ASEC 2024 Conference

The 5th International Electronic Conference on Applied Sciences

04–06 December 2024 | Online

MDPI

Modelling the Quantitative Structure-Activity Relationships of 1,2,4-Triazolo[1,5-a]pyrimidin-7-amine Analogues in the Inhibition of *Plasmodium falciparum*

Inalegwu S. Apeh[1,2], Thecla O. Ayoka[3], Charles O. Nnadi[1], Wilfred O. Obonga[1]*

[1]Department of Pharmaceutical and Medicinal Chemistry, Faculty of Pharmaceutical Sciences, University of Nigeria Nsukka, 410001 Enugu Nigeria

[2]Department of Pharmacy, Benue State Hospitals Management Board, 972261 Otukpo Benue State, Nigeria

[3]Department of Science Laboratory Technology (Biochemistry Unit), Faculty of Physical Sciences, University of Nigeria, Nsukka, 410001, Enugu State, Nigeria

INTRODUCTION & AIM

Several measures such as chemotherapy, vaccine development and environmental (insecticide-treated mosquito nets and indoor residual spraying) interventions have been deployed to prevent or treat malaria; some of which have faced numerous challenges. This calls for a clear need to adopt machine-learning models to predict the antimalarial activity of compounds and optimize the available ones for better alternatives. Triazolopyrimidine represents an important scaffold in medicinal chemistry drug research; combining the dual functionality of triazoles and pyrimidines as well as the multiple isomeric potential. Apart from its antiprotozoal activity potential, several other activities such as antibacterial, antihypertensive, anti-tuberculosis, antifungal, herbicidal anti-inflammatory and antitumor have also been attributed to this pharmacophore. The majority of the triazolopyrimidines known today have been synthetically obtained. Specifically, essramycin isolated from *Streptomyces* species has remained one of the few existing naturally occurring (1,2,4)-triazolo[1,5-a]pyrimidine antibiotic leaving semi-synthetic as the only source of antiplasmodial agents. Importantly, the (1,2,4)-triazolo[1,5-a]pyrimidine analogues are unique due to the presence of the heterocyclic ring as a bioisostere candidate of purines scaffolds, carboxylic acid and *N*-acetylated lysine. Against these backdrops, the study modelled the QSAR of 125 derivatives of (1,2,4)-triazolo[1,5-a]pyrimidine using 306 molecular features representing different functionalities that provide necessary insights into the mechanism of antiplasmodial activity and guide synthetic chemist in harnessing and deploying the predictive potential of machine learning-based modelling

METHOD

The dataset, obtained from the ChEMBL database, was made up of 125 molecules of the 1,2,4-triazolo[1,5-a]pyrimidin-7-amines with their corresponding endpoints (pIC_{50}) values tested against *P. falciparum*. The 306 molecular features of the compounds were obtained from the Padel software representing both 2D and 3D descriptors.

Recursive feature elimination (RFE) was used to drop insignificant molecular features automatically. Linear regression function from Sklearn for RFE was used. The number of features to be considered to build the model was placed at 11 using $m > n^2$ where m is the number of molecules, and n is the number of features. The features marked 'True' were selected. The Statsmodel was used to get the detailed statistics and summary of the model. Features with $p < 0.05$ are significant and were selected for the model. Variance inflation factor (VIF) was also used to select features that fall within the acceptable range.

The X- and Y-matrix were split into train sets (99 molecules) and test sets (25 molecules), using a split ratio of 0.20, where 80 % is assigned to the train set and 20 % is assigned to the test set. The size of the training dataset was denoted as X-train, Y-train, while the size of the test dataset was X-test, Y-test. The training set was used to train the model, while 25 molecules belonging to the test set were used to validate the models. The models were trained on the training set using the fit method. The hyper-parameters of the models were adjusted on the test dataset to obtain the best hyper-parameter configuration. This was done using a random search because their hyper-parameters were continuous.

Residue analysis of the error terms was checked, to ascertain their normal distribution. The histogram of the error term was plotted. Normal distribution is one of the major assumptions of multiple linear regression.

Six machine learning sci-kit-learn algorithms: Multiple Linear Regression (MLR), *k*-Nearest Neighbours (*k*NN), Support Vector Regressor (SVR), Random Forest Regressor (RFR) RIDGE regression and LASSO were deployed to establish the relationship between the actual and predicted pIC_{50} values of the molecules.

Different evaluation metrics; coefficient of determination (R^2), mean squared error (MSE), mean absolute error (MAE) and root mean squared error (RMSE) were deployed to assess the performance of the models. The *p*-values, F-statistic, and variance inflation factor (VIF) were also used.

RESULTS & DISCUSSION

The chemical data set used for this study consisted of 125 triazolopyrimidin-7-amine analogues substituted at positions 2, 5 and 7-NH_2 (**Figure 1**).

