

Math 156: Machine Learning

Group 9 Project Proposal

Andrew Zhang, Aryan Dalal, Jiwoo Hyun, Steven Villanueva
University of California, Los Angeles (UCLA)

Fall 2025

Problem Description

We want to predict the secondary structure of protein sequences given their primary structure. Starting with vector embeddings of amino acids from Meta's ESM-2 inference model, we will observe how increasing the size of the ESM-2 model improves the performance of common classification methods in protein structure prediction.

Dataset Description

Our data consists of 1,440,460 rows of amino acids. For each amino acid, we know the primary structure of the protein it originates from, its position in the protein's primary structure, and the protein's secondary structure, which we hope to predict. There are nine possible secondary structures in our dataset.

Problem Type

We treat the task as a multi-class classification problem.

Methods, Error Metrics and Algorithm for Evaluation

We will use the 35-million, 150-million, and 650-million parameter versions of the ESM-2 model as starting points for our classification. From there, we will train and compare supervised learning models, including Multiclass Logistic Regression, a two-layer Neural Network, and SVM. After training, we will evaluate the models by computing the confusion matrix to observe class-specific performance.

Since secondary structure datasets are typically imbalanced, we will use the macro-averaged F1 score as the primary evaluation metric. Macro F1 gives equal weight to each class and therefore reflects performance across all secondary structure types more fairly than micro F1.

Comments & Concerns