Regular Article

PHYSICAL
CHEMISTRY
RESEARCH
Published by the
Iranian Chemical Society
www.physchemres.org
info@physchemres.org

Phys. Chem. Res., *Vol. 6, No. 3, 479-492, September 2018*
DOI: 10.22036/pcr.2018.115878.1457

A Novel QSAR Model for the Evaluation and Prediction of (E)-N'-Benzylideneisonicotinohydrazide Derivatives as the Potent Anti-mycobacterium Tuberculosis Antibodies Using Genetic Function Approach

E. Shola Adeniji*, S. Uba and A. Uzairu
Department of Chemistry, Ahmadu Bello University, Zaria-Nigeria

(Received 22 January 2018, Accepted 17 April 2018)

A dataset of (E)-N'-benzylideneisonicotinohydrazide derivatives as the potent anti-mycobacterium tuberculosis antibodies has been investigated utilizing quantitative structure-activity relationship (QSAR) techniques. Genetic function algorithm (GFA) and multiple linear regression analysis (MLRA) were used to select the descriptors and to generate the correlation QSAR models that correlate the minimum inhibitory concentration (MIC) values against mycobacterium tuberculosis with the molecular structures of the active molecules. The models were validated, and the best model with squared correlation coefficient (R^2) of 0.9202, adjusted squared correlation coefficient (R_{adj}) of 0.91012, and leave one out (LOO) cross validation coefficient (Q^2_{CV}) value of 0.8954 was selected. R^2 pred of 0.8842 was achieved for the external validation set used for confirming the predictive power of the model. Stability and robustness of the model obtained by the validation test indicate that the model can be used to design and synthesize the other (E)-N'-benzylideneisonicotinohydrazide derivatives with improved anti-mycobacterium tuberculosis activity.

Keywords: Anti-tuberculosis, Descriptors, Genetic function algorithm, QSAR, Validation

INTRODUCTION

Tuberculosis (TB) is infectious disease caused by *Mycobacterium tuberculosis*. About 2.5 billion people were infected with tuberculosis worldwide and mortality of approximately 1.5 million people were reported annually [1,2]. Despite the availability of tuberculosis first-line drug; rifampicin (RIF), pyrazinamide (PZA), isoniazid (INH), ethambutol (EMB) and streptomycin (STP) the increase in the incidence of both multidrug-resistant (MDR-TB) and extensively drug-resistant tuberculosis (XDR-TB) are observed [3,4]. Furthermore, treatment requiring the use of these drugs causes serious side effects such as: thrombocytopenia occurring mostly due to rifampicin (RIF) [5], neuropathy induced by isoniazid while biggest problem associated with tuberculosis treatment causes hepatitis [6]. Considering these effect, the synthesis of new compounds with anti-tuberculosis activity has been the target of many medicinal chemistry and pharmacist. New synthetized (E)-N'-benzylideneisonicotinohydrazide derivative compounds demonstrates tuberculosis inhibition activity similar to the control drug Streptomycin [7].

Synthesis of novel compounds are developed using a trial and error approach, which is time consuming and expensive. The application of quantitative structure activity relationship (QSAR) technique to this problem has a potential to minimize the effort and time required to discover new compounds or to improve current compounds in terms of their efficiency.

QSAR establishes the mathematical relationship between physical, chemical, biological or environmental activities of interest and measurable or computable parameters such as physicochemical, topological, stereo chemical or electronic indices called molecular descriptors

*Corresponding author. E-mail: shola4343@gmail.com

[8].

The aim of this research is to develop a QSAR model. To do so, several statistical tools, such as genetic function algorithm (GFA), and multiple linear regression (MLR) method, were employed for variable selection and prediction the activity of (E)-N'-benzylideneisonicotinohydrazide derivatives as potent anti-mycobacterium tuberculosis antibodies.

MATERIALS AND METHOD

Data Collection

Data set of (E)-N'-benzylideneisonicotinohydrazide derivatives as potent anti-mycobacterium tuberculosis antibodies used in this study were obtained from the literature [9].

Biological Activities (pMIC)

The biological activities of (E)-N'-benzylideneisonicotinohydrazide derivatives against mycobacterium tuberculosis measured in MIC (μM) were converted to logarithm unit (pMIC) using the Eq. (1) in order to increase the linearity of activity values and approach normal distribution. The observed structures and the biological activities of these compounds are presented in Fig. 1 and Table 1, respectively.

$$pMIC = -\log(MIC) \tag{1}$$

Optimization

The 2D structures of the compounds presented in Table 1 were drawn by the chemdraw program software [10]. The spatial conformations of the compounds were exported from 2D structure to 3D format using the Spartan 14 V1.1.4 Wave Function programming package. All 3D structures were geometrically optimized by minimizing energy. The chemical structures were initially minimized by molecular mechanics force field (MMFF) count to remove strain energy. Density functional theory (DFT) method was later employed using the Becke's three parameter exchange functional (B3) hybrid with Lee, Yang and Parr correlation functional (LYP), termed as B3LYP hybrid functional, for complete geometric optimization of the structures. The Spartan files of all the optimized molecules were then saved in SD file format, which is the recommended input format in PaDEL-Descriptor software V2.20.

