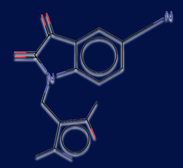


Improving Drug Discovery through Machine Learning

Improving Drug Discovery by Utilizing Regression based Machine Learning Models and Biological Activity Data of Target Proteins

pIC50

Lagnajeet Panigrahi

Abstract

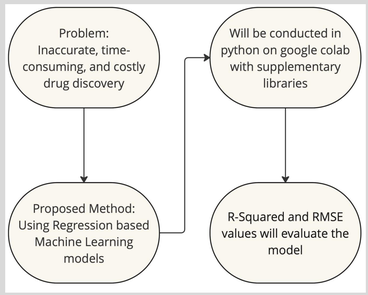
Humans have been battling the age-old battle against illness for centuries. Only recently have humans made significant progress in averting illness through the creation of drugs and the field of drug discovery. To create a drug and deem it usable requires a rigorous process that examines one attribute in particular: safety. Unfortunately, ninety percent of drug development fails despite the many successful strategies used. Current chemical methodologies rely on a hit-and-miss approach where large amounts of drugs are analyzed for their properties by hand. This process is expensive, time-consuming, and often inaccurate. This project aims to address this through the implementation of various machine-learning algorithms that can predict the IC50 and pIC50 values of possible drug candidates. These values are essential in determining the quantity of a drug needed to inhibit a biological process by half but also are important in determining the toxicity of a drug and how it impacts patients. Using regression-based machine learning models and bioactivity data of compounds and target proteins, these values can be predicted and outputted. Upon completion of this project, we developed multiple regression models for the target protein of the SARS coronavirus. After statistical analysis, the best model was chosen: DecisionTreeRegressor. This model had a root mean squared error score of 0.34 and an R-squared score of 0.9019. This implies that this model fits the situation and makes accurate predictions. We conclude that drug discovery can become quicker, more accurate, and cost-effective through the implementation of machine learning algorithms.

Keywords: Drug discovery, IC50, pIC50, machine learning, regression, bioactivity, r-squared score, root mean squared error score

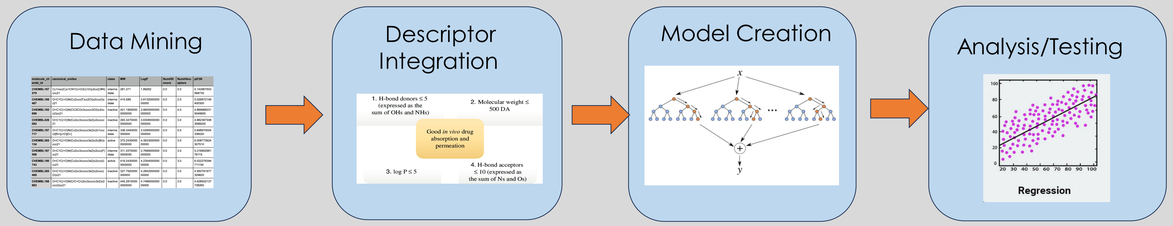
Introduction

In this expanding world where illness and disease are increasingly common, efficient, accurate, and cost-effective approaches to drug discovery are becoming more necessary. Inaccuracies and other sources of error can greatly delay drug development which can have disastrous impacts on individuals around the world. It was calculated that there were 29 life-years lost in North America alone per one hour of delay in drug approval (Helwick, 2015) and 90 percent of clinical drug development fails (Sun et al., 2022). A major step in drug development is determining the toxicity of the drug, its potency, and overall, its safety. Current methods require careful experimentation and individual assessment of large amounts of potential drug candidates by hand (Blanco-González et al., 2023). Although current methods are good, they are prone to errors, are costly, and are time-consuming (Blanco-González et al., 2023). Consequently, together, the high likelihood of failure combined with the inherent problems with current methods create a system that fails to perform effectively and efficiently. The proposed method that this project will employ to solve this problem is the use of regression-based machine learning algorithms in Python with preprocessed data to give information on drug toxicity and potency by predicting IC50 and pIC50 values. To build the model, any target protein can be utilized (the only difference between models of different target proteins is what they are trained and tested on). For this project, the SARS coronavirus 3C-like proteinase target will be used. The model's success will be determined by statistical analysis.

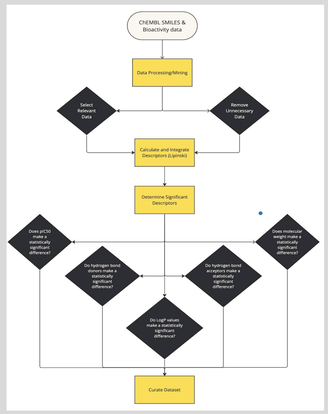
Visual Overview

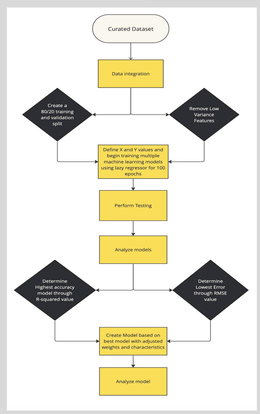


Methodology



Detailed Procedure





Data Mining & Descriptor Integration: After the target protein search is performed from ChEMBL and the bioactivity data is collected, the data must be preprocessed. First, the unnecessary data that is not relevant to the project is removed. Then the Lipinski descriptors are integrated and calculated. Once this is done, statistical analysis in the for of a U-test will be performed to determine which features are the most useful in the determination of a drug's toxicity and how it is absorbed into the body. Finally, the data set will be curated.

