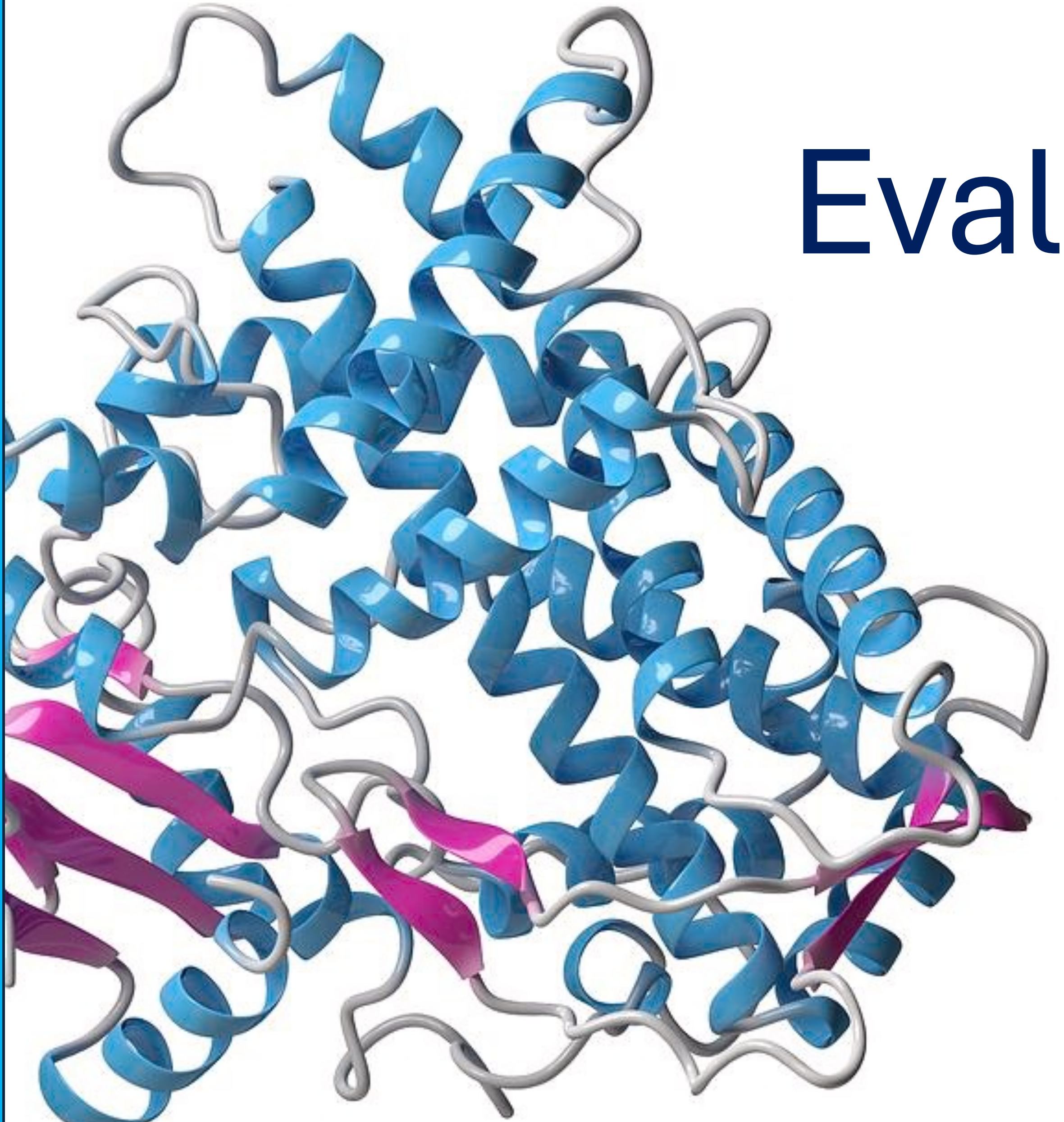
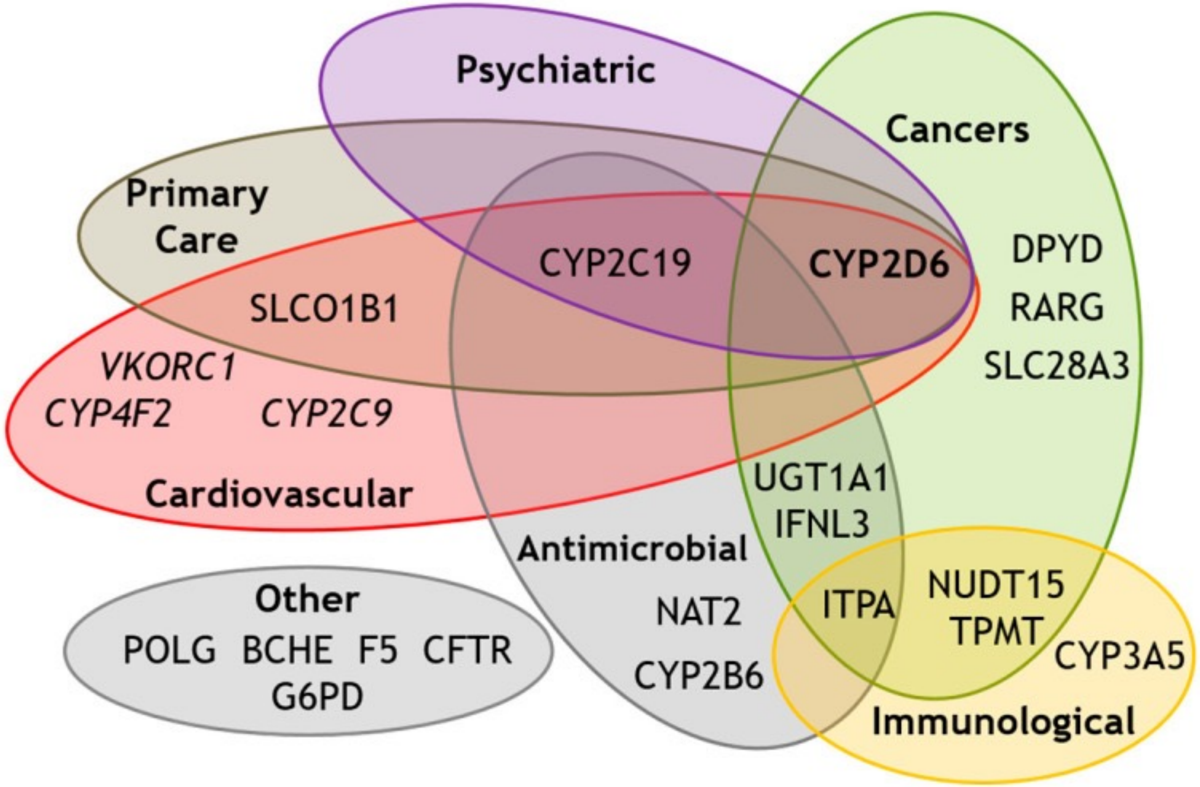


