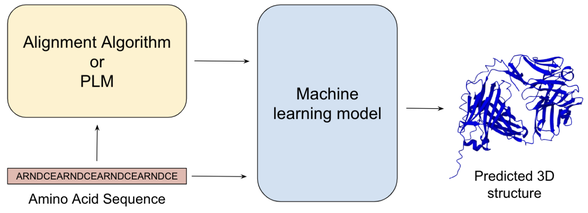
Adapting an ESM protein language model based multimer protein structure prediction model

Hong Zhi Ee
Advised by: Prof. Brian Mak

Introduction

The protein folding problem is the challenge of predicting a protein's three-dimensional structure solely based on its amino acid sequence. Modern solutions to protein folding rely on machine learning models. To generate predictions, models need to produce alignments from the input amino acid sequence.

The main alignment method is to use Multiple Sequence Alignment (MSA) algorithms which generate the highest quality predictions. Alternatively, synthetic alignments can be generated using pre-trained protein language models (PLM), a less resource-intensive method which comes at the cost of reduced prediction accuracy.



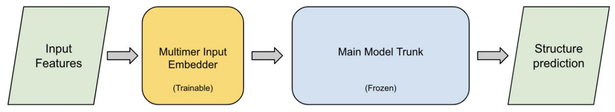
Objectives

While open-source PLM-based models exist for single-chain protein prediction, no open-sourced solution currently exists for multi-chain proteins. The goal of this project is to develop a PLM-based multi-chain protein prediction model by fine-tuning existing single-chain PLM-based models.

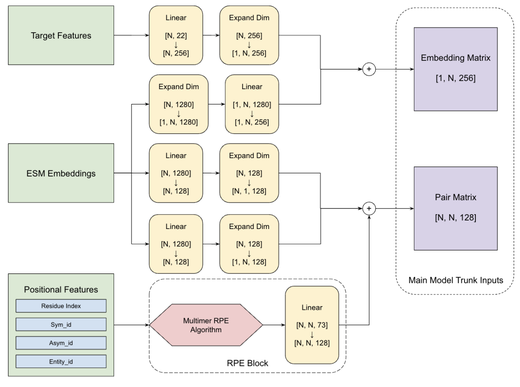
Model architecture

The model architecture is based on the architecture of OpenFold-ESM, an existing open-source PLM-based single-chain prediction model. To support multi-chain predictions, the relative position encoding algorithm in the input embedder will be swapped out with a multi-chain relative position encoding algorithm. The entire input embedder will then be trained using multi-chain protein structures from the protein data bank.

Main Model Architecture



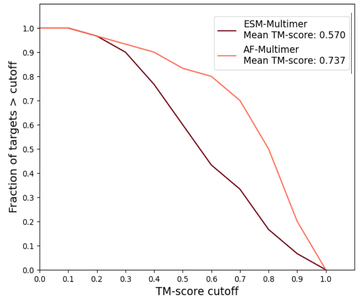
Multimer Input Embedder Architecture



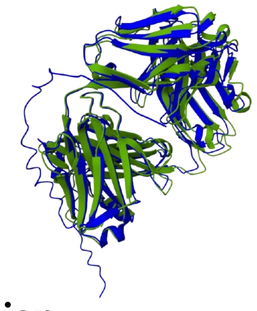
Training and results

The model is trained on 6000 protein structures from the protein data bank (PDB) and evaluated on a test set of 60 proteins evaluated using the Template Modelling (TM) Score ranging from 0 (worst) to 1 (best).

The graph below compares the performance of the trained model (ESM-multimer) against a SOTA structure prediction model based on the Multiple Sequence Alignment (MSA) method (AF-multimer).



Sample ESM-Multimer prediction (blue) superposed on ground truth structure (green)



Conclusion

While there remains a significant performance gap between the PLM multi-chain protein structure prediction model and the current SOTA MSA models. The current model demonstrates the potential of PLM-based multi-chain protein structure prediction models. The findings provide a solid proof of concept for further exploration of PLM-driven multi-chain protein structure prediction techniques.