Prediction of anti-cancerous activity of Chalcones in colon cancer using machine learning approach

Gaurav kumar singh
IIT Hyderabad
Hyderabad, Telangana
bm22mtech11002@iith.ac.in

Abstract

Cancer is a complex disease and the second most common cause of death in the US, with no proper treatment and drug. Hence, shortlisting potential drugs candidate for cancer is a major area of research in current time. Predicting the IC50 value of compounds can play an important role in determining the potency of the drug, as lower the IC50 value the potency of drug is higher. The experimental prediction of IC50 value is cumbersome, costly and time-consuming process. Hence, machine learning approach can be used to overcome these shortcomings of experimental prediction. In this project machine learning approach is used for prediction of anti-cancerous activity of Chalcones in colon & Colo-rectal cancer where prediction of IC50 value is carried out using molecular descriptors, The XGBoost regression model is used to predict this value and found to give highest accuracy (95.98%) among other models like RF, LR, SVM, AdaBoost which gave 55.42%, 65%, 45% and 95.48% accuracy respectively.

1. Introduction

Cancer is an umbrella term for group of diseases specified by the uncontrolled growth and proliferation of abnormal cells. It can occur in any tissue or organ of the body and may takeover nearby tissues and organs or metastasize to other parts of the body via bloodstream or lymphatic system. Cancer starts off by genetic mutations or DNA alterations which is inside the nucleus of the cells. These mutations can be inherited from a person's parents or can be acquired during a person's lifetime due to various factors such as exposure to carcinogens (substances that can cause cancer), lifestyle habits, viral infection, radiation exposure, hormonal imbalance, and other unknown factors. The exact cause of many cancers is still not understood completely, and research is ongoing (CB Blackadar, 2016) over it. There are over 100 different types of cancer, which are typically named after the type of cells or the organs, from which they originate. Some commonly investigated types of cancer include breast cancer, lung cancer, colorectal cancer and many more. Each type of cancer has its own unique characteristics, behaviour, and the treatment options (Steven A. Frank, 2010).

Chalcone is considered as a privileged structure in Medicinal Chemistry. They belong to a class of natural flavonoid family, which are widely distributed in the plant kingdom. Their distinctive chemical entities, which includes a central three-carbon, -unsaturated ketone bridge that gives them their distinctive yellow colour, distinguishes them from other molecules. Chalcones are established to exhibit a wide range of biological activities, including anti-inflammatory, antioxidant, anti-cancer, anti-microbial, anti-viral, anti-diabetic, and anti-obesity properties, among others.

Colon cancer being the third most common cancer diagnosed and prevalent in both men and women worldwide (Baojun Duan, 2022) and the target found for this cancer was HCT116 cells.

Chalcones have chemical formula C15H12O2 and a basic structure of three aromatic rings connected by an α,β-unsaturated ketone (also known as a chalcone) in the middle. The three aromatic rings are usually referred to as ring A, ring B, and ring C, and they can be further substituted with various functional groups, resulting in a wide range of chalcone derivatives with diverse chemical properties and biological activities.

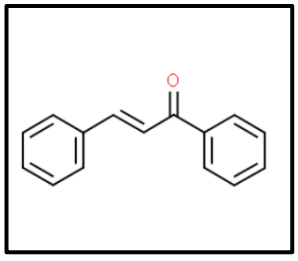


Fig1. Chemical structure of chalcones.
(Source – Wikipedia.com)

Chalcones are commonly found in various plant sources, including fruits, vegetables, herbs, and spices. Some of the common dietary sources include apples, grapes, soybeans, green tea. Chalcones are also found in many medicinal plants used in traditional medicine

systems (Chunlin Zhuang, 2017).

