CAFA 5 Protein Function Prediction Competition, May 2023
EXECUTIVE SUMMARY

Team AstroBios: Eamon Byrne, Ness Mayker Chen, Dustin Nguyen, Salma Bassem
Mentor: Alexis Johnson
Github/Kaggle: https://github.com/eamonbyrne/CAFA5

Overview

The goal of the competition is to predict the biological function of a set of proteins.

Label organisation: We are provided with gene ontology (GO) data (GO ID terms) along with amino acid sequences. The gene ontology data is in the form of a directed acyclic graph, where the protein functions are in "parent"/"child" relationships with three top-level categories: Molecular Function (MF), Biological Process (BP), Cellular Component (CC).
The training data are very large, with 142,246 unique proteins, 31,466 unique labels and 5.3 million protein+label pairs. The most-labelled protein has 815 labels; the most-proteined label has 92,912 proteins.

Approach

In order to speed up the training and to avoid overfitting the data, we limited the labels to train on to the 1500 most frequently occuring labels in the dataset. All of these labels had >400 instances (proteins) in the training data. We took advantage of pre-trained embeddings that had been generated by masked language models (MLMs) applied to large sets (100s of millions) of protein sequences. These MLMs did not use human labelling (they were "self-supervised").
We evaluated the performance of three different protein embeddings (T5, ProtBERT, ESM2) as input features with three different models (Ridge Regression, Decision Tree Regression, and a simple Neural Network using TensorFlow), all on the same set of the most common 1500 labels.

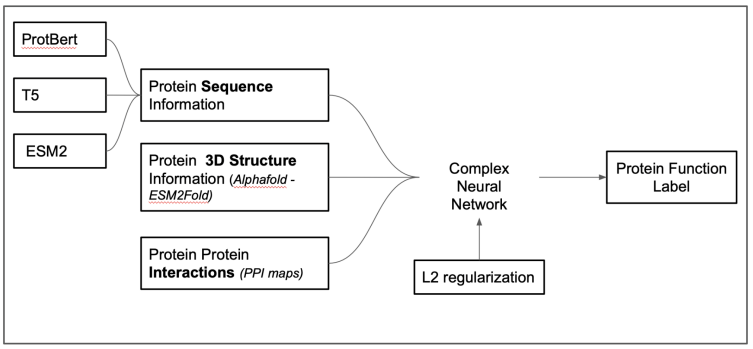
Results & Strategies

		Embeddings		
		T5	**ESM2**	**ProtBERT**
Models	**DTR**	0.29178	0.28556	0.27450
	Ridge	0.47144	0.44286	0.35624
	Neural Network	0.44306	0.43475	0.37331

We found that T5 outperformed the other two embeddings, ESM2 and ProtBERT, for all models. Across each embedding, the Ridge model performed the best. The Ridge model trained on T5 embeddings leading to an F-Measure score of 0.47144. This places our team, "Erdos AstroBio", in 259th place out of 499 teams (as of June 2nd, 2023).

Future Iterations

With approximately 3 months until the end of the competition, there is still time to make improvements. Moving forward, we have a number of strategies that we will test further, including: use different subsets of labels (not just the N most common labels); use multiple MLM embeddings in one pipeline; include embeddings from different starting data (e.g. protein-protein interactions); split up the labels into three categories (MF, BP, CC) and train them separately.



As a preliminary investigation on these future models, we tested using a subset of labels. For a T5 embedding + neural network model, instead of using 1500 top labels, we tried using only 1200 top labels and then using 300 randomly selected remaining labels. This actually increased the score, indicating that there is a performance saturation in using only top labels. As we gain a deeper understanding of the structure of the training data (via hypothesis testing) and the embeddings, we anticipate that we will be able to produce increasingly predictive models.