Journal of Computational Methods in Molecular Design, 2015, 5 (4):61-76

Scholars Research Library
(http://scholarsresearchlibrary.com/archive.html)

ISSN : 2231- 3176
CODEN (USA): JCMMDA

Quantitative structure and activity relationship modeling study of anti-HIV-1 RT inhibitors: Genetic function approximation and density function theory methods

Emmanuel Israel Edache*, Adamu Uzairu and Stephen Eyije Abechi

Department of Chemistry, Ahmadu Bello University, Zaria, Kaduna State-Nigeria

---

ABSTRACT

In the present work, quantitative structure activity relationship studies were performed to explore the structural and physicochemical requirements of 1-[(2-hydroxyethoxy)methyl]-6-(phenylthio)thymine (HEPT) derivatives for anti-HIV activity. QSAR models have been developed using steric, electronic and thermodynamic descriptors. Statistical techniques like genetic function approximation-multiple linear regression (GFA-MLR) as the data preprocessing step were applied to identify the structural and physicochemical requirements for anti-HIV activity. The generated equations were statistically validated using leave-one-out technique and the best models were also subjected to leave-5-out cross-validation. The quality of fit and predictive ability of equations obtained from GFA-MLR is of acceptable statistical range (explained variance of 91.74%, while predicted variance of 74.14%). The robustness of the best models was checked by Y–randomization test and identified as good predictive models. The coefficient of ALogP, ATSm5 and CrippenLogP shows that the activity increases with increase in ALogP, ATSm5 and Crippen LogP of molecules. The coefficient of C2SP3, VPC-4, SsI, ETA_AlphaP, ETA_Epsilon_1, nAtomP, Petitjean Number and Wlambda2.unity shows that the activity decreases with increase in volume and Wlambda1.polar of the molecules is detrimental to activity. The information generated from the present study may be useful in the design of more potent HEPT derivatives as anti HIV agents.

Keywords: Anti-HIV-1, QSAR, Validation, Internal validation, External validation, Randomization, Applicability domain.

INTRODUCTION

Human immunodeficiency infection (HIV) is the chief reason for (AIDS). In the most recent decades, numerous anti-HIV medications have been created, yet the most these medications have issues like the genuine unfavorable symptoms of the accessible medications and the rise of medication resistance, in light of the fact that exploration to find and create extra novel non-nucleoside reverse transcriptase inhibitors (NNRTIs) drugs with different atomic frameworks for more viable treatment and potential AIDS prevention [1].

Quantitative structure activity relationships (QSARs) [2] have turned into an imperative part in the compound outline and movement process since they speak to a much less expensive, quick different option for the medium throughput in vitro and low throughput in vivo measures which are by and large limited to later in the revelation course [3]. A QSAR is basically a numerical comparison that is resolved from an arrangement of molecules with known exercises utilizing computational methodologies [4]. The careful type of the relationship in the middle of

structure and action can be resolved utilizing an assortment of measurable techniques and processed sub-atomic descriptors and this mathematical statement is then used to foresee the action of new particles. Early QSARs spearheaded by Hanch and Fugita [5, 6] comprised of moderately little number of molecules of a given chemotype being utilized to infer a basic direct comparison to foresee the following particle in the arrangement to be combined. The upside of this methodology was that the terms in the mathematical statement were for the most part straightforward and effectively interpretable, while the sorts of atoms being anticipated were for the most part fundamentally the same to those that were at that point incorporated, giving the client more prominent trust in the model expectations. Conversely, over the previous decade an expanding number of QSARs have been accounted for taking into account extensive, assorted datasets, normally termed worldwide models, which are viewed as more dependable at anticipating differing structures than QSARs based on little datasets of low differences [7, 8]. A few QSAR Studies have been performed by different creators, which give important bits of knowledge in configuration and advancement of HIV-1RT inhibitors [9, 10]. As a piece of progressing exertion the present work is planned to infer some factually huge QSAR models for HEPT subsidiaries to associate against HIV-1 RT action to its physicochemical properties. The outcomes got may add to further plan and improvement of novel antiretroviral specialism.

MATERIALS AND METHODS

Data set

In present studies, a series of HEPT derivatives, reported by [11] as potent anti-HIV, was selected. One hundred and six compounds were divided (using Kennard-Stone Selection) into training and test set, the former set consisting of seventy four compounds and the remaining thirty two compounds were taken as the test set.

Biological activities

Structures of all the compounds used for QSAR analysis and their anti-HIV activity (EC50, molar concentration of the drug required achieving 50% protection of MT-4 cells against the cytopathic effect of virus) are given in Table 1. For every compound of the series, the experimental values of biological activity are used in the negative logarithmic scale (pEC50) to achieve normal distribution.

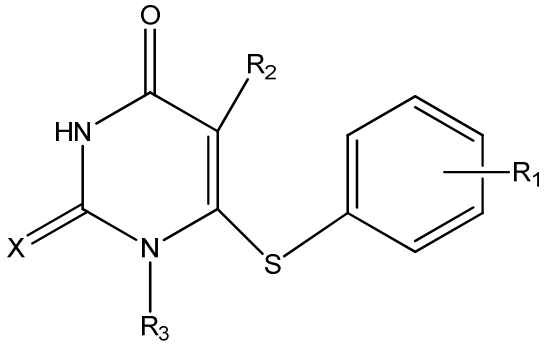

Figure 1: The chemical structure of the compounds used in this study

Table 1: The observed activity data of the compounds used in this study

No.	R1	R2	R3	X	Obs.	Ypred	Residual
1*	3-CN	Me	$CH_2OCH_2CH_2OH$	O	5.000	3.979	1.021
2	3-COMe	Me	$CH_2OCH_2CH_2OH$	O	5.140	5.494	-0.354
3*	3-COOMe	Me	$CH_2OCH_2CH_2OH$	O	5.100	5.777	-0.677
4	3,5-Cl_2	Me	$CH_2OCH_2CH_2OH$	O	5.890	6.232	-0.342
5	3,5-Me_2	Me	$CH_2OCH_2CH_2OH$	O	6.590	6.092	0.498
6	3-OMe	Me	$CH_2OCH_2CH_2OH$	O	4.660	4.389	0.271
7	3-OH	Me	$CH_2OCH_2CH_2OH$	O	4.090	4.498	-0.408
8*	3-NO_2	Me	$CH_2OCH_2CH_2OH$	O	4.470	3.618	0.852
9	3-I	Me	$CH_2OCH_2CH_2OH$	O	5.000	5.157	-0.157
10	3-Br	Me	$CH_2OCH_2CH_2OH$	O	5.240	5.598	-0.358
11	3-Cl	Me	$CH_2OCH_2CH_2OH$	O	4.890	5.060	-0.170
12	3-F	Me	$CH_2OCH_2CH_2OH$	O	5.480	4.788	0.692
13	3-CF_3	Me	$CH_2OCH_2CH_2OH$	O	4.350	4.097	0.253
14	3-t-Bu	Me	$CH_2OCH_2CH_2OH$	O	4.920	4.882	0.038