Model Creation & Analysis: First, the curated dataset will be iterated through a program that will remove all low variance features to ensure accurate training. Next the data will be split into training and testing sets following a 80/20 split. Next, define X and Y values and begin training and testing. Finally, once models are trained and tested, a statistical analysis of R-squared values and RMSE values will be taken.

Data Mining & In-Silico Model Testing Results

Figure 1a. Amount Active and Inactive Molecules in dataset

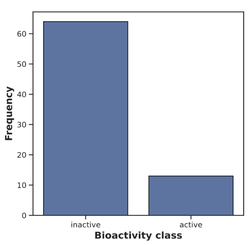


Figure 1a displays the quantity of inactive and active molecules within the dataset. 13 molecules are active at the target protein and 65 inactive molecules. There was a total of 78 molecules.

Table 1a. Statistical Significance of pIC50

	Descriptor	Statistics	p	alpha	Interpretation
0	pIC50	832.0	1.5880944752282E-08	0.05	Different distribution (reject H0)

This table displays the result of the Mann-Whitney U test for pIC50. There was a statistically significant difference between the pIC50 value for active and inactive molecules and drugs. This can be used as a distinguisher in the machine learning regression model.

Table 1b. Statistical Significance for Number of Hydrogen Bond Donors

	Descriptor	Statistics	p	alpha	Interpretation
0	NumHDonors	178.5	0.0006202477871294170	0.05	Different distribution (reject H0)

This table displays the result of the Mann-Whitney U test for the number of hydrogen bond donors. There was a statistically significant difference between the number of hydrogen bond donors for active and inactive molecules and drugs. This can be used as a distinguisher in the machine learning regression model.

Table 1c. Statistical Significance for Number of Hydrogen Bond Acceptors

	Descriptor	Statistics	p	alpha	Interpretation
0	NumHAcceptors	248.0	0.020963970330247300	0.05	Different distribution (reject H0)

This table displays the result of the Mann-Whitney U test for the number of hydrogen bond acceptors. There was a statistically significant difference between the number of hydrogen bond acceptors for active and inactive molecules and drugs. This can be used as a distinguisher in the machine learning regression model.

Table 1d. Statistical Significance of Molecular Weight

	Descriptor	Statistics	p	alpha	Interpretation
0	MW	249.0	0.015566956435349300	0.05	Different distribution (reject H0)

This table displays the result of the Mann-Whitney U test for molecular weight. There was a statistically significant difference between the molecular weight of active and inactive molecules and drugs. This can be used as a distinguisher in the machine learning regression model.

Table 1e. Statistical Significance of LogP

	Descriptor	Statistics	p	alpha	Interpretation
0	LogP	425.0	0.9079043598683140	0.05	Same distribution (fail to reject H0)

This table displays the result of the Mann-Whitney U test for druglikeness (LogP). There was not a statistically significant difference between the LogP of active and inactive molecules and drugs. This cannot be used as a distinguisher in the machine learning regression model.

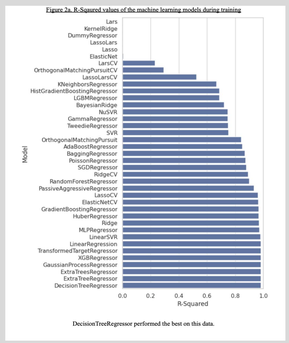
Table 2a. R-squared and RMSE values of the best model during training

Model	R-Squared Score	RMSE Score
DecisionTreeRegressor	0.98	0.13

Table 3a. R-squared and RMSE values of the best model during testing

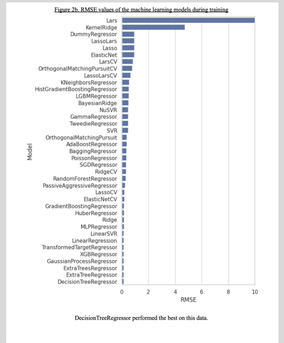
Model	R-Squared Score	RMSE Score
DecisionTreeRegressor	0.9019	0.34

Figure 2a. R-Squared values of the machine learning models during training



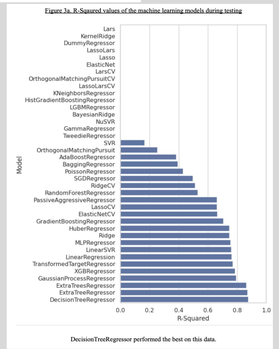
DecisionTreeRegressor performed the best on this data.

