INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH & DEVELOPMENT

ISSN 2278 – 0211 (Online)

Multivariate QSAR Study of Indoleβ- Diketo Acid, Diketo Acid and Carboxamide Derivatives as Potent Anti-HIV Agents

Emmanuel Israel Edache
Student, Department of Chemistry, Ahmadu Bello University, Zaria, Nigeria
Adamu Uzairu
Professor, Department of Chemistry, Ahmadu Bello University, Zaria, Nigeria
Stephen Eyije Abechi
Lecturer, Department of Chemistry, Ahmadu Bello University, Zaria, Nigeria

Abstract:
In this study, a set of novel synthesized β- diketo acid, diketo acid and carboxamide derivativesas HIV-1 integrase (HIV-1 IN) was subjected to multivariate QSAR study. Two different variable selection approaches, namely, genetic function approximation (GFA) and multiple linear regression (MLR) used to build the regression models were compared to predict the HIV-1IN inhibition activity. Based on prediction, the best validation model for 5-variable 3' processing inhibition activity with squared correlation coefficient (R^2)= 0.9477, cross validated correlation coefficient (Q^2)= 0.9202 and external prediction ability pred_R^2= 0.8654. Thisshows that nlowhighest atom weighted BCUTS (BCUTw-1h), minimum E-State for (Strong) Hydrogen bond donors (minHBd), Maximum E-State descriptors of strength for potential Hydrogen Bonds of path length (maxHBint3), Fraction of sp3 carbons to sp2 carbons (HybRatio)and Non-directional WHIM, weighted by atomic masses (WD.mass) were the positive contributors, whereas for 6-variables 3' processing inhibition activity, parameters R^2= 0.9588, Q^2= 0.9212 and pred_R^2= 0.7364 showed VPC-4, VPC-5, maxHBd, maxwHBa, maxHBint9 andWD.mass contributed positively to the activity. The binding mode pattern of the compounds to the binding site of integraseenzyme was confirmed by two novel parametersr^{2}m(test) and R^2p. Y-randomization methods confirmed the model robustness. The results of the present study is useful for designing more potent HIV-1IN inhibitors.

Keywords: *QSAR, β- diketo acid, diketo acid and carboxamide derivatives, MLR, PM3, HIV*

1. Introduction

The HIV epidemic is still a major concern. Infection with the human immunodeficiency virus type-1 (HIV-1) causes increasing destruction of immunity, which finally results in the development of the immunodeficiency syndrome (AIDS) [i]. Up to 19 different drugs have been approved for the treatment of HIV-infected individuals, including 7 nucleoside reverse transcriptase (RT) inhibitors (NRTIs), 1 nucleotide RT inhibitors (NtRTI), 3 non-nucleoside RT inhibitors (NNRTIs), 7 protease inhibitors (PIs) and 1 fusion inhibitor [ii]. Virtually every country in the world has seen new infections in 1998, and the epidemic is out of control in many places according to the World Health Organization (WHO) and the Joint United Nations Programme on HIV/AIDS (UNAIDS) [iii,iv]. Human immunodeficiency virus type1 (HIV–1) Integrase is an enzyme required for viral replication. HIV Integrase catalyzes integration of viral DNA into host genome I two separate but chemically similar reactions known as 3'processing and DNA strand transfer. In 3' processing IN removes a dinucleotide next to conserved cytosine–adenine sequence from each 3'– end of the viral DNA. IN then attaches the processed 3'– end of the viral DNA to the host cell DNA in the strand transfer reaction. As there is no known human counterpart of HIV Integrase, IN is an attractive target for anti–retroviral drug design [v].

During the past two decades an increasing number of quantitative structure-activity/property relationship (QSAR/QSPR) models have been studied using theoretical molecular descriptors for predicting biomedical, activity, toxicological and technological properties of chemicals [vi]. QSAR/QSPR includes all statistical methods, by which biological activities are related with structural elements, physicochemical properties or fields [vii].

QSAR studies of anti-HIV activity represent an emerging and exceptionally important topic in the area of computed-aided drug design. The present research aimed to describe the structure-property relationships ofβ- diketo acid, diketo acid and carboxamide derivatives and developed a QSAR model on these compounds with respect to their inhibitory activity (IC_{50}).The results obtained may contribute to further designing novel anti-HIV IN agents.

2. Materials and Methods

2.1. Dataset

A dataset of β −Diketo acid, Diketo acid and Carboxamide derivatives containing 44 compounds with well-defined activity [viii, ix], was selected for QSAR study. The compounds which do not have well defined activity were excluded from dataset. The biological activity data in the form of IC_{50} (molar concentration of the drug leading to 50% inhibition of enzyme Integrase) value were converted into negative logarithmic dose in moles (pIC50) for QSAR Analysis (Table 1).

2.1.1. Table 1: Structures and Biological Activity of Training and Test Set.

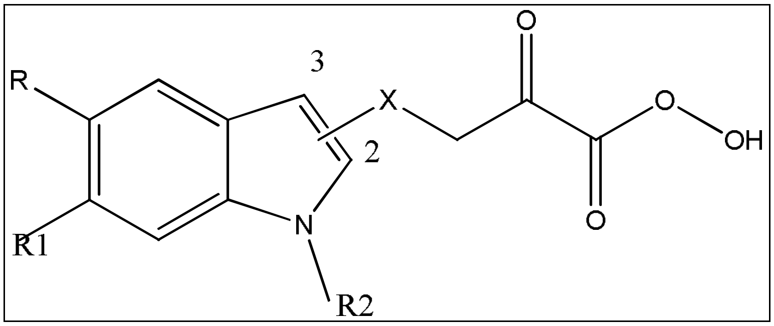

Figure 1

Compd No	R	R1	R2	X	Log IC_{50}
1	H	H	CH_3	2-CO	0.7780
2	OCH_2O		CH_3	2-CO	0.3010
3	H	H	CH_2CH_3	2-CO	0.2040
4	OCH_2O		CH_2CH_3	2-CO	0.6990
5	H	H	Bn	2-CO	0.0000
6	OCH_2O		Bn	2-CO	0.3010
7	H	H	CH_3	3-CO	0.3010
8	OCH_2O		CH_3	3-CO	0.4770
9	H	H	CH_2CH_3	3-CO	0.4770
10	OCH_2O		CH_2CH_3	3-CO	0.4770
11	H	H	Bn	3-CO	0.0000

Table 1a

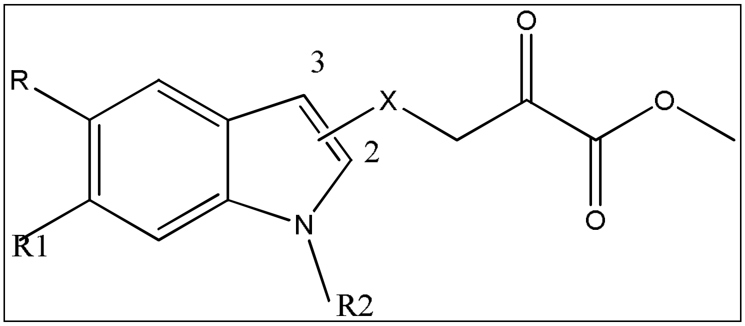

Figure 2

Compd No	R	R1	R2	X	Log IC_{50}
12	H	H	CH_3	2-CO	1.6530
13	OCH_2O		CH_3	2-CO	1.6990
14	OCH_2O		CH_2CH_3	2-CO	1.8130
15	OCH_2O		CH_3	3-CO	1.7780
16	H	H	CH_2CH_3	3-CO	1.4150

Table 1b

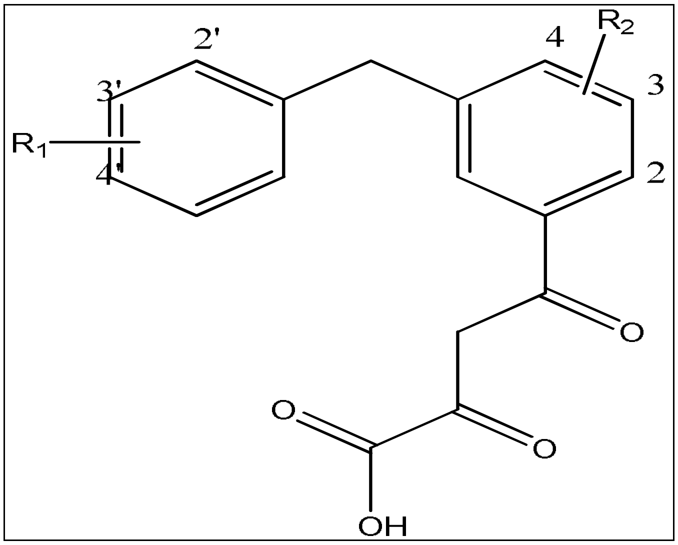

Figure 3

Compd No	R1	R2	R3	IC50
17	4'-Cl	-	-	0.000
18	3'-F	-	-	0.602
19	-	4-OCH_3	-	0.824
20	-	3-OCH_3	-	0.854