Molecular Descriptor Calculation

Molecular descriptors are mathematical values describing the properties of a molecule. Quantum chemical descriptors calculation for all the 50 molecules of (E)-N'-benzylideneisonicotinohydrazide derivatives was calculated using PaDEL-Descriptor software V2.20. A total of 1876 molecular descriptors were calculated.

Normalization and Data Pretreatment

The descriptors' value were normalized using Eq. (2) in order to give each variable the same opportunity at the onset to influence the model [11].

$$X = \frac{X_1 - X_{\min}}{X_{\max} - X_{\min}} \tag{2}$$

Where Xi is the value of each descriptor for a given molecule, Xmax and Xmin are the maximum and minimum values for each column of descriptors X. The normalized data were subjected to pretreatment using Data Pretreatment software obtained from drug theoretical and cheminformatics laboratory (DTC Lab) in order to remove noise and redundant data.

Data Division

In order to obtain validated QSAR models, the dataset was divided into training and test sets using Data Division software obtained from DTC Lab by employing Kennard and Stone's algorithm [12]. This algorithm has been applied with a great success in many recent QSAR studies and has been highlighted as one of the best ways to build training and test sets [13-17]. In this algorithm, two compounds with the largest Euclidean distance apart were initially selected for the training set. The remaining compounds for the training set were selected by maximizing the minimum distance between these two compounds and the rest of the compounds in the dataset. This process continues until the desired number of compounds needed for the training set are selected, then, the remaining compounds in the dataset are used as the test set [12].

The algorithm employs Euclidean distance ED_X (p,q),

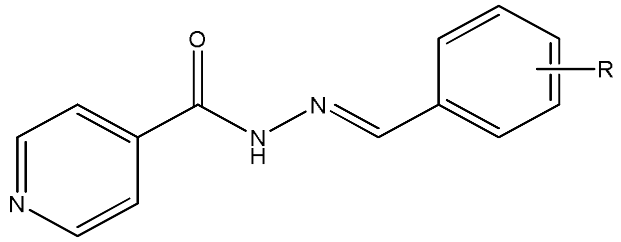

Fig. 1. General structure of (E)-N'-benzylideneisonicotinohydrazide derivatives.

Table 1. Molecular Structure of (E)-N'-benzylideneisonicotinohydrazide Derivatives as the Potent Anti-mycobacterium Tuberculosis Antibodies and their Activities

Molecule	R	Activity pMIC
1[a]	H	2.34
2	2-Br	3.00
3	3-Br	3.11
4	4-Br	3.12
5	2-Cl	2.91
6[a]	3-Cl	3.05
.7	4-Cl	2.86
8[a]	2-F	2.15
9[a]	3-F	2.48
10	4-F	2.34
11	2-CN	2.22
12	3-CN	2.26
13	4-CN	2.28
14	2-NO_2	2.01
15	3-NO_2	1.95
16	4-NO_2	2.00
17	2-OCH_3	2.38
18[a]	3-OCH_3	2.44
19	4-OCH_3	2.49
20	2-OCH_2CH_3	2.88
21[a]	3--OCH_2CH_3	2.96
22[a]	3-OH	2.54

Where superscript a represent the test set.

between the x vectors of each pair (p,q) of samples to ensure a uniform distribution of such a subset along the x data space;

$$ED_X\,(p,q) = \sqrt{\sum_{j=1}^{N}[x_p(j) - x_q(j)\,2_{p,q}\;\varepsilon[1,M]} \tag{3}$$

N is the number of variables in x, and M is the number of samples while x_p *(j)* and x_q *(j)* are the *jth* variable for samples p and q, respectively. The training set was used to generate the model, while the test set was used for the external validation of the model

Model Development

MLR is a strategy utilized for displaying direct relationship between a dependent variable Y (pMIC) and independent variable X (atomic descriptors). The model fits well such that sum of the square difference between the experimental and predicted values of set biological activity is minimized. In regression analysis, contingent mean of dependent variable (pMIC) Y relies on

(Descriptors) X. MLR examination extends this idea to incorporate more than one autonomous variables, and regression equation takes the form:

$$Y = b_1x_1 + b_2x_2 + b_3x_3 + C \tag{4}$$

where Y is dependent variable, 'b's are regression coefficients for corresponding '*x*'s (independent variables), and 'C' is a regression constant or intercept.

Validation of Model

Validation of the model was carried out using Material Studio software version 8 using genetic function approximation (GFA) method. The numbers of descriptors in the regression equation were three, and Population and Generation were set to 10000 and 10000, respectively. The number of top equations returned was ten. Mutation probability was 0.1, and the smoothing parameter was 0.5. The models were scored based on Friedman's LOF. In GFA algorithm, an individual or model was represented as one-dimensional string of bits. It was a distinctive characteristic of GFA that could create a population of models rather than a single model.