15	3-Et	Me	$CH_2OCH_2CH_2OH$	O	5.570	4.901	0.669
16	3-Me	Me	$CH_2OCH_2CH_2OH$	O	5.590	5.366	0.224
17	2-OMe	Me	$CH_2OCH_2CH_2OH$	O	4.720	5.285	-0.565
18	2-NO_2	Me	$CH_2OCH_2CH_2OH$	O	3.850	4.529	-0.679
19	2-Me	Me	$CH_2OCH_2CH_2OH$	O	4.150	4.434	-0.284
20	H	Et	$CH_2OCH_2CH_2OH$	O	6.920	6.688	0.232
21	H	i-Pr	$CH_2OCH_2CH_2OH$	O	7.200	6.819	0.381
22	3,5-Me_2	Et	$CH_2OCH_2CH_2OH$	O	7.890	6.730	1.160
23	3,5-Me_2	i-Pr	$CH_2OCH_2CH_2OH$	O	8.570	7.658	0.912
24	3,5-Cl_2	Et	$CH_2OCH_2CH_2OH$	O	7.850	7.356	0.494
25	H	Me	$CH_2OCH_2CH_2OH$	O	5.150	4.842	0.308
26	H	I	$CH_2OCH_2CH_2OH$	O	5.440	5.667	-0.227
27	H	$CH=CPH_2$	$CH_2OCH_2CH_2OH$	O	6.070	6.649	-0.579
28	4-F	Me	$CH_2OCH_2CH_2OH$	O	3.600	3.822	-0.222
29	4-Cl	Me	$CH_2OCH_2CH_2OH$	O	3.600	3.466	0.134
30	4-OH	Me	$CH_2OCH_2CH_2OH$	O	3.560	3.562	-0.002
31	3-$CONH_2$	Me	$CH_2OCH_2CH_2OH$	O	3.510	2.899	0.611
32*	H	COOMe	$CH_2OCH_2CH_2OH$	O	5.180	4.844	0.336
33	H	CONHPh	$CH_2OCH_2CH_2OH$	O	4.740	4.160	0.580
34*	H	SPh	$CH_2OCH_2CH_2OH$	O	4.840	5.119	-0.279
35	H	CCH	$CH_2OCH_2CH_2OH$	O	4.740	5.010	-0.270
36	H	CCPh	$CH_2OCH_2CH_2OH$	O	5.470	4.990	0.480
37*	H	$COCHMe_2$	$CH_2OCH_2CH_2OH$	O	4.920	5.908	-0.988
38	H	COPh	$CH_2OCH_2CH_2OH$	O	4.890	4.959	-0.069
39*	H	CCMe	$CH_2OCH_2CH_2OH$	O	4.720	5.862	-1.142
40	H	F	$CH_2OCH_2CH_2OH$	O	4.000	4.326	-0.326
41	H	Cl	$CH_2OCH_2CH_2OH$	O	4.520	3.710	0.810
42	H	Br	$CH_2OCH_2CH_2OH$	O	4.700	4.637	0.063
43	2-Cl	Me	$CH_2OCH_2CH_2OH$	O	3.890	3.813	0.077
44	3-CH_2OH	Me	$CH_2OCH_2CH_2OH$	O	3.530	3.726	-0.196
45	4-NO_2	Me	$CH_2OCH_2CH_2OH$	O	3.720	3.691	0.029
46	4-CN	Me	$CH_2OCH_2CH_2OH$	O	3.600	4.092	-0.492
47	4-OMe	Me	$CH_2OCH_2CH_2OH$	O	3.600	4.514	-0.914
48	4-COMe	Me	$CH_2OCH_2CH_2OH$	O	3.960	4.230	-0.270
49*	4-COOH	Me	$CH_2OCH_2CH_2OH$	O	3.450	3.418	0.032
50*	3-NH_2	Me	$CH_2OCH_2CH_2OH$	O	3.600	4.267	-0.667
51	H	Pr	$CH_2OCH_2CH_2OH$	O	5.470	6.296	-0.826
52	4-Me	Me	$CH_2OCH_2CH_2OH$	O	3.660	4.644	-0.984
53*	H	$CH=CH_2$	$CH_2OCH_2CH_2OH$	O	5.690	6.617	-0.927
54	H	CH=CHPh	$CH_2OCH_2CH_2OH$	O	5.220	5.520	-0.300
55*	H	CH_2Ph	$CH_2OCH_2CH_2OH$	O	4.370	6.843	-2.473
56*	H	Me	$CH_2OCH_2CH_2OAc$	O	5.170	5.169	0.001
57*	H	Et	CH_2OCH_2Me	O	7.720	6.630	1.090
58*	H	Et	CH_2CH_2Ph	O	8.230	8.625	-0.395
59	3,5-Cl_2	Et	CH_2CH_2Me	O	8.130	8.689	-0.559
60	H	Me	$CH_2OCH_2CH_2OC_5H_{11}$	O	4.460	4.666	-0.206
61	H	Me	$CH_2OCH_2CH_2OCH_2Ph$	O	4.700	5.730	-1.03
62	H	Me	H	O	3.600	3.875	-0.275
63	H	Me	Me	O	3.820	3.925	-0.105
64	H	c-Pr	CH_2OCH_2Me	O	7.000	6.954	0.046
65	H	Et	CH_2O-i-Pr	O	6.470	6.430	0.04
66*	H	Et	CH_2O-c-Hex	O	5.400	5.177	0.223
67	H	Et	CH_2OCH_2-c-Hex	O	6.350	5.943	0.407
68	H	Et	$CH_2OCH_2CH_2Ph$	O	7.020	7.041	-0.021
69*	H	Me	CH_2OMe	O	5.680	5.721	-0.041
70	H	Me	CH_2OBu	O	5.330	5.445	-0.115
71*	H	Me	Et	O	5.660	5.416	0.244
72	H	Me	Bu	O	5.920	5.611	0.309
73	H	i-Pr	CH_2OCH_2Me	O	7.990	7.496	0.494
74*	H	i-Pr	CH_2OCH_2Ph	O	8.510	8.426	0.084
75*	3,5-Me_2	Et	CH_2OCH_2Ph	O	8.550	8.617	-0.067
76	3,5-Me_2	Et	CH_2OCH_2Me	O	8.240	8.680	-0.44
77*	H	Me	$CH_2OCH_2CH_2OMe$	O	5.060	5.214	-0.154
78	H	Me	$CH_2OCH_2CH_2OCOPh$	O	5.120	5.023	0.097
79	H	Me	CH_2OCH_2Me	O	6.480	5.704	0.776
80*	H	Me	$CH_2OCH_2CH_2Cl$	O	5.820	5.569	0.251

81	H	Me	$CH_2OCH_2CH_2N_3$	O	5.240	4.924	0.316
82	H	Me	$CH_2OCH_2CH_2F$	O	5.960	5.773	0.187
83	H	Me	$CH_2OCH_2CH_2Me$	O	5.480	6.084	-0.604
84	H	Me	CH_2OCH_2Ph	O	7.060	6.718	0.342
85*	H	Et	CH_2OCH_2Me	S	7.580	7.289	0.291
86	H	i-Pr	CH_2OCH_2Me	S	7.890	8.026	-0.136
87*	H	i-Pr	CH_2OCH_2Ph	S	8.140	9.365	-1.225
88*	3,5-Cl_2	Et	CH_2OCH_2Me	S	7.890	8.576	-0.686
89	H	Et	CH_2O-i-Pr	S	6.660	7.041	-0.381
90*	H	Et	CH_2O-c-Hex	S	5.790	5.757	0.033
91	H	Et	CH_2OCH_2-c-Hex	S	6.450	6.979	-0.529
92	H	Et	$CH_2OCH_2C_6H_4$(4-Cl)	S	7.920	6.602	1.318
93*	H	Et	$CH_2OCH_2CH_2Ph$	S	7.040	7.855	-0.815
94	H	c-Pr	CH_2OCH_2Me	S	7.020	7.542	-0.522
95	3,5-Me_2	Me	$CH_2OCH_2CH_2OH$	S	6.660	6.629	0.031
96*	H	Pr	$CH_2OCH_2CH_2OH$	S	5.000	6.794	-1.794
97	3,5-Me_2	i-Pr	$CH_2OCH_2CH_2OH$	S	8.300	8.597	-0.297
98*	H	Et	CH_2OCH_2Ph	S	8.090	8.700	-0.61
99	3,5-Me	Et	CH_2OCH_2Ph	S	8.140	9.268	-1.128
100*	3,5-Me_2	Et	CH_2OCH_2Me	S	8.300	8.171	0.129
101	H	Et	$CH_2OCH_2CH_2OH$	S	6.960	6.524	0.436
102	H	i-Pr	$CH_2OCH_2CH_2OH$	S	7.230	7.355	-0.125
103*	3,5-Me_2	Et	$CH_2OCH_2CH_2OH$	S	8.110	7.584	0.526
104*	3,5-Cl_2	Et	$CH_2OCH_2CH_2OH$	S	7.370	7.936	-0.566
105	H	Me	$CH_2OCH_2CH_2OH$	S	6.010	5.318	0.692
106*	H	$CH_2CH=CH_2$	$CH_2OCH_2CH_2OH$	O	5.600	6.594	-0.994