Since chalcones is found to have anti-cancerous properties, chalcones have been extensively studied for their potential role as anticancer drugs. They possess several properties that make them attractive candidates for cancer therapy, including their ability to inhibit cancer cell proliferation, induce cell cycle arrest, promote apoptosis (programmed cell death). Furthermore, chalcones had been shown to sensitize cancer cells to chemotherapy and radiation therapy, making them potential adjuvants in combination cancer therapy (Klaus Gundertofte, 2000). They can enhance the anticancer effects of conventional chemotherapy drugs and reduce drug resistance in cancer cells, thus improving the overall therapeutic outcome. Moreover, chalcones have been found promising in overcoming multidrug resistance, a common challenge in cancer treatment, as they can inhibit drug efflux pumps in cancer cells, which are responsible for pumping out chemotherapy drugs and reducing their intracellular accumulation. This can enhance the efficacy of chemotherapy drugs and overcome drug resistance in cancer cells (P Larsson, 2020).

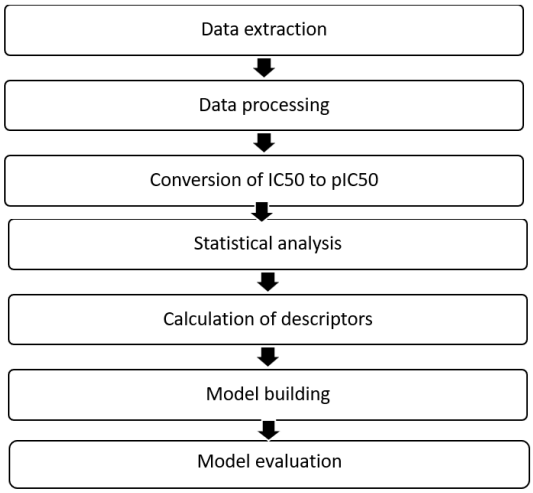
The process of selecting drugs to treat various types of cancer and improving therapeutic efficacy has emerged as a significant challenge in the development and research o f cancer drugs. Currently, the method of assessment and d eveloping new drugs through experiment is quite slow an d demands a lot of manpower and material assets.

To predict the potency of drugs, IC_{50} value which stands for Half-maximal inhibitory concentration (IC_{50} gives the measure of potency of drug) plays an important role (Yifeng He, 2016), since lower the value of IC_{50} higher the potency of drug and vice-versa. Thus, to determine IC_{50} we must perform investigations depending on experimental conditions, such as the specific assay used, the cell type or tissue used, and the experimental protocol. Therefore, it is essential to carefully design and execute experiments to ensure accurate and reproducible IC_{50} values (Markossian S, et.al). This problem can be overcome by evaluating the efficacy of drug computationally by developing a tool with the help of machine learning. With the help of this, we can overcome the challenges of conventional drug design and save our time, money, and effort to determine the IC_{50} value. Based on these values, we can go for drug design (Vinita Periwal, 2022). In recent years, numerous web servers have came up to develop and apply Quantitative Structure Activity Relationship (QSAR) models (Capuzzi, et al., in 2017).

To make it more time and cost efficient, the method of predicting bioactivity (IC_{50}) of compounds using machine learning models can be used to monitor and identify potentially active compounds (Klaus Gundertofte, 2000). In this project the prediction of IC_{50} value of chalcones against colon and colon cancer cell line is carried out.

2. Materials and methods

All The following steps are followed in this project-



2.1 Data Extraction:

All the chalcones data are taken from ChalconeDB. ChalconeDB is a database of chalcone which consists of only chalcone compounds around 2,52,653 numbers of compounds are present in the database. The Compounds which don't have IC_{50} values are removed, out of all these compounds only approximately only 35000 compounds have IC_{50} value of all the assays is extracted with the help of the code (which contains requests library of python). Then only MTT assays are selected (used to check cell viability) and with the help of these assays different cancer cell lines are generated. Different cancers identified such as skin, breast, blood, lung, liver, colon, kidney etc. These different cancers have different targets, so targeting particular cancer cell line we are taking into consideration specific targets for specific cell line.

Colon cancer is the third most common cancer diagnosed in both men and women in the worldwide. The data collected for colorectal cancer is approximately 3350 molecules, for these molecules the target is HCT116 cells.

