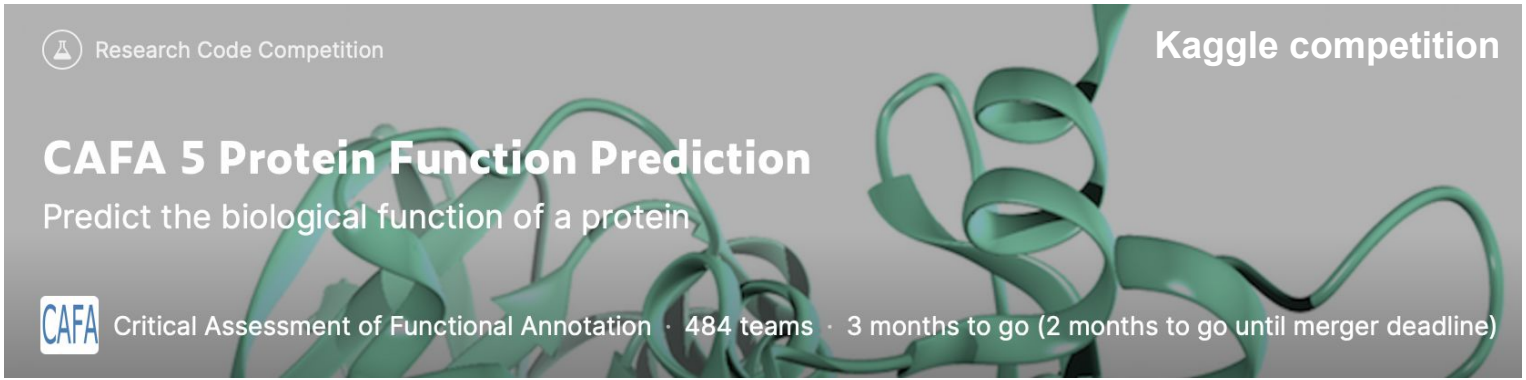
CAFA 5
Protein Function Prediction

Eamon Byrne, Ness Mayker Chen,
Dustin Nguyen, Salma Bassem

Erdos Institute
Data Science Bootcamp, May 2023

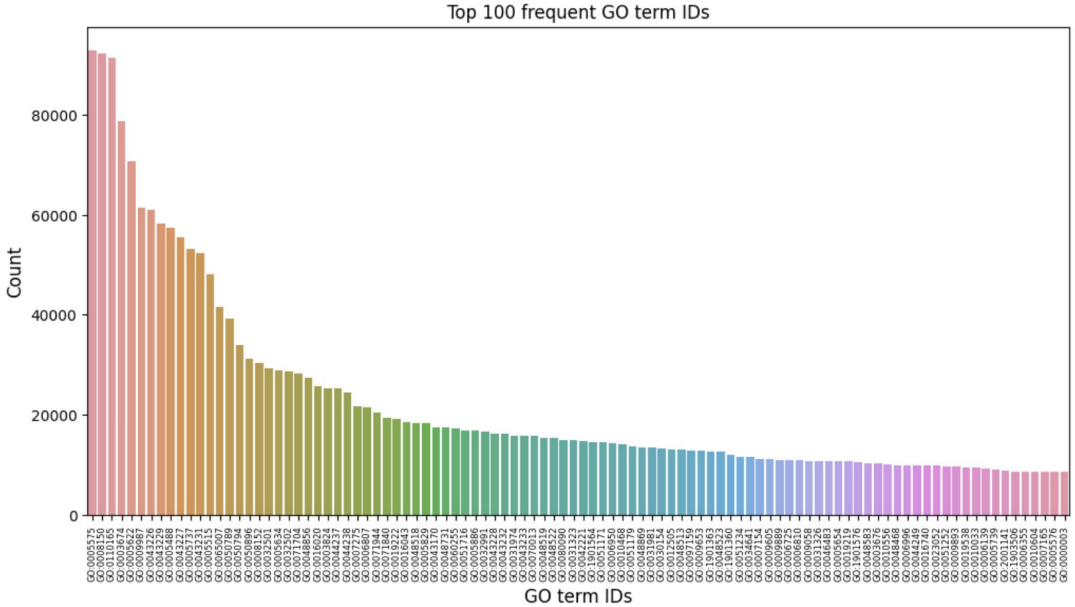
Introduction

Motivation: Proteins are important. Accurate assignment of biological function to a specific protein is critical to understanding life at the molecular level. This is ultimately required for curing diseases and making developments towards improving overall human and animal health.



Problem: Proteins can have multiple functions, and may be dependent on interaction with other proteins. There is a huge amount of complexity in assigning function with respect to sequence or structure. Fundamentally, proteins consist of sequences of amino acids, similar to words in a sentence.

- 142,246 proteins
- 31,466 functions
- Most labeled-protein has 815 functions
- Most proteined-function has 92,912 proteins



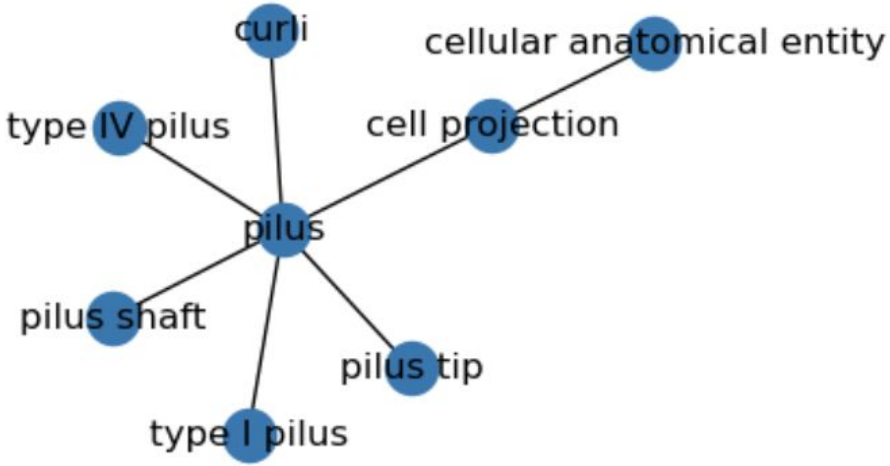
Goal: Develop a model that is able to predict the biological function of a set of proteins based on their amino acid sequence and other data.

