Journal of Advances in Medical and Pharmaceutical Sciences
5(2): 1-18, 2016, Article no.JAMPS.22227
ISSN: 2394-1111

SCIENCEDOMAIN *international*
www.sciencedomain.org

Development and Estimation of an in Silico Model for Anti-HIV-1 Integrase Inhibitor Using Genetic Function Approximation

Emmanuel Israel Edache[1*], Adamu Uzairu[1] and Stephen Eyije Abechi[1]

[1]Department of Chemistry, Ahmadu Bello University, Zaria, Nigeria.

Authors' contributions

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 manuscript.

Article Information

DOI: 10.9734/JAMPS/2016/22227
Editor(s):
(1) Esam Z. Dajani, President, IDDC Corporation, USA, Adjunct Professor of Medicine, Loyola University Chicago, USA.
Reviewers:
(1) Yahya Nural, Mersin University, Turkey.
(2) Jadwiga Soloducho, Wrocław University of Technology, Poland.
Complete Peer review History: http://sciencedomain.org/review-history/12116

Short Research Article

Received 23rd September 2015
Accepted 16th October 2015
Published 6th November 2015

ABSTRACT

Aim: Integrase inhibitors are an essential enzyme required for replication of the acquired immunodeficiency syndrome virus. It is a potent target for anti- HIV therapy. A QSAR study is performed on the series indole β-diketo, diketo acid and carboxamide derivatives in order to analyze the physicochemical requirements of integrase inhibitors and to provide structural insight into the binding mode of the molecules to the enzyme. This will help in the design of these molecules as integrase inhibitors and predicting the inhibitory activity of the newly designed analogues.
Materials and Methods: All the derivatives in the series were sketched using ChemDraw ultra v12.0.2 module of ChemOffice 2010 and the sketched structures were consequently used for the calculation of molecular descriptors available in QSAR software Spartan'14 and PaDEL-Descriptors software. Quantum, constitutional, topological and functional group descriptors for all molecules were calculated using Spartan'14 v1.1.2, 2013 and PaDEL-Descriptors software v2.18, 2011 and correlation between the biological activity and molecular descriptors was found through genetic

*Corresponding author: E-mail: inalegwu334real@yahoo.com;

function approximation adopted by statistical program material studio v7.0.

Results: The generated QSAR models revealed that SsF, minHBint3, minHdsCH, FPSA-1 and RHSA descriptors have good correlation to the integrase inhibitors activity.

Conclusion: The results obtained by regression analysis indicated that minHBint3, minHdsCH and RHSA is negatively contributing to inhibitory activity thus; enhancement of inhibitory activity can be achieved by decreasing the respective descriptors. Positive contribution of SsF and FPSA-1 specifies that increase of sum of atom-type E-state: -F and PPSA-1/total molecular surface area, will impart positive influence on activity.

Keywords: Quantitative Structure–Activity Relationships (QSAR); GFA approach; density function theory indole β-diketo; diketo acid and carboxamide derivatives; integrase inhibitors.

1. BACKGROUND

A needed condition in the retroviral life cycle is the integration of the viral double-stranded DNA into the host chromosome. HIV-1 integrase (IN) enzyme removes a dinucleotide next to a conserved cytosine–adenine sequence from each 3'-end of the viral DNA [1]. Formerly, IN catalyzes linking of the processed viral 3'-ends to the 5'-ends of strand disruptions in the host DNA. HIV-1 IN enzyme has no equal in host cell and is also an essential enzyme for effective viral replication [2]. Inhibitors of this enzyme are of paramount importance for the treatment of HIV infection [3]. A deep research is being carried out on HIV-1 IN protein, but only one US-FDA-approved drug 'Raltagravir' is available in market (http://www.fda.gov) [4], which is administered in combination with other antiretroviral agents [5] and one more, Elvitegravir [6], is now entering phase III clinical trials. Therefore, current situation warrants more HIV-1 IN inhibitors with good potency.

Numerous molecular modeling aspects have been involved in the development of potent HIV-1 IN inhibitors, e.g., QSAR, pharmacophore mapping, and docking studies. A number of 3D QSAR studies were done to obtain understandings into the structural requirement of HIV-1 IN inhibitors, which can be useful in the enhancement of HIV-1 inhibitory activity [7]. Likewise, 2D QSAR was done on different series of molecules and found that topological indices [8]. Recently published 2D QSAR analysis of coumarin derivatives result showed that Henry's law Constant, Partition Coefficient and Dipole moment-Z component significantly affect the inhibition of HIV-1 IN activity [9].

The scarcity of new affordable drugs has not only complicated the clinical management of HIV-1 in pervasive areas, but has also resulted in an increase in the mortality rate [10]. This situation emphasizes needs for urgent discovery of new anti-HIV agents. Nevertheless lower abundance of Raltagravir and related products encourage the medicinal chemists to search for new chemical pharmacophores which may prove effective as anti-HIV. Several experimental methods available for screening the biological activity of chemicals (e.g., in vivo and in vitro assay tests). These methods have been applied widely to rat and mouse [11]. However, these methods are costly, time-consuming, and can potentially produce toxic side products. The efficient way to obtain a complete set of the data, without the necessity of performing expensive laboratory experiments is apply quantitative structure–activity relationship (QSAR) techniques. The QSAR is one of the most important areas in chemometrics, and is a valuable tool that is used extensively in drug design and medicinal chemistry [12, 13]. Chemical and biological effects are related closely to molecular physicochemical properties by QSAR technique [14]. Once a reliable QSAR model is established, the activities of molecules can be predicted, and the structural features that play a significant role in the biological process can be identified. The advances in QSAR studies have therefore, widened the scope of rational drug design as well as the search for the mechanisms of drug actions. Many different methodologies, such as multiple linear regression (MLR), partial least squares (PLS), principal component analysis (PCA), support vector machine (SVM), heuristic method (HM), and different types of artificial neural networks (ANN), can be applied for QSAR development. Genetic function approximation (GFA) has gained great popularity in QSAR research. The main aim of the present work is to establish a new QSAR model for predicting anti-HIV activity of 44 indole β -diketo, diketo acid and carboxamide derivatives using GFA technique.

2. DATA SET AND METHODS

2.1 Data Set

In this study, a data set of 44 indole β-diketo, diketo acid and carboxamide derivatives was collected from the literature [15, 16]. The chemical structures and anti-HIV activity (IC_{50}) of these 44 molecules are presented in Table 1. The IC_{50} values were converted into its logarithmic scale $-\log(IC_{50}) = pIC_{50}$, to reduce the skewness of the data set, which was then used for subsequent QSAR analysis as the response variable. It is essential to assess the predictive power of QSAR models by using a test set of molecules according to the following criteria:

(1) The anti-HIV activity values of the test set should span the training set several times;
(2) The biological assay methods for both the training set and test set should be the same or comparable;
(3) The test set should represent a balanced number of both active and inactive molecules for uniform sampling of the data set. The remaining molecules are taken as the training set in order to create an efficient QSAR model [17].

