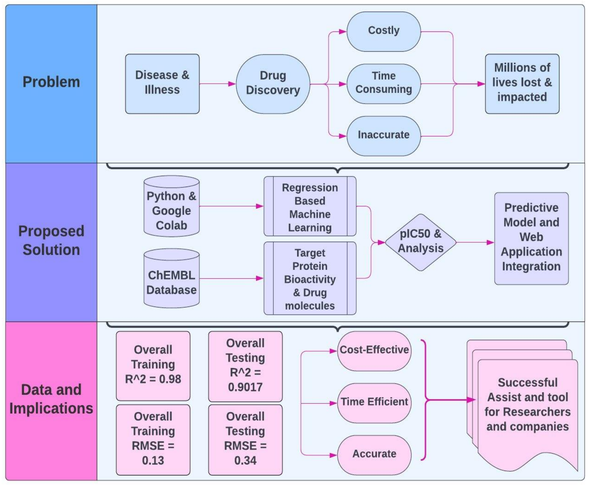
Improving Drug Discovery through Machine Learning

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

Lagnajeet Panigrahi

pIC50

Visual Abstract



Introduction

Drug Discovery

Drug Discovery is the field of science that analyzes and produces new chemical entities that have therapeutic uses. This field is crucial in the fight against disease and illness.

Current Methods

A major step in drug development is determining the toxicity of the drug, its potency, and overall, its safety. Current methods utilize biological assays for property detection.

H2L and Lead Discovery

Lead Discovery and Hit-two-lead are the stages in drug discovery in which assays are performed to predict certain characteristics of molecules and identify optimal characterisitcs.

Problem

90% **90 percent of clinical drug development fails**

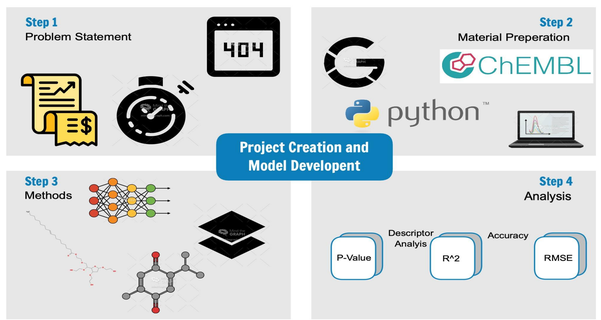
Clinical Trials can cost from 30-300 million dollars

29 life-years lost in North America alone per one hour of delay in drug approval

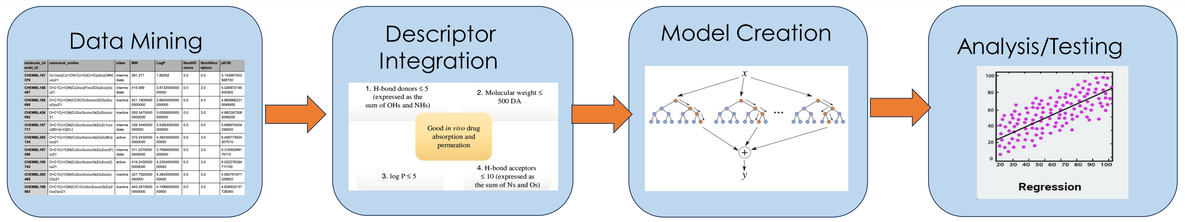
Hypothesis/Solution

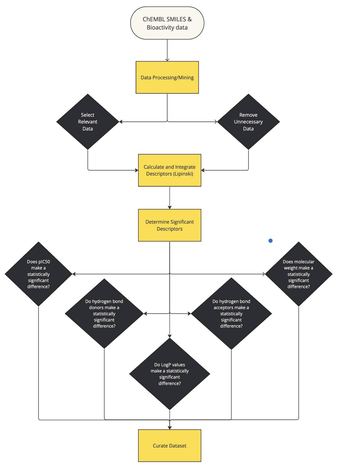
To solve this problem, regression-based machine learning algorithms with preprocessed data of pIC50 values and bioactivity will be employed. The SARS coronavirus 3C-like proteinase target will be used in testing of the model. However, this same model can be applied to any target protein.

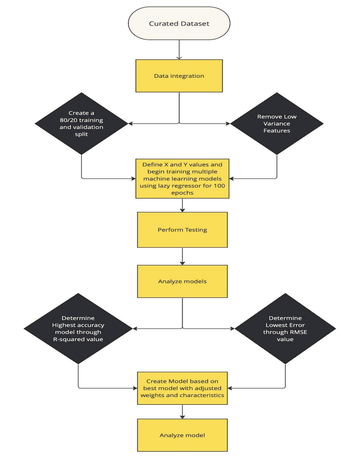
Visual Overview



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

Table 1a. Statistical Significance of pIC50

	Descriptor	Statistics	p	alpha	Interpretation
0	pIC50	832.0	1.5880944752824E-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 Accepters

	Descriptor	Statistics	p	alpha	Interpretation
0	NumHAcceptors	248.0	0.020963972338247300	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	0.023566956483549300	0.05	Different distribution (reject H0)

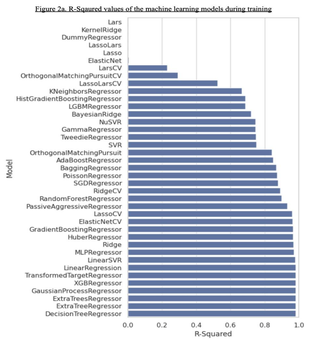
This table displays the result of the Mann-Whitney U test for molecular weight. There was a statistically significant difference between the molecular weight of active and inactive molecules and drugs. This can be used as a distinguisher in the machine learning regression model.

