Data driven predictive machine learning modeling for PARP1 inhibition

Ryan Friedrich[IN], Amanda Paulson Ph.D.[CA]
Acknowledgements: Dr. Caleb Class, Da Shi Ph.D., Sarangan Ravichandran Ph.D., Susan Mertis Ph.D.
Butler University, Indianapolis, IN
ATOM Consortium, San Francisco, CA

Abstract

The PARP1 enzyme plays a role in DNA damage repair (DDR) by recruiting DDR proteins that allow tumors to remain viable. Computational modeling to target PARP1 is an important piece of the drug discovery process. While *in vitro* models are costly and time-consuming, computational modeling can reduce such costs. A model that can predict inhibitory strength allows for selection of the most potent PARP1 inhibiting molecules. These inferences can help identify molecular characteristics of the strongest inhibitors. We used public databases to gather SMILES strings, Ki or IC50 values of PARP1 inhibition, and the corresponding assay descriptions. The assays were analyzed to ensure comparable data was used. The datasets were curated by removing outliers, standardizing values, and averaging duplicate values. The data were combined into two final datasets: one for IC50's and another for Ki's. Regression and classification models were made using neural networks, random forests, and XGBoost. Descriptors used to characterize the molecules include ECFP, RDKit, Mordred, MOE, and graph convolutional neural networks. The best regression model for Ki's produced an R^2 value 0.7457 using ECFP features and random forest, while for IC50 values the best R^2 was 0.6443 using Mordred descriptors with random forest. The best classification model for Ki's produced an ROC AUC score of 0.978 using a graph convolutional neural network, while for IC50 values the best ROC AUC was 0.8985 using ECFP with random forest. These PARP1 targeting models will be used in the ATOM GMD loop for PARP1 inhibitor development.

Data Collection and Curation

The databases used to gather data include ChEMBL, Drug Targeting Commons, PubChem, and Excape. This data was gathered experimentally as Ki and IC50 values and published onto these public databases. The IC50 and Ki values were converted to pKi and pIC50 values within Jupyter for ease of use. Key information gathered includes Ki or IC50 values, molecular structures represented as a SMILES string, and the description of the experimental assay used. Each individual dataset was curated within Jupyter. The curation process involved ensuring each molecule had all relevant information, removing outlier values over a set standard deviation, and averaging values when multiple were available for the same molecule. The individual curated datasets were then combined into two final datasets, one for pKi values and another for pXC50 values. Both final datasets were eventually used to build several types of classification and regression models.

Curation Challenges - Excape

All data from the Excape database was cut from the final data. The cut occurred because Excape was reporting numerous value types including Ki and IC50 but did not provide a way to differentiate the value types. I originally believed the values were all IC50 values, therefore after this discovery the data was no longer useful. This discovery was made by comparing IC50 and Ki values for the same molecules. All other datasets included a clear description of the value type being measured.

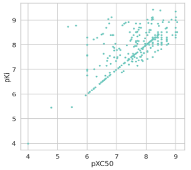


Figure 1. Scatter plot comparing pXC50 vs pKi values for molecules with both values. It is clear many values correlate far too closely. Graph exposed the Excape data's flaw. The diagonal line values represent values from Excape in the XC50 data that were really Ki values and therefore were identical to Ki values reported by the other datadatabases for same molecule.

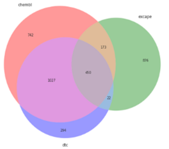


Figure 2. Escape data would have provided a total of 876 additional unique molecules and could have improved molecular diversity within both datasets signigicantly.

Compounds

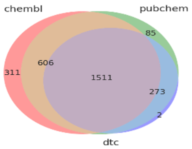


Figure 3. Venn diagram of molecules from each dataset. Many shared molecules from in all three original sets. The final XC50 data includes 2,818 unique molecules while the Ki data includes 1,088 unique molecules.

Acknowledgements

I would like to specifically thank Amanda Paulson, Caleb Class, Da Shi, Susan Mertis, and Sarangan Ravichandran for their help throughout the course of this project. I would not have been able to complete this project without their assistance and I am very grateful for everything they have taught me over the last couple months.

Final Datasets

An active label was placed on all molecules with a pXC50 exceeding 7 or a pKi exceeding 7.7. These threshold values were chosen to create an even split of active and inactive molecules for each dataset. Each original dataset from each database provided far more XC50 values than Ki values.The final XC50 dataset consists of 2,818 molecules and the final Ki dataset consists of 1,088 molecules.. The final Ki data is composed of 545 active molecules and 543 inactive based on a threshold of 7.7 while the XC50 final data is composed of 1404 active and 1414 inactive molecules based on threshold of 7.

Figure 4.

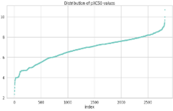


Figure 5.

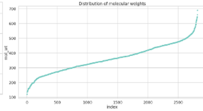


Figure 4 shows the XC50 datasets distribution of pXC50 values. All values fall between 2.5 and 10.5 with the majority of values falling between 6 and 8. Figure 5. shows the distribution of molecular weights. 2,622 molecules have molecular weights under 500 which is a positive sign for developability. The Ki data shared these trends with XC50 data. Although, the pKi values where shifted slightly higher than the XC50 values with all values falling between a pKi of 5 to 10 and the bulk of values ranging from 7 to 9.

