Original Article

INTERNATIONAL JOURNAL OF RESEARCH IN PHARMACEUTICAL SCIENCES

Published by IJRPS Home Page: https://ijrps.com/

Identification of a 2-aminothiazole framework using classical QSAR model targeting chloroquine-sensitive *Plasmodium falciparum*

Ravindran Karuppaiyan[1], Anitha Gunavel[2], Kasthuri Bai Solai[1]*, Prabha Thangavelu[3]

[1]Department of Anaesthesiology, Government Medical College and Hospital, Pudukkottai-622004, Tamil Nadu, India

[2]Coimbatore Medical College, Coimbatore-641018, Tamil Nadu, India

[3]Department of Pharmaceutical Chemistry, Nandha College of Pharmacy, Affiliated with The Tamil Nadu Dr. MGR Medical University-Chennai, Erode-638052, Tamil Nadu, India

Article History

Received on: 18 Nov 2023
Revised on: 11 Jan 2024
Accepted on: 15 Jan 2024

Keywords

Plasmodium falciparum, Antimalarial effect, Chloroquine-sensitive strain, 2-aminothiazole framework, QSARINS

Abstract

Chloroquine-sensitive *Plasmodium falciparum* is the most deadly form of human malaria. It is associated with a number of mutations in *P. falciparum*. Chloroquine-resistant transporter is a protein that serves as a transporter in the parasite's digesting vacuole membrane. In order to combat chloroquine-sensitive *P. falciparum* strains (NF54), this study employs QSAR modelling to examine possible structural alterations of 2-amino-thiazole derivatives. The traditional QSAR model was built using the PaDEL descriptor via QSARINS software. The model was found to have an internal cross-validation value of Q^2loo = 0.7890 and an external validation parameter of RMSE ext = 0.6938. The predicted pIC_{50} values from the QSAR techniques for the case study chemicals were compared and found to be well fitted to the model and well predicted for the external set of compounds. The outcome demonstrates the value of using the suggested method in the creation of new medication candidates could fill the critical gap in scientific knowledge and open up novel possibilities for pharmaceutical development.

*Corresponding Author

Name: Kasthuri Bai Solai
Phone: +91 9655649030
Email: kasthuri.anaes@gmail.com

eISSN: 0975-7538
DOI: https://doi.org/10.26452/ijrps.v15i1.4659

Production and hosted by
IJRPS | www.ijrps.com

INTRODUCTION

Millions of people die from malaria each year, making it one of the most common diseases on Earth. The advent of plasmodium strains that exhibit sensitivity and/or resistance to established chemotherapeutic drugs has exacerbated the health issue created by malaria, one of the most deadly parasite illnesses. Furthermore, to make more challenging medical treatment for malaria in endemic regions, the lack of cutting-edge technology, and reasonably priced medications has increased the fatality rate [1]. This scenario emphasizes how vital it is to find novel anti-malarial drugs. However, medical chemists are driven to find novel chemical substances that could

be useful as antimalarials due to the decreasing availability and expensive price of artemisinin and similar drugs. While there are some experimental techniques for determining whether a chemical has antimalarial activity (such as in vivo and in vitro assay tests), none of them are particularly helpful or cheap, and they may result in harmful byproducts from the current experimental procedures. All of these techniques have been tested using biological materials and receptors that are preferable to humans, rats, and mice [2]. Using quantitative structure-activity relationship (QSAR) approaches is an effective way to collect a full set of data without having to conduct costly laboratory tests.

To overcome this hurdle faced by researchers, many hybrid *in silico* technologies are now employed to expedite the drug discovery process. Modifying the structure of natural products with well-established activities and then using statistical techniques (QSAR) to determine correlations among chemical structures and their corresponding biological activities is an effective approach that could be taken into consideration to find potential drugs quickly [3]. Recently, to design relevant structural information for designing medicines with greater biological activities, the use of machine learning has grown into an important tool in drug discovery due to the rise of chemical data [4][5]. Once a link between structure properties and activity is discovered, any number of compounds, including ones that have not been synthesised yet, may easily be tested for the selection of compounds with the desired features [6][7]. Additionally, some computer techniques, such as partial least squares regression (PLS) and multiple linear regression (MLR), have been demonstrated to be effective in establishing these correlations [8].

The method is used in a statistical analysis to produce several QSAR models and choose appropriate descriptors. The best model created can be used to forecast test set chemicals that have not been included in the training set chemicals after the analysis of QSAR models has been done. The model's appropriate validation is ensured by randomization tests run on it at varied intervals of confidence levels. The primary goal of this work is to create a new QSAR model that will predict the anti-malarial activity of derivatives of 2-aminothiazole compounds. It has been previously shown that thiazole compounds have a wide range of biological actions, including antimycobacterial and antimalarial activity [9][10][11][12][13].

This study aimed to develop a robust QSAR model by employing 2-aminothiazole compounds as an antagonist of chloroquine-sensitive NF54 strains. This work is a continuation of our attempts to build a QSAR model, which is necessary in today's drug discovery process. Herein, we report a QSAR model developed by using traditional QSARINS software for the statistical analysis and validation of 2-aminothiazole derivatives. Following the proper preprocessing, the PaDEL descriptors have been calculated and utilised to construct QSAR models to determine the association between the biological activity (measured in μM/ml) and the properties of 2-aminothiazole, as indicated by the molecular descriptors [14][15]. Our proposed study could address a significant gap in scientific knowledge and open up novel possibilities for pharmaceutical development.

METHODOLOGY

Software and interactive computing platform used

The 2D QSAR prediction model is developed using QSARINS software. The compounds' 2D structures were drawn using ACD/Labs ChemSketch. Using Chem3D Pro version 12.0.2.1076, the MM2 force field was utilised to minimise the energy of the three-dimensional structure. The molecular descriptors were created using the PaDEL descriptor software (version 2.20) [16]. Using ordinary least squares, the MLR-QSAR prediction models were created via QSARINS software. MLR and genetic algorithms (GA) serve as components of a hybrid technique employed by QSARINS [17][18]. It aids in the development of highly predictive yet simple QSAR models.

