IJBPAS, December, 2021, 10(12): 4585-4600

ISSN: 2277–4998

International Journal of Biology, Pharmacy
and Allied Sciences (IJBPAS)
'A Bridge Between Laboratory and Reader'

www.ijbpas.com

MACHINE LEARNING APPROACHES TO IDENTIFY POTENTIAL DNA Gyr B INHIBITORS FOR MYCOBACTERIUM TUBERCULOSIS

SURENDRAN S*, PUSHPA VL, MANOJ KB AND ACHUTHA AS

P G and Research Department of Chemistry, Sree Narayana College, Kollam, Kerala, India 691001

***Corresponding Author: Suchitra Surendran: E Mail: suchitrasreeraj@gmail.com**

Received 23rd Feb. 2021; Revised 24th April 2021; Accepted 10th June 2021; Available online 1st Dec. 2021

https://doi.org/10.31032/IJBPAS/2021/10.12.5774

ABSTRACT

Tuberculosis, the deadly airborne disease that has now re-emerged as Multi-drug resistant tuberculosis, poses a major health threat to mankind. It has created a global concern for researchers to find a new drug that can fight against this dreadful bacterium. The bacterial DNA Gyrase B has been identified as an important drug target for the disease. Even though high-throughput screens to identify the anti-tubercular activity of small molecules are now available, they are expensive and time-consuming. Computational methods including Machine Learning approaches, QSARmodelling, and Docking process are less time-consuming as well as efficient. This attracts us to build classifiers for virtual high-throughput screening and to select molecules from large libraries for further analysis. We developed four supervised classification models (SMO, Random Forest, Naive Bayes, and MLR) using Weka to classify the molecules as actives and inactives and four regression models to predict the inhibitory activities of these active molecules. The filtered molecules were further docked to the Gyrase B protein (PDB ID: 4B6C) to understand the interacting residues and the output poses were also rescored to understand the stability of the ligand-protein complex. Finally, we have screened 6 phytomolecules from a phytochemical database IMPAAT by applying these developed models and predicted their inhibitory activities. The pharmacokinetics and toxicity of these molecules were also studied. These molecules could be modified to increase the potency and thus can be used to develop drugs against Mycobacterium tuberculosis.

Keywords: DNA Gyrase, Machine learning,Docking,QSAR, ADMET, Virtual Screening

1. INTRODUCTION

Tuberculosis (TB) is a deadly disease triggered by Mycobacterium tuberculosis (Mtb). Despite effective short-course chemotherapy the tubercle bacillus continues to claim more lives than any other single infectious agent **[1]**. Current treatment strategies are strongly associated with side-effects including liver toxicity and also the emergence of multidrug-resistant strains. Again this does not consider the effect on the reactivation of the disease, which significantly contributes to the global incidence of TB. Therefore, new pharmaceutical agents are urgently required to control TB and reduce the increasing incidence of Multidrug resistant Tuberculosis MDR-TB and Extensively drug resistant Tuberculosis XDR-TB

The bacterial topoisomerases DNA gyrase is validated targets for antibacterial chemotherapy. It comprises two subunits gyrase A (GyrA) and gyrase B (GyrB). The role of subunit A is cleavage and recombination of double stranded DNA, while B subunit bear the ATPase activity that helps in ATP hydrolysis and furnishes a sufficient amount of energy for supercoiling **[2]**. Fluoroquinolones based drugs have found to inhibit the catalytic activity of GyrA, but the resistance developed due to their mutation in, GyrA domain makes DNAGyrB an interesting drug target against Mtb **[3]**. This has resulted in the identification of a wide range of bioactive compounds. These molecules belong to various classes such as pyrrolamides, pyrazolthiazoles, bithiazoles, aminopyrazinamides, N-linked amino-piperidine, thiazoleaminopiperidine etc. But the only drug approved by FDA for the target is an aminocaumarin, Novobiocin, this also has been withdrawn from the market for safety concerns. At present, there is no marked inhibitor for DNA Gyrase B. In search of a potential lead for the GyrB domain, different computational methods like machine learning approaches, QSAR studies, molecular docking, and virtual screening have been utilised.

Machine learning technique is a versatile tool for finding bioactive compounds against the GyrB domain more quickly **[4]**. Quantitative structure activity relationship (QSAR) modelling represents a promising approach from computational chemistry to predict the structural requirements needed for biological activity. Based on the types of predicted properties, QSAR models can be classified into qualitative prediction (classification) models and quantitative prediction (regression) models. From a series of compounds whose biological activities are known, these models can unravel the relationships between structural features and their biochemical properties and then predict the activity of new compounds based on this knowledge. So QSAR models can serve as a practical screening tool for those researchers who are synthesizing new compounds **[5]**.