Table 1. Structures and biological activity of training and test set

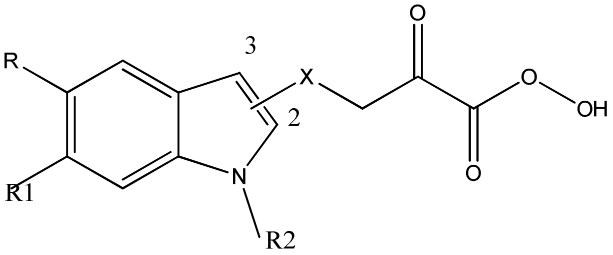

Fig. A

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

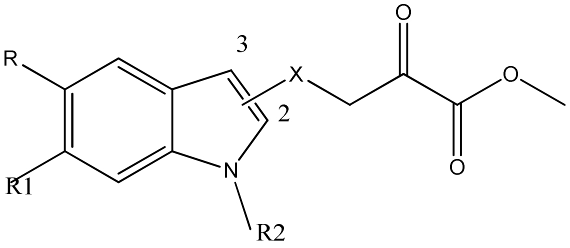

Fig. B

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

Compd No	R	R1	R2	X	Log IC_{50}
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

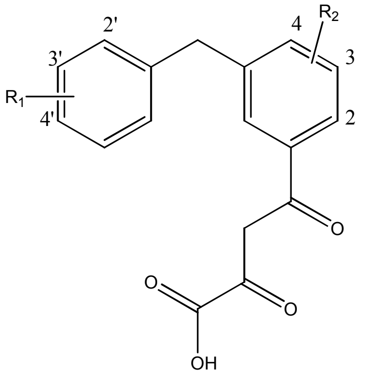

Fig. C

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

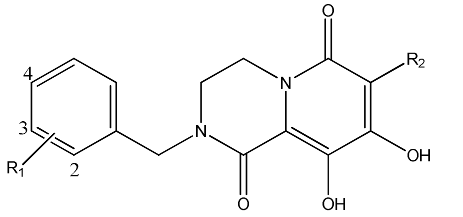

Fig. D

Compd No.	R1	R2	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

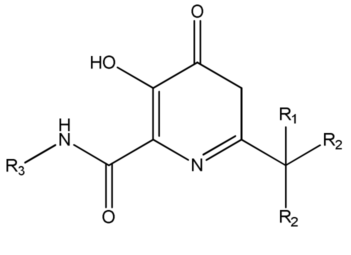

Fig. E

Compd No.	R1	R2	R3	LogIC50
30	$NHCOCH_3$	CH_3	4-fluorotoluene	2.1555
31	$NH\text{-}SO_2\text{-}CH_3$	CH_3	4-fluorotoluene	2.097
32	$NHCO\text{-}N(CH3)_2$	CH_3	4-fluorotoluene	1.745
33	$NHSO2\text{-}N(CH3)_2$	CH_3	4-fluorotoluene	1.921
34	$NHCOCO\text{-}N(CH3)_2$	CH_3	4-fluorotoluene	2.000
35	$NHCOCO\text{-}OCH_3$	CH_3	4-fluorotoluene	1.824
36	NHCOCO-OH	CH_3	4-fluorotoluene	2.398
37	$N(CH3)COCO\text{-}N(CH3)_2$	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

2.2 Descriptor Calculation

All of the molecules were drawn into the ChemDraw ultra version 12.0.2 software and transferred to Spartan’14 version 1.1.2 to create the three-dimensional (3D) structure, pre-optimized using the MM+ molecular mechanics force field. Then a more precise optimization was performed with the density functional theory (DFT) with Becke’s three-parameter hybrid functional [18] using LYP correlation functional [19]. The standard Pople’s 6-311G* using basis set was used. Descriptors were calculated by using the Spartan’14 v1.1.2 and PaDEL-Descriptor version 2.18 software package [20] which include: constitutional, topological, geometrical, electrostatic, charged partial surface area, quantum-chemical, molecular orbital and thermodynamic descriptors. Before commencing with the development of the QSAR model, the correlation matrix of about 2000 descriptors was calculated and highly correlated descriptors, with correlation values above 0.98, were removed. Furthermore, descriptors with constant values as well as those with poor correlation with the anti-HIV activity were discarded; some descriptors having zero value were also discarded. Finally, remained descriptors were considered for statistical fitting using the GFA method.

2.3 Computational Methods

Density functional theory (DFT) were used in this study. These methods have become very popular in recent years because they can reach similar precision to other methods in less time and less cost from the computational point of view. In agreement with the DFT results, energy of the fundamental state of a polyelectronic system can be expressed through the total electronic density, and in fact, the use of electronic density instead of wave function for calculating the energy constitutes the fundamental base of DFT [21] using the B3LYP functional [18,19] and a 6-311G* basic set. The B3LYP, a version of DFT method, uses Becke’s three-parameter functional (B3) and includes a mixture of HF and DFT exchange terms associated with the gradient correlation functional of Lee Yang and Parr (LYP). The geometry of all species under investigation was determined by optimizing all geometrical variables without any symmetry constraints [22].

2.4 Genetic Function Approximation for Descriptor Selection

Genetic function approximation (GFA) are governed by biological evolution rules [23]. The GFA, which are based on the principles of Darwinian evolution, have emerged as robust optimization and search methods [24]. In a GFA feature selection procedure, potential solutions for the problem being studied are subsets of molecular descriptors. They are represented as data structures called chromosomes, which are binary strings of length N (the total number of available features), with a zero or one in position i indicating the absence or presence of feature i in the set. The initial population of chromosomes is usually generated randomly. After that, GA runs in cycles. The fitness of each chromosome is evaluated by the fitness function. The fitness function used here was the leave-one-out, cross-validated correlation coefficient (Q^2_{LOO}). New chromosomes are then created by genetic operators such as crossovers and mutations. Crossover occurs when two parent chromosomes exchange parts of their corresponding elements. Mutations induce sporadic alterations of randomly selected chromosome elements. In each cycle, a new chromosome (feature set) is produced either by mutation or crossover on the selected parents, and it is compared with the worst member of the existing population. If the new one is better, it becomes a member of the population, and the original worst one is discarded; if not, the new one is discarded, and GFA goes into next generation with the population unchanged. The GFA cycle is repeated until a satisfactory descriptor set is found or a pre-set limit of generation is reached. The GFA program was perform in Material Studio version 7.0.

2.5 The Need for Defining the Limitation

1) If there is a measurement error in the experimental data, it is very likely that false correlations may arise.
2) If the training dataset is not large enough, the data collected may not reflect the complete property space. Consequently, many QSAR results cannot be used to confidently predict the most likely compounds of the best activity.
3) The third aspect is the chemical domain. There are always chemicals which do not follow the given simple relationship between dependent (pIC50) and descriptors.

