



# QSAR study of PETT derivatives as potent HIV-1 reverse transcriptase inhibitors

Razieh Sabet<sup>a</sup>, Afshin Fassihi<sup>a,b,\*</sup>, Behzad Moeinifard<sup>c</sup>

<sup>a</sup> Department of Medicinal Chemistry, Faculty of Pharmacy, Isfahan University of Medical Sciences, 81746-73461, Isfahan, Iran

<sup>b</sup> Isfahan Pharmaceutical Sciences Research Center, 81746-73461, Isfahan, Iran

<sup>c</sup> Department of Chemistry, Islamic Azad University, Shahreza Branch, Shahreza, Iran

## ARTICLE INFO

### Article history:

Received 24 January 2009

Received in revised form 13 May 2009

Accepted 19 May 2009

Available online 23 May 2009

### Keywords:

HIV-1 reverse transcriptase

PETT derivatives

QSAR

PCRA

FA-MLR

## ABSTRACT

A series of phenylethylthiazolylthiourea (PETT) derivatives was subjected to quantitative structure–activity relationship (QSAR) analysis to find the structural requirements for ligand binding. The structural invariants used in this study were those obtained from whole molecular structures: chemical, quantum, topological, geometrical, constitutional and functional groups. Four chemometrics methods including multiple linear regressions (MLRs), factor analysis-MLR (FA-MLR), principal component regression analysis (PCRA) and partial least squares combined with genetic algorithm for variable selection (GA-PLS) were employed to make connections between structural parameters and enzyme inhibition. Using the pool of all types of calculated descriptors a QSAR model was derived for selected calibration set compounds indicating the importance of geometrical and chemical parameters on the Human Immunodeficiency Virus Type-1 (HIV-1) reverse transcriptase inhibitory activity. The results of FA-MLR analysis revealed the effects of geometrical and chemical indices on the inhibitory activity too. GA-PLS analysis showed the constitutional and geometrical indices to be the most significant parameters on inhibitory activity. A comparison between the different statistical methods employed indicated that PCRA represented superior results and it could explain and predict 74% and 79% of variances in the  $pIC_{50}$  data, respectively.

© 2009 Elsevier Inc. All rights reserved.

## 1. Introduction

HIV-1 (Human Immunodeficiency Virus Type-1), the main etiological agent for the transmission of acquired immunodeficiency syndrome (AIDS) is a retrovirus of the lentivirus family and contains a reverse transcriptase (RT) enzyme that makes a DNA copy of the viral RNA template [1,2]. Looking at the anti-HIV drugs in the market one can see that the most FDA approved anti-HIV drugs belong to one of the two classes of HIV-1 RT inhibitors: nucleoside reverse transcriptase inhibitors (NRTIs) and non-nucleoside reverse transcriptase inhibitors (NNRTIs). AZT, 3TC, ddI and ddC are typical examples of NRTIs and nevirapine, delavirdine, and efavirenz are NNRTIs that have been approved by FDA. The non-nucleoside HIV-1 RT inhibitors under investigation are structurally different entities: phenylethylthiazolylthiourea (PETT), tetrahydro-imidazo[4,5,1-jk][1,4]-benzodiazepin-2(1H)-one and -thione (TIBO), 1-(2-hydroxyethoxymethyl)-6-(phenylthio)-thymine (HEPT), quinoxalines (with efavirenz in the market),  $\alpha$ -anilino-phenylacetamide ( $\alpha$ -APA), 2',5'-bis(O-(tert-

butyldimethylsilyl)-3'-spiro-5''-(4''-amino-1'',2''-oxathiole-2'',2''-dioxide) pyrimidine (TSAO), bis (heteroaryl) piperazine derivatives (BHAP) (with delavirdine approved by FDA), and dipyrindodiazepinone (with nevirapine in the market) derivatives [3–10].