The goal of the present work is to develop classification and regression models for categorizing GyrB inhibitors and predicting the activity of these inhibitors using Weka (Waikato Environment for Knowledge Analysis). Even though QSAR models were earlier developed for some of the congeneric series of DNA Gyrase B inhibitors, it was only used to predict the activity of similar molecules. In our work we have taken a large diverse dataset of 286 Gyrase B inhibitors for training our model, hence all the significant structural features that influence Gyrase B activity were captured and the model was developed. Computational studies including data pre-processing, feature selection, and QSAR model generation was done using the software Weka, version 3-7-9. In this work, a combination of QSAR and molecular docking was applied to screen out bioactive compounds that inhibit DNA Gyrase B activity. According to the results obtained by this combined computational approach the phytomolecules of Indian Medicinal Plants, Phytochemistry And Therapeutics (IMPPAT) **[6]** was screened. The pharmacokinetics and toxicity of these molecules also were analyzed to confirm the safety of these candidates.

2.MATERIAL AND METHODS

2.1. Dataset generation

The dataset of 286 GyrB inhibitors **[7-14] (Supplementary file, Figure 1 Table S1)** consisting of diverse scaffolds along with their MsmGyrB ATPase inhibitory activity was collected from the literature. The inhibitory activity (IC50 value) of the dataset was converted into pIC50 (-log IC50). The values of pIC50 ranged from 4.146 M to 9.301 M. The 3D structure of all the molecules was sketched and the optimization process was done using LigPrep **[15]** module of Schrodinger. Optimization, as well as low energy conformers, were, generated using OPLS3eforce field **[16]**. These datasets were then employed to develop classification and regression models to screen inhibitors for GyrB.

2.2. Molecular descriptors calculation

Molecular descriptors are distinct mathematical measurements of compounds that are explored as attributes to develop a correlation between molecular structure and biological activity. Molecular descriptors were calculated by employing PaDEL-Descriptor software, version 2.9 **[17]** .

2.3.Descriptor selection

After the generation of descriptors, the dataset was further optimized by feature selection methods. This was done to enhance the performance of the model by avoiding overfitting of attributes. At first unwanted descriptors were eliminated by utilising the remove useless filter of Weka. This removes descriptors having identical values for all the compounds in the dataset. The correlation-based feature selection (CFS) of Weka was employed to reduce the dimensionality of the dataset by removing irrelevant features **[18]**. This estimates the extent of correlation between the attributes and the target class, as well as inter-correlations between the features. The relevance of the attributes increases with the increase in correlation between attributes and classes and decreases with intercorrelation between the attributes **[19]**. CFS has been applied to determine the best attribute subset. The selected descriptors were then explored as input variables for developing classification models and regression models.

2.4.QSAR models

Optimization of descriptors and selection of the appropriate statistical algorithm for machine learning approaches play an important role in developing reliable predictive models. Different machine learning algorithms like Naive Bayes, Support vector machine, Multilayer perceptron, Random Forest, Gaussian Process, and Linear Regression which are accessible from Weka were incorporated for building predictive models.

2.5 Validation of models

The performance of the models was validated by two validation methods. Ten-fold cross-validation techniques were used at first. In this technique, the training set was randomly divided into ten non-overlapping pairs of training and test sets. Training and testing were carried out ten times in such a way that each time one set was used for testing and the remaining (n-1) sets were used for training. The internal validation does not evaluate the model performance on a new set of data, hence the whole data set was divided into training and test sets, and the models developed from the training set were validated externally using the test set. Finally, the fitness of the model was assessed using various standard parameters like sensitivity (SN), specificity (SP), accuracy, ROC, and F measure. Sensitivity defines the true positive rate while specificity defines the proportion of the true negatives. F measure is a measure of test's accuracy and ROC is a graphical plot between true positive rate and false positive rate used to assess the performance of a binary classifier. Statistical parameters like sensitivity and specificity of classification models are calculated as:

$SN = TP / (TP + FN)$

$SP = TN / (TN + FP)$

Here TP, FP, TN, and FN denote the number of true positives, false positives, true negatives, and false negatives, respectively. In this study, we have also evaluated the performance of the models using the balanced accuracy (BA) also which is considered as the correct classification rate and is calculated as **[20]**

BA = (SN + SP) /2

The accuracy of the regression models was evaluated using the root mean square error (RMSE), the squared correlation coefficient R^2, and the coefficient of determination Q^2. The supremacy of a QSAR model is understood not only by internal validation but also by external validation on a set of test molecules that were not considered while building the model **[21]**.

To determine the applicability of the developed models, validation was again done on another external dataset of 22 Gyr B inhibitor molecules that were not included for the model development taken from the study of R Janupally *et al* **[25] (Structures given in supplementary file Figure 2)** and the drug molecule Novobiocin.

2.6. Docking Studies

The X-ray crystallographic structure of MsmGyrB ATPase domain in complex with aminopyrazinamide resolved at 2.2 A ° (PDB ID: 4B6C) was retrieved from the protein data bank of RCSB. Preparation of the protein has performed by the Protein Preparation Wizard tool **[22]** of Schrodinger. This was done by removing water molecules resulting from the crystallization process and other heteroatoms. The hydrogen atoms were also added according to pH 7.0. The loop regions were refined and energy minimization was done by the application of OPLS3e force field **[23]**.

