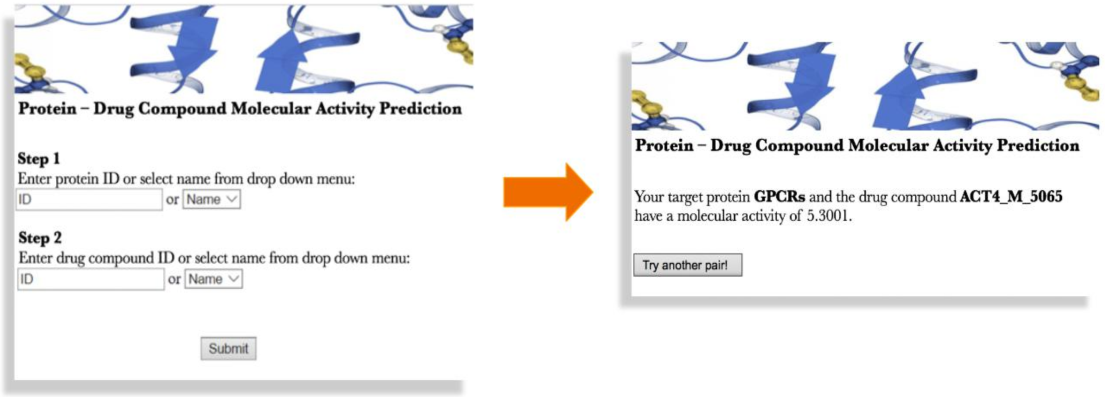
2-paragraph “news-like” story:



Protein – Drug Compound Molecular Activity Prediction

Step 1
Enter protein ID or select name from drop down menu:

Step 2
Enter drug compound ID or select name from drop down menu:

Protein – Drug Compound Molecular Activity Prediction

Your target protein **GPCRs** and the drug compound **ACT4_M_5065** have a molecular activity of 5.3001.

Traditionally, drug discovery is the area of research and development that used the most amount of time and money. The time frame can easily range from 3 to 20 years and costs can range between several billion to tens of billions of dollars for the research teams to find the molecules that highly active toward the target through myriad experiments. Therefore, Merck, the largest pharmaceutical company in the world, proposed a machine learning challenge to help develop safe and effective medicines by predicting molecular activity.

To address this challenge, this project aims to utilize machine learning as a cost-effective tool for predicting biological activities of different molecules, both on- and off-target, given numerical descriptors generated from their chemical structures.

There is a total of 15 datasets included in this challenge, each for a biologically relevant target. Each row of the data corresponds to a molecule and its molecular descriptors that derived from the chemical structure. Given, the time and scope of this project, we choose to work on the data set with largest number of molecules and largest number of numerical descriptors.

Exploratory Data Analysis gave us valuable insights about the data. We found that 68% values of the output fall into the range of 4.3-4.7, which leads to imbalanced data distribution. Moreover, we found that there are no missing values in the dataset. Additionally, we realised that high dimensional data can be reduced to 441 features via near zero variance method with 90% cumulative variance of total variance.

An artificial neural network with 2 hidden layers showed us that we could achieve a mean absolute percentage prediction error of 6-7%. However, there was significant variance in the predictions. We clustered the data and identified certain groups that could be causing the large amounts of variance.

A random forest baseline model with 500 trees and the maximum features equal to the square root of the input dimensions was built. The model was runned with 5-folds cross-validation. After evaluating the performance of the baseline random forest model, we trained the model on 3 datasets created by different pre-processing methods as well as utilized the random search method to tune the parameters. The best model, which can reach the lowest mean absolute percentage error (MAPE) and mean absolute error (MAE) as well as highest R^2score, was trained by the 50 features selected by the correlation. The corresponding MAPE is 5.58%.

Since we were making this platform not only for research institutes and researcher who have scientific background but also for those who have no background in drug development but are interested in learning about protein and drug binding, we designed a user interface that is as straightforward and easy to navigate as possible.

Major challenges and bottlenecks we faced during the course of the project were related to data sparsity, data dimensionality, data sampling and the difficulty to improve performance of ANN past 6-7% MAPE value.

From future perspective, minimal or no preprocessing might help to improve the performance of the ANN. It may require deeper nets and a dedicated GPU. Additionally, ensemble models and more iterations of the neural network architecture might improve the performance of the model. Lastly, testing the model on the other Merck datasets to evaluate generalizability and linking the compound name to compound ID to determine biosimilars could be the next steps.