2.2 Data Processing:

The SMILES (Simplified Molecular Input Line Entry System) either through PubChem identifier Exchange or through code. SMILES notation provides a compact and standardized way to represent the chemical structure of molecules. It uses a simple string of characters that represents the atoms, bonds, and connectivity of a molecule, allowing it to be easily stored, exchanged, and processed in databases, software tools, and algorithms. It is used in cheminformatics and computational chemistry for data analysis, visualization, and machine learning tasks. It allows for the extraction of molecular features, generation of chemical fingerprints, and development of predictive models for various chemical properties, such as toxicity,

bioactivity, and physicochemical properties.

Then Lipinski descriptors are calculated with the help of RDkit. RDKit is an open-source software toolkit for cheminformatics and molecular modelling. It provides a wide range of tools and functions for handling and manipulating chemical structures, as well as performing various computational chemistry tasks, it is widely used in the field of cheminformatics and drug discovery for a variety of applications, including drug design, virtual screening, molecular modelling, and chemical data analysis. With the help of SMILES, Lipinski descriptors are calculated. These Lipinski descriptors also known as the "Lipinski's rule of five" are a set of four empirical guidelines used in drug discovery to assess the likelihood of a small molecule drug candidate to have favourable pharmacokinetic properties. The four Lipinski descriptors are as follows (i) Molecular weight (MW): The molecular weight of the compound should be less than 500 Da. (ii) Lipophilicity (logP): The calculated octanol/water partition coefficient (logP) of the compound should be less than 5. (iii) Hydrogen bond donors (HBD): The compound should not have more than 5 hydrogen bond donors. (iv)Hydrogen bond acceptors (HBA): The compound should not have more than 10 hydrogen bond acceptors.

2.3 Conversion of IC_{50} to pIC_{50}:

pIC_{50} is a negative logarithm of IC_{50}. It is a transformed version of IC_{50} values that is often used to simplify data analysis and comparison of compound potencies. The conversion of IC_{50} to pIC_{50} involves taking the negative logarithm (base 10) of the IC_{50} value, which results in a unitless value that represents the potency of a compound on a logarithmic scale. According to Collaborative Drug Design (CDD), this type of conversion is important because (i) Linearization of potency data: Converting IC_{50} values to pIC_{50} values linearizes the potency data, making it easier to compare and analyse the relative potencies of different compounds. (ii) Simplification of data analysis: Using pIC_{50} values simplifies data analysis, as it reduces the need to work with large numbers and allows for more straightforward statistical calculations. (iii) Facilitation of structure-activity relationship (SAR) analysis: Converting IC_{50} to pIC_{50} facilitates the analysis of structure-activity relationships (SAR), which involves studying the relationship between the chemical structure of a compound and its biological activity. (iv) Standardization of potency data: Converting IC_{50} to pIC_{50} allows for the standardization of potency data, as pIC_{50} values are unitless and independent of the concentration units used for IC_{50} values. This makes it easier to compare and integrate potency data from different sources and studies.

The formula for conversion of IC_{50} to pIC_{50} is as follows:

$$IC_{50} = 10(x - pIC_{50}) \quad \text{and} \quad pIC_{50} = x - \log10\,(IC_{50})$$

Where, x = 3 for millimolar, 6 for micromolar, 9 for nanomolar, and 12 for picomolar concentrations

2.4 Statistical Analysis:

The Shapiro-Wilk test is a statistical test used for assessing the normality of a data set. It is a widely used test for checking the assumption of normality, which is a common assumption in many statistical methods and techniques that rely on the assumption of normal distribution, such as parametric tests like t-tests, ANOVA. The Shapiro-Wilk test is a numerical method that computes a test statistic based on the sample data and compares it to critical values from a known distribution (the Shapiro-Wilk distribution) to determine if the data follows a normal distribution.

If the test statistic is smaller than the critical value, the null hypothesis of normality is rejected, indicating that the data does not follow a normal distribution. If the test statistic is larger than the critical value, the null hypothesis of normality is not rejected, indicating that the data may follow a normal distribution.

