KIMIKA Volume 24, Number 2, pp. 2-17 (2013)

ISSN 0115-2130 (Print); 2508-0911 (Online)
https://doi.org/10.26534/kimika.v24i2.2-17

MLR-Based QSAR Models for Predicting Inhibitory Activity of Reverse Transcriptase by HEPT Derivatives using GETAWAY Descriptors

Alex A. Tardaguila[1*], Jennifer C. Sy[1], Marielyn R. Omañada[1], and Eric R. Punzalan[2]

[1] *Department of Physical Science, Pamantasan ng Lungsod ng Maynila, Intramuros, Manila, Philippines*
[2] *Department of Chemistry, De La Salle University, Malate, Manila, Philippines*

In this study, quantitative structure-activity relationship (QSAR) models for non-nucleoside reverse transcriptase inhibitors based on 1-[(2-hydroxyethoxy)-methyl]-6-(phenylthio)thymine (HEPT) derivatives were generated. The structures of the compounds and their activities were obtained from the literature. The data set were divided into two sets: training set (N=91) and validating set (N=10). All 3-D structures of these inhibitors were optimized by semi-empirical method, AM1 prior to calculations of 3-D molecular descriptors, GETAWAY. Multiple linear regression (MLR) using stepwise method was applied to determined significant descriptors. Out of 197 GETAWAY descriptors, 4-14 molecular descriptors have significant relationships with the activities (expressed as log $(1/EC_{50})$) of HEPT. The MLR method generated 14 models. The predictive power of these models were evaluated internally by applying the following statistical parameters for the training set and test set: root-mean-square error for prediction (RMSE), correlation coefficient (R), squared correlation coefficient (R^2), adjusted squared correlation coefficient (R^2_{adj}), difference between R^2 and R^2_{adj} ($R^2 - R^2_{adj}$), squared cross-validation correlation coefficient (Q^2). External validation was performed by employing Golbraikh and Tropsha criteria. Moreover, residual analysis was performed. Internal validation of Model XX (N = 91) revealed that it has the highest predictive power (RMSE = 0.4288, R = 0.9393, R^2 = 0.882, R^2_{adj} = 0.8620, R^2 - R^2_{adj} = 0.0203, Q^2 = 0.8317). However, external validation (using the validating set, N=10) showed that Model XII has the highest predictive power (R^2 = 0.961, R^2_0 = 0.9565, k = 0.8648, k' = 0.9800, [R^2 - R^2_0] = 0.0066, [R^2 - R^2_0] / R^2 = 0.0069, R^2_{pred} = 0.9481) based on Golbraikh and Tropsha criteria. Residual analysis confirmed that both models are valid.

Keywords: *QSAR, HIV, reverse transcriptase, HEPT, MLR*

INTRODUCTION

Retroviruses such as HIV-1 requires reverse transcriptase (RT) to convert RNA into proviral DNA that can be inserted into host DNA for HIV production (Le Grice, 1993; Basu et al.,1992). The HIV-1 RT enzyme is an asymmetric heterodimer composed of p66

* Authors to whom correspondences should be addressed; email: einstein_alpha@yahoo.com, punzalane@gmail.com

(560 amino acids) and p51 subunits (440 amino acids) (Kohlstaedt et al., 1992). Because of this process, RT has become target for anti-HIV drug discovery. The strategy is to inhibit the RT using candidate molecules. Example of these molecules are 1-[(2-hydroxyethoxy)-methyl]-6-(phenylthio) thymine (HEPT) derivatives (De Clercq, New, 2001; Baba et al.; 1989, Miyasaka et al., 1989; Tiwari et al., 2006). HEPT are non-nucleoside inhibitors that block RT by binding to a compartment adjacent to the catalytic site of the enzyme and thereby interrupt the conformation of several amino acids crucial for proper RT function (De Clercq, 1998). The binding site contains five aromatic (Tyr-181, Tyr-188, Phe-227 and Trp-229), six hydrophobic (Pro-59, Leu-100, Val-106, Val-179, Leu-234 and Pro-236) and five hydrophilic (Lys-101, Lys-103, Ser-105, Asp-132 and Glu-224) amino acids that belong to the p66 subunit and two amino acids (Ile-135 and Glu-138) belonging to the p51 subunit (Kohlstaedt et al., 1992). However, drug development is time-consuming and expensive. Thus, it is necessary to develop a method that can predict efficacy of the candidate drug prior to laboratory or clinical stage to reduce drug development cost (Kola and Landis, 2004).

Quantitative structure-activity relationship (QSAR) models have been used intensively in predicting biological activities, chemical properties and toxicities of many organic molecules (Bazoui et al., 2002; Duda-Seiman et al., 2007; Verma et al., 2010). QSAR models are generated by finding significant relationships between activities/properties and molecular descriptors. Furthermore, QSARs studies are vital part of drug development because they shorten the amount of time to find a better lead compounds by allowing us to predict activities/toxicities of these compounds using molecular descriptors without undergoing tedious animal and laboratory testing (Verma et al., 2010). Currently, there are more than 3000 molecular descriptors that are used in QSAR studies (Todeschini and Consonni, 2009). These are categorized into different types: 0D, 1D, 2D, and 3D molecular descriptors. In this work, GETAWAY descriptors were used for QSAR modelling.

GETAWAY stands for **GE**ometry, **T**opology, and **A**tom **W**eights **A**ssemblY. V. Consonni, et al. (Consonni, Theory, 2002) proposed this set of novel molecular descriptors based on a leverage matrix, called Molecular Influence Matrix (MIM). GETAWAY is a set of 3D molecular descriptors that matches 3D molecular geometry provided by MIM and atom relatedness by topology with chemical information by different atomic weighting schemes such as unit weights, mass, polarizability, electronegativity. GETAWAY descriptors have low or no degeneracy, which avoids getting the same value for a descriptor for more than one compound sharing the same structural features which are often observed in topological descriptors. The H is defined by

$$H=M\cdot\left(M^{T}\cdot M\right)^{-1}\cdot M^{T} \qquad \text{(Equation 1)}$$

where M is the molecular matrix which represents the location of each atom (A) of an optimized molecule using Cartesian coordinates x, y, z. The resultant A x A matrix is invariant to rotation of the molecular coordinates. The superscript T refers to the transposed of the matrix. The diagonal elements h_{ij} are leverages and represent the influence of each atom in determining the shape of the molecule. Each off diagonal element $h_{ij,}$ represents the degree of accessibility of the j'th atom to interactions with the i'th atom. GETAWAY descriptors are classified into two types: H – GETAWAY and R – GETAWAY. We illustrate here some of the descriptors derived by V. Consonni, et al. (Consonni, Theory, 2002).

