Research Article

Journal of Chemistry Letters
journal homepage: www.jchemlett.com
ISSN (online) 2717-1892 (print) 2821-0123

In-silico modeling of inhibitory activity and toxicity of some indole derivatives towards designing highly potent dengue virus serotype 2 NS4B inhibitors

Samuel Ndaghiya Adawara [a,]*, Gideon Adamu Shallangwa [b], Paul Andrew Mamza [b] , Ibrahim Abdulkadir [b]

[a] *Department of Pure and Applied Chemistry, Faculty of Science, University of Maiduguri, P.M.B. 1069, Maiduguri, Borno State, Nigeria*

[b] *Department of Chemistry, Faculty of Physical Sciences, Ahmadu Bello University, P.M.B. 1044, Zaria, Kaduna State, Nigeria*

ARTICLE INFO

Article history:
Received 10 April 2022
Received in revised form 15 June 2022
Accepted 23 June 2022
Available online 25 June 2022

Keywords:
QSAR
Descriptors
MLR
GFA
ALogP

ABSTRACT

The global prevalence of dengue virus (DENV) infection has become a source of great concern to humanity. As such, infection, if left untreated, could progress to a life-threatening stage called dengue hemorrhagic fever or dengue shock syndrome. A large percentage of the world's population could be at risk of being infected by the dengue virus. The DENV NS4B receptor is essential in viral replication and hence could in principle be suitable as a therapeutic target in the treatment of dengue viral infection. The augmentation of existing agents that could inhibit the dengue virus is important. In this research, various classes of molecular descriptors were generated. Quantitative structure-activity relationship studies (QSARs) have been conducted to correlate the molecular properties of some indole derivatives with their anti-dengue activity and toxicity. The inhibitory activity and toxicity prediction models were statistically valid and robust, with acceptable statistical validation factors such as predicted $R^2_{pred.}$, adjusted $R^2_{adj.,}$ cross-validated Q^2 and R^2 regression coefficient, etc. ($R^2_{pred.} = 0.64448$, $R^2_{adj.} = 0.59223$, $cR_p^2 = 0.57134$, $Q^2_{CV} = 0.64448$, $R^2 = 0.63201$) and ($R^2_{pred.} = 0.81813$, $R^2_{adj.} = 0.56015$, $cR_p^2 = 0.5386$, Q2CV = 0.50548, $R^2 = 0.60645$), respectively. The models revealed that the average Broto-Moreau autocorrelation-lag 7/weighted by first ionization potential (AATS7i), number of hydrogen bond acceptors (nHBAcc) for activity and 3D topological distance-based autocorrelation-lag 9/weighted by van der Waals volumes (TDB9v) descriptors were found to strongly influence the anti-dengue biological activity (pEC_{50}) and toxicity (pCC_{50}) of the indole derivatives, respectively. The indole derivatives were predicted to be orally bioavailable with excellent gastrointestinal absorption (94.044–90.219%). The DENV-2 NS4B inhibitory activity, as well as the cytotoxicity of indole derivatives with no experimental data, could be predicted with high precision using the models developed, which could further lead to a cut in experimental cost as well as the design of highly potent and less toxic derivatives.

1. Introduction

Dengue virus (DENV) infection results from diseased mosquito bites, specifically the female *Aedes aegypti or Aedes albopictus,* of the Aedes genus due to a virus called dengue virus [1][2]; an associate of the *Flavivirus,* primarily, found within the tropical and sub-tropical areas around the world [3]. The alarming prevalence of this mosquito-borne viral dengue infection calls for global concern. Every year, about 390 million dengue infections are recorded globally with a large percentage of the incidences taking place in the tropical

* Corresponding author. Tel.: +2348067811759; e-mail: agapalawa@gmail.com

and subtropical region, with about a quarter of the cases developing clinical symptoms [4].

Dengue virus infection is associated with symptoms such as fever, joint pain and in some cases could worsen to a severe stage of the infection that could lead to hemorrhagic fever or shock syndrome; a more devastating stage of dengue infection that could lead to death [5][6][7].

There is still no licensed antiviral drug for the cure of dengue virus disease in the face of the widespread of the disease and the endangerment associated with it [8] [9] [10].

The dengue virus is classified into four serotypes (serotypes 1-4), with each of the serotypes having seven non-structural (NS) proteins (NS1, NS2A-2B, NS3, NS4A-NS4B, and NS5). Infection caused by any of the four serotypes is lethal generally, but infection caused by serotype 2 is associated with the most devastating outcomes such as hemorrhagic fever and dengue shock syndrome. [11][12].

The significance of NS4B for viral replication gives it credence as a model target for developing anti-dengue virus agents for the treatment of diseases instigated by *Flavivirus* [13]. Due to its high hydrophobicity, neither the crystal nor NMR structure of Flavivirus NS4B is currently available [13].

In a particular study, a 3-acyl indole derivative was identified from a phenotypic screening using a DENV-2 induced CPE assay. It was revealed that this class of compound resulted in multiple mutations in the NS4B protease of the virus. Furthermore, it was concluded that such potency is associated with the methoxy group in the meta position of the aniline moiety[13].

It has been reported that clinical development of dengue inhibitors has been hindered due to poor ADMET (absorption, distribution, metabolism, excretion, and toxicity) in animal model [14] [15]. Studies on the indole derivatives were also carried out to ascertain the drug-likeness, bioavailability as well as the toxicity of the this important class of compounds with high therapeutic potential against DENV NS4B that is essential in the viral replication cycle, thus avoiding drug development failure.

This research was targeted at building robust QSAR models for predicting the anti-dengue activity and toxicity of some indole derivatives as dengue virus serotype 2 NS4B inhibitors, as well as predicting the inhibitory activity and toxicity of newly designed or synthesized indole derivatives.

The findings of this study will be used to provide structural information for the development of potent anti-dengue virus agents with lower toxicity.

2. Methods

2.1 Data collection

Some indole derivatives were obtained from recently published scientific literature [13], with reported experimentally measured DENV-2 NS4B inhibitory biological activity expressed as the concentration of the compounds where 50% of their maximal effect were observed (EC_{50}), as well as their 50% cytotoxic concentration (CC_{50}).