2.5 Descriptors Calculation:

The descriptors are calculated with the help of PaDEL. PaDEL is a freely available, open-source software tool for calculating molecular descriptors and fingerprints from chemical structures. It is widely used in cheminformatics and drug discovery research for predicting various properties and activities of chemical compounds. It offers a wide range of molecular descriptors and fingerprints, including topological, physicochemical, constitutional, and electronic descriptors, as well as popular fingerprints such as Extended Connectivity Fingerprints (ECFP), PubChem fingerprints and MACCS fingerprints. These descriptors are used to describe small compounds' biological activity, pharmacokinetics, and toxicity in the drug discovery and development process. These descriptors is frequently used for virtual screening, lead optimization, and drug creation in both the pharmaceutical business and academic research.

PubChem fingerprints are generated by the PubChem database for chemical compounds that is a set of binary fingerprints. PubChem fingerprints are used to represent the structural features and properties of chemical compounds in a binary format, which can be used for similarity searching, virtual screening, and other cheminformatics applications. They are generated using a set of predefined rules that encode different molecular features and properties, such as atom types, bond types, ring systems, substructures, and functional groups. These rules are applied to the chemical structure of a compound, resulting in a binary fingerprint that represents the presence or absence of each predefined feature or property. Each fingerprint bit is assigned a value of 1 if the corresponding feature or property is present in the compound, and 0 if it is absent.

Around 900 different types of descriptors are calculated

on the basis of their structure, and then low variance data is removed so that there should be no bias during the prediction. The threshold for removing the low variance data is 0.8 means at least there should be 20% variance between the molecular descriptors. So out of 881 only 228 descriptors were selected, on these 228 descriptors the model was build.

2.6 Model Building:

In this we are going to predict the pIC_{50} hence, regression model is used. Different types of Machine learning model is used such as Random forest, Linear regression, Support vector Machine, XGBoost, AdaBoost.

2.6.1 Random Forest:

Random Forest is an popular algorithm for machine learning that is used for both classification and regression tasks. It combines multiple decision trees to create a robust predictive model that is more accurate that is called an ensemble learning method. In a Random Forest, a collection of decision trees is built on random subsets of the training data, and the predictions from these trees are combined to make the final prediction. The predictions from the individual decision trees in the Random Forest are combined using various methods, such as averaging (for regression tasks) or voting (for classification tasks), to obtain the final prediction. The Random Forest algorithm is known for its ability to handle high-dimensional data, handle missing values, and reduce the risk of overfitting as compared to individual decision trees.

2.6.2 Linear Regression:

Linear regression is a statistical technique that is used to model the relationship between one or more independent variables and a dependent variable. It is a widely used and a simple method for predicting the value of a variable that is dependent variable based on the values of one or more variables that are independent variable. The basic idea of linear regression is to fit a straight line (or a hyperplane in case of more number of independent variables) through a set of data points in such a way that it minimizes the differences of sum of the squared between the observed values of the dependent variable and the predicted values (i.e., the residuals). It is a simple and interpretable method that can provide insights into the relationship between variables and can be used for prediction, inference, and understanding the underlying patterns in data.

2.6.3 Support Vector Machine:

Support Vector Machine (SVM) regression is a type of machine learning algorithm that is used for regression tasks. It is an extension of the original SVM algorithm, which was originally developed for binary classification tasks. In SVM regression, the aim is to find a function that can model the relationship between input features and target output values, which could be continuous or numeric. SVM regression targets to evaluate a hyperplane that well fits the data, while also maximizing the margin among the data points and the hyperplane. Its regression is known for its ability to model non-linear relationships in data using kernel functions, which allows it to capture complex patterns in the data. It is also effective in handling high-dimensional data and dealing with outliers, since its main goal is to identify the optimal hyperplane with the greatest margin among the data points and the hyperplane. However, SVM regression can be computationally expensive, especially with large datasets, and it requires careful hyperparameter tuning to achieve optimal performance.