Table 1e. Statistical Significance of LogP

	Descriptor	Statistics	p	alpha	Interpretation
0	LogP	425.0	0.9079804356663140	0.05	Same distribution (fail to reject H0)

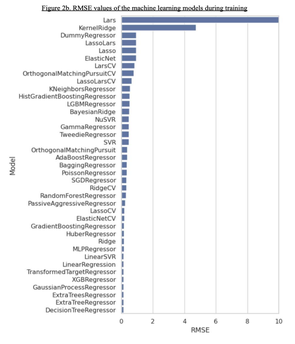
This table displays the result of the Mann-Whitney U test for druglikeness (LogP). There was not a statistically significant difference between the LogP of active and inactive molecules and drugs. This cannot be used as a distinguisher in the machine learning regression model.

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



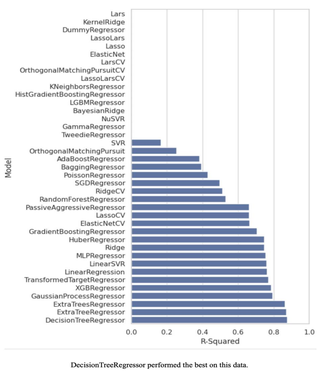
DecisionTreeRegressor performed the best on this data.

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



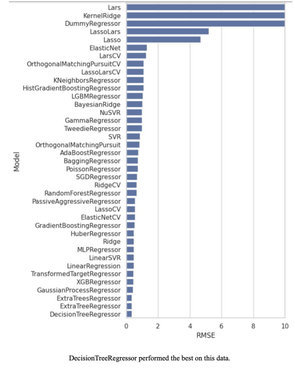
DecisionTreeRegressor performed the best on this data.

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.

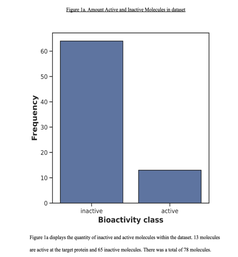
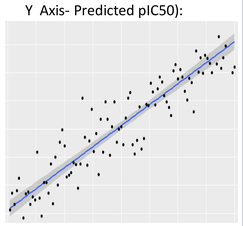


Figure 1a. Amount Active and Inactive Molecules in dataset

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.

Scatterplot of Predictions versus Experimental Values (X Axis- Experimental pIC50, Y Axis- Predicted pIC50):



R-squared and RMSE values of DecisionTreeRegressor

	Training	Testing
R-squared	0.98	0.9019
RMSE	0.13	0.34

Statistical Analysis

1 **$P < 0.05$**

Statistical Significance of Descriptors

The statistical tests confirm that all descriptors except for druglikeness (LogP) contribute (they create a statistically significant difference as their p-values are less than 0.05) and have an impact on pIC50 and IC50 values. **Therefore, theses descirptors can be used for accurate prediction of IC50 and pIC50 values.**

2 90.19%

Accuracy Metrics

Using LazyRegressor and the dataset, many models were created and analyzed by their RMSE and R-squared values. The best model after testing and modifications was determined to be the DecisionTreeRegressor machine learning model. **This model had a RMSE value of 0.34 and a R-squared value of 0.9019.**

3

Overall Takeaway

A large R-squared value is favored as that implies that the model fits the situation well 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; less variance. **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

Project Goal

This project aimed to develop a regression-based machine learning model that could predict the pIC50 of drugs while being accurate, fast, and cost-effective. This project focused on SARS Coronavirus 3C-like proteinase but is applicable to any target protein

Model Description

The model 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.

Overall Results

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.

Constraints

The results could have been improved with the inclusion of more data about SARS Coronavirus 3C-like proteinase. Unfortunately, there is not a lot of data on this target protein as it is fairly new compared to other viruses and diseases.

Usage & Overall Conclusion

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

Accuracy Augmentation

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.

Usability

To improve the usability of these models, web applications can be formed to streamline the process of entering data and outputting a prediction. This allows the project to be accessible to more individuals as well

GAN Utilization

GAN (Generative Adversial Networks) is a machine learning framework that allows for the creation/generation of novel entities. The data used in this project in along with the GAN framework can create a platform for the creation of new drugs and chemicals. The molecules can then be analyzed by this model to determine effectiveness. This creates a positive beneficial feedback loop/process for the creation and analysis of new drugs, ultimately creating a very positive impact on the industry.

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