The models were estimated using the LOF, measured using a slight variation of the original Friedman formula, so that the best fitness score can be received. In Materials Studio version 8, LOF is measured using a slight variation of the original Friedman formula [18] .The revised formula is:

$$LOD = \frac{SEE}{\left(1 - \frac{c + d \times p)}{M}\right)^2} \tag{5}$$

where SEE is the Standard Error of Estimation which is equivalent to the models standard deviation. It is a measure of model quality, and a model is said to be a better model if this term has the low SEE value. SEE is defined by equation below;

$$SEE = \sqrt{\frac{(Y_{\exp} - Y_{pred})^2}{N - P - 1}} \tag{6}$$

c is the number of terms in the model other than the constant term, d is a user defined smoothing parameter, p is the total number of descriptors contained in the model and M is the number of data in the training set [21].

The square of the correlation coefficient (R^2) describes the fraction of the total variation attributed to the model. The closer the value of R^2 is to 1.0, the better the regression equation explaining the Y variable. R^2 is the most commonly used internal validation indicator and is expressed as follows:

$$R^2 = 1 - \left[\frac{\sum(Y_{\exp} - Y_{pred})^2}{\sum(Y_{\exp} - \bar{Y}_{training})^2}\right] \tag{7}$$

where:

$Y_{\exp}$, Y_{pred} and $\bar{Y}_{training}$ are the experimental activity, the predicted activity and the mean experimental activity of the samples in the training set, respectively. R^2 value varies directly with the increase in number of repressors, i.e., descriptors, thus, R^2 cannot be a useful measure for the stability of model. Therefore, R^2 is adjusted for the number of explanatory variables in the model. The adjusted R^2 is defined as:

$$R^2_{adj} = \frac{R^2 - P\,(n-1)}{n - p + 1} \qquad (8)$$

where p is the number of independent variables in the model.

The capability of the QSAR equation to predict bioactivity of the new compounds was determined using the leave-one-out cross validation method. The cross-validation regression coefficient (Q^2_{CV}) was calculated with the equation below:

$$Q^2_{CV} = 1 - \left[\frac{\sum (Y_{pred} - Y_{\exp})^2}{\sum (Y_{\exp} - \bar{Y}_{training})^2}\right] \qquad (9)$$

where
Y_{pred}, Y_{exp}, and $\bar{Y}_{training}$ are the predicted, experimental and mean values of experimental activity of the training set.
The coefficient of determination for the test set R^2_{test} was calculated with the equation below;

$$R^2_{test} = 1 - \frac{\sum (Y_{pred_{test}} - Y_{\exp_{test}})^2}{\sum (Y_{pred_{test}} - \bar{Y}_{training})^2} \qquad (10)$$

where $Y_{pred_{test}}$ and $Y_{\exp_{test}}$ are the predicted and experimental activity test set, while $\bar{Y}_{training}$ is mean values of experimental activity of the training set.

Y-Randomization Test

To guarantee that the created QSAR model is strong and not inferred by chance, the Y-randomization test was performed on the training set data as suggested by [19]. Random MLR models are generated by randomly shuffling the dependent variable (activity data) while keeping the independent variables (descriptors) unaltered. The new QSAR models are expected to have significantly low R^2 and Q^2 values for several trials confirming that the developed QSAR models are robust. Another parameter calculated is cR^2_p which should be more than 0.5 for passing this test;

$$cR^2_p = R \times [R^2 - (R_r)^2]^2 \qquad (10)$$

where

cR^2_p is the coefficient of determination for Y-randomization, R is the coefficient of determination for Y-randomization and Rr is the average ‘R’ of random models.

Evaluation of the Applicability Domain of the Model

The built QSAR model was evaluated based on applicability domain approach to verify that the model is robust and reliable to predict the activities of the inhibitor compounds [19]. The leverage approach was employed in defining and describing the applicability domain of the QSAR models built [20]. Leverage of a given chemical compound, hi, is defined as follows:

$$hi = X_i (X^T X)^{-1} X_i^T \qquad (11)$$

where *Xi* is training compounds of the *i* matrix. *X* is the m × k descriptor matrix of the training set compound and X^T is the transpose matrix of *X* used to build the model. The warning leverage (h^*) is the boundary of values for *X* outliers and is defined as:

$$h^* = 3\frac{(d+1)}{m} \qquad (12)$$

where *m* is the descriptors and *d* is the compound that made up the training set.

Quality Assurance of the Model

The fitting ability, stability, reliability and predictive ability of the developed models were evaluated by internal and external validation parameters. The validation parameters were compared with the minimum value recommended for a generally acceptable QSAR model [20] shown in Table 2.

RESULTS AND DISCUSSION

A QSAR examination was performed to investigate the structure activity relationship of 22 compounds as the potent anti-mycobacterium tuberculosis antibodies. The nature of models in a QSAR study is expressed by its fitting and forecast capacity. In order to assemble a decent QSAR

Table 2. Minimum Recommended Value of Validation Parameters for a Generally Acceptable QSAR Model

Symbol value	Name	Value
R^2	Coefficient of determination	≥ 0.6
$P_{(95\%)}$	Confidence interval at 95% confidence level	< 0.05
Q^2_{CV}	Cross validation coefficient	> 0.5
$R^2 - Q^2_{CV}$	Difference between R^2 and Q^2_{CV}	≤ 0.3
$N_{\text{ext. test set}}$	Minimum number of external test set	≥ 5
cR^2_p	Coefficient of determination for Y-randomization	> 0.5

model for anti-mycobacterium tuberculosis with good predictive power for the selected test set Kennard-Stone algorithm was used to divide the dataset of 22 compounds into a training set of 15 compounds used to develop the model and a test set of 7 compounds applied to assess the predictive ability of the model built.