2.6.4 AdaBoost Regression:

AdaBoost is a well-known algorithm of machine learning that can be implied for both regression tasks and classification. It uses an ensemble learning technique that combines multiple weak learners (typically decision trees) into a powerful predictive model. The principle of AdaBoost is to sequentially train a series of weak learners on the same dataset, with each weak learner giving more importance to the misclassified samples from the previous learners. This allows AdaBoost to focus on the samples that are difficult to classify and improve the overall prediction accuracy of model. The key steps in the AdaBoost algorithm are as follows: (i) Initialize sample weights: Each sample in the training dataset is assigned an initial weight, which is typically set to a uniform value. (ii) Train weak learner: A weak learner, such as a decision tree, is trained on the training dataset with the current sample weights. The weak learner produces a prediction for each sample. (iii) Update sample weights: The weights of sample are updated on the basis of misclassifications made by the weak learner. Misclassified samples are given higher weights, while correctly classified samples are given lower weights. (iv) Repeat: Steps 2 and 3 are repeated for a specified number of iterations or until a criterion to stop is met.

2.6.5 XGBoost Regression:

XGBoost (eXtreme Gradient Boosting) is a well-known algorithm of machine learning that is used for both regression tasks and classification. Unlike AdaBoost, it uses an ensemble learning method that uses gradient boosting to combine multiple weak learners (typically decision trees) into a strong predictive model. The main features of XGBoost that make it popular are its ability to handle large and complex datasets, its efficiency in terms of both computation and memory usage, and its high predictive accuracy. Gradient boosting: XGBoost uses gradient boosting, which is an iterative technique that sequentially adds new trees to the ensemble to correct the errors made by previous trees. This helps improve the

accuracy of predictive capability of the model by combining the predictions of multiple trees in a weighted manner. (i) Regularization techniques: XGBoost includes built-in support for various regularization techniques, such as L1 (Lasso) and L2 (Ridge) regularization, to limit overfitting and improve model generalization. (ii) Parallelization: XGBoost is designed to be highly efficient in terms of computation and memory usage, and it includes support for parallel processing, which allows for faster training of large datasets. (iii) Feature importance: XGBoost provides feature importance scores that help identify the most important features for making accurate predictions, which can be used for feature selection or feature engineering. (iv) Tree pruning: XGBoost includes techniques for pruning trees during the boosting process, which helps prevent overfitting and improves model generalization.

After using different types of models, model selection is done with the help of different parameters such as Accuracy, Mean Square Error, R^2.

Accuracy: The overall correctness of a classification model is measured to evaluate accuracy of the model. It is defined as the ratio of Number of predictions done correctly to the Number of total predictions, presented as a percentage. The formula for accuracy is:

Accuracy = (Number of predictions done correctly) / (Number of total predictions) * 100

Mean Squared Error (MSE): it is a measure of the average squared difference between actual values and predicted value in a regression model. It is evaluated by taking the average of the squared differences between predicted values and actual values. The formula for MSE is:

MSE = (1/n) * Σ[(predicted value - actual value)^2]

where n is the number of samples in the dataset.

MSE is typically used as a loss function in regression tasks, where the aim is to predict continuous numeric values. It gives higher weight to larger errors, making it sensitive to outliers.

R-squared (R^2): R-squared is a measure of how well a regression model fits the data, expressed as a value between 0 and 1. It indicates the proportion of the variance in the dependent variable (i.e., the target variable) that is defined by the features (i.e., independent variables) in the model. R-squared value of Zero shows that the model does not validates any variance, while an R-squared value of 1 shows that the model validates all the variance in the data. R^2 is a widely used metric for assessing the goodness-of-fit of a regression model. Higher R^2 values indicate a better fit, while lower R^2 values indicate a poorer fit.

2.7 Model Deployment:

Model deployment is known as the process of integrating a model of machine learning into a production environment, making it available for use in real-world applications. After a machine learning model has been trained and evaluated, the next step is to deploy it so that it can be used for making predictions or generating insights in a live or operational environment.

Deploying a machine learning model with Streamlit involves creating a Streamlit app as a web-based user interface for the model, loading the model in the app, deploying the app to a web server or cloud-based hosting platform, monitoring and maintaining the app, and testing and validating the app for accuracy and reliability. Streamlit provides a simple and efficient way to deploy machine learning models as interactive web applications, making them accessible to users for real-world use.

