Using Linear Regression to Detect the Binding Efficiency of Ligands for Effective p53-MDM2 Inhibition

Hoshita Undella

Abstract

Did you know that each day about 100-130 billion potentially cancerous cells undergo cell death? Every day, each person's body prevents tumors from growing, and it's all thanks to the p53 protein suppressor. Coded by the TP53 gene, the p53 tumor suppressor protein is responsible for signaling the cell repair and/or cell apoptosis pathway when damaged DNA growth is detected. In over 50% of human cells, the code for the p53 protein is corrupted, reducing its ability to inhibit tumor growth. When tumorous cells are present, there can be an overexpression of the MDM2 protein. The MDM2 protein is a protein that suppresses the activity of p53 making it even harder for controlling the potential development of cancer.

This research project targets the interaction between the MDM2 and p53 proteins to find out the most efficient ligands, or small molecules, that can bind to MDM2 and prevent the inhibition of p53 so as to stimulate the opportunity for p53 to signal for cell repair/death. This was accomplished by pre-processing a dataset of known MDM2-p53 inhibition ligands from chEMBL using numpy/pandas and adding necessary features into the dataset from rdkit. The processed data was then ran through a sklearn regression model to accurately predict the efficiency of the ligands or the negative log of their Standard Value (binding efficiency).

Overall, the model was able to perform quite well after optimization; however, it would produce even more accurate results if the images consisting of each ligand's molecular structure could be processed as well since it would allow for a greater variety of features to draw patterns between.

For further research, this project will be expanded into predicting new ligands through given features based on the efficiency of known ligands with specific features. This, if accurate, would aid in discovering new ligands for cancer therapy and drug administration.

I. Introduction

The research question is how can Linear Regression be used to accurately predict the binding efficiency of ligands for the MDM2-p53 interaction. Each day about 100-130 billion potentially cancerous cells undergo cell death. Every person has P53 is a tumor suppressor protein that signals for cell repair or cell apoptosis when DNA damage within the cell is detected. This allows p53 to keep over 50% of cancers under control by preventing damaged DNA growth. However, to regulate the activity of p53 so as to not induce accidental cell death, MDM2 is a protein that binds to p53. When

this happens, p53 activity is inhibited, and this is especially bad in tumors that overexpress MDM2. The objective of the research project is to prevent the inhibition of p53 by MDM2 through introducing an effective small molecule called a ligand. The model that has been worked with is a regression model due to the need for calculating the negative aLogP value of the given ligand in order to determine its binding efficiency. The aLogP value is a number for how likely something is to bind to a molecule but it is often in smaller decimal numbers, so taking the negative log of it allows us to gauge the actual percentage. This can help us distinguish between 10% efficient versus 25% efficient. There are four levels of categorization: highly effective (70% or higher), moderately effective (between 40% and 70%), low effectiveness (between 10% and 40%), and ineffective (below 10%).

II. Background

There are very few research projects for detecting the efficiency of ligands specifically through machine learning models. However, some articles discuss detection through a lab setup, bridging to light specific features to use in this project's data such as RO5 violations. RO5 violations are values that tell us how orally active an administered drug will be. The higher the violations, the less active and thus the less likely to bind as well. Another study that was researched tested the effectiveness of already tested ligands on mice under a lab setup by recording the cytotoxicity and radioactivity experienced by the mice. The radiation protection and cytotoxicity are the two things that were being evaluated and run through regression equations in order to figure out which compounds will properly help functioning of p53 inhibitors. The max cell death - min cell death calculation had to be above the 20% threshold in order to be considered as strongly toxic while the Candidate compounds were applied in varying levels to test the cell death rate (higher the death means lower the radiation protection). Overall, they tried to predict ligands with high radiation protection and low toxicity. A drawback with that study in relation to this project is that no ready access to a lab setup was nor permissions to test on animals were acquirable. Additionally, datasets regarding the cytotoxicity and radioactive effects of specific ligands are unavailable unless they are derived from direct experimentation. The last study talked about the measures of efficiency within a ligand and how each one was specifically derived. The mathematical explanations behind the results of the study allowed for a better understanding of what "y value" to use to accurately record ligand efficiency. The efficiencies talked about included BEI, SEI, LE, LEE, and AlogP with the last one being the easiest to find the difference in binding efficiency between two ligands.