Dataset collection and transformation

A dataset of 48 diverse derivatives of 2-amino thiazole with their IC_{50} inhibitory values (μM) against the malarial cloroquine-sensitive NF54 strain has been retrieved from the previously reported work [19][20]. To convert the experimental activity values of the chemicals in the dataset, which were represented as μg/ml concentration, the logarithmic scale of their reciprocal values was calculated using the

Table 1 Structures of 2-aminothiazole along with their IC_{50} values

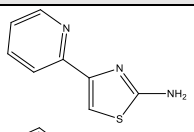
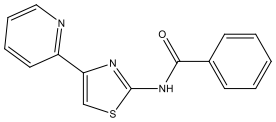
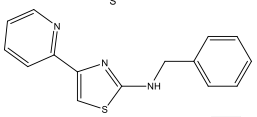
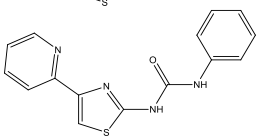
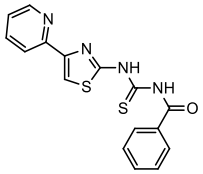
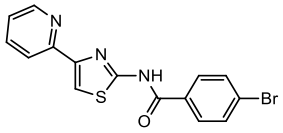
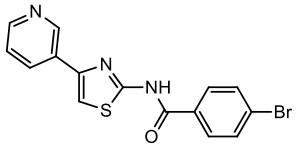
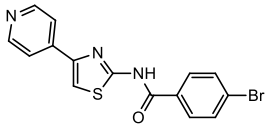
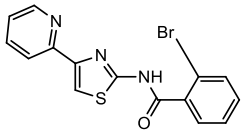
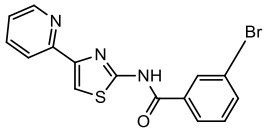
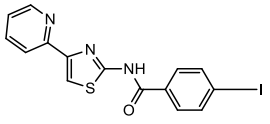
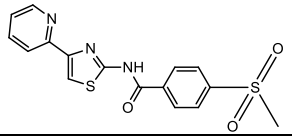
S.No.	Code	STRUCTURE	IUPAC NAME	IC_{50} (μM)
1.	3		4-(pyridin-2-yl)thiazol-2-amine	47.8
2.	4a		N-(4-(pyridin-2-yl)thiazol-2-yl)benzamide	1.0
3.	4b		N-benzyl-4-(pyridin-2-yl)thiazol-2-amine	3.2
4.	4c		1-phenyl-3-(4-(pyridin-2-yl)thiazol-2-yl)urea	1.7
5.	4d		N-((4-(pyridin-2-yl)thiazol-2-yl)carbamothioyl)benzamide	2.7
6.	5a		4-bromo-N-(4-(pyridin-2-yl)thiazol-2-yl)benzamide	5.3
7.	5b		4-bromo-N-(4-(pyridin-3-yl)thiazol-2-yl)benzamide	15.9
8.	5c		4-bromo-N-(4-(pyridin-4-yl)thiazol-2-yl)benzamide	3.5
9.	6		2-bromo-N-(4-(pyridin-2-yl)thiazol-2-yl)benzamide	3.2
10.	7		3-bromo-N-(4-(pyridin-2-yl)thiazol-2-yl)benzamide	1.6
11.	8		4-iodo-N-(4-(pyridin-2-yl)thiazol-2-yl)benzamide	3.9
12.	9		4-(methylsulfonyl)-N-(4-(pyridin-2-yl)thiazol-2-yl)benzamide	5.9

Table 1 Structures of 2-aminothiazole along with their IC_{50} values (Continued)

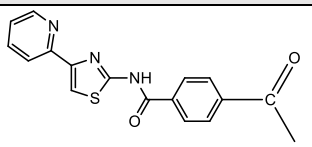
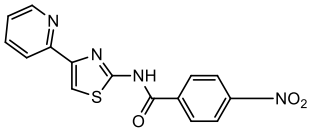
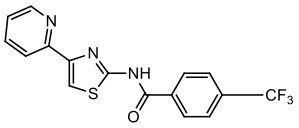
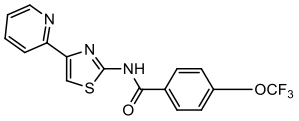
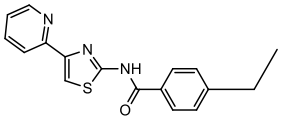
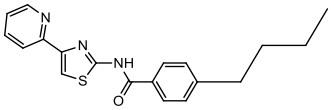
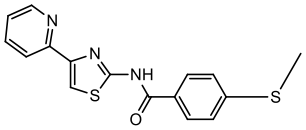
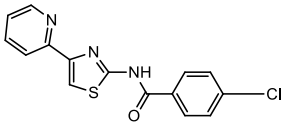
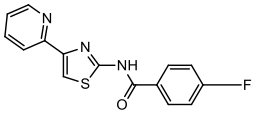
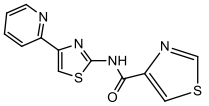
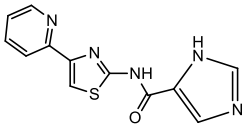
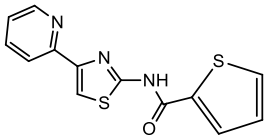
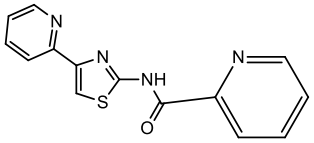
S.No.	Code	STRUCTURE	IUPAC NAME	IC_{50} (µM)
13.	10		4-acetyl-N-(4-(pyridin-2-yl)thiazol-2-yl)benzamide	14.8
14.	12		4-nitro-N-(4-(pyridin-2-yl)thiazol-2-yl)benzamide	6.1
15.	13		N-(4-(pyridin-2-yl)thiazol-2-yl)-4-(trifluoromethyl) benzamide	18.6
16.	14		N-(4-(pyridin-2-yl)thiazol-2-yl)-4-(trifluoromethoxy) benzamide	5.5
17.	15		4-ethyl-N-(4-(pyridin-2-yl)thiazol-2-yl)benzamide	6.5
18.	16		4-butyl-N-(4-(pyridin-2-yl)thiazol-2-yl)benzamide	7.4
19.	17		4-(methylthio)-N-(4-(pyridin-2-yl)thiazol-2-yl)benzamide	7.6
20.	18		4-chloro-N-(4-(pyridin-2-yl)thiazol-2-yl)benzamide	1.3
21.	19		4-fluoro-N-(4-(pyridin-2-yl)thiazol-2-yl)benzamide	1.8
22.	20		N-(4-(pyridin-2-yl)thiazol-2-yl)thiazole-4-carboxamide	13.8
23.	21		N-(4-(pyridin-2-yl)thiazol-2-yl)-1H-imidazole-5-carboxamide	35.0
24.	22		N-(4-(pyridin-2-yl)thiazol-2-yl)thiophene-2-carboxamide	0.9
25.	23		N-(4-(pyridin-2-yl)thiazol-2-yl)picolinamide	35.6

