British Journal of Applied Science & Technology
13(1): 1-20, 2016, Article no.BJAST.21791
ISSN: 2231-0843, NLM ID: 101664541

SCIENCEDOMAIN *international*
www.sciencedomain.org

QSAR Study of Series of HEPT (1-[(2-hydroxyethoxy) methyl]-6-(phenylthio)thyio)thymine) Derivatives Using Genetic Function Approximation as Anti-HIV-1 Agents

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. Authors EIE and SEA 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/BJAST/2016/21791
Editor(s):
(1) Lesley Diack, School of Pharmacy and Life Sciences, Robert Gordon University, UK.
Reviewers:
(1) Aurea Regina Telles Pupulin, State University of Maringa, Brazil.
(2) Kadima Ntokamunda Justin leonard, University of Rwanda, Rwanda.
Complete Peer review History: http://sciencedomain.org/review-history/12156

Short Research Article

Received 3rd September 2015
Accepted 8th October 2015
Published 7th November 2015

ABSTRACT

Aims/Objectives: Studies were performed to correlate the biological activity of the HEPT (1-[(2-hydroxyethoxy) methyl]-6-(phenylthio)thyio)thymine) 107 sets of compound with the independent descriptor to know the structural requirement of the drug receptor binding interaction.
Methodology: Genetic function approximation algorithm (GFA) approach has been applied to linearly correlate dependent biological activities and independent descriptors. Genetic function approximation algorithm (GFA) has been widely used when the number of samples surpass the amount of descriptors.
Results: The result obtained from the regression analysis is good and statistical values of multiple correlation coefficient R^2= 0.9118 and standard error of estimation (Se) = 0.4449, Fisher ratio (F) = 65.1139, Q^2_{LOO} = 0.8830 and Q^2_{L5O} = 0.8816 proves that the obtained mathematical model from the 107 sets of HEPT derivatives is the best.

**Corresponding author: E-mail: edacheson2004@gmail.com, edacheson2004@yahoo.com;*

Conclusion: The role of RotBtFrac, VPC-5, SP-4 and SHaaCH is important to reduce the required concentration of the drug and so as LogP and Weta2.volume also play vital role in this concern.

Keywords: Anti HIV; biological activity; drug design; NNRTIs; QSAR; regression analysis.

1. INTRODUCTION

Acquired immunodeficiency syndrome (AIDS), initiated by infection from the human immunodeficiency virus type 1 (HIV-1), remains a severe international health problem. Even if there is no definite cure for HIV infection, a number of drugs slow or halt disease progression [1]. After years of hard work, a number of inhibitors of reverse transcriptase (RT), integrase (IN) and protease (PR) are discovered and introduced in clinical practice [2,3]. Unluckily, all the mono therapies using either RT, IN or PR inhibitors have failed owing to the rapid emergence of HIV-resistant strains, and the long-term goal of eradicating the virus from infected cells is still unattained [4]. However, the use of both RT, IN and PR inhibitors have resulted in significant increases in disease-free survival [1,5]. This numerous outbreak is more effective, blocking two different steps of the virus replication cycle and causing a delay in the emergence of resistant strains [6,7,8]. Therefore, it is evident that the development of new inhibitors targeted toward other viral proteins is of paramount importance [9].

Two main categories of HIV RT inhibitors have been discovered to date. The first category of inhibitors is nucleoside analogues and the second category of inhibitors is non-nucleoside analogues [10]. Nucleoside analogues cause chain termination when they are incorporated within newly synthesized DNA. Non-nucleoside inhibitors block RT binding to a pocket adjacent to the catalytic site of the enzyme and thereby interrupt the conformation of several amino acids essential for proper RT function [11]. Reverse transcriptase (RT) plays a central role in the replication of HIV because of its specificity and its low cytotoxicity [12]. A number of RT-inhibitors active against both HIV-1 and HIV-2 RT or only against HIV-1 RT have been discussed in the literature [13,14,15]. Structure-Activity Relationships (SARs) and Quantitative Structure Activity Relationships (QSARs), jointly referred to as (Q)SARs, are theoretical models that relate the structure of chemicals to their biological activities. (Q)SARs are used to predict the physicochemical, biological and fate properties of molecules from knowledge of chemical structure [16]. In a QSAR study, generally, the quality of a model is expressed by its fitting ability and prediction ability, and of these the prediction ability is the most important. The QSAR studies enable the scientists to establish reliable quantitative relationship to derive the QSAR model and predict the activity of potent, novel and non-toxic molecules prior to their synthesis. These studies reduce the trial and error element in the design of compounds by establishing mathematical relationships between biological activities of interest and measurable or computable parameters such as physicochemical, electronic, topological, or thermodynamic. The main success of the QSAR method is the possibility to estimate the characteristics of new chemical compounds without the need to synthesize and test them. This analysis represents an attempt to relate structural descriptors of compounds with their physicochemical properties and biological activities. This is widely used for the prediction of physicochemical properties in the chemical, pharmaceutical, and environmental spheres. This method included data collection, molecular descriptor selection, correlation model development, and finally mode evaluation. QSAR are certainly a major factor in contemporary drug design. Therefore, it is quite clear why a large number of users of QSAR are located in industrial research units [10].

There is a series of statistical model studies that are used to develop a QSAR model, which include multiple linear regression (MLR), principle component analysis (PCA), partial least square (PLS), genetic function algorithm (GFA).

A QSAR is a quantitative relationship between a biological activity and one or more molecular descriptors that are used to predict the activity [17]. A molecular descriptor is a structural or physicochemical property of a molecule, or part of a molecule, which specifies a particular characteristic of the molecule and is used as an independent variable in a QSAR [18].

QSAR analyses of HIV-1 reverse transcriptase inhibitors [10], HIV-1 protease inhibitors [19], HIV- 1 integrase inhibitors [16, 20] and gp 120

envelope glycoprotein [21] were reported. The present group of authors has developed a few quantitative structure-activity relationship models to predict anti-HIV activity of different group of compounds [22,23,24]. Although several QSAR studies on HIV reverse transcriptase, protease and integrase inhibitors have been reported [4,19,25-32] using MLR, PLS, PCA, GFA and ANN, the QSAR study on HIV-1 reverse transcriptase using the GFA method has been lacking in literature. Such a kind about the GFA method might provide a new starting point for the design of novel inhibitors against HIV-1. The main purpose of this work is to find out how accurate the QSAR analysis predicted the activities of compounds that were already synthesized in comparison to their experimental biological activities.

2. MATERIALS AND METHODS

2.1 Data Set

The HEPT derivatives selected with their activities [33] are listed in Table 1 and the parent structure of the HEPT derivatives is given in the Fig. 1. The molecular structures of the compounds in the selected series were sketched using ChemBioDraw ultra 12.0 module of CambridgeSoft 2010 molecular modeling software. The sketched structures were then transferred to Spartan'14 version 1.1.2 for complete geometry optimization with the semi-empirical Parameterized Model 3 (PM3) method was performed. The geometries of generated structures were pre-optimized using MM2 force field as implemented in the PaDEL-Descriptor version 2.18 software.

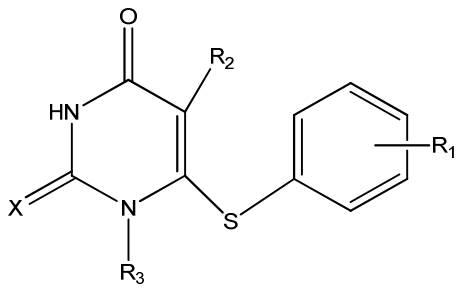

Fig. 1. Structure of training and test set

A QSAR model was therefore, used to analyze some potential (1-[(2-hydroxyethoxy) methyl]-6-(phenylthio)thyio)thymine) Derivatives HIV-1 reverse transcriptase inhibitors. The list of the structures of 107 inhibitors employed in this study and their experimental inhibitory concentration (EC_{50}) effective against HIV-1 RT enzyme was taken from literature [33] (Table 1). It was observed that in each case 500 crossovers and smoothing factor d = 0.5 resulted in optimum internal and external predictivity (Table 2).