**Test set*

Computational Details

The two-dimensional structures of molecules were drawn in Spartan'14 version 1.1.2 [12] software, converted to 3D and also many numbers of theoretical molecular descriptors such as HOMO, LUMO, Aqueous Energy, Energy, volume, Gibb's Energy, log P, formation enthalpy and other quantum descriptors have been computed with the Spartan'14 version 1.1.2. ALOGP Descriptor, APol Descriptor, Aromatic Atoms Count Descriptor, Aromatic Bonds Count Descriptor, Atom Count Descriptor and other descriptors have been computed with the PaDEL-Descriptors version 2.18 [13]. These descriptors used to modeling Quantitative structure–activity relationship of HEPT derivatives. The equilibrium geometries of all HEPT derivatives were fully optimized using the DFT/B3LYP method [14] with the 6-311G* basis set. No molecular symmetry constraint was applied, rather full optimization of all bond lengths and angles was carried out. The calculated descriptors for each molecule are summarized in Table 2. The GFA-multiple linear regression statistic technique is used to study the relation between one dependent variable and several independent variables. It is a mathematic technique that minimizes the differences between actual and predicted values. The GFA-multiple linear regression model (GFA-MLR) was generated using the software Material Studio to predict EC50.

Model development

A model's predictive accuracy and confidence for different unknown chemicals differs according to how well the training set signifies the unknown chemicals and how robust the model is in generalizing beyond the chemistry space defined by the training set. So, the selection of the training set is significantly important in QSAR analysis. Predictive potential of a model on the new data set is influenced by the similarity of chemical nature between training set and test set [15]. The test set molecules will be predicted well when these molecules are very similar to the training set compounds. The reason is that the model has represented all features common to the training set molecules. In this paper, for the development of models for a particular data set, Kennard-Stone method were employed. This approach (Kennard-Stone method) ensures that the similarity principle can be employed for the activity prediction of the test set. Based on Kennard-Stone, each data set was divided into training and test sets. In each case, 70% of the total compounds were selected as training set and remaining 30% were selected as test set. Models were developed from a training set using GFA-MLR and the best model was selected from the population of models obtained based on lack-of-fit score. The selected model was then validated internally by leave-one-out method and then externally by predicting the activity values of the corresponding test set. Based on the results obtained from multiple models which are derived based on different combinations of training and test sets, we have tried to evaluate performance of different validation parameters.

Statistical methods

GFA

In GFA, a particular number of equations (set at 100 by default) are generated randomly. Then p appearances of "parent" equations are chosen randomly from this set of 100 equations and "crossover" operations were performed at random. The number of crossing over was set by default at 5000. The goodness of each progeny equation is assessed by Friedman's lack of fit (LOF) score, which is given by following formula:

$$LOF = \frac{LSE}{\left(1-\frac{c+d\times p}{M}\right)^2} \quad (1)$$

Where LSE is the least-squares error, c is the number of basis functions in the model, d is smoothing parameter, p is the number of descriptors and m is the number of observations in the training set. The smoothing parameter, which controls the scoring bias between equations of different sizes, was set at default value of 0.5 and GFA crossover of 5000 were set to give reasonable convergence. The length of equation was fixed to twelves terms, the population size was established as 2000, the equation term was set to linear polynomial and the mutation probability was specified as 0.1. It has been shown that a high value of statistical characteristics r and F and low value of s and LOF need not be the proof of a highly predictive model. Hence, in order to evaluate the predictive ability of the QSAR model, the method described by Roy et al [16] and Golbraikh and Tropsha [17] and for external predictability was used. It was determined by calculating the value of predictive R^2 (R^2_{pred}) using the following equation.

Validation parameters

R^2

The coefficient of determination (R^2) indicates the quality of fit and is calculated as:

$$R^2 = 1 - \frac{\sum(Y_{obs}-Y_{calc})^2}{\sum(Y_{obs}-\bar{Y}_{obs})^2} \quad (2)$$

In the above equation, Yobs stands for the observed response value, while Ycalc is the model-derived calculated response and $\bar{Y}_{obs}$ is the average of the observed response values. For the ideal model, the sum of squared residuals being 0, the value of R^2 is 1. As the value of R^2 deviates from 1, the fitting quality of the model deteriorates. The square root or R^2 is the multiple correlation coefficient (R).

Total sum of squares (TSS) and MSE

It is the total variance that a regression model can explain and is used as a reference quantity to calculate standardized quality parameters. Also denoted as *SSY*, it is the sum of the squared differences between the experimental responses and the average experimental response:

$$TSS = SSY = \sum(Y_{obs} - \bar{Y}_{train})^2 \quad (3)$$

TSS is assumed as a theoretical reference model where for each experimental response a constant value is calculated as the average experimental response. *TSS* depends on the measure unit used for the response. *MSE* represents the standard distance data values far from the regression line. For a given study, the better the equation predicts the response, the lower *MSE*.

$$MSE = \frac{\sum(Y_{obs}-\bar{Y}_{train})^2}{N-p} \quad (4)$$

R^2_{adj} and F-test

The statistical qualities of the equations were judged by the parameters such as explained variance (R^2_{adj}), determination coefficient (R^2) and variance ratio (F) at specified degrees of freedom (df) [18]. R^2_{adj} is defined as

$$R^2_{adj} = \frac{(N-1)R^2-p-1}{N-P-1} \quad (5)$$

If one goes on increasing the number of descriptors in a model for a fixed number of observations, R^2 values will always increase, but this will lead to a decrease in the degree of freedom and low statistical reliability. Therefore, a high value of R^2 is not necessarily as indication of a good statistical model that fits well the available data. To reflect

the explained variance (the fraction of the data variance explained by the model) in a better way. In the above expression, p is the number of predictor variables used in the model development. *F*-ratio test is the most well-known statistical tests, this is defined as:

$$F = \frac{\frac{\sum(Y_{cal}-\bar{Y}_{train})^2}{p}}{\frac{\sum(Y_{obs}-Y_{cal})^2}{N-p-1}} \quad (6)$$

The F value has two degrees of freedom: p, N − p − 1. The computed F value of a model should be significant at P < 0.05. For overall significance of the regression coefficients, the F value should be high.

Standard error of estimate (SEE) and Quality index or factor (Q)
For a good model, the standard error of estimate of Y should be low and this is defined as follows:

$$SEE = \sqrt{\frac{(Y_{obs}-Y_{calc})^2}{N-p-1}} \quad (7)$$

It has a degree of freedom of N − p − 1. In 1994, a quality factor Q [19, 20]] for regression was defined as:

$$Q = \frac{R}{SEE} \quad (8)$$

"This quality factor Q is defined as the ratio of the correlation coefficient (*R*) to the standard error of estimate. This factor accounts for the predictive power of the model." As it can be easily observed, none of the parameters in the quality index definition is in some way related to the prediction power of the model, but is (of course) related to *R*.