Experimental and predicted activity for (E)-N'-benzylideneisonicotinohydrazide derivatives as the potent anti-mycobacterium tuberculosis antibodies and the residual values are presented in Table 3. The low residual value between experimental and predicted activity indicates that the model is of high predictability.

Univariate analysis of the activity values of the training and test set data reported in Table 4 shows that range of test set value (2.15 to 3.06) is within the range of training set value (2 to 3.12). Also, the mean and standard deviations of the test set activity value (2.3757 and 0.3413) are approximately similar to those of the training set values (2.6093 and 0.3916). This indicates that the test set is interpolative within the training set and the spread or point distribution of the two set were comparable implying that Kennard and Stone algorithm employed was able to provide a test set that is a good reflection of the training set data.

The genetic algorithm- multi linear regression (GA-MLR) investigation led to the selection of three descriptors which were used to assemble a linear model for calculating predictive activity on mycobacterium tuberculosis. Ten QSAR models were built using GFA, however, due to the statistical significance, model 1 was selected, reported and its s parameters were calculated as well.

pMIC = -0.301042455 (AATSC2m) + 0.354110306 (C1C4) - 0.007605618 (GG17) - 5.484268669

$N_{train} = 15$, $R^2 = 0.96013700$, $R_{adj} = 0.94926500$, $Q^2_{CV} = 0.92752300$ and the external validation for the test set was found to be R^2pred = 0.8863.

All the corresponding values reported in Table 7 were in agreement with parameters presented in Table 2, actually confirming the robustness of the model.

The QSAR model generated in this research was compared with that in the model obtained in the literature [21,22], as shown in equations/models …. and…;

pMIC = 4.77374(+/-0.03903) -0.18609(+/-0.04924) AATS4i +0.50382(+/-0.05235) SCH-3 - 0.44712(+/-0.06573) AVP-1 - 0.22376(+/-0.05623) maxHCsats - 0.18403(+/-0.04374)PSA

$N_{train} = 16$, $R^2 = 0.9184$, $Q^2_{CV} = 0.84987$ and R^2pred = 0.79343. [21]

Table 3. Experimental, Predicted and Residual Values of (E)-N-benzylideneisonicotinohydrazide Derivatives

S/N (Molecule)	Experimental activity	Predicted activity	Residual
1[a]	3	3.034207	-0.03421
2	3.11	3.207185	-0.09719
3	3.12	3.039861	0.080139
4	2.91	2.872538	0.037462
5	3.05	3.041451	0.008549
6[a]	2.86	2.790466	0.069534
7	2.15	2.18537	-0.03537
8[a]	2.48	2.516716	-0.03672
9[a]	2.22	2.182069	0.037931
10	2.26	2.147217	0.112783
11	2.28	2.358591	-0.07859
12	2	2.063555	-0.06356
13	2.38	2.334004	0.045996
14	2.44	2.589991	-0.14999
15	2.88	2.776778	0.103222
16	2.34	2.436288	-0.09629
17	2.34	2.349393	-0.00939
18[a]	2.01	1.829552	0.180448
19	1.95	1.919356	0.030644
20	2.49	2.506329	-0.01633
21[a]	2.96	2.84742	0.11258
22[a]	2.54	2.763758	-0.22376

Superscript a represents the test set.

$$\text{pIC50} = -2.040810634 \times \text{nCl} - 19.024890361 \times \text{MATS2m} + 1.855704759 \times \text{RDF140s} + 6.739013671$$

$N_{train} = 27$, $R^2 = 0.9480$, $R_{adj} = 0.9350$, $Q^2_{CV} = 0.87994$ and R^2pred = 0.76907. [22]

Table 4. Univariate Analysis of the Inhibition Data

Statistical parameters	Activity	
	Training set	Test set
Number of sample points	15	7
Range	1.12	1.01
Maximum	3.12	3.06
Minimum	2	2.15
Mean	2.61	2.38
Median	2.48	2.34
Variance	0.1431	0.0998
Standard deviation	0.39162	0.3413
Mean absolute deviation	0.3553	0.2465
Skewness	-0.0144	0.2846
Kurtosis	-1.7381	-1.2296

Table 5. List of some Descriptors Used in the QSAR Optimization Model

S/NO	Descriptors symbols	Name of descriptor (s)	Class
1	AATSC2m	Average Broto-Moreau autocorrelation - lag2/weighted by mass	2D
2	C1C4	Complementary information content index (neighborhood symmetry of 4-order)	2D
3	GIG7	Topological charge index of order 7	2D