H – GETAWAY descriptors are calculated based on the information provided by the MIM. The H – GETAWAY autocorrelation descriptors use the geometric information inherent in the leverage values and atomic weighting schemes to make a new set of descriptors. For example, the HATS indices are defined using a weight vector w' for each atom in a molecule:

$$w'_i = w_i \cdot h_{ii} \qquad \text{(Equation 2)}$$

Therefore,

$$HATS_0(w) = \sum_{i=1}^{A}(w_i \cdot h_{ii})^2 \qquad \text{(Equation 3)}$$

and the $HATS_k$ is written as

$$HATS_k(w) = \sum_{i=1}^{A-1}\sum_{J>i}(w_i \cdot h_{ii}) \cdot (w_j \cdot h_{jj}) \cdot \delta(k; d_{ij}) \qquad \text{(Equation 4)}$$

where d_{ij} is the topological distance between the i'th and j'th atoms and k = 1, 2,...,d with

$$\delta(k; d_{ij}) = \begin{cases} 1 & if \quad d_{ij} = k \\ 0 & if \quad d_{ij} \neq k \end{cases}$$

Hence, the HATS total index is

$$HATS(w) = (W')^T \cdot U \cdot W' = \sum_{i=1}^{A}(w_i \cdot h_{ii})^2 + 2\sum_{i=1}^{A-1}\sum_{j>1} w_i \cdot h_{ii} \cdot w_j \cdot h_{jj}$$

$$= HATS_0(w) + 2\sum_{k=1}^{d} HATS_k(w) \qquad \text{(Equation 5)}$$

A second set, called R – GETAWAY, combines information with geometric interatomic distances in the molecule. R and R+ descriptors are obtained from the leverage/geometry matrix. The matrix R is defined as

$$[R]_{ij} = \left[\frac{\sqrt{h_{ii} \cdot h_{jj}}}{r_{ij}}\right]_{ij} \qquad \text{(Equation 6)}$$

where r_{ij} is the geometric distance between the two atoms. Also, the R – GETAWAY autocorrelational indices are defined analogously to the H – GETAWAY autocorrelational indices. Hence, we have the w weighted k'th order autocorrelation index (R_k)

$$R_k(w) = \sum_{i=1}^{A-1}\sum_{j>1}\frac{\sqrt{h_{ii} \cdot h_{jj}}}{r_{ij}} w_i \cdot w_j \cdot \delta(k; d_{ij}) \qquad \text{(Equation 7)}$$

The R total index is defined as

$$RT(w) = W^T \cdot R \cdot W = 2 \cdot \sum_{i=1}^{A-1}\sum_{j>1}\frac{\sqrt{h_{ii} \cdot h_{jj}}}{r_{jj}} \cdot w_i \cdot w_j = 2 \cdot \sum_{k=1}^{d} R_k(w) \qquad \text{(Equation 8)}$$

V.Consonni et al. (2002) evaluated extensively the prediction ability of GETAWAY descriptors by analyzing the regressions of these descriptors for selected properties of some reference compound classes. Also, they studied the general performance of these descriptors in QSAR/QSPR with respect to other well-known sets of molecular descriptors. They concluded that GETAWAY descriptors provide more predictive models when the property to be modeled depends strictly on the 3D features of the molecule,

e.g. biological activities. Furthermore, when the GETAWAY descriptors are added to other types of descriptors the predictive ability of the model improved.

Multiple linear regressions (MLR) have been widely used method in QSAR research (Verma et al., 2010). MLR has an advantage for its simplicity and ease of interpretation because the model assumes a linear relationship between activity and molecular descriptors. The MLR model is expressed as

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \quad \text{(Equation 9)}$$

where β_0 is model constant, $x_1,...,x_k$ are molecular descriptors with their corresponding coefficients $\beta_1,\ldots\beta_k$. These coefficients are calculated by least-squares method, which minimizes the sum of squared residuals. The magnitudes of the coefficients indicate the extent of influence of the corresponding descriptors on the activity (y).

Several QSAR studies involving HEPT derivatives as inhibitors of HIV-RT using 0D, 1D, 2D and 3D molecular descriptors has been reported in the literature. Castro and colleagues (2005) developed QSAR for HEPT using Morgan Extended Connectivity in Graph of Atomic Orbitals (GAO). The best model was based on correlation weights of Morgan extended connectivity of first order in the GAO (N = 59, R^2 = 0.91, s = 0.41, F = 577). Application of the model to a test set (N= 20) gave a good prediction power based on coefficient of determination (R^2 = 0.91, s = 0.42, F = 183).

Kobra Zarei and Morteza Atabati (2009) generated five QSAR models using MLR with 5-13 descriptors for HEPT inhibitors of HIV-1 wild type and mutant strains. They used several sets descriptors, namely, topological, molecular walk counts, 2D autocorrelation, geometrical, WHIM, GETAWAY, RDF, BCUT, 3D-MoRSE, physico-chemical. Internal validation of the five models showed high prediction power (N = 23, R^2 = 0.940-0.999). Moreover, the five models have high prediction power when applied to HEPT derivatives in the test set (N = 7, R^2 = 0.849-0.992).

Bazoui H., et al. (2002) performed QSAR studies on 103 HEPT derivatives. Their results showed that the anti-HIV activity of HEPT derivatives was strongly dependent on hydrophobic character and also steric factors of substituents. These findings were confirmed by C. Duda – Seiman, et.al. (2007).They used topological, π – Hansch hydrophobic substituent constant, the ES – Taft steric constant and the B1 (STERIMOL) as descriptors. The best model they obtained has nine descriptors and high coefficient of determination (N = 79, R^2 = 0.949; s = 0.44; Q^2 = 0.745).

In the recent study of D. Ivan, et al. (2013), the best MLR-based QSAR they obtained has five descriptors of different types (topological, connectivity index, and GETAWAY). The model has high coefficient of determination (N = 91, R^2 = 0.826, s = 0.682, F = 84.4) and satisfactorily predicted the activities of HEPT derivatives in the test set (N =20, R^2 =0.675, R^2_{pred} = 0.663).

The objective of this research was to generate QSAR models using the GETAWAY descriptors to predict the inhibitory activities of HEPT derivatives against RT. We performed extensive statistical analyses to validate the models. Furthermore, we decided to use these descriptors because they represent 3D molecular chemical information of the molecules under study. The weighing schemes used in GETAWAY descriptors allow the researcher to easily design or manipulate the structure of the molecule to have better activity once the QSAR model is established. Moreover, MLR – based QSAR study on HEPT derivatives using only GETAWAY descriptors has not yet been reported in the literature. Thus, this research adds to the present knowledge about the applicability of GETAWAY descriptors in QSAR studies.

METHODOLOGY

We used the data set of 101 HEPTs (Figure 1) with inhibitory activities against HIV-RT reported in the literature (Table 1). The data were divided into two sets: training set (N = 91) and test set (N = 10). All 3-D structures of these HEPTs were optimized by MM+ followed by semi – empirical method, AM1, applying the Polak – Rabiere conjugate gradient with restricted Hartree – Fock (RHF) spin pairing, 0.001 convergence limit in vacuo and RMS gradient of 0.001 kcal/A mol. AM1 method allows a large number of structures to be optimized in a shorter timeframe, and for calculations involving large molecular systems. Also, AM1 has been applied for optimization of HEPT structures in the literature (Hannongbua, Structure, 1996; Zarei and Atabati, 2009; Ivan, et al., 2013). All calculations were performed using Hyperchem (Hypercube Inc.). After geometry optimization, 197 GETAWAY were calculated using Dragon software (Talete srl).