3. RESULTS AND DISCUSSION

A QSAR analysis was done to discover the structure–activity relationship of different 44 indole β-diketo, diketo acid and carboxamide derivatives acting as anti-HIV. In a QSAR study, normally, the quality of a model is expressed by its fitting and prediction ability. In order to build and test model, a data set of 44 compounds was separated into a training set of 30 compounds, which was used to build model and a test set of 11 compounds, which was applied to evaluate the built model. The GFA analysis led to the derivation of five model, with five descriptors. With the selected descriptors, we have built the linear model using the training set data, and obtained the following equation (Table 2):

Table 2. The linear model using the training set data

No.	Equation	Definitions
1	Y = 0.0269 * X152 - 0.0982* X157 - 1.2987* X165 +5.2783* X295- 2.9555* X310+ 0.5865	X152 : SsF; X157 : minHBint3; X165 : minHdsCH X295 : FPSA-1; X310 : RHSA
2	Y = - 0.0977* X157- 1.5300* X165+ 0.4739* X178 - 0.0087* X289+ 0.0053* X309- 2.9440	X157 : minHBint3; X165 : minHdsCH; X178 : maxHBa; X289 : PNSA-1; X309 : TPSA
3	Y = 7.7558* X83- 0.1224* X157- 1.3135* X165+ 3.6339* X295- 3.6413* X310+ 2.1059	X83 : VC-6; X157 : minHBint3; X165 : minHdsCH X295 : FPSA-1; X310 : RHSA
4	Y = - 0.0909* X157- 1.6131* X165+ 0.5841* X178 + 5.9132* X295- 0.0043* X308 - 7.1316	X157 : minHBint3; X165 : minHdsCH; X178 : maxHBa; X295 : FPSA-1; X308 : THSA
5	Y = - 1.9310* X37- 1.8798* X153- 0.9923* X165+ 1.7979* X296- 5.3707* X310+ 5.0087	X37 : ATSc1; X153 : minHBd; X165 : minHdsCH; X296 : FPSA-2; X310 : RHSA

Table 3. Summary of input data for genetic function approximation

Number of rows requested	30
Number of rows used	30
Number of rows omitted due to invalid row description	0
Number of rows omitted due to invalid data	0
Number of columns requested	380
Number of columns used	380
Number of columns omitted due to invalid column description	0
Number of columns omitted due to invalid data	0
Number of cells omitted due to invalid data	0
Number of cells replaced by default value	0

In this equation, R^2 is the squared correlation coefficient, Q^2_{LOO}, and Q^2_{LNO} are the squared cross-validation coefficients for leave one out and leave many out respectively, F is the Fisher F statistic, and RMSE is the root mean square error (Table 4). The built model was used to predict the test set data, and the prediction results are given in Table 6. The predicted values for pIC_{50} for the compounds in the training and test sets using equation 1 were plotted against the experimental pIC_{50} values in Figs. 1, 2 and 3. As can be seen from Table 2 and Fig. 3, the predicted values for the pIC_{50} are in good agreement with those of the observed values.

Also, the plot of the residual for the predicted values of pIC_{50} for both the training and test sets against the observed pIC_{50} values are shown in Fig. 4. As can be seen the model did not show any proportional and systematic error, because the propagation of the residuals on both sides of zero is random.

3.1 QSAR Model Validation

The effectiveness of QSAR models is not just their capability to reproduce known data, proved by their fitting power (R^2), but primarily is their potential for predictive application. For this reason, the internal reliability of the training set was confirmed by using leave-one-out (LOO) cross-validation method to ensure the robustness of the model. The high calculated Q^2_{LOO} value, 0.9636 suggests a good internal validation. A second validation method was also developed on the basis of a leave-many-out (LNO) internal cross-validation method. In this case, a group of compounds including 17% of the training data set were left out and predicted later by the model obtained with the remaining 83% of the data. This process was repeated 100 times for each one of the 100 unique subsets selected at random. The overall mean for this process (17% full-leave out cross-validation), Q^2_{L5O} = 0.8816 indicates the robustness and stability of the built model (Table 4). The difference between the R^2 and R^2_{adj} value is less than 0.3 indicates that the number of descriptors involved in the QSAR model is acceptable. The number of descriptors is not acceptable if the difference is more than 0.3. For good predictability $R^2 – Q^2$ value is less than 0.3.The results in (Table 6), has shown that the selected model presented high external predictability, considering the proposed limits. The values of K or K' and the relation $\left|r_0^2 - r_o'^2\right|$ are inside the acceptable range [25,26].

Table 4. Comparison of statistical quality and internal validation parameters of different models

	Parameter	Equation 1	Equation 2	Equation 3	Equation 4	Equation 5
1	Friedman LOF	0.0683	0.0718	0.0737	0.0741	0.0754
2	R-squared	0.9782	0.9771	0.9765	0.9764	0.9760
3	Adjusted R-squared	0.9737	0.9723	0.9716	0.9714	0.9709
4	Cross validated R-squared	0.9636	0.9654	0.9627	0.9624	0.9652
5	Significant Regression	Yes	Yes	Yes	Yes	Yes
6	Significance-of-regression F-value	215.4925	204.7373	199.4729	198.3318	194.7928
7	Critical SOR F-value (95%)	2.6441	2.6441	2.6441	2.6441	2.6441
8	Replicate points	0	0	0	0	0
9	Leave five out Cross validated R-squared	0.8816	0.8786	0.8721	0.8777	0.8816
10	Lack-of-fit points	24	24	24	24	24
11	Min expt. error for non-significant LOF (95%)	0.1001	0.1027		0.1043	0.1052
12	Root mean square error (RMSE)	0.1105	0.1131	0.1145	0.1149	0.1162