Table 1c

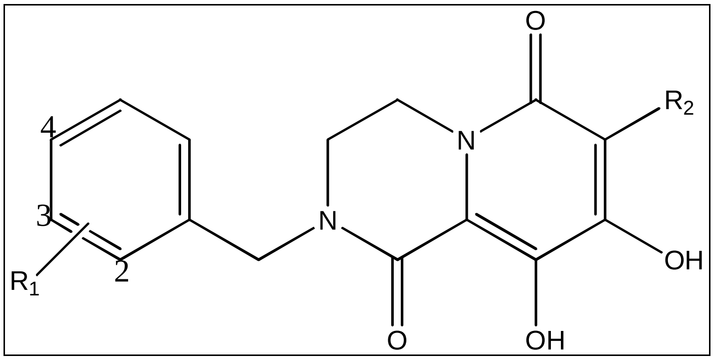

Figure 4

Compd No.	R1	R2	R3	LogIC50
21	4-F	-	-	1.000
22	H	-	-	0.638
23	2-Cl	-	-	0.432
24	3-Cl	-	-	1.398
25	4-Cl	-	-	0.420
26	4-F, 3-Cl	-	-	1.398
27	4-F	CN	-	1.699
28	4-F	Br	-	1.523
29	4-F	I	-	1.699

Table 1d

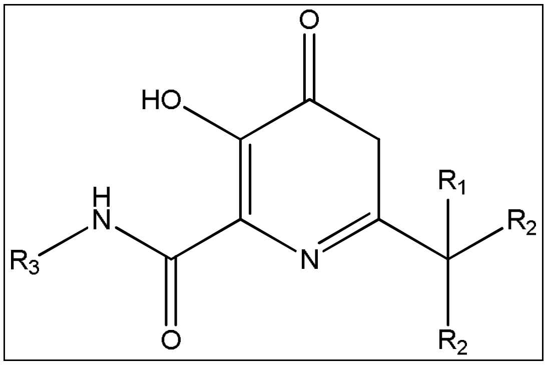

Figure 5

Compd No.	R1	R2	R3	LogIC50
30	NHCOCH$_3$	CH$_3$	4-fluorotoluene	2.1555
31	NH-SO$_2$-CH$_3$	CH$_3$	4-fluorotoluene	2.097
32	NHCO-N(CH3)$_2$	CH$_3$	4-fluorotoluene	1.745
33	NHSO2-N(CH3)$_2$	CH$_3$	4-fluorotoluene	1.921
34	NHCOCO-N(CH3)$_2$	CH$_3$	4-fluorotoluene	2.000
35	NHCOCO-OCH$_3$	CH$_3$	4-fluorotoluene	1.824
36	NHCOCO-OH	CH$_3$	4-fluorotoluene	2.398
37	N(CH3)COCO-N(CH3)$_2$	CH$_3$	4-fluorotoluene	1.824
38	NHCO-pyridine	CH$_3$	4-fluorotoluene	1.699
39	NHCO-pyridazine	CH$_3$	4-fluorotoluene	1.824
40	NHCO-pyrimidine	CH$_3$	4-fluorotoluene	2.155
41	NHCO-oxazole	CH$_3$	4-fluorotoluene	2.155
42	NHCO-thiazole	CH$_3$	4-fluorotoluene	2.097
43	NHCO-1H imidazole	CH$_3$	4-fluorotoluene	2.222
44	NHCO-1,3,4-oxadiazole	CH$_3$	4-fluorotoluene	1.824

Table 1e

2.2. Molecular Modeling and Generation of Molecular Descriptors

The molecular modeling study was performed using MSOffice 2007 software. Structure ofall the compounds were drawn using ChemDraw Ultra [x]version 12.0.2 moduleof the program and transferred to Spartan'14 [xi] version 1.1.2 to create the three-dimensional (3D) structure. These structures were then subjected to energy minimization. Energy minimized molecules were subjected to optimization via paraterization method (PM3) and also transferred to PaDEL-Descriptor [xii] version 2.18 and were subjected to re-optimization MM2 force field. Most stable structure for each compound was generated and used for calculating various physicochemical parameters like thermodynamic, steric and electronic descriptors (Table S1 in Supplementary material).

2.3. Variable Selection and Model Generation

Though many molecular descriptors are available, only a subset of them is statistically significant in terms of correlation with biological activity. Therefore, it is very important to address the variable selection method for originating the optimal QSAR model. GFA [xiii] and MLR approaches were adopted to select the best possible variables as well as for the generation of QSAR models.

2.3.1. GFA Method

Genetic function approximation(GFA) algorithm are governed by biological evolution rules [xiv,xv]. GFA, which is based on the principles of Darwinian evolution [xvi],is a search method to find exact or approximate solutions to optimization and search problems. GFA is conceived from
(1) Genetic algorithm and
(2) Friedman's multivariate adaptive regression splines (MARS) algorithm.
The following steps were performed:
(1) Initial population of equations were generated by a random number of descriptors,
(2) Pairs from the population of equations were chosen at random, crossovers were performed and offspring equations were generated,
(3) The fitness of each progeny equation was assessed by lack of fit (LOF) score that automatically penalizes models with too many features. A peculiar feature of GFA is that it generates a population of equations rather than a single equation as do most other statistical methods. The range of variations in this population gives added information on the quality of fit and importance of the descriptors [xvii]. The fitness function, i.e., lack-of-fit used here was the leave one-out cross validated correlation coefficient (Q_{LOO}^2) and is calculated by

$$LOF = \frac{LSE}{\left\{1 - \left[\frac{c+d*p}{m}\right]\right\}^{2}}$$

Where c is the number of basic functions, d is the smoothing parameter, M is the number of samples in the training set, LSE is the least square error and p is the total number of features contained in all basis functions. Selected descriptors are given in supplementary material (Table S2). GFA technique was used for generating QSAR models for both classes with 5000 crossovers and the smoothness value (d) of 1.0 was used during the equation generation

2.3.2. MLR

Multiple linear regression analysis of molecular descriptors was carried out using the Microsoft Excel for Windows. Multiple linear regression (MLR) is a method used to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to the observed and was employed to correlate the binding affinity and molecular descriptors [xviii]. This method has been widely applied in many QSAR studies, and has upheld to be a useful linear regression method to build QSAR models that may explore forthright the properties of the chemical structure in combination with its ability of inducing a pharmacological response [xix]. The advantage of MLR is its simple method and easily interpretable mathematical expression.The multi-collinearity among variables was identified using variance inflation factor (VIF) [xx]. The VIF for the regression coefficient is expressed as follws:

$$VIF = \frac{1}{1 - R_i^2}$$

Where R^2 is the correlation coefficient of the multiple regression between the variables within the model. If VIF equals 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 [xxi-xiii]

2.4. Validation of QSAR Models

The QSAR models were developed by GFA and MLRmethods and evaluated using the following statistical parameters:In the MLR equations, the figures in the parentheses are the standard errors of the regression coefficients, N is the total number of compounds in the data set, $N_{training}$ is the number of compound in the training set, $N_{test}$is the number of compound in the test set, *R* is the correlation coefficient, R^2 is the determination coefficient, Q^2 is the leave many out(LOO) cross validated,The cross-validated Q^2 in each case was found to be very close to the value of R^2 for the entire data set and hence these models can be labelled as statistically significant. Cross validation provides the values of PRESS, SSY and Q^2cv and *RMSEP* from which we can test the predictive power of the proposed model. It is argued that PRESS (the predictive residual sum of the squares), is a good estimate of the real predictive error of the model and if it is smaller than SSY the model predicts better than chance and can be considered statistically significant. F is the significance test (F-test). The F-test reflects the ratio of the variance explained by the model and the variance due to the error in the regression. High values of the F-test indicate that the model is statistically significant, *RMSECV* is the root mean square error of cross validation(training set), and *RMSEP* is the root mean square error of prediction(external validation set) and is more directly related to the uncertainty of the predictions. The *RMSEP* values also support our results. Se is the standard error of estimate represents standard deviation which is measured by the error mean square, which expresses the variation of the residuals or the variation about the regression line. Therefore, standard deviation is an absolute measure of quality of fit and should have low value for the regression to be significant. R^2_{pred} is the correlation coefficient of multiple determination (external validation set). F-test values are for all equation statistically significant at 95% level probability.