The active site was defined based on the position of the crystal ligand of MsmGyrB (4B6C). The centroid of the residues predicted by x-ray co-crystallized ligand was defined as the grid box. Van der Waals scaling factor 1.00, charge partial cutoff 0.25, and OPLS3e force filed were applied for receptor grid generation. Preliminary validation of the active site pocket was performed by redocking the crystal ligand, 6-(3,4-dimethyl phenyl)-3-[[4-[3-(4-ethylpiperazin-1-yl)propoxy]phenyl]amino]pyrazine-2-carboxamidewith the target protein in the extra precision mode (XP) using the module GLIDE of Schrodinger **[24]**. Structure-based virtual screenings for a set of active phytomolecules were performed by Glide module in standard (SP) and extra (XP) precision.

2.7. Virtual screening of phytomolecules

Virtual screening was first carried out based on the developed machine learning classification models. The active molecules were then docked into the binding site of MsmGyrB protein 4B6C.

For screening and identifying better lead molecules, the cut-off for the docking score was set as-8.1(docking score of crystal ligand). Phytomolecules which were predicted actives from the classification models and having a docking score lower than cut-off was further filtered for activity prediction using the developed regression based QSAR model.

2.8.ADMET Studies

ADMET properties deal with its absorption, distribution, metabolism, excretion, and toxicity in the human body. It constitutes the pharmacokinetic profile of a molecule. Poor pharmacokinetics and toxicity in the biological system lead to failure in drug development. I*n silico* ADMET analysis is the fastest approach to find this. In this study, we have used the admetSAR prediction tool for this purpose.

3.RESULTS AND DISCUSSION

3.1 Molecular descriptors calculation and selection

By employing PaDEL-Descriptor software a total of 15345 molecular descriptors(1D, 2D, 3D descriptors, and fingerprints) have been generated for all the 286 GyrB inhibitors. Their pIC50 values were appended to the last column. For a classification model, the dataset was divided into two classes highly active considered as actives, and least active considered as inactives. Molecules were assigned as actives if pIC50 is greater than or equal to 5M and as inactives if pIC50 is less than 5M. To build the model 75% of the dataset were randomly selected as a training set and the remaining 25 % dataset was used for testing the model. Even though it was a random selection, care was taken so that the highest active molecule and the least active molecule of each dataset remained in the training set to ensure the training is done over a wide range. The number of molecules in the training sets was 215 and that of the test set was 71 molecules (The train and test files used for the generating machine learning based models have been provided as supplementary file **Table S1** considering the reproducibility of the results.) All the attributes that do not vary at all or that vary too much were considered useless and were removed using the remove useless filter of Weka. This resulted in the reduction of the number of descriptors to 4923. This dataset was further optimized by feature selection methods to improve the quality of the model. Weka's attribute selection CfsSubsetEval-Best first was explored to select important descriptors from the dataset. This resulted in 44physically significant descriptors for the classification model and 21 descriptors for the regression model. The list of descriptors filtered for model building is given in the Supplementary file **Table S2**.

3.2.Classification model

The QSARclassification models for the training set were constructed by applying four machine learning methods, namely Random Forest (RF), Naive Bayes(NB), Multi-layer Perceptron (MLP), and Sequential Minimal Optimisation (SMO). The best model was selected by analysing the statistical parameters of the different models given in **Table 1**.

It can be seen that all models build on the 44physically significant attributes, showed high accuracies. Model NB and SMO showed almost similar accuracies of 78.50% and 79.44% respectively for the training set but NB showed slightly higher accuracy of 75% for the test set when compared to SMO. We have also used balanced accuracy, besides accuracy, to introduce a correct balance in the sensitivity and specificity and to give a more accurate measure of the performance of the models. Balancedaccuracy value for the SMO and NB classifier was almost the same for the training set but NB showed higher balanced accuracy for the test set. The ROC values were also the highest NB which clearly states that the NB model outperforms the other classifiers. The percentage of F measure which is the measure of test accuracy was also highest for NB. Thus NB model was selected as the most accurate classifier for this dataset since it had fairly good values for all these parameters. Applicability of the model was further validated using an external set consisting of the drug molecule novobiocin and 22 molecules. Out of 23 molecules 18 were correctly classified as actives **(Table 4)**. Hence the model could be used to classify any diverse set of molecules. A total of 1040 phytomolecules was downloaded from the curated database IMPPAT **[26]** and classified as actives and inactives according to the Naive Bayes classification model. Out of the 1040 phytomolecules 436 molecules were classified as actives.

3.3 Structure-based virtual screening using molecular docking

Molecular Docking simulations were performed by Glide module in standard (SP) and extra (XP) precision. 436active phytomolecules were subjected to SP docking and all those molecules having docking scores less than -8.1were subjected to XP docking and their interactions were studied.

