Classifying and Modeling of Selective and Common Actives and Inactives Against GSK3α and GSK3β

Team 36: **Joe Bosco, Julia Roy**

Technical Advisors: **Arthur J. Campbell (Broad Institute), Sumaiya Iqbal (Broad Institute)**

Inhibitors of glycogen synthase kinase-3 are implicated in the treatment of Alzheimer's Disease, Fragile X syndrome, and bipolar disorder symptoms. However, most known GSK3 inhibitors are non-specific, affect both paralogs, and as a result, incur undesirable side effects. Hence, machine learning classification of GSK3 inhibitors into alpha-specific and beta-specific binding categories is fundamental in narrowing down compounds for more precise drug development, which in turn can guide the way towards improved patient outcomes. As in any machine learning problem, the first step of this project was to gather a large dataset (labeled, in this case, for supervised training) and to filter out duplicate and unlabeled data points. Additional preprocessing included feature selection and generation of Morgan fingerprints from SMILES keys. Our machine learning model consisted of multiple different classifiers. Integration of multiple algorithms to increase predictive power is common practice for many applications in the machine learning world, such as natural language processing and image analysis. Studies have shown that multi-layered models can decrease variance, avoid overfitting, and improve classification accuracy of ligands as drug-like and non-drug like. In this project, we integrate random forest, SVM, and MLP algorithms to predict the ligand activity of chemical inputs to GSK3α and GSK3β separately.