DL Final Project: Check In #3

Ethan Ye, Sophia Li, Russell Chiu, Kelley Tu

As of the week of April 21st, our group has made significant progress and is close to wrapping up our model. Currently, we have completed the task of collecting and preprocessing data and the task of implementing our model. The key task that remains involves randomly batching our data for training and testing purposes, as well as fine-tuning our model to reach the highest possible accuracy.

At our current stage, the most challenging part of the project so far has been collecting enough data in a format that we can parse and preprocess. In particular, since our model relies on predicting the secondary structure of a sequence of amino acids, we struggled immensely with obtaining datasets that contain both MSA data AND corresponding secondary structures for each sequence. Not only that, but furnishing a simple and efficient way to parse this data proved highly challenging. We eventually addressed this issue by spending hours manually finding datasets that contained comfortable amounts of MSA data and corresponding secondary structure data. Additionally, to address the issue of parsing and preprocessing data, we took inspiration from the original code that came with the paper we are attempting to implement. We currently still need to collect more data for training and testing, as well as parsing the testing data. Because the paper we are trying to implement did not actually have much source code, it has been a lot of back and forth with shape errors and drawing out the logic for our model to actually work with the data we have been collecting ourselves, but because we have basically had to implement almost everything from scratch, I feel like we have learned a lot from trying to preprocess the raw data to actually implementing the model using pytorch.