Table 1. The HEPT derivatives selected with their activities

Cpd no	R1	R2	R3	X	Obs. pEC50
1[b]	3-CN	Me	$CH_2OCH_2CH_2OH$	O	5.000
2[a]	3-COMe	Me	$CH_2OCH_2CH_2OH$	O	5.140
3[a]	3-COOMe	Me	$CH_2OCH_2CH_2OH$	O	5.100
4[a]	3,5-Cl_2	Me	$CH_2OCH_2CH_2OH$	O	5.890
5[a]	3,5-Me_2	Me	$CH_2OCH_2CH_2OH$	O	6.590
6[a]	3-OMe	Me	$CH_2OCH_2CH_2OH$	O	4.660
7[a]	3-OH	Me	$CH_2OCH_2CH_2OH$	O	4.090
8[a]	3-NO_2	Me	$CH_2OCH_2CH_2OH$	O	4.470
9[b]	3-I	Me	$CH_2OCH_2CH_2OH$	O	5.000
10[a]	3-Br	Me	$CH_2OCH_2CH_2OH$	O	5.240
11[a]	3-Cl	Me	$CH_2OCH_2CH_2OH$	O	4.890
12[b]	3-F	Me	$CH_2OCH_2CH_2OH$	O	5.480
13[a]	3-CF_3	Me	$CH_2OCH_2CH_2OH$	O	4.350
14[b]	3-t-Bu	Me	$CH_2OCH_2CH_2OH$	O	4.920
15[a]	3-Et	Me	$CH_2OCH_2CH_2OH$	O	5.570
16[a]	3-Me	Me	$CH_2OCH_2CH_2OH$	O	5.590
17[a]	2-OMe	Me	$CH_2OCH_2CH_2OH$	O	4.720
18[a]	2-NO_2	Me	$CH_2OCH_2CH_2OH$	O	3.850

Cpd no	R1	R2	R3	X	Obs. pEC50
19[a]	2-Me	Me	$CH_2OCH_2CH_2OH$	O	4.150
20[a]	H	Et	$CH_2OCH_2CH_2OH$	O	6.920
21[a]	H	i-Pr	$CH_2OCH_2CH_2OH$	O	7.200
22[a]	3,5-Me_2	Et	$CH_2OCH_2CH_2OH$	O	7.890
23[b]	3,5-Me_2	i-Pr	$CH_2OCH_2CH_2OH$	O	8.570
24[b]	3,5-Cl_2	Et	$CH_2OCH_2CH_2OH$	O	7.850
25[a]	H	Me	$CH_2OCH_2CH_2OH$	O	5.150
26[a]	H	I	$CH_2OCH_2CH_2OH$	O	5.440
27[a]	H	$CH=CPH_2$	$CH_2OCH_2CH_2OH$	O	6.070
28[a]	4-F	Me	$CH_2OCH_2CH_2OH$	O	3.600
29[a]	4-Cl	Me	$CH_2OCH_2CH_2OH$	O	3.600
30[a]	4-OH	Me	$CH_2OCH_2CH_2OH$	O	3.560
31[a]	3-$CONH_2$	Me	$CH_2OCH_2CH_2OH$	O	3.510
32[a]	H	COOMe	$CH_2OCH_2CH_2OH$	O	5.180
33[a]	H	CONHPh	$CH_2OCH_2CH_2OH$	O	4.740
34[b]	H	SPh	$CH_2OCH_2CH_2OH$	O	4.840
35[a]	H	CCH	$CH_2OCH_2CH_2OH$	O	4.740
36[a]	H	CCPh	$CH_2OCH_2CH_2OH$	O	5.470
37[a]	H	$COCHMe_2$	$CH_2OCH_2CH_2OH$	O	4.920
38[a]	H	COPh	$CH_2OCH_2CH_2OH$	O	4.890
39[a]	H	CCMe	$CH_2OCH_2CH_2OH$	O	4.720
40[b]	H	F	$CH_2OCH_2CH_2OH$	O	4.000
41[a]	H	Cl	$CH_2OCH_2CH_2OH$	O	4.520
42[b]	H	Br	$CH_2OCH_2CH_2OH$	O	4.700
43[a]	2-Cl	Me	$CH_2OCH_2CH_2OH$	O	3.890
44[a]	3-CH_2OH	Me	$CH_2OCH_2CH_2OH$	O	3.530
45[a]	4-NO_2	Me	$CH_2OCH_2CH_2OH$	O	3.720
46[a]	4-CN	Me	$CH_2OCH_2CH_2OH$	O	3.600
47[a]	4-OMe	Me	$CH_2OCH_2CH_2OH$	O	3.600
48[a]	4-COMe	Me	$CH_2OCH_2CH_2OH$	O	3.960
49[a]	4-COOH	Me	$CH_2OCH_2CH_2OH$	O	3.450
50[a]	3-NH_2	Me	$CH_2OCH_2CH_2OH$	O	3.600
51[a]	H	Pr	$CH_2OCH_2CH_2OH$	O	5.470
52[a]	4-Me	Me	$CH_2OCH_2CH_2OH$	O	3.660
53[a]	H	$CH=CH_2$	$CH_2OCH_2CH_2OH$	O	5.690
54a	H	CH=CHPh	$CH_2OCH_2CH_2OH$	O	5.220
55a	H	CH_2Ph	$CH_2OCH_2CH_2OH$	O	4.370
56a	H	Me	$CH_2OCH_2CH_2OAc$	O	5.170
57[b]	H	Et	CH_2OCH_2Me	O	7.720
58[a]	H	Et	CH_2CH_2Ph	O	8.230
59[b]	3,5-Cl_2	Et	CH_2CH_2Me	O	8.130
60[a]	H	Me	$CH_2OCH_2CH_2OC_5H_{11}$	O	4.460
61[b]	H	Me	$CH_2OCH_2CH_2OCH_2Ph$	O	4.700
62[a]	H	Me	H	O	3.600
63[a]	H	Me	Me	O	3.820
64[b]	H	c-Pr	CH_2OCH_2Me	O	7.000
65[a]	H	Et	CH_2O-i-Pr	O	6.470
66[b]	H	Et	CH_2O-c-Hex	O	5.400
67[b]	H	Et	CH_2OCH_2-c-Hex	O	6.350
68[b]	H	Et	$CH_2OCH_2CH_2Ph$	O	7.020
69[b]	H	Me	CH_2OMe	O	5.680
70[b]	H	Me	CH_2OBu	O	5.330

Cpd no	R1	R2	R3	X	Obs. pEC50
71[a]	H	Me	Et	O	5.660
72[b]	H	Me	Bu	O	5.920
73[a]	H	i-Pr	CH_2OCH_2Me	O	7.990
74[a]	H	i-Pr	CH_2OCH_2Ph	O	8.510
75[a]	3,5-Me_2	Et	CH_2OCH_2Ph	O	8.550
76[b]	3,5-Me_2	Et	CH_2OCH_2Me	O	8.240
77[b]	H	Me	$CH_2OCH_2CH_2OMe$	O	5.060
78[b]	H	Me	$CH_2OCH_2CH_2OCOPh$	O	5.120
79[b]	H	Me	CH_2OCH_2Me	O	6.480
80[a]	H	Me	$CH_2OCH_2CH_2Cl$	O	5.820
81[a]	H	Me	$CH_2OCH_2CH_2N_3$	O	5.240
82[a]	H	Me	$CH_2OCH_2CH_2F$	O	5.960
83[b]	H	Me	$CH_2OCH_2CH_2Me$	O	5.480
84[a]	H	Me	CH_2OCH_2Ph	O	7.060
85[a]	H	Et	CH_2OCH_2Me	S	7.580
86[b]	H	i-Pr	CH_2OCH_2Me	S	7.890
87[a]	H	i-Pr	CH_2OCH_2Ph	S	8.140
88[a]	3,5-Cl_2	Et	CH_2OCH_2Me	S	7.890
89[b]	H	Et	CH_2O-i-Pr	S	6.660
90[a]	H	Et	CH_2O-c-Hex	S	5.790
91[a]	H	Et	CH_2OCH_2-c-Hex	S	6.450
92[a]	H	Et	$CH_2OCH_2C_6H_4$(4-Me)	S	7.110
93[a]	H	Et	$CH_2OCH_2C_6H_4$(4-Cl)	S	7.920
94[a]	H	Et	$CH_2OCH_2CH_2Ph$	S	7.040
95[a]	H	c-Pr	CH_2OCH_2Me	S	7.020
96[b]	3,5-Me_2	Me	$CH_2OCH_2CH_2OH$	S	6.660
97[b]	H	Pr	$CH_2OCH_2CH_2OH$	S	5.000
98[b]	3,5-Me_2	i-Pr	$CH_2OCH_2CH_2OH$	S	8.300
99[a]	H	Et	CH_2OCH_2Ph	S	8.090
100[a]	3,5-Me	Et	CH_2OCH_2Ph	S	8.140
101[b]	3,5-Me_2	Et	CH_2OCH_2Me	S	8.300
102[a]	H	Et	$CH_2OCH_2CH_2OH$	S	6.960
103[a]	H	i-Pr	$CH_2OCH_2CH_2OH$	S	7.230
104[b]	3,5-Me_2	Et	$CH_2OCH_2CH_2OH$	S	8.110
105[b]	3,5-Cl_2	Et	$CH_2OCH_2CH_2OH$	S	7.370
106[b]	H	Me	$CH_2OCH_2CH_2OH$	S	6.010
107[a]	H	$CH_2CH=CH_2$	$CH_2OCH_2CH_2OH$	O	5.600