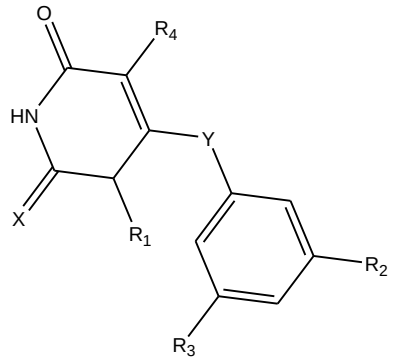

Figure 1. *General structure of HEPT derivatives.*

Multiple linear regression (MLR) using stepwise method was applied to determined significant descriptors. Internal validation of these models was evaluated by applying the following statistical parameters for the training set and test set: root – mean – square error for prediction (RMSEP), coefficient of determination (R^2), adjusted R^2 (R^2_{adj}).

$$\mathrm{RMSEP} = \sqrt{\frac{\sum(\hat{y}_i - y_i)^2}{n}} \qquad \text{(Equation 10)}$$

$$R^2 = \frac{SSR}{SST} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} \qquad \text{(Equation 11)}$$

$$Q^2 = 1 - \frac{\sum(\hat{y}_i - y_i)^2}{\sum(y_i - \bar{y})^2} = 1 - \frac{PRESS}{SS} \qquad \text{(Equation 12)}$$

$$R^2_{adj} = \frac{R^2 - (1 - R^2)(p-1)}{(n-p)} \qquad \text{(Equation 13)}$$

where:

n = number of samples
SSR = Regression Sum of Squares
= sum of the squared deviations of the predicted value $(\hat{y}_i)$ about the mean $(\bar{y})$
SST = Total Sum of Squares
= sum of the squared deviations of the observed variable (y_i) about the mean
PRESS = predictive residual sum of squares
= the difference between the predictive values and observed values
SS = sum of squares
= the difference between the observed values and their mean

Generally, an acceptable QSAR model generated from training set is considered to have a higher predictive if the following criteria are met: $R > 0.8$, $R^2 > 0.6$, $R^2_{adj} > 0.6$, $[R^2 - R^2_{adj}] < 0.3$, $Q^2 > 0.65$.

For external validation, the following Golbraikh–Tropsha criteria (Golbraikh et al., 2003; Golbraikh and Tropsha, 2002) were evaluated to determine the predictive ability of the MLR-based QSAR model applied to a test set (N =10).

- $Q^2 > 0.5$ (squared cross-validation correlation coefficient)
- $R^2 > 0.6$ (the squared correlation coefficient R between the predicted and observed activities)
- $(R^2 - R_0^2) / R^2 < 0.1$ (the coefficients of determination for the predicted vs. the observed activities R_0^2)
- $0.85 \leq k \leq 1.15$ or $0.85 \leq k' \leq 1.15$ (slopes k and k' of the regression lines through the origin)
- $R^2 - R_0^2 < 0.3$.

The predictive power of QSAR models were also tested by predictive parameter (R^2_{pred}). For a predictive QSAR model, the value of R^2_{pred} should be higher than 0.5

$$R^2_{pred} = 1 - \frac{SSE}{SST} = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

(Equation 14)

In addition, we performed correlation analysis among the independent variables (e.g. variation inflation factor (VIF), and tolerance) (Kutner et al., 2004). Paired t-test was employed to determine the significant difference between the experimental and predicted inhibitory activities by the models. All statistical analyses were performed using SPSS software (SPSS Inc.).

RESULTS AND DISCUSSION

QSAR analysis was carried out in order to explore properties, which might be responsible for the interaction of molecules with HIV – RT receptors. The structures of molecules, which were employed in this study, are shown in Table 1. MLR using stepwise method was employed to select significant descriptors and to generate QSAR models. Twenty QSAR models were generated using the training set, N = 91(Table 2). These models have four to 14 GETAWAY descriptors which have significant correlations with the actual activities of HEPT derivatives (Table 3). The number of independent variables considered in the QSAR models was five to six times less than the number of molecules in the training set, which is the acceptable number of variables for MLR modeling for a particular database size used in this study (Tropsha et al., 2003).

Table 1. Structures of the Molecules used in this study (Zahouily, et al. 2007).

ID	R_1	R_2	R_3	R_4	X	Y	Expt'l Log $(1/EC_{50})$	Predicted Log $(1/EC_{50})$ *Model XII*	*Model XX*
Training Set									
1	$CH_2O(CH_2)_2OMe$	H	H	Me	O	S	5.060	5.764	5.337
2	CH_2OMe	H	H	Me	O	S	5.677	5.765	6.097
3	CH_2OEt	H	H	Me	O	S	6.481	5.721	5.887
4	CH_2OPr	H	H	Me	O	S	5.443	5.292	5.402
6	CH_2OCH_2Ph	H	H	Me	S	S	7.055	6.777	6.841
7	CH_2OEt	H	H	Et	S	S	7.585	7.225	7.212
9	CH_2OEt	Cl	Cl	Et	S	S	7.886	7.798	7.852
10	CH_2O-i-Pr	H	H	Et	S	S	6.657	6.570	6.833
11	CH_2O-c-Hex	H	H	Et	S	S	5.795	5.747	5.765
13	CH_2OCH_2Ph	H	H	Et	S	S	8.096	7.626	7.919
15	$CH_2OCH_2C_6H_4$(4-Me)	H	H	Et	S	S	7.107	7.555	7.724
16	$CH_2OCH_2C_6H_4$(4-Cl)	H	H	Et	S	S	7.920	7.245	7.521
17	$CH_2OCH_2CH_2Ph$	H	H	Et	S	S	7.041	7.319	7.436
18	CH_2OEt	H	H	i-Pr	S	S	7.835	7.632	7.712
20	CH_2OEt	H	H	c-Pr	S	S	7.022	6.887	7.065
21	CH_2OEt	H	H	Et	O	S	7.721	6.991	7.088
23	CH_2OEt	Cl	Cl	Et	O	S	8.154	8.594	8.761
24	CH_2O-i-Pr	H	H	Et	O	S	6.468	7.074	6.884