Table 5. Predicted and Residual values of training and test set

No.	pIC50	Eq1:Prd	Eq1: Resi	Eq2: Prd	Eq2: Resi	Eq3: Prd	Eq3: Resi	Eq4: Prd	Eq4: Resi	Eq5: Prd	Eq5: Resi
2	0.301	0.3508	-0.0498	0.3171	-0.0161	0.4837	-0.1827	0.2329	0.0681	0.3488	-0.0478
3	0.204	0.3053	-0.1013	0.3259	-0.1219	0.2963	-0.0923	0.3307	-0.1267	0.2288	-0.0248
4	0.699	0.6282	0.0708	0.6008	0.0982	0.6048	0.0942	0.578	0.1210	0.4745	0.2245
5	0.00	0.0451	-0.0451	0.0998	-0.0998	-0.0121	0.0121	0.1805	-0.1805	0.0625	-0.0625
6	0.301	0.3128	-0.0118	0.3162	-0.0152	0.3016	-0.0006	0.3485	-0.0475	0.3498	-0.0488
9	0.477	0.5267	-0.0497	0.4450	0.032	0.4190	0.058	0.4718	0.0052	0.2792	0.1978
10	0.477	0.5427	-0.0657	0.6266	-0.1496	0.6033	-0.1263	0.5942	-0.1172	0.512	-0.0360
12	1.653	1.5749	0.0781	1.5483	0.1047	1.5966	0.0564	1.6128	0.0402	1.5632	0.0898
13	1.699	1.6278	0.0712	1.6040	0.0950	1.7074	-0.0084	1.6025	0.0965	1.6457	0.0533
14	1.813	1.8768	-0.0638	1.9034	-0.0904	1.885	-0.072	1.9135	-0.1005	1.8916	-0.0786
15	1.778	1.7526	0.0254	1.8221	-0.0441	1.8399	-0.0619	1.845	-0.067	1.8143	-0.0363
17	0.000	0.0722	-0.0722	0.0154	-0.0154	0.1981	-0.1981	-0.0519	0.0519	0.332	-0.3320
18	0.602	0.4780	0.1240	0.6024	-0.0005	0.3507	0.2513	0.6128	-0.0108	0.5847	0.0173
19	0.824	0.7638	0.0602	0.7174	0.1066	0.7143	0.1097	0.7351	0.0889	0.8184	0.0056
20	0.854	0.7810	0.0730	0.7039	0.1501	0.7294	0.1246	0.7053	0.1487	0.8401	0.0139
21	2.155	2.2291	-0.0741	2.2743	-0.1193	2.0845	0.0705	2.2951	-0.1401	2.0981	0.0569
24	1.921	1.9172	0.0038	1.9212	-0.0002	1.9368	-0.0158	1.9990	-0.0780	1.8954	0.0256
25	2.00	2.2170	-0.2170	2.1659	-0.1659	2.1317	-0.1317	2.0773	-0.0773	2.1171	-0.1171
27	2.398	2.1899	0.2081	2.2775	0.1205	2.3171	0.0809	2.2072	0.1908	2.3956	0.0024
28	1.824	1.6518	0.1722	1.6907	0.1333	1.8578	-0.0338	1.7426	0.0814	1.7547	0.0693
29	1.699	1.8761	-0.1771	1.784	-0.085	1.7854	-0.0864	1.7844	-0.0854	1.9083	-0.2093
30	1.824	1.7711	0.0529	1.7087	0.1153	1.7352	0.0888	1.7368	0.0872	1.7991	0.0249
31	2.155	1.9126	0.2424	1.9011	0.2539	1.9070	0.248	1.9009	0.2541	2.0159	0.1391
33	2.097	2.0919	0.0051	2.0279	0.0691	2.114	-0.017	1.9919	0.1051	2.0907	0.0063
35	1.824	1.9200	-0.0960	1.8957	-0.0717	1.9381	-0.1141	1.8790	-0.0550	1.8668	-0.0428
36	1.000	1.0860	-0.0860	0.9865	0.0135	0.8936	0.1064	0.9862	0.0138	0.9838	0.0162
38	0.432	0.4737	-0.0417	0.5452	-0.1132	0.5414	-0.1094	0.5677	-0.1357	0.5007	-0.0687
40	0.420	0.2912	0.1288	0.3206	0.0994	0.4187	0.0013	0.2984	0.1216	0.3677	0.0523
42	1.699	1.8968	-0.1978	1.9632	-0.2642	1.8609	-0.1619	1.9273	-0.2283	1.8356	-0.1366
43	1.523	1.4901	0.0328	1.5420	-0.090	1.4129	0.1100	1.5474	-0.0244	1.2778	0.2452

Test set

No.	pIC50	Eq1:Prd	Eq1: Resi	Eq2: Prd	Eq2: Resi	Eq3: Prd	Eq3: Resi	Eq4: Prd	Eq4: Resi	Eq5: Prd	Eq5: Resi
7	0.301	0.6276	-0.3266	-0.1253	0.4263	0.5412	-0.2402	0.5003	-0.1993	0.3998	-0.0988
8	0.477	0.4046	0.0724	0.9028	-0.4258	0.5407	-0.0637	0.4425	0.0345	0.4225	0.0545
11	0	0.1967	-0.1967	-0.1566	0.1566	0.1138	-0.1138	0.1601	-0.1602	0.1913	-0.1913
16	1.415	1.7782	-0.3632	0.2419	1.1731	1.7074	-0.2924	1.9038	-0.4888	1.6469	-0.2319
22	2.097	2.1128	-0.0158	1.3559	0.7411	2.1365	-0.0395	2.1398	-0.0428	2.3839	-0.2869
23	1.745	2.1175	-0.3725	0.6467	1.0983	1.9694	-0.2244	2.2199	-0.4749	2.042	-0.297
26	1.824	2.2232	-0.3992	0.7881	1.0359	2.1587	-0.3347	2.1059	-0.2819	2.0422	-0.2182
32	2.155	2.1964	-0.0414	1.0199	1.1351	2.1491	0.0059	2.2088	-0.0538	2.2619	-0.1069
34	2.222	2.1981	0.0239	0.8526	1.3694	2.1371	0.0849	2.1273	0.0947	2.446	-0.224
37	0.638	0.7283	-0.0903	-0.2682	0.9062	0.7501	-0.1121	0.8108	-0.1728	0.7548	-0.1168
44	1.699	1.3893	0.3097	1.7864	-0.0874	1.325	0.3740	1.4491	0.2499	1.1638	0.5352

pIC50: Actual values; Eq1: predicted value for equation 1, Residual1: residual value for equation 1; Eq2: predicted value for equation 2, Residual2: residual value for equation 2; Eq3: predicted value for equation 3, Residual3: residual value for equation 3; Eq4: predicted value for equation 4, Residual4: residual value for equation 4; Eq5: predicted value for equation 5, Residual5: residual value for equation 5

Table 6. External validation parameters of equation 1

Parameter	value	Threhold value
r^2	0.9246	≥ 0.5
r_o^2	0.9222	-
$r_0^{;2}$	0.9096	-
$r_{m(LOO)}^2$	0.9559	≥ 0.5
$r_{m(test)}^2$	0.8794	≥ 0.5
$r_{m(overall)}^2$	0.9468	≥ 0.5
R_{pred}^2	0.8985	≥ 0.5
Q_{f2}^2	0.8961	≥ 0.5
$\lvert r_0^2 - r_o^{'2} \rvert$	0.0126	≤ 0.3
K	0.9245	$0.85 \leq K \leq 1.15$
$r^2 - r_o^2 / _{r^2}$	0.0026	< 0.1
$K^{'}$	1.0601	$0.85 \leq K \leq 1.15$
$r^2 - r_o^{'2} / _{r^2}$	0.0162	< 0.1
$PRESS$	0.6104	Less value is better
$RMSEP$	0.2356	Less value is better
R_p^2	0.8928	≥ 0.5