2.2 The response variables (Biological activities)

The obtained response values (EC_{50} and CC_{50}) of the indole derivatives reported in micro-Molar units (μM) were converted to a Molar unit and subsequently transformed to their respective logarithm units (pEC_{50}) and (CC_{50}) with the aid of Equation 1 (Eq.1) to obtain a normal statistical distribution of the values [16]. The chemical structures of the indole derivatives are presented in Supplementary Table 1(Table SM1).

$$pEC_{50}= -\log (EC_{50}) \qquad (1)$$

2.3 Molecular geometry optimization and descriptors generation

Molecular structures of the indole derivatives in Table SM1 were sketched with Chemdraw and successively optimized to obtain their equilibrium geometries at ground state with density functional method (DFT/B3LYP/6-31G*). This is normally carried out to achieve conformation with the lowest stable energy [17]. The molecular descriptors (0D-3D) which are the numerical expression of information entrenched in any chemical structure were generated using the paDel-Descriptor software tool and combined with the Spartan 14 V1.1.4 software-generated descriptors [18].

2.4 Dataset pretreatment

The generated descriptors of all the compounds were in an Excel file containing the respective numerical values of the descriptors. In this, all descriptor columns having a constant or null value were all deleted, as well as those descriptors having a correlation coefficient of more than 0.8.

The pretreatment is targeted at removing redundant values to facilitate the generation of a robust model [16] [19].

2.5 Dataset division for model building

The sets of the compounds were grouped into a training set for building the model and a test set for testing the predictive power of the model. Dataset Division software package was used for the division [20]. About 70 % of the compounds were used for the model building and 30 % for testing the predictive power of the models in each case [21].

2.6 MLR-GFA Model building

The models for both inhibitory activity and toxicity were developed from their respective training set of the compounds using multiple linear regression (MLR) statistical method of the genetic function algorithm (GFA) implemented in Material Studio. This algorithm selects the best arrangements of descriptors that best describe the variation in the bioactivity of the compounds. This method has the capacity of generating several collections of descriptors that characterized a model. It also uses a lack-of-fit (LOF) function to identify over-fitting and moderate redundancy [21] in a model. The lesser the LOF value, the better the quality of the model, and it is evaluated using the mathematical expression:

$$LOF = \frac{LSE}{\left(1-((c+ dp)/M)\right)^2} \qquad (2)$$

From Eq. 2, the numerator represents the model's error of least squares, while c, d, p, and M in the denominator represent the number of basic terms, smoothing factor, model's descriptors sum, and training compounds involved in building the model respectively [22].

2.7 Model quality tools and assessment

Statistical validation parameters of the built models were appraised to determine the models' fitting ability, consistency, predictive capability, and robustness [23].

The quality of a developed model is satisfactory if the results agree with the generally satisfactory QSAR standard threshold value suggested in Table 1 [24].

2.8 QSAR model validation

After developing the models, internal validations were performed to select the respective initial QSAR models, passing the required criteria are the internal validations; the predictive capability of the models are evaluated by the external validations using the test sets.

The various internal and external validation parameters that characterize the models' robustness were evaluated with the use of MLRplusValidation and Y-Randomization software [20], which evaluates R^2 (square correlation coefficient), Q^2_{CV} (cross-validation coefficient), R^2 pred. (external test set correlation coefficient) and cRp2 (coefficient of determination for Y-Randomization) which were all obtained from the package.

2.8.1 Multi-collinearity detection test

The presence of over correlation between the descriptors was examined with a factor called variance inflation factor (VIF) value for the respective descriptor in the model is evaluated with the expression in Eq. 3:

$$VIF_i = \left(\frac{1}{1-R_{ij}^2}\right) \qquad (3)$$

From Eq. 3, R^2_{ij} represents the correlation value of the multiple regression of particular descriptor i concerning others j within the model [25].

2.8. 2 Model's applicability domain analysis

The prediction space or region of the model called the domain of applicability was examined using the extrapolation method [26]. It is significant in checking the use of dissimilar compounds in developing the model that could lead to the prediction of biological activity of compounds that is out of their domain [27]. This method involves the use of the compounds leverage (hi) values and standardized residual (SDR) of the model [28]. The leverages hi are evaluated as the transverse component of the matrix H:

$$H = R.(R^T R)^{-1}.R^T \qquad (4)$$

From Eq. 4, R represents the matrix column of the descriptors and its transpose by R^T and SDR was determined as follows:

$$SDR = \frac{\hat{y}- y}{\sqrt{\frac{\sum_{i=1}^{n}(\hat{y}- y)^2}{m}}} \qquad (5)$$

From Eq. 5, the observed and predicted activities are denoted by y and $\hat{y}$ respectively, while m denotes the training or test number of compounds. The AD of the model is set to predict within a boundary limit of $0 < h_i < h^*$ and an SDR limit of ±3. The cautionary leverage h* is computed by Equation 6:

$$h^* = \frac{3.(j+1)}{m} \qquad (6)$$

From Eq. 6, j symbolizes the model's descriptors amount, while m represents the number of compounds used in building the model. A graphical illustration of the AD is a plot of SDR against h_i (leverages) known as William's plot was obtained [29].

2.8.3 Mean effect

Every descriptor in the models generated has its level of contribution in the prediction of the biological activity concerning other descriptors within the models, such levels of contribution by each of the descriptors relative to other was evaluated by calculating a parameter called mean effect represented by Eq. 7.

From Eq. 7, MEj is the mean effect of a specific descriptor j, whereas bj is the constant of the descriptor j, Rij indicates the descriptors numerical of respective molecule, and m is the total number of descriptors in the model [16].

$$MEj = \frac{\beta_j \sum_{i=1}^{i=n} R_{ij}}{\sum_{j}^{m}.(\beta_j \sum_{i}^{n} R_{ij})} \qquad (7)$$

2.9 In silico ADMET prediction

The ADMET parameters of some indole derivatives were assessed with the use of the Swiss-ADME [15] and pkCSM – pharmacokinetics [30] free online tools to overcome challenges associated with poor pharmacokinetics.

Compounds 12, 33, 37, 45, 51, 53, and 101 were selected based on their antiviral potency for a detailed evaluation of their *in silico* ADMET properties.