The output poses after XP Glide docking were rescored and the binding energy of all the compounds was calculated using Prime MM-GBSA with water as the solvent medium. The low binding energy value signifies the stability of the ligand-protein docked complex **[27]**. The binding energies of these phytomolecules towards the GyrB domain were found to be a large negative value **(Table 2)** suggesting the high stability of the complex.

It is seen that out of the 12 molecules 6 molecules showed the interaction with Asp79 which is a key interaction for GryB inhibitors **[11]**. The interaction diagram of the phytomolecule ID:5492482 having docking score -11.142is shown in **Figure 1(a)**. The figure shows that the molecule forms hydrogen bonds as hydrogen bond donors with Asp79, Asp55, Glu56 and Asn52 and as hydrogen bond acceptors with Arg82 and Gly83. There are also hydrophobic interactions with Val125, Val99, Met100, Val128, Val49, Ile171, Pro85, Ile84, Ala53, Met58 and Ala 59. Docking analysis of the crystal ligand aminopyrazinamide with the protein also shows similar hydrophobic interaction with Val125, Val49, Ile171, Pro85, Ile84, Ala53. and hydrogen bonded interaction with Asp79, Gly83 **(Figure 1 b)**.

3.4 Regression models

The regression models are built to predict the activity of the screened molecules and to identify important descriptors that contribute to this activity. Various models were built using GaussianProcess, Linear Regression, MLP, and SMOreg algorithms. The robustness and stability of the model were revealed by the correlation coefficients between the observed and predicted activity of the training set R^2, external predictability of the model was assessed by cross-correlation coefficient, Q^2. For a model to be good, the R^2 and Q^2 have to be close and greater than 0.6 **[28]**. The results of the different models are given in **Table 3**. From the table, it is clear that all developed models have acceptable statistical performance. Therefore, these models could be useful for the prediction of GyrB activity. GaussianProcess, SMOreg, and linear regression methods gave a similar R^2 value of 0.94 but the Q^2 of the linear regression method was slightly greater than the other methods. RMSE of 0.45 also suggests for best performance of the linear regression model and is thus used to predict the activities of the screened molecules.

3.4.1.Linear regression model

Linear regression is a tool used to figure out the relationship amongst various variables. The linear equation obtained explains the effect of independent variables on the dependent variable and thus activities of the molecules can be predicted based on the equation. In our model the biological activity is related to the descriptors as

$$\mathbf{pic50 = -1.2868 * VCH\text{-}5 + -2.0371 * RPCG + 0.0344 * Wlambda2.unity + 0.1862 * FP47 + 1.6121 * FP110 + 0.313 * FP450 + 0.2934 * ExtFP112 + 0.5874 * ExtFP114 + 0.361 * ExtFP388 + -0.1582 * ExtFP896 + -0.2394 * GraphFP815 + 0.46 * GraphFP960 + 5.0777}$$

The 3D Descriptor Wlambda2.unity (Directional WHIM, weighted by unit weights), the fingerprints FP47, FP110, FP450, extended fingerprint describing ring features ExtFP112, ExtFP114, ExtFP388, and graph only fingerprint which does not take bond orders into accountGraphFP960 has positive coefficients which suggest that increase in the values of these descriptors increases the GyrB activity whereas 2D descriptor VCH-5(Valence chain), 3D Descriptors RPCG (Relative positive charge--most positive charge / total positive charge), extended fingerprint ExtFP896 and graph only fingerprint GraphFP815 have negative coefficients which suggest that decrease in the values of these descriptors decreases the GyrB activity. When the applicability of the model was checked for another external set of 22 molecules and drug molecule novobiocin, it was seen that 15 molecules were predicted the activity with an error of less than 1 **(Table 4)**. Hence the model was used to predict the activities of the six phytomolecules screened after docking. The predicted activity along with their source is given in **Table 5**. Structures of these molecules are given in **Figure 2**.

3.5.ADMET Predictions

A good drug candidate should be absorbed in the required time and well distributed throughout the human body for effective metabolism and action. Toxicity is another very important factor that often overshadows the ADME behaviour. **Table 6** illustrates the various ADMET properties obtained from the admet SAR tool.

