Data Preprocessing/Feature Engineering

Molecules were featurized using RDKit, specifically physicochemical properties, ECFP, and MACCS keys were chosen for the features. Features were normalized/standardized to ensure smooth input to a variety of test models. Both ECFP and MACCS in combination/concatenated were shown in literature to show increased performance (See here: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7819282/). We believe this aided in some of the heavy lifting with the imbalance/lack of data.

Trials and Tribulations

We gathered the wild type and mutant protein files, except for the V560G mutant file, which was not available through the Protein Data Bank. To generate the 3D structure of this mutant, we used AlphaFold2. We then proceeded to dock all structures with SMINA. However when we examined results, there was a poor correlation between binding affinity and docking score. Additionally, some ligands failed in docking due to RDKit structural issues from SMILES. We attempted to perform faster docking, with a method that utilized GPUs, so we set up input scripts for Autodock GPU. However, we did not have a chance to incorporate these results in the allocated time frame. Our future plan is to calculate structural interaction fingerprints between the docked protein and ligand pose.

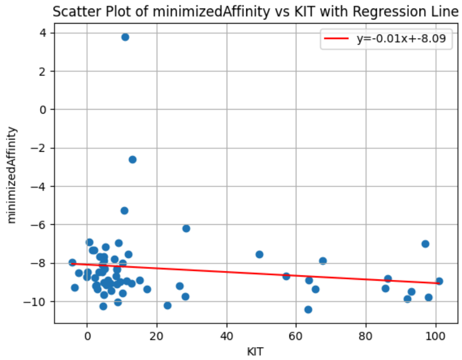

Figure 1. *Binding affinity vs SMINA docking score for KIT protein target.*

Model Selection

Multiple basic models (classifiers and regressors) were tested against the compiled dataset. Notably: LightGBM, Catboost, XGBoost, Random Forest, Nearest Neighbors, Neural Network.

Of these, LightGBM performed the best. A RandomSearch hyperparameter tuning was performed and the best parameters were chosen for the final model.