Table 1 Structures of 2-aminothiazole along with their IC_{50} values (Continued)

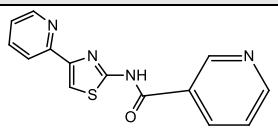
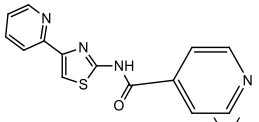
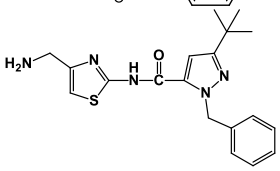
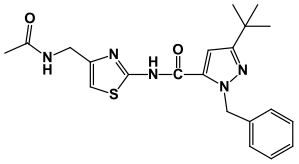
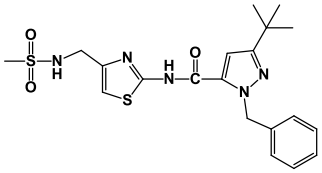
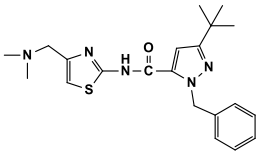
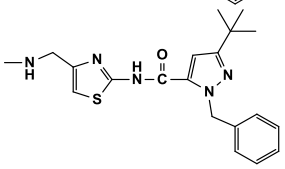
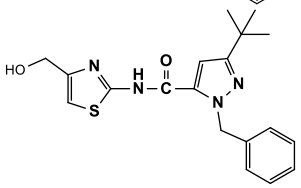
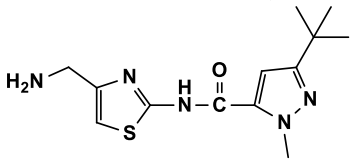
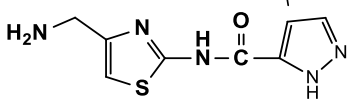
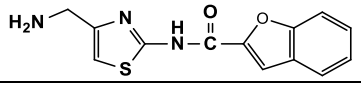
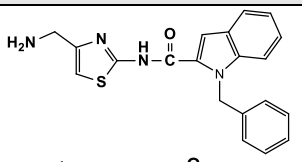
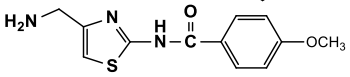
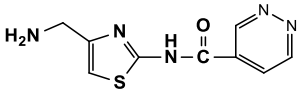
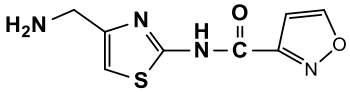
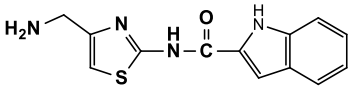
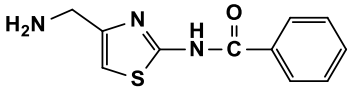
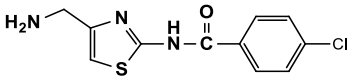
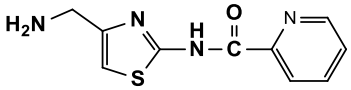
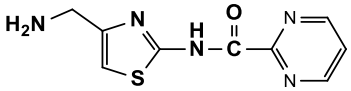
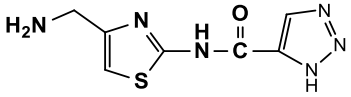
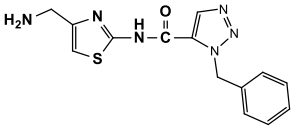
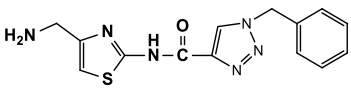
S.No.	Code	STRUCTURE	IUPAC NAME	IC_{50} (μM)
26.	24		N-(4-(pyridin-2-yl)thiazol-2-yl)nicotinamide	36
27.	25		N-(4-(pyridin-2-yl)thiazol-2-yl)isonicotinamide	1.9
28.	3		N-(4-(aminomethyl)thiazol-2-yl)-1-benzyl-3-(tert-butyl)-1H-pyrazole-5-carboxamide	0.08
29.	11		N-(4-(acetamidomethyl)thiazol-2-yl)-1-benzyl-3-(tert-butyl)-1H-pyrazole-5-carboxamide	4.27
30.	12		1-benzyl-3-(tert-butyl)-N-(4-(methylsulfonamidomethyl)thiazol-2-yl)-1H-pyrazole-5-carboxamide	2.93
31.	13		1-benzyl-3-(tert-butyl)-N-(4-((dimethylamino)methyl)thiazol-2-yl)-1H-pyrazole-5-carboxamide	1.48
32.	14		1-benzyl-3-(tert-butyl)-N-(4-((methylamino)methyl)thiazol-2-yl)-1H-pyrazole-5-carboxamide	0.19
33.	15		1-benzyl-3-(tert-butyl)-N-(4-(hydroxymethyl)thiazol-2-yl)-1H-pyrazole-5-carboxamide	0.67
34.	18		N-(4-(aminomethyl)thiazol-2-yl)-3-(tert-butyl)-1-methyl-1H-pyrazole-5-carboxamide	34.11
35.	19		N-(4-(aminomethyl)thiazol-2-yl)-1H-pyrazole-5-carboxamide	45
36.	20		N-(4-(aminomethyl)thiazol-2-yl)benzofuran-2-carboxamide	36.62

Table 1 Structures of 2-aminothiazole along with their IC_{50} values (Continued)