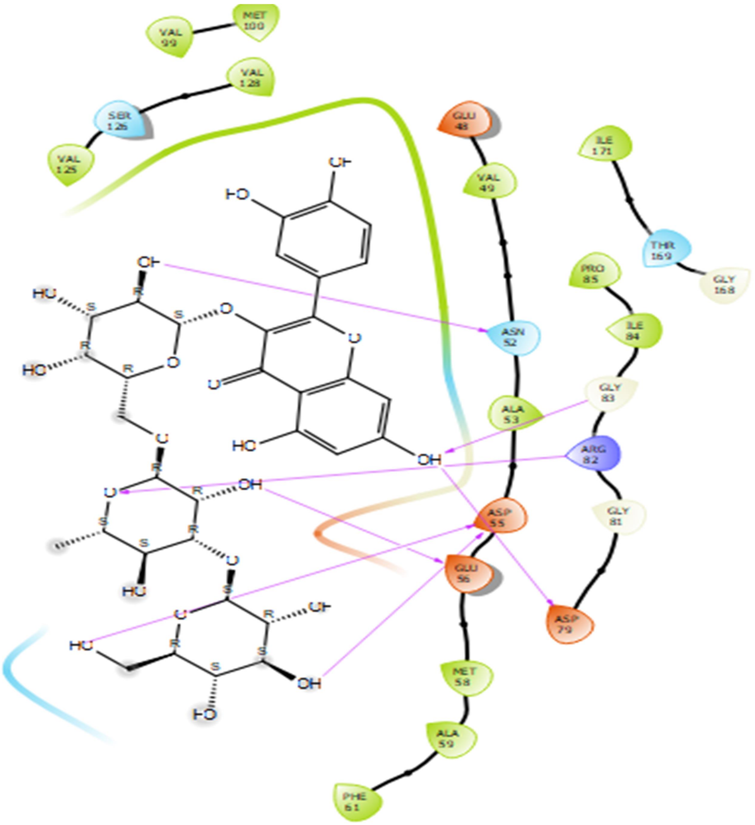
Greater Human Intestinal Absorption (HIA) denotes that the compound is better absorbed from the intestinal tract after oral administration. From the **Table 6**, it is clear that the phytochemical CID:5492482 has the highest HIA and penetration through the Blood-Brain Barrier (BBB). Cytochrome P450 is an enzyme responsible for the metabolism of the drug in a biological system and its clearance from the body. Its inhibition can affect the drug metabolism and increase toxicity. Here all the molecules are a noninhibitor of CYP450 means that the molecule will not obstruct the biotransformation of drugs metabolized by the CYP450 enzyme. The carcinogenic profile also revealed that the ligands were noncarcinogenic. Here all the molecules screened are noninhibitors of hERG inhibition suggesting that these are not cardiotoxic

Table 1: Statistical parameters for different classification models

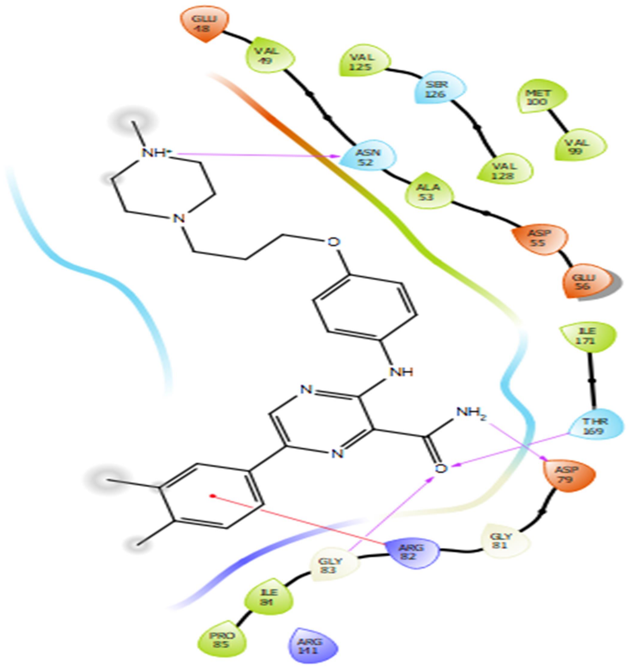
Method	Sensitivity%		Specificity%		Balanced Accuracy%		F Measure%		ROC%		Accuracy%	
	Training	Test	Training	Test	Training	Test	Training	Test	Training	Test	Training	Test
RF	83.65	85.42	63.64	66.67	73.65	76.04	73.00	79.00	81.00	79.00	73.36	79.17
NB	99.04	87.50	59.09	50.00	79.06	68.75	78.00	74.00	85.00	71.00	78.50	75.00
MLP	69.23	60.42	73.64	58.33	71.43	59.38	72.00	61.00	84.00	76.00	71.50	59.72
SMO	90.38	77.08	69.09	58.33	79.74	67.71	79.00	71.00	79.00	68.00	79.44	70.83

Table 2:Docking scores of GyrB active phytochemicals

Phytomolecule ID	Docking Score	MM-GBSA bind	Interacting residues
3084341	-10.843	-47.25	ASP79,ASP55,GLY83,GLU48,ARG82
5492482	-11.142	-51.56	ASP79, ASP55, ARG82, ASN52, GLU56, GLY83
10813969	-8.222	-49.09	ASP79, GLU56, ASP55, ASN52, GLU48, THR169
11980943	-9.404	-55.65	ASP79,ASP51,ASP55,ASN52,GLU48
12004528	-10.079	-52.31	ASP79, ASN52, THR169, GLY83
12304093	-9.332	-48.75	ARG82, GLY83
14889736	-8.333	-41.26	ASP55,ASN52,GLU56
44256718	-11.435	-64.72	ASP55,GLU48,ARG82
44256720	-10.91	-58.29	GLU56, ASN52, VAL99
44256732	-10.014	-55.08	ASP55,ARG82,ASP51,GLU48,VAL99,GLU56
44258427	-10.613	-68.13	GLU56, ASN52, VAL99
44259146	-10.193	-56.79	ASP79, ASN52, ASP55, GLU56, GLU48