Figure 2b. RMSE values of the machine learning models during training



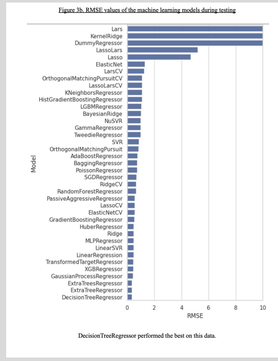
DecisionTreeRegressor performed the best on this data.

Figure 3a. R-Squared values of the machine learning models during testing



DecisionTreeRegressor performed the best on this data.

Figure 3b. RMSE values of the machine learning models during testing



DecisionTreeRegressor performed the best on this data.

Statistical Analysis

The statistical tests show that all descriptors except for druglikeness (LogP) contribute (they create a statistically significant difference as their p-values are less than 0.05) to whether or not a molecule will be active or inactive and thereby have an impact on pIC50 and IC50 values. Using LazyRegressor and the dataset, many models were created and analyzed by their RMSE and R-squared values. The best model after testing was determined to be the DecisionTreeRegressor machine learning model. This model had the lowest RMSE value (0.34) and the highest R-squared value (0.9019). Having a large R-squared value is favored as that implies that the model fits the situation well and the predictions are accurate because the predictions are closer to the line of best fit (the actual values). RMSE is the average distance that all of the predictions are from the actual values. Having lower RMSE values is better because it would imply that the predictions are not very far off from the actual values. Together the low RMSE value and R-squared value convey that the developed model is accurate and fast whilst also being cost-effective.

Conclusion/Discussion

The ultimate goal of this project was to develop and test regression-based machine learning models that could make predictions about drugs and molecules designed for SARS Coronavirus 3C-like proteinase while also being accurate, fast, and cost-effective. The models would predict the potency and toxicity by outputting an IC50 value and pIC50 value for the molecule. These models made predictions based on SMILES notation and bioactivity data of the molecules and drugs from the database ChEMBL. In the end, the best regression-based machine learning model, DecisionTreeRegressor achieved an RMSE value of 0.34 and a R-squared value of 0.9019. This shows that the model was fairly accurate while also being cost-effective and quick. The results could have been improved with the inclusion of more data about SARS Coronavirus 3C-like proteinase. Despite the lack of data and an average accuracy of less than 95 percent, we deem this project to be a success. This is because although it did not reach a very high accuracy it was still fairly accurate considering the constraints. In addition, this project proved that using these regression-based machine learning models is a possible avenue in improving the field of drug discovery. At the very least, the model developed in this project can be used as more of an aid and supplement to scientists to make preliminary judgments and to confirm that the data they receive is accurate. By continuing to build on the work conducted in this project we can continue to improve the field of drug discovery and save millions of lives.

Future Work

In the future, similar regression-based machine learning models can be expanded to other fields and even areas of drug discovery rather than focusing on the target protein SARS Coronavirus 3C-like proteinase. Work can also be done towards optimizing the current model or even testing new models with the same situation to look for ways to improve accuracy and decrease prediction error. The main action that can be taken to improve the accuracy of these types of models is gathering more data. Machine learning algorithms rely heavily on training and testing data. Without ample data, it is not possible to create models that reach accuracies of 97 percent higher. Finally, to improve the usability of these models, web applications can be formed to streamline the process of entering data and outputting a prediction.

Key References

Alvarellos, M. (2023, October 5). *What are the current challenges of drug discovery?* Www.lifebit.ai. https://www.lifebit.ai/blog/current-challenges-of-drug-discovery

Aykul, S., & Martinez-Hackert, E. (2016). Determination of half-maximal inhibitory concentration using biosensor-based protein interaction analysis. *Analytical Biochemistry, 508*(1), 97–103. https://doi.org/10.1016/j.ab.2016.06.025

Berrouet, C., Dorilas, N., Rejniak, K. A., & Tuncer, N. (2020). Comparison of Drug Inhibitory Effects (IC50) in Monolayer and Spheroid Cultures. *Bulletin of Mathematical Biology, 82*(6). https://doi.org/10.1007/s11538-020-00746-7

ChEMBL. (n.d.). *ChEMBL Database*. Www.ebi.ac.uk. https://www.ebi.ac.uk/chembl/

Code Ocean. (n.d.). *Code Ocean*. Codeocean.com. Retrieved February 1, 2024, from https://codeocean.com/explore/capsules?query=tag:data-curation

Laerd Statistics . (2013). *Mann-Whitney U Test in SPSS Statistics*. Laerd Statistics. https://statistics.laerd.com/spss-tutorials/mann-whitney-u-test-using-spss-statistics.php

Liu, P., Li, H., Li, S., & Leung, K.-S. (2019). Improving prediction of phenotypic drug response on cancer cell lines using deep convolutional network. *BMC Bioinformatics, 20*(1). https://doi.org/10.1186/s12859-019-2910-6