Table 6. Pearson's Correlation Matrix and Statistics for the Descriptor Used in the QSAR Model

	Inter-correlation			Statistics		
Descriptors	*AATSC2m*	*CIC4*	*GGI7*	t-Stat	VIF	P-value
AATSC2m	1			-6.6580	1.0162	2.67E-08
CIC4	-0.10553	1		-13.829	2.7360	2.11E-06
GGI7	-0.12039	0.02855	1	8.9930	2.7829	2.89E-06

Table 7. Validation of the Genetic Function Approximation from the Material Studio Program Package

S/NO		Equation
1	Friedman LOF	0.0346
2	R-squared	0.96014
3	Adjusted R-squared	0.9493
4	Cross validated R-squared (Q^2_{CV})	0.9275
5	Significant Regression	Yes
6	Significance-of-regression F-value	88.3142
7	Critical SOR F-value (95%)	3.7487
8	Replicate points	0
9	Computed experimental error	0.0000
10	Lack-of-fit points	11
11	Min expt. error for non-significant LOF (95%)	0.0658

Based on the models presented above, the validation parameters reported in this work and those reported in the literature are all in agreement with parameters presented in Table 2 confirming the robustness of the model generated.

The names and symbols of the descriptors used in the QSAR optimization model are included in Table 5. The presence of the three 2D descriptors in the model suggests that these types of descriptors are able to characterize better anti-mycobacterium tuberculosis activities of the compounds. Table 6 includes Pearson's correlation matrix and statistics of the three descriptors employed in the QSAR model This table clearly indicates that the correlation coefficients between each pair of descriptors is very low that can be inferred that there exists no significant inter-correlation among the descriptors used in building the model. The absolute t-statistics value for each descriptor is greater than 2 at 95 % significant level which indicates the selected descriptors were good. The estimated variance inflation factor (VIF) values for all the descriptors were less than 4 implying that the model generated was statistically significant and the descriptors were orthogonal. The p-value is a probability that measures the evidence against the null hypothesis. Lower probabilities provide stronger evidence against the null hypothesis. The null hypothesis implies that there is no association between the descriptors and the activities of the molecules. The P-values of all the descriptors in the model at 95% confidence level shown in Table 6 are less than 0.05. This implies that the alternative hypothesis is accepted. Hence, there is a relationship between the descriptors used in the model and the activities of molecules which take preference over the null hypothesis.

Y- Randomization parameter test is reported in Table 8. The low R^2 and Q^2 values for several trials confirm that the developed QSAR model is robust, while the cR_p^2 value greater than 0.5 affirms that the created model is powerful and not inferred by chance.

Interpretation of the Selected Descriptors

AATSC2m is Average Broto-Moreau autocorrelation -lag2/weighted by mass descriptor. It is based on spatial dependent autocorrelation function which measures the strength of the relationship between observations (atomic or

Table 8. Y-Randomization Parameters Test

Model	R	R^2	Q^2
Original	1	1	1
Random 1	0.3454	0.1193	-1.0841
Random 2	0.4868	0.2370	-1.0985
Random 3	0.4408	0.1943	-0.9815
Random 4	0.5575	0.3108	-0.5503
Random 5	0.2957	0.0874	-1.1088
Random 6	0.5562	0.3093	-0.7285
Random 7	0.7724	0.5966	0.0328
Random 8	0.2752	0.0757	-1.1166
Random 9	0.74823	0.5598	-0.0362
Random 10	0.5557	0.3088	-0.4448
	Random Models Parameters		
Average r:	0.5034		
Average r^2:	0.2799		
Average Q^2:	-0.7117		
cRp^2:	0.8640		

molecular properties) and space separating them (lag). This descriptor is obtained by taking the molecule atoms as the set of discrete points in space and an atomic property as the function evaluated at those points. When this descriptor is calculated on molecular graph, the lag coincides with the topological distance between any pair of the vertices. AATSC2m is defined on the molecular graphs using atomic masses (m), Sanderson electronegativity (e) and inductive effect respectively of pairs of atoms 2 bond apart as the weighting scheme. These observations suggested that atomic masses and electronic distribution of atoms that made up the molecule had significant effect on the anti-tubercular activity of the dataset. In addition, the signs of the regression coefficients for each descriptor indicated the direction of influence of the descriptors in the models such that negative regression coefficient associated to a descriptor will diminish the activity of the compound.

C1C4 corresponds to complementary information content index (neighborhood symmetry of 4-order**.** It has positive coefficient indicating that an increase in the weight of molecule leads to a decrease in its anti-tubercular activity.