(a)



(b)

Figure 1: (a) Molecular interaction diagram of screened phytomolecule with GyrB protein 4B6C; (b) Interaction of the crystal ligand with the target protein 4B6C

Table 3: Statistical parameters for regression models

Method	Training			Test		
	R^2	MAE	RMSE	Q^2	MAE	RMSE
GaussianProcess	0.94	0.35	0.45	0.90	0.41	0.59
Linear Regression	0.94	0.35	0.45	0.91	0.39	0.60
MLP	0.91	0.46	0.58	0.84	0.72	0.94
SMOreg	0.94	0.33	0.45	0.90	0.41	0.60

Table 4: Validation of the QSAR models using an external set

External test molecules	IC50 (µM)	pIC50 (M)	Predicted by the classification model	Predicted by the regression model	error
Novobiocin	180nM	6.7447275	Active	5.404	-1.341
4	1.75	5.756962	Active	4.782	-0.975
5	2.6	5.5850267	Active	4.794	-0.791
6	9.8	5.0087739	Active	4.785	-0.224
7	15.7	4.8041003	Active	4.82	0.016
8	23.9	4.6216021	Active	5.091	0.469
9	5.85	5.2328441	Active	4.779	-0.454
10	3.56	5.44855	Active	5.002	-0.447
11	9.8	5.0087739	Active	4.848	-0.161
12	8.07	5.0931265	Active	4.863	-0.23
13	8.707	5.0601315	Active	4.85	-0.21
14	23.31	4.6324577	Inactive	4.9	0.267
15	21.51	4.6673596	Active	5.178	0.51
16	17.85	4.7483618	Inactive	4.869	0.121
17	4.538	5.3431355	Active	5.065	-0.278
18	2.136	5.6703988	Active	6.775	1.104
19	3.6	5.4436975	Active	6.741	1.297
20	6.4	5.19382	Active	6.781	1.587
21	1.5	5.8239087	Active	6.776	0.952
22	7.5	5.1249387	Inactive	6.832	1.707
23	1.5	5.8239087	Active	6.842	1.018
24	13.05	4.8843895	Inactive	6.838	1.953
25	7.6	5.1191864	Inactive	6.86	1.741

Table 5: Screened Phytochemicals with their predicted activity

Phytochemical Identifier	Indian Medicinal Plant	Common Name	Phytochemical name	Predicted activity
CID:3084341	**Phyllanthus *emblica***	**Amla**	**Terchebin**	**8.108**
CID:5492482	**Solanum *nigrum***	**Black Nightshade**	**Quercetin 3-glucosyl(1-3)rhamnosyl(1-6)galactoside**	**8.234**
CID:10813969	**Asparagus *racemosus***	**Satawari**	**isoquercitin**	**8.319**
CID:11980943	**Phyllanthus *emblica***	**Amla**	**Theaflavin**	**6.651**
CID:12004528	**Asparagus *racemosus***	**Satawari**	**Miquelianin**	**8.268**
CID:44259146	**Solanum *nigrum***	**Black Nightshade**	**Quercetin 3-gentiobioside**	**8.367**