[a]Training set; [b]Test set

2.2 Calculation of the Parameters

All the physicochemical properties viz. S (Entropy), PSA (Polar Surface Area), E(aq) (Aqueous Energy), Acc. Area (Accessible Area), LogP (Partition Coefficient), HBD count (Hydrogen Bond Donar) etc. were calculated by Spartan'14 version 1.1.2 software (Wavefunction/spartan14v1.1.2). All other descriptors were calculated by PaDEL-Descriptor version 2.18. A total of 238 descriptors were calculated using the fore mentioned molecular modeling package. A list of the descriptors used are summarized in Table 5.

2.3 The GFA Approach

In this study, we define the application of QSAR models based on GFA approach. GFA is an experimental search method used for finding optimal solutions to a problem where the possible solution space is too large to be systematically computed. The GFA approach has a number of significant benefits, which comprise: ability to build multiple models rather than a single model, as do most other statistical methods, it produces a population of models (e.g., 100). The range of variation this population gives added information on the quality of fit and

importance of descriptors (Table 7). For example, the frequency of use of a particular descriptor in the population of equations (Table 8) may indicate how relevant the descriptor is to the prediction of activity [34]; automatic selection of features to be used in its basic functions and to determine the suitable number of basic functions to be used by testing full-size models rather than incrementally building them; reliable discovery of combinations of basic functions that take advantage of correlations between features; ability to include the lack of fit (LOF) error measure developed by Friedman [35] that resists over fitting and allows user control over the smoothness of fit (in this case, 0.5); use of larger variety of basic functions in building of its models, preferred model length and useful partitions of the data set, automatic removal of outliers and finally, provision of additional information not existing from other statistical standard regression analysis. The GFA has been applied to three published data sets to demonstrate it is an effective tool for doing both QSAR and QSPR [36-40].

Table 2. Summary of GFA analysis

Analysis type	Genetic function approximation
Response column	ID : pC50
Number of rows in model	74
Population	5000
Maximum generations	500
Initial terms per equation	10
Maximum equation length	10
Constant equation length	Yes
Number of top models returned	5
Scoring function	Friedman LOF
Scaled LOF smoothness parameter	0.50000000
Mutation probability	0.10000000
Linear splines	No
Quadratic splines	No
Random number seed	9999
Minimum prediction fraction for term inclusion	1.000000e-004
Number of variables requested for plot	5

3. MODEL VALIDATION

The final model was systematically validated using a set of measures suggested in the literature [17,41-43]. The statistical parameters listed in (Tables 3, 5 and 9) were used to evaluate the quality of the model. For the internal quality, the recommended limits are $R^2 > 0.6$ and $Q^2_{LOO} > 0.5$ [17,44]. The SEE, RMSECV and SDEP should be lower as possible. The F-value and the Q value [45,46] should be higher.

The robustness of the optimized model was examined by leave-N-out cross-validation procedure. The average value of each Q^2_{LNO} (leave-many-out cross-validation) is expected to be close to Q^2_{LOO} (leave-one-out cross-validation) with standard deviation close to zero.

The parameter R^2_{pred} was used as a measure of the predictive power of the QSAR model. For this work, it was used the recommended limited of $R^2_{pred} > 0.6$ [17]. However, this is not a sufficient condition to guarantee that the model is really predictive. It is also recommended to check:

1) The slopes 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 $0.85 \leq K \leq 1.15$ or $0.85 \leq K' \leq 1.15$;
2) The absolute value of the difference between the coefficients of multiple determination, R^2_0 and R'^2_0 smaller than 0.3 [41];
3) r^2_m (overall) and R^2_p are ≥ 0.5 (or at least near 0.5) [42].

4. RESULTS AND DISCUSSION

Dissimilar QSAR equations were produced using the GFA algorithm in Material Studio V7.0 for a series HEPT (1-[(2-hydroxyethoxy) methyl]-6-(phenylthio)thyio)thymine) anti-HIV derivatives. A total of 107 compounds (Table 1) were used for QSAR model generation. It is essential to assess the predictive power of models by using a test set of compounds. This was achieved by setting aside 33 compounds as a test set such that it represented the various functional groups included in the training set and had a regularly distributed biological data. The mean of the biological activity of the training and test set was 5.4715 and 6.3964, respectively.

The selection of the best model was based on the value of correlation coefficient (*R*), the squared correlation coefficient (R^2), the *F*–test (Fischer's value) for statistical significance F, the standard error of estimation (*Se*), lack of fit (LOF) and the quality of fit (*Q*). The squared correlation

coefficient (or coefficient of multiple determination) 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. The correlation coefficient values closer to 1.0 represent the better fit of the regression. 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. Standard deviation is measured by the error mean square, which expresses the variation of the residuals or the variation about the regression line. Thus, standard deviation is an absolute measure of quality of fit and should have a low value for the regression to be significant. The positive value of quality factor (*Q*) for this QSAR model suggest its high predictive power and lack of over fitting.

Table 3. The 5 different equations derived from the QSAR model

Model	Equation	Definitions	R	R_a	SE	Q
1	Y = 0.338701317 * X13 - 0.875272810 * X14 - 0.986051104 * X49 + 1.631029432 * X51 + 2.175537287 * X53 - 1.800692936 * X56 - 0.007561581 * X61 + 0.228107397 * X72 + 9.340522738 * X159 + 2.260431080 * X217 + 2.871039151	X13 : M : LogP X14 : N : HBD count X49 : AW : SPC-6 X51 : AY : VPC-5 X53 : BA : SP-4 X56 : BD : VP-3 X61 : BI : ECCEN X72 : BT : SHaaCH X159 : FC : RotBtFrac X217 : HI : Weta2.volume	0.9549	0.8978	0.4449	2.1463
2	Y = 0.308084659 * X13 - 0.963731145 * X14 - 1.006817442 * X49 + 1.929380444 * X51 + 2.151143192 * X53 - 2.229128728 * X56 + 8.122243032 * X159 + 4.588403111 * X199 - 0.145656116 * X219 - 3.157365610 * X234 + 4.169923122	X13 : M : LogP X14 : N : HBD count X49 : AW : SPC-6 X51 : AY : VPC-5 X53 : BA : SP-4 X56 : BD : VP-3 X159 : FC : RotBtFrac X199 : GQ : Weta3.unity X219 : HK : WT.volume X234 : HZ : Weta1.polar	0.9546	0.8972	0.4463	2.1389
3	Y = 0.427039437 * X13 - 0.815890850 * X14 + 3.830810703 * X44 - 11.048338687 * X46 - 1.560143956 * X49 + 3.443199958 * X51 + 3.347430388 * X53 - 3.415283583 * X57 - 0.011762241 * X61 + 8.814113810 * X159 + 1.831259072	X13 : M : LogP X14 : N : HBD Count X44 : AR : SC-5 X46 : AT : VC-5 X49 : AW : SPC-6 X51 : AY : VPC-5 X53 : BA : SP-4 X57 : BE : VP-4 X61 : BI : ECCEN X159 : FC : RotBtFrac	0.9543	0.8966	0.4475	2.1325
4	Y = 0.343092405 * X13 - 0.870560056 * X14 - 0.914133690 * X49 + 1.648081562 * X51 + 2.311402825 * X53	X13 : M : LogP X14 : N : HBD count X49 : AW : SPC-6 X51 : AY : VPC-5 X53 : BA : SP-4	0.9543	0.8965	0.4478	2.131