Evaluating Compounds Targeting CYP2D6

A Cheminformatics and Machine Learning Project

Christian Geils

Motivation: CYP2D6 in Drug Metabolism and Disease



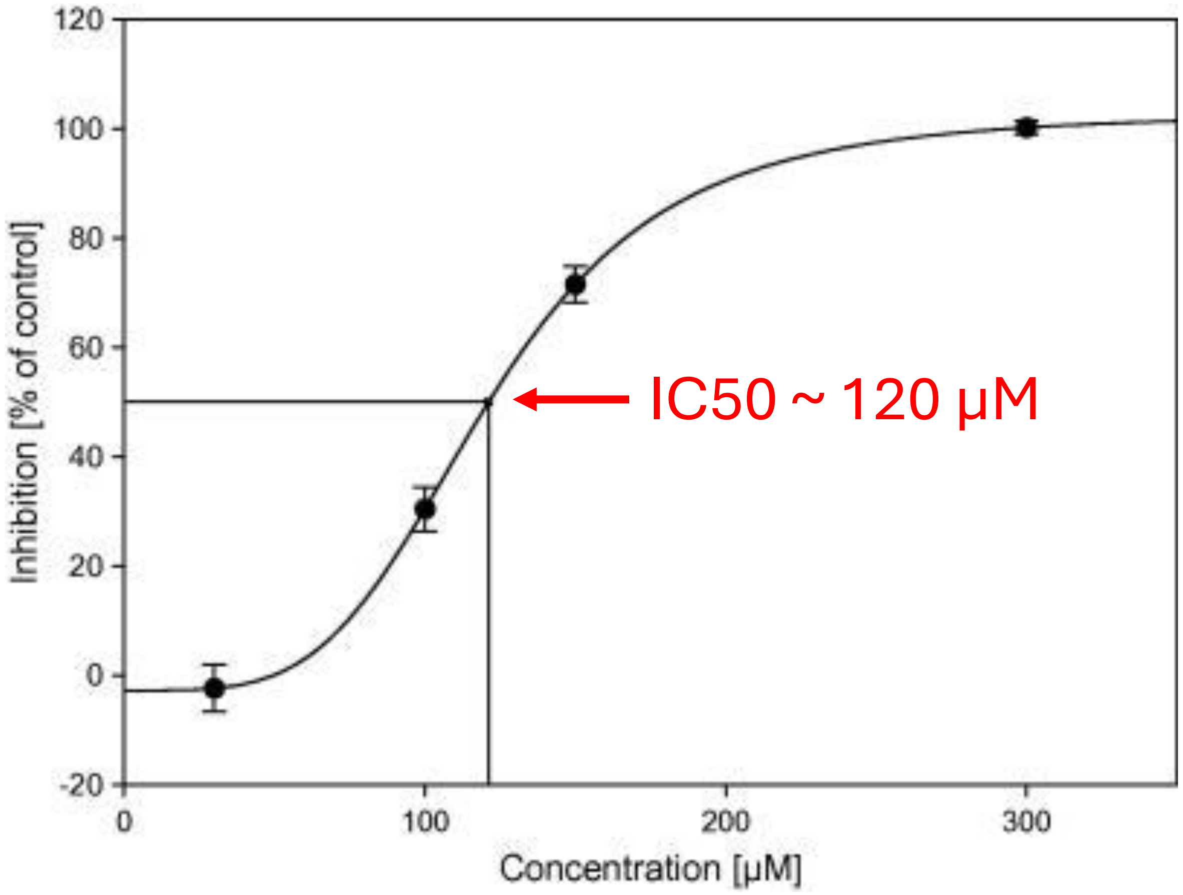
Taylor C, Crosby I, Yip V, Maguire P, Pirmohamed M, Turner RM. A Review of the Important Role of CYP2D6 in Pharmacogenomics. Genes (Basel). 2020;11(11):1295. Published 2020 Oct 30. doi:10.3390/genes11111295