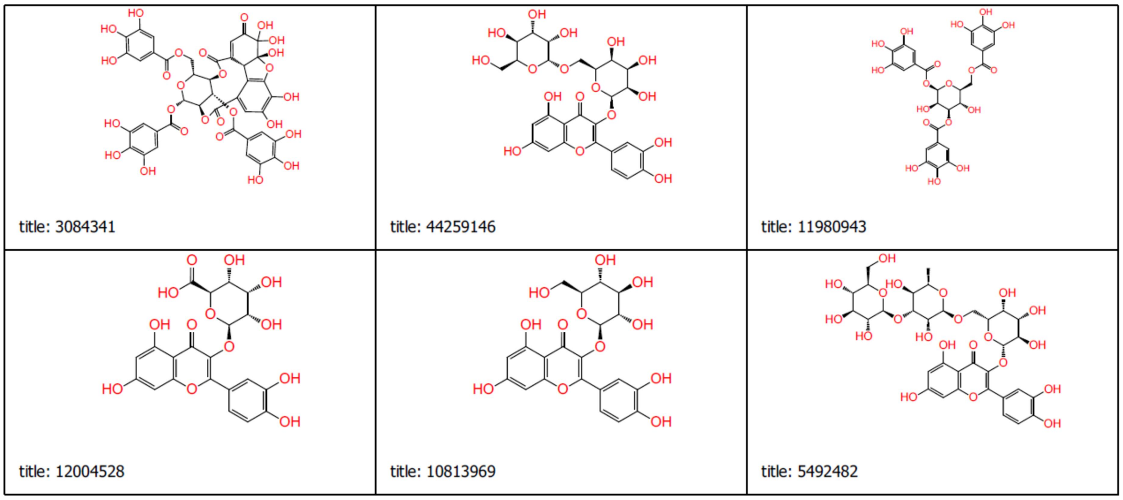


Figure 2: Final screened phytomolecules

Table 6: ADMET Properties of the screened phytomolecules

ADMET PROPERTY	Human Intestinal Absorption	Blood-Brain Barrier	CYP450 Inhibition	Carcinogenicity	hERG inhibition
CID:3084341	**0.677**	**0.541**	**Non-substrate/ Non-inhibitor**	**Non-carcinogens**	**Non-inhibitor**
CID:5492482	**0.803**	**0.859**	**Non-substrate/ Non-inhibitor**	**Non-carcinogens**	**Non-inhibitor**
CID:10813969	**0.786**	**0.698**	**Non-substrate/ Non-inhibitor**	**Non-carcinogens**	**Non-inhibitor**
CID:11980943	**0.726**	**0.839**	**Non-substrate/ Non-inhibitor**	**Non-carcinogens**	**Non-inhibitor**
CID:12004528	**0.737**	**0.605**	**Non-substrate/ Non-inhibitor**	**Non-carcinogens**	**Non-inhibitor**
CID:44259146	**0.701**	**0.725**	**Non-substrate/ Non-inhibitor**	**Non-carcinogens**	**Non-inhibitor**

4. CONCLUSIONS

Various classification models were developed to classify molecules as DNA GyrB actives and inactives. The model using the Naïve Bayes algorithm was chosen as the best model and was used to screen the phytochemical database IMPPAT. Docking studies were carried out on the screened active molecules and 12 molecules showing appreciable higher docking scores were screened out. Out of these 6 molecules showed the key interaction with Asp79. Large negative values of binding energy obtained for these molecules with target protein also explained the stability of the complex. Activities of all these molecules were then predicted by a linear regression model developed using Weka. From the regression model, the descriptors showing positive contributions and those showing negative contributions were also identified. The predicted pIC50 values of five phytomolecules were greater than 8.0 which suggests them to be potential GyrB inhibitors. Compound CID: 5492482, Quercetin 3-glucosyl(1-3)rhamnosyl(1-6)galactoside (IUPAC name: 3-[(2*S*,5*R*)-6-[[(2*R*,4*S*,5*S*)-3,5-dihydroxy-6-methyl-4-[(2*S*,4*S*,5*S*)-3,4,5-trihydroxy-6-(hydroxymethyl)oxan-2-yl]oxyoxan-2-yl]oxymethyl]-3,4,5-trihydroxyoxan-2-yl]oxy-2-(3,4-dihydroxyphenyl)-5,7-dihydroxychromen-4-one) present in plant Solanum nigrum showed highest docking score and inhibitory activity against DNA GyrB. It also has the highest HIA and BBB suggesting good absorption of the molecule in our body. The toxicity studies also revealed them to be non-toxic. This molecules can be further modified to make it pharmacologically more active.

REFERENCES

[1] Country Profiles. *OECD SME Entrep. Outlook 2019* 189–249 (2019). doi:10.1787/10f0b36a-en

[2] Barančoková, M., Kikelj, D. & Ilaš, J. Recent progress in the discovery and development of DNA gyrase B inhibitors. *Future Med. Chem.* **10**, 1207–1227 (2018).

[3] Durcik, M. *et al.* ATP-competitive DNA gyrase and topoisomerase IV inhibitors as antibacterial agents. *Expert Opin. Ther. Pat.* **29**, 171–180 (2019).

[4] Subbarao, N., Chandra, S., Tiwari, N. & Kumari, M. Comparative analysis of machine learning based QSAR models and molecular docking studies to screen potential anti-tubercular inhibitors against InhA of mycobacterium tuberculosis. *Int. J. Comput. Biol. Drug Des.* **11**, 209 (2018).

[5] Qin, Z., Xi, Y., Zhang, S., Tu, G. & Yan, A. Classification of Cyclooxygenase - 2 Inhibitors Using Support Vector Machine and Random Forest Methods. *J. Chem. Inf. Model.* **59**, 1988–2008 (2019).

[6] Mohanraj, K. *et al.* IMPPAT: A curated database of Indian Medicinal Plants, Phytochemistry and Therapeutics. *Sci. Rep.* **8**, 1–17 (2018).

[7] Indrasena, K. *et al.* Bioorganic & Medicinal Chemistry An efficient synthesis and biological screening of benzofuran and benzo [d] isothiazole derivatives for Mycobacterium tuberculosis DNA

GyrB inhibition. *Bioorg. Med. Chem.* **22**, 6552–6563 (2014).