Modeling Methods

The machine learning models used throughout this project included random forests, neural networks, and XgBoost. These models were used for both regression and classification modeling. Descriptors used to characterize the molecules include ECFP, RDKit, Mordred, MOE, and graph convolutional neural networks.. Every model type was used in combination with each featurizer and descriptor (except graphconv only with neural network). After initial models were produced, the results were analyzed to adjust parameters to improve the performance of each model..Initial modeling was performed with fairly random parameters, but over time these choices were altered to improve each model's performance.

Ki data vs XC50 data

Figure 6.

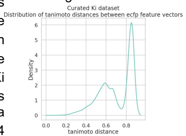
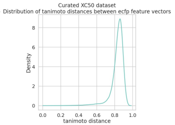


Figure 7.



The above graphs show the distribution of tanimoto distances between ecfp features of molecules from both datasets. Figure 7 represents the XC50 data and has its only peak around 0.85 signifying a large degree of structural diversity among the data's molecules. Figure 6 represents the Ki dataset and has a peak at 0.85 and another at 0.6. The existence of this peak at 0.6 shows that the Ki dataset is less diverse than the XC50 dataset. Less diversity in the Ki data was expected with a smaller dataset.

Figure 8.

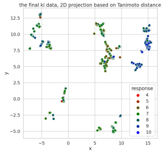
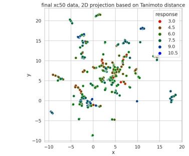


Figure 9.



The above graphs are 2D projections of the tanimoto distances of each dataset. Figure 8 represents the Ki data while Figure 9 represents the XC50 data. Notice the XC50 dats's projection has points that are spread out all over the grid with large distance between many molecules. Alternatively, the Ki data's projection shows points that are mostly located in noticeable groups. These groupings also show how the Ki data is less diverse. Most molecules in the Ki data can be considered part of a specific cluster of molecules with highly similar structures and therefore similar Ki values. Notice these clusters are largely colored the same which shows those pKi values are very close to each other.

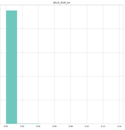


Figure 12. Histogram shows very small standard deviation values in Ki dataset. This highlights the consistency of the Ki measurement.

Conclusion - Results

The highest performing regression models for Ki values produced an r2 value 0.7457 using ECFP features with a random forest model, while for XC50 values the best r2 was 0.6443 using mordred descriptors with random forest. The highest performing classification models for Ki values produced an roc auc score of 0.978 using graphconv features with a neural network, while for XC50 values the best roc auc was 0.8985 using ECFP with Random Forest.

Figure 10.

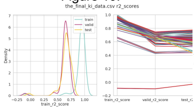
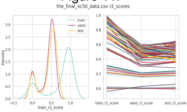


Figure 11.



The above two figures show most of the results for regression models from each dataset with Figure 10 representing Ki data and Figure 11 for XC50 data. Models for the Ki data consistently outperformed models for the XC50 data. Most r2 scores produced from pKi values range from 0.6 to 0.74 while most r2 scores from XC50 data fall between 0.4 and 0.6. The pKi values had very small standard deviations which may show it was a more consistent measure than IC50 leading to better results as values could have related closer to structure. Also, a lack of diversity in Ki data could have led to models memorizing values from train set compounds. Then making predictions based on the train values from that cluster of highly similar molecules rather than predicting by structure.

Future Aims

My first goal is to build a model that predicts both pKi and pIC50 values for molecules which would allow for using the measurements together. In the future I would like to gather more information on PARP1 inhibiting molecules especially Ki values. I would like to see how an increase in diversity of those molecules effects the results. I would also like to find a way identify Excape data's value types to use those unique molecules. Most importantly I want to use results and insights from these models to help predict new PARP1 inhibitors.

References

IMalyuchenko, N. V., Kotova, E. Y., Kulaeva, O. I., Kirpichnikov, M. P., & Studitskiy, V. M. (2015). *PARP1 Inhibitors: antitumor drug design*. Acta naturae. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4610162/.

Slade, D. (2020, March 1). *PARP and PARG inhibitors in cancer treatment*. Genes & development. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7050487/.

Pommier, Y., O'Connor, M. J., & Bono, J. de. (2016, October 26). *Laying a trap to kill cancer cells: PARP inhibitors and their mechanisms of action*. Science Translational Medicine. https://stm.sciencemag.org/content/8/362/362ps17.full.

Fu, L., Wang, S., Wang, X., Wang, P., Zheng, Y., Yao, D., Guo, M., Zhang, L., & Ouyang, L. (2016, December 5). *Crystal structure-based discovery of a novel synthesized PARP1 inhibitor (OL-1) with apoptosis-inducing mechanisms in triple-negative breast cancer*. Nature News. https://www.nature.com/articles/s41598-016-0007-2.

Wang, S., Han, L., Han, J., Li, P., Ding, Q., Zhang, Q.-J., Liu, Z.-P., Chen, C., & Yu, Y. (2019, October 28). *Uncoupling of PARP1 trapping and inhibition using selective PARP1 degradation*. Nature News. https://www.nature.com/articles/s41589-019-0379-2.

atomscience.org | @ATOM_consortium | #ATOMscience