- CYP2D6 is one of the most well-studied drug-metabolizing enzymes
- In the case of certain drugs, such as atomoxetine, metabolism by CYP2D6 results in adverse effects
- Pharmacological inhibition of CYP2D6 is thus an important component in preventing drug-drug interactions for many diseases (see figure)

Bertilsson L, Dahl ML, Dalén P, Al-Shurbaji A. Molecular genetics of CYP2D6: clinical relevance with focus on psychotropic drugs. Br J Clin Pharmacol. 2002;53(2):111-122. doi:10.1046/j.0306-5251.2001.01548.x
Cicali EJ, Smith DM, Duong BQ, Kovar LG, Cavallari LH, Johnson JA. A Scoping Review of the Evidence Behind Cytochrome P450 2D6 Isoenzyme Inhibitor Classifications. Clin Pharmacol Ther. 2020;108(1):116-125. doi:10.1002/cpt.1768

Half-Maximal Inhibitory Concentration

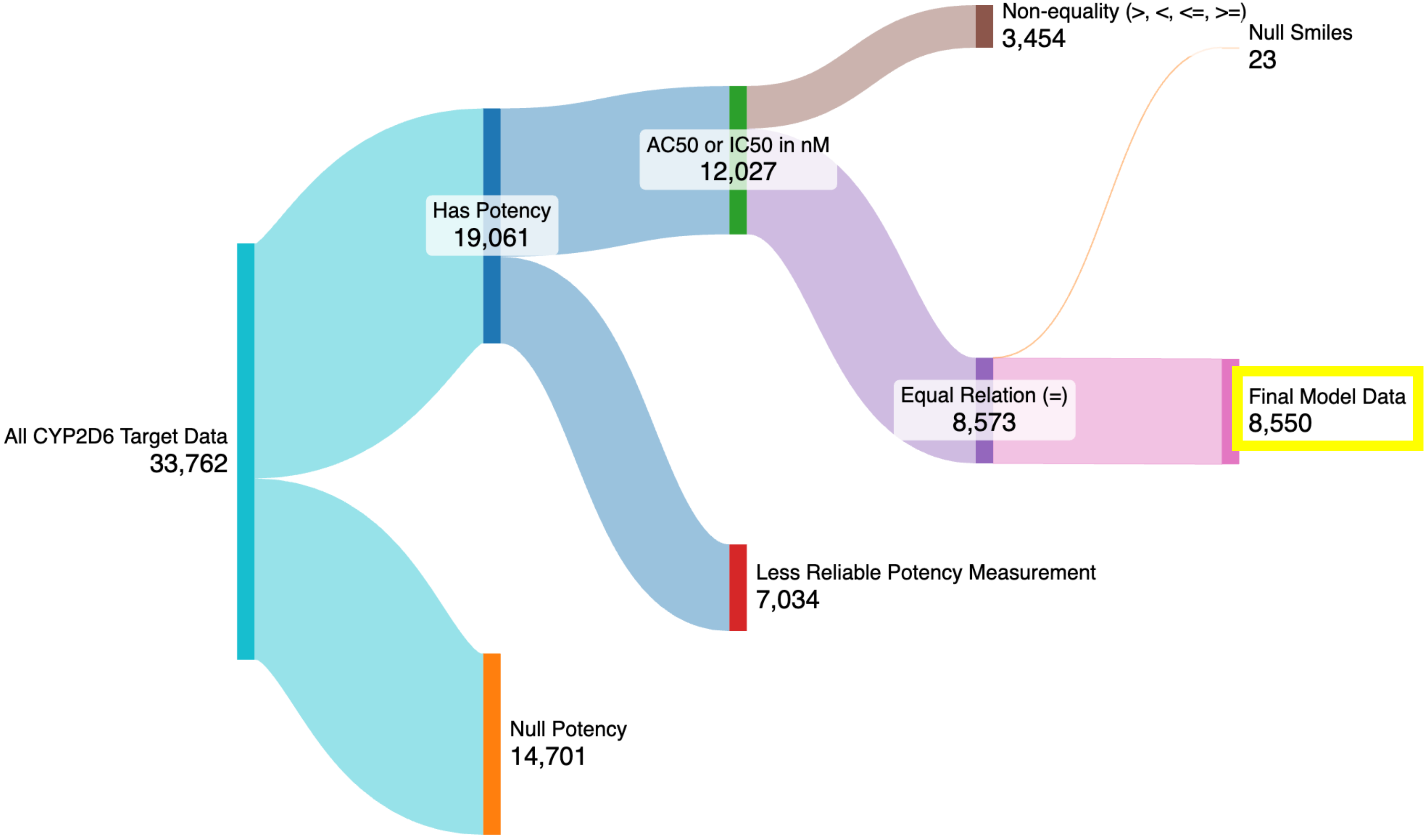
- Represents the concentration at which 50% of normal protein activity is observed
- Provides a basis for comparing the inhibitory potency of different substances