An illustration of the results obtained for each combination studied is given in Table 4. The Q^2 value obtained for all the models are well above the stipulated value of 0.5 with equation 1 showing the highest Q^2 values of 0.9636. However, the external validation of the models showed a wide range of variation in the values of R^2_{pred}. The parameter, $r^2_{m(overall)}$, was used which penalizes a model for large differences in observed and predicted activity. A model may be considered satisfactory when $r^2_{m(overall)}$ is greater than 0.5. The value of $r^2_{m(overall)}$ takes into consideration prediction of both training and test set compounds and maintains a balance between the values of Q^2 and R^2_{pred}. The $r^2_{m(LOO)}$ parameter for a giving model indicates the extent of deviation of the Q^2_{LOO} predicted activity values from the observed ones for the training set compounds. This implies that equation 1, despite having an acceptable Q^2, is capable of accurately predicting the activities of the some training set molecules and this is reflected in the value of $r^2_{m(LOO)}$. Remarkably, equation 1 has the maximum Q^2 value (0.9737) and $r^2_{m(LOO)}$ value of this model is 0.9559. the $r^2_{m(test)}$ parameter determines the extent of deviation of the predicted activity from the observed activity values of test set compounds where the predicted activity is calculated on the basis of the model developed using the corresponding training set. Equation 1 show acceptable values of R^2_{pred} and $r^2_{m(test)}$. From these model the difference between the value of R^2_{pred} and $r^2_{m(test)}$ is very low (less than 0.1) indicating that the predicted activity values of the test set compounds obtained from the corresponding models are very close to the corresponding observed activities of the compounds.

The developed models were further validated by the process randomization technique. The values of R^2_r and R^2 were determined which were then used for calculating the value of R^2_p. models with R^2_p values greater than 0.5 are considered to be statistically robust. If the value of R^2_p is less than 0.5, then it may be concluded that the outcome of the models is merely by chance and they are not at all well predictive for truly external datasets. The values of R^2_p in equation 1 crossed the threshold value of 0.5 and therefore, equation 1 may be considered to be statistically robust. These result suggest that this combination of training and test sets is the best one.

3.2 Euclidean Based Applicability Domain (AD)

Applicability domain (AD) is the physicochemical, structural or biological space, knowledge or information 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. Euclidean based application 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, 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. biological activity) including both training and test set (Table 7). 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 [25,26].

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

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

Where R^2 is the correlation coefficient of the multiple regression between the variables within the model which is defined as:

$$R^2 = 1 - \frac{\sum\left(y_{pred} - y_{actual}\right)^2}{\sum(y_{actual} - y_{mean})^2}$$

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 [28]. The corresponding VIF values of the five descriptors are presented in Table 2. As can be seen from this table, all the variables have VIF values of less than five, indicating that the obtained model has statistical significance, and the descriptors were found to be reasonably orthogonal.

Table 7. Equation 1 Euclidean based applicability domain (AD)

Compound No.	Distance score	Mean distance	Normalized mean distance
2	289.95	9.664999	0.53059
3	288.3358	9.611193	0.505584
4	288.9239	9.630797	0.514695
5	287.6763	9.589211	0.495368
6	287.7724	9.592412	0.496856
9	288.4801	9.616002	0.507819
10	289.0854	9.636179	0.517196
12	288.1643	9.605476	0.502927
13	288.0949	9.603163	0.501852
14	288.1314	9.60438	0.502418
15	288.1231	9.604103	0.502289
17	272.4481	9.081605	0.25946
18	320.251	10.67503	1
19	272.3741	9.079136	0.258312
20	272.3721	9.07907	0.258282
21	256.6197	8.553991	0.014253
24	259.208	8.640268	0.05435
25	262.6979	8.756595	0.108413
27	256.8328	8.561094	0.017554
28	291.715	9.723833	0.557933
29	258.2923	8.609742	0.040163
30	258.0338	8.601128	0.03616
31	258.0615	8.602049	0.036588
33	281.321	9.377368	0.396915
35	257.5266	8.584221	0.028302
36	259.0893	8.63631	0.05251
38	283.4106	9.44702	0.429285
40	283.4058	9.44686	0.429211
42	256.9077	8.563592	0.018715
43	255.6997	8.523323	0
Test set			
Compound No.	**Distance score**	**Mean distance**	**Normalized mean distance**
7	289.1057	9.636858	0.517512
8	289.9612	9.665373	0.530764
11	287.7555	9.59185	0.496594
16	288.3279	9.61093	0.505462
22	258.2999	8.609997	0.040282
23	257.4336	8.58112	0.026861
26	257.0176	8.567254	0.020417
32	257.6564	8.588546	0.030312
34	257.8537	8.595122	0.033369
37	283.4515	9.448382	0.429918
44	255.7337	8.524457	0.000527

Table 8. Specification of entered descriptors in genetic function approximation

Descriptors	Definition	VIF*	p-value**	t-value***
SsF	Sum of atom-type E-State: -F	1.885	2.86E-06	6.0708
minHBint3	Minimum E-State descriptors of strength for potential hydrogen bonds of path length 3	2.515	6.13E-12	-12.4178
minHdsCH	Minimum atom-type H E-State: =CH-	1.078	3.67E-11	-11.3875
FPSA-1	PPSA-1 / total molecular surface area	1.670	7.59E-10	9.7793
RHSA	THSA / total molecular surface area	1.010	1.34E-06	-6.3823

**Variation inflation factor; **p-value was introduced for compare under the confidence level 95%; ***t-test was introduced for compare under the confidence level 95%*

In order to assess the robustness of the model, the Y-randomization test was applied in this study [25]. Y-randomization test confirms whether the model is obtained by chance correlation, and is a true structure–activity relationship to validate the adequacy of the training set molecules. The steps followed during the randomization test are:

(I) repeatedly scrambling the activity data in the training set molecules,
(II) using the randomized data to generate QSAR equations, and
(III) Comparing the resulting scores with the score of the original QSAR equation generated with non-randomized data. If the activity prediction of the random model is comparable to that of the original equation, the set of observations is not sufficient to support the model.

The new QSAR models (after several repetitions) would be expected to have low R^2 and Q^2_{LOO} values (Table 3). If the opposite happens, then an acceptable QSAR model cannot be obtained for the specific modeling method and data. The results of Table 3 indicate that an acceptable model is obtained by GA–MLR method, and the model developed is statistically significant and robust.

3.3 Interpretation of the Selected Descriptors

The GFA model is useful in predicting the binding affinity of indole β -diketo, diketo acid and carboxamide derivatives. The brief descriptions of descriptors are shown in Table 2. To examine the relative importance as well as the contribution of each descriptor in the model, the statistical results for the selected descriptors in this model are given in Table 3. In this model, a Student’s t-test was performed at a confidence level of 95% to confirm the significance of each descriptor. All the P-values of the descriptors were less than 0.05, indicating that the selected descriptors were statistically significant at the 95% level. Moreover, the multi-colinearity of the descriptors was evaluated using the variation inflation factor (VIF). A VIF value larger than 10 indicates that a descriptor is highly correlated with one or more of the remaining independent variables. In this model, all the VIF values were less than 2.515, revealing that the descriptors were fairly independent of each other.

