Proof of Concept for DeepTracer 3.0

A custom pipeline combining the c-alpha results from DeepTracer 2.0 with GNN model refinement from Model-Angelo

Aditya Duggirala

Computer Science and Software Engineering

The University of Washington Bothell

Kirkland, WA, King County

adityaduggirala2001@gmail.com

ABSTRACT

DeepTracer 2.0 uses a CNN model showing a 2D image of a protein structure in Euclidian space. Only image classification is possible, and nodes, edges, and graphs can’t be predicted since those parts of the protein are shown in non-euclidian space with three dimensions. This is why we want a new version called DeepTracer 3.0 which utilizes a GNN model. Proof of concept must be established to show that adding a GNN model is worth implementing in the code base for DeepTracer 2.0. With help from my mentor Jason Chen, we made a custom pipeline using the output from DeepTracer 2.0 and GNN model refinement from Model Angelo. The output from this pipeline is compared to the Model Angelo pipeline. After evaluating metrics, we can see if it’s worth moving forward with DeepTracer 3.0 or if we must make changes. If metrics show the pipeline output is the same or worse than Model Angelo’s pipeline, then an improvement must be proposed. If the results are better, then DeepTracer 3.0 can be implemented. This is done by converting the protein database file from DeepTracer 2.0 prediction into a CIF file. After this, the file path is put in the Initial C-alpha Prediction in Model Angelo’s build. This is done so Model Angelo uses Deep Tracer’s prediction. The GNN model refinement from Model Angelo’s build will use the C-Alpha prediction from DeepTracer 2.0 as the structure of the GNN output. Finally, the evaluation command will be used to compare differences between both pipelines with the input folder of a protein containing the fasta sequence, the MRC file, and the protein data bank map file. This is done by comparing how many protein structures (including secondary structures) and fasta sequences match the input folder. Which pipeline’s output is closer to matching the input folder is the more accurate between the two and determines if improvements need to be made. After gathering results, the custom pipeline matches the Model Angelo output or has higher accuracy than the Model Angelo output. If we add a similar GNN model, we will have the same results for some proteins or better outcomes for others.