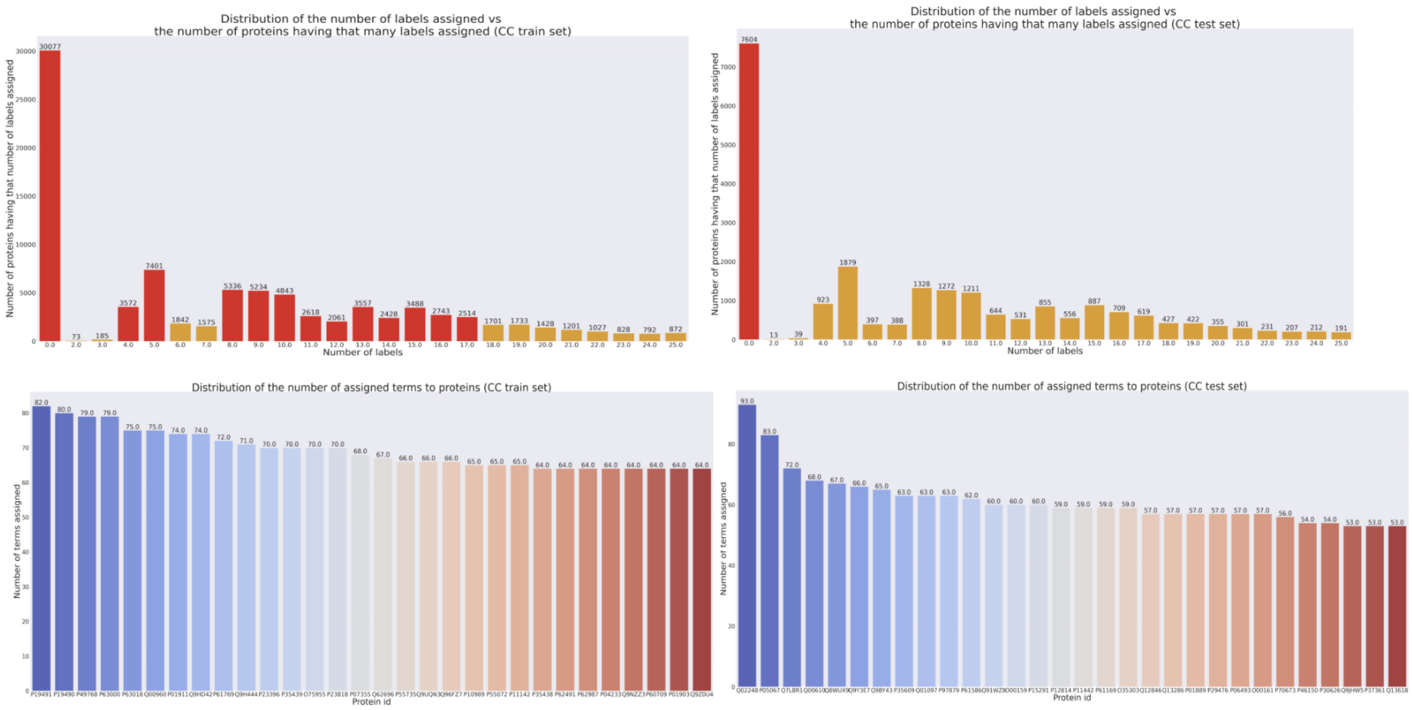

Figure 7: Distributions of numbers of assigned labels to proteins in train and test set for Cellular Component

3.3.2 Model Evaluation

Post-training, the model undergoes evaluation on both the training and validation sets. Metrics such as precision, recall, and F1-score provide insights into the model's performance. Additionally, ROC curves are plotted to visualize the trade-off between true positive and false positive rates.

3.3.3 CAFA Evaluation

To assess the model's performance in a community-wide context, CAFA evaluation is conducted as an illustration over the Cellular Component dataset. Predictions are formatted and submitted for evaluation using the CAFA evaluator[2]. Results, including precision, recall, and F1-score, are recorded for comprehensive analysis.