III. Methodology

Dataset

The dataset was derived from ChEMBL, which is an open database of bioactive molecules listed with different drug-like properties. The dataset acquired without any processing consisted of 1540 known ligand

molecules who's efficiency was calculated (BEI, SEI, LE, LEE) along with the derivation of 45 different properties: 'Molecule ChEMBL ID', 'Molecule Name', 'Molecule Max Phase', 'Molecular Weight', '#RO5 Violations', 'AlogP', 'Compound Key', 'Smiles', 'Standard Type', 'Standard Relation', 'Standard Value', 'Standard Units', 'pChEMBL Value', 'Data Validity Comment', 'Comment', 'Uo Units', 'Potential Duplicate', 'Assay ChEMBL ID', 'Assay Description', 'Assay Type', 'BAO Format ID', 'BAO Label', 'Assay Organism', 'Assay Tissue ChEMBL ID', 'Assay Tissue Name', 'Assay Cell Type', 'Assay Subcellular Fraction', 'Assay Parameters', 'Assay Variant Accession', 'Assay Variant Mutation', 'Target ChEMBL ID', 'Target Name', 'Target Organism', 'Target Type', 'Document ChEMBL ID', 'Source ID', 'Source Description', 'Document Journal', 'Document Year', 'Cell ChEMBL ID', 'Properties'.

Pre-Processing

Firstly, NumPy, and Pandas were imported in order to process the dataset. From here, data preprocessing allowed the dataset to be narrowed down to 1250 molecules by 26 properties. Firstly, the following data columns about the molecules from the original dataset were kept: Molecular Weight, #RO5 Violations, Smiles, Standard Value. I further narrowed molecular weight into molecules that were between 100 and 600 since that is the optimal weight for ligands. With this, the number of molecules reduced from 1540 to 1250. The significance of keeping the RO5 violations column is that the lower the number of violations, the easier the ligand is to permeate into the cell and bind with the targeted proteins. Smiles strings are the encoded strings unique for each molecule which were later processed using rdkit. The Standard Value is the column that is to be predicted in order to gauge the efficiency of ligands in prohibiting the p53-MDM2 interaction. The Standard Value is the LogP value of a ligand which reveals how easily it will bind to a given protein. By predicting the negative log of this value, it gives a readable percentage such as 10% bindable. After the original dataset was narrowed down, all the nAn values were replaced with None and those were replaced with a 0.0. This made sure that all values in the dataset were floats, creating uniformity for feeding into a regression model. The only remaining non-float values were the Smiles strings. Using rdkit, a library containing chemical information for a variety of molecules based on smiles strings, the columns containing smiles strings were processed into various features. An additional 22 columns were added into the dataset through rdkit, with each converted into a float value to preserve uniformity. The features added were: 'HeavyAtomMolWt', 'FpDensityMorgan1', 'FpDensityMorgan2', 'FpDensityMorgan3', 'FMaxAbsPartialCharge', 'MaxPartialCharge', 'MinAbsPartialCharge', 'MinPartialCharge', 'NumRadicalElectrons', 'NumValenceElectrons', 'HeavyAtomCount', 'NHOHCount', 'NOCount', 'NumHDonors', 'NumHeteroatoms', 'NumRotatableBonds', 'NumAmideBonds', 'NumRings', 'RingCount', 'FractionCSP3', 'NumSpiroAtoms', 'NumBridgeheadAtoms'. The Fp Density values allow for evaluating

how easily a smaller ligand molecule would bind to MDM2 to inhibit the process. The other values such as partial charge, different types of electrons, bonds, and various atoms allowed for the model to compare these values between all 1250 molecules in order to group the ones with similar values. As a result, those with similar values in those groups and high Fp density, less RO5 violations, and high Standard Value would prevail in the model as highest efficiency. After all of these columns were filled in with values, they were encoded through Label Encoder to further keep uniformity,