Model	Equation	Definitions	R	R_a	SE	Q
	- 1.896511453 * X56 - 0.007232540 * X61 + 6.216826562 * X159 + 2.248251987 * X217 - 2.278087732 * X234 + 4.124022113	X56 : BD : VP-3 X61 : BI : ECCEN X159 : FC : RotBtFrac X217 : HI : Weta2.volume X234 : HZ : Weta1.polar				
5	Y = 0.411808582 * X13 - 0.906458365 * X14 - 7.135987171 * X46 - 1.133728780 * X49 + 3.003270876 * X51 + 3.064555264 * X53 - 3.337058906 * X57 - 0.010962312 * X61 + 8.767199680 * X159 + 2.072447104 * X217 + 1.119078704	X13 : M : LogP X14 : N : HBD count X46 : AT : VC-5 X49 : AW : SPC-6 X51 : AY : VPC-5 X53 : BA : SP-4 X57 : BE : VP-4 X61 : BI : ECCEN X159 : FC : RotBtFrac X217 : HI : Weta2.volume	0.9538	0.8955	0.4498	2.1205

Model 2, 3, 4 and 5 showed lower Q^2cv and high R^2_{pred} values than Model 1 which means that cross validated Q^2 ability of Model 1 was much better. The quality factor (Q) was performed to access the robustness and statistical confidence. Higher value of R^2, RMSEP, Q and F and lower value of Se, and RMSECV of Model 1 in comparison to Model 2, 3, 4 and 5 revealed that Model 1 was robust and promising. In the developed Model the value of coefficient of correlation was significantly high supporting reliability and goodness. Based on the above results Model 1 was considered as the best validation model for 107 inhibition activity. The accuracy of the Model 1 was ascertained by correlation coefficient (R = 0.9549), statistical significance more than 99% (against tabulated value F= 65.1139) and low standard error of estimate (0.4449).

The model shows that parameter LogP, VPC-5, SP-4, SHaaCH, RotBtFrac and Weta2.volume showed positive contribution. The regression model has small residuals that can be seen in (Table 3). LOO cross-validation analysis revealed that R^2-Q^2_{LOO} < 0.3 (0.9118-0.8890 = 0.0288). The robustness of the model was justified According to Golbraikh and Tropsha [44], the proposed QSAR model is predictive as it satisfies this conditions like R^2_{pred} > 0.5, R^2 > 0.6, r^2-r^2_o/r^2 < 0.1, r^2-r'^2o/r^2 < 0.1 and $0.85 \leq k \leq 1.15$ or $0.85 \leq k' \leq 1.15$, but this model satisfy the following criteria R^2_{pred}=0.8526 > 0.5, and R^2=0.9118 > 0.6 (Tables 3 and 5). So this QSAR model is predictive as it's satisfy this conditions reported by Golbraikh and Tropsha, [44]. The internal validation parameter of the model (Q^2_{cv} = 0.8830) was also good.

4.1 Y-Randomization Tests

The Y-randomization test is useful to verify the possibility that the explained and predicted variances by the obtained model may suffer from chance correlation [41]. The statistical significance of the relationship between the anti-HIV activity and chemical structure descriptors which was confirmed by randomization procedure. The test was done by:

(1) Frequently permuting (100 trail) the activity values of the data set,
(2) Using the permuted values to generate QSAR models and
(3) Relating the resulting scores with the score of the original QSAR model generated from non-randomized activity values.

If the original QSAR model is statistically significant, its score should be significantly better than those from permuted data. The R, R^2 and Q^2 value of the original model was much higher than any of the trials using permuted data. It can be observed in Table 9 that the results obtained for all randomized models are of bad quality when compared to the real model. Hereafter, model 1 is statistically significant and robust.

Table 4. Observed pIC50 and GFA predicted pIC50 for training set

Actual values	Equation 1		Equation 2		Equation 3		Equation 4		Equation 5	
pC50	Predicted values 1	Residual values 1	Predicted values 2	Residual values 2	Predicted values 3	Residual values 3	Predicted values 4	Residual values 4	Predicted values 5	Residual values 5
5.14	4.24	0.9	4.26	0.88	4.52	0.62	4.15	0.99	4.31	0.83
5.1	4.37	0.73	4.7	0.4	4.34	0.76	4.54	0.56	4.44	0.66
5.89	6.1	-0.21	6.15	-0.3	5.79	0.1	6.39	-0.5	6.01	-0.12
6.59	5.76	0.83	6.34	0.25	6.04	0.55	6.05	0.54	5.86	0.73
4.66	4.79	-0.13	5.04	-0.4	4.74	-0.1	4.82	-0.2	4.91	-0.25
4.09	3.85	0.24	3.87	0.22	3.83	0.26	3.88	0.21	3.81	0.28
4.47	3.89	0.58	4.13	0.34	3.93	0.54	3.97	0.5	3.75	0.72
5.24	5.2	0.04	5.16	0.08	5.17	0.07	5.35	-0.1	5.13	0.11
4.89	5.22	-0.33	5.13	-0.2	5.05	-0.2	5.37	-0.5	5.24	-0.35
5.48	4.96	0.52	5.06	0.42	4.88	0.6	4.95	0.53	4.95	0.53
4.35	4.31	0.04	4.51	-0.2	4.85	-0.5	4.27	0.08	4.14	0.21
4.92	5.02	-0.1	5.13	-0.2	4.72	0.2	4.86	0.06	4.71	0.21
5.57	5.02	0.55	5.15	0.42	5.28	0.29	5.09	0.48	5.36	0.21
5.59	5.21	0.38	5.23	0.36	5.19	0.4	5.33	0.26	5.27	0.32
4.72	5.06	-0.34	4.88	-0.2	5.12	-0.4	4.92	-0.2	5.11	-0.39
3.85	4.02	-0.17	4.19	-0.3	4.08	-0.2	4.1	-0.2	4.01	-0.16
4.15	4.45	-0.3	4.41	-0.3	4.57	-0.4	4.58	-0.4	4.56	-0.41
6.92	6.35	0.57	6.16	0.76	6.32	0.6	6.3	0.62	6.25	0.67
7.2	7.06	0.14	6.84	0.36	6.84	0.36	7.05	0.15	6.83	0.37
7.89	7.42	0.47	7.75	0.14	7.54	0.35	7.69	0.2	7.5	0.39
5.15	4.89	0.26	4.76	0.39	4.77	0.38	4.83	0.32	4.78	0.37
5.44	5	0.44	4.99	0.45	5.06	0.38	4.77	0.67	5.28	0.16
6.07	6.1	-0.03	5.66	0.41	5.92	0.15	5.69	0.38	6.2	-0.13
3.6	4.19	-0.59	4.42	-0.8	4.01	-0.4	4.15	-0.5	4.15	-0.55
3.6	4.19	-0.59	4.02	-0.4	4.15	-0.5	4.28	-0.7	4.47	-0.87
3.56	3.13	0.43	3.03	0.53	2.96	0.6	3.04	0.52	3.07	0.49
3.51	3.26	0.25	3.4	0.11	3.37	0.14	3.42	0.09	3.19	0.32
5.18	5.19	-0.01	5.31	-0.1	5.16	0.02	5.31	-0.1	5.27	-0.09
4.74	4.57	0.17	4.35	0.39	4.65	0.09	4.54	0.2	4.6	0.14
4.84	5.27	-0.43	4.51	0.33	4.95	-0.1	5.01	-0.2	4.91	-0.07