4 Results and Discussion

The model demonstrated promising results during training, achieving high performance metrics on the training set. However, it is essential to note that the validation set's performance was comparatively lower, indicating potential overfitting. Further analysis, including ROC curves, confusion matrices, and precision-recall curves, was conducted to gain insights into the model's behavior.

[2] https://github.com/BioComputingUP/CAFA-evaluator

Tables 1 and 2 represent the evaluation metrics results on the train and validation sets for each aspect respectfully. The micro, macro and weighted averages over all classes are reported for the metrics: Precision, Recall and F1-Score. Looking at the weighted averaged F1-Score values, for cellular component we obtain 0.74 on the train and 0.59 on the test set, for biological process we obtain 0.58 and 0.38 accordingly, and for molecular function 0.72 with 0.50. From these results we may observe that there is a difference of around 0.2 among the metrics on train and test set, which in turn may indicate potential model overfitting.

Having the discussions in the previous sections, we take into account that there may be multiple GO terms that can be assigned to a protein and also that anyways the most common case is no assigned labels for the proteins. Therefore we further plot and analyse the ROC curves as well as the confusion matrices of the most and least commonly assigned GO Terms.

Table 1: Train Set Evaluation Metrics

Component	Metric	Precision	Recall	F1-Score	Support
CC	Micro Avg	0.66	0.87	0.75	842359
	Macro Avg	0.69	0.72	0.69	842359
	Weighted Avg	0.66	0.87	0.74	842359
BP	Micro Avg	0.56	0.64	0.60	2046266
	Macro Avg	0.61	0.45	0.49	2046266
	Weighted Avg	0.56	0.64	0.58	2046266
MF	Micro Avg	0.63	0.86	0.72	418193
	Macro Avg	0.71	0.82	0.75	418193
	Weighted Avg	0.64	0.86	0.72	418193

Table 2: Test Set Evaluation Metrics

Aspect	Metric	Precision	Recall	F1-Score
CC	Micro Avg	0.53	0.68	0.59
	Macro Avg	0.36	0.33	0.33
	Weighted Avg	0.52	0.68	0.58
BP	Micro Avg	0.40	0.43	0.41
	Macro Avg	0.34	0.23	0.26
	Weighted Avg	0.38	0.43	0.38
MF	Micro Avg	0.46	0.59	0.51
	Macro Avg	0.43	0.44	0.42
	Weighted Avg	0.45	0.59	0.50

Figures 8 and 9 display the Confusion matrices for first 4 most common and 4 of the least common labels in the cellular component validation set. These representations are similar in the case of molecular function and biological process. Therefore the same observations as in the case of cellular component follow. What may be observed here is that in the case of more common labels, that is matrices 1-4, the model misclassified most of the zero labels as 1, which is especially evident in the first two most common labels. That is, higher percentage of False Positives is observed. As the labels become less common, this issue disappears, and on the contrary the model classifies most of the zero labels correctly. This is due to the fact that these terms are mostly labeled with zeroes since are less common and in that way it is much easier for the model to learn rather than learning which term to label as 1 in the case of more commonly occurring annotations. In addition, looking at the last matrix in Fig. 8 we may observe nearly perfect classification, due to the small number of 1 labels for the least common term. Similar scenario is observed on the validation set as well, such that in this case the model fails to correctly assign labels 1 to the least common terms, indicating lower validation set performance. From the previous data analysis steps, we observed that the distributions of GO Terms is pretty much the same in both train and test set. But still there are many GO Terms that are associated with small numbers of proteins. Therefore in order for the model to be able to perform better, more complex model architectures including various embeddings sources or additional data to provide more insights for example should be further explored.