GIG7 is a topological charge index of order 7 descriptor contained in the model. It is negatively correlated to the anti-tubercular activity meaning that decrease in its value augments the activity of the studied compounds. The descriptor measures the strength of the connection between atomic charges 7-bonds apart. The number of ring in the

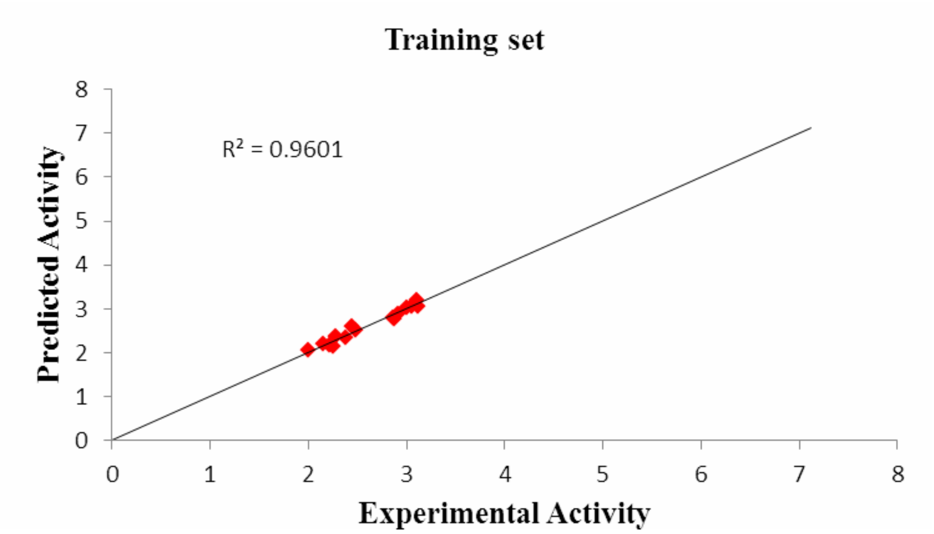

Fig. 2. Plot of the predicted activity against the experimental activity of training set.

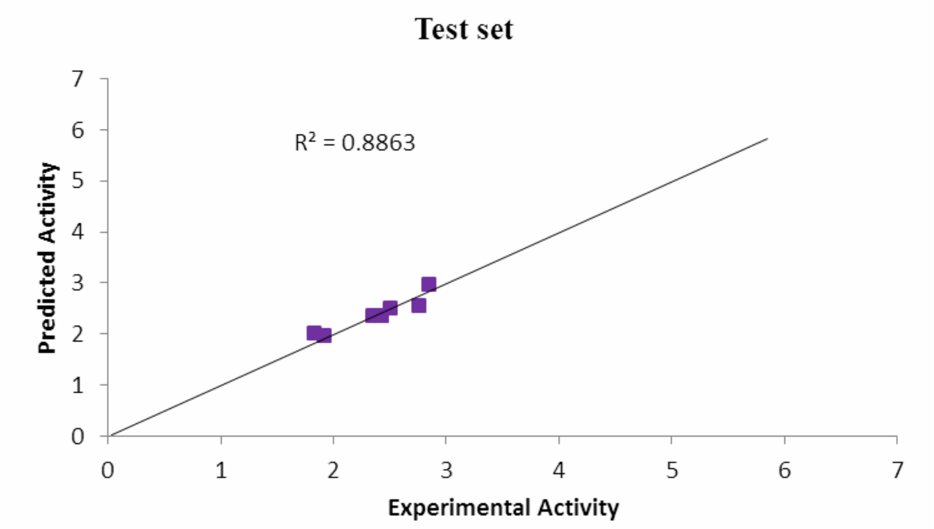

Fig. 3. Plot of predicted activity against the experimental activity of test set.

molecular system tends to increase the values of this descriptor. This may be due to increase in the amount of π-electrons in the molecular system bringing about increase in the charge difference between atoms 7-bonds apart.

Plot of predicted activity against experimental activity of training and test set are shown in Figs. 2 and 3, respectively. The R^2 value of 0.9601 for training set and R^2 value of 0.8863 for test set recorded in this study are in agreement with GFA derived R^2 value reported in Table 2. This confirms the reliability of the model. Plot of standardized residual *versus* experimental activity shown in Fig. 4 indicates that there was no systemic error in model development as the spread of residuals was pragmatic on both sides of zero [23].

The leverage values for the entire compounds in the dataset were plotted against their standardized residual values leading to discovery of outliers and influential compound in the models. The Williams plot of the

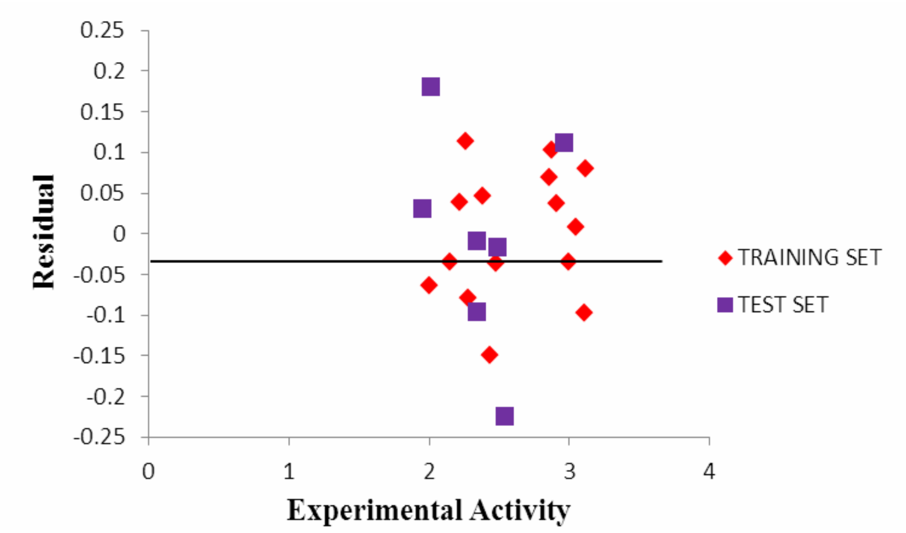

Fig. 4. Plot of residual activity *versus* experimental activity.