3. Results & Discussion

When the Lipinski's descriptors are calculated, there was no correlation or very less correlation between the four descriptors they were independent of each other. The matrix is given bellow in the table.

	cid	MW	LogP	numHD	numHA	pIC_{50}
cid	1.00	-0.04	-0.05	-0.01	-0.001	0.107
MW	-0.04	1.00	0.43	0.44	0.69	-0.05
LogP	-0.53	0.43	1.00	-0.33	-0.65	0.02
numHD	-0.014	0.44	-0.33	1.00	0.39	-0.06
numHA	-0.0011	0.48	-0.06	0.39	1.00	-0.70
pIC_{50}	0.0107	-0.05	0.02	0.068	-0.07	1.00

As mentioned in the above table, we can see there is no correlation or very less correlation between the descriptors and pIC_{50}.

3.1 Statistical Analysis:

The Shapiro-Wilk test is a statistical test used to assess whether a given dataset follows a normal distribution. The test produces a test statistic and a p-value, which can be used to make a decision about the normality of the data.

Statistics=0.958, p=0.000 (for pIC_{50})

In this case, the test statistic is 0.958 and the p-value is 0.000. The null hypothesis (H0) in the Shapiro-Wilk test is that the data follows a normal distribution. Since the p-value is less than the significance level (commonly set at 0.05), which is typically used for hypothesis testing, it

suggests that there is sufficient evidence to reject the null hypothesis. This means that the sample which have been tested does not appear to be normally distributed based on the Shapiro-Wilk test.

3.2 Model Evaluation:

For colon cancer cell lines different models were used and check the accuracy of the models. The accuracy for models Random Forest, Linear regression, Support vector Machine, XGBoost, AdaBoost were evaluated as 66%, 55%, 45%. 94.72% and 94.3% respectively.

Cancer	RF	LR	SVM	XGBoost	AdaBoost
Colon cancer	66	55	45	94.72	94.3

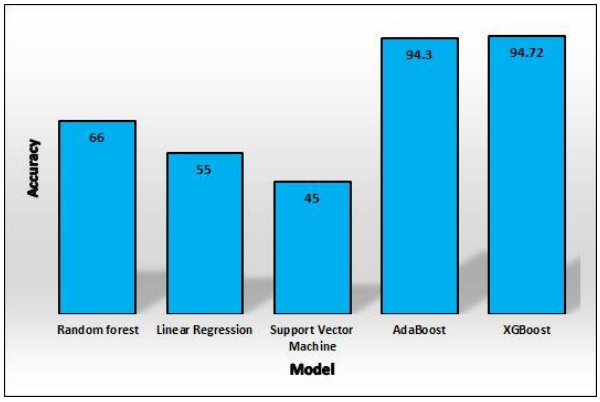


Fig- Accuracy of different models.

Since, XGBoost perform well in Colon cancer and Colo-rectal cancer cell lines model and the accuracy is more as compared to different models. So XGBoost model is selected for further evaluations such as Mean Square Error, R^2. The Mean Square Error, R^2 for colon cancer is 0.145 and 0.72 respectively while mean absolute error is 0.27 and Root-mean-square deviation is 0.36.

Validation Type	Score
Accuracy	94.72
R2 Value	0.72
MSE	0.145
MAE	0.27
RMSE	0.36

Lower MSE (Mean squared error) values indicate better model performance Whereas higher MSE value indicates higher error rate which indicates low performance of model. Lower MAE (Mean absolute error) values indicate better model performance. Lower RMSE (Root-mean-square deviation) values indicate better model performance. Higher R^2 value indicates better performance of model and lower the R^2 value indicates less correlation between the model that means randomness is more.

4. Conclusion:

Drug potency: The predicted IC50 values can provide insights into the potency of the tested compounds or drugs against Colon cancer cell lines. Lower IC50 values indicate higher potency, meaning that lower concentrations of the drug are needed to inhibit 50% of cell growth or viability. This information can be used to prioritize or select compounds with higher potency for further development or testing.