3. Results

3.1 QSAR model quality

Using a multiple linear regression of genetic algorithm, QSAR models for the predictions of both activity and toxicity of indole derivatives were developed. Each of the models contains four (4) descriptors. Eq. 8 and 9, respectively represent the models:

Activity model

pEC_{50} = 2.30487(+/-1.27191) -0.34692(+/-0.13721) **ALogP** + 5.68264(+/-1.32172) **MATS7c** -0.57907(+/-0.11402) **nHBAcc** +0.00425(+/-0.00099) **TDB9v** (8)

Where, $R^2_{adj.}$ = 0.59223, $R^2_{pred.}$ = 0.64448, cR_p^2= 0.57134, Q^2_{CV} = 0.64448, N_{train} =41, N_{set} =20, R^2 = 0.63201

Toxicity model

pCC_{50} = 19.39932(+/-4.11029) +0.12922(+/-0.04698) **ALogP** -0.09695(+/-0.02659) **AATS7i** +0.17277(+/-0.04282) **SM1_Dzs** -0.1827(+/-0.05436) **RPCS** (9)

Where, $R^2_{adj.}$ = 0.56015, $R^2_{pred.}$ = 0.81813, cR_p^2= 0.5386, Q^2_{CV} =0.50548, $N_{train\ set}$ =38 N_{test} set =18, R^2 = 0.60645

N is the number of compounds, R^2 is the squared correlation coefficient, R^2adj signifies the adjusted R^2 while Q^2_{CV} represents the leave-one-out cross-validation value.

Eq. 8 and 9 contain four descriptors each that are most significant to the activity (pEC_{50}) and toxicity (pCC_{50}) which includes ALogP, AATS7i, SM1_Dzs, RPCS and ALogP, MATS7c, nHBAcc, TDB9v respectively.

Figures 1 and 2 depict the plots of predicted activity (pEC_{50}) and toxicity (pCC_{50}) against experimental activity (pEC_{50}) and toxicity (pCC_{50}), respectively. It could be observed from both plots (Figures 1 and 2) that there is close agreement between the predicted (pEC_{50}) and (pCC_{50}) of the test sets and those of the train sets in each case.

The numerical values of the calculated descriptors involved in developing both activity and the toxicity models are presented in Supplementary Tables 2 and 3 (Tables SM2 and SM3) as well as the residual values of their predictions. The presence of low residual values entails a good predictive capacity of the models.

The statistically recommended validation factors for the acceptable QSAR model as well as the validation factors for the developed activity and toxicity models are presented in Table 1, which shows the validity of the models as such built models conformed to all the required validation factors for acceptability.

The Y-randomization test results for both models (pEC_{50} and pCC_{50}) are presented in Tables 1, SM4, and SM5 signifying robust models demonstrated by Y-randomization parameters.

The Williams' plots for identifying the region of applicability of the activity and the toxicity models' predictions are represented by Figures 3 and 4 respectively, with most compounds having leverage values less than the critical leverage values and within the required standardized residual value of ±3.

The complete explanation of the descriptors in each of the models, as well as their reliability in terms of chance correlation and degree of contribution, are presented in Tables 2 and 3. From Tables 2 and 3, both models are having variation inflation factors below the value of five (5), which is indicative of the absence of chance correlation.

The absence of systematic error in the development of the models is also supported by the random distribution of the variables within the plots represented by Figures 5 and 6. The results of the predicted ADMET of some selected derivatives (12, 33, 37, 45, 51, 53 and 101) of the indole are presented in Table 4.

4. Discussion

4.1 QSAR model predictive quality analysis

QSAR models for the prediction of biological activity (pEC_{50}) and toxicity (**pCC_{50})** were built from the set of 61 indole derivatives using the GA-MLR statistical method.

The models were developed from the training set using the multiple linear regression statistical method of the Genetic algorithm. About 30 % of the total number of compounds called the test set was used to validate the predictive power of the models in each case.

In each of the models, four descriptors were found to be associated with the activity (pEC_{50}) and toxicity (pCC_{50}) of the compounds. These descriptors include ALogP, AATS7i, SM1_Dzs, RPCS and ALogP, MATS7c, nHBAcc, TDB9v, respectively. The models for the anti-dengue activity and toxicity predictions reported in the study are represented by Eq. 8 and 9, respectively, characterized by the following recommended QSAR model validation factors, ($R^2_{adj.}$ = 0.59223, $R^2_{pred.}$ = 0.64448, cR_p^2= 0.57134, Q^2_{CV} = 0.64448, R^2 = 0.63201) and ($R^2_{adj.}$ = 0.56015, $R^2_{pred.}$ = 0.81813, cR_p^2= 0.5386, Q^2_{CV} = 0.50548, R^2 = 0.60645), respectively.

The models were used to predict the activity (pEC_{50}) and toxicity (pCC_{50}) of both training and test reported in Tables SM2 and SM3 respectively.

The graph of predicted activity (pEC_{50}) and toxicity (pCC_{50}) against their respective experimental activity (pEC_{50}) and toxicity (pCC_{50}) of the training set for the models (Figures 1 and 2) showed that a direct relationship exists between the two variables and the models had good internal extrapolation fitness [16][19].

In addition, Figures 1 and 2 showed the predicted activity (pEC_{50}) and toxicity (pCC_{50}) of the train and test sets against their experimental activity (pEC_{50}) and toxicity (pCC_{50}). It could be seen from Figures 1 and 2 that the calculated activity (pEC_{50}) and toxicity (pCC_{50}) of the training were in good agreement with those of the test sets in each case, respectively.

Statistical validation factors [24][27] that characterize an acceptable QSAR model are presented in Table 1. The statistical validation factor obtained for the developed models (Equations 8 and 9) presented in Table 1 is excellently acceptable based on the recommended standard of QSAR model acceptability.

The data obtained show that the predictive effectiveness of the models is very good since R^2 greater than 0.5 shows that if these models are evaluated by test data, the accuracy of the prediction could be very reliable, which confirms the validity of the obtained results [16][19].

Detailed of the statistical validation parameters computed for the model presented in Table 1 showed that values for $R^2_{adj.}$ = 0.59223, $R^2_{pred.}$ = 0.64448, cR^2_p= 0.57134, Q^2_{CV} = 0.64448, R^2 = 0.63201 for the activity model and $R^2_{adj.}$ = 0.56015, $R^2_{pred.}$ = 0.81813, cR^2_p= 0.5386, Q^2_{CV} = 0.50548, R^2 =0.60645 for the toxicity model, respectively, are all within the recommended threshold values as suggested in Table 1 [24] [27]. Consequently, the models had outstanding internal and external prediction ability and it is not a product of coincidental correlation [16]. The model also passed all the recommended benchmarks for a predictive model [24][27].