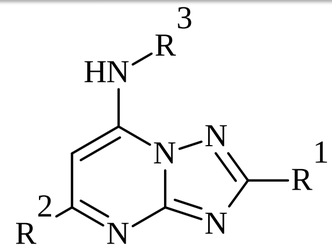

Figure 1. Chemical structure of triazolopyrimidin-7-amine template

The number of significant variables to be considered to build the model was placed at 11 following the RFE analysis. The following variables (npr1, PEOE_VSA_FHYD, PEOE_VSA_FPNEG, PEOE_VSA_FPOL, PEOE_VSA_FPPOS, pmi3, Q_VSA_FNEG, Q_VSA_FPOS, slogP, vsurf_CW2, and vsurf_W2) were selected by RFE and are ranked 'True'. They were, therefore, considered significant for the model. The Statsmodel function was further used to check the detailed statistics and summary of the model from the selected variables using the RFE. Variables with $p < 0.05$ were considered significant for the machine learning (ML) modelling of the QSAR. The significant variables were npr1, pmi3, slogP, vsurf_CW2, and vsurf_W2. The summary of the Statsmodel is represented in **Table 1**

Table 1. Selected significant features following RFE

Features	Coeff	SE	t	p > |t|	[0.025 - 0.975]	
constant	5.8964	0.060	98.522	0.000	5.778	6.015
npr1*	-0.7146	0.085	-8.360	0.000	-0.884	-0.545
pmi3*	-1.5210	0.281	-5.407	0.000	-2.078	-0.964
SlogP**	0.8752	0.094	9.349	0.000	0.690	1.061
vsurf_CW2*	-0.5733	0.204	-2.808	0.006	-0.978	-0.169
vsurf_W2*	1.1120	0.312	3.570	0.001	0.495	1.729

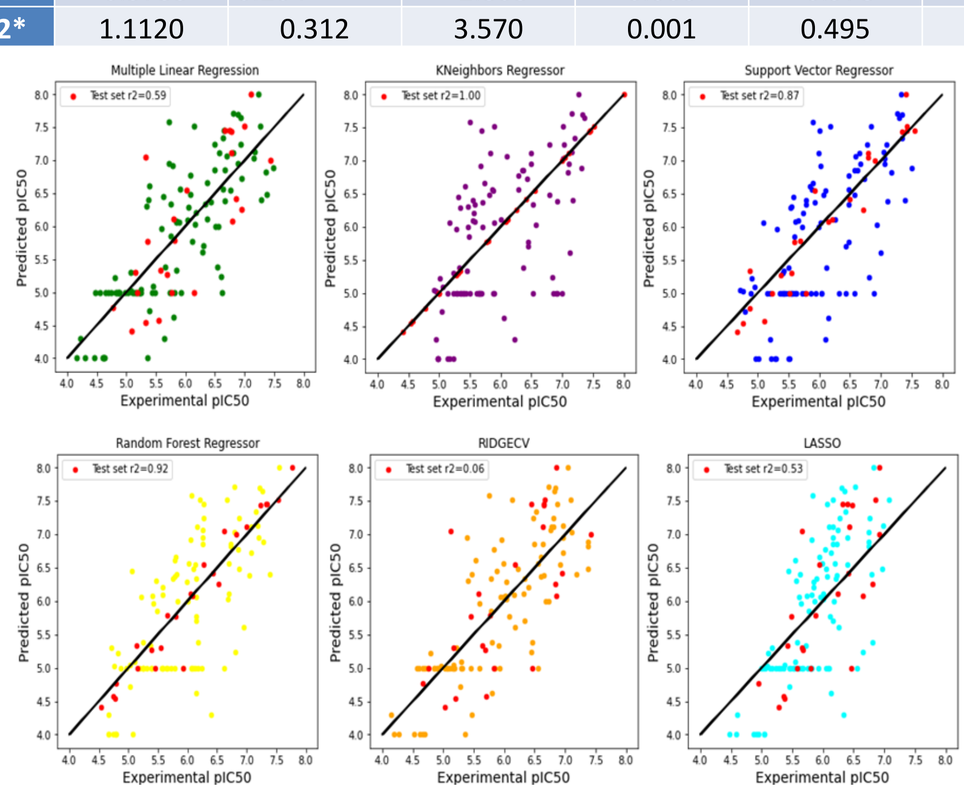

Figure 2: Correlation of the predicted and experimental pIC_{50} values.

The correlations of the predicted and experimental pIC_{50} values are shown in **Figure 2**. The R^2 indicates how closely the data resemble the regression line and how well the data fit the regression line. The regression equation is $pIC_{50} = 5.90 – 0.71npr1 – 1.52pmi3 + 0.88slogP – 0.57vsurf\text{-}CW2 + 1.11vsurf\text{-}W2$

CONCLUSION

The study developed a model with a predictive regression (SVR) equation of $pIC_{50} = 5.90 – 0.71npr1 – 1.52pmi3 + 0.88slogP – 0.57vsurf\text{-}CW2 + 1.11vsurf\text{-}W2$. This provides insights into understanding a single target mechanism of antiplasmodial activity of 1,2,4-triazolo[1,5-a]pyrimidin-7-amine analogues

FUTURE WORK / REFERENCES

Further work could involve exploring hybrid models that combine SVR with feature engineering techniques to enhance prediction accuracy. Expanding the dataset or integrating additional molecular descriptors could also improve model robustness, particularly for neural networks or deep learning approaches that typically benefit from large datasets.

Yap CW. PaDEL-descriptor: An open-source software to calculate molecular descriptors and fingerprints. J. Comput. Chem. 2011; 32(2011): 1466–1474. https://doi.org/10.1002/jcc.21707