25	CH_2O-c-Hex	H	H	Et	O	S	5.397	6.593	6.308
26	CH_2OCH_2-c-Hex	H	H	Et	O	S	6.346	6.240	6.251
27	CH_2OCH_2Ph	H	H	Et	O	S	8.221	7.284	7.558
28	CH_2OCH_2Ph	Me	Me	Et	O	S	8.522	7.737	7.860
29	$CH_2OCH_2CH_2Ph$	H	H	Et	O	S	7.017	6.477	6.910
30	CH_2OCH_2OEt	H	H	i-Pr	O	S	7.920	7.577	7.035
31	CH_2OCH_2Ph	H	H	i-Pr	O	S	8.522	8.216	8.189
32	CH_2OEt	H	H	c-Pr	O	S	7.000	7.273	7.307
33	Et	H	H	Me	O	S	5.657	6.082	6.218
34	n-Bu	H	H	Me	O	S	5.920	5.571	5.414
35	$CH_2OCH_2CH_2OH$	Me	H	Me	O	S	5.585	5.375	5.371
37	$CH_2OCH_2CH_2OH$	t-Bu	H	Me	O	S	4.920	5.860	5.738
38	$CH_2OCH_2CH_2OH$	CF_3	H	Me	O	S	4.346	4.827	4.689
39	$CH_2OCH_2CH_2OH$	F	H	Me	O	S	5.481	4.525	4.384
40	$CH_2OCH_2CH_2OH$	Cl	H	Me	O	S	4.886	5.155	5.108
41	$CH_2OCH_2CH_2OH$	Br	H	Me	O	S	5.244	5.240	5.493
42	$CH_2OCH_2CH_2OH$	I	H	Me	O	S	5.000	4.584	4.590
43	$CH_2OCH_2CH_2OH$	NO_2	H	Me	O	S	4.468	3.897	4.159
44	$CH_2OCH_2CH_2OH$	OH	H	Me	O	S	4.086	5.066	4.822
45	$CH_2OCH_2CH_2OH$	OMe	H	Me	O	S	4.657	5.418	5.381
46	$CH_2OCH_2CH_2OH$	Me	H	Me	O	S	6.585	6.222	6.361
48	$CH_2OCH_2CH_2OH$	Me	Me	Me	S	S	6.657	6.775	6.750
49	$CH_2OCH_2CH_2OH$	COOMe	H	Me	O	S	5.102	4.716	5.013
50	$CH_2OCH_2CH_2OH$	COMe	H	Me	O	S	5.136	5.539	5.544
51	$CH_2OCH_2CH_2OH$	CN	H	Me	O	S	5.000	4.687	4.900
52	$CH_2OCH_2CH_2OH$	H	H	Et	O	S	6.958	6.494	6.717
53	$CH_2OCH_2CH_2OH$	H	H	i-Pr	S	S	7.229	7.696	7.661
54	$CH_2OCH_2CH_2OH$	Me	Me	Et	S	S	8.090	8.291	8.206
55	$CH_2OCH_2CH_2OH$	Me	Me	i-Pr	S	S	8.301	8.151	8.174
56	$CH_2OCH_2CH_2OH$	Cl	Cl	Et	S	S	7.366	7.759	7.757
57	$CH_2OCH_2CH_2OH$	H	H	Et	O	S	6.920	6.494	6.706
58	$CH_2OCH_2CH_2OH$	H	H	i-Pr	O	S	7.200	7.170	7.261
59	$CH_2OCH_2CH_2OH$	Me	Me	Et	O	S	7.886	7.603	7.656
60	$CH_2OCH_2CH_2OH$	Me	Me	i-Pr	O	S	8.522	8.280	7.962
61	$CH_2OCH_2CH_2OH$	Cl	Cl	Et	O	S	7.853	7.309	7.421
62	$CH_2OCH_2CH_2OH$	H	H	$CH{=}CH{-}CH_2$	O	S	5.602	5.569	5.282
63	$CH_2OCH_2CH_2OH$	H	H	I	O	S	5.366	5.507	5.075
64	$CH_2OCH_2CH_2OH$	H	H	$CH{=}CH_2$	O	S	5.327	5.235	5.508
65	$CH_2OCH_2CH_2OH$	H	H	n-Pr	S	S	5.000	6.186	5.908
66	$CH_2OCH_2CH_2OH$	H	H	n-Pr	O	S	5.468	6.015	5.835
67	$CH_2OCH_2CH_2OH$	H	H	H	O	S	5.154	5.392	5.107
68	$CH_2OCH_2CH_2OH$	H	H	H	S	S	6.008	6.299	5.888

70	CH_2CH=CH-Ph	H	H	Me	O	S	5.602	6.180	6.144
71	CH_2CH=CH-Ph	Me	Me	Et	O	S	7.408	7.540	7.618
72	CH_2CH=CH-thiényl	H	H	Et	O	S	7.096	7.222	7.548
73	CH_2CH=CH-thiényl	Me	Me	Et	O	S	7.397	7.301	7.009
74	CH_2CH=CH-furyl	H	H	Et	O	S	6.769	6.475	6.744
75	CH_2CH=CH-furyl	Me	Me	Et	O	S	7.397	7.327	7.213
76	$CH_2CH=CH_2$-pyridyl	H	H	Et	O	S	7.000	6.768	7.006
77	$(CH_2)_3$Ph	H	H	Et	O	S	6.537	6.899	6.735
78	$CH_2OCH_2CH_2OH$	H	H	Et	O	CH_2	6.455	6.575	6.567
80	CH_2OEt	H	H	Et	O	CH_2	7.387	6.670	6.717
81	CH_2OEt	Me	Me	Et	O	CH_2	8.795	8.019	8.241
82	$CH_2OCH_2CH_2OH$	H	H	i-Pr	O	CH_2	7.200	7.241	7.048
83	$CH_2OCH_2CH_2OH$	Me	Me	i-Pr	O	CH_2	8.568	8.407	8.163
84	CH_2OEt	H	H	i-Pr	O	CH_2	7.376	7.909	7.960
85	CH_2OEt	Me	Me	i-Pr	O	CH_2	9.221	8.987	8.923
86	n-Bu	H	H	Et	O	CH_2	6.677	6.591	6.737
87	n-Bu	H	H	i-Pr	O	CH_2	7.376	7.508	7.741
88	CH_2CH_2OMe	H	H	Et	O	CH_2	6.602	6.681	6.872
89	CH_2CH_2OMe	H	H	i-Pr	O	CH_2	7.284	7.432	7.708
90	CH_2O CH_2OMe	H	H	Me	O	CH_2	4.638	4.827	4.637
91	$CH_2OCH_2CH_2OH$	c-Hex	H	Me	O	S	5.455	5.277	5.009
92	$CH_2OCH_2CH_2OH$	H	H	Me	O	O	5.050	5.134	5.104
93	$CH_2OCH_2CH_2OH$	H	H	Me	O	S	5.244	4.906	4.919
94	CH_2SCH_3	H	H	Et	O	CH_2	8.689	7.930	7.989
95	$CH_2SCH_2CH_3$	H	H	Et	O	CH_2	7.398	7.119	7.024
96	CH_2SCH_3	Me	Me	Et	O	CH_2	7.301	8.570	8.218
97	$CH_2SCH_2SCH_3$	Me	Me	Et	O	CH_2	8.398	8.520	8.411
98	CH_2SCH_3	H	H	i-Pr	O	CH_2	7.699	7.737	7.187
99	$CH_2SCH_2SCH_3$	H	H	i-Pr	O	CH_2	8.222	8.773	8.747
100	$CH_2OCH_2CH_2Cl$	H	H	Me	O	S	5.820	6.042	5.887
101	$CH_2OCH_2CH_2OCOPh$	H	H	Me	O	S	5.120	5.575	5.469
Test Set									
5	CH_2OBu	H	H	Me	O	S	5.327	5.219	5.075
8	CH_2OEt	Me	Me	Et	S	S	8.397	8.347	8.406
12	CH_2OCH_2-c-Hex	H	H	Et	S	S	6.455	6.369	6.103
14	CH_2OCH_2Ph	Me	Me	Et	S	S	8.154	7.904	7.803
19	CH_2OCH_2Ph	H	H	i-Pr	S	S	8.154	7.692	7.977
22	CH_2OEt	Me	Me	Et	O	S	8.301	7.733	7.675
36	$CH_2OCH_2CH_2OH$	Et	H	Me	O	S	5.568	5.630	5.509
47	$CH_2OCH_2CH_2OH$	Cl	Cl	Me	O	S	5.886	6.012	6.119
69	CH_2CH=CH-Ph	H	H	Et	O	S	6.721	7.055	7.238
79	$CH_2OCH_2CH_2OH$	Me	Me	Et	O	CH_2	7.886	7.658	7.844

Table 2. Internal Validation of 20 QSAR models using Training Set (N = 91).