IC50 ~ 120 μM

Data Collection and Processing

- Data were obtained from ChEMBL
- Initial filter was for all compounds targeting human CYP2D6
- Isolated for compounds with IC50 or AC50 data, as these measurements of potency can be reliably compared.
  - These values are effectively the same and are used interchangeably
- Some potency measurements were not equalities (for example ‘greater than 10,000’), and these were excluded as it was unclear what the true value was
- The final model dataset included 8550 compounds after filtering out null SMILES strings



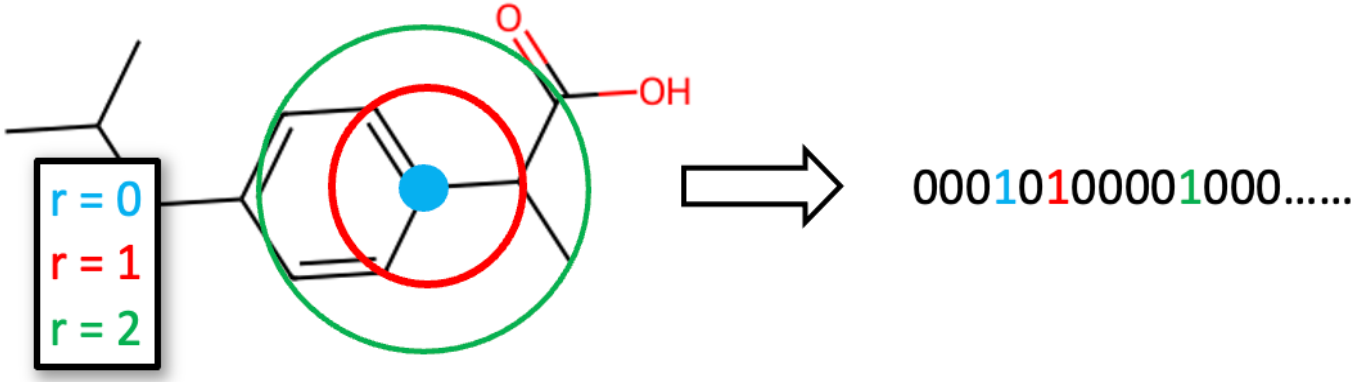
Made at SankeyMATIC.com

Data Generation: Descriptors and Fingerprints

- Because I wanted to build an in-silico model to predict IC50, I elected not to use any experimental data for the compounds other than the IC50.
- My reasoning for doing this is that my resulting model could be tested on molecules generated computationally via enumeration, where molecules identified to have high potency (leads) are modified with additional functional groups to increase their efficacy (virtual screening)
- Predictions were made using either Morgan Fingerprints or molecular descriptors generated using RDKit