R^2, Q^2, *RMSECV*, *Q*, and *RMSEP* of a model can be obtained from:

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

R^2 is a measure of explained variance. Each additional X variable added to a model increases R^2. R^2 is a relative measure of fit by the regression equation. Correspondingly, it represents the part of the variation in the observed data that is explained by the regression. Calculation of Q^2 (cross-validated R^2) confirm the validity of the modelsis called an internal validation.

$$Q^2 = 1 - \frac{\sum\left(Y_{obs} - Y_{pred}\right)^2}{\sum(Y_{obs} - \bar{Y})^2}$$

$$RMSECV = \sqrt{\sum \frac{\left(Y_{obs} - Y_{pred}\right)^2}{N}}$$

Where, Y_{obs} ,Y_{pred} and *N* indicate observed, predicted activity values and number of samples in the training set respectively and $\bar{Y}$ indicates mean activity value. A model is considered acceptable when the value of Q^2 exceeds 0.5.

External validation or predictability of the models are performed by calculating predictive $R^2(R^2_{pred})$.$R^2_{pred}$was calculated for evaluating the prediction ability of the models.

$$R^2_{pred} = 1 - \frac{\sum\left(Y_{pred(Test)} - Y_{Test}\right)^2}{\sum\left(Y_{(Test)} - \bar{Y}_{training}\right)^2}$$

$$RMSEP = \sqrt{\sum \frac{\left(Y_{pred(Test)} - Y_{Test}\right)^2}{M}}$$

Where,$Y_{pred(Test)}$, $Y_{(test)}$ and M indicate predicted, observed activity values and number of samples respectively of the test set compounds and $\bar{Y}_{training}$ indicates mean of observed activity values of the training set. For a predictive QSAR model, the value of R^2_{pred} should be more than 0.5 [vii, xxiv,xxv].

However, this is not a sufficient condition to guarantee that the model is really predictive. It is also recommended to check: 1) the slope K or K' of the linear regression lines between the observed activity and the predicted activity in the external validation, where the slopes should be 0.85≤K≤1.15 or 0.85≤K'≤1.15 and 2) the absolute values of the difference between the coefficients of multiple determination, R^2_o and R'^2_o smaller than 0.3 [xxvi].

Q is the quality factor [xxvii,xxviii]. The quality factor Q is used to decide the predictive potential of the models. The quality factor Q is defined as the ratio of correlation coefficient to the standard error of estimation. We found it to be a good parameter to explain the predictive potential of the models proposed by us. The higher the value of Q the better is the predictive potential of the models [xxvii-xxix].

$$Q = \frac{R}{SE}$$

R^2_a takes into account the adjustment of R^2. $R^2_a$is a measure of the percentage explained variation in the dependent variable that takes into account the relationship between the number of cases and the number of independent variables in the regression model, whereas R^2 will always increase whenan independent variable is added. R^2_awill decrease if the addedvariable does not reduce the unexplained variable enough tooffset the loss of decrease of freedom.For reliability of the model, probable errorof correlation (PE) was also calculated. If the value of correlation coefficient (R) is more than six times of PE then the expression is good and reliable [xxx].

$$P.E = 0.6745(S.E)$$

Where SE is the standard error of estimate, and be calculate as follows:

$$S.E = \frac{1 - R^2}{\sqrt{N}}$$

Where R is the coefficient of correlation and N is the number of training set.

2.5. Training and Test Set Selection

The main target of any QSAR modeling is that the built model should be robust enough to be capable of making accurate and reliable predictions of biological activities of new compounds [xxxi]. So, QSAR models derived from a training set should be validated using new chemical entities for checking the predictive capacity of the constructed models. The validation strategies check the reliability of the models for their possible application on a new data set, and so confidence in the prediction can be judge [xxxii,xxxiii]. As a result for the division of the data set into training and test sets, the compounds were ranked according to the IC50 values and every alternate compound was assigned to the test set. 70% compounds were selected for the training set and 30% for the test set. In our present work, the total data set consisted of 44 compound.

3. Results and Discussion

3.1. QSAR Study

The model generated for 3' processing inhibition activity by GFA algorithm was Model 1.

5-variable

3.1.1. Model 1

$$pIC_{50} = -0.0019(ECCEN) - 1.4578(minHsOH) + 0.1282(maxHBint9) + 4.1615(maxHaaCH) - 0.7802(ELUMO) - 0.6178$$

$$N_{total} = 44, N_{training} = 30, N_{test} = 10, outlier = 4, R = 0.9718, R^2 = 0.9444, R_a = 0.9328, Q^2_{cv} = 0.9110, SE = 0.1985, F = 81.5595, LOF = 0.1765, SSY = 0.9454, PRESS = 5.4177, Q = 4.8957, RMSECV = 0.1775, RMSEP = 0.7361, R^2_{pred} = 0.1003$$

3.1.2. Model 2

$$pIC_{50} = -0.0627(BCUTw-1h) - 6.9460(minHBd) + 0.4484(maxHBint3) - 7.0400(ETA\_EtaP\_F\_L) + 2.4092(WD.mass) + 2.8472$$

$$N_{total} = 44, N_{training} = 30, N_{test} = 10, outlier = 4, R = 0.9718, R^2 = 0.9444, R_a = 0.9328, Q^2_{cv} = 0.9162, SE = 0.1985, F = 81.5730, LOF = 0.1768, SSY = 0.9453, PRESS = 0.9212, Q = 4.8957, RMSECV = 0.1775, RMSEP = 0.3035, R^2_{pred} = 0.8470$$

3.1.3. Model 3

$$pIC_{50} = -1.5272(minHsOH) - 0.3178(minsssN) - 2.2221(maxwHBa) + 6.8629(petitjeanNumber) + 2.9963(WD.polar) - 0.3218$$

$$N_{total} = 44, N_{training} = 30, N_{test} = 10, outlier = 4, R = 0.9720, R^2 = 0.9448, R_a = 0.9333, Q^2_{cv} = 0.9002, SE = 0.1978, F = 82.1555, LOF = 0.1753, SSY = 0.9389, PRESS = 2.0407, Q = 4.9141, RMSECV = 0.1769, RMSEP = 0.4517, R^2_{pred} = 0.6610$$

3.1.4. Model 4

$$pIC_{50} = -1.6057(minHsOH) - 0.3088(minsssN) - 2.1030(maxwHBa) + 7.2083(petitjeanNumber) + 3.2553(WD.volume) - 0.7588$$

$$N_{total} = 44, N_{training} = 30, N_{test} = 10, outlier = 4, R = 0.9722, R^2 = 0.9452, R_a = 0.9338, Q^2_{cv} = 0.8984, SE = 0.1971, F = 82.7590, LOF = 0.1741, SSY = 0.9325, PRESS = 1.9814, Q = 4.9325, RMSECV = 0.1763, RMSEP = 0.4451, R^2_{pred} = 0.6710$$

3.1.5. Model 5

$$pIC_{50} = -0.0584(BCUTw-1h) - 7.0812(minHBd) + 0.4684(maxHBint3) + 1.5843(HybRatio) + 2.5225(WD.mass) + 0.3460$$

$$N_{total} = 44, N_{training} = 30, N_{test} = 10, outlier = 4, R = 0.9735, R^2 = 0.9477, R_a = 0.9369, Q^2_{cv} = 0.9202, SE = 0.1924, F = 87.0528, LOF = 0.1660, SSY = 0.8889, PRESS = 0.8103, Q = 5.0598, RMSECV = 0.1721, RMSEP = 0.2847, R^2_{pred} = 0.8654$$

Validation was performed by dividing dataset into trainingset and test set. The best model generated for 3'processing inhibition activity using GFA method was Model 6

3.1.6. Model 6

6-variable

$$pIC_{50} = -1.3314(VPC-4) + 1.0971(VPC-5) - 1.6374(maxHBd) - 1.4925(maxwHBa) + 0.1492(maxHBint9) + 0.8821(WD.mass) + 3.5344$$

$$N_{total} = 44, N_{training} = 30, N_{test} = 10, outlier = 4, R = 0.9792, R^2 = 0.9588, R_a = 0.9481, Q^2_{cv} = 0.9212, SE = 0.1745, F = 89.2670, LOF = 0.1416, SSY = 0.7003, PRESS = 1.5872, Q = 5.6115, RMSECV = 0.1528, RMSEP = 0.3984, R^2_{pred} = 0.7364$$

MLR analysis resulted in several significant modelswith respect to inhibition of 3'processing and integration activity, respectively. Model 3 was selected for 3'processing inhibition activity.