S.No.	Code	STRUCTURE	IUPAC NAME	IC_{50} (µM)
37.	21		N-(4-(aminomethyl)thiazol-2-yl)-1-benzyl-1H-indole-2-carboxamide	6.98
38.	22		N-(4-(aminomethyl)thiazol-2-yl)-4-methoxybenzamide	11.10
39.	23		N-(4-(aminomethyl)thiazol-2-yl)pyridazine-4-carboxamide	42.54
40.	24		N-(4-(aminomethyl)thiazol-2-yl)isoxazole-3-carboxamide	45
41.	25		N-(4-(aminomethyl)thiazol-2-yl)-1H-indole-2-carboxamide	22.21
42.	26		N-(4-(aminomethyl)thiazol-2-yl)benzamide	42.87
43.	27		N-(4-(aminomethyl)thiazol-2-yl)-4-chlorobenzamide	37.45
44.	28		N-(4-(aminomethyl)thiazol-2-yl)picolinamide	43
45.	29		N-(4-(aminomethyl)thiazol-2-yl)pyrimidine-2-carboxamide	44
46.	30		N-(4-(aminomethyl)thiazol-2-yl)-1H-1,2,3-triazole-5-carboxamide	45
47.	31		N-(4-(aminomethyl)thiazol-2-yl)-1-benzyl-1H-1,2,3-triazole-5-carboxamide	3.19
48.	32		N-(4-(aminomethyl)thiazol-2-yl)-1-benzyl-1H-1,2,3-triazole-4-carboxamide	3.19

following equation: The experimental activity levels, therefore, resemble the usual distribution. The structures of 2-aminothiazole along with their IC_{50} values in µM are listed in **Table 1**.

$pIC_{50} = -\log10\ [IC_{50}]$

Computation of molecular descriptors

The structure was stored as a '.mol' file format after optimisation, and it was then loaded into the PaDEL descriptor programme to compute around 1445 descriptors for each molecule. The dataset employed in this study consists of the biological activity of the compounds and the accompanying chemical descriptors. However, we have ultimately arrived at 917 descriptors for our machine learning QSAR model protocol after performing preprocessing such as deleting highly correlated descriptors, zeros, and missing values.

Optimised retrieval of descriptors

The first step in QSAR analysis is choosing the most appropriate descriptors from the libraries of available descriptors. The most important descriptors were chosen for this investigation using the genetic algorithm (GA) approach included in the QSARINS program with respect to an objective function (the cross-validation

correlation coefficient of leave-one-out, abbreviated as R^2cv).

Model development using QSARINS

The QSAR model via QSARINS has been created using the multiple linear regression technique. Molecular descriptors, the independent variable in this study, and dependent variable Y (pIC_{50}) were found to be linearly correlated using the MLR approach. The basic principle is least squares modelling, which fits the model to minimise the sum-of-squares of the variations between the observed and predicted values. Regression coefficient values (R^2) are estimated by MLR using the least squares fitting of the curve technique. The model best resembles all of the separate points of data by creating an interaction that seems to be a straight line (linear). This is how a regression equation looks:

$Y = m1x1 + m2x2 + m3x3 + c$

where 'Y' is biological activity, all 'm's are regression coefficients for the corresponding independent variable, 'x' is a regression constant, and 'c' is an intercept.

Y-scrambling test

As suggested by Roy and Mitra (2011) [21], the information from the training set was subjected to the y-scrambling, or y-randomization test, to make sure the generated QSAR model is reliable and not obtained by chance. It acts as an external validation measure to evaluate the generated QSAR model's validity and reliability. In the present investigation, the independent variable remains constant while the dependent variable is randomly scrambled for building MLR models. The created models are flexible, as evidenced by the predicted significantly low R^2 and leave-one-out cross-validated correlation coefficient Q^2loo values for multiple runs in the derived models.

Model applicability domain

It has been suggested to evaluate the applicability domain of the model in order to determine if a QSAR model is reliable for making predictions within the chemical space in which it has been designed. A QSAR model can have its applicability domain defined in several ways. However, the most popular leverage technique was applied in the present study [17]. The warning leverage ($h*$) is a predictive tool that is defined as $h* = 3(P + 1)/n$, where 'P' is the number of descriptors in the model and 'n' is the number of training compounds. It represents the limit of the normal range for 'X' outliers. For test compounds with leverages $hi < h*$, predictions made by the model have been accepted as reliable.

Williams' plot, which shows the standardised residuals against leverage levels, provides a graphic depiction of the QSAR model's applicability domain. It was applied in the present study to interpret the produced model's applicability domain. For chemicals in the external test set, the domain of reliable prediction was characterized as those with leverage levels within the threshold range of $hi < h*$ and standardized residuals no larger than ±3 standard deviation values. A test set was considered a Y outlier if its standard deviation units were more than ±3. Additionally, the training set contains molecules that are influential, that is, compounds having the largest structural effect ($h > h*$), which were identified using the Williams plot and employed in the model's development. With the use of the QSARINS programme, further statistical analyses, including the Insurbia plot and the Y-scrambling analysis plot, were completed and included in this article. Professor Gramatica's pioneering work—the ones that developed this QSARINS software—provides complete details on the methods used in these analyses [18].

Model quality assessment and validation

The following statistical metrics were used to assess the robustness, fitting criteria, predictive efficacy, stability, and reliability of the QSAR models created in this study. F test (Fischer's value) for statistical significance; R^2 (the squared correlation coefficient); the variables that are acquired by the randomised test are: α (the statistical importance parameter derived from the randomization test); $randR^2$ (the highest value of R^2 in the randomization test); pred R^2 (R^2 for the external test set); and CV R^2 (the cross-validated correlation coefficient).

R^2 is the regression coefficient, which is a relative indicator of how well the regression equation fits the data. $R^2 > 0.6$, $Q^2 > 0.6$, and $R^2pred > 0.5$ are the requirements that must be met for a QSAR model to be considered accurate [17][18]. The F-test displays the ratio of the variance accounted for by the regression error to the variance explained by the model. The model is statistically significant

when the F-test has high scores. The model's absolute fitness quality is demonstrated by the low standard errors of pred R^2se, Q^2se, and R^2se [22]. However, the QSAR model needs to meet the minimum recommended statistical values listed by Professor Gramatica's research guidelines to be considered acceptable [17][18].