The result of the evaluation for Q^2_{CV} is 0.64448 and 0.50548 respectively; higher Q^2_{CV} value evidences the ability of the presented models in support of their respective internal validation.

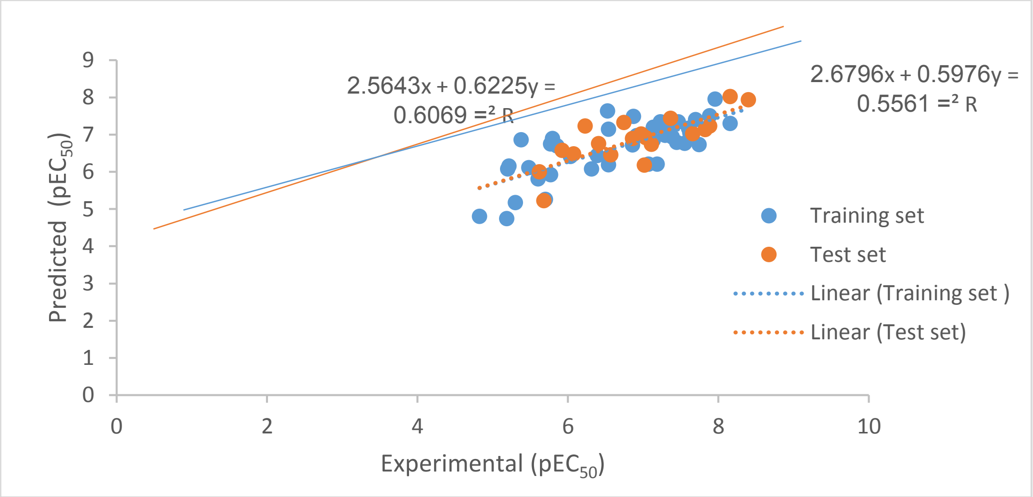

Figure 1. Graphical illustration of predicted against experimental activity by GA-MLR (anti-dengue activity).

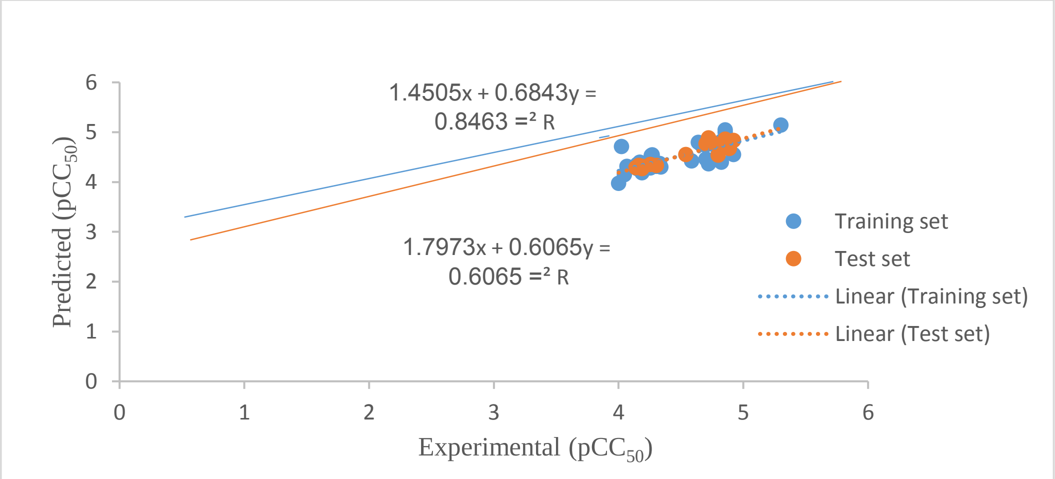

Figure 2. Graphical representation of predicted against experimental toxicity MLR

Table 1. Model authentication factors for a generally satisfactory QSAR model

Parameter	**Equation**	**Threshold Score**	pEC50 (Activity) **Model score**	pCC50 (Toxicity) **Model score**	**Remark**
		Internal validation			
$\mathbf{R^2}$	$\frac{\left[\sum\left\{(\mathrm{Y}-\overline{\mathrm{Y}})\times\left(\hat{\mathrm{Y}}-\overline{\hat{\mathrm{Y}}}\right)\right\}\right]^2}{\sum\left(\mathrm{Y}-\hat{\mathrm{Y}}\right)^2\times\sum\left(\hat{\mathrm{Y}}-\overline{\hat{\mathrm{Y}}}\right)^2}$	$R^2 > 0.6$	0.63201	0.60645	Good
$\mathbf{R^2_{adj}}$	$\frac{(N-1)\times R^2 - p}{N-1-p}$	$R^2_{adj} > 0.6$	0.59223	0.56015	Good
$\mathbf{Q^2}$	$1-\frac{\sum\left(\mathrm{Y}-\hat{\mathrm{Y}}_{loo}\right)^2}{\sum(\mathrm{Y}-\overline{\mathrm{Y}})^2}$	$Q^2 > 0.5$	0.55303	0.50548	Good
$\mathbf{F_{(4,37)}}$	$\frac{\sum(\mathrm{Y}-\overline{\mathrm{Y}})^2}{p}\Big/\frac{\sum\left(\mathrm{Y}-\hat{\mathrm{Y}}\right)^2}{N-p-1}$	$F_{(4,37)} > 2.09$	15.886	13.0984	Good
	Random model Parameter (Y-Randomization) for robustness				
$\overline{\boldsymbol{R}}_{\boldsymbol{r}}$	The average of the coefficient of correlation for randomized data	$\overline{R} < 0.5$	0.381252	0.396145	Good
$\overline{\boldsymbol{R}}^2_{\boldsymbol{r}}$	The average of the coefficient of determination for randomized data	$\overline{R}^2_r < 0.5$	0.169464	0.180819	Good
$\overline{\boldsymbol{Q}}^2_{\boldsymbol{r}}$	The average of leave one out cross-validated determination coefficient for randomized data	$\overline{Q}^2_r < 0.5$	-0.09626	-0.05495	Good
$\mathbf{{}^cR^2_p}$	$R^2\times\left(1-\sqrt{\lvert R^2-\overline{R}^2_r\rvert}\right)$	${}^cR^2_p > 0.5$	0.57134	0.538618	Good
	External validation for predictability and stability of the model				
$\mathbf{R^2_{Pred.}}$	$1-\frac{\sum\left(\mathrm{Y}_{Ext.}-\hat{\mathrm{Y}}_{Ext.}\right)^2}{\sum(\mathrm{Y}_{Ext.}-\overline{\mathrm{Y}})^2}$	$R^2_{pred} > 0.6$	0.64448	0.81813	Good
$\mathbf{r^2}$	Coefficient of determination for the plot of predicted versus observed for test set	$r^2 > 0.6$	0.60685	0.84631	Good
$\mathbf{r^2_0}$	r^2 at zero intercept		0.60619	0.81556	Good
$\mathbf{r'^2_0}$	r^2 for the plot of experimental versus predicted activity for the test set at zero intercept		0.6061	0.81556	Good
$\lvert\mathbf{r^2_0-r'^2_0}\rvert$		$\lvert r^2_0-r'^2_0\rvert < 0.3$	0.21141	0.1478	Good
$\boldsymbol{k}$	The slope of the plot of predicted versus experimental activity for test set at zero intercept	$0.85 < k < 1.15$	1.00682	1.00133	Good
$\mathbf{\frac{r^2-r^2_0}{r^2}}$		$\frac{r^2-r^2_0}{r^2} < 0.1$	0.00109	0.03634	Good
$\mathbf{k'}$	The slope of the plot of experimental versus predicted activity at zero intercept	$0.85 < k' < 1.15$	0.9881	0.9979	Good
$\mathbf{\frac{r^2-r'^2_0}{r^2}}$		$\frac{r^2-r'^2_0}{r^2} > 0.1$	0.34946	0.21098	Good