5-variables

3.1.7. Model 7

$$pIC_{50} = -20.8484(\pm4.1229) - 0.2040(\pm0.0435)LogP + 0.0045(\pm0.0024)ZPE. -0.0325(\pm0.0094)Area + 16.9125(\pm3.850)Ovality + 0.0792(\pm0.0177)\mathrm{minLocIonPot}$$

$$N_{total} = 44, N_{training} = 30, N_{test} = 10, outlier = 4, R = 0.9065, R^2 = 0.8217, R_a = 0.7846, Q^2_{cv} = 0.7067, SE = 0.3555, F = 22.1243, LOF = 0.5663, SSY = 3.0324, PRESS = 4.9870, Q = 2.5499, RMSECV = 0.3179, RMSEP = 0.4636, R^2_{pred} = 0.6431$$

6-variable

3.1.8. Model 8

$$pIC_{50} = -25.0603(\pm4.3489) - 0.2634(\pm0.0530)LogP + 0.0121(\pm0.0047)Acc.P - Area(75) - 0.0213(\pm0.0056)Area + 0.0710(\pm0.0189)MinLocIonPot - 0.0227(\pm0.0086)PSA + 20.3641(\pm4.2267)Ovality$$

$N = 27, N_{training} = 27, N_{test} = 10, R = 0.9229, R^2 = 0.8518, R_{adj} = 0.8131, SE = 0.3311, F = 22.0327, PRESS =$ $4.7394, RMSECV = 0.2898, RMSEP = 0.3121, Q = 2.7874, R^2_{pred} = 0.8383,$

Based on the statistical significance and validation parameters, a comparison was done between the validation models “Model 5 and Model 7 for 3' anti-HIV-1 IN activity” generatedby GFA and MLR methods (Table 2).Model 7 showed lower Q^2 and R^{2}pred values than Model 5which means that prediction ability of Model 5 was much better. Statistical analysis was performed to access the robustness and statistical confidence. Higher value of Q and lower value of RMSECV, Y-randomization test and RMSEP of Model 5 in comparison to Model 7 revealed that Model 5 was robust and promising.In the developed Model the value of coefficient of correlation was significantly higher than the value of PE (0.1298) supporting reliability and goodness. Based on the above results Model5 was considered as the best validation model for 3’ processing inhibition activity. The accuracy of the Model 5 was ascertained by correlation coefficient (R= 0.9735), statistical significance more than 99% “against tabulated value F= 87.0528and low standard error of estimate = 0.1924”. The model shows that BCUT descriptor (BCUTw-1h), Electrotopological state atom type descriptor (minHBd and maxHBint3), Hydridization ratio descriptor (HybRatio) and WHIMdescriptor (WD.mass) showed positive contribution. The correlation matrix between the physic-chemical parameters and the biological activity is presented in Table S1 (supplementary material). Here the negative values BCUTw-1h and minHBd indicates that the decrease in BCUT and Electrotopological state atom type descriptors will favor the exhibition of the anti-HIV activity. The brief description of the descriptors is given in Table S2 (supplementary material). The robustness of the model was justified by the magnitude of a modified r^2 (r$^2_{m(test)}$ = 0.8432), and the novel parameter R^2_p = 0.8700, which was near to the conventional R^2(0.9477). The internal validation parameter of the model (Q^{2}cv = 0.9202) was also good. The scatterplot of observed activity versus predicted activity is shown in Figure. 6a and 6b.

	Model 5	**Model 6**	**Model 7**	**Model 8**
R^2	0.9477	0.9588	0.8217	0.8518
Q^2_{cv}	0.9202	0.9212	0.7068	0.7214
R^2_{pred}	0.8654	0.7364	0.6431	0.8383
PE	0.1298	0.1177	0.2398	0.2233
Q	5.0598	5.6115	2.5499	2.7874
$RMSECV$	0.1721	0.1528		0.2898
$RMSEP$	0.2847	0.3984	0.4636	0.3121
LOF	0.1660	0.1416	-	-
F	87.0528	89.2670	22.1243	22.0327
K	0.9347	0.9154	0.9538	0.9373
K'	1.0347	1.0300	0.9581	1.0288
r^2	0.8837	0.8354	0.6346	0.8466
$/r_o^2 - r_o'^2/$	0.023	0.0698	0.0871	0.0205
$\frac{r^2 - r_0^2}{r^2}$	0.0024	0.0134	0.0078	0.00001
$\frac{r^2 - r_o'^2}{r^2}$	0.0284	0.0970	0.1451	0.0243
$r^2_{m(test)}$	0.8432	0.7470	0.5899	0.8445
R^2_p	0.8700	0.8626	0.7272	0.7430
SE	0.1924	0.1745	0.3555	0.3311
R_{yrand}	0.3862	0.4276	0.4221	0.4513
R^2_{yrand}	0.1613	0.1950	0.1957	0.2163
Q^2_{yrand}	-0.3382	-0.4171	-0.2867	-0.3592

Table 2: Predicted values of training (internal cross-validation) and test set (external cross-validation) and results of statistical parameters.