The Predicted residual sum of squares (PRESS) and Standard deviation error of prediction (SDEP)
The *PRESS* (predicted residual sum of squares) statistic appears to be the most important parameter for a good estimate of the real predictive error of the models. Its small value indicates that the model predicts better than chance and can be considered statistically significant. It is calculated by following equation [21]:

$$PRESS = \sum\left(Y_{obs} - Y_{pred}\right)^2 \quad (9)$$

$$SDEP = \sqrt{\frac{PRESS}{N}} \quad (10)$$

N refers to the number of observation

$Q^2_{(LOO)}$ and $Q^2_{(LMO)}$
In case of leave-one-out (LOO) cross-validation, each member of the sample in turn is removed, the full modeling method is applied to the remaining *n*-1 members, and the fitted model is applied to the holdback member. The LOO approach perturbs the data structure by removing 1/Nth compound in each cross-validation round, thus, accomplishing an increasingly smaller perturbation with increasing N. Hence, the Q2 value of LOO approaches to that of R2, which is highly unsatisfactory [22]. Cross-validated squared correlation coefficient R2 (LOO-Q2) is calculated according to the formula:

$$Q^2_{LOO} = 1 - \frac{\sum(Y_{pred}-Y)^2}{\sum(Y-\bar{Y})^2} \quad (11)$$

Ypred and Y indicate predicted and observed activity values respectively and *Y* indicate mean activity value. A model is considered acceptable when the value of Q^2 exceeds 05. The model predictivity is judged using the predicted residual sum of square (PRESS) and Q^2 for the model while the value of standard deviation of error of prediction (SDEP) is calculated from PRESS.

The basic principle of the leave-many-out (LMO) technique or leave-Group-out (LGO) technique is that a definite portion of the training set is held out and eliminated in each cycle. For each cycle, the model is constructed based on the remaining molecules (and using the originally selected descriptors) and then the activity of the deleted

compounds is predicted using the developed model. After all the cycles have been completed, the predicted activity values of the compounds are used for the calculation of the LMO-Q^2.

R^2_{pred} and $Q^2_{(F2)}$

Cross validation provides a reasonable approximation of ability with which the QSAR predicts the activity values of new compounds. However, external validation gives the ultimate proof of the true predictability of a model. In many cases, truly external data points being unavailable for prediction purpose, original data set compounds are divided into training and test sets [16], therefore enabling external validation. This division of the data set can be accomplished in many ways, but approximately similar ranges of the biological responses and structural properties and all available structural and/or physicochemical features should be represented in both training and test sets. Equations are generated based on training set compounds and predictive capacity of the models is judged based on the predictive R^2 (R^2_{pred}) values calculated according to the following equation:

$$R^2_{pred} = 1 - \frac{\sum(Y_{pred(test)} - Y_{(test)})^2}{\sum(Y_{(test)} - \bar{Y}_{train})^2} \quad (12)$$

Ypred(test) and Y(test) indicate predicted and observed activity values respectively of the test set compounds and *Y* training indicates mean activity value of the training set. For a predictive QSAR model, the value of R^2_{pred} should be more than 0.5. $Q^2_{(F2)}$ is based on prediction of test set compounds ($Q^2_{(F2)}$) proposed by Schüürmann et al. [23] as given by equation below:

$$Q^2_{(F2)} = 1 - \frac{\sum(Y_{obs(test)} - Y_{pred(test)})^2}{\sum(Y_{obs(test)} - \bar{Y}_{test})^2} \quad (13)$$

Here, $\bar{Y}_{test}$ refers to the mean observed data of the test set compounds. A threshold value 0.5 is defined for this parameter.

$r2m$

It has been earlier shown [24] that $R2pred$ may not be sufficient to indicate external predictivity of a model. It may not truly reflect the predictive capability of the model on a new dataset. Besides this, a good value of squared correlation coefficient ($r2$) between observed and predicted values of the test set compounds does not necessarily mean that the predicted values are very near to corresponding observed activity (there may be considerable numerical difference between the values though maintaining an overall good inter-correlation). So, for better external predictive potential of the model, a modified $r2$ [$r2m(test)$] was introduced by the following equation [24]:

$$r^2_{m(test)} = r^2 \times \left(1 - \sqrt{r^2 - r_0^2}\right) \quad (14)$$

Where $r02$ is squared correlation coefficient between the observed and predicted values of the test set compounds with intercept set to zero. The value of $r2m(test)$ should be greater than 0.5 for an acceptable model. Initially, the concept of $r2m$ was applied only to the test set prediction [15], but it can as well be applied for training set if one considers the correlation between observed and leave-one-out (LOO) predicted values of the training set compounds [25]. More interestingly, this can be used for the whole set considering LOO-predicted values for the training set and predicted values of the test set compounds. The $r2m(overall)$ statistic may be used for selection of the best predictive models from among comparable models.

The $r2m$ metric for internal validation

An acceptable value of $Q^2_{(LOO)}$ does not inevitably indicate that the predicted activity data lie in close proximity to the observed ones although there may exist a good overall correlation between the values. Therefore, to avoid this problem and to better indicate the model predictability, the $r2m$ metrics introduced by Roy et al. [16] may be computed by the following equations:

$$\bar{r}^2_m = \frac{(r^2_m + r'^2_m)}{2} \quad (15)$$

$$\Delta r_m^2 = \left| r_m^2 - r_m'^2 \right| \quad (16)$$

The $\bar{r}_m^2$ is the average value of r_m^2 and$r_m'^2$, and Δr_m^2 is the absolute difference between r_m^2 and$r_m'^2$. In case of internal validation of the training set, the $\bar{r}_{m(LOO)}^2$ and $\Delta r_{m(LOO)}^2$ parameters can be employed and it has been shown that the value of $\Delta r_{m(LOO)}^2$ should preferably be lower than 0.2 provided that the value of $\bar{r}_{m(LOO)}^2$ is more than 0.5. Roy et al [16].

R^2_p

Further statistical significance of the relationship between activity and the descriptors can be checked by randomization test (Y-randomization) of the models. This method is of two types: process randomization and model randomization. In case of process randomization, the values of the dependent variable are randomly scrambled and variable selection is done freshly from the whole descriptor matrix. In case of model randomization, the Y column entries are scrambled and new QSAR models are developed using same set of variables as present in the unrandomized model. For an acceptable QSAR model, the average correlation coefficient (*Rr*) of randomized models should be less than the correlation coefficient (R) of non-randomized model. We have used a parameter *R2p* [26] in the present paper, which penalizes the model *R2* for the difference between squared mean correlation coefficient (Rr2) of randomized models and squared correlation coefficient (R2) of the non-randomized model. The above mentioned novel parameter can be calculated by the following equation:

$$R_p^2 = R^2 \times \sqrt{R^2 - R_r^2} \quad (17)$$

This novel parameter Rp2 ensures that the models thus developed are not obtained by chance. We have assumed that the value of Rp2 should be greater than 0.5 for an acceptable model.