Table 9. R_{train}, R^2_{train} and $Q^2_{(LOO)train}$ values after several Y-randomization tests

Model	R_{train}	R^2_{train}	$Q^2_{(LOO)train}$
Original	0.989045	0.978211	0.963637
Random 1	0.384344	0.14772	-0.49419
Random 2	0.444433	0.197521	-0.32954
Random 3	0.47697	0.2275	-0.21534
Random 4	0.416674	0.173617	-0.32322
Random 5	0.382305	0.146157	-0.29027
Random 6	0.517191	0.267487	-0.11088
Random 7	0.382771	0.146514	-0.33458
Random 8	0.403669	0.162949	-0.34735
Random 9	0.382032	0.145948	-0.38038
Random 10	0.252255	0.063633	-0.54145
Random models parameters			
Average R_{train}:			0.4043
Average R^2_{train} :			0.1679
Average $Q^2_{(LOO)train}$:			-0.3367

Table 10. The descriptors relevance for the variables used in the model proposed

Variable	Abbreviation	Occurrences in population	Variable	Abbreviation	Occurrences in population
NP : pIC50	Y		HB : ETA_dEpsilon_D	X210	161
M : P-Area(75)	X13	553	HD : ETA_Shape_P	X212	119
N : P-Area(100)	X14	88	HM : ETA_Beta_ns_d	X221	182
O : P-Area(125)	X15	119	HN : ETA_BetaP_ns_d	X222	87
P : Acc.P-Area(75)	X16	91	HP : ETA_EtaP	X224	111
U : HBD	X21	135	IG : nAtomP	X241	115
AS : ATSm5	X45	116	IM : MDEC-14	X247	84
BP : C1SP3	X68	62	IQ : MDEC-33	X251	951
BS : SCH-6	X71	376	JD : nRing	X264	404
BT : SCH-7	X72	67	JE : n5Ring	X265	61
BV : VCH-6	X74	221	JV : WTPT-2	X282	168
CP : SP-6	X94	377	JZ : PPSA-1	X286	180
DB : nHBd	X106	83	KB : PPSA-3	X288	55
DU : ndssC	X125	108	KC : PNSA-1	X289	118
EV : SsF	X152	903	KE : PNSA-3	X291	307
EW : minHBd	X153	4031	KF : DPSA-1	X292	250
FA : minHBint3	X157	1561	KH : DPSA-3	X294	363
FB : minHBint4	X158	68	KI : FPSA-1	X295	279
FG : minHsOH	X163	337	KL : FNSA-3	X298	854
FI : minHdsCH	X165	1095	KQ : WNSA-3	X303	61
FM : minaaCH	X169	87	KT : RPCS	X306	101
FN : minaasC	X170	218	KU : RNCS	X307	1359
FR : minsOH	X174	99	KV : THSA	X308	270
FU : minsCl	X177	2533	KW : TPSA	X309	2112
			KX : RHSA	X310	1913

Because high Q^2 values appear to be a necessary but not sufficient condition for high predictive power, the predictiveness of the model was further evaluated using an internal validation set and external prediction test set. The robustness, predictiveness, and applicability of the MLR model were demonstrated by a high Q^2 value (Q^2 = 0.8138), internal predictive squared correlation coefficient (R^2 = 0.9782) (Table 4), and predictive squared correlation coefficient (R^2_{pred} = 0.8985) (Table 6). The indole β-diketo, diketo acid and carboxamide activities predicted by the GFA-MLR model are listed in Table 5, and Fig. 1 shows the plot of experimental activities versus the predicted activities. The model in equation 1 indicates that inhibitory activity of compounds against HIV-1 IN depends on electro topological state atom type and CPSA descriptors. Table 10 above shows the descriptor relevance for the variables used in the proposed model.

The QSAR study revealed that minHBint3, minHdsCH, and RHSA descriptors have negative contribution to the integrase activity while SsF and FPSA-1 have positive contribution to the activity. The E-State value for a given non hydrogen atom in a molecule is given by its intrinsic state plus the sum of the perturbations on that atom by all the other atoms in the molecule [29]. If one looks into drug like molecules and uses C, N, O, S and the halogens as the main building blocks, Kier and Hall use 35 atom types to calculate E-states. The symbols associated with the atom types are *s* for single bond, *d* for double, *t* for triple and *a* for aromatic. The attraction and proposed advantage of E-states over simple counts of the equivalent atom types is that E-states values for each atom in a given molecule 'reflect' the steric and electronic effects of the surrounding atoms and as such, could be best described as information rich atomic descriptors. Therefore, for example, if two different molecules have one phenol group, simple phenolic OH counts would not differentiate between two different substitution patterns that the phenolic group might have, while E-states would [30]. Electro topological state atom type descriptor SsF, represents Sum of atom-type E-State: -F; This descriptor contributes positively which indicates that inhibitory activity of indole β-diketo, diketo acid and carboxamide derivatives will increases with sum of atom-type E-state: -F. Negative contribution of the minHBint3 (Minimum E-State descriptors of strength for potential Hydrogen Bonds of path length 3) and minHdsCH (Minimum atom-type H E-State: =CH-) indicates that inhibitory activity of indole β-diketo, diketo acid and carboxamide derivatives will increases with decrease of the molecular descriptors.

The charged partial surface area, or CPSA descriptors were originally designed for use in structure-physical relationship studies to capture information about the features of molecules responsible for polar intermolecular interactions. Since their development, they have found applications in a broad variety of both structure-property and structure-activity relationship studies. The CPSA descriptors have been found to be practically useful in the study of acute aquatic toxicity where they appear to provide an alternative to LUMO energy level measures for describing global and local electrophilicity in cases of non-covalent molecular interactions [31]. CPSAs, were found to be necessary to provide separation between reactivity patterns for agonists and antagonists, all having high binding affinity to estrogen receptor [32]. CPSA descriptors FPSA-1, define as Partial positive surface area -- sum of surface area on positive parts of molecule/ total molecular surface area contribute positively which indicates that inhibitory activity of indole β-diketo, diketo acid and carboxamide derivatives will increases with FPSA-1. RHSA, represent Sum of solvent accessible surface areas of atoms with absolute value of partial charges less than 0.2/ total molecular surface area; this descriptors contribute negatively indicates that increase in number of aromatic bonds in the molecule are not conducive to the integrase activity of indole β -diketo, diketo acid and carboxamide derivatives.