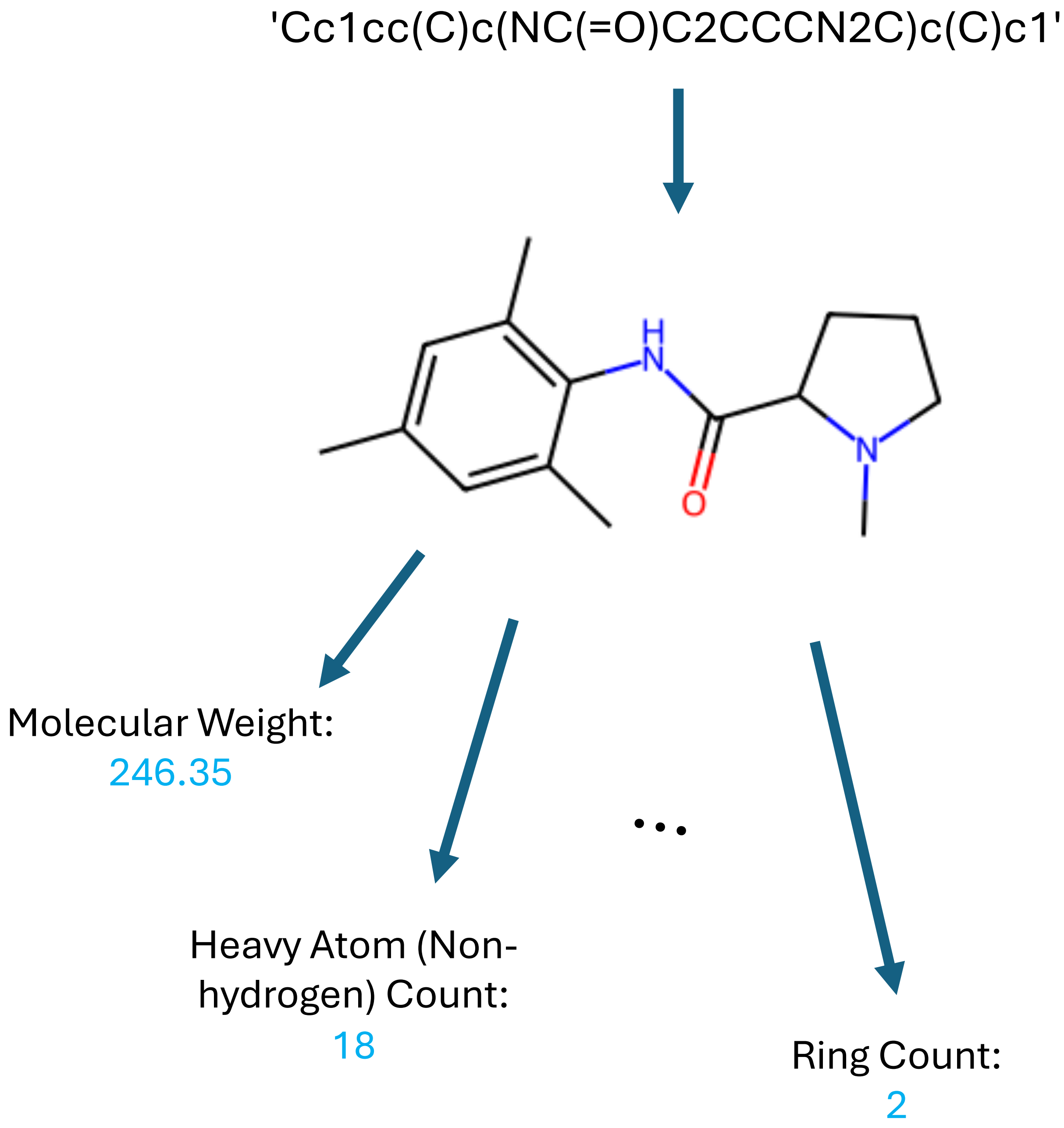
Morgan Fingerprints

- Morgan fingerprints, also known as circular fingerprints, encode molecules into bitvectors (literally lists of ones and zeroes)
- Each ‘1’ corresponds to a substructure, which is extracted by considering all atoms bound at a certain bond radius from each atom (see figure)
- In my case, I chose a radius of 2 and a length of 512 to reduce dimensionality (the number of features in my dataset)



Molecular Descriptors

- Molecular descriptors were generated computationally using RDKit based on SMILES Strings (a text-based representation of a molecule)
- A total of 210 were calculated



'Cc1cc(C)c(NC(=O)C2CCCN2C)c(C)c1'

Molecular Weight: 246.35

Heavy Atom (Non-hydrogen) Count: 18

...