Golbraikh and Tropsha's criteria [17] proposed a set of parameters for determining the external predictability of QSAR model. According to Golbraikh and Tropsha, models are considered satisfactory, if all of the following conditions are satisfied

i) $Q_{LOO}^2 > 0.5$
ii) $R_{pred}^2 > 0.6$
iii) $\frac{r^2 - r_0^2}{r^2} < 0.1$ and $0.85 \leq k \leq 1.15$
iv) $\frac{r^2 - r_0'^2}{r^2} < 0.1$ and $0.85 \leq k' \leq 1.15$
v) $\left| r_0^2 - r_0'^2 \right| < 0.3$

RESULTS AND DISCUSSION

For the selection of the most important descriptors, GFA multiple regression techniques were used. Firstly, the GFA-MLR analysis selection and the variables elimination was employed to model the QSARs with a different set of descriptors. In order to build and test model, a data set of 106 compounds was separated (using Kennard-Stone method) [27] into a training set of 74 compounds (70%), which was used to build model and a prediction set of 32 compounds (30%), which was applied to test the built model. The selection of the test set molecules was with respect to distribution in the range of the biological data for the whole set, and their structure diversity. The GFA-MLR analysis led to the derivation of two model, with twelve (12) variables, the next to the ratio of five training molecules for each descriptor [28] with low generality and prediction ability for the test set. It is described by the following equation:

Model 1

$$pC50 = 47.40398(\pm 4.23076) + 0.56731(\pm 0.0973)\ ALogP + 0.11102(\pm 0.00929)\ ATSm5 - 0.21722(\pm 0.07907)\ C2SP3 - 2.26655(\pm 0.32454)\ VPC - 4 + 0.58171(\pm 0.10746)\ CrippenLogP - 5.95183(\pm 1.49219)\ SsI - 40.28928(\pm 6.29647)\ ETA\_AlphaP - 26.31724(\pm 2.21578)\ ETA\_Epsilon\_1 -$$

$0.26092(\pm 0.02824)\ nAtomP\ -\ 11.8987(\pm 2.85944)\ PetitjeanNumber\ -$
$0.18179(\pm 0.07087)\ Wlambda2.unity\ -\ 0.13964(\pm 0.01609)\ Wlambda1.polar.$

$N\ =\ 74, LOF = 0.9812,$
$SEE\ =\ 0.4634, r^2\ =\ 0.9174, r^2 adjusted\ =\ 0.9011, F\ =\ 56.4427\ (DF: 12, 61), Q2(LOO) =$
$0.8840, PRESS\ =\ 18.3834, SDEP(LOO) =\ 0.4984, r2m^{(Loo)} =\ 0.8651, rm^{2'(Loo)} =$
$0.8012, average\ rm^{2(LOO)} =\ 0.8331, delta\ rm^{2(LOO)} =\ 0.0639, Q^2(L5O) = 0.8729, SDEP\ (L5O) =$
$0.1357, R_p^2 = 0.8289.$

External Validation Parameters (Without Scaling):
$r^2 = 0.76686, r0^2 = 0.75244, reverse\ r0^2 = 0.7522, rm^{2(test)} = 0.67476, reverse\ rm^{2(test)} =$
$0.674, average\ rm^{2(test)} = 0.67438, delta\ rm^{2(test)} = 0.00075, rmsep = 0.83033, rpred^2 =$
$0.74141, Q2f1 = 0.74141, Q2f2 = 0.70698.$

Overall Parameters:
$rm\text{^}2(overall)\ =\ 0.78227, reverse\ rm\text{^}2(overall)\ =\ 0.77117, average\ rm\text{^}2(overall)\ =$
$0.77672, delta\ rm\text{^}2(overall)\ =\ 0.0111.$

Golbraikh and Tropsha acceptable model criteria's:
1. $Q^2\ =\ 0.88404, Passed\ \ (Threshold\ value\ Q^2 > 0.5)$
2. $r^2\ =\ 0.76686, Passed\ \ (Threshold\ value\ r^2 > 0.6)$
3. $\left|r0^2 - r'0^2\right| =\ 0.00024, Passed\ \ \left(Threshold\ value\ \left|r0^2 - r'0^2\right| < 0.3\right)$
4. $k\ =\ 0.95057,$
$[(r\text{^}2 - r0\text{^}2)/r\text{^}2]\ =\ 0.01881\ OR\ k'\ =\ 1.03673\ [(r\text{^}2 - r'0\text{^}2)/r\text{^}2]\ =$
$0.01912 Passed\ (Threshold\ value: [0.85 < k < 1.15\ and\ ((r\text{^}2 - r0\text{^}2)/r\text{^}2) < 0.1\ ]\ OR\ [0.85 < k' <$
$1.15\ and\ ((r\text{^}2 - r'0\text{^}2)/r\text{^}2) < 0.1]\ )$

Model 2
$pEC50\ =\ 45.13546(\pm 3.86674)\ +\ 0.59311(\pm 0.09875)\ ALogP\ +\ 0.10689(\pm 0.00898)\ ATSm5\ -$
$0.2022(\pm 0.07942)\ C2SP3\ -\ 2.28321(\pm 0.32573)\ VPC\ -\ 4\ +\ 0.49616(\pm 0.10548)\ CrippenLogP\ -$
$6.32113(\pm 1.53778)\ SsI\ -\ 36.86178(\pm 5.74698)\ ETA\_AlphaP\ -\ 25.89084(\pm 2.18645)\ ETA\_Epsilon\_1\ -$
$0.26897(\pm 0.02847)\ nAtomP\ -\ 11.24469(\pm 2.85923)\ PetitjeanNumber\ -$
$0.12202(\pm 0.05051)\ Wlambda2.mass\ -\ 0.12557(\pm 0.01477)\ Wlambda1.polar.$
$N = 74, LOF = 0.9921, SEE\ =\ 0.46595, r2\ =\ 0.91646, r2\ adjusted\ =\ 0.90003, F\ =\ 55.76538\ (DF:$
$12, 61), Q2: 0.88512, PRESS\ =\ 18.2119, SDEP\ =\ 0.49609, rm\text{^}2(Loo)\ =\ 0.86761, rm\text{^}2'(Loo)\ =$
$0.80139, average\ rm\text{^}2(LOO)\ =\ 0.8345, delta\ rm\text{^}2(LOO)\ =\ 0.06622.$

External Validation Parameters (Without Scaling):
$r^2 = 0.75421, r0^2 = 0.73646, reverse\ r0^2 = 0.73932, rm^{2(test)} = 0.65372, reverse\ rm^{2(test)} =$
$0.66218, average\ rm^{2(test)} = 0.65795, delta\ rm^{2(test)} = 0.00846, rmsep = 0.8666, rpred^2 =$
$0.71833, Q2f1 = 0.71833, Q2f2 = 0.68082.$

Overall Parameters:
$rm^{2(overall)} = 0.77242, reverse\ rm^{2(overall)} = 0.76688, average\ rm^{2(overall)} = 0.76965, delta\ rm^{2(overall)} =$
$0.00554.$

Golbraikh and Tropsha acceptable model criteria's:
1. $Q\text{^}2 = 0.88512 Passed\ (Threshold\ value\ Q\text{^}2 > 0.5).$
2. $r^2 = 0.75421 Passed\ (Threshold\ value\ r^2 > 0.6)$
3. $\left|r0^2 - r'0^2\right| = 0.00286, Passed\ \left(Threshold\ value\ \left|r0^2 - r'0^2\right| < 0.3\right)$
4. $k\ =\ 0.94558\ [(r\text{^}2 - r0\text{^}2)/r\text{^}2]\ =\ 0.02354\ OR\ k'\ =\ 1.04122, [(r\text{^}2 - r'0\text{^}2)/r\text{^}2]\ =$
$0.01974, Passed\ (Threshold\ value: [0.85 < k < 1.15\ and\ ((r\text{^}2 - r0\text{^}2)/r\text{^}2) < 0.1\ ]\ OR\ [0.85 < k' <$
$1.15\ and\ ((r\text{^}2 - r'0\text{^}2)/r\text{^}2) < 0.1]\ )$