[8] Ghorpade, S. R. *et al.* Thiazolopyridine Ureas as Novel Antitubercular Agents Acting Through Inhibition of DNA Gyrase B. (2013). doi:10.1021/jm401268f

[9] Jeankumar, V. U. *et al.* Development of novel N-linked aminopiperidine-based mycobacterial DNA gyrase B inhibitors: Scaffold hopping from known antibacterial leads. *Int. J. Antimicrob. Agents* 4–13 (2014). doi:10.1016/j.ijantimicag.2013.12.006

[10] Renuka, J. *et al.* Design , synthesis , biological evaluation of substituted benzofurans as DNA gyrase B inhibitors of Mycobacterium tuberculosis. *Bioorg. Med. Chem.* (2014). doi:10.1016/j.bmc.2014.06.041

[11] Kale, R. R. *et al.* Bioorganic & Medicinal Chemistry Letters Thiazolopyridone ureas as DNA gyrase B inhibitors : Optimization of antitubercular activity and efficacy. *Bioorg. Med. Chem. Lett.* **24**, 870–879 (2014).

[12] Saxena, S. *et al. Development of 2-amino-5-phenylthiophene-3-carboxamide derivatives as novel inhibitors of Mycobacterium tuberculosis DNA GyrB domain. BIOORGANIC & MEDICINAL CHEMISTRY* (Elsevier Ltd, 2015). doi:10.1016/j.bmc.2015.02.032

[13] Medapi, B. *et al.* 4-Aminoquinoline derivatives as novel Mycobacterium tuberculosis GyrB inhibitors: Structural optimization, synthesis and biological evaluation. *Eur. J. Med. Chem.* **103**, 1–16 (2015).

[14] Medapi, B. *et al.* Bioorganic & Medicinal Chemistry Design and synthesis of novel quinoline – aminopiperidine hybrid analogues as Mycobacterium tuberculosis DNA gyraseB inhibitors. *Bioorg. Med. Chem.* **23**, 2062–2078 (2015).

[15] Schrödinger Release 2019-2: LigPrep, Schrödinger, LLC, New York, NY, 2019. No Title.

[16] Harder, E. *et al.* OPLS3: A Force Field Providing Broad Coverage of Drug-like Small Molecules and Proteins. *J. Chem. Theory Comput.* **12**, 281–296 (2016).

[17] Yap, C. W. (2010). PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. Journal of Computational Chemistry, 32(7), 1466–1474. doi:10.1002/jcc.21707.

[18] Li, T., Zhang, C. & Ogihara, M. A comparative study of feature selection and multiclass classfication methods for tissue classification based on gene expression. *Bioinformatics* **20**, 2429–2437 (2004).

[19] Kohavi, R. & John, H. Artificial Intelligence Wrappers for feature subset selection. *Artif. Intell.* **97**, 273–324 (1997).

[20] Kovalishyn, V. *et al.* Rational design of isonicotinic acid hydrazide derivatives with antitubercular activity: Machine learning, molecular docking, synthesis and biological testing. *Chem. Biol. Drug Des.* **92**, 1272–1278 (2018).

[21] Divya, V., Pushpa, V. L., Sarithamol, S. & Manoj, K. B. Computational approach for generating robust models for discovering novel molecules as Cyclin Dependent Kinase 4 inhibitors. *J. Mol. Graph. Model.* **82**, 48–58 (2018).

[22] Madhavi Sastry, G., Adzhigirey, M., Day, T., Annabhimoju, R. & Sherman, W. Protein and ligand preparation: Parameters, protocols, and influence on virtual screening enrichments. *J. Comput. Aided. Mol. Des.* **27**, 221–234 (2013).

[23] Roos, K. *et al.* OPLS3e: Extending Force Field Coverage for Drug-Like Small Molecules. *J. Chem. Theory Comput.* **15**, 1863–1874 (2019).

[24] Friesner, R. A. *et al.* Glide : A New Approach for Rapid , Accurate Docking and Scoring . 1 . Method and Assessment of Docking Accuracy. 1739–1749 (2004).

[25] Jeankumar, V. U. *et al.* Exploring the Gyrase ATPase domain for tailoring newer anti-tubercular drugs : Hit to lead optimization of a novel class of thiazole inhibitors. *Bioorg. Med. Chem.* (2014). doi:10.1016/j.bmc.2014.12.001

[26] Saxena, S., Renuka, J., Yogeeswari, P. & Sriram, D. Discovery of Novel Mycobacterial DNA Gyrase B Inhibitors : In Silico and In Vitro Biological Evaluation. 597–609 (2014). doi:10.1002/minf.201400058

[27] Sarithamol, S. & Lalithabhai, V. Comparative QSAR model generation using pyrazole derivatives for screening Janus kinase-1 inhibitors. *Chem. Biol. Drug Des.* 1–17 (2020). doi:10.1111/cbdd.13667

[28] Achutha, A. S., Pushpa, V. L. & Suchitra, S. Theoretical Insights into the Anti-SARS-CoV - 2 Activity of 2 Chloroquine and Its Analogs and In Silico Screening of Main 3 Protease Inhibitors. *J. Proteome Res.* (2020). doi:10.1021/acs.jproteome.0c00683