Ring Count: 2

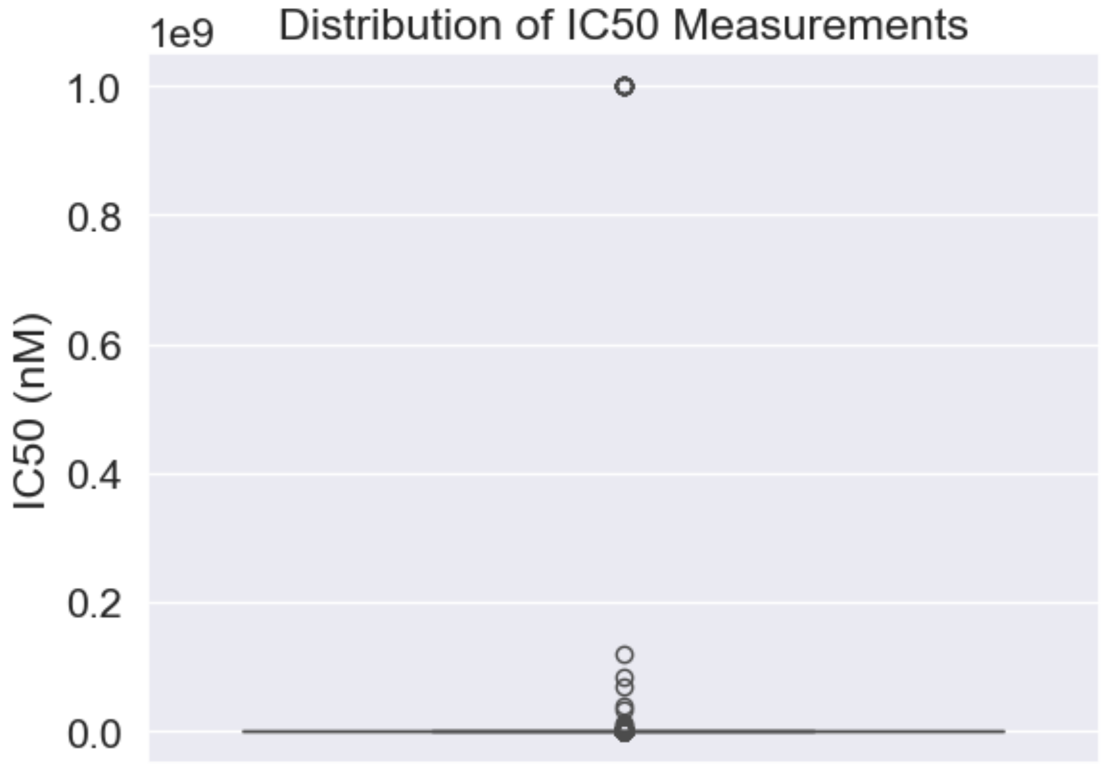
Classification of Molecule Potency

- Because of a few extreme outliers in the IC50 measurements, I thought it would be advantageous to convert this to a classification problem
- The classes were defined as follows:

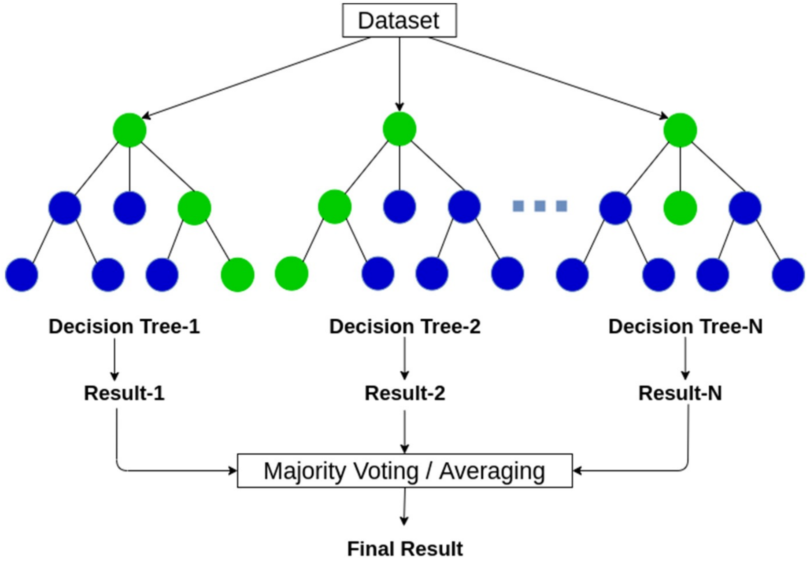
0-150 nM: strong inhibition

151-400 nM: weak inhibition

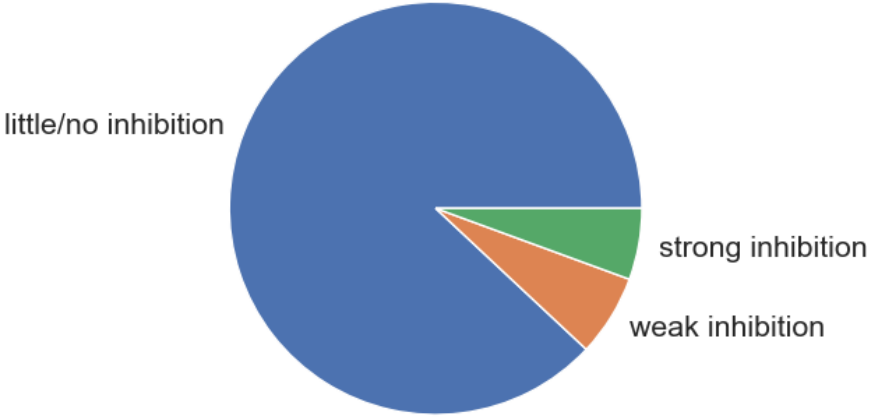
> 400 nM: little/no inhibition



Selection of Model Type



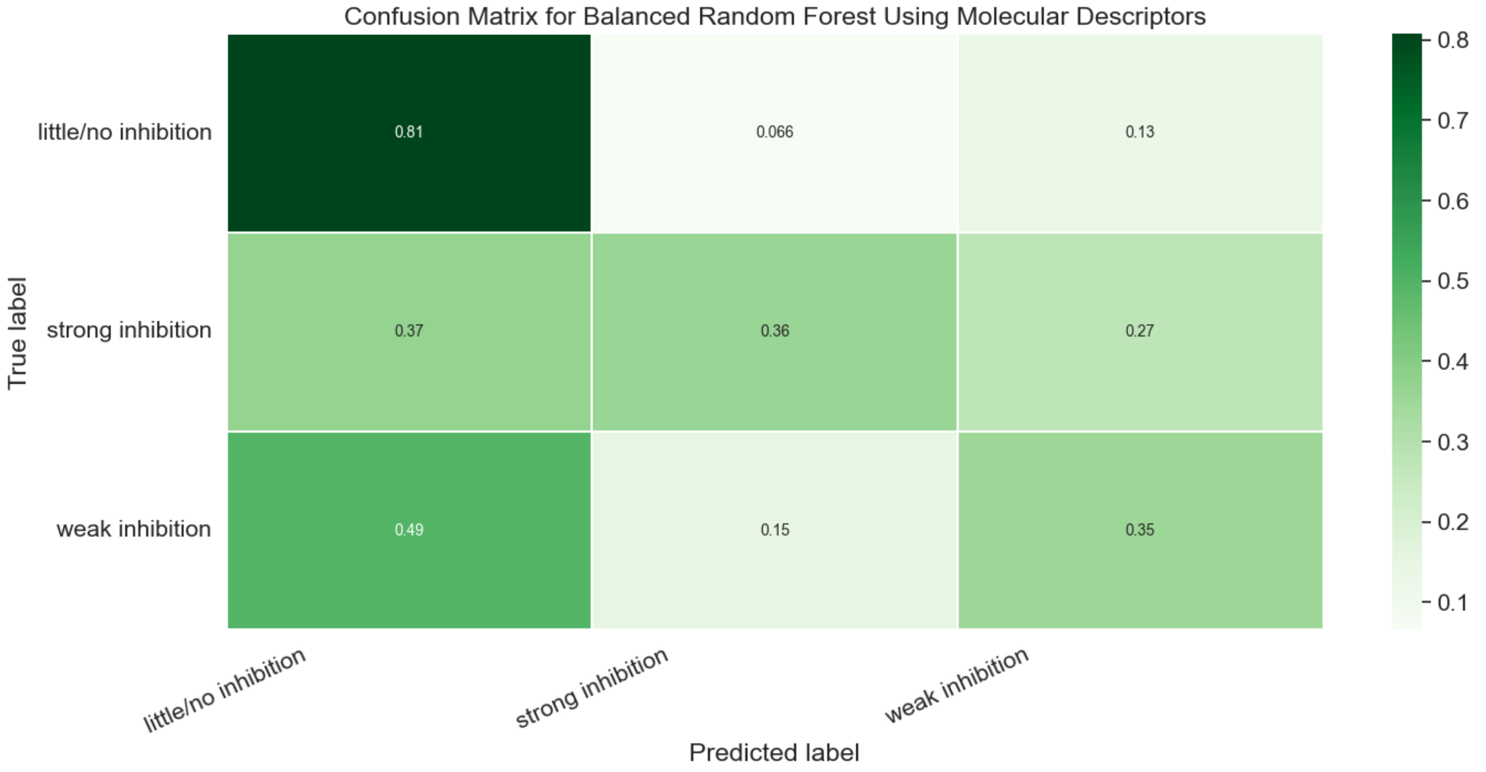
Breakdown of Inhibition Classes in Model Data



- Due to the high dimensionality of the data for both fingerprints and descriptors (512 and 210 features, respectively), coupled with the sparsity of the fingerprint data, I went with Random Forest Classification
  - Random forest classification a method which aggregates results from many randomly-generated decision *trees* (hence ‘random forest’)
- Specifically, because my data were *unbalanced* (there were many more inactive than active compounds), I used Balanced Random Forest Classification