Table 2: Description about selected variables

Descriptors	Definition	VIF*	MF**
ALogP	Ghose-Crippen LogKow	2.40446	-0.02121
ATSm5	ATS autocorrelation descriptor, weighted by scaled atomic mass	3.234091	-0.14507
C2SP3	Singly bound carbon bound to two other carbons	2.974552	0.002948
VPC-4	Valence path cluster, order 4	2.149982	0.094357
CrippenLogP	Crippen's LogP	3.754902	-0.0327
SsI	Sum of sI E-states	1.150876	-0.00016
ETA_AlphaP	Sum of alpha values of all non-hydrogen vertices of a molecule relative to molecular size	3.261837	0.466105
ETA_Epsilon_1	A measure of electronegative atom count	2.477218	0.396611
nAtomP	Number of atoms in the largest pi system	1.630305	0.054386
PetitjeanNumber	Petitjean number	1.30256	0.134718
Wlambda2.unity	Directional WHIM, weighted by unit weights	1.925111	0.018152
Wlambda1.polar	Directional WHIM, weighted by atomic polarizabilities	1.462233	0.031869

**Variation Inflation Factor; **Mean effect*

In this equation, N is the number of compounds, R^2 is the squared correlation coefficient, SDEP is the standard deviation error of prediction, Q^2(LOO) and Q^2(LMO) are the squared cross-validation coefficients for leave one out and leave many out. F is the Fisher F statistic. The genetic algorithm was used to select the best set of variables. The best model has twelve parameters because the increase in the number of molecular descriptors has no significant effect on the accuracy of the best model. After the selection of the most important descriptors by genetic algorithm, MLR was performed to build the linear model. This equation and its statistical parameters are presented in model 1 and 2. With the test set, the prediction results were obtained. The experimental and predicted values based on the GA-MLR model are shown in Table 1. Also, Fig. 2 shows the predicted versus experimental pIC_{50} for all of the 106 compounds studied, the training set and the test set. As can be seen, the predicted values for the pEC_{50} are in good agreement with those of the experimental values. As can be seen from model (1) and (2), the R^2_{pred} Q^2_{f1} *and* Q^2_{f2} values in test set improved from 0.7183, 0.7183 and 0.6808 respectively by GA-MLR model. The results illustrated show successful variable selection procedure is adequate to generate an efficient QSAR model for predicting the pEC_{50} of compounds.

Figure 2: Shows the predicted versus observed pEC_{50} for all of the 106 compounds studied, the training set and the test set

Evaluation of the GA-MLR model

The quality of the QSAR model was characterized by the number of compounds used in the study (N), coefficient of determination (R^2), root mean square error (RMSE), and variance ratio (F). For a more exhaustive testing of the predictive power of the model, validation of the model was also carried out using the leave one out (LOO) and the leave many out (LMO) cross-validation techniques on the training set of compounds. For LOO cross-validation, a data point is removed from the set, and the model is recalculated. The predicted pIC_{50} for that point is then

compared with its actual value. This is repeated until each data point has been omitted once. For LMO, 8% of the data points are removed from the dataset and the model was refitted; the predicted values for those points were then compared with the experimental values. The results produced by the LOO (Q^2 = 0.8840) and the LMO (Q2LMO =0.8729) cross-validation tests illustrated the quality of the obtained model. Because all of the validation techniques show the obtained model 1 is a valid model so, it can be used to predict the inhibition activity of the components.

Euclidean based applicability domain (AD)

The principle of Applicability Domain helps the users to specify the scope of their proposed models therefore, defining the model limitations with respect to its structural domain and response space. If an external compound is beyond the defined scope of a given model, it is considered outside that model's Applicability Domain (AD) and cannot be associated with a reliable prediction. The resulting model can be reliably applicable for only those compounds which are inside this domain. Euclidean based applicability domain helps to ensure that the compounds of the test set are representative of the training set compounds used in model development. It is based on distance scores calculated by the Euclidean distance norms. At first, normalized mean distance score for training set compounds are calculated and these values ranges from 0 to 1(0 = least diverse, 1 = most diverse training set compound). Then normalized mean distance score for test set are calculated (Table 3), and those test compounds with score outside 0 to 1 range are said to be outside the applicability domain. This can also be checked by plotting a 'Scatter plot' (normalized mean distance vs. respective activity) including both training and test set as shown in Figure 3. If the test set compounds are inside the domain/area covered by training set compounds that means these compounds are inside the applicability domain otherwise not [29, 30].

Table 3: Calculated normalized mean distance score for training and test set

Training Set				Test Set			
No.	Distance Score	Mean Distance	Normalized Mean Distance	No.	Distance Score	Mean Distance	Normalized Mean Distance
2	985.2709	13.31447	0.185163	1	755.8312	10.21393	0.019847
4	1537.977	20.78348	0.583397	3	960.3047	12.97709	0.167174
5	728.9325	9.850439	0.000466	8	737.4291	9.965258	0.006588
6	751.3568	10.15347	0.016624	32	1005.782	13.59165	0.199942
7	801.2018	10.82705	0.052538	34	1030.543	13.92625	0.217782
9	2116.177	28.59699	1	37	979.0617	13.23056	0.180689
10	1204.272	16.27395	0.342957	39	780.6075	10.54875	0.037699
11	738.457	9.979148	0.007329	49	748.6123	10.11638	0.014646
12	753.1498	10.1777	0.017915	50	808.6955	10.92832	0.057937
13	774.2701	10.46311	0.033133	53	987.4589	13.34404	0.186739
14	754.9798	10.20243	0.019234	55	799.9877	10.81064	0.051663
15	787.4674	10.64145	0.042642	56	916.5486	12.38579	0.135647
16	821.6045	11.10276	0.067238	57	821.2313	11.09772	0.066969
17	741.4916	10.02016	0.009516	58	1019.434	13.77613	0.209778
18	817.8876	11.05253	0.06456	66	931.3225	12.58544	0.146292
19	935.7124	12.64476	0.149455	69	1153.782	15.59164	0.306578
20	989.2855	13.36872	0.188055	71	1405.845	18.99791	0.488194
21	766.2568	10.35482	0.027359	74	921.1002	12.4473	0.138927
22	837.6937	11.32019	0.078831	75	1009.56	13.6427	0.202663
23	1034.127	13.97469	0.220364	77	934.5424	12.62895	0.148612
24	1960.147	26.48847	0.887578	80	910.7247	12.30709	0.131451
25	977.6947	13.21209	0.179704	85	828.1272	11.19091	0.071938
26	1451.917	19.62051	0.52139	87	1274.512	17.22313	0.393565
27	1767.103	23.87977	0.748486	88	2473.945	33.4317	1.257778
28	898.6704	12.14419	0.122766	90	1055.782	14.26733	0.235967
29	836.3896	11.30256	0.077891	93	1159.102	15.66354	0.310411
30	918.1233	12.40707	0.136782	96	876.8551	11.84939	0.107047
31	755.6354	10.21129	0.019706	98	970.8061	13.119	0.174741
33	1481.807	20.02442	0.542926	100	1094	14.78378	0.263504
35	763.6945	10.3202	0.025513	103	1135.921	15.35028	0.293709
36	995.4436	13.45194	0.192492	104	2564.136	34.65048	1.322762
38	1329.607	17.96766	0.433263	106	737.5251	9.966556	0.006658
40	897.6838	12.13086	0.122055				
41	750.544	10.14249	0.016038				
42	1082.96	14.63459	0.255549				
43	752.3112	10.16637	0.017311				
44	764.6559	10.33319	0.026206				