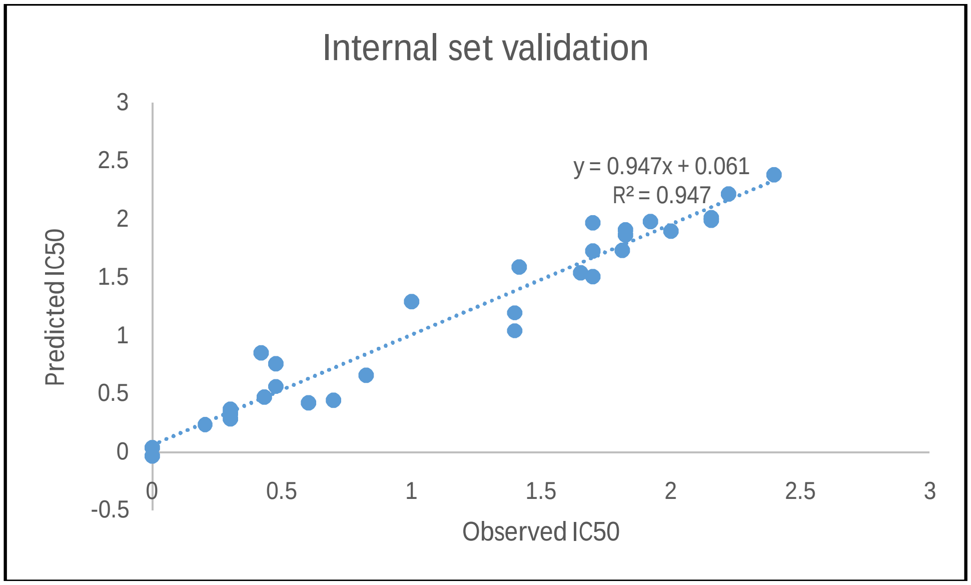

Figure 6a: Scatter plot between the observed and predicted IC50 of training set

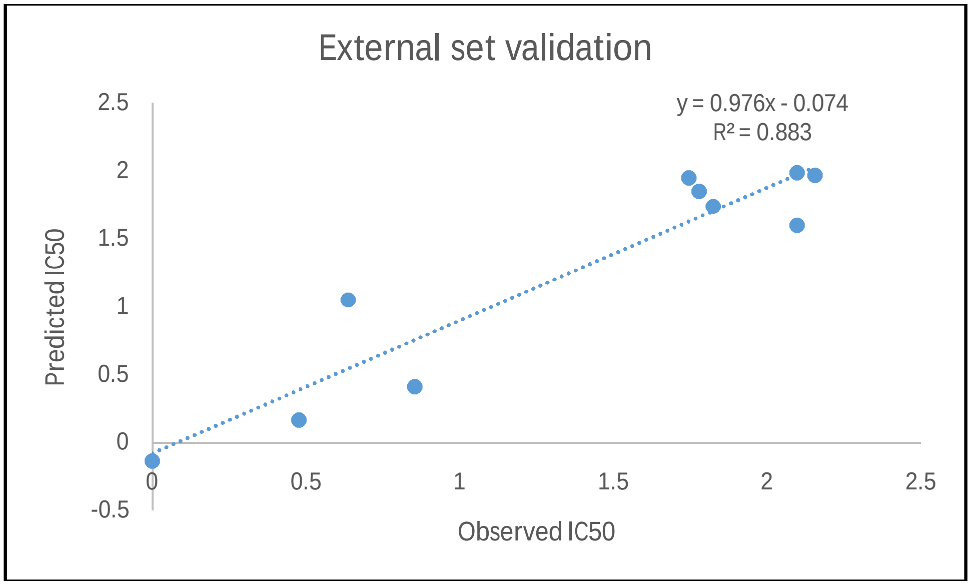

Figure 6b: Scatter plot between the observed and predicted IC50 of test set

	Training Set				
		Model 5	**Model 6**	**Model 7**	**Model 8**
No.	LogIC50	*Predicted LogIC50*	Predicted LogIC50	Predicted LogIC50	*Predicted LogIC50*
2	0.301	0.329803	0.409022	0.473759	0.329189
3	0.204	0.236325	0.275987	0.447183	0.226848
4	0.699	0.44512	0.505158	0.697852	0.524859
5	0	-0.03252	0.093622	0.664315	0.427789
6	0.301	0.28511	0.238816	0.021393	-0.05825
7	0.301	0.369664	0.280338	0.210907	0.446324
8	0.477	0.760988	0.553206	0.52708	0.615096
9	0.477	0.564935	0.324286	0.455747	0.602123
12	1.653	1.544044	1.481148	1.228877	1.141114
13	1.699	1.73033	1.709368	1.419593	1.445649
14	1.813	1.734899	1.771609	1.688969	1.695976
16	1.415	1.589804	1.484529	1.23088	1.297639
17	0	0.040174	0.038348	0.44302	0.69355
18	0.602	0.426671	0.449258	0.164721	0.358612
19	0.824	0.663393	0.76562	0.739912	0.456137
21	1	1.295131	1.339194	1.255071	1.154853
23	0.432	0.47352	0.643179	1.060842	0.818633
24	1.398	1.043577	1.00065	1.044196	0.94288
25	0.42	0.852245	0.732477	1.046013	0.95126
26	1.398	1.198014	1.399681	1.039	1.349106
27	1.699	1.509179	1.539825	1.276858	1.674029
30	2.155	2.015955	2.329163	2.39157	2.299321
33	1.921	1.98423	1.821612	2.25202	2.222018
34	2	1.900735	2.001545	2.268632	2.123117
36	2.398	2.384736	2.253868	2.280222	2.029935
37	1.824	1.869168	1.737919	1.64589	1.574366
38	1.699	1.971709	1.784684	1.774617	1.948498
39	1.824	1.911861	1.970123	1.617896	1.782821
40	2.155	1.994816	2.184289	1.914539	2.203716
43	2.222	2.217382	2.192477	2.029428	2.033793
	Test Set				
10	0.477	0.1663	0.329384	1.2408	1.106169
11	0	-0.135	0.026372	-0.01725	0.164073
15	1.778	1.8526	1.747529	1.261405	1.535098
20	0.854	0.4142	0.086399	0.694975	0.694061
22	0.638	1.0517	0.73405	1.107069	0.570962
31	2.097	1.9881	2.808833	2.085442	2.399209
32	1.745	1.9511	1.800076	2.50273	2.29985
41	2.155	1.9702	2.247528	1.728638	1.949455
42	2.097	1.6002	1.573561	1.565314	1.96431
44	1.824	1.74	2.239993	1.757981	1.940356
	Outliers				
1	0.778				
28	1.523				
29	1.699				
35	1.834				

Table 3: Observed and predicted activity of model 5, 6, 7 and 8.

From Table 2 it is evident that Model 6 showed better valuefor Q^2cv (0.9212) than Model 8 (0.8042) but a high value for RMSEP. A high value of Q^2cv alone is an insufficient criterion for a QSAR model to be highly predictive [xxxiv,xxxv]. Based on prediction ability, Model 6 was selected as the best validation model for 3' processing inhibition activity. Model 6 shows a good correlation between descriptors (VPC-4, VPC-5, maxHBd, maxwHBa, maxHBint9 and WD.mass) and integration inhibition activity. The correlation matrix between the physicochemical parameters and the biological activity is given in Table S3 (supplementary material). Correlation coefficient (R= 0.9792), squared correlation coefficient (R^2= 0.9588), Low standard error of estimate(0.1745) of the model and a statistical significance more than 99% (F value = 89.2670) demonstrate the accuracy of the model. Positive contribution of VPC-5, maxHBint9 and WD.mass indicated favorable interactions that were responsible for the enhancement of HIV-1IN inhibition activity. The scatter plot between calculated and predicted activities of the training set and test set compounds is given in Figure.7a and 7b. To confirm the robustness of the derived best validation models, a y-randomization test was performed by scrambling the experimental activity at 100 random numbers of trial considering the same number and definition of descriptors. The results so obtained show that original model was not obtained due to a chance correlation.Low value for LOF for all the models suggested that selected models for both activities were robust. The predicted biological activities of training and test set molecules are given in Table 3. From Table 3, it is evident that the predicted activities of all compounds in the training set and test set are in good agreement with their corresponding experimental activities. The robustness of the model was justifiedby the magnitude of a modified r^2 ($r^2_{m(test)}$ = 0.7470), and the novel parameter R^2_p = 0.8626, which wasnear to the conventional R^2(0.9477).