Models

After pre-processing, the objective was to predict a float value, meaning that this project is a Regression problem (not immediately classifying into separate categories). So, Sklearn's Linear Regression Model was imported, which means that the model's data will be fit by drawing correlations between the depend and independent variables. It will predict the Standard Value/Binding Efficiency (dependent variable) of a molecule based on the relation to the other independent variables (Weight, Bonds, FPDensity, RO5 Violations, etc). This model was chosen because there are many variables that the binding efficiency of a molecule depends on, and it is imperative to get an accurate efficiency reading so as to make sure the ligand will be able to inhibit MDM2 and p53 interaction in case of including it within drug administration. All of the columns except the Standard Value were fed into the X values and the Standard Value into the Y value. Test and Train sets were split into 20 and 80 respectively and the model was fit accordingly. To test the models performance, sklearn's metrics were imported and the mean_squared_error, r2_score, as well as explained_variance_score were printed. The end goal of the model regarding efficiency was to get the last two metrics as close to 1 as possible while the first one is as close to 0 as possible. In order to achieve this end goal, additional models were run on the data such as Random Forest Generator and XG Boost. For Random Forest Generator, the following parameters were tuned to the corresponding values: max_depth=10, random_state=0. Max_depth relates to the number of splits or subroots that a decision tree can have. For XGBoost, the following parameters were tuned to the corresponding values: n_estimators=88, reg_lambda=25, gamma=0, max_depth=10. N_estimators is the number of trees the model will use to learn, and a recommended number is between 50-300. Meanwhile, reg_lambda is a way to decrease the size of the weights which will allow the model to learn at a slower rate, preventing loss of information during transfer within trees. After optimizing through all these models through manually backpropagating based on metrics outputs, the predictions will be categorized into four categories: highly effective (70% or higher), moderately effective (between 40% and 70%), low effectiveness (between 10% and 40%), and ineffective (below 10%).

IV. Results and Discussion

To generate results for the models in order to gauge the differences between each model, Seaborn was imported for the scatterplot functionality.

Figure 1 below shows the results of the Linear Regression model on the processed dataset. The metrics for this model are as follows:

Mean_absolute_error	0.8278395956864693
Mean_squared_error	1.0807838305000221
R2_score	0.696487957657997
Explained_variance_score	0.6965297674173426

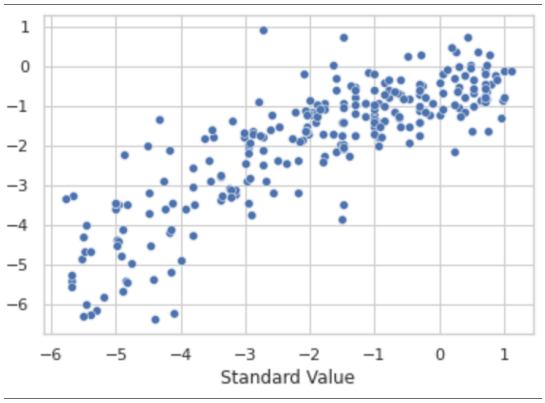


Figure 2 below shows the results after employing Random Forest Generator.

Mean_absolute_error	0.6242325020094242
Mean_squared_error	0.6523346733009341
R2_score	0.8168075581844433
Explained_variance_score	0.817875755152205

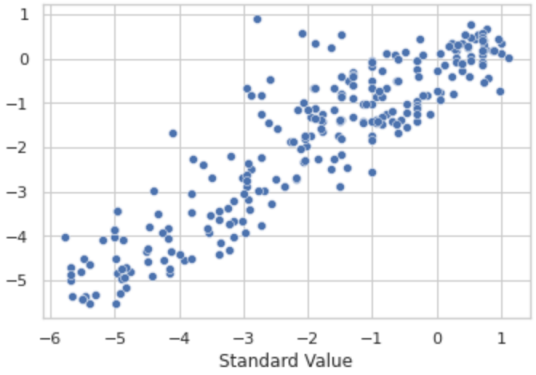
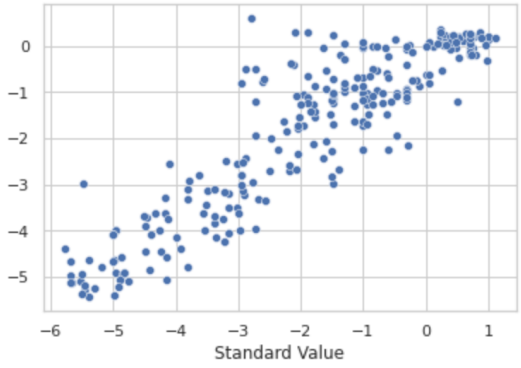