MODEL	No. of GETAWAY Descriptors	INTERNAL VALIDATION (N = 91)					
		RMSE	R	R^2	R^2_{adj}	R^2-R^2_{adj}	Q^2
I[a]	4	0.8472	0.8462	0.7161	0.6720	0.0441	0.6860
II	4	0.6763	0.8410	0.7073	0.6910	0.0163	0.6687
III	4	0.6763	0.8410	0.7073	0.6940	0.0133	0.6775
IV	5	0.6536	0.8524	0.7266	0.7110	0.0156	0.6891
V	6	0.6074	0.8740	0.7639	0.7470	0.0169	0.7281
VI	7	0.5761	0.8875	0.7876	0.7700	0.0176	0.7474
VII	8	0.5624	0.8931	0.7976	0.7780	0.0196	0.7506
VIII	9	0.5423	0.9010	0.8118	0.7910	0.0208	0.7645
IX	10	0.5172	0.9104	0.8288	0.8070	0.0218	0.7793
X	11	0.5026	0.9156	0.8384	0.8160	0.0224	0.7852
XI	12	0.4884	0.9205	0.8474	0.8240	0.0234	0.7941
XII[b]	**13**	**0.4731**	**0.9256**	**0.8568**	**0.8330**	**0.0238**	**0.8013**
XIII	14	0.4586	0.9303	0.8654	0.8410	0.0244	0.8059
XIV	13	0.4612	0.9294	0.8639	0.8410	0.0229	0.8080
XV	12	0.4685	0.9271	0.8595	0.8380	0.0215	0.7980
XVI	13	0.4552	0.9313	0.8674	0.8450	0.0224	0.8159
XVII	12	0.4556	0.9312	0.8672	0.8470	0.0202	0.8150
XVIII	13	0.4431	0.9351	0.8743	0.8530	0.0213	0.8224
XIX	14	0.4287	0.9393	0.8824	0.8610	0.0214	0.8275
XX[b]	**13**	**0.4288**	**0.9393**	**0.8823**	**0.8620**	**0.0203**	**0.8317**

[a] *Model I has one outlier (#65)*
[b] *Selected models used in this study.*

Internal validation of the training set (N = 91) using RMSE, R, R^2, ($R^2 - R^2_{adj}$), Q^2 revealed that model XX is the best 13 – descriptor QSAR model (Table 3). The model has an acceptable values of root – mean – squared error (RMSE) < 1.00, correlation coefficient (R) > 0.80, squared correlation coefficient (R^2) > 0.65, adjusted squared correlation coefficient (R^2_{adj}) > 0.6 , difference between R^2 and R^2_{adj} ($R^2 - R^2_{adj}$) < 0.30, which, all show that the results are not based on chance correlation. The model's Q^2 > 0.65 supports the predictive ability and significance of the model.

Model XX has two mass –, three electronegativity –, two van der Waals volume –, one polarizabilty – related, and five unweighted GETAWAY descriptors. In this model the most (negative) influential variable to Log ($1/EC_{50}$) is R_2u^+ (R maximal autocorrelation of lag 2 / unweighted), which is related to three dimensional geometry of the molecule (Consonni et al., 2002). However, after we performed correlation analysis, we found several multicollinear descriptors that may complicate the interpretation of the model. To test if there is complication exists, we subsequently calculated the variance inflation factor (VIP) for each variable. VIP provides an index that measures how much the variance of an estimated regression coefficient is increased because of collinearity (Kutner et al., 2004). Results indicate that the variances of all regression coefficients of the model are not severely affected by collinearity (VIP). Moreover, the 13 variables in the model explained 88.2% variation in Log ($1/EC_{50}$) and they are statistically significant ($P < 0.01$). The F statistic ($F = 44.416$, $P < 0.0001$) shows that the regression equation/model is statistically significant. The

Table 3. List of significant descriptors used in Model XII and XX.

Descriptor	Name
$\mathbf{R_4e^+}$	R maximal autocorrelation of lag 4 / weighted by atomic Sanderson electronegativities
$\mathbf{H_1e}$	H autocorrelation of lag 1 / weighted by atomic Sanderson electronegativities
$\mathbf{H_6m}$	H autocorrelation of lag 6 / weighted by atomic masses
$\mathbf{R_8v}$	R autocorrelation of lag 8 / weighted by atomic van der Waals volumes
$\mathbf{H_8u}$	H autocorrelation of lag 8 / unweighted
$\mathbf{HATS_6u}$	leverage-weighted autocorrelation of lag 6 / unweighted
$\mathbf{R_8e}$	R autocorrelation of lag 8 / weighted by atomic Sanderson electronegativities
$\mathbf{R_3u}$	R autocorrelation of lag 3 / unweighted
$\mathbf{R_7e}$	R autocorrelation of lag 7 / weighted by atomic Sanderson electronegativities
$\mathbf{R_1v^+}$	R maximal autocorrelation of lag 1 / weighted by atomic van der Waals volumes
$\mathbf{R_4m}$	R autocorrelation of lag 4 / weighted by atomic masses
$\mathbf{H_3u}$	H autocorrelation of lag 3 / unweighted
$\mathbf{R_2u^+}$	R maximal autocorrelation of lag 2 / unweighted
$\mathbf{R_5e^+}$	R maximal autocorrelation of lag 5 / weighted by atomic Sanderson electronegativities
$\mathbf{RTp^+}$	R maximal index / weighted by atomic polarizabilities
$\mathbf{R_8u}$	R autocorrelation of lag 8 / unweighted
$\mathbf{H_7v}$	H autocorrelation of lag 7 / weighted by atomic van der Waals volumes

values of t statistic for all regression coefficients are statistically significant (non – zero) ($P < 0.001$) and all descriptors can be used to explain the dependent variable Log ($1/EC_{50}$).

Model XX

$$\begin{aligned}\text{Log}\,(1/EC_{50}) = {} & 22.656 - 6.304\ (\pm 1.365)\ H_1e - 9.374\ (\pm 1.406)\ H_6m \\ & + 6.019\ (\pm 2.172)\ R_8v + \\ & 6.404\ (\pm 1.221)\ HATS_6u - \\ & 7.721\ (\pm 2.695)\ R_8e - 5.648 \\ & (\pm 0.801)\ R_3u + 3.011\ (\pm 0. \\ & 546)\ R_4m + 2.051\ (\pm 0.\ 434) \\ & H_3u - 50.338\ (\pm 9.509)\ R_2u^+ \\ & + 18.628\ (\pm 6.401)\ R_5e^+ - \\ & 14.008\ (\pm 2.885)\ RTp^+ + \\ & 7.621\ (\pm 2.901)\ R_8u + \\ & 10.914\ (\pm 2.155)\ H_7v\end{aligned}$$

$N = 91$, $R = 0.939$, $R^2 = 0.882$, $F = 44.416$, $SE = 0.4662$, $P < 0.001$

To further validate the model, we employed residual analysis on the residuals calculated from the difference between the experimental and predicted inhibitory activities by Model XX. For the MLR model to be valid there are three assumptions to be checked on the residuals: (1) no outliers; (2) the data points must be independent; (3) the distribution of the residuals must be normal with mean = 0, and constant variance (Chan, 2004). The first assumption is satisfied by the model. For the second, we used Durbin-Watson estimate to check for independence. Values near two indicate that data points are independent. The value we obtained for Model XX is 1.869 which satisfies the independence assumption. Furthermore, it can be seen in Figure 2 (A – C) that the distribution of the standardized residual is normal (mean = 0) and the scatter of the points has no clear pattern indicating that the variance is constant. Therefore the third assumption is satisfied. These results conformed to our internal validation performed for Model XX, which, suggests the model is valid.