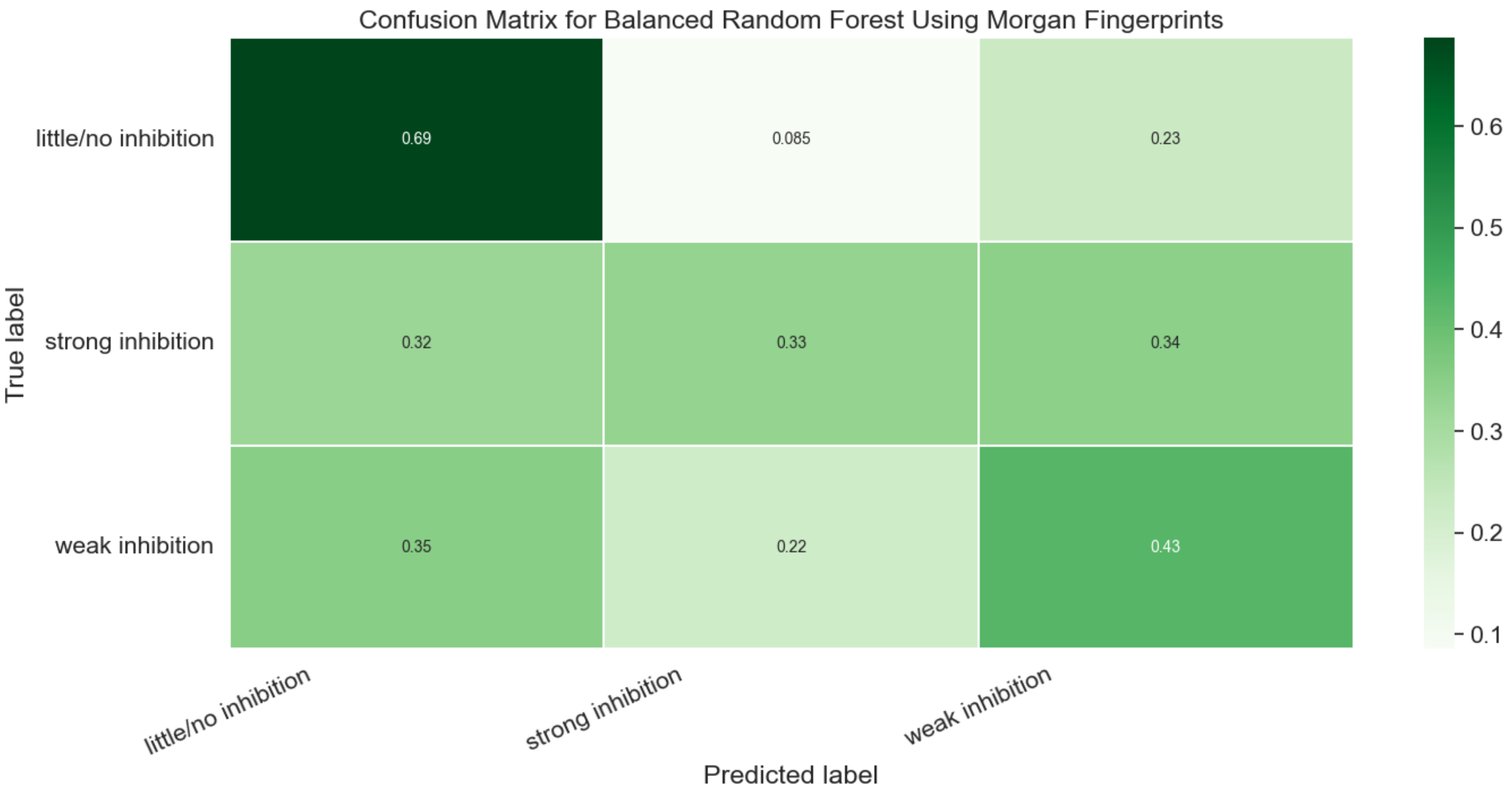
Model Performance: Molecular Descriptors

- Accuracy: 0.7542
  - **Accuracy is based on 20% test split
- Although accuracy was acceptable, the model did not perform well on classes other than ‘little/no inhibition’ as shown in the confusion matrix.



Model Performance: Morgan Fingerprints

- Accuracy: 0.6518
- Accuracy for Morgan Fingerprints was inferior to molecular descriptors
- Again, the model biases towards the little/no inhibition class, likely because of the unbalanced dataset



Conclusions

- These models did not perform as well as I would have liked, although I did learn a lot about the random forest method.
- In the future, I’d like to perform more comprehensive variable selection to try and improve model performance
- Experimenting with certain hyperparameters including the number of trees, length of fingerprint, etc. may be beneficial
- I’d also like to expand to other model types, specifically graph convolutional neural networks which are the new state-of-the-art in molecular property prediction
- Once that is completed, I’d like to use the resulting model to evaluate enumerated compounds (virtual screening)

Technology/Packages Used

- Scikit-Learn
- Numpy
- Pandas
- Seaborn
- RdKit
- Jupyter Notebooks
- Python
- Conda Environments