Actual values	Equation 1		Equation 2		Equation 3		Equation 4		Equation 5	
pC50	Predicted values 1	Residual values 1	Predicted values 2	Residual values 2	Predicted values 3	Residual values 3	Predicted values 4	Residual values 4	Predicted values 5	Residual values 5
5.47	5.05	0.42	5.73	-0.3	5.56	-0.1	5.37	0.1	5.18	0.29
4.92	5	-0.08	4.75	0.17	4.86	0.06	5	-0.1	4.95	-0.03
4.89	5.5	-0.61	5.18	-0.3	5.46	-0.6	5.43	-0.5	5.59	-0.7
4.72	5.08	-0.36	5.23	-0.5	5.03	-0.3	5.3	-0.6	5.07	-0.35
4.7	4.79	-0.09	4.5	0.2	4.8	-0.1	4.59	0.11	4.91	-0.21
3.89	4.17	-0.28	4.1	-0.2	4.37	-0.5	4.04	-0.1	4.27	-0.38
3.53	4.03	-0.5	4.08	-0.6	4.01	-0.5	4.05	-0.5	4.09	-0.56
3.72	3.56	0.16	3.85	-0.1	3.94	-0.2	3.56	0.16	3.63	0.09
3.6	3.97	-0.37	4	-0.4	3.95	-0.4	4.18	-0.6	4.18	-0.58
3.6	4.27	-0.67	4.24	-0.6	4.3	-0.7	4.11	-0.5	4.41	-0.81
3.6	3.61	-0.01	3.7	-0.1	3.56	0.04	3.69	-0.1	3.54	0.06
5.47	6.07	-0.6	6.06	-0.6	5.95	-0.5	5.77	-0.3	5.83	-0.36
3.66	4.18	-0.52	4.39	-0.7	4.29	-0.6	4.04	-0.4	4.46	-0.8
5.69	5.87	-0.18	5.75	-0.1	5.94	-0.3	5.92	-0.2	5.89	-0.2
5.17	5.4	-0.23	4.99	0.18	5.34	-0.2	5.05	0.12	5.28	-0.11
7.72	7.25	0.47	6.97	0.75	7.29	0.43	7.18	0.54	7.33	0.39
8.13	8.19	-0.06	8.19	-0.1	8.17	-0	8.18	-0.1	8.07	0.06
4.46	4.5	-0.04	4.61	-0.2	4.78	-0.3	4.85	-0.4	4.63	-0.17
4.7	5.09	-0.39	5.15	-0.5	4.91	-0.2	5.03	-0.3	4.77	-0.07
3.6	3.87	-0.27	3.9	-0.3	4.03	-0.4	3.77	-0.2	3.68	-0.08
3.82	4.02	-0.2	3.95	-0.1	4.21	-0.4	3.91	-0.1	4.01	-0.19
6.35	6.12	0.23	6.06	0.29	5.86	0.49	6.41	-0.1	6.05	0.3
5.68	5.71	-0.03	5.49	0.19	5.41	0.27	5.59	0.09	5.66	0.02
5.66	5.8	-0.14	5.72	-0.1	5.69	-0	5.98	-0.3	5.79	-0.13
5.92	5.71	0.21	5.6	0.32	5.56	0.36	5.52	0.4	5.71	0.21
5.12	5.39	-0.27	4.82	0.3	4.85	0.27	5.29	-0.2	5.09	0.03
6.48	5.64	0.84	5.6	0.88	5.7	0.78	5.69	0.79	5.72	0.76
5.82	5.79	0.03	5.84	-0	5.78	0.04	5.83	-0	5.75	0.07
5.24	5.18	0.06	5.49	-0.2	5.13	0.11	5.27	-0	5.16	0.08
5.48	5.85	-0.37	5.78	-0.3	5.89	-0.4	5.78	-0.3	5.83	-0.35
8.14	8.12	0.02	8.51	-0.4	8.24	-0.1	8.02	0.12	8.02	0.12
7.89	8.69	-0.8	8.26	-0.4	8.82	-0.9	8.52	-0.6	8.8	-0.91

Actual values	Equation 1		Equation 2		Equation 3		Equation 4		Equation 5	
pC50	**Predicted values 1**	**Residual values 1**	**Predicted values 2**	**Residual values 2**	**Predicted values 3**	**Residual values 3**	**Predicted values 4**	**Residual values 4**	**Predicted values 5**	**Residual values 5**
5.79	6.22	-0.43	6.49	-0.7	6.33	-0.5	6.6	-0.8	6.28	-0.49
6.45	6.31	0.14	6.79	-0.3	6.32	0.13	6.5	-0.1	6.27	0.18
7.11	6.94	0.17	6.96	0.15	6.99	0.12	6.85	0.26	7.22	-0.11
7.92	7.05	0.87	7.07	0.85	6.85	1.07	7.04	0.88	7.33	0.59
7.04	7.22	-0.18	6.91	0.13	7.42	-0.4	6.93	0.11	7.2	-0.16
7.02	6.6	0.42	6.57	0.45	6.48	0.54	6.78	0.24	6.46	0.56
8.3	8.44	-0.14	8.37	-0.1	8.49	-0.2	8.51	-0.2	8.4	-0.1
8.09	7.4	0.69	7.98	0.11	7.78	0.31	7.33	0.76	7.44	0.65
8.14	8.58	-0.44	8.96	-0.8	8.9	-0.8	8.8	-0.7	8.76	-0.62
6.96	6.76	0.2	6.33	0.63	6.78	0.18	6.43	0.53	6.66	0.3
7.23	7.66	-0.43	7.33	-0.1	7.3	-0.1	7.5	-0.3	7.41	-0.18
5.6	6.14	-0.54	6.07	-0.5	5.78	-0.2	6	-0.4	6.03	-0.43
Test set										
5	4.34	0.66	4.27	0.73	4.37	0.63	4.27	0.73	4.51	0.49
5	5.69	-0.7	5.31	-0.3	5.39	-0.4	5.9	-0.9	5.56	-0.6
8.57	7.94	0.63	8.1	0.47	8.01	0.56	8.18	0.39	7.91	0.66
7.85	7.31	0.54	7.67	0.18	7.28	0.57	7.23	0.62	7.25	0.6
4.74	5.27	-0.5	5.48	-0.7	5.39	-0.7	5.5	-0.8	5.33	-0.6
4	4.41	-0.4	4.72	-0.7	4.43	-0.4	4.41	-0.4	4.11	-0.1
4.52	4.76	-0.2	4.32	0.2	4.63	-0.1	4.47	0.05	4.67	-0.2
3.96	3.97	-0	4.22	-0.3	4.53	-0.6	4.1	-0.1	4.17	-0.2
3.45	3.17	0.28	3.27	0.18	3.71	-0.3	3.24	0.21	3.29	0.16
5.22	6.04	-0.8	6.09	-0.9	6.04	-0.8	6.01	-0.8	6.07	-0.9
4.37	5.86	-1.5	5.12	-0.8	5.75	-1.4	5.63	-1.3	5.81	-1.4
8.23	7.02	1.21	6.88	1.35	7.13	1.1	7.07	1.16	6.99	1.24
7	6.27	0.73	6.22	0.78	6	1	6.53	0.47	6.12	0.88
6.47	7.5	-1	7.36	-0.9	7.55	-1.1	7.41	-0.9	7.73	-1.3
5.4	5.67	-0.3	6.15	-0.7	5.85	-0.5	6.09	-0.7	5.75	-0.3
7.02	7.2	-0.2	7.08	-0.1	6.93	0.09	7.28	-0.3	7.13	-0.1
5.33	5.2	0.13	5.55	-0.2	5.96	-0.6	5.59	-0.3	5.91	-0.6
7.99	7.94	0.05	7.96	0.03	7.84	0.15	7.79	0.2	7.9	0.09

8.51	8.08	0.43	8.03	0.48	7.76	0.75	8.15	0.36	7.93	0.58
8.55	8.31	0.24	9	-0.4	8.41	0.14	8.75	-0.2	8.47	0.08
8.24	8.39	-0.1	8.19	0.05	8.58	-0.3	8.48	-0.2	8.67	-0.4
5.06	5.62	-0.6	5.58	-0.5	5.72	-0.7	5.61	-0.5	5.67	-0.6
5.96	5.92	0.04	6.13	-0.2	5.91	0.05	5.92	0.04	5.92	0.04
7.06	6.06	1	5.87	1.19	5.81	1.25	6.16	0.9	6.06	1
7.58	7.55	0.03	6.95	0.63	7.74	-0.2	7.27	0.31	7.65	-0.1
7.89	8.44	-0.6	8.55	-0.7	8.3	-0.4	8.37	-0.5	8.4	-0.5
6.66	7.91	-1.2	7.49	-0.8	8.01	-1.4	7.62	-1	8.15	-1.5
6.66	6.14	0.52	6.5	0.16	6.48	0.18	6.28	0.38	6.25	0.41
5	6.65	-1.6	6.26	-1.3	6.4	-1.4	6.36	-1.4	6.4	-1.4
8.3	8.72	-0.4	8.5	-0.2	9.05	-0.7	8.8	-0.5	9.01	-0.7
8.11	7.45	0.66	7.88	0.23	7.98	0.13	7.95	0.16	7.82	0.29
7.37	7.74	-0.4	7.34	0.03	7.74	-0.4	7.55	-0.2	7.67	-0.3
6.01	5.3	0.71	4.95	1.06	5.22	0.79	5.11	0.9	5.21	0.8

Predicted1: Predicted value for equation 1, Residual1: Residual value for equation 1
Predicted2: Predicted value for equation 2, Residual2: Residual value for equation 2
Predicted3: Predicted value for equation 3, Residual3: Residual value for equation 3
Predicted4: Predicted value for equation 4, Residual4: Residual value for equation 4
Predicted5: Predicted value for equation 5, Residual5: Residual value for equation 5