Y is the experimental activity for a train set,$\overline{\mathrm{Y}}$, the average of the experimental activity for the train set $\hat{\mathrm{Y}}$, Predicted activity for a train set, $\hat{\mathrm{Y}}_{loo.}$ leave one out cross-validation predicted activity for the train, $\mathrm{Y}_{Ext.}$ experimental activity for the test set, and $\hat{\mathrm{Y}}_{ext}$ predicted activity for the test set

4.2 Applicability domain

The values of 0.37 and 0.39 were obtained as the precautionary leverage (h*) for the built activity (pEC_{50}) and toxicity (pCC_{50}) models using Equation 6, respectively. It could be realized from Figures 3 and 4 that about 98 – 99 % of the compounds are within the plot area boundary of 0< h* < 0.37, 0.39 and SDR of ±3 as depicted by the models William's plot (Figures 3 and 4). Such observations entail highly reliable predictions [16].

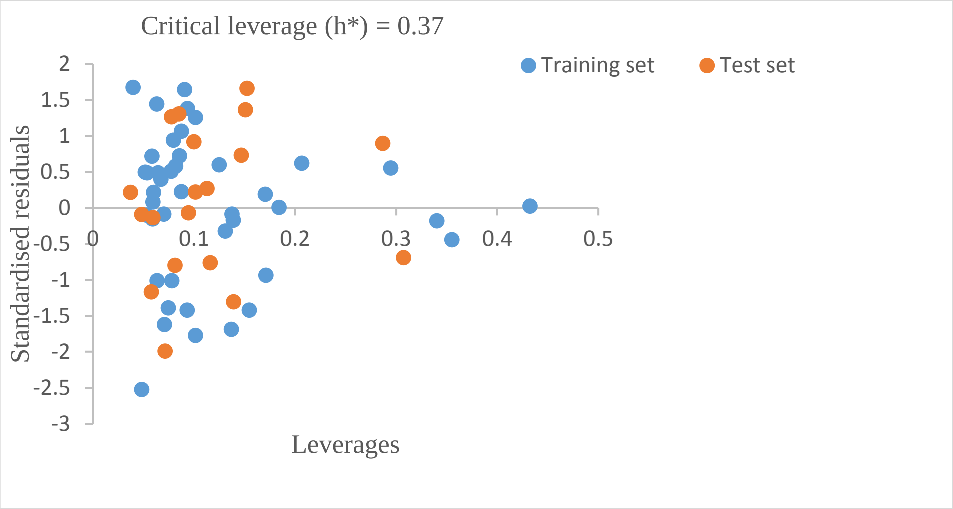

Figure 3. William's plot: A graphical illustration of standardized residual against leverages of activity (pEC_{50})

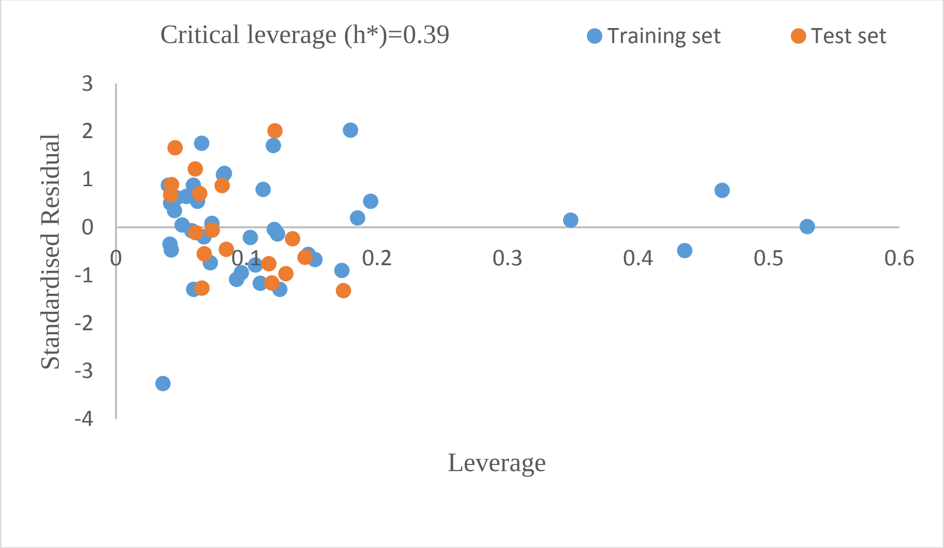

Figure 4. William's plot: A graphical illustration of standardized residual against leverage of toxicity (pCC_{50})

Table 2. Definition of molecular descriptors with their corresponding mean effects and collinearity study in the activity model

Descriptors	Definition	Mean effect	Class	VIF
ALogP	Ghose-Crippen LogKow	-0.03215	2D	1.259308
MATS7c	Moran autocorrelation - lag 7 / weighted by charges	0.152607	2D	1.075805
nHBAcc	Number of hydrogen bond acceptors	-0.51101	2D	1.185843
TDB9v	3D topological distance-based autocorrelation - lag 9 / weighted by van der Waals volumes	1.390546	3D	1.371894

Table 3. Definition of molecular descriptors with their resultant mean effects and collinearity evaluation for the toxicity model.