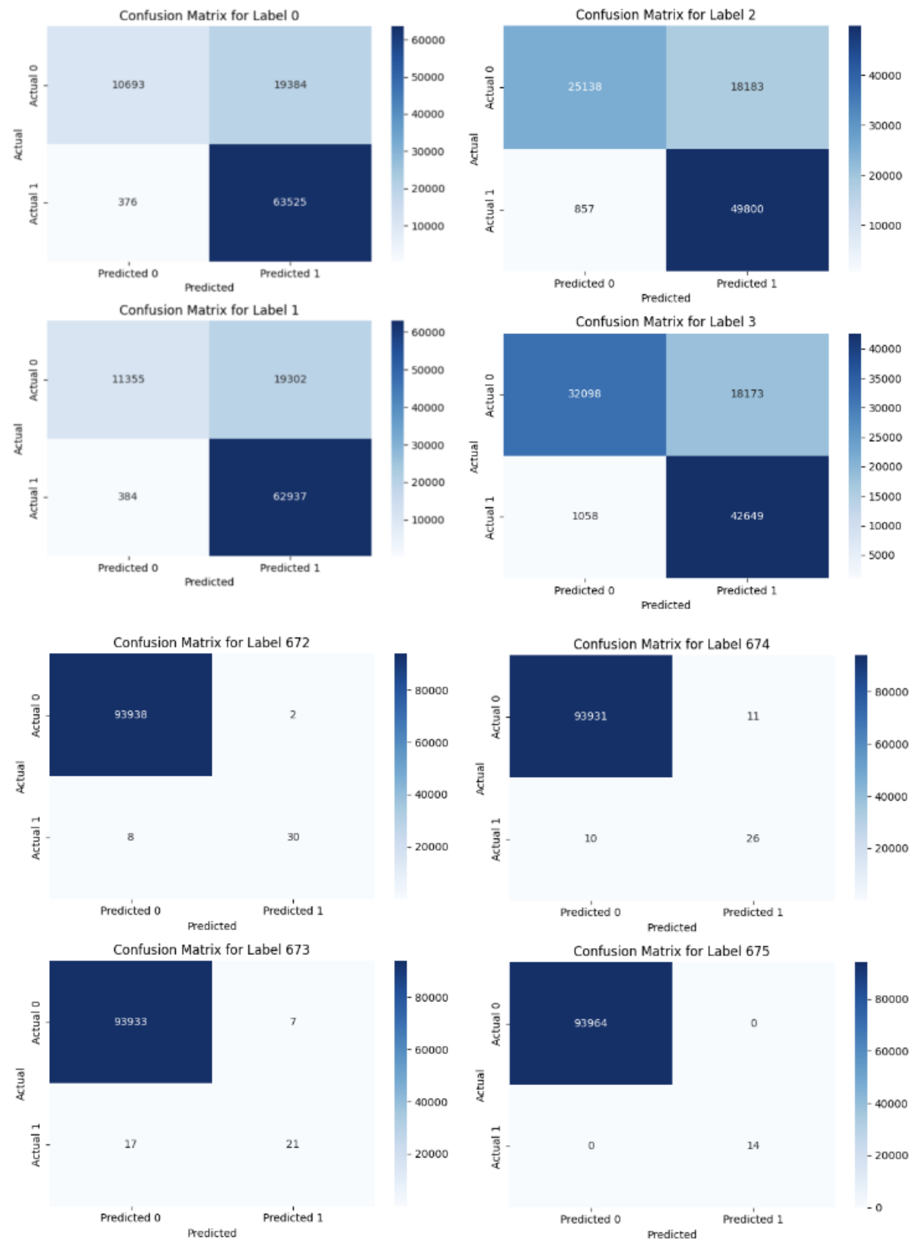

Figure 8: Confusion matrices for first 4 most common and 4 of the least common labels in the CCO train set

Figure 10 displays the ROC curves for the 15 most and least common GO Terms in validation sets for each aspect respectfully. These plots once again confirm the findings that False Positive Rate is considerably smaller in the cases of the least common GO Terms (see right column on Fig. 10). There are bigger variations in the True Positive Rate.

The model's predictions were also evaluated using CAFA Evaluation tools, comparing the results against ground truth annotations. An illustrative plot for the results over the Cellular

Component dataset is given in 11 These evaluations provide additional perspectives on the model's performance. For example, having these evaluations we observed that the max f-score value of 0.712, was obtained with threshold tau of 0.23.

In summary, while the model exhibited strong predictive capabilities during training, careful consideration and fine-tuning may be required to enhance generalization to unseen data.