Table 5. Summary of the five best QSAR models generated

	Equation 1	Equation 2	Equation 3	Equation 4	Equation 5
Friedman LOF	0.878001	0.88338	0.888121	0.889254	0.897552
R-squared	0.911782	0.911241	0.910765	0.910651	0.909817
Adjusted R-squared	0.897779	0.897153	0.896601	0.896469	0.895503
Cross validated R-squared	0.883022	0.880003	0.873742	0.879057	0.882876
Significant Regression	Yes	Yes	Yes	Yes	Yes
Significance-of-regression F-value	65.11385	64.679	64.30007	64.21014	63.55825
Critical SOR F-value (95%)	1.989118	1.989118	1.989118	1.989118	1.989118
Replicate points	0	0	0	0	0
Computed experimental error	0	0	0	0	0
Lack-of-fit points	63	63	63	63	63
Min expt. error for non-significant LOF (95%)	0.388461	0.389649	0.390693	0.390942	0.392762
RMSECV	0.4105	0.4118	0.4129	0.4131	0.4151
RMSEP	0.6951	0.6452	0.7134	0.6446	0.7132
R^2_{pred}	0.8526	0.8730	0.8447	0.8733	0.8447
Q^2_{LNO}	0.8816	0.8786	0.8721	0.8777	0.8816
SDEP	0.1236	0.1252	0.1285	0.1257	0.1236
K	0.9866	0.9907	0.9782	0.9816	0.9786
K’	1.0025	0.9998	1.0108	1.0094	1.0103
/R^2o-R’^{2}o/	0.0389	0.0299	0.0566	0.0324	0.0358
$R^2 - R_o^2/R^2$	0.0486	0.0362	0.0711	0..0389	0.0454
$R^2 - R_o'^2/R^2$	0.0002	0.0001	0.0002	0.0001	0.0006
$r^2_{m(test)}$	0.7926	0.8215	0.7876	0.8265	0.7811
R^2_p	0.8464	0.8423	0.8392	0.8348	0.8390

4.2 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 (Table 10). 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 [41,47,48].

Table 6. Summary of input data for genetic function approximation

Number of rows requested	74
Number of rows used	74
Number of rows omitted due to invalid row description	0
Number of rows omitted due to invalid data	0
Number of columns requested	238
Number of columns used	238
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

Table 7. The frequency of use of a particular descriptor in the population

Variable ID : pC50	Abbreviation Y	Occurrences in population
M : LogP	X13	4963
N : HBD Count	X14	4134
Z : ATSc3	X26	528
AT : VC-5	X46	259
AW : SPC-6	X49	4972
AY : VPC-5	X51	4990
BA : SP-4	X53	4994
BD : VP-3	X56	4707
BE : VP-4	X57	304
BI : ECCEN	X61	2058
BT : SHaaCH	X72	305
BZ : SaasC	X78	100
CC : SdO	X81	107
CK : minHBint7	X89	561
CO : minHaaCH	X93	127
CW : mindO	X101	166
DH : maxssCH2	X112	148
DU : ETA_dEpsilon_D	X125	765
EC : ETA_Beta_ns_d	X133	313
EK : Kier2	X141	214
FA : RotBFrac	X157	117
FB : nRotBt	X158	1821
FC : RotBtFrac	X159	4574
FE : topoDiameter	X161	381
FT : DPSA-2	X176	159
GG : MOMI-XZ	X189	740
GH : geomRadius	X190	108
GP : Weta2.unity	X198	496
GQ : Weta3.unity	X199	137
GS : WD.unity	X201	107
HE : Wlambda2.volume	X213	143
HF : Wlambda3.volume	X214	441
HI : Weta2.volume	X217	639
HK : WT.volume	X219	310
HL : WK.volume	X220	116
HZ : Weta1.polar	X234	368

Table 8. Table of all descriptors used in this study

Family	Descriptor	Description	Class
ChiClusterDescriptor			
	SC-5	Simple cluster, order 5	2D
	VC-5	Valence cluster, order 5	2D
ChiPathClusterDescriptor			
	SPC-6	Simple path cluster, order 6	2D
	VPC-5	Valence path cluster, order 5	2D
ChiPathDescriptor			
	SP-4	Simple path, order 4	2D
	VP-3	Valence path, order 3	2D
	VP-4	Valence path, order 4	2D
EccentricConnectivityIndexDescriptor			
	ECCEN	A topological descriptor combining distance and adjacency information	2D

ElectrotopologicalStateAtomTypeDescriptor			
	SHaaCH	Sum of atom-type H E-State: :CH:	2D
PaDELRotatableBondsCountDescriptor			
	RotBtFrac	Fraction of rotatable bonds, including terminal bonds	2D
WHIMDescriptor			
	Weta3.unity	Directional WHIM, weighted by unit weights	3D
	Weta2.volume	Directional WHIM, weighted by van der Waals volumes	3D
	WT.volume	Non-directional WHIM, weighted by van der Waals volumes	3D
	Weta1.polar	Directional WHIM, weighted by atomic polarizabilities	3D
StructuralDescriptors			
	HBD count	Number of hydrogen bond donors	2D
Thermodynamic			
	LogP	Partition Coefficient	3D

Table 9. The average R, R^2 and Q^2_{LOO} values after several Y-Randomization

Model no.	R_{yrand}	R^2_{yrand}	Q^2_{yrand}
1	0.3550	0.1312	-0.2189
2	0.3641	0.1392	-0.2009
3	0.3709	0.1439	-0.2109
4	0.3657	0.1385	-0.2022
5	0.3690	0.1409	-0.2087

Table 10. Euclidean based application domain for Model 1 and 2

No.	Training set: Model: 1			Training set: Model: 2		
	Distance score	Mean distance	Normalized mean distance	Distance score	Mean distance	Normalized mean distance
2	6417.245	86.72	0.023	260.37	3.519	0.007
3	8134.74	109.929	0.114	367.363	4.964	0.117
4	5988.343	80.924	0	264.626	3.576	0.011
5	5988.195	80.922	0	266.309	3.599	0.013
6	6052.786	81.794	0.003	282.424	3.817	0.029
7	6375.277	86.152	0.021	299.036	4.041	0.046
8	6422.607	86.792	0.023	339.914	4.593	0.089
10	6369.603	86.076	0.02	255.471	3.452	0.002
11	6368.698	86.063	0.02	253.979	3.432	0
12	6370.446	86.087	0.02	273.506	3.696	0.02
13	7022.548	94.899	0.055	298.721	4.037	0.046
14	7024.635	94.928	0.055	331.599	4.481	0.08
15	6053.396	81.803	0.003	293.023	3.96	0.04
16	6368.989	86.067	0.02	256.229	3.463	0.002
17	5998.776	81.065	0.001	278.337	3.761	0.025
18	6052.373	81.789	0.003	365.344	4.937	0.115
19	6434.475	86.952	0.024	285.106	3.853	0.032
20	6511.143	87.988	0.028	286.836	3.876	0.034
21	6080.442	82.168	0.005	264.599	3.576	0.011
22	6032.777	81.524	0.002	271.129	3.664	0.018
25	7275.349	98.316	0.068	282.147	3.813	0.029
26	7275.184	98.313	0.068	336.868	4.552	0.086