Non-nucleoside inhibitors are recommended as the first-line therapy for the treatment of HIV-1-infection [11,12]. In fact, after the attachment of the virus to CD4 and chemokine receptors on the host cells, copies of the RNA genome are released into the cytoplasm [13,14]. Both RNA and DNA can be used as a template for DNA synthesis by HIV-1 reverse transcriptase [15,16]. The synthesized DNA is then integrated into the host cell genome by another viral enzyme called HIV integrase [17,18]. The key function of HIV-1 RT in viral cell cycle and replication makes it a major target for drug development [11,19]. Unfortunately rapid development of resistant mutations has diminished the effectiveness of current approved NNRTIs in the market and put an urgent need for a second generation of NNRTIs in the clinic [20,21].

There are a large number of literature reports on the application of computational methods for describing the activity of biologically active compounds [22–26]. Quantitative structure–activity relationship (QSAR) studies are the most extensively used methods in computational chemistry. Appropriate representation of the structural and physicochemical features of chemical agents is an essential key to the successful application of QSAR models [27–31]. QSAR studies play a fundamental role in predicting the biological

\* Corresponding author at: Department of Medicinal Chemistry, Faculty of Pharmacy, Isfahan University of Medical Sciences, 81746-73461, Isfahan, Iran. Tel.: +98 311 7922562; fax: +98 311 6680011.

E-mail address: [fassihi@pharm.mui.ac.ir](mailto:fassihi@pharm.mui.ac.ir) (A. Fassihi).

activity of new compounds and identifying ligand–receptor interactions [32–37]. The first step in constructing the QSAR models is finding one or more molecular descriptors that represent variation in the structural property of the molecules by a number [38]. Structural descriptors have been classified into different categories according to different approaches including physiochemical, constitutional, geometrical, topological, and quantum chemical descriptors. Currently, more than 1000 molecular descriptors can be easily calculated using available softwares such as Dragon [39].

There are different variable selection methods available including multiple linear regression (MLR), genetic algorithm (GA), principal component or factor analysis (PCA/FA) and so on. The mathematical relationships between molecular descriptors and activity are used to find the parameters affecting the biological activity and/or estimate the property of other molecules.

HIV-1 RT inhibitor compounds have been the subject of QSAR studies to find statistical models describing the relationship between the structure and biological activity. Among these inhibitors PETT compounds have been also studied. Ravichandran and co-workers described QSAR study on this class by using COMFA and COMSIA approach [40,41]. Their results showed that steric and electrostatic properties predicted by CoMFA contours and the steric, electrostatic, hydrophobic, and hydrogen-bond acceptor and donor properties predicted by CoMSIA contours are related to anti-HIV-1 RT activity. In another study this research group had shown that the molecular weight, valence connectivity index and critical pressure play important roles in the HIV-1 RT inhibitory activities [42].

In the present paper we used the structural invariants obtained from whole molecular structures of a series of 61 derivatives of the PETT class of HIV-1 RT inhibitors. We exploited four different chemometrics methods to make connections between structural

parameters and HIV-1 RT inhibition. These methods included multiple linear regressions, factor analysis-MLR (FA-MLR), principal component regression analysis (PCRA) and partial least squares combined with genetic algorithm for variable selection (GA-PLS).

## 2. Data and methodology

### 2.1. Equipment

A Pentium IV personal computer (CPU at 3.06 GHz) with windows XP operating system was used. The two-dimensional structures of molecules were drawn using Hyperchem 7.0 software. The final geometries were obtained with the semi-empirical AM1 method in Hyperchem program. The molecular structures were optimized using Polak–Ribiere algorithm until the root mean square gradient was  $0.01 \text{ kcal mol}^{-1}$ . The resulted geometry was transferred into Dragon program, which was developed by Milano Chemometrics and QSAR Group [39]. Z-matrices of the structures were provided by the Hyperchem software and transferred to Gaussian 98 program. Complete geometry optimization was performed taking the most extended conformation as starting geometries. Semi-empirical molecular orbital calculation (AM1) of the structures was preformed using the Gaussian 98 program [43].

### 2.2. Activity data and descriptor generation

The biological data used in this study were HIV-1 reverse transcriptase inhibitory activity (in terms of  $-\log \text{IC}_{50}$ ) of a set of 61 PETT derivatives [44,45]. The data set was already used by Ravichandran and co-workers for QSAR studies [40–42]. The structural features and biological activity of these compounds are listed in Table 1 and then used for subsequent QSAR analysis as

**Table 1**

Chemical structures of PETT analogues used in this study and their experimental and predicted activity for HIV-1 reverse transcriptase inhibition.