RESULTS

It remains a fascinating scientific effort to find novel heterocycles with potency for numerous biological targets. The objective of the present investigation is to demonstrate how the variation in structure impacts the antimalarial efficacy and to build a QSAR model of novel 2-aminothiazole derivatives for their antimalarial activity. Considering IC_{50} values ranging from 0.08 µM to 47.8 µM, the 2-aminothiazole derivatives demonstrated remarkable antimalarial efficacy. To study and investigate the structural properties of 2-aminothiazole analogues, a QSAR model has been created. Its parameters were well-fit, and extensive validation had been done following the OECD regulatory requirements. Additionally, the applicability domain was established. We carried out ligand-based *in silico* investigations to understand the interaction between the structural characteristics and the biological activity of the compounds.

QSAR model development and validation

For our comparative QSAR modeling investigation, the 2-aminothiazole derivatives and their antimalarial activity (pIC_{50}) were utilized. Both the machine learning programme and the QSARINS software were used to produce an MLR-QSAR model. A total of 48 compounds having antimalarial activity against the chloroquine sensitive strain were gathered from the literature for our research. The PaDEL descriptor software was used to compute the molecular descriptors. For each chemical, it computed about 917 molecular descriptors.

While building the model using QSARINS, about 465 molecular descriptors with more than 80% constant values and more than 95% intercorrelated molecular descriptors were eliminated via the molecular descriptor step's processing. Two sets were created from the gathered data set: the training set (70%) and the test set (30%), also known as the prediction set. The training dataset (35 compounds) was used for QSAR model development, and the test/prediction dataset (13 compounds) was used for validation of the developed QSAR model. Whereas, when building the QSAR model using a machine learning programme, all 917 descriptors were utilised; later, following principle component analysis, the selected descriptors were used for the model equation.

Interpretation of result obtained via QSARINS software

The different statistical parameters for the generated QSAR models are listed in **Table 2**. The fitting criteria and internal and external validation parameters of the model are shown in **Table 3**. The correlation matrices between the chemical descriptors in the generated QSAR model are shown in **Table 4**. The generated QSAR model's applicability domain was determined with the help of the Williams plot. The Williams plot of the model's constituent compounds appears in Figure 4. In accordance with the Williams plot that describes the model's applicability domain, the model successfully predicted the activity of each compound in the dataset. For the QSAR model, the applicability domain was set within a squared area of leverage threshold of ± 2.5 standard residuals in the y-axis and h* = 0.300 in the x-axis.

Table 2 Model statistics and validation parameters

Variable	Coeff.	Std.Coeff.	Co.Int.95%
Intercept	13.9652		2.7841
VE1_Dzv	-3.5391	-0.4174	1.7340
ASP-7	-161.4155	-0.4786	70.9777
maxHaaCH	-1.2847	-0.2220	1.0671

It can be observed in **Table 2** that the model's most significant contributions are contributed by VE1_Dzv; ASP-7; maxHaaCH. Since these descriptors are positively associated with activity, adding more of them to a chemical would improve its bioactivity. Although the fitting parameters show that the model incorporates the individual contributions of the descriptors, the internal validation parameter Q^2loo shows that the value of 0.7890 is above the threshold value of 0.5, suggesting the capability of the model to analyse the activity of the compounds concerning the descriptors. The external validation value in Table 3 demonstrates that RMSEext = 0.6938, suggesting the model can accurately predict the chemicals

Table 4 Characteristic Fitting criteria with internal and external validation of the model

(Fitting criteria)				
R^2: 0.8414	R^2adj: 0.8195	R^2-R^2adj: 0.0219	CCC tr: 0.9139	
s: 0.3082	F: 38.4673			
(Internal validation criteria)				
Q2loo: 0.7890	R2-Q2loo: 0.0524	RMSE cv: 0.3284		
CCC cv: 0.8872	Q2LMO: 0.7597	R2Yscr: 0.1202		
(External validation criteria)				
RMSE ext: 0.6938	MAE ext: 0.5890	PRESS ext: 6.2586	R2ext: 0.0409	
Q2-F1: -0.5710	Q2-F2: -0.6379	Q2-F3: 0.0579	CCC ext: -0.1631	
r2m aver.: -0.0153	r2m delta: 0.0503			
Predictions by LOO:				
Exp(x) vs. Pred(y): R2: 0.7909	R'2o: 0.7577	k': 0.9992	Clos':0.0420	r'2m: 0.6469
Pred(x) vs. Exp(y): R2: 0.7909	R2o: 0.7891	k: 0.9992	Clos: 0.0024	r2m: 0.7568
External predictions by model equation:				
Exp(x) vs. Pred(y): R2: 0.0409	R'2o: -3.9251	k': 1.0143	Clos':97.0790	r'2m: -0.0405
Pred(x) vs. Exp(y): R2: 0.0409	R2o: -0.5354	k: 0.9794	Clos: 14.1065	r2m: 0.0098

that fall within its applicability domain but were not specifically employed in its development.

Table 3 The correlation matrices between the chemical descriptors

	VE1_Dzv	ASP-7	maxHaaCH
VE1_Dzv	1.0000		
ASP-7	-0.6093	1.0000	
maxHaaCH	-0.4391	0.3850	1.0000

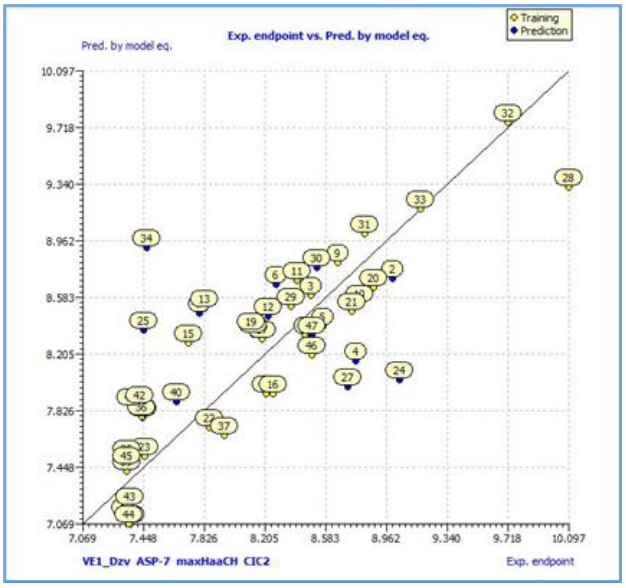


Figure 1 Scatter plot of experimental pIC50

The experimental activity of the MIC values is shown against the predicted activity (pIC_{50}) determined by the constructed QSAR model in **Figure 1**. The figure indicates that around 90% of the data were located along the plot's tread line. The QSAR model was developed using the data, which were scattered uniformly over the plot. The experimental value leave-one-out cross-validation plot is displayed in **Figure 2**. The figure not only demonstrates that the training and test sets were distributed equally over the tread line, but it also demonstrates how well the model was able to predict the experimental activity.