27	22275.86	301.025	0.864	486.675	6.577	0.24
28	5989.994	80.946	0	275.193	3.719	0.022
29	5988.516	80.926	0	257.58	3.481	0.004
30	5994.973	81.013	0	299.722	4.05	0.047
31	6419.362	86.748	0.023	324.2	4.381	0.072
32	6005.294	81.153	0.001	277.332	3.748	0.024
33	13044.73	176.28	0.374	453.384	6.127	0.206
34	8975.711	121.293	0.158	363.44	4.911	0.113
36	12089.76	163.375	0.324	408.259	5.517	0.159
37	6190.679	83.658	0.011	290.447	3.925	0.038
38	9725.494	131.426	0.198	338.04	4.568	0.087
39	6045.544	81.697	0.003	263.146	3.556	0.009
42	7274.631	98.306	0.068	272.71	3.685	0.019
43	6434.763	86.956	0.024	265.533	3.588	0.012
44	6054.034	81.811	0.003	295.952	3.999	0.043
45	7429.039	100.392	0.076	339.057	4.582	0.088
46	6654.738	89.929	0.035	265.588	3.589	0.012
47	6655.47	89.939	0.035	297.147	4.015	0.045
50	6380.379	86.221	0.021	316.866	4.282	0.065
51	6045.197	81.692	0.003	298.405	4.033	0.046
52	5988.894	80.931	0	269.927	3.648	0.016
53	6511.96	87.999	0.028	273.289	3.693	0.02
56	7935.983	107.243	0.103	444.326	6.004	0.196
57	9122.86	123.282	0.166	317.775	4.294	0.066
59	9124.343	123.302	0.166	393.824	5.322	0.144
60	17104.49	231.142	0.59	1223	16.527	1
61	23040.69	311.361	0.905	1080.162	14.597	0.853
62	16457.8	222.403	0.555	594.312	8.031	0.351
63	15520.03	209.73	0.506	534.024	7.217	0.289
67	10784.64	145.738	0.254	492.931	6.661	0.247
69	12293.85	166.133	0.334	446.488	6.034	0.199
71	14469.71	195.537	0.45	459.258	6.206	0.212
72	10051.69	135.834	0.216	391.298	5.288	0.142
78	24840.04	335.676	1	1079.839	14.592	0.852
79	10051.31	135.829	0.216	317.654	4.293	0.066
80	7274.939	98.31	0.068	292.66	3.955	0.04
81	7054.95	95.337	0.057	280.63	3.792	0.028
83	7275.295	98.315	0.068	291.507	3.939	0.039
87	11874.86	160.471	0.312	374.41	5.06	0.124
88	7000.693	94.604	0.054	316.385	4.275	0.064
90	7829.074	105.798	0.098	412.573	5.575	0.164
91	10787.27	145.774	0.255	420.425	5.681	0.172
92	14344.41	193.843	0.443	478.471	6.466	0.232
93	14344.14	193.84	0.443	455.852	6.16	0.208
94	14728.37	199.032	0.464	598.696	8.09	0.356
95	7332.329	99.086	0.071	322.853	4.363	0.071
98	6262.021	84.622	0.015	342.7	4.631	0.092
99	10786.3	145.761	0.255	344.603	4.657	0.094
100	13506.42	182.519	0.399	537.795	7.268	0.293
102	6515.135	88.042	0.028	288.235	3.895	0.035
103	6085.631	82.238	0.005	329.324	4.45	0.078
107	6044.618	81.684	0.003	311.872	4.214	0.06
	Test Set:	Equation: 1		Test Set:	Equation: 2	
No.	Distance Score	Mean Distance	Normalized Mean Distance	Distance Score	Mean Distance	Normalized Mean Distance

1	6052.392	81.789	0.003	275.79	3.727	0.023
9	6373.226	86.125	0.02	269.353	3.64	0.016
23	6258.549	84.575	0.014	313.207	4.233	0.061
24	6032.54	81.521	0.002	308.087	4.163	0.056
35	6514.128	88.029	0.028	280.669	3.793	0.028
40	7281.659	98.401	0.069	299.03	4.041	0.046
41	7276.445	98.33	0.068	279.32	3.775	0.026
48	7426.01	100.351	0.076	316.669	4.279	0.065
49	7427.383	100.37	0.076	302.828	4.092	0.05
54	12089.61	163.373	0.324	362.038	4.892	0.112
55	8976.403	121.303	0.159	285.218	3.854	0.032
58	7828.431	105.79	0.098	308.876	4.174	0.057
64	7329.028	99.041	0.071	296.171	4.002	0.044
65	8019.866	108.377	0.108	292.175	3.948	0.039
66	7825.859	105.755	0.097	337.891	4.566	0.087
68	14728.88	199.039	0.464	546.805	7.389	0.302
70	6009.081	81.204	0.001	351.49	4.75	0.101
73	8251.606	111.508	0.12	368.028	4.973	0.118
74	11874.43	160.465	0.312	403.69	5.455	0.154
75	13505.95	182.513	0.399	402.862	5.444	0.154
76	6998.368	94.573	0.054	355.503	4.804	0.105
77	5999.233	81.071	0.001	308.683	4.171	0.056
82	7275.301	98.315	0.068	308.232	4.165	0.056
84	9816.197	132.651	0.203	382.816	5.173	0.133
85	9125.161	123.313	0.166	386.296	5.22	0.137
86	8254.067	111.541	0.12	385.994	5.216	0.136
89	8022.266	108.409	0.108	324.558	4.386	0.073
96	5995.88	81.025	0	284.772	3.848	0.032
97	6052.282	81.788	0.003	287.4	3.884	0.034
101	7003.77	94.646	0.054	333.223	4.503	0.082
104	6044.345	81.68	0.003	296.586	4.008	0.044
105	6036.658	81.576	0.003	287.493	3.885	0.035
106	7274.306	98.301	0.068	285.18	3.854	0.032

5. CONCLUSION

In the present investigation, a QSAR model for a set of HEPT derivatives that have the capability of inhibiting in vitro strain of HIV. The leave-one-out (LOO) and leave-many-out (LNO) cross-validation methods, the Y-randomization technique, and the external validation indicated that the model is significant, robust and has good internal and external predictability. The inhibitory activity of the investigated compounds was described based in descriptors: LogP, HBD count, SPC-6, VPC-5 SP-4, VP-3, ECCEN, SHaaCH, RotBTFrac and Weta2.volume. Thermodynamic, ChiPath, ChiCluster, ChiPathCluster, EccentricConnectivityIndex, ElectroTopological State atom type, Rotatable bonds count, Structural and WHIM descriptors play a significant role in explaining the activity of the data set. The results indicated that the activity against strain of HIV is favoured by higher partition coefficient, valence path cluster, order5, simple path, order 4, sum of atom-type H E-state: CH:, directional WHIM, weighted by van der waals volumes, smaller number of hydrogen bond donors, simple path cluster, order 6, valence path, order 3, decreased topological descriptor combining distance and adjacent information. The mechanism of action is related with structural and thermodynamic aspects of the compounds, which can explained by the descriptors that were selected in the QSAR model proposed. The study indicates that the increase of LogP, VPC-5, SP-4, SHaaCH, RotBtFrac and Weta2.volume would be contributing for biological activity. It's important the synthesis of 1-[(2-hydroxyethoxy) methyl]-6-(phenylthio)thyio)thymine) with these descriptors for verify the authenticity of the facts. The proposed model may provide a better understanding of the ant-HIV activity of 1-[(2-hydroxyethoxy) methyl]-6-(phenylthio)thyio)thymine) and can be used as guidance for proposition of new chemo-preventive agents.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. Ettari R, Pinto A, Micale N. Synthesis and anti-HIV activity evaluation of new phenyl ethyl thiourea (PET) derivatives. ARKIVOC. 2009;14:227-234.
2. Artico M. Non-nucleoside anti-HIV-1 reverse transcriptase inhibitors (NNRTIs) a chemical survey from lead compounds to selected drugs for clinical trials. Farmaco. 1996;51:305–331.
3. Jain VS, Sonawane VL, Patil RR, Bari SB. Pharmacophore modeling of some novel indole B-diketo acid and coumarin-based derivatives as HIV integreas inhibitors. Journal of Medicinal Chemistry Research. 2012;21(2):165-173.
4. Ojha LK, Chaturvedi AM, Bhardwaj A, Thakur M, Thakur A. QSAR analysis of some TIBO derivatives as HIV-1 reverse transcriptase inhibitors. International Research Journal of Pure & Applied Chemistry. 2013;3(4):417-427.
5. Pommier Y, Johnson AA, Marchand C. Integrase inhibitors to treat HIV/AIDS. Nature Reviews Drug Discovery. 2005; 4(3):236-248.
6. Cole SR, Herna´n MA, Anastos K, Jamieson BD, Robins JM. Determining the effect of highly active antiretroviral therapy on changes in Human Immunodeficiency Virus type 1 RNA viral load using a marginal structural left-censored mean model. American Journal of Epidemiology. 2007;166(2):1-9.
7. Lee P, Knight R, Smit JM, Wilschut J, Griffin DE. A single mutation in the E2 glycoprotein important for neurovirulence influences binding of sindbis virus to neuroblastoma cells. Journal of Virology. 2002;76(12):6302–6310.
8. Liu Q, Liu D, Yang Z. Characteristics of human infection with avian influenza viruses and development of new antiviral agents. Acta Pharmacologica Sinica. 2013;34:1257–1269.
9. Huang Y, Wang X, Yu X, et al. In hibitoy activity of 9-phenylcyclohepta[d] pyrimidinedione derivatives against different strains of HIV-1 as non-nucleoside reverse transcriptase inhibitors. Virology Journal. 2011;8:230.
10. Zahouily M, Rakik J, Lazar M, Bahlaoui MA, et al. Exploring QSAR of non-nucleoside reverse transcriptase inhibitors by artificial neural networks: HEPT derivatives. Journal of ARKIVOC. 2007;14: 245-256.
11. Mitsuya H, Border S. Inhibition of the in vitro infectivity and cytopathic effect of human T-Lymphotrophic virus type III/lymphadenopathy-associated virus (HTLV-III/LAV) by 2', 3'-dideoxynucleosides. Proceeding of the National Academy of Science of the United State America. 1986;83(6):1911-1915.
12. De Clercq E. Perspective of non-nucleoside reverse transcriptase inhibitors (NNRTIs) in the therapy of HIV-1 infection. II Farmaco. 1999;54:26-45.
13. Thakur A, Tiwari BK, Thakur M, Thakur S, Pandey ND, Narvi SS, 2D, 3D modeling of Inhibition activity of reverse transcriptase-1 by HEPT derivatives. Asian Journal of Biochemistry. 2007;2(2):84-100.
14. Ivan D, Crisan L, Funar-Timofei S, Mracec M. 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. 2013;78(4): 495-506.
15. Shaik B, Zafar T, Agrawal VK. Estimation of anti-HIV activity of HEPT analogues using MLR, ANN, and SVM techniques. International Journal of Medicinal Chemistry. 2013;2013:1-8.
16. Li B, Chiang C, Hsu LY. QSAR studies of 3,3'-(substituted-benzylidene)-bis-4-hydroxycoumarin, potential HIV-1 integrase inhibitor. Journal of the Chinese Society. 2010;57:742-749.
17. Ravichandran V, Harish R, Abhishek J, et al. Validation of QSAR models – strategies and importance. International Journal of Drug Design and Discovery. 2011;2(3):511-519.
18. Karelson M, Lobanov VS. Quantum-chemical descriptors in QSAR/QSPR studies. Chemical Reviews. 1996;96: 1027-1043.
19. Agrawal VK, Singh J, Mishra KC, Khadikar PV, Jaliwala YA. QSAR Studies on the use of 5,6-dihydro-2-pyrones as HIV-1 protease inhibitors. ARKIVOC, (ii). 2006; 162-177.