45	741.8753	10.02534	0.009792
46	822.5552	11.11561	0.067923
47	870.0964	11.75806	0.102177
48	766.3808	10.3565	0.027449
51	744.2993	10.0581	0.011538
52	947.2455	12.80061	0.157765
54	980.7949	13.25399	0.181938
59	1679.681	22.69839	0.685497
60	1402.82	18.95702	0.486014
61	1225.021	16.55433	0.357906
62	2033.867	27.48468	0.940694
63	1821.807	24.61902	0.787902
64	759.8092	10.26769	0.022714
65	792.6499	10.71148	0.046376
67	929.8999	12.56621	0.145267
68	912.0834	12.32545	0.13243
70	985.2961	13.31481	0.185181
72	1207.873	16.32261	0.345551
73	784.4533	10.60072	0.04047
76	993.5976	13.42699	0.191162
78	1209.989	16.3512	0.347076
79	1056.677	14.27942	0.236612
81	1111.755	15.02371	0.276296
82	968.0279	13.08146	0.172739
83	999.0982	13.50133	0.195125
84	826.6621	11.17111	0.070882
86	1016.226	13.73279	0.207467
89	820.0085	11.0812	0.066088
91	1063.23	14.36797	0.241333
92	1215.749	16.42904	0.351226
94	998.7944	13.49722	0.194907
95	884.6314	11.95448	0.11265
97	1486.12	20.0827	0.546033
99	1417.359	19.1535	0.496489
101	810.5394	10.95324	0.059266
102	1055.089	14.25796	0.235468
105	728.2851	9.841691	0

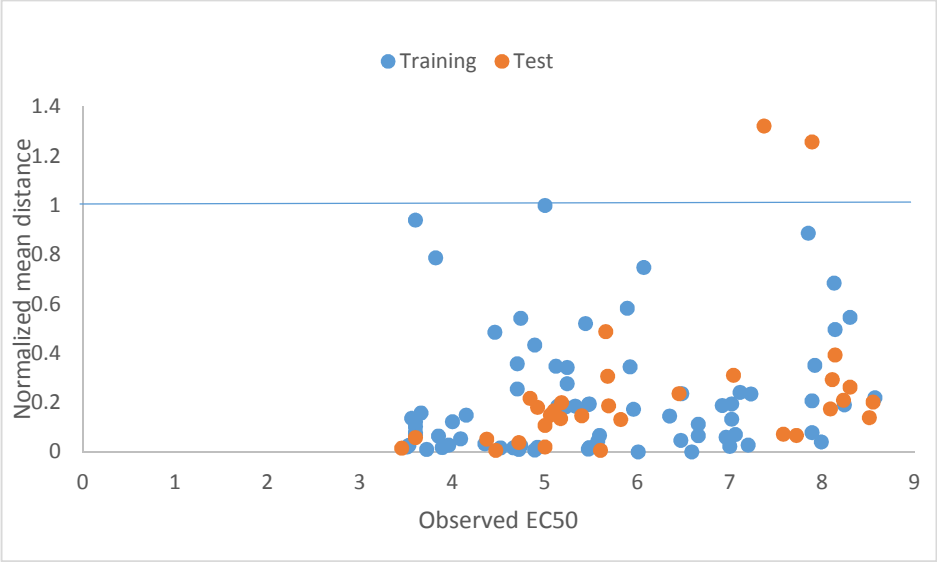

Figure 3: Euclidean Based applicability domain plot, the plot of the normalized mean distance vs. observed EC50

Compound No. 88 and 104 are said to be outside the applicability domain as shown in Figure 3.

The Williams plot, the plot of the standardized residuals versus the leverage, was exploited to visualize the applicability domain (AD) [31]. Leverage indicates a compound's distance from the centroid of *X*. The leverage of a compound in the original variable space is defined as:

$$h_i = X_i^T (X^T X)^{-1} X_i \quad (18)$$

where x_i is the descriptor vector of the considered compound and *X* is the descriptor matrix derived from the training set descriptor values. The warning leverage (h^*) is defined as:

$$h^* = \frac{3(p+1)}{N} \quad (19)$$

Where *N* is the number of training compounds, *p* is the number of predictor variables. From the Williams plot (Fig. 4), it is obvious that all compounds in the test set fall inside the domain of the GA-MLR model (the warning leverage limit is 0.53). There are only two chemicals (No. 9 and No. 26 in the training set and No. 66 and No. 90 in the test set) which have the leverage higher than the warning h^* value, so they can be regarded as structural outliers. Luckily, in this case the data predicted by the model are good for compound numbers 9, 26, 66 and 90, therefore, they are "good leverage" chemicals. For all the compounds in the training and test sets, their standardized residuals are smaller than three standard deviation units (3δ).

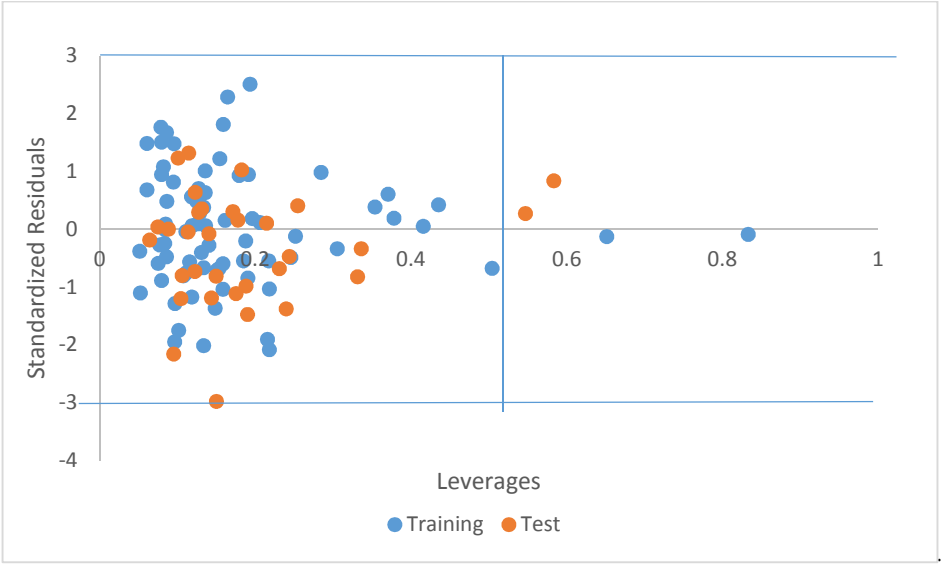

Figure 4: The Williams plot, the plot of the standardized residuals vs. leverages

The GA-MLR model was further validated by applying *Y*-randomization. Several random shuffles of the *Y* vector (pEC_{50}) were performed and the low R^2 and Q^2 values that were obtained showing that the good results in the original model is not due to a chance correlation or structural dependency of the training set. The results of the *Y*-randomization test are presented in Table 2.

Table 4: Y-randomization test for training set

Model	R	R^2	Q^2
Original	0.957799	0.917379	0.88404
Random 1	0.375985	0.141364	-0.20232
Random 2	0.369551	0.136568	-0.40213
Random 3	0.506992	0.257041	-0.16724
Random 4	0.422673	0.178652	-0.1396
Random 5	0.486196	0.236387	-0.07539
Random 6	0.31597	0.099837	-0.37354
Random 7	0.355047	0.126059	-0.27841
Random 8	0.399184	0.159348	-0.20639
Random 9	0.407199	0.165811	-0.59454
Random 10	0.465549	0.216736	-0.19884
Random Models Parameters			
Average R :			0.410435
Average R^2 :			0.17178
Average Q^2 :			-0.26384

The brief description of the selected descriptors by GA-MLR model is summarized in Table 3. The correlation matrix of the twelve selected descriptors is less than 0.621, which means the descriptors are independent in the analysis.

The multi-collinearity between the above twelve descriptors was detected by calculating their variation inflation factors (VIF), which can be calculated as follows:

$$VIF = \frac{1}{1-R^2} \quad (20)$$

Where the R2 is the correlation coefficient of the multiple regression between the variables within the model. If VIF equal to 1, then no inter-correlation exists for each variable; if VIF falls into the range of 1-5, the related model is acceptable; and if VIF is larger than 10, the related model is unstable and a recheck is necessary [32]. The corresponding VIF values of the twelve descriptors are presented in Table 2. As can be seen from the Table, all the descriptors have VIF values of less than five (5), indicating that the obtained model has statistical significance, and the descriptors were found to be reasonably orthogonal.