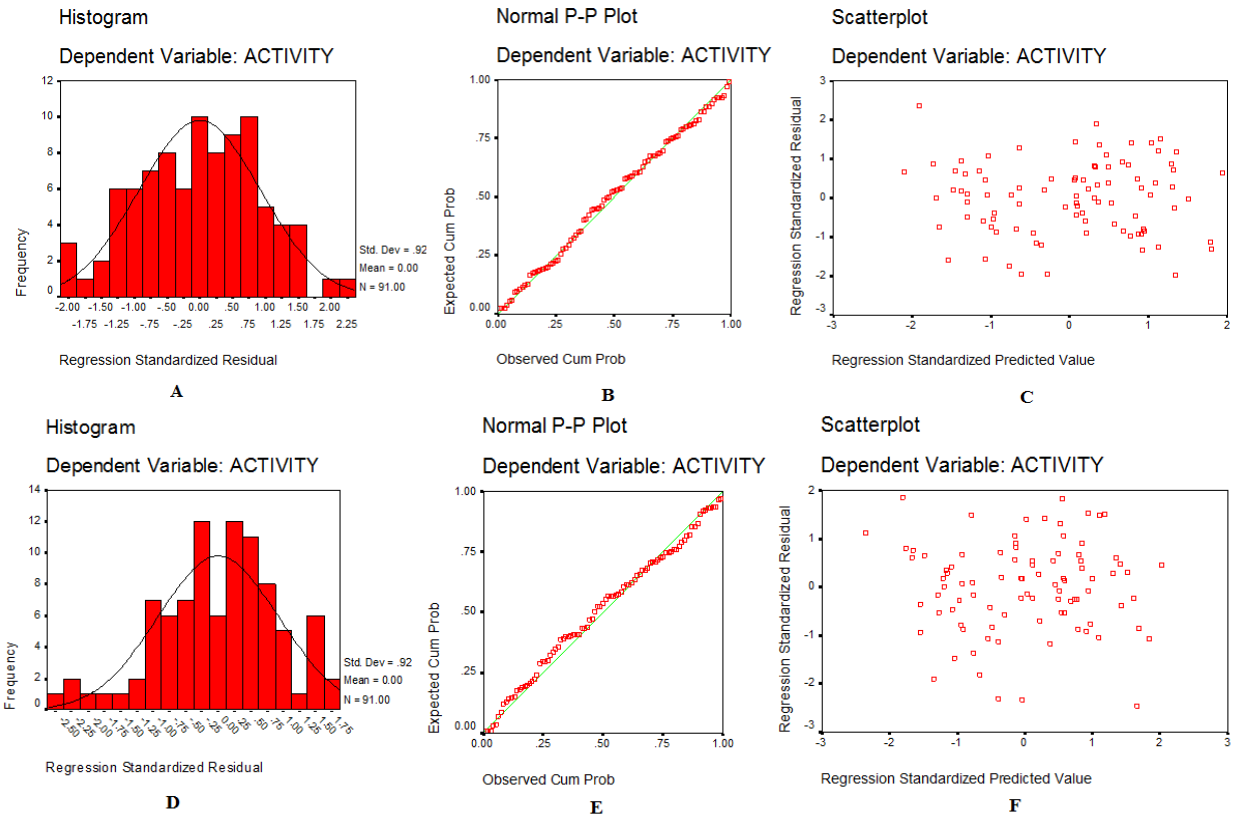

Figure 2. *Distribution and variance of standardized residuals calculated from the difference between experimental and predicted inhibitory activities by Model XX (A-C) and Model XII (D – F).*

Moreover, to estimate the true predictive power of a QSAR model, it must be able to predict the activities of the molecules in the test set and compare the predicted with the experimental values (Veerasamy et al., 2011). We applied the Golbraikh and Tropsha criteria to estimate the predictive power of Model XX using the test set (N = 10). Based on R^2, R^2_0, k, k', $[R^2 - R^2_0]$, $[R^2 - R^2_0] / R^2$, R^2_{pred} presented in Table 4, Model XX satisfactorily predicted the activities of the HEPT derivatives in the test set. However, the calculated value of k for Model XX (k = 0.7864) is not within the range of Golbraikh and Tropsha criteria ($0.85 \leq k \leq 1.15$). Hence, we applied the same criteria for other models to find a better model.

After applying the same procedure for the other 19 models and we found that Model XII has the best prediction performance when applied to test set, although, it is not the best QSAR model based on the internal validation. This has been expected since it was reported (Kubinyi et al., 1997; Kubinyi, Three, 1998) that in general there is no relationship between internal and external validations of the model. Furthermore, our results confirmed the findings (Kubinyi et al., 1997) in the literature that low internal predictivity may result to high external predictivity and *vice versa*.

Table 4. Predictive power results for the external test set; Golbraikh and Tropsha criteria.

Model	XX	XII
R^2	0.8575	0.9631
R_0^2	0.8395	0.9565
k	**0.7864**	0.8648
k'	0.9260	0.9800
$\lvert R^2 - R_0^2 \rvert$	0.0180	0.0066
$\frac{R^2 - R_0^2}{R^2}$	0.0210	0.0069
R^2_{pred}	0.9315	0.9481

Model XII has two mass –, four electronegativity –, two van der Waals volume – related and five unweighted GETAWAY descriptors. The two most influential variables in this model are R_1v^+ (R maximal autocorrelation of lag 1/weighted by atomic van der Waals volumes) and R_2u^+ (R maximal autocorrelation of lag 2 / unweighted) which are both related to the 3D geometry of the HEPT derivatives. Akin to model XX, model XII has muticolinearity among its variables; however, VIPs calculations suggest that this has no significant threat to predictivity of the model (VIP < 10, tolerance > 0.10). The 13 variables explained 85.7% of variance in Log ($1/EC_{50}$). Moreover, the F statistic (F = 35.430, P < 0.0001) indicates that the regression equation/model is significant. Furthermore, the values of t statistic for all regression coefficients are all significant and can be used to explain Log ($1/EC_{50}$).

Moreover, residual analysis on the residuals calculated between the experimental and predicted inhibitory activities by Model XII revealed that the training set used to generate model has no outlier. The model's Durbin – Watson estimate value is 1.693. It can be observed in Figure 2 (D – F), the distribution of its residual is normal (mean = 0) and the variance is constant. The results indicate that the Model XII is valid. The graph of correlations between the predicted and experimental values of inhibitory activity of HEPT (in the training set) using model XII and XX are presented in Figure 3.

The comparison of predictive performances of model XII and XX using the test set based on Golbraikh and Tropsha criteria are shown in Table 4. Unlike model XX, model XII conformed to the criteria. The graph of correlations between the predicted and experimental values of inhibitory activities of HEPT (in the test set) using model XII and XX are presented in Figure 4. These results show that all calculated statistical parameters indicate that both models have good predictive power. Nevertheless, the values obtained for model XII are more satisfactory than that for Model XX.