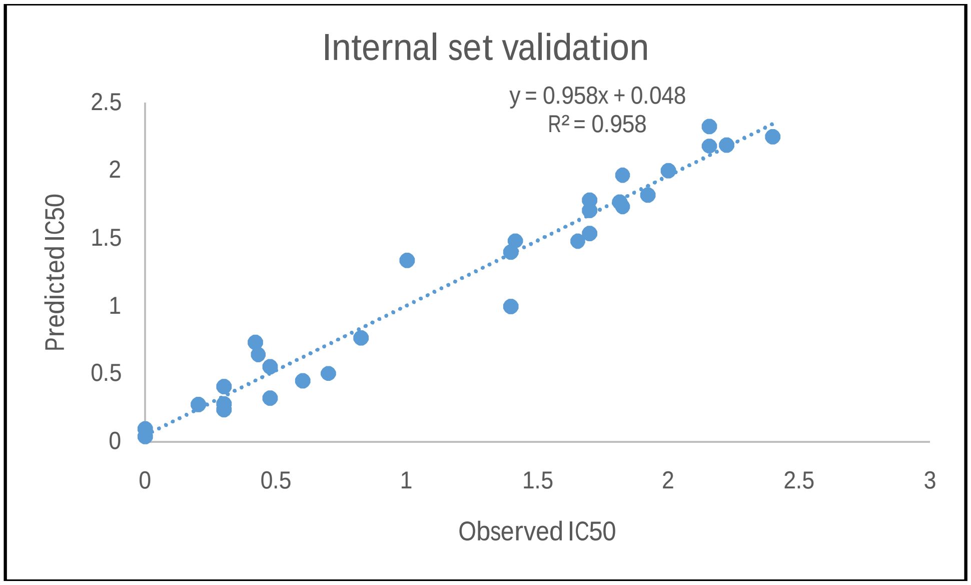

Figure 7a: Scatter plot between the observed and predicted IC50 of training set

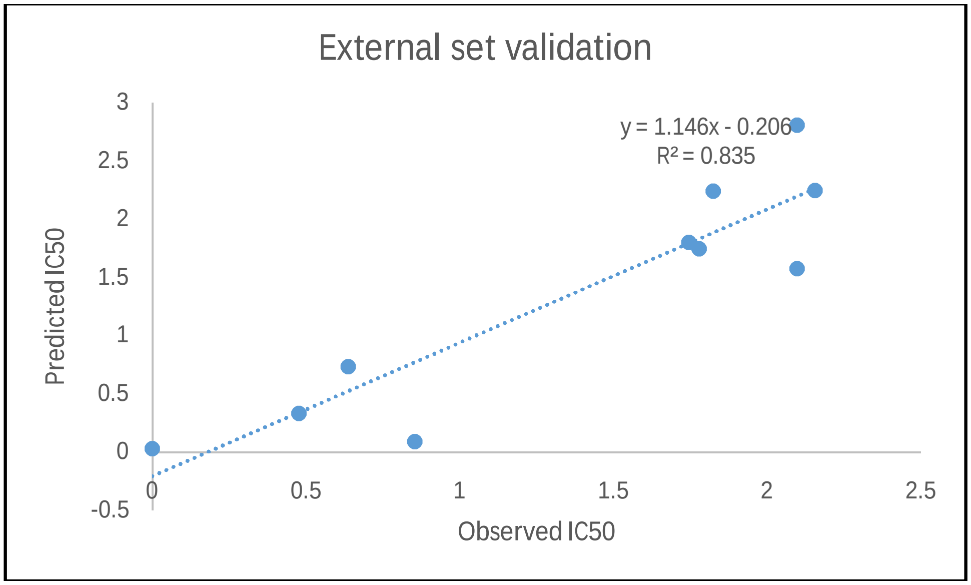

Figure 7b: Scatter plot between the observed the predicted IC50 of training set

3.3. Applicability Domain

The use of (Q)SAR models for chemical risk management and regulatory purposes have increased steadily (in the EU: Registration, Evaluation, Authorization and Restriction of Chemicals).It is of crucial importance to be able to judge the reliability of predictions. The chemical descriptor space covered by a particular training set of chemicals is called Applicability Domain. It offers the opportunity to assess whether a compound can be reliably predicted [xxxvi]. Applicability domain (AD) is the physicochemical, structural or biological space, knowledge orinformation on which the training set of the model has been developed. The resulting model can be reliably applicable for only those compounds which are inside this domain [vii,xxxvii]. AD helps to ensure that the compounds of the test/external set are representative of the training set compounds used in model development [xxxviii]. 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, and those test compounds with score outside 0 to 1 (Table 4 and 5) 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/property) including both training and test set. 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[xxxv].

Training Set:	**Model 5**			**Training Set:**	**Model 6**		
Compound No.	**Distance Score**	**Mean Distance**	**Normalized Mean Distance**	**Compound No.**	**Distance Score**	**Mean Distance**	**Normalized Mean Distance**
2	214.019	7.134	0.144	2	88.642	2.955	0.015
3	213.638	7.121	0.143	3	90.482	3.016	0.03
4	213.381	7.113	0.142	4	86.734	2.891	0
5	215.128	7.171	0.148	5	92.932	3.098	0.05
6	212.515	7.084	0.139	6	90.035	3.001	0.027
7	213.566	7.119	0.142	7	89.917	2.997	0.026
8	213.809	7.127	0.143	8	87.306	2.91	0.005
9	212.571	7.086	0.139	9	88.201	2.94	0.012
12	288.653	9.622	0.417	12	95.563	3.185	0.071
13	288.641	9.621	0.417	13	94.779	3.159	0.065
14	289.051	9.635	0.418	14	96.436	3.215	0.078
16	288.9	9.63	0.417	16	93.586	3.12	0.055
17	448.44	14.948	1	17	97.001	3.233	0.083
18	186.485	6.216	0.044	18	92.843	3.095	0.049
19	202.024	6.734	0.1	19	115.408	3.847	0.23
21	180.557	6.019	0.022	21	87.004	2.9	0.002
23	439.703	14.657	0.968	23	87.151	2.905	0.003
24	439.2	14.64	0.966	24	87.85	2.928	0.009
25	439.701	14.657	0.968	25	88.804	2.96	0.017
26	439.214	14.64	0.966	26	89.498	2.983	0.022
27	180.723	6.024	0.022	27	93.063	3.102	0.051
30	176.105	5.87	0.006	30	157.003	5.233	0.565
33	377.515	12.584	0.741	33	155.455	5.182	0.552
34	175.195	5.84	0.002	34	153.068	5.102	0.533
36	188.891	6.296	0.052	36	211.137	7.038	1
37	177.947	5.932	0.012	37	157.089	5.236	0.566
38	174.729	5.824	0.001	38	150.44	5.015	0.512
39	174.563	5.819	0	39	151.171	5.039	0.518
40	176.325	5.877	0.006	40	152.118	5.071	0.526
43	176.761	5.892	0.008	43	153.296	5.11	0.535
Test Set:				Test Set:			
Compound No.	**Distance Score**	**Mean Distance**	**Normalized Mean Distance**	**Compound No.**	**Distance Score**	**Mean Distance**	**Normalized Mean Distance**
10	213.492	7.116	0.142	10	87.764	2.925	0.008
11	215.237	7.175	0.149	11	93.072	3.102	0.051
15	288.947	9.632	0.418	15	95.555	3.185	0.071
20	202.181	6.739	0.101	20	91.837	3.061	0.041
22	210.486	7.016	0.131	22	86.639	2.888	-0.001
31	378.154	12.605	0.743	31	157.593	5.253	0.57
32	175.036	5.835	0.002	32	153.135	5.105	0.534
41	176.186	5.873	0.006	41	154.285	5.143	0.543
42	378.208	12.607	0.744	42	153.282	5.109	0.535
44	174.46	5.815	0	44	154.82	5.161	0.547

Table 4: GFA Applicability domain results for model 5 and 6

Training Set:	**Model 7**			**Training Set:**	**Model 8**		
Compound No.	Distance Score	Mean Distance	Normalized Mean Distance	Compound No.	Distance Score	Mean Distance	Normalized Mean Distance
2	5363.47	178.782	0.18	2	2131.22	71.041	0.127
3	4602.01	153.4	0.081	3	2162.29	72.076	0.144
4	4004.28	133.476	0.004	4	1974.51	65.817	0.041
5	4402.09	146.736	0.055	5	2070.5	69.017	0.094
6	4868.63	162.288	0.116	6	2320.34	77.345	0.232
7	6594.26	219.809	0.34	7	2637.86	87.929	0.407
8	5323.29	177.443	0.175	8	2293.25	76.442	0.217
9	4605.79	153.526	0.082	9	2196.49	73.216	0.163
12	4593.52	153.117	0.08	12	2647.1	88.237	0.412
13	4076.69	135.89	0.013	13	2004.42	66.814	0.057
14	4133.41	137.78	0.021	14	2024.34	67.478	0.068
16	4018.95	133.965	0.006	16	2177.94	72.598	0.153
17	4102.64	136.755	0.017	17	1923.78	64.126	0.013
18	3976.11	132.537	0	18	1900.5	63.35	0
19	4380	146	0.053	19	2047.86	68.262	0.081
21	4076.7	135.89	0.013	21	2120.78	70.693	0.122
23	4046.23	134.874	0.009	23	2185.66	72.855	0.157
24	4021.34	134.045	0.006	24	2093.42	69.781	0.106
25	4020.07	134.002	0.006	25	2088.26	69.608	0.104
26	4233.64	141.121	0.034	26	1933.42	64.447	0.018
27	3974.61	132.487	0	27	2182.61	72.754	0.156
30	5904.59	196.82	0.25	30	2447.9	81.597	0.302
33	8040.92	268.031	0.528	33	3562.93	118.764	0.917
34	8812.02	293.734	0.628	34	3384.31	112.81	0.819
36	5532.96	184.432	0.202	36	3075.47	102.516	0.648
37	11682.7	389.425	1	37	3712.63	123.754	1
38	8015.27	267.176	0.524	38	3327.98	110.932	0.788
39	7323.44	244.115	0.434	39	3705.82	123.527	0.996
40	7302.32	243.411	0.432	40	3531.59	117.72	0.9
43	7012.93	233.764	0.394	43	3233.46	107.782	0.736
Test Set:				Test Set:			
Compound No.	**Distance Score**	**Mean Distance**	**Normalized Mean Distance**	**Compound No.**	**Distance Score**	**Mean Distance**	**Normalized Mean Distance**
10	3978.34	132.611	0	10	1984.96	66.165	0.047
11	4439.78	147.993	0.06	11	2072.72	69.091	0.095
15	4060.97	135.366	0.011	15	1979.77	65.992	0.044
20	4378.83	145.961	0.052	20	1927.53	64.251	0.015
22	4163.68	138.789	0.025	22	2417.46	80.582	0.285
31	5878.62	195.954	0.247	31	3409.78	113.659	0.833
32	8068.04	268.935	0.531	32	2919.48	97.316	0.562
41	6504.24	216.808	0.328	41	3133.1	104.437	0.68
42	6456.99	215.233	0.322	42	3135.82	104.527	0.682
44	6035.73	201.191	0.267	44	3499.93	116.664	0.883

Table 5: MLR Applicability domain results for model 7 and 8

3.4. Descriptors Contribution

Makhija and Kulkarni [xxxix] 2002, reported that molar refractivity, desolvation free energy for energy for octanol, non-common overlap steric volume, principal moment of inertia Y-component, difference volume, number of hydrogen bond acceptors, and sum of atomic polarizabilities are descriptors responsible for the HIV integrase inhibitory activities. Sahu et al., [5] 2008, reported that heat of formation, partition coefficient, lowest unoccupied molecular orbital, solvent accessible surface area and shape index play an important role for the HIV integrase inhibitory activities. Gupta and coworker [xl] 2012, reported that Moran autocorrelation-lag 4/weighted by atomic masses, Geary autocorrelation-log 7/weighted atomic masses, 3D-MoRSE signal 17/weighted by atomic masses, (R-CR----X)-represents an aromatic bond, Lovasz-pelikan index and neighborhoods information content play an important role in the activity. Recently, Adebimpe and co-worker [9] 2014, reported that radius of gyration, Zagreb index, wiener index and minimized energy play an important role in the HIV-1 integrase inhibition.