4.1 Future Work

In future work, exploring more complex model architectures holds promise for enhancing the performance of our system. Architectures such as transformer-based models and deep neural networks with increased depth could be investigated to ascertain their effectiveness in capturing intricate patterns within the data. The exploration of these models may contribute to mitigating overfitting issues by leveraging their inherent capacity for feature extraction and representation learning. Investigating a spectrum of sophisticated architectures will be crucial in unlocking the full potential of deep learning techniques for protein function prediction.

Additionally, we propose investigating the application of a label diffusion approach for further enhancing results. Homology-based label diffusion method leverages the concept that proteins seldom operate in isolation, forming communities with shared functions. By adopting some type of label diffusion algorithm, incorporating ground truth annotations from the training set, and employing BLAST for similarity search, we could model overlapping community effects. This strategy, performed exclusively in the test phase, has the potential to refine initial function predictions and contribute to a more comprehensive understanding of protein functionality.

5 Conclusion

Machine-learning based protein function prediction shows great promise, as demonstrated by the satisfactory accuracy produced by the models in this work. Continuing to explore and experiment with more complex and powerful architectures can improve this performance in the future. Rather than relying on individual features, utilizing the combination of protein sequence embeddings and domain and family information contributed to the performance and could further be improved on by incorporating more information, such as protein sequence embeddings calculated by models other than ProtT5 and information accretion.

Notably, our findings underscored challenges in predicting less common GO terms, where the model struggled with false positives. Future work should focus on exploring more intricate model architectures, possibly incorporating transformer-based models, and experimenting with label diffusion approaches for improved predictions. Additionally, the integration of diverse sources of information and features may contribute to a more comprehensive understanding of protein functionality.

References

[1] Wyatt T. Clark and Predrag Radivojac. Information-theoretic evaluation of predicted ontological annotations. *Bioinformatics*, 29(13):i53–i61, 06 2013.

[2] Wyatt T Clark and Predrag Radivojac. Information-theoretic evaluation of predicted ontological annotations. *Bioinformatics*, 29(13):i53–i61, 2013.

[3] The Gene Ontolgoy Consortium. The gene ontology (go), 2024.

[4] Ahmed Elnaggar et al. Prottrans: Towards cracking the language of life's code through self-supervised learning. *bioRxiv*, 2021.

[5] Radivojac et al. A large-scale evaluation of computational protein function prediction. *Nature Methods*, 2013.

[6] Robert D Finn, Penelope Coggill, Ruth Y Eberhardt, Sean R Eddy, Jaina Mistry, Alex L Mitchell, Simon C Potter, Marco Punta, Matloob Qureshi, Amaia Sangrador-Vegas, et al. The pfam protein families database: towards a more sustainable future. *Nucleic acids research*, 44(D1):D279–D285, 2016.

[7] European Bioinformatics Institute. Uniprotkb trembl statistics 01/2024, 2024.

[8] Alex L Mitchell, Teresa K Attwood, Patricia C Babbitt, Matthias Blum, Peer Bork, Alan Bridge, Shoshana D Brown, Hsin-Yu Chang, Sara El-Gebali, Matthew I Fraser, et al. Interpro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic acids research*, 47(D1):D351–D360, 2019.

[9] D Piovesan. Cafa-evaluator. `https://github.com/BioComputingUP/CAFA-evaluator`, 2024.