Model XII

$$\text{Log } (1/EC_{50}) = 23.434 (\pm 3.717) - 20.414 (\pm 8.377)\ R_4e^+ - 6.592 (\pm 1.243)\ H_1e - 6.609 (\pm 1.233)\ H_6m + 12.352 (\pm 2.965)\ R_8v + 2.804 (\pm 0.585)\ H_8u + 5.610 (\pm 1.365)\ HATS_6u - 4.026 (\pm 1.484)\ R_8e - 3.633 (\pm 0.850)\ R_3u + 2.299 (\pm 1.086)\ R_7e - 24.506 (\pm 6.663)\ R_1v^+ + 1.675 (\pm 0.507)\ R_4m + 1.196 (\pm 0.449)\ H_3u - 25.569 (\pm 11.379)\ R_2u^+$$

N = 91, R^2 =0.857, F = 35.430, SE = 0.5143, P < 0 .0001

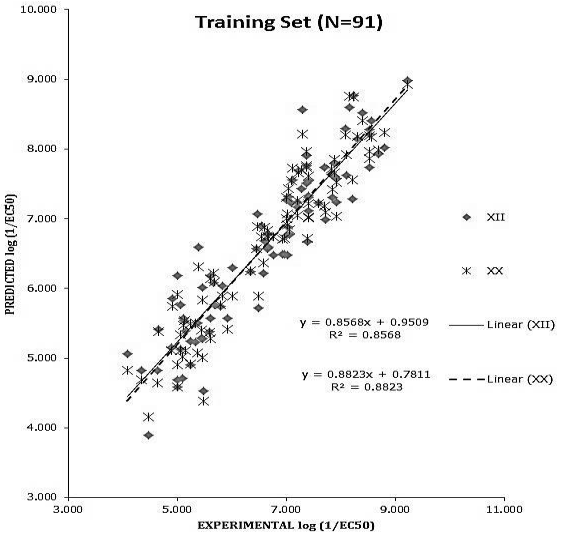

Figure 3*. Correlations between the experimental and predicted values of inhibitory activity of HEPT using model XII and XX using the training set (N = 91)*

We further explored the predictive performances of the two selected models using the compounds in the test set by employing paired t test and residual analysis (Table 5). Results show that there is no significant difference between the predicted and experimental activities ($t_{(XII)}$ = 1.444, *P* = 0.183; $t_{(XX)}$ = 0.998, *P* = 0.344, respectively). Likewise, the residuals distributions are normal ($mean_{(XII)}$ = 0.1230 and $mean_{(XX)}$ = 0.1488) and have constant variance for both models.

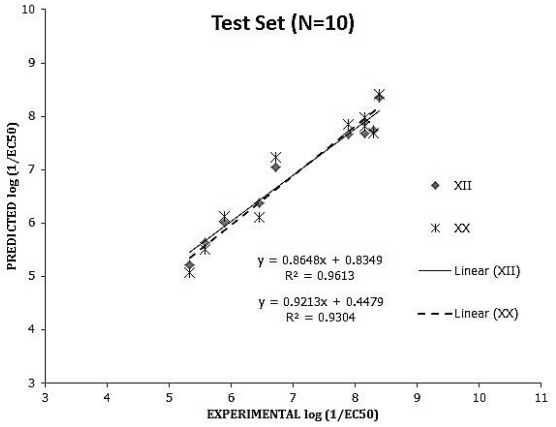

Figure 4*. Correlations between the predicted and experimental values of inhibitory activity of HEPT using model XII and XX using the test set (N = 10).*

The mean relative error of predictions exhibited by model XII is 3.064 (SD 2.0958) % and model XX is 4.899 (SD 3.7468) %. The results show that the prediction performances of the two models have high accuracy. Also, these confirmed the validity of the models.

The predictive powers (internal and external validations) of the present models are generally better than the recent QSAR model with three types of descriptors (topological, connectivity index, and GETAWAY) generated by D. Ivan, et al. (2013). In contrast, the model obtained by C. Duda – Seiman, et al. (2007) is more satisfactory than the present models. They used topological, hydrophobic and steric –related descriptors. However, they did not use a test set to validate their model.

Table 5. Comparison of the predictive performances of model XII and XX.

ID	Expt'l	XX	XII	Residual		% Relative Error	
				XX	XII	XX	XII
79	7.886	7.624	7.658	0.262	0.228	3.328	2.890
5	5.327	5.333	5.219	-0.006	0.108	-0.120	2.028
8	8.397	7.869	8.347	0.528	0.050	6.287	0.595
12	6.455	6.970	6.369	-0.515	0.086	-7.974	1.327
14	8.154	8.358	7.904	-0.204	0.250	-2.503	3.065
19	8.154	7.389	7.692	0.765	0.462	9.378	5.672
22	8.301	7.404	7.733	0.897	0.568	10.801	6.837
36	5.568	5.559	5.630	0.009	-0.062	0.168	-1.114
47	5.886	5.737	6.012	0.149	-0.126	2.527	-2.145
69	6.721	7.118	7.055	-0.397	-0.334	-5.902	-4.970

Nevertheless, the five models presented by Kobra Zarei and Morteza Atabati (2009) have higher prediction power than the present models, but, the models were based on smaller training set (N = 23) and test set (N = 7). In addition, most of the HEPT QSAR studies reported used 2D and 3D molecular descriptors, which might complicate the interpretation of the models (Ivan et al., 2013; Zarei and Atabati, 2009). In contrast, the use of one type of descriptors (particularly 3D molecular descriptors) simplifies the information about the possible interaction of the ligand to the enzyme under study. This makes drug design more manageable albeit the structure of the receptor or enzyme is unknown.

The GETAWAY descriptors selected in the present models suggest that the geometry, net electronegativity and size of the molecules play an important role in their activities. These conform to the results of 3D-QSAR studies on HIV RT-HEPT interaction, that the steric and electrostatic interactions for the HIV-1 RT-HEPT affect inhibitory activities (Hannongbua et al., 2001). Moreover, these observations agree well with the experimental studies on the crystal structure of HIV-1-HEPT complex. The increase in potency of the HEPT is due to

the interaction between residue Tyr181 in the enzyme and the 6-benzyl ring of the inhibitors, which stabilizes the structure of the complex (Hopkins et al., 1996). Also, our models support the conclusion in the literature that GETAWAY descriptors provide more predictive models for biological activities that are dependent on 3D features of the molecule (Consonni et al., 2002).

In this work, we performed extensive statistical analyses to validate Model XII and XX for their prediction performances. Our results showed that the best model, XII was realized after performing validation methods and applying the models to predict the activities using the test set.

Validating procedures are necessary to establish the predictive performances of the QSAR models. Most of the researches in QSAR modeling relied only on several statistical parameters (e.g. R, R^2, Q^2) to validate their MLR – based QSAR models. Hawkins D.M., et al. (2003) argued that for sample size holding a portion of it back for testing is wasteful, and it is much better to use the cumbersome leave – one – out cross – validation. In contrast, A. Golbraikh, and A Tropsha (2002) suggested that this assumption is generally incorrect and there is no correlation between the values of Q^2 for the training set and predictive ability for the test set. Hence, the high value of Q^2 is necessary, but not an indication that the model has high predictive power. Still, the best method of validating a model is external validation (Golbraikh and Tropsha, 2002; Veerasamy et al., 2011). Furthermore, applying all available validation strategies to check the robustness of the model is necessary (Veerasamy et al., 2011).

In summary, the present QSAR models, XII and XX, due to their high predictive power, can be used as an alternative method to the costly and time – consuming experiments for determining the inhibitory activities of new HEPT derivatives. The two models have five H – and eight R – GETAWAY descriptors. These descriptors are related to the geometry, distributions of electronegativity, van der Waals, and atomic mass in a molecule. The results conform to the idea that many biological activities are dependent on the three – dimensional arrangement of atoms in a molecule (Schuur et al., 1996; Vracˇko, 2002). Furthermore, this study demonstrated that the use of GETAWAY descriptors successfully predicted the inhibitory activity of HEPT against HIV-RT.

CONCLUSION

The MLR – based QSAR models for the inhibitory activity of 91 HEPT derivatives have been obtained using GETAWAY descriptors. The best MLR model is XII and has 13 GETAWAY descriptors. The present models have good stability, robustness, and predictability when verified by internal validation. Application of the model XII to HEPT derivatives in a test set, satisfactorily predicted their inhibitory activities. Moreover, the model suggests that the three dimensional distributions of atomic electronegativity, mass, volume, and geometry of the molecule have considerably effect on the inhibitory activities of HEPT derivatives. Furthermore, these models offer new theoretical tools for drug design and development.