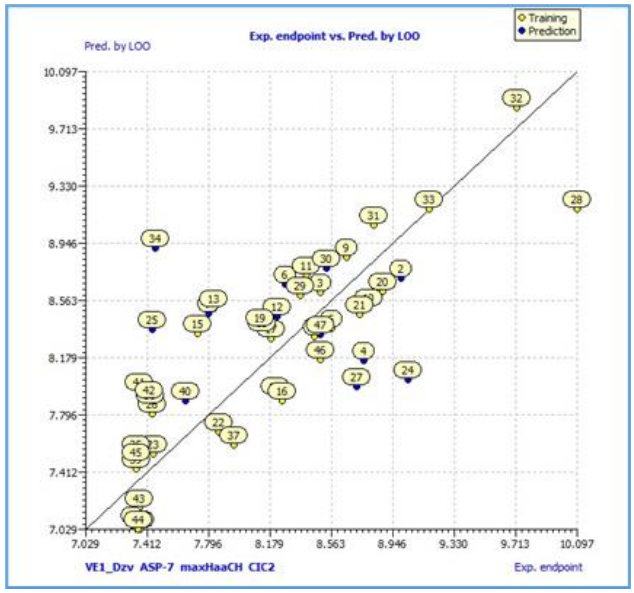


Figure 2 Scatter plot of experimental pIC50

Figure 3 shows the outlier detection graph and also notes the presence of low residual values between the observed bioactivities and predicted values, which should be within ±1. It also demonstrates the lack of compounds with a significant degree of error in their validation as

compounds. Although the chemical structures of these compounds differed from those of the remaining compounds in the dataset, the model was still able to predict all of them. The applicability domain plot of the Hat diagonal leverage value against the predicted model is displayed in **Figure 4**. It aids in identifying the QSAR model's outliers; it reveals the existence of no outliers with standard residuals greater than 2.5. The y-randomization plot is displayed in **Figure 5**, which demonstrates how error-free and reliable the model performed.

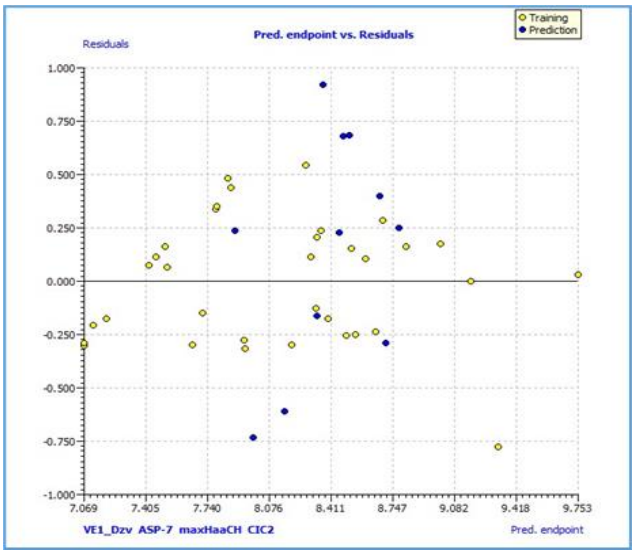


Figure 3 Residual plot of predicted endpoint versus standard residuals

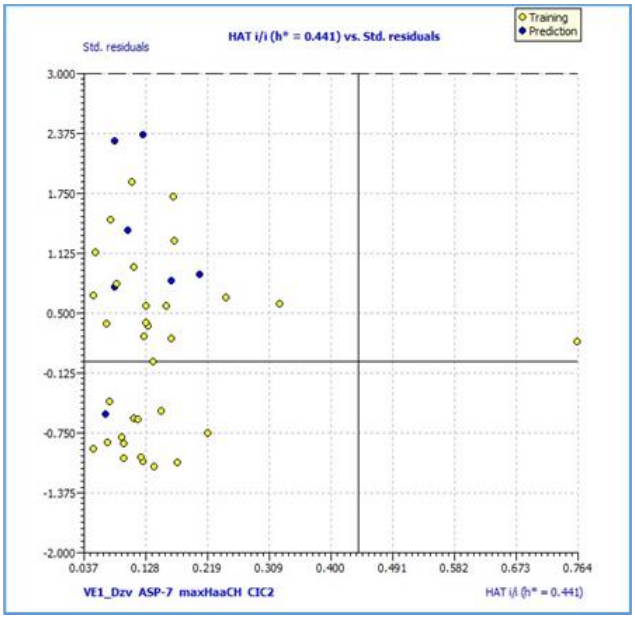


Figure 4 Applicability domain plot against the predicted endpoint

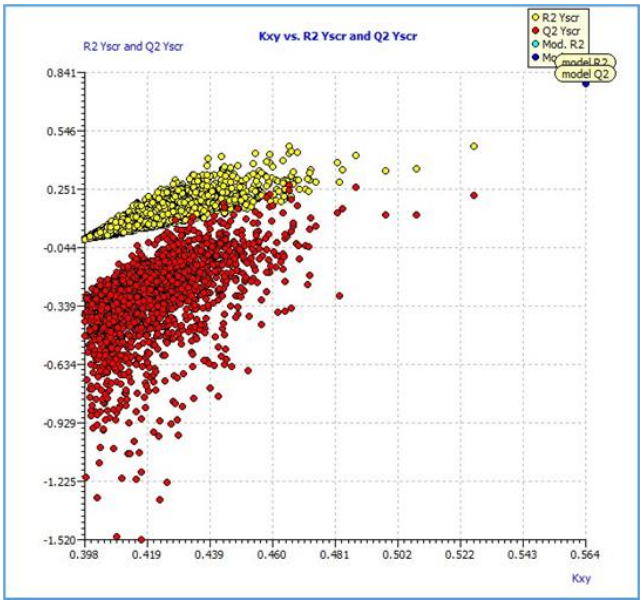


Figure 5 Scatter plot of Y-scrambled model compared with the original model

QSAR Mathematical formula derived from QSARINS software

pIC_{50} = 13.9652 -3.5391 (VE1_Dzv) -161.4155 (ASP-7) -1.2847 (maxHaaCH)