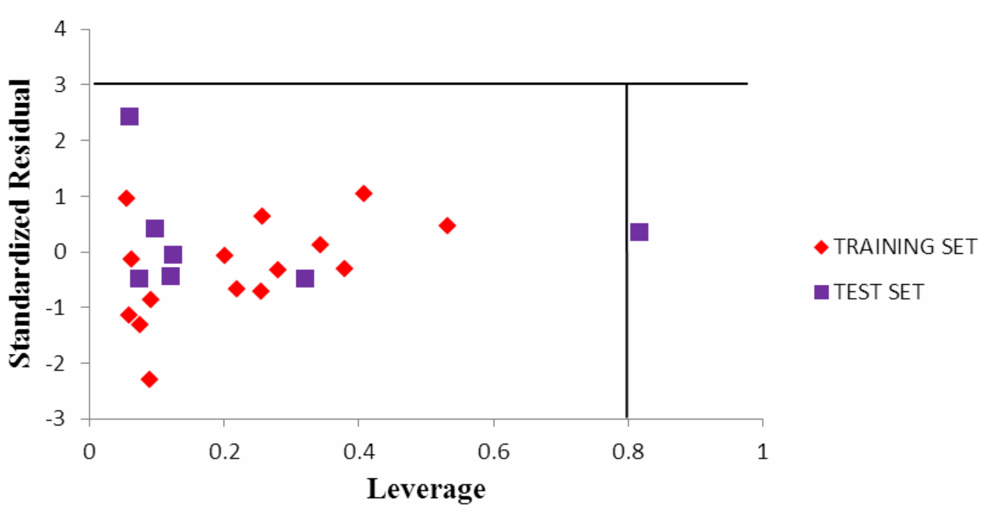

Fig. 5. The Williams plot of the standardized residual *versus* the leverage value.

standardized residuals versus the leverage value is shown in Fig. 5. Our results evidently show that all the compounds are within the square area ±3 of standardized cross-validated residual produced by the model. Therefore, no compound is said to be an outlier. However, only one compound is said to be an influencing compound since its leverage value is greater than the warning leverage (h^* = 0.80). This was attributed to difference in its molecular structure compared to other compounds in the dataset.

CONCLUSIONS

This work addresses the quantitative structure activity relationship (QSAR) between (E)-N'-benzylideneisonicotinohydrazide derivatives and their pMIC against mycobacterium tuberculosis. Results from the optimal model showed that the pMIC of the studied molecules against mycobacterium tuberculosis is affected by the AATS2m, C1C4 and G1G7 descriptors. The robustness and applicability of QSAR equation were established by the internal and external validation techniques. Stability and robustness of the model obtained by the validation test indicate that the model can be used to design other (E)-N'-benzylideneisonicotinohydrazide derivatives with improved anti-mycobacterium tuberculosis activity.

REFERENCES

[1] Lönnroth, K.; Castro, K. G; Chakaya, J. M.; Chauhan, L. S.; Floyd, K.; Glaziou, P., Tuberculosis control and elimination 2010-50: cure, care, and social development. *Lancet.* **2010**, *75*, 1814-1829, DOI: 10.1016/S0140-6736(10)60483-7

[2] Jhamb, S. S.; Goyal, A.; Singh, P. P., Determination of the activity of standard anti-tuberculosis drugs against intramacrophage Mycobacterium tuberculosis, *in vitro*: MGIT 960 as a viable alternative for BACTEC 460. *Brazilian J. Infect. Dis.* **2014**, *18*, 336-40, DOI: 10.1016/ j.bjid.2013.12.004.

[3] Aziz, M. A.; Wright, A.; Laszlo, A.; De Muynck, A.; Portaels, F.; Van Deun, A., WHO/international union against tuberculosis and lung disease Global Project on anti-tuberculosis drug resistance surveillance. Epidemiology of antituberculosis drug resistance (the global project on anti-tuberculosis drug resistance surveillance): an upd. *Lancet.* **2006**, *368*, 2142-54, DOI: 10.1016/S0140-6736(06)69863-2.

[4] Balabanova, Y.; Ruddy, M.; Hubb, J.; Yates, M.; Malomanova, N.; Fedorin, I., Multidrug-resistant tuberculosis in Russia: clinical characteristics, analysis of second-line drug resistance and development of standardized therapy. *Eur. J. Clin. Microbiol. Infect. Dis.* **2005**, *24*, 136-9, DOI: 10.1007/s10096-004-1268-4.