ACKNOWLEDGEMENT

We would like to thank Commission on Higher Education (CHEd) for financial assistance.

REFERENCES

Baba M, Tanaka H, De Clercq E, Pauwels R, Balzarini J, Schols D et al. Highly specific inhibition of human immunodeficiency virus type 1 by a novel 6-substituted acyclouridine derivative. Biochem. Biophys. Res. Commun. 1989; 165(3): 1375 - 81.

Basu A, Basu S, and Modak MJ. Structure-activity analyses of HIV-1 reverse transcriptase. Biochem. Biophys. Res. Commun. 1992; 183(3): 1131 - 38.

Bazoui H, Zahouily M, Boulajaaj S, Sebti S, and Zakarya D. QSAR for anti-HIV activity of HEPT derivatives.SAR QSAR Environ Res. October 2002; 13(6): 567 - 577.

Castro EA, Toropov AA, Toropova AP, and Mukhamedzhanove DV. QSAR Modeling of Anti-HIV-1 Activity of HEPT Derivatives. Optimization of Correlation Weights of Morgan Extended Connectivity in Graph of Atomic Orbitals. J. Argent. Chem. Soc. 2005; 93(4 - 6): 109 - 121.

Chan YH. Biostatistics 201: Linear Regression Analysis. Singapore Med. J. 2004; 45(2): 55-61.

Consonni V, Todeschini R and Pavan M. Structure/Response Correlations and Similarity/Diversity Analysis by GETAWAY Descriptors, 1. Theory of the Novel 3D Molecular Descriptors. J. Chem. Inf. Comp. Sci.2002;: 682-692.

Consonni V, Todeschini R, Pavan M, and Gramatica P. Structure/Response Correlations and Similarity/Diversity Analysis by GETAWAY Descriptors, 2. Application of the Novel 3D Molecular Descriptors to QSAR/QSPR Studies. Chem. Inf. Comp. Sci. 2002; 42: 693 - 705.

De Clercq E. The role of nonnucleoside reverse transcriptase inhibitors (NNRTIs) in the therapy of HIV-1 infection. Antiviral Res. 1998; 38: 153–179.

De Clercq E. New developments in anti-HIV chemotherapy. Pure Appl. Chem. 2001; 73(1): 55–66.

Duda-Seiman C, Duda-Seimana D, Putz MV and Ciubotariub D. QSAR Modelling of Anti-HIV Activity with HEPT Derivatives. Digest Journal of Nanomaterials and Biostructures. June 2007; 2(2): 207-219.

Golbraikh A and Tropsha A. Beware of q2!. J Mol Graph Model. January 2002; 20(4): 269-276.

Golbraikh A and Tropsha A. Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. J Comput Aided Mol Des. May 2002; 16(5 -6): 357-369.

Golbraikh A, Shen M, Xiao Z, Xiao YD, Lee KH and Tropsha A. Rational selection of training and test sets for the development of validated QSAR models . J Comput Aided Mol Des. February 2003; 17(2 - 4): 241-253.

Hannongbua S, Lawtrakul L and Limtrakul J. Structure-activity correlation study of HIV-1 inhibitors: Electronic and molecular parameters. J Comput Aided Mol Des. April 1996; 10(2):145-152.

Hannongbua S, Nivesanond K, Lawtrakul L, Pungpo P, and Wolschann P. 3D-quantitative structure-activity relationships of HEPT derivatives as HIV-1 reverse transcriptase inhibitors, based on Ab initio calculations. J Chem Inf Comput Sci. May - Jun 2001; 41(3): 848 - 55.

Hawkins DM, Basak SC and Mills D. Assessing model fit by cross-validation. J. Chem. Inf. Comput. Sci. 2003; 43(2): 579-86.

Hopkins A, Ren J, Esnouf R, Willcox B, Jones E, Ross C et al. Complexes of HIV-1 reverse transcriptase with inhibitors of the HEPT series reveal conformational changes relevant to the design of potent non-nucleoside inhibitors. J Med Chem. April 1996; 12(39): 1589 - 600.

Ivan D, Crisan L, Funar-Timofei S, and Mracec M. A quantitative structure – activity relationships study for the anti - HIV - 1activities of 1-[(2-hydroxyethoxy)methyl]-6- -(phenylthio)thymine derivatives using the multiple linear regression and partial least squares methodologies. J Serb Chem. Soc. 2013; 78,(4): 495 - 506.

Kohlstaedt L, Wang J, Friedman J, Rice P and Steitz T. Crystal structure at 3.5 A resolution of HIV-1 reverse transcriptase complexed with an inhibitor. Science. 1992; 256: 1783-1790

Kola I and Landis J. Can the pharmaceutical industry reduce attrition rates? Nat Rev Drug Discov. 2. 2004; 3: 711–716.

Kubinyi H, van de Waterbeemd H, Testa B and Folkers G, editors. In Computer-Assisted Lead Finding and Optimization. Basel, Weinheim: VHChA and VCH; 1997.

Kubinyi H, Hamprecht FA and Mietzner T. Three-dimensional quantitative similarity-activity relationships (3D QSiAR) from SEAL similarity matrices. J Med Chem. 1998; 41: 2553-2564.

Kutner MH, Nachtsheim CJ and Neter J. Applied Linear Regression Models. 4th. McGraw-Hill Irwin; 2004.

Le Grice S F J and A M S S P G, editors. New York: Cold Spring Harbor laboratory press; 1993.

Miyasaka T, Tanaka H, Walker RT, Balzarini J, De Clercq E. A novel lead for specific anti-HIV-1 agents: 1-[(2-hydroxyethoxy)methyl]-6-(phenylthio)thymine. J Med Chem. 1989;32.

Schuur JH, Selzer P and Gasteiger J. The coding of the three-dimensional structure of molecules by molecular transforms and its application to structure-spectra correlations and studies of biological activity. J Chem Inf Comput Sci. 1996; 36: 334–344.

Tiwari BK, Thakur A, Thakur M, Pandey ND, Narvi SS and Thakur S. Modeling of cytotoxicity on some non - nucleoside reverse transcriptase inhibitors of HIV-1: role of physicochemical parameters. Arkivoc. 2006: 213 - 225.

Todeschini R and Consonni V. Molecular Descriptors for Chemoinformatics (2 volumes). Weinheim (Germany): WILEY-VCH; 2009.

Verma J, KhedkarVM and Coutinho EC. 3D-QSAR in drug design--a review. Current Topics in Medicinal Chemistry. 2010; 10: 95-115.

Veerasamy R, Rajak H, Jain A, Sivadasan S, Varghese CP, and Agrawal RK. Validation of QSAR Models - Strategies and Importance. Int J of Drug Disc. July - September 2011;2(3).

Vracˇko, M and Gasteiger J. A QSAR study on a set of 105 flavonoid derivatives using descriptors derived from 3D structures. Internet Electron J Mol Des. 2002; 1: 527–544.

Zahouily M, Rakik J, Lazar M, Bahlaoui MA and Rayadh A and Komiha N. Exploring QSAR of non-nucleoside reverse transcriptase inhibitors by artificial neural networks: HEPT derivatives . ARKIVOC. 2007; 14: 245-256.

Zarei K, and Morteza A. QSAR Study of Anti-HIV Activities Against HIV-1 and Some of Their Mutant Strains for a Group of HEPT Derivatives. Jnl Chinese Chemical Soc. 2009; 56: 206-213.