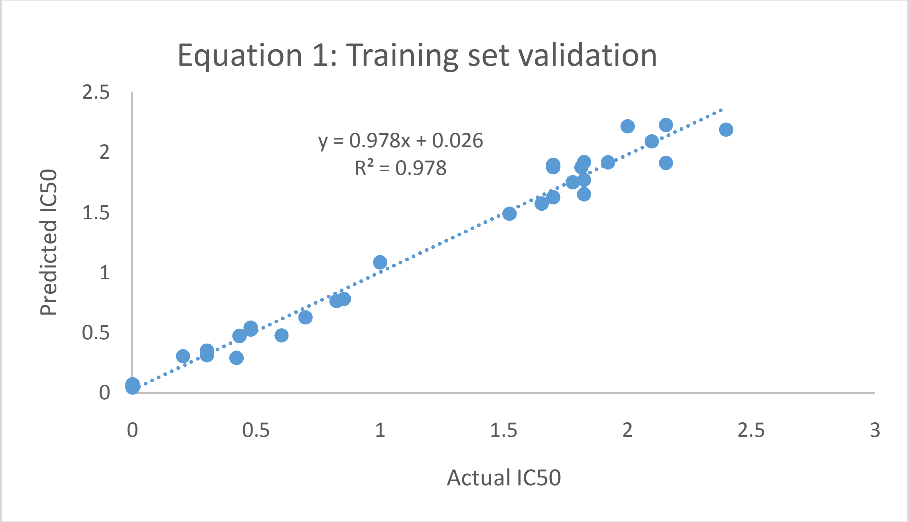

Fig. 1. Scatter plot (Training set) of actual activities vs. the predicted activities

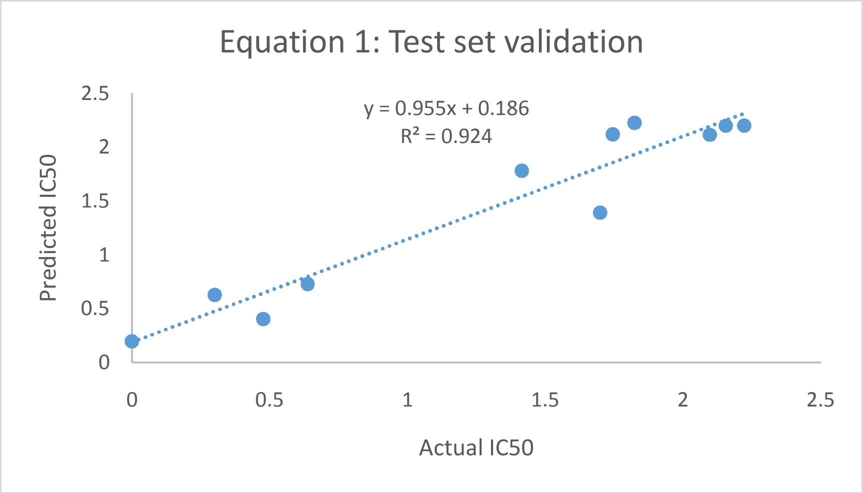

Fig. 2. Scatter plot (Test set) of Actual activities vs. the predicted activities

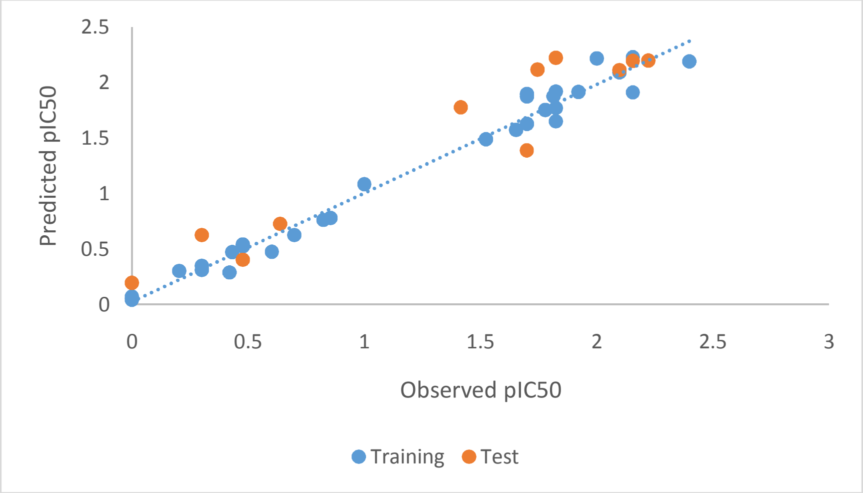

Fig. 3. The predicted pIC50 against the observed values for the training and test set

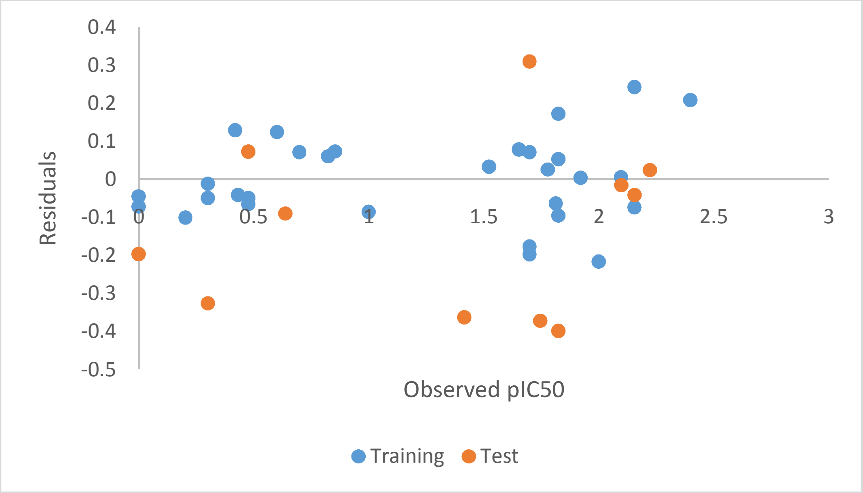

Fig. 4. The residuals vs. observed pIC50 values for the training and test sets

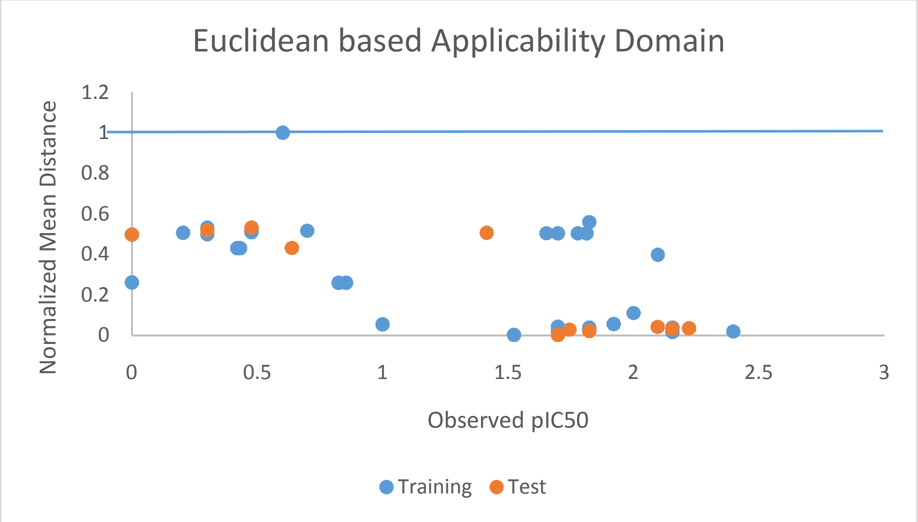

Fig. 5. plot of normalized mean distance vs. observed pIC50 for training and test set

4. CONCLUSION

The aim of the present work was to develop a QSAR study and predict the anti-HIV activities of indole β-diketo, diketo acid and carboxamide derivatives. Spartan'14 and PaDEl-Descriptor Software and selected by Genetic Function Approximation. The built GFA model was judged systematically (internal and external validations), and all the validations indicate that the QSAR model we built is robust and satisfactory. Selection of five variables showed that Sum of atom-type E-State: -F, Minimum E-State descriptors of strength for potential Hydrogen Bonds of path length 3, Minimum atom-type H E-State: =CH-, PPSA-1 / total molecular surface area, and THSA / total molecular surface area of the molecule play a main role in the anti-HIV activity of the compounds.