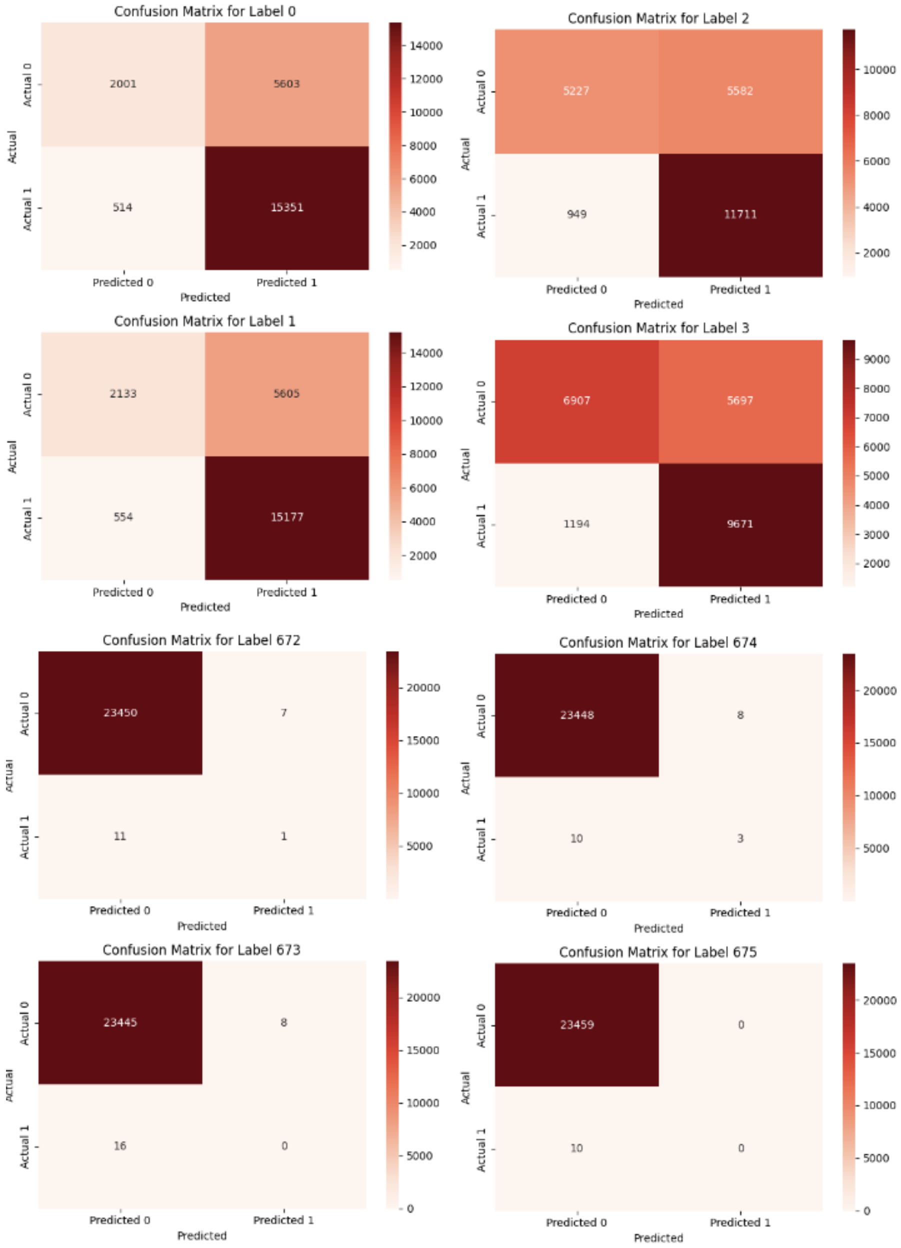

Figure 9: Confusion matrices for first 4 most common and 4 of the least common labels in the CCO validation set