Figure 3 below shows the results after employing XGBoost.

Mean_absolute_error	0.6169271437924931
Mean_squared_error	0.6577549251455131
R2_score	0.8152854113305298
Explained_variance_score	0.816908294214389



Comparing the three scatterplots, it can be seen that the line of best fits becomes clearer and more defined as more hyperparameters are tuned and weights are added. In the end, the end goals were achieved in that the r2_score and explained_variance_score showed high correlation between the variables in the model while the mean_absolute_error and mean_squared_error showed that on average, the square root of the predicted

values are 0.65 away from the actual value. These values could be optimized further by adding more data to allow the model to correlate between more features, making it more accurate.

V. Conclusion

The research question was how Linear Regression could be used to accurately predict the efficiency of different ligands for effective p53-MDM2 inhibition. A database from chEMBL consisting of 1540 ligand molecules was pre processed and more columns of molecule features from rdkit were added. In order to output predictions, data was split into 80% training and 20% testing; This dataset was then run through a Linear Regression model from sklearn and the efficiency of the model was calculated through the mean_squared_error, r2_score, and explained_variance_score. The goal was to acquire a mean_squared_error of less than 0.7 as well as an r2_score and an explained_variance_score of over 0.8. At its most optimized state, the model was able to achieve a mean_absolute_error of 0.6169271437924931, mean_squared_error of 0.6577549251455131, an r2_score of 0.8168075581844433, and an explained_variance_score of 0.817875755152205. The model performed well due to the optimization of the hyperparameters and optimization of the metrics by adding weights. In order to predict with a wider variety of features, further steps such as analyzing the molecular structures of ligands using 3D and 2D models would be helpful. This would also allow for a wider derivation of features from rdkit, meaning that more independent variables would be considered within the model, ultimately making it more accurate. More models such as Elastic Net Regression can also be experimented with to see if better results can be outputted. Lastly, This project can be expanded into predicting new ligands through given features based on the efficiency of known ligands with specific features. This, if accurate, would aid in the discovery of new ligands for cancer therapy and drug administration.

VI. Acknowledgments

I am thankful for my parents who have constantly supported me through all my endeavors. I would also like to thank my mentor, Ayush Pandit, for helping me narrow down my project, constantly guiding me through it, and readily providing me with as many resources as he could possibly find.

VII. References

1) https://www.ebi.ac.uk/chembl/g/#browse/activities/filter/target_chembl_id%3ACHEMBL1907611%20AND%20standard_type%3A(%22IC50%22
2) https://pubmed.ncbi.nlm.nih.gov/27269808/
3) https://www.thermofisher.com/us/en/home/life-science/antibodies/antibodies-learning-center/antibodies-resource-library/cell-signaling-pathways/p53-mediated-apoptosis-pathway.html
4) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2676446/#:~:text=Because%20the%20interaction%20between

%20MDM2,the%20focus%20of%20this%20review.
5) https://inspirit11men-pal1838.slack.com/files/U03TJ2Y1RPV/F042FE0BRNK/predicting_radiation_protection_and_toxicity_of_p53_targeting_radioprotectors_using_machine_learning.pdf
6) https://pubmed.ncbi.nlm.nih.gov/25407396/
7) https://www.youtube.com/watch?v=2RG9caushI0
8) https://www.youtube.com/watch?v=81NCnoRlbGI
9) https://www.youtube.com/watch?v=w5scJuhoGs8
10) https://www.youtube.com/watch?v=nd_SMrd1oag