Descriptors	Definition	Class	Me	VIF
ALogP	Ghose-Crippen LogKow	2D	-0.0032	1.203
AATS7i	Average Broto-Moreau autocorrelation - lag 7 / weighted by first ionization potential	2D	1.0168	1.392
SM1_Dzs	Spectral moment of order 1 from Barysz matrix / weighted by I-state	2D	-0.0219	1.381
RPCS	Relative positive charge surface area -- most positive surface area * RPCG	3D	0.0084	1.172

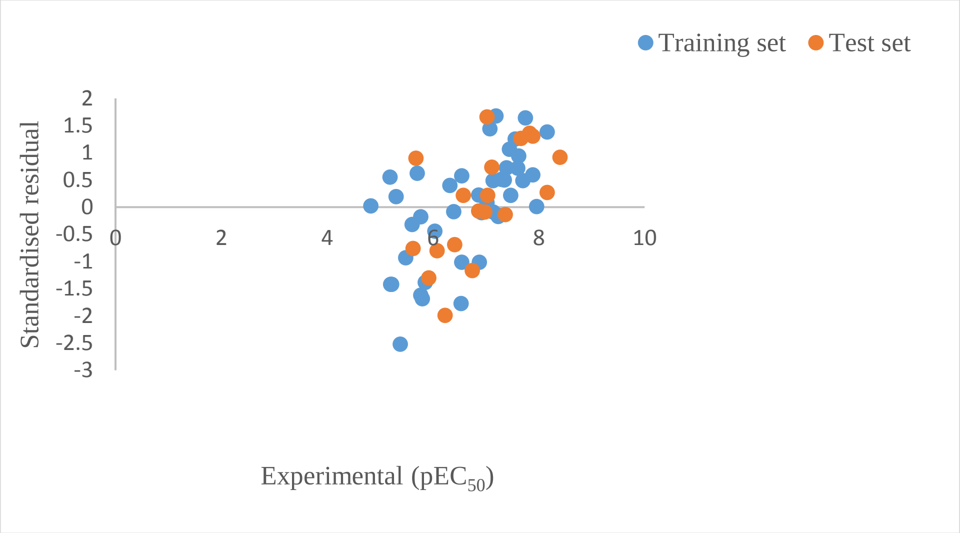

Figure 5. Graphical representation of standardized residual against experimental activity (anti-dengue activity)

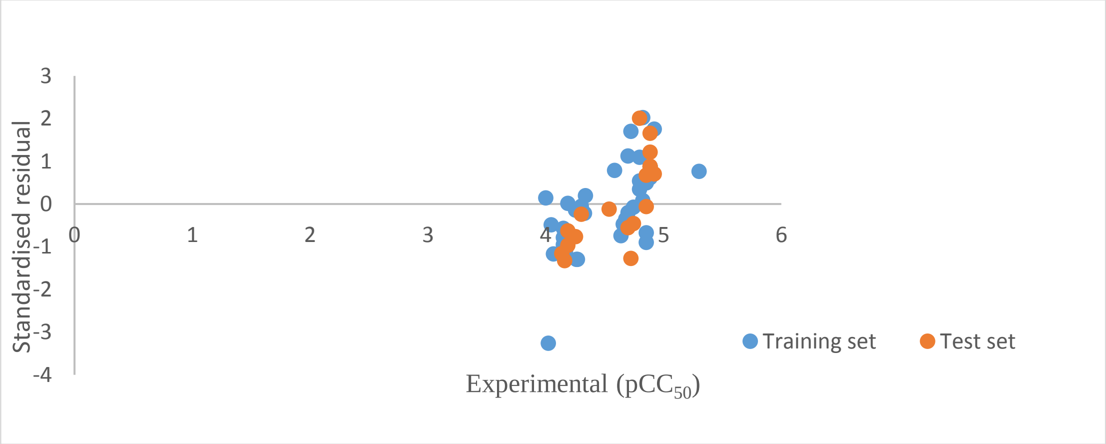

Figure 6. Graphical representation of standardized residual against experimental toxicity (pCC_{50})

4.3 Variance inflation and systematic error analysis

The multi-co-linearity results showed that the VIF value for each of the descriptors was less than the value of 5, entailing a statistically satisfactory model void of the multi-co-linearity and hence, not coincidentally obtained [16]. The confirmation of the lack of systematic error from the models is shown in Figures 5 and 6, the uniform dispersal of the data points around the line where the standardized residual equal zero signifies the absence of systematic error in building the models.

4.4 Descriptors analysis and implication

The understanding of the information encoded in the molecular descriptors contained in the models (Eq. 8 and 9), provides insights into how the chemical functional groups translate into the activity (pEC_{50}) and toxicity (pCC_{50}) of the DNV-2 NS4B inhibitors considered. Hence, a suitable understanding of the descriptors is significant. The descriptors ALogP, MATS7c, nHBAcc, TDB9v had mean effect values of -0.03215, 0.152607, -0.51101, and 1.39054 for the activity while the descriptors ALogP, AATS7i, SM1_Dzs, and RPCS had mean effect values of -0.0032, 1.0168, -0.0219, 0.0084 for the toxicity.

From both the activity and toxicity models (Equations 8 and 9) respectively. The Ghose-Crippen LogKow descriptor (ALogP) is common to both models with a negative mean effect in both models entails a decrease in activity and increase in toxicity since higher pEC_{50} entails better activity while a higher pCC_{50} entails higher toxicity value, but its percentage contributions in both models are significantly low.

