Final Proposal

Proposal Summary

The general premise of my final project will be to use protein databases and information about their respective ligands to characterize protein activity and identify binding site similarity between different proteins.

To go into more detail, I will be working with the PDBbind+ database, which contains a plethora of proteins and their respective ligands. In addition, the database contains physicochemical details of the ligands such as number of atoms, molecular weight, number of hydrogen bond donors, etc. I could also get access to various descriptors of the protein secondary structure, such as the number of helices, number of beta turns, etc. By inputting the physicochemical properties of the ligands into an unsupervised clustering algorithm, I could analyze similarities among different ligands. In addition, I can perform a separate clustering analysis on the protein properties to characterize similarities between proteins. By determining which proteins and ligands are most similar to each other, I intend to use this information to glean novel insights into how binding pocket similarities are associated with the similarity between protein structures and ligand properties.

To build more upon this work, I will also generate a supervised model that will attempt to predict binding affinity based on the residues near the ligand binding pocket as well as the ligand properties. Looking at the dataset information provided, there is a file provided that stores all of the protein residues that are within 10 angstroms of the ligand. Using these files for each protein, I plan to analyze these amino acids using a long short-term memory (LSTM) model. An LSTM model enables the analysis of sequential data like amino acid sequences because it is able to process long-term dependencies. Furthermore, I would then input the ligand properties separately to generate a separate network architecture. My intention is that I would then combine the LSTM model with the ligand network model into one input that would be funneled into another network that would use the information extracted from both models to predict the binding affinity, which would be predicted in the form of IC50.

General Applications

The overall objective of generating these models is to use machine learning techniques to elucidate protein properties, which can be useful for predicting the effects of drug discovery campaigns.

For the unsupervised clustering model, the intention of generating analyses of the protein binding pockets and characterizing how similar they are to each other is to be able to predict off-target drugging effects. For example, if one intends to target a protein that is found in one cluster, then it would be prudent for them to also be wary of how their drug affects the other proteins found in that cluster (provided there are other proteins in that cluster that are from humans). This has very prominent implications in drug discovery campaigns because a primary issue that comes with drugging certain proteins is the off-target effects, so being able to predict those effects and then addressing them before mass production of a drug could make campaigns more efficient.

In a similar manner, the purpose of creating a supervised model that predicts the binding affinity of a protein to ligands with certain properties definitely has its applications in virtual screening. By being able to predict the binding affinity, one can virtually screen many different candidate molecules against a target protein to determine whether those candidates are worth pursuing in a longer term campaign.

By undertaking this project, I aim to bridge the gap between computational methods and practical applications in drug discovery. The integration of unsupervised clustering with supervised learning models not only aligns with the core principles of this course but also holds the potential to influence drug development pipelines. Ultimately, this project represents an attempt to leverage machine learning techniques to address complex biological challenges.