R^2: 0.8414; R^2adj: 0.8195; F: 38.4673; Q^2loo: 0.7890; RMSE cv: 0.3284; CCC cv: 0.8872; R^2Yscr: 0.1202; RMSE ext: 0.6938

DISCUSSION

We examined the effectiveness and applicability of the machine learning approach in medicinal chemistry by comparing it with other modern QSAR approaches, such as the machine learning approach, along with more conventional QSAR methods, like PLS and MLR. It has been shown that machine learning has remarkable prediction accuracy when the parameters are changed [23]. In contrast to other 3D-QSAR approaches, PLS and MLR are techniques used for manipulating huge quantities of data and facilitating the easy creation of models. The models in the present research were generated by carefully integrating the same set of data and descriptions. All of the chemicals in the collection had their bioactivity predicted using the QSAR model, and a Williams plot was used to determine the applicability domain. The OECD Work Programme on QSARs' principle [24] emphasises the critical necessity for an application domain to potentially find significant compounds and outliers.

We created QSAR models using QSARINS software that exhibited strong model fitting and adhered to the OECD guidelines for producing reliable predictive QSAR models. This model's Q^2value is 0.7890, indicating that it has good internal predictive potential. The created model's correlation coefficient (R^2 = 0.8414) indicated its quality of fit. The prediction power of the generated QSAR model was assessed using an external correlation coefficient (R^2adj = 0.8043). The suggested QSAR models are not the result of random correlation, as indicated by the lower R^2Yscr values. A minimum variance of 0.1202 (less than 0.3) in R^2 and R^2adj supported the significance of the number of chemical descriptors in the built QSAR model. The generated QSAR model's robustness was validated by the lowest R^2-Q^2loo variation of 0.0524 (less than 0.3).

Using this approach, we omitted a single compound, built a model using the leftover compounds, and then estimated the activity of the omitted compound. The leave-many-out (LMO) approach, which excludes 30% of the compounds in order to build the prediction model and examine the model's behaviour, is a subsequent, more potent method included in the QSARINS. Table 1 contains all of the model's statistical parameters. The model's regression coefficient of determination, called R^2, is 0.8414 and helps determine the model's quality of fit. The R^2 value of an acceptable model needs to be higher than 0.6.

The adjusted coefficient of determination, denoted R^2adj, provides information on the acceptability of adding new chemical descriptors to the QSAR model. R^2adj must be more than 0.6 for a model to be regarded as excellent. Friedman's lack of fit criteria, or LOF for short, evaluates overfitting in the QSAR model. A model's LOF must be less than 0.3 to be regarded as excellent. The total correlation between the descriptors is represented by Kxx. Delta K represents the correlation difference between the response (kxy) and the descriptor (kx), as well as the correlation difference between both of them. In the training set, RMSEtr refers to root-mean-square error. An acceptable model's root mean square error (RMSE) should be less than 0.3.

A model's performance is regarded as good if its mean absolute error (MAE), which is the correction calculated within the training set, is less than 0.3. The letter "S" represents the standard estimate error. A 'S' value of less than 0.3 is indicative of an effective model. Furthermore, it is important to minimise both R^2Yscr and Q^2Yscr to demonstrate that the model that was built was not the result of random correlation. All of the necessary statistical variables are satisfied by the model. The generated model's predictive ability may be measured using the Q^2-F1, Q^2-F2, and Q^2-F3 values.

The concordance correlation coefficient, abbreviated as CCC, needs to be close to 1 for a model to qualify as acceptable. The degree of similarity between the experimental and predicted results is measured by this CCC number and found to be close to 1 (**Table 3**). The most dependable and consistent variables are Q^2F3 and CCC. Fisher's F value, which should be greater than the theoretical threshold, is shown by the value of F. William's plot illustrated the selected model's applicability domain. It demonstrated that the majority of the compounds fell within the application domain of the model's chemical space. The antimalarial activity in the model is influenced by several molecular descriptors, including VE1_Dzv, ASP-7, and maxHaaCH. Of these, all three descriptors impacted negatively, and these contributed well to the antimalarial activity. Where the VE1_Dzv (Coefficient sum of the last eigenvector from Barysz matrix / weighted by van der Waals volumes) descriptor is related to the barysz matrix, ASP-7 (Average simple path, order 7) is related to the chi path, and maxHaaCH (Maximum atom-type H E-State:CH:) is related to the atom type electrotopological state.

CONCLUSIONS

A QSAR model has been created and compared using the QSARINS programme and a machine learning algorithm. Employing 48 potent 2-aminothiazole inhibitors of the MTB strain with 917 PaDEL descriptors showed that the three molecular descriptors, *viz*., VE1_Dzv, ASP-7, and maxHaaCH, are included in the model by QSARINS and have contributed the most to the model. R^2ext = 0.8414, the external validation parameter, was more than the minimal value needed to create a QSAR model. Additional findings, including the Williams plot and Insurbia data, helped to optimise the selection procedure for the further compounds that were used to create the more active molecules.

The findings of this study therefore supported our research hypothesis that the newly developed 2-aminothiazole derivatives might serve as a valuable "lead" molecule for the eventual development of antimalarial agent. It is advised that these compounds be subjected to further laboratory testing, including *in vitro* and *in vivo* analyses, or biological assessment. The model generated by QSAR methods given in this article satisfied every statistical requirement and confirmed the requirements established by the OECD principles. Hence, our article presents the characteristics (internal robustness, removal of chance correlation, external prediction, and applicability domain) for QSAR validation of the model and its ability to predict the activities of future compounds.

ACKNOWLEDGEMENTS

We wish to acknowledge Dr. Paolo Gramatica, who leads the QSAR Research Division at Insubria University in Italy and provided QSARINS software with an academic license.

Authors' contributions

RV framed the concept of this study; KS collected the data for our research; RV and KS drafted the manuscript; PT executed the QSAR model. All authors read and approved the manuscript.

