Team: Laura McDonnell (PhD), David Antolick (Master's)

Title: Evaluating the Functional Information Content of Protein Structure Prediction Model Embeddings using Deep Mutational Scanning Data

Research Question: Do the representations learned by state-of-the-art protein structure prediction models, specifically per-residue embeddings, contain information that correlates with the functional consequences of amino acid mutations?

Proposed Project: The aim of our project is to evaluate whether embeddings generated by models like AlphaFold2 [1] or ESMFold [2] can provide insight into mutational effects measured by Deep Mutational Scanning (DMS). First, we'll choose one or two proteins (like Beta-lactamase or GFP), that has already been studied with Deep Mutational Scanning (DMS), giving us data on how thousands of mutations affect its function. This information will likely be sourced from MaveDB [3]. We will obtain the predicted structure and corresponding per-residue embeddings for the wild-type sequence using readily available tools e.g., ColabFold for AlphaFold2, HuggingFace Transformers library for ESM-2 embeddings. For each position analyzed in the DMS dataset, we will extract the corresponding embedding vector(s) from the wild-type prediction. Then, we will then investigate the correlation between features derived from these embeddings and the experimentally measured functional scores reported in the DMS data. We will perform statistical analysis and simple machine learning models in python to see if there's a correlation between the embedding information and the measured functional scores from the DMS data. We'll compare how well the embeddings predict function compared to baseline predictors, such as those based on amino acid substitution matrices (e.g., BLOSUM62). Also, we will look at the structures generated and see whether structural deviations (or reductions in confidence - pLDDT) correlate to changes in function.

References

[1] Jumper, J., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873), 583–589.
[2] Lin, Z., et al. (2022). Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv*.
[3] Esposito, D., et al. (2019). MaveDB: an open-source platform for massive assay data. *Genome Biology*, 20(1), 269.