Exploratory Data Analysis

Gene Ontology data maps out protein functions as “parent”/”child” relationships

Challenge: accurately assign the function of each protein, including all parent functions

- pilus ← part_of ← pilus shaft
- pilus ← part_of ← pilus tip
- pilus ← is_a ← type IV pilus
- pilus ← is_a ← curli
- pilus ← is_a ← type I pilus



Outline of Pipeline

Each instance is a protein with a particular amino acid sequence. The features are the embeddings that have been pre-trained upon the amino acid sequences. Each label is a biological function that has been experimentally determined.

1. Labels

Generate a (truncated) array of labels (multi-one-hot encoded) from the starting data (a list of protein+function pairings).

2. Features

Load pre-trained embeddings as features.

3. Model

Train a multilabel classification model using the embeddings as inputs.

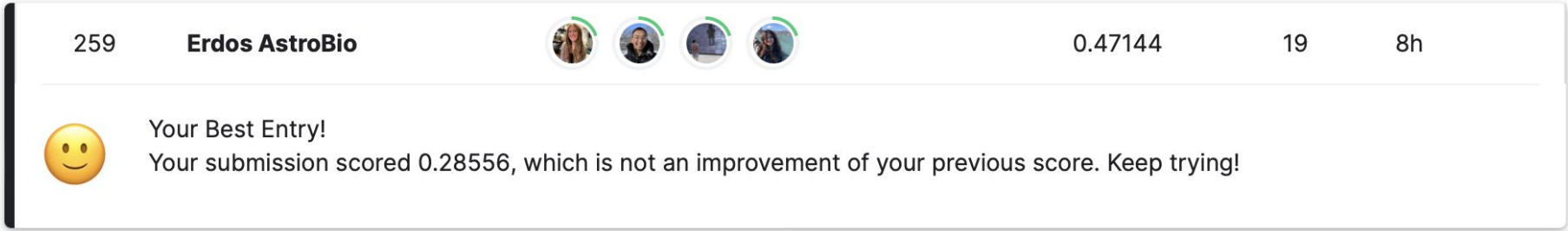
4. Evaluation

Evaluate model on held-out set of protein sequences.

Embeddings x Model Test Grid - Results

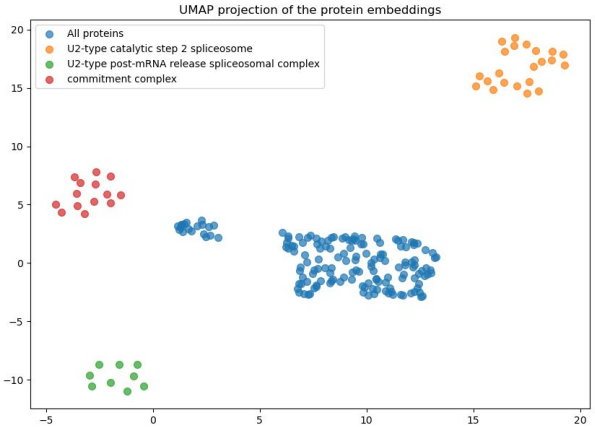
Table 1: F-measure evaluation on Test set (generated by Kaggle competition after submission) for 3 embeddings and 3 models.

		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

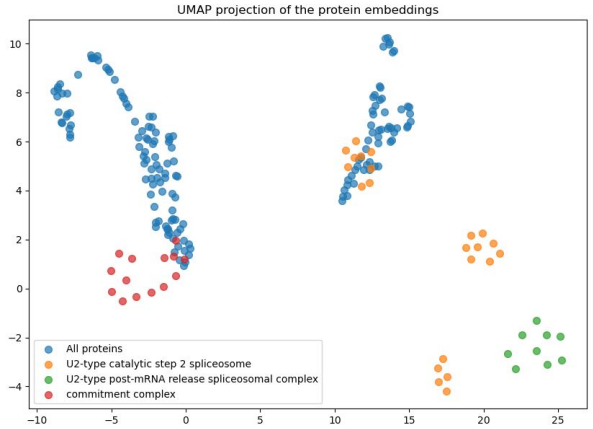


Visualizing the Embeddings for Spliceosome Complex Proteins

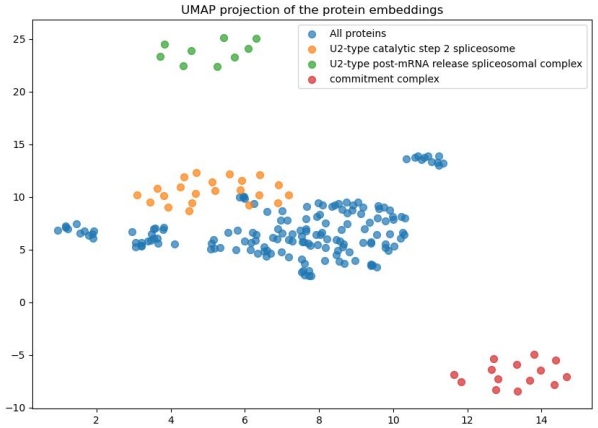
T5 Embeddings



ESM2 Embeddings



ProtBert Embeddings



Next Steps and Future Model Diagram