The present QSAR study, reveals that valence path cluster, order 4, maximum E-states for (strong) Hydrogen Bond donor, maximum E-states for weak Hydrogen bond acceptors, which are used in model 6 contribute negatively in the activity of HIV integrase inhibitors, which means decreasing the value of this physiochemical produce higher biological activity of the compound. valence path cluster, order 5, maximum E-states descriptors of strength for potential hydrogen bond of path length 9, and non-directional WHIM, weighted by atomic masses used in Model 6 contribute positively to the activity. Increasing the value of this descriptors produce higher activity of the compound. Partition coefficient, molecular surface area, ovality, polar surface area, accessible polar area corresponding to absolute values of the electrostatic potential greater than 75 and minimum values of the local ionization potential (as mapped on to an electron density surface) used in model 8 play important role in the HIV-1 integrase inhibition.

4. Conclusion

This study obtained a multivariate QSAR model for a set of β −Diketo acid, Diketo acid and Carboxamide derivatives that have the capability of inhibiting in vitro strain of anti-HIV-1 IN. The LOO cross validation, the Y-randomization technique, and the external validation indicated that the model is significant, robust and has good internal and external predictability. QSAR was performed using robust statistical technique GFA and MLR, coupled with the of different classes of descriptors. The QSAR model was obtained from GFA (Model 5 and 6) with explain variance and predicted variance 94.77%, 86.54% and 95.88%, 73.64% respectively. The quality of models obtained from MLR (Model 10 and 12) are of comparable range with explain variance 87.68% and 89.32% and predicted variance 72.14% and 81.06% respectively. All the developed QSAR models haveWD.mass (GFA) and LogP, P-Area(75) and PSA (MLR) that indicates that these variables are more important to explain the anti-HIV activity of β −Diketo acid, Diketo acid and Carboxamide derivatives. The negative coefficient of LogP and maxEIPot indicate that these parameters are detrimental to activity when increased. The positive coefficient of WD.mass and P-Area(75) indicates that these parameters are conducive to activity when increased. In conclusion, the QSAR study of β −Diketo acid, Diketo acid and Carboxamide compounds with the volume of partition coefficient (LogP) should be less while the Non-directional WHIM, weighted by atomic masses (WD.mass) should be high for their anti-HIV activity. The information generated from the present is useful in the design of more potent β −Diketo acid, Diketo acid and Carboxamide derivatives as anti-HIV agents.

4.1. Acknowledgement

The authors are thankful to Mr. OluwaseyiAdebirin for given valuable suggestions and Mr. David Ebuka Arthur to provide the softwares.

4.2. Conflict of Interest

They are no conflict of interest

5. References

i. Daniela I., Luminita C., Simona F., and Mircea M., (2013). A quantitative structure–activity relationships study for the anti-HIV-1 activities of 1-[(2-hydroxyethoxy)methyl]-6--(phenylthio)thymine derivatives using the multiple linear regression and partial least squares methodologies. Journal of the Serbian Chemical Society 78 (4) 495–506.

ii. Jan, B., Miguel S., Erik De C., et al. (2005). Pyridine N-oxide derivatives: unusual anti-HIV compounds with multiple mechanisms of antiviral action. Journal of Antimicrobial Chemotherapy 55, 135-138.

iii. UNAIDS, (2011). World AIDS Day Report 2011, accessed on 02/05/ 2012, http://www.unaids.org/en/media/unaids/contentassets/documents/unaidspublication/2011/JC2216_WorldAIDSday_report_2011_en.pdf.

iv. Ole S Pedersen and Erik B Pedersen (1999). Non-nucleoside reverse transcriptase inhibitors: the NNRTI boom. Antiviral Chemistry & Chemotherapy 10:285–314.

v. Sahu K. K., V. Ravichandran, Prateek K. J., Simant S., et al. (2008). QSAR Analysis of Chicoric Acid Derivatives as HIV–1 Integrase Inhibitors. ActaChim. Slov.,55, 138–145.

vi. RuslinHadanu and Syamsudin (2013). Quantitative Structure-Activity Relationship Analysis of Antimalarial Compound of Mangostin Derivatives Using Regression Linear Approach. Asian Journal of Chemistry 25 (11) 6136-6140.

vii. Ravichandran V., Harish R., Abhishek J., et al. (2011). Validation of QSAR models – Strategies and importance. International Journal of Drug Design and Discovery 2 (3) 511-519.

viii. Sechi M (2004) Design and synthesis of novel indole a-diketo acid derivatives as HIV-1 integrase inhibitors. Journal of Medicinal Chemistry 47: 5298–5310.

ix. ArodolaOlayideAdebimpe, RadhaCharan Dash and Mohmoud E.S. Soliman (2014). QSAR Study on Diketo Acid and Carboxamide Derivatives as Potent HIV-1 Integrase Inhibitor. Letters in Drug design & Discovery, volume 11, No. 5, pp 000-000

x. CS Chem3D Ultra Cambridge Soft Corporation, Cambridge, USA

xi. Wavefunction, (2013), Inc. Spartan'14, version 1.1.2, Irvine, California, USA.

xii. Yap Chun Wei. Inc. PaDEL-Descriptor, version 2.18. A software to calculate molecular descriptors and fingerprints.http://padel.nus.edu.sg.

xiii. D. Rogers, and A. J. Hopfinger, (1994). Application of Genetic Functional Approximation to Quantitative Structure-Activity Relationships and Quantitative Structure-Property Relationships. Journal of Chemical Information and Computer Science; vol.4, pp854-866.

xiv. Hunger, J., Huttner, G., (1999). Optimization and analysis of force field parameters by combination of genetic algorithms and neural networks. Journal Computational Chemistry; vol. 20, pp 455–471.

xv. Ahmad, S., Gromiha, M.M., (2003). Design and training of a neural network for predicting the solvent accessibility of proteins. Journal of Computational Chemistry; vol. 24, pp 1313–1320.

xvi. Holland, J.H., (1975). Adaptation in Natural and Artificial Systems. The University of Michigan Press, Ann Arbor, MI. (2nd ed., (1992) Boston, MA: MIT Press).

xvii. V.K. Srivastav, M. Tiwari, (2013). QSAR and docking studies of coumarin derivatives as potent HIV-1 integrase inhibitors. Arabian Journal of Chemistry vol. 7, pp 1-14.

xviii. Nitin S. Sapre, TarangBhati, Swagata Gupta, NilanjanaPancholi, UrmilaRaghuvanshi, DivyaDubey, VandanaRajopadhyay, NeelimaSapre, (2011). Computational modeling studies on anti-1 non-nucleoside reverse transcriptase inhibition by dihydroalkoxybenzyloxopyrimidines analogues: an electropological atomistic approach. Journal of Biophysical Chemistry; vol 2, No. 3, pp 361-372.

xix. Wang, H.Y., Li, Y., Ding, J., Wang, Y. and Chang, Y.Q. (2008) Prediction of binding affinity for estrogen receptor alpha modulators using statistical learning approaches. Journal of Molecular Diversity, 12, 93-102.

xx. Myers, R.H. (1990). Classical and Modern Regression with Applications. 2nd ed. Duxbury Advanced Series in Statistics. PWS-Kent Publishing Co. Boston, MA.

xxi. Shapiro, S., Guggenheim, B., 1998. Inhibition of oral bacteria by phenolic compounds. Part 1. QSAR analysis using molecular connectivity. Journal of Quantitative Structure-Activity Relationship; vol. 17, pp 327–337.

xxii. Jaiswal, M., Khadikar, P.V., Scozzafava, A., Supuran, C.T., (2004). Carbonic anhydrase inhibitors: the first QSAR study on inhibition of tumor-associated isoenzyme IX with aromatic and heterocyclic sulfonamides. Bioorganic and Medicinal. Chemistry Letters; vol. 14, pp3283–3290.