The descriptors MATS7c is the Moran autocorrelation - lag 7 / weighted by charges, nHBAcc is the number of hydrogen bond acceptors, while TDB9v is the 3D topological distance-based autocorrelation - lag 9 / weighted by van der Waals volumes. As presented in Table 2 positive mean effect value for the activity model entails improvement in the inhibitory activity with an increase in the value of such descriptor while a negative mean effect entails a decrease in the activity. MATS7c and TDB9v descriptors had positive mean effects while ALogP and nHBAcc had negative mean effect values, with TDB9v having the highest contribution.

For the toxicity model (Equation 9), the descriptor AATS7i is the average Broto-Moreau autocorrelation - lag 7 / weighted by first ionization potential, while SM1_Dzs is the spectral moment of order 1 from Barysz matrix / weighted by I-state and RPCS is the relative positive charge surface area -- most positive surface area * RPCG. The AATS7i and RPCS descriptors both have positive mean effect values, with AATS7i, having the largest contribution and positive mean effect value of 1.0168, which means decreasing the value of such factor will greatly reduce the toxicity of such compound.

The MATS7c and AATS7i descriptors in both models belong to the autocorrelation descriptors (ATS) that are mathematically evaluated using Eq. 10.

$$ATSdw=\sum_{i=1}^{A}.\sum_{j=1}^{A}.\delta_{ij}.w_{id}.w_{jd} \qquad (10)$$

From Eq. 10, the number of atoms is denoted by A, whereas δ_{ij} and d denote the Kronecker function and the autocorrelation lag factors respectively, while atomic properties such as atomic mass or electronegativity for atoms i and j are represented by wi and wj respectively [31].

4.5 Prdicted ADMET and drug-likeness

The ADMET and drug-likeness results of the selected indole derivatives are presented in Table 4. From Table 4, it is shown that the compounds all passed the oral bioavailability criteria recommended by Lipinski. Non-violation of such criteria by any compound entails oral bioavailability and drug-likeness [15].

Furthermore, the high gastrointestinal absorption predicted for the selected indole derivatives could be attributed to non-violation of the Lipinski's rule [15]. A compound with gastrointestinal absorption of less than 30% is regarded as having poor gastrointestinal absorption. The gastrointestinal absorption predicted for the selected indole derivatives ranged between 94.044 and 90.219%, hence the compound could be said to have high or excellent gastrointestinal absorption as shown in Table 4 [30].

A compound with positive AMES toxicity is considered carcinogenic or mutagenic [30]. The selected compounds were predicted to have negative AMES toxicity except compounds 12 and 53 with positive AMES toxicity.

Table 4: ADMET predicted drug-likeness of some selected indole derivatives (12, 33, 37, 45, 51, 53 and 101) with high potency against the NS4B receptor

Compound ID	12	33	37	45	51	53	101
MW	386.44	404.43	387.43	397.43	404.43	390.86	434.46
#Rotatable bonds	7	7	7	6	7	6	9
#H-bond acceptors	3	4	4	4	4	2	5
#H-bond donors	2	2	2	2	2	2	3
TPSA	63.35	63.35	76.24	80.15	63.35	54.12	83.58
Consensus Log P	4.04	4.35	3.3	3.94	4.36	4.59	3.9
GI absorption	High	High	High	High	High	High	High
Lipinski #violations	0	0	0	0	0	0	0
Intestinal absorption (human)	92.834	92.566	94.044	93.14	92.785	91.854	90.219
AMES toxicity	Yes	No	No	No	No	Yes	No

5. Conclusions

The current finding describes the QSAR study of the anti-dengue inhibitory activity, as well as the cytotoxicity of indole derivatives. The models developed for the prediction of the anti-dengue activity and the toxicity of the indole derivatives towards the design of non-toxic and potent derivatives of the indole derivative proved to be excellent, as such, models were statistically valid. The biological activity of indole derivatives was shown to be influenced by ALogP, MATS7c, nHBAcc, and TDB9v descriptors while the toxicity was also revealed to be determined by ALogP, AATS7i, SM1_Dzs molecular descriptors.

Based on the reported results, it will be right to conclude that the anti-dengue activity, as well as toxicity of newly synthesized indole derivatives could be predicted with certainty, coupled with the means of information obtained from the models. It is also promising to design compounds with improved anti-dengue activity against the multiple serotypes of dengue NS4B viral receptor with an improved pharmacokinetic profile.

Acknowledgements

The authors acknowledge the financial support of the University of Maiduguri and Ahmadu Bello University, Zaria, for providing the platform for this research.

References

[1] Z. Fatima, M. Idrees, M. A. Bajwa, Z. Tahir, O. Ullah, M. Q. Zia, A. Hussain, M. Akram, B. Khubaib, S. Afzal and S. Muni, Serotype and genotype analysis of dengue virus by sequencing followed by phylogenetic analysis using samples from three mini outbreaks-2007-2009 in Pakistan. *BMC microbiology*, 11 (2011) 1-8.

[2] M. G. Guzman, M. Alvarez and S. B. Halstead, Secondary infection as a risk factor for dengue hemorrhagic fever/dengue shock syndrome: an historical perspective and role of antibody-dependent enhancement of infection. *Archives of virology*, 158 (2013) 1445-59.

[3] M E. Beatty, A. Stone, D. W. Fitzsimons, J. N. Hanna, S. K. Lam, S. Vong, M. G. Guzman, J. F. Mendez-Galvan, S. B. Halstead, G. W. Letson, J. Kuritsky, Best practices in dengue surveillance: a report from the Asia-Pacific and Americas Dengue Prevention Boards. *PLoS neglected tropical diseases*, 4 (2010) e890.

[4] S. P. Lim, C. G. Noble, C. C. Seh, T. S. Soh, A. El Sahili, G. K. Chan, J. Lescar, R. Arora, T. Benson, S. Nilar and U. Manjunatha, Potent allosteric dengue virus NS5 polymerase inhibitors: mechanism of action and resistance profiling. *PLoS pathogens*, 12 (2016) e1005737.

[5] A. Samimi, S. Zarinabadi, A. Bozorgian, Optimization of Corrosion Information in Oil and Gas Wells Using Electrochemical Experiments, *International Journal of New Chemistry,* 8 (2021), 149-163.

[6] S. P.Lim, Q. Y. Wang, C. G. Noble, Y. L. Chen, H. Dong, B. Zou, F. Yokokawa, S. Nilar, P. Smith, D. Beer and J. Lescar, Ten years of dengue drug discovery: progress and prospects. *Antiviral research*, 100 (2013) 500-19.

