Predicting Drug Potency for the TRPM8 Cold Receptor: A Data Science Approach

Adedolapo Ojoawo, Carmen Al Masri, Jaehyoun Seiler, Jessica Pan

Introduction

Our goal for this project was to use Quantitative structure-activity relationship (QSAR) to predict inhibitors for transient receptor potential cation channel subfamily M member 8 (TRPM8), an ion channel that mediates both cold and pain. QSAR enables the prediction of the molecule's potency based on its physical, chemical and structural properties. In this project, we developed a QSAR workflow, and then chose TRPM8 as a representative protein to test our process. We chose TRPM8 because there are no approved drugs for it. Creating a model that would allow us to screen thousands of molecules and computationally score their inhibitor ability would be immensely useful, as it cuts down on the costly and lengthy process of screening these molecules in the lab.

Dataset and Preprocessing

Our dataset (1168 compounds) was taken from the chEMBL database where we extracted drug-like molecules that interact with TRPM8. The database includes the IC50, which is a score for how potent the molecule is, and is what we're trying to predict. We cleaned up the dataset to remove molecules with empty values, duplicates, etc. yielding a total of 654 compounds. We then calculated descriptors for all the molecules (using RDKit python package for 3D, topological (2D), and constitutional (1D) descriptors, and performed QM calculations to get the quantum features). A total of 1008 features were obtained. The datasets with calculated features were split using 15% for testing (external validation) and the rest for training.

Feature Selection & Model Training

After filtering out features with low variance or high correlation, we performed 4 types of feature selection on the training set and used them as inputs to train three regression and three classification models. The feature selection methods were: 1) PCA, 2) Factor analysis, 3) Genetic algorithm, 4) Random Forest with recursive elimination. The regression and classification models were 1) SVM, 2) Random Forest, and 3) XGBoost.

Using PCA-based feature reduction, we were able to see good separation between the high and low potency molecules but medium potency overlaps with high potency. This indicates that the features present in our selected principal components contribute better at predicting low potency/affinity and high potency/affinity. This is also reflected in our results, where the models performed better at predicting low and high-potency molecules compared to the medium.

Results

Most feature selection and model combinations scored remarkably high on both regression and classification tasks on the test set. For regression, we obtained a pearson correlation coefficient between predicted and experimental IC50s that ranged between 0.7 and 0.8, and for classification the accuracies ranged between 80% and 95%. We found that the best model for regression was random forest with random forest recursive elimination feature selection. For classification, it was SVM with genetic algorithm feature selection.

Overall, the workflow was effective and would generalize well to any protein we wish to target.