The mean effect (MF) shown in Table 2 indicates the relative importance of a descriptor, compared with the other descriptors in the model. Its sign indicates the variable direction in the values of the activities as a result of increase or decrease of the descriptor values and the value of the mean effect can be calculated as follows:

$$MF = \frac{\beta_j \sum_{i=1}^{i=n} d_{ij}}{\sum_j^m \beta_j \sum_i^n d_{ij}} \quad (21)$$

Where MF represents the mean effect for the descriptor j, βj is the coefficient of the descriptor j, dij is the value of the interested descriptors for each molecule and m is the number of descriptors in the model [33].

A negative mean effect of this descriptor illustrates that the activity increases with decreasing the value of Ghose-Crippen LogKow (ALogP), ATS autocorrelation descriptor, weighted by scaled atomic mass (ATSm5), Crippen's LogP (CrippenLogP) and Sum of sI E-states (SsI). The Singly bound carbon bound to two other carbons (C2SP3), Valence path cluster, order 4 (VPC-4), Sum of alpha values of all non-hydrogen vertices of a molecule relative to molecular size (ETA_AlphaP), A measure of electronegative atom count (ETA_Epsilon_1), Number of atoms in the largest pi system (nAtomP), Petitjean number (PetitjeanNumber), Directional WHIM, weighted by unit weights (Wlambda2.unity) and Directional WHIM, weighted by atomic polarizabilities (Wlambda1.polar) mean effect has a positive sign. This sign suggest that the anti-HIV activity is directly related to this descriptors.

CONCLUSION

The aim of the present work was developing a QSAR study and predicting the anti-HIV activities of HEPT derivatives. Various theoretical molecular descriptors were calculated by Spartan's14 and PaDEL software and

selected by Genetic function approximation. The built GFA-MLR model was assessed comprehensively, and all the validations indicate that the QSAR model we built is robust and satisfactory. Selection of the twelve descriptors showed that the play a main role in the anti-HIV activity of the compounds.

Acknowledgements
Emmanuel Israel Edache wishes to thank Emmanuel Oduma Edache and Ameh Joshua for its financial support. We also greatly acknowledge David Ebuka Arthur and Adedirin Oluwaseye for sending valuable information for the development of this work.

CONFLICT OF INTEREST
The authors declare that there is no conflict of interests regarding the publication of this paper. Also, they declare that this paper or part of it has not been published elsewhere.

CONTRIBUTION OF THE AUTHORS
This work was carried out in collaboration between all authors. Author EIE designed the study, performed the statistical analysis, wrote the protocol, wrote the first draft of the manuscript and managed literature searches. Authors EIE, AU and SEA managed the analyses of the study and literature searches. All authors read and approved the final version.

REFERENCES

[1] DC Meadows; J Gervay-Hague, *J. Chem. Med. Chem.* **2006**, 1, 16–29.
[2] JM Tronchet; M Seman, *Curr. Top. Med. Chem.* **2003**, 3, 1496–1511.
[3] M Artico; S Massa; A Mai, *Antivir. Chem. Chemother*. **1993**, 4, 361–368.
[4] M De Béthune, *Antivir. Res.* **2010**, 85, 75–90.
[5] C Hansch, T Fujita J. *Am. Chem. Soc.*, **1964**, 86 (8), pp 1616–1626.
[6] S Massa; A Mai; M Artico; G Sbardella; E Tramontano; A Loi; P Scano; P la Colla, *Antivir. Chem. Chemother*. **1995**, 6, 1–8.
[7] A Mai; M Artico; G Sbardella; S Quartarone; S Massa; AG Loi; A de Montis; F Scintu; M Putzolu; P la Colla, *J. Med. Chem*. **1997,** *40*, 1447–1454.

[8] A Mai; M Artico; G Sbardella; S Massa; E Novellino; G Greco; AG Loi; E Tramontano; ME Marongiu; P la Colla, *J. Med. Chem.* **1999**, *42*, 619–627.
[9] S Cheng; D Maier; D Neubueser; DR Hipfner, *Dev. Biol.*, **2010**, 337(1): 99--109.
[10] H Yuan; AL Parrill, *Bioorg. Med. Chem.* **2002**, 10(12), 4169-83.
[11] JM Luco and FH Ferretti, *J. chem. Infor. & comput. Sci.*, **1997**, 37(2), 392-401.
[12] Wavefunction, (**2014**) Inc. Spantan'14, Irvin, California, USA
[13] CW Yap, *J. Comput. Chem*., **2011**, 37(7), 1466-1474.
[14] RG Parr and W Yang, Density functional theory of atoms and molecules. Oxford University Press, Oxford. **1989.**
[15] P Gramatica, *QSAR Comb Sci*., **2007**, 26, 694–701
[16] k Roy; I Mitra; S Kar; PK Ojha; RN Das; H Kabir, *J Chem Inf Model*., **2012**, 52, 396–408.
[17] A Golbraikh; A Tropsha, *J Mol Graph Mod*., **2002**, 20, 269–276
[18] GW Snedecor; WG Cochran, Statistical Methods, Oxford & IBH Publishing Co. Pvt. Ltd., New Delhi, **1967**, 381–418.
[19] L Pogliani, *Amino Acids*., **1994**, 6, 141-153.
[20] L Pogliani, *J. phys. Chem*., **1996,** 100, 18065-18077.
[21] DC Montgomerey; EA Peck; GG Vinig, *Introduction to Linear Regression Analysis*, Wiley, New York**, 2001**.
[22] R Roy; S Kar, How to judge predictive quality of classification and regression based QSAR models? In: Haq Z, Madura JD (eds) Frontiers in computational chemistry. Bentham Science Publishers, Sharjah, **2014**.
[23] G Schuurmann; RU Ebert; J Chen; B Wang; R Kuhne, *J Chem Inf Model*., **2008**, 48, 2140–2145.
[24] PP Roy; K Roy, *QSAR Comb. Sci.* **2008**, *27*, 302-313.
[25] D Rogers; AJ Hopfinger, *J. Chem. Inf. Comput. Sci*., **1994**, *34*, 854-866.
[26] K Roy; S Paul, *QSAR Comb. Sci.*, **2008**, 28, 406-425.
[27] RW Kennard; LA Stone, *Technometrics*, **1969**, 11:1, 137-148.

[28] C Hansch; J Taylor; P Sammes, Comprehensive Medicinal Chemistry: The Rational Design, Mechanistic Study & Therapeutic Application of Chemical Compounds. Pergamon, New York, **1990.**
[29] J Jaworska; N Nikolova-Jeliazkova; T Aldenberg, A review. *Altern. Lab. Anim.* **2005**, *33*, 445–459
[30] A Tropsha; P Gramatica; V Gombar, *QSAR & Comb. Sci.*, **2003**, *22*, 69–7
[31] TI Netzeva; AP Worth; T Aldenberg; R Benigini; MTD Cronin; P Gramatica; JS Joworska; S Kahn; G Klopman; CA Marchant; G Myatt; N Nikolova-Jeliazkova; GY Patlewicz; R Perkins; DW Roberts; TW Schultz; DT Stanton; JJM Van De Sandt; W Tong; G Veith; C Yang, *Altern. Lab. Anim.*, **2005**, 33, 155-173.
[32] M Jaiswal; PV Khadikar; A Scozzafava; CT Supuran, *Bioorg. Med. Chem. Lett*., **2004**, 14, 3283-3290.
[33] S Riahi; E Pourbasheer; MR Ganjali; P Norourzi, *J. Hazard. Mater*., **2009**, 166, 853-859.