[7] T. T. Nguyen, S. Lee, H. K. Wang, H. Y. Chen, Y. T. Wu, S. C. Lin, D. W. Kim and D. Kim, In vitro evaluation of novel inhibitors against the NS2B-NS3 protease of dengue fever virus type 4. *Molecules*, 18 (2013) 15600-12.

[8] R. E. Blanton, L. K. Silva, V. G. Morato, A. R. Parrado, J. P. Dias, P. R. Melo, E. A. Reis, K. A. Goddard, M. R. Nunes, S. G. Rodrigues and P. F. Vasconcelos, Genetic ancestry and income are associated with dengue hemorrhagic fever in a highly admixed population. *European Journal of Human Genetics*, 16 (2008) 762-5.

[9] A. Guzman and R. E. Istúriz, Update on the global spread of dengue. *International journal of antimicrobial agents*, 36 (2010) S40-2.

[10] P. D. Zanotto, E. A. Gould, G. F. Gao, P. H. Harvey and E. C. Holmes, Population dynamics of flaviviruses revealed by molecular phylogenies, Proceedings of the National Academy of Sciences, 93 (1996) 548-53.

[11] A. Balasubramanian, T. Teramoto, A. A. Kulkarni, A. K. Bhattacharjee and R. Padmanabhan, Antiviral activities of selected antimalarials against dengue virus type 2 and Zika virus. *Antiviral research*, 137 (2017) 141-50.

[12] A. Bozorgian, B. Raei, Thermodynamic modeling and phase prediction for binary system dinitrogen monoxide and propane, *Journal of Chemistry Letters* 1 (2020) 143-148.

[13] D. Bardiot, M. Koukni, W. Smets, G. Carlens, M. McNaughton, S. Kaptein, K. Dallmeier, P. Chaltin, J. Neyts and A. Marchand, Discovery of indole derivatives as novel and potent dengue virus inhibitors. *Journal of Medicinal Chemistry*, 61 (2018) 8390-401.

[14] M. Bagheri Sadr, A. Bozorgian, An Overview of Gas Overflow in Gaseous Hydrates, *Journal of Chemical Reviews* 3 (2021), 66-82

[15] A. Daina, O. Michielin and V. Zoete, SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. *Scientific reports*. 7 (2017) 1-3.

[16] S. N. Adawara, G. A. Shallangwa, P. A. Mamza and A. Ibrahim, Molecular docking and QSAR theoretical model for prediction of phthalazinone derivatives as new class of potent dengue virus inhibitors. *Beni-Suef University Journal of Basic and Applied Sciences*, 9(2020) 1-7.

[17] W. J. Hehre and W. W. Huang, Chemistry with computation: an introduction to SPARTAN. Wavefunction, Incorporated, 1995.

[18] A. Bozorgian, Investigation of the effect of Zinc Oxide Nano-particles and Cationic Surfactants on Carbon Dioxide Storage capacity, Advanced Journal of Chemistry, Section B: Natural Products and Medical Chemistry, 3 (2021) 54-61.

[19] S. N. Adawara, G. A. Shallangwa, P. A. Mamza and A. Ibrahim, In Silico Studies of Oxadiazole Derivatives as Potent Dengue Virus Inhibitors. *Chemistry Africa*, 4 (2021) 861-8.

[20] P Ambure, R. B. Aher, A. Gajewicz, T. Puzyn and K. Roy, “NanoBRIDGES” software: open access tools to perform QSAR and nano-QSAR modeling. *Chemometrics and Intelligent Laboratory Systems*, 15 (2015) 1-3.

[21] S. B. Olasupo, A. Uzairu, G. A. Shallangwa and S. Uba, QSAR modeling, molecular docking and ADMET/pharmacokinetic studies: a chemometrics approach to search for novel inhibitors of norepinephrine transporter as potent antipsychotic drugs. *Journal of the Iranian Chemical Society*, 17 (2020) 1953-66.

[22] J. H. Friedman, Multivariate adaptive regression splines. *The annals of statistics*, 19 (1991) 1-67.

[23] R. Veerasamy, H. Rajak, A. Jain, S. Sivadasan, C. P. Varghese and R. K. Agrawal, Validation of QSAR models-strategies and importance. *Int. J. Drug Des. Discov,* 3 (2011) 511-9.

[24] A. Bozorgian, Investigation of Hydrate Formation Kinetics and Mechanism of Effect of Inhibitors on it, a Review, *Journal of Chemical Reviews*, 3 (2021), 50-65.

[25] D. Weintrop, E. Beheshti, M. Horn, K. Orton, K. Jona, L. Trouille and U. Wilensky, Defining computational thinking for mathematics and science classrooms. *Journal of Science Education and Technology,* 25 (2016) 127-47.

[26] N. Nikolova and J. Jaworska, Approaches to measure chemical similarity–a review. *QSAR & Combinatorial Science*, 22 (2003) 1006-26.

[27] A. Golbraikh, X. S. Wang, H. Zhu and A. Tropsha, Predictive QSAR modeling methods and applications in drug discovery and chemical risk assessment. *Handbook of computational chemistry,* 1 (2012) 1309-42.

[28] T. I. Netzeva, A. P. Worth, T. Aldenberg, R. Benigni, M. T. Cronin, P. Gramatica, J. S. Jaworska, S. Kahn, G. Klopman, C. A. Marchant and G. Myatt, Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships: The report and recommendations of ecvam workshop 52. *Alternatives to Laboratory Animals,* 33(2): (2015) 155-73.

[29] S. Dimitrov, G. Dimitrova, T. Pavlov, N. Dimitrova, G. Patlewicz, J. Niemela and O. Mekenyan, A stepwise approach for defining the applicability domain of SAR and QSAR models. *Journal of chemical information and modeling*, 45(4) (2005) 839-49.

[30] D. E. Pires, T. L. Blundell, and D. B. Ascher, pkCSM: predicting small-molecule pharmacokinetic and toxicity properties using graph-based signatures. *J. Med. Chem*., 58 (2015) 4066-72.

[31] V. Consonni and R. Todeschini, Molecular Descriptors for Chemoinformatics: Volume I: Alphabetical Listing/Volume II: Appendices, References. John Wiley & Sons; 2009.