20. De Melo EB, Ferreira MMC. Multivariate QSAR study of 4,5-dihydroxypyrimidine carboxamides as HIV-1 integrase inhibitors. European Journal of medicinal Chemistry. 2009;44:3577-3583.
21. Debnath AK, Jiang S, Strick N, Lin K, Haberfield P. Three-dimensional structure-activity analysis of a series of porphyrin derivatives with Anti-HIV-1 activity targeted to the V3 loop of the gp120 envelope glycoprotein of the human immunodeficiency virus type 1. Journal of Medicinal Chemistry. 1994;37:1099-1108.
22. Zarei K, Atabati M. QSAR study of anti-HIV activities against HIV-1 and some of their mutant strains for a group of HEPT derivatives. Journal of the Chinese Chemical Society. 2009;56:206-213.
23. Adebimpe AO, Dash RC, Soliman MES. QSAR study on Diketo and Carboxamide derivatives as potent HIV-1 integrase inhibitor. Journal of Latters in Drug Design & Discovery. 2014;11.
24. Thakur M, Thakur A, Ojha L. Surface area gggrid in modeling of anti HIV activity of TIBO derivatives. International Journal of Research and Development in PHARMACY and Life Science. 2014;3(3): 983-992.
25. Mahani NM, Sabermahani F, Mahdavi SA. A DFT based QSAR study of novel 4-substituted 1,5-diarylanilines as potent Hiv-1 agents using quantum chemical descriptors. International Journal of Pharmaceutical Chemical and Biological Sciences. 2014;4(2):250-255.
26. Sapre NS, Bhati T, Gupta S, Pancholi N, Raghuvanshi U, Dubey D, Rajopadhyay V, Sapre N. Computational modeling studies on anti-HIV-1 non-nucleoside reverse transcriptase inhibition by dihydroalkoxybenzyloxopyrimidines analogues: An electrotopological atomistic approach. Journal of Biophysical Chemistry. 2011;2(3):361-372.
27. Adimi M, Salimi M, Nekoei M, Pourbasheer E, Beheshti A. A quantitative structure–activity relationship study on histamine receptor antagonists using the genetic algorithm–multi-parameter linear regression method. Journal of the Serbian Chemical Society. 2012;77(5):639–650.
28. Douali L, Villemin D, Cherqaoui D. Exploring QSAR of non-nucleoside reverse transcriptase inhibitors by neural networks: TIBO derivatives. International Journal of Molecular Sciences. 2004;5:48-55.
29. Leonard JT, Roy K. Exploring molecular shape analysis of styrylquinoline derivatives as HIV-1 integrase inhibitors. European Journal of Medicinal Chemistry. 2008;43:81-92.
30. Gupta P, Roy N, Garg P. Docking-based 3D-QSAR study of HIV-1 integrase inhibitors. European Journal Medicinal Chemistry. 2009;44:4276-4287.
31. Dessalew N. Investigation of the structural requirement for inhibiting HIV integrase: QSAR study. Acta Pharmaceutica. 2009;59:31-43.
32. Sharma H, Cheng X, Buolamwini JK. Homology model-guided 3D-QSAR studies of HIV-1 integrase inhibitors. Journal of Chemical Information and Modeling. 2012;52:515-544.
33. Luco JM, Ferretti FH. QSAR based on multiple linear regression and PLS methods for the anti-HIV activity of a large group of HEPT derivatives. Journal of Chemical Information and Computer Science. 1997;37(2):392-401.
34. Karki RG, Kulkarni VM. Three-dimensional quantitative structure–activity relationship (3D-QSAR) of 3-Aryloxazolidin-2-one antibacterials. Bioorganic & Medicinal Chemistry. 2001;9:3153–3160.
35. Friedman JH. Multivariate adaptive regression splines. Annals of Statistics. 1991;19:1-67.
36. Hadizadeh F, Vahdani S, Jafarpour M. Quantitative structure-activity relationship studies of 4-Imidazolyl- 1,4-dihydropyridines as calcium channel blockers. Iranian Journal of Basic Medical Sciences. 2013;16:910-916.
37. Patel V, Chhabria MT, Brahmkshatriya PS, Mahajan B, Patel S. QSAR study of series of 4-amino-3,5-di(substituted)thiazol-2(3H)-thione using Genetic function approximation (GFA) as antitubercular agents. PHARMAGENE. 2013;1(2):1-6.
38. Khaled KF. Modeling corrosion inhibition of iron in acid medium by genetic function approximation method: A QSAR model. Corrosion Science. 2011;53:3457–3465.
39. Khaled KF, Abdel-Shafi NS. Quantitative structure and activity relationship modeling study of corrosion inhibitors: Genetic function approximation and molecular dynamics simulation methods. International Journal of Electrochemical Science. 2011;6:4077–4094.

40. Niculescu SP. Artificial neural networks and genetic algorithms in QSAR. Journal of Molecular Structure (Theochem). 2003;622:71–83.
41. 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.
42. Roy PP, Paul S, Mitra I, Roy K. On two novel parameters for validation of predictive QSAR models. Molecules. 2009;14(5):1660-1701.
43. Kiralj R, Ferreira MMC. Basic validation procedures for regression methods in QSAR and QSPR studies: Theory and application. Journal of Brazilian Chemical Society. 2009;20(4):770-787.
44. Golbraikh A, Tropsha A. Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. Journal Computer.-Aided Molecular Design. 2002;16(5-6):357-369.
45. Pogliani L. Structure property relationships of amino acids and some dipeptides. Amino Acids. 1994;6(2):141–153.
46. Pogliani L. Modeling with special descriptors derived from a medium-sized set of connectivity indices. Journal of Physical Chemistry. 1996;100(46): 18065–18077.
47. 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.
48. Jaworska J, Nikolova-Jeliazkova N, Aldenberg T. QSAR applicabilty domain estimation by projection of the training set descriptor space: A review. Alternative Laboratory Animals. 2005;33:445–459.

Peer-review history:
The peer review history for this paper can be accessed here:
http://sciencedomain.org/review-history/12156