[5] Yakar, A.; Yakar, F.; Yildiz, N.; Kılıçaslan, Z., Isoniazid-and rifampicin-induced thrombocytopenia. *Multidiscip Respir Med.* **2013**, *8*, 13-16, Doi: 10.1186/2049-6958-8-13.

[6] Abideen, P. S.; Chandrasekaran, K.; Uma V. A.; Kalaiselvan, V., Implementation of self reporting pharmacovigilance in anti tubercular therapy using knowledge based approach. *J. Pharmacovigil.* **2013**, *6*, 33-38, DOI: 10.4172/2329-6887.1000101.

[7] Tatarczak-Michalewska, M.; Flieger, J.; Wujec, M.; Swatko-Ossor, M., Isolation and quantitative determination of new tuberculostatic 1,2,4-triazole derivative in urine and plasma samples. *J. Anal. Bioanal. Tech.* **2014**, *5*, 1-2.

[8] Rathod, A., Antifungal and antibacterial activities of Imidazolylpyrimidines derivatives and their QSAR Studies under Conventional and Microwave-assisted. *Int. J. Pharm. Tech. Res.* **2011**, *3*, 1942-1951.

[9] da Silva M. C.; de Lima F. M.; de Souza, M. N.; Peralta, M. A.; Vasconcelos, T. A.; das Graças, M. O., Synthesis and anti-mycobacterial activity of (E)-N′-(monosubstituted-benzylidene) isonicotino-hydrazide derivatives. *Eur. J. Med. Chem.* **2008**, *43*, 1344-7, DOI: 10.1016/j.ejmech.2007.08.003.

[10] Li, Z.; Wan, H.; Shi,Y.; Ouyang, P., Personal experience with four kinds of chemical structure drawing software: review on ChemDraw, ChemWindow, ISIS/Draw, and Chem. Sketch. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1886-90, DOI: 10.1021/ci049794h.

[11] Singh, P., Quantitative Structure-activity relationship study of substituted-[1,2,4] oxadiazoles as S1P1 agonists. *J. Curr. Chem. Pharm. Sci.* **2013**, *3*, 64-79.

[12] Kennard, R. W.; Stone, L. A.; Computer aided design of experiments. *Technometrics. J. Sci. Res.* **1969**, *11*, 137-48, DOI: 10.1080/ 00401706.1969.10490666.

[13] Melagraki, G.; Afantitis, A.; Makridima, K.; Sarimveis, H.; Igglessi-Markopoulou, O., Prediction of toxicity using a novel RBF neural network training methodology. *J. Mol. Model.* **2006**, *12*, 297-305.

[14] Afantitis, A.; Melagraki, G.; Sarimveis, H.; Koutentis, P. A.; Markopoulos, J.; Igglessi-Markopoulou, O., A novel QSAR model for predicting induction of apoptosis by 4-aryl-4H-chromenes. *Bioorg Med Chem.* **2006**, *14*, 6686-6694.

[15] Chakraborti, A. K.; Gopalakrishnan, Sobhia, M. E.;

Malde, A., 3D-QSAR studies of indole derivatives as phosphodiesterase IV inhibitors. *Eur. J. Med. Chem.* **2003**, *38*, 975-82, DOI: 10.1016/ j.ejmech.2003.09.001.

[16] Wu, W.; Walczak, B.; Massart, D. L.; Heuerding, S.; Erni, F., Last, I. R., Artificial neural networks in classification of NIR spectral data: design of the training set. *Chemom. Intell. Lab. Syst.* **1996**, *33*, 35-46.

[17] Khaled, K. F., Modeling corrosion inhibition of iron in acid medium by genetic function approximation method: A QSAR model. *Corros. Sci.* **2011**, *53*, 3457-3465.

[18] Friedman, J. H., Multivariate adaptive regression splines. *Ann. Stat.* 1991, 1-67.

[19] Tropsha, A.; Gramatica, P.; Gombar, V. K., The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. *Mol. Inform.*. **2003**, *22*, 69-77, DOI: 10.1002/qsar.200390007.

[20] Veerasamy, R.; Rajak, H.; Jain, A.; Sivadasan, S.; Varghese, C. P.; Agrawal, R. K., Validation of QSAR models-strategies and importance. *Int. J. Drug Des. Discov.* **2011**, *3*, 511-519.

[21] Eric, G. M.; Uzairu, A.; Mamza, P. A., A quantitative structure-activity relationship (QSAR) study of the anti-tuberculosis activity of some quinolones. *J. Sci. Res.* **2016**, *10*, 1-15.

[22] Ogadimma, A. I; Adamu, U., Analysis of selected chalcone derivatives as mycobacterium tuberculosis inhibitors. *Open Access Library J.* **2016**, *3*, 1-13, DOI: 10.4236/oalib.1102432.

[23] -Jalali-Heravi, M.; Kyani, A., Use of computer-assisted methods for the modeling of the retention time of a variety of volatile organic compounds: a PCA-MLR-ANN approach. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1328-1335, DOI: 10.1021/ci0342270.