CONSENT

It is not applicable.

ETHICAL APPROVAL

It is not applicable.

ACKNOWLEDGEMENTS

EIE wishes to thank Emmanuel Oduma Edache and Ameh Joshua for its financial support.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. Yoder KE, Bushman FD. Repair of gaps in retroviral DNA integration intermediates. J Virol. 2000;74:11191–11200.
2. Brin E, Yi J, Skalka AM, Leis J. Modeling the late steps in HIV-1 retroviral integrase catalyzed DNA integration. J Biol Chem. 2000;275:39287–39295.
3. Bujacz G, Alexandratos J, Wlodawer A, Merkel G, Andrake M, Katz RA, Skalka AM. Binding of different divalent cations to the active site of Avian Sarcoma virus integrase and their effects on enzymatic activity. J Biol Chem. 1997;272:18161–18168.
4. U.S. Food and Drug Administration. Available:http://www.fda.gov
5. Hicks C, Gulick RM. Raltegravir: The first HIV type 1 integrase inhibitor. Clin Infect Dis. 2009;48:931–939.
6. Mehellou Y, De Clercq E. Twenty-Six years of Anti-HIV drug discovery: Where do we stand and where do we go? J. Med. Chem. 2010;53:521-538.
7. Gupta P, Garg P, Roy N. Identification of novel HIV-1 integrase inhibitors using shape-based screening, QSAR, and

docking approach. Chem Biol Drug Des. 2012;79:835–849

8. Agrawal VK, Singh J, Mishra KC, Padmakar VK, Jaliwala YA, QSAR studies on the use of 5,6-dihydro-2-pyrones as HIV-1 protease inhibitors. ARKIVOC. 2006; 2:162-177.

9. Srivastav VK, Tiwari M. QSAR and docking studies of coumarin derivatives as potent HIV-1 integrase inhibitors. Arabian J. Chem; 2013.

10. Rastelli, G, Pacchioni, S, Sirawaraporn W, Sirawaraporn, R, Parenti MD, Ferrari AM. Docking and database screening reveal new classes of *Plasmodium falciparum* dihydrofolate reductase inhibitors. J. Med. Chem. 2003;46:2834–2845.

11. Hill, D.L. The Biochemistry and Physiology of Tetrahymena. Academic Press, New York; 1972.

12. Manly CJ, Louise-May S, Hammer JD. The impact of informatics and computational chemistry on synthesis and screening. Drug Discovery Today. 2001;6:1101–1110.

13. Pourbasheer E, Riahi S, Ganjali MR, Norouzi P. QSAR study of C allosteric binding site of HCV NS5B polymerase inhibitors by support vector machine. Mol. Divers. 2011;15:645–653.

14. Burger A, Abraham DJ. Burger’s Medicinal Chemistry and Drug Discovery. Wiley, Hoboken, NJ; 2003.

15. Sechi M. Design and synthesis of novel indole a-diketo acid derivatives as HIV-1 integrase inhibitors. J. Med. Chem. 2004; 47:5298–5310.

16. Arodola OA, Radha CD, Mohmoud ESS. QSAR study on diketo acid and carboxamide derivatives as potent HIV-1 integrase inhibitor. Letters in Drug design & Discovery. 2014;11(5).

17. Beheshti A, Pourbasheer E, Nekoei M, Vahdani S. QSAR modeling of antimalarial activity of urea derivatives using genetic algorithm-multiple linear regressions, Journal of Saudi Chemical Society; 2012.

18. Becke AD, Density-functional thermochemistry 3. The role of exact exchange. J. Chem. Phys. 1993;98:5648.

19. Lee C, Yang W, Parr RG. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. Phys Rev. 1988;37:785–789.

20. Yap CW. PaDel –Descriptor: An open source software to calculate molecular descriptors and fingerprints. J. Comput. Chem. 2011;32(7):1466-1474.

21. Chtita S, Ghamali M, Larif M, Adad A, Rachid H, Bouachrine M and Lakhlifi T. Prediction of biological activity of imidazo{1,2-a}pyrazine. International Journal of Innovative Research in Science, Engineering and Technology. 2013;2(12): 7951-7962.

22. Adad A, Hmamouchi R, Taghki AI, Abdellaoui A, Bouachrine M, Lakhlifi T. Atmospheric half-lives of persistent organic pollutants (POPs) study combining DFT and QSPR results. J. of Chemical and Pharmaceutical Research. 2013;5(7):28-41.

23. Ahmad S, Gromiha MM. Design and training of a neural network for predicting the solvent accessibility of proteins. J. Comput. Chem. 2003;24:1313–1320.

24. Holland JH. Adaptation in natural and artificial systems. The University of Michigan Press, 1975; Ann Arbor, MI. (2nd ed., (1992) Boston, MA: MIT Press).

25. Tropsha A, Gramatica P, Gombar V. The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR Models. QSAR & Combinatorial Science. 2003;22:69–77.

26. Melagraki G, Afantitis A, Sarimveis H, Koutentis PA, Markopolus J, Igglesi-Markopoulou OJ. Comput. Aided Mol. Des. 2007;21:251-267

27. Dimitrov S, Dimitrova G, Pavlov T, Dimitrova N, Patlewicz G, Niemela J, Mekenyan OA. Stepwise approach for defining the applicability domain of SAR and QSAR models. Journal of Chemical Information and Modeling. 2005;45:839–49.

28. Jaiswal M, Khadikar PV, Scozzafava A, Supuran CT. Carbonic anhydrase inhibitors: the first QSAR study on inhibition of tumor-associated isoenzyme IX with aromatic and heterocyclic sulfonamides. Bioorg. Med. Chem. Lett. 2004;14:3283–3290.

29. Kier LB, Hall LH. Molecular Structure Descriptors: The Electrotopological State. Academic Press: New York; 1999.

30. Darko B. Performance of Kier-Hall E-state descriptors in Quantitative Structure

Activity Relationship (QSAR) studies of multifunctional molecules. Molecules. 2004;9:1004-1009.

31. Stanton DT, Jurs PC. Development and use of charged partial surface area structural descriptors in computer assisted quantitative structure property relationship studies. Analytical Chemistry. 1990; 62:2323-2329.

32. Stanton DT, Dimitrov S, Grancharov V, Mekenyan OG. Charged partial surface area (CPSA) descriptors QSAR applications. SAR QSAR Environ Res. 2002;13(2):341-51.

Peer-review history:
The peer review history for this paper can be accessed here:
http://sciencedomain.org/review-history/12116