Conflict of Interest

The authors declare no conflict of interest, financial or otherwise.

Funding Support

The authors declare that they have no funding for this study.

REFERENCES

[1] Rastelli G, Pacchioni S, Sirawaraporn W, Sirawaraporn R, Parenti MD, Ferrari AM. Docking and database screening reveal new classes of Plasmodium falciparum dihydrofolate reductase inhibitors. Journal of Medicinal Chemistry. 2003 Jul 3;46(14):2834-45.

[2] Hill DG. The biochemistry and physiology of Tetrahymena. Elsevier; 2012 Dec 2..

[3] Guha, R., & Jurs, P. C. (2004). Development of QSAR models to predict and interpret the biological activity of artemisinin analogues. Journal of chemical information and computer sciences, 44(4), 1440-1449.

[4] Katsila, T., Spyroulias, G. A., Patrinos, G. P., & Matsoukas, M. T. (2016). Computational approaches in target identification and drug discovery. Computational and structural biotechnology journal, 14, 177-184.

[5] Cherkasov, A., Muratov, E. N., Fourches, D., Varnek, A., Baskin, I. I., Cronin, M., ... &Tropsha, A. (2014). QSAR modeling: where have you been? Where are you going to?. Journal of medicinal chemistry, 57(12), 4977-5010.

[6] Prabha, T., Sivakumar, T. Design, Synthesis, and Docking of Sulfadiazine Schiff Base Scaffold for their Potential Claim as INHA Enoyl-(Acyl-Carrier-Protein) Reductase Inhibitors. Asian J Pharm Clin Res.,2018, 11(10), 233-237.

[7] Macalino, S. J. Y., Billones, J. B., Organo, V. G., & Carrillo, M. C. O. (2020). In silico strategies in tuberculosis drug discovery. Molecules, 25(3), 665.

[8] U., Safder, K., Nam, Kim, D., Shahlaei, M., and Yoo, C. (2018). Quantitative structure-property relationship (QSPR) models for predicting the physicochemical properties of polychlorinated biphenyls (PCBs) using deep belief network. Ecotoxicology and environmental safety, 162, 17-28.

[9] Siddiqui N, Arya SK, Ahsan W, Azad B. Diverse biological activities of Thiazoles: A Retrospect. Int J Drug Dev Res. 2011 Oct;3(4):55-67.

[10] Prabha T, Aishwaryah P, Manickavalli E, Chandru R, Arulbharathi G, Anu A, Sivakumar T. A Chalcone Annulated Pyrazoline Conjugates as a Potent Antimycobacterial Agents: Synthesis and in Silico Molecular Modeling Studies. Research Journal of Pharmacy and Technology. 2019;12(8):3857-65.

[11] Ozdemir A, Turan-Zitouni G, Kaplancikli ZA, Revial G, Güven K. Synthesis and antimicrobial activity of 1-(4-aryl-2-thiazolyl)-3-(2-thienyl)-5-aryl-2-pyrazoline derivatives. Eur J Med Chem. 2007 Mar;42(3):403-9.

[12] Cohen A, Verhaeghe P, Crozet MD, Hutter S, Rathelot P, Vanelle P, Azas N. Tandem synthesis and in vitro antiplasmodial

evaluation of new naphtho [2, 1-d] thiazole derivatives. European journal of medicinal chemistry. 2012 Sep 1;55:315-24.

[13] González Cabrera D, Douelle F, Feng TS, Nchinda AT, Younis Y, White KL, Wu Q, Ryan E, Burrows JN, Waterson D, Witty MJ. Novel orally active antimalarial thiazoles. Journal of medicinal chemistry. 2011 Nov 10;54(21):7713-9.

[14] Gramatica, P., &Sangion, A. (2016). A historical excursus on the statistical validation parameters for QSAR models: a clarification concerning metrics and terminology. Journal of chemical information and modeling, 56(6), 1127-1131.

[15] Roy, K., & Narayan Das, R. (2014). A review on principles, theory and practices of 2D-QSAR. Current drug metabolism, 15(4), 346-379.

[16] Mulatsari E, Mumpuni E, Nurhidayati L, et al. Chemical molecular visualization training with chemsketch software for senior high school students. JurnalPengabdianMasyarakat 2021;2(1):102‐112.

[17] Gramatica, P. (2007). Principles of QSAR models validation: internal and external. QSAR & combinatorial science, 26(5), 694-701.

[18] Gramatica, P., Chirico, N., Papa, E., Cassani, S., &Kovarich, S. (2013). QSARINS: A new software for the development, analysis, and validation of QSAR MLR models.Journal of Computational Chemistry,34 (24), 2121‐2132.

[19] González Cabrera D, Douelle F, Feng TS, Nchinda AT, Younis Y, White KL, Wu Q, Ryan E, Burrows JN, Waterson D, Witty MJ. Novel orally active antimalarial thiazoles. Journal of medicinal chemistry. 2011 Nov 10;54(21):7713-9.

[20] Mjambili F, Njoroge M, Naran K, De Kock C, Smith PJ, Mizrahi V, Warner D, Chibale K. Synthesis and biological evaluation of 2-aminothiazole derivatives as antimycobacterial and antiplasmodial agents. Bioorganic & medicinal chemistry letters. 2014 Jan 15;24(2):560-4.

[21] Roy, K., & Mitra, I. (2011). On various metrics used for validation of predictive QSAR models with applications in virtual screening and focused library design. Combinatorial chemistry & high throughput screening, 14(6), 450-474.

[22] Papa, E., Dearden, J. C., &Gramatica, P. (2007). Linear QSAR regression models for the prediction of bioconcentration factors by physicochemical properties and structural theoretical molecular descriptors. Chemosphere, 67(2), 351-358.

[23] Jayaprakash, V., Saravanan, T., Ravindran, K., Prabha, T., Selvaraj, J., Jayapalan, S., ... & Sivakumar, T. (2023). Relevance of Machine Learning to Predict the Inhibitory Activity of Small Thiazole Chemicals on Estrogen Receptor. Current Computer-Aided Drug Design, 19(1), 37-50.

[24] OECD (2014), Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models, OECD Series on Testing and Assessment, No. 69, OECD Publishing, Paris, https://doi.org/10.1787/9789264085442-en.