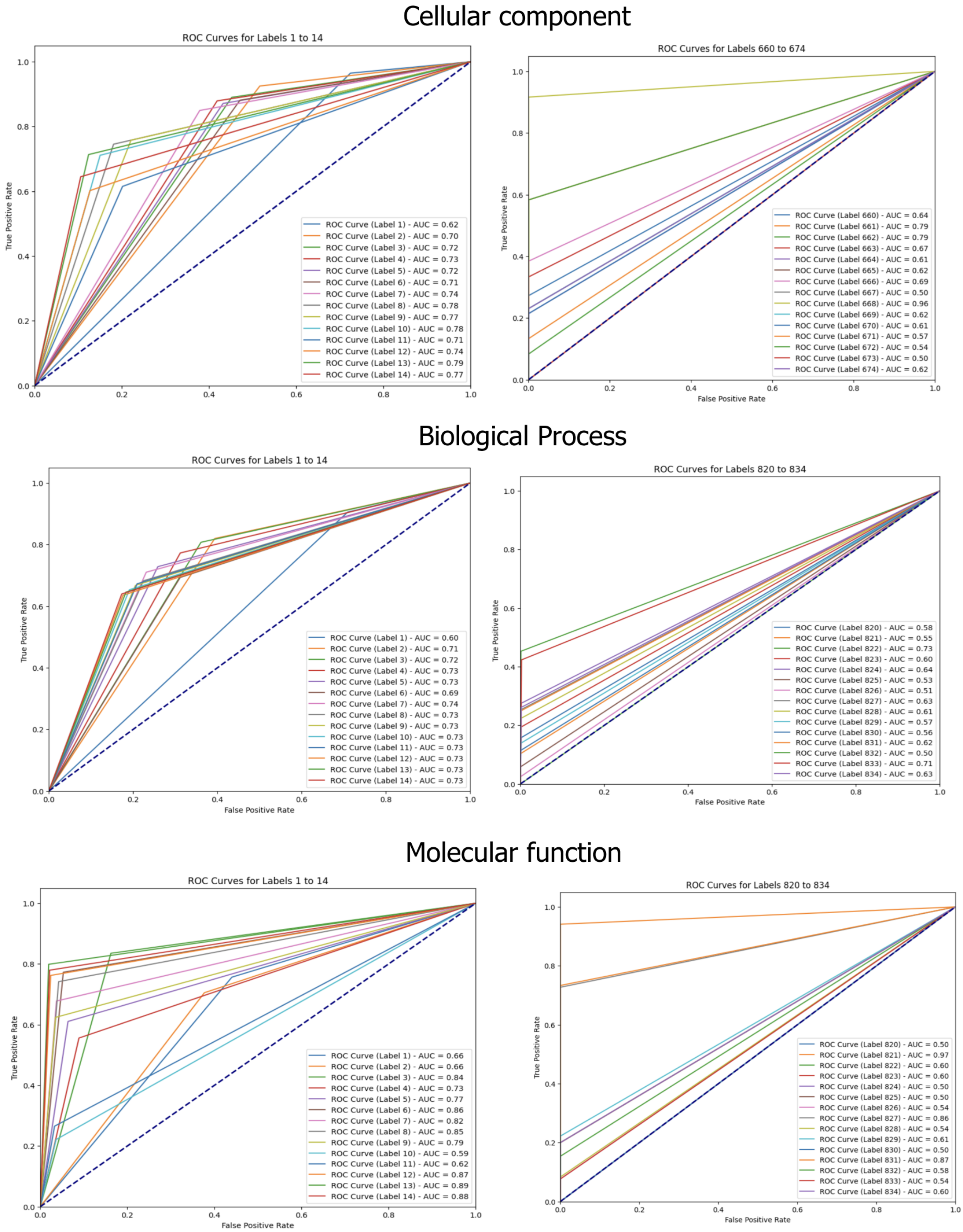

Figure 10: ROC-AUC curves results on the validation set for each aspect

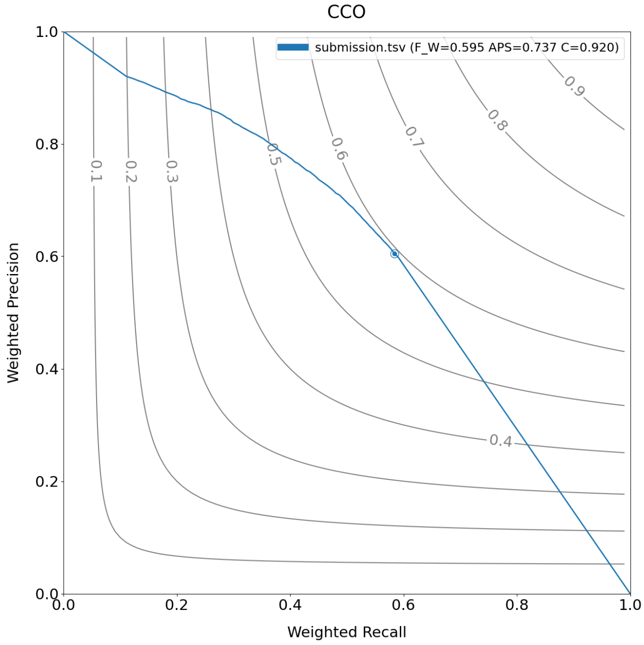

Figure 11: CafaEval plot displaying the Precision/Recall and F1_weighted score for the cellular component model. (Plotted with CafaEval plot.ipynb [9])