Sensitivity or resistance: The predicted IC50 values can also reveal the sensitivity or resistance of Colon & Colo-rectal cancer cell lines to the tested compounds or drugs. Cell lines with lower predicted IC50 values may be more sensitive to the compounds, indicating potential efficacy against those cancer types. On the other hand, cell lines with higher predicted IC50 values may be more resistant to the compounds, suggesting potential challenges in using the drugs against those cancer types.

Comparative analysis: The predicted IC50 values can be used for comparative analysis between Colon & Colo-rectal cancer cell line, compounds, or drug candidates. This can help identify patterns, trends, or correlations in the data, such as identifying cancer types that are sensitive to certain compounds or identifying compounds that consistently show higher or lower potency across different cell lines.

Further experimentation: The predicted IC50 values can serve as a basis for designing further experiments, such as in vitro or in vivo validation studies, to confirm the predicted potency and efficacy of the compounds or drugs. Experimental validation is crucial to confirm the reliability and accuracy of the prediction model or method used, and to provide more robust evidence for the conclusions drawn from the predicted IC50 values.

In conclusion, the prediction of IC50 values for Colon cancer and Colo-rectal cancer cell lines can provide valuable insights into the potency, sensitivity, or resistance of compounds or drugs, and can inform decision-making in colon cancer and Colo-rectal cancer research and drug discovery. However, it is important to interpret the results in the context of the specific prediction model or method used, and to validate the predictions through further experimentation before drawing definitive conclusions.

References

[1] Ávila HP, Smânia ED, Delle Monache F, Júnior AS. Structure–activity relationship of antibacterial chalcones. Bioorganic & medicinal chemistry. 2008 Nov 15;16(22):9790-4.

[2] Berrouet C, Dorilas N, Rejniak KA, Tuncer N. Comparison of drug inhibitory effects (IC 50) in monolayer and spheroid cultures. Bulletin of mathematical biology. 2020 Jun;82(6):68.
[3] Chennamadhavuni A, Lyengar V, Shimanovsky A. Continuing Education Activity.
[4] Cooper GM, Hausman RE. The development and causes of cancer. The cell: A molecular approach. 2000;2:719-28.
[5] Ekins S, Mestres J, Testa B. In silico pharmacology for drug discovery: methods for virtual ligand screening and profiling. British journal of pharmacology. 2007 Sep;152(1):9-20.
[6] Gundertofte K, Jørgensen FS, editors. Molecular modeling and prediction of bioactivity. Springer Science & Business Media; 2012 Dec 6.
[7] He Y, Zhu Q, Chen M, Huang Q, Wang W, Li Q, Huang Y, Di W. The changing 50% inhibitory concentration (IC50) of cisplatin: a pilot study on the artifacts of the MTT assay and the precise measurement of density-dependent chemoresistance in ovarian cancer. Oncotarget. 2016 Oct 10;7(43):70803.
[8] Kausar S, Falcao AO. An automated framework for QSAR model building. Journal of cheminformatics. 2018 Dec;10(1):1-23.
[9] Muthukumaran P, Rajiniraja M. MIA-QSAR based model for bioactivity prediction of flavonoid derivatives as acetylcholinesterase inhibitors. Journal of Theoretical Biology. 2018 Dec 14;459:103-10.
[10] Oduor RO, Ojo KK, Williams GP, Bertelli F, Mills J, Maes L, Pryde DC, Parkinson T, Van Voorhis WC, Holler TP. Trypanosoma brucei glycogen synthase kinase-3, a target for anti-trypanosomal drug development: a public-private partnership to identify novel leads. PLoS neglected tropical diseases. 2011 Apr 5;5(4):e1017.
[11] Periwal V, Bassler S, Andrejev S, Gabrielli N, Patil KR, Typas A, Patil KR. Bioactivity assessment of natural compounds using machine learning models trained on target similarity between drugs. PLOS Computational Biology. 2022 Apr 25;18(4):e1010029.
[12] Thakur A, Kumar A, Sharma VK, Mehta V. PIC50: An open source tool for interconversion of PIC50 values and IC50 for efficient data representation and analysis. bioRxiv. 2022:2022-10.
[13] Yap CW. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. Journal of computational chemistry. 2011 May;32(7):1466-74.