xxiii. Jin Soo Song, Taesung Moon, Kee Dal Nam, Jae Kyun Lee, Hoh-Gyu Hahn, Eui-JuChoic,and Chang No Yoona, (2008). Quantitative structural–activity relationship (QSAR) study for fungicidal activities of thiazoline derivatives against rice blast. Bioorganic & Medicinal Chemistry Letters vol. 18, pp 2133–2142

xxiv. R. Veerasamy, S. Ravichandran, A. Jain, H. Rajak, R.K. Agrawal, (2009). QSAR studies on novel anti-HIV agents using FA-MLR, FA-PLS and PCRA techniques. Digest Journal of Nanomaterials and Biostructures: vol.4; pp823-834.

xxv. Doreswany and Chanabasyya M. Vastrad (2012). Predictive comparative QSAR analysis of sulfathiazole analogues as mycobacterium tuberculosis H37RV. Journal of advanced bioinformatics applications and research: vol. 3; issue 3, pp 379-390.

xxvi. LuizFrederico Motta and Wanda Poreira Almeida, (2011). Quantitative structure-activity relationship (QSAR) of a series of Ketone derivatives as anti-candida albicans. International Journal of Drug Discovery, vol. 3, issue 2, pp 100-117.

xxvii. Pogliani L., (1994). “Structure property relationships of amino acids and some dipeptides,” Amino Acids, vol. 6, no. 2, pp. 141–153.

xxviii. Pogliani L., (1996). “Modeling with special descriptors derived from a medium-sized set of connectivity indices,” Journal of PhysicalChemistry, vol. 100, no. 46, pp. 18065–18077.

xxix. Chaterjee, S.; Hadi, A.S.; Price, B. (2000). Regression analysis by examples, 3rd Ed. Wiley; New York.

xxx. Surendra Kumar, Vineet Singh, Meena Tiwari, (2011). Qsar Modeling of the Inhibition of Reverse Transcriptase Enzyme with Benzimidazolone Analogs. Medicinal Chemistry Research; vol. 20, pp 1530-1541.

xxxi. Roy, K. (2007), "On some aspects of validation of predictive quantitative structure-activity relationship models", Expert OpinionDrug Discovery. 2 (12): 1567–1577.

xxxii. Leonard JT, Roy K. (2006). On Seleection of Training and Test Sets for the Development of Predictive QSAR Models. QSAR Combinatorial Science. Volume 25, pp235-251.

xxxiii. Roy ParthaPratim and Roy Kunal. (2008). On Some Aspects of Variable Selection for Partial Least Squares Regression Models. QSAR & Combinatorial Science, volume 27, issue 3, pp302-313.

xxxiv. Golbraikh A. and A. Tropsha, (2002). “Beware of q2!.” Journal of Molecular Graphics and Modelling, vol. 20, no. 4, pp. 269-276.

xxxv. Tropsha A, Gramatica P, Gombar VK, (2003). The importance of being Earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. QSAR Comb.Sci.; vol, 22: pp 69-77.

xxxvi. E. C. Ibezim1, P. R. Duchowicz, N. E. Ibezim, L. M. A. Mullen, I. V. Onyishi4, S. A. Brown and E. A. Castro, (2009). Computer - aided linear modeling employing QSAR for drug discovery Scientific Research and Essay Vol. 4 (13), pp. 1559-1564.

xxxvii. SupratikKar and Kunal Roy, (2011). Development and validation of a robust QSAR model for prediction of carcinogenicity of drugs. Indian Journal of Biochemistry & Biophysics; vol. 48, pp 111-122.

xxxviii. Jaworska, J.; Nikolova-Jeliazkova, N.; Aldenberg, T. (2005). QSAR applicabilty domain estimation by projection of the training set descriptor space: A review. Altern. Lab. Anim. 33, 445–459.

xxxix. Mahindra T. Makhija and Vithal M. Kulkarni, (2002). QSAR of HIV-1 Integrase Inhibitors by Genetic Function Approximation Method. Bioorganic & Medicinal Chemistry 10, pp1483–1497.

xl. Pawan Gupta1, PrabhaGarg and Nilanjan Roy, (2012). Identification of Novel HIV-1 Integrase Inhibitors Using Shape-Based Screening, QSAR, and Docking Approach. ChemBiol Drug Des 2012; 79: 835–849.

Annexure

	LogIC50	LogP	ZPE	Area	MinLocIonPot	Ovality	BCUTw-1h	minHBd	maxHBint3	HybRatio	WD.mass
LogIC50	1										
LogP	-0.606	1									
ZPE	0.652	-0.436	1								
Area	0.619	-0.453	0.981	1							
MinLocIonPot	0.529	-0.178	0.251	0.181	1						
Ovality	0.635	-0.398	0.947	0.975	0.1094	1					
BCUTw-1h	-0.068	-0.208	-0.05	-0.034	0.259	-0.097	1				
minHBd	-0.718	0.257	-0.287	-0.206	-0.743	-0.206	-0.083	1			
maxHBint3	-0.547	0.115	-0.076	0.019	-0.707	0.0313	-0.071	0.963	1		
HybRatio	0.599	-0.439	0.469	0.396	0.0981	0.4557	-0.062	-0.49	-0.413	1	
WD.mass	-0.19	-0.299	-0.166	-0.127	-0.106	-0.213	0.6868	0.385	0.3711	-0.251	1

Table S1: The correlation matrix between the physicochemical parameters and the biological activity.

Abbreviation	Description	Class
BCUTDescriptor		
BCUTw-1h	nlow highest atom weighted BCUTS	2D
ElectrotopologicalStateAtomTypeDescriptor		
minHBd	Minimum E-States for (strong) Hydrogen Bond donors	2D
maxHBd	Maximum E-States for (strong) Hydrogen Bond donors	2D
maxwHBd	Maximum E-States for weak Hydrogen Bond donors	2D
maxHBint3	Maximum E-State descriptors of strength for potential Hydrogen Bonds of path length 3	2D
maxwHBa	Maximum E-States for weak Hydrogen Bond acceptors	2D
maxHBint9	Maximum E-State descriptors of strength for potential Hydrogen Bonds of path length 9	2D
HybridizationRatioDescriptor		
HybRatio	Fraction of sp3 carbons to sp2 carbons	2D
WHIMDescriptor		
WD.mass	Non-directional WHIM, weighted by atomic masses	3D
ChiPathClusterDescriptor		
VPC-4	Valence path cluster, order 4	2D
VPC-5	Valence path cluster, order 5	2D
Thermodynamic Descriptor		
LogP	partition Coefficient	3D
Area	Molecular Surface Area	
minLocIonPot	min. values of the local ionization potential (as mapped on to an electron density surface)	3D
PSA	polar surface area	
Acc.P-Area(75)	Accessible polar area corresponding to absolute values of the electrostatic potential greater than 75	3D
Ovality	Ovality	3D
ZPE	Zero-point energy	3D

Table S2: the Brief description of the descriptors

	LogIC50	***LogP***	***Acc. P-Area (75)***	***Area***	***MinLocIonPot***	***Ovality***	***PSA***	***VPC-4***	***VPC-5***
LogIC50	1								
LogP	-0.606	1							
Acc. P-Area (75)	0.488	-0.427	1						
Area	0.619	-0.453	0.654	1					
MinLocIonPot	0.529	-0.178	-0.126	0.1806	1				
Ovality	0.635	-0.398	0.682	0.9752	0.109	1			
PSA	0.493	-0.568	0.8214	0.8082	-0.16	0.846	1		
VPC-4	0.628	-0.631	0.5412	0.8434	0.176	0.7971	0.6711	1	
VPC-5	0.592	-0.718	0.4413	0.7207	0.143	0.6544	0.5829	0.947	1
maxHBd	-0.596	0.115	0.2037	-0.063	-0.71	-0.059	0.2627	-0.15	-0.1766
maxwHBa	-0.757	0.6484	-0.758	-0.712	-0.11	-0.741	-0.77	-0.73	-0.6502
maxHBint9	0.728	-0.499	0.6938	0.8937	0.241	0.915	0.8327	0.763	0.6095
WD.mass	-0.19	-0.299	0.057	-0.127	-0.11	-0.213	-0.044	0.072	0.0775

Table S3: The correlation matrix between the physicochemical parameters and the biological activity.

maxHBd	***maxwHBa***	***maxHBint9***	***WD.mass***
1			
0.0861	1		
-0.07	-0.7777	1	
0.4332	0.0892	-0.096	1

Table S3: Cont;d