Chemical structures of PETT analogues are shown. The structures are labeled 1-24, 25-35, and 36-41. The structures show the general scaffold of the PETT analogues, with R<sub>1</sub> and R<sub>2</sub> representing substituents.

Compound	R <sub>1</sub>	R <sub>2</sub>	Experimental pIC <sub>50</sub> <sup>a</sup>	Predicted pIC <sub>50</sub>			
				MLR	PLS	FA-MLR	PCR
1 <sup>b</sup>	Phenyl	–	0.046	1.030	0.642	0.575	0.396
2	2-Fluorophenyl	–	1.222	1.042	0.995	1.109	0.852
3	3-Fluorophenyl	–	0.824	1.078	0.978	1.109	0.805
4	2-Methoxyphenyl	–	1.398	1.094	0.984	0.945	0.811
5	3-Methoxyphenyl	–	0.824	0.847	0.786	0.945	0.728
6	4-Methoxyphenyl	–	0.455	0.745	0.751	0.942	0.689
7	2-Methylphenyl	–	1.096	1.191	0.807	0.969	0.735
8 <sup>b</sup>	2-Nitrophenyl	–	0.824	0.772	0.971	0.987	0.740
9	2-Hydroxyphenyl	–	–0.041	0.663	0.858	1.109	0.653
10	2-Chlorophenyl	–	0.222	1.088	0.892	1.109	0.680
11	3-Ethoxyphenyl	–	1.221	0.979	0.769	0.897	0.728
12	3-Propoxyphenyl	–	0.698	1.079	0.751	0.848	0.636
13	3-Isopropoxyphenyl	–	0.398	1.068	0.626	0.856	0.678
14	3-Phenoxyphenyl	–	–0.041	0.629	0.078	0.366	0.586

Table 1 (Continued)

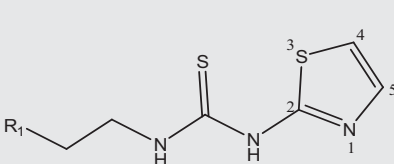
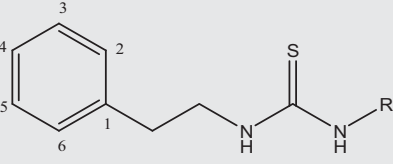
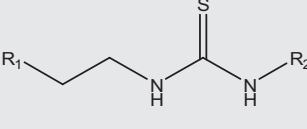
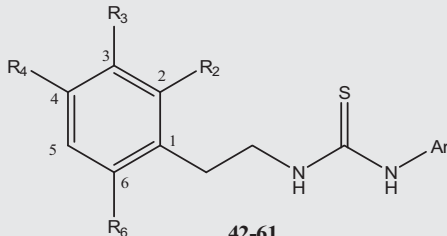
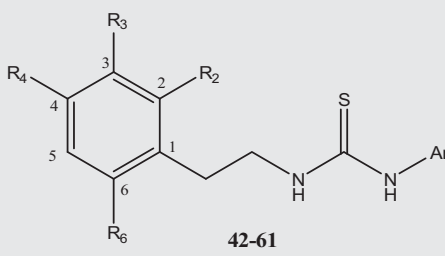
<div><div><p>1-24</p></div><div><p>25-35</p></div><div><p>36-41</p></div></div>										
Compound	R <sub>1</sub>	R <sub>2</sub>	Experimental pIC <sub>50</sub> <sup>a</sup>	Predicted pIC <sub>50</sub>						
				MLR	PLS	FA-MLR	PCR			
15	2,6-Dimethoxyphenyl	–	1.046	0.983	1.029	0.904	0.879			
16	2,5-Dimethoxyphenyl	–	0.699	0.831	1.077	1.115	0.818			
17	3-Bromo-6-methoxyphenyl	–	1.522	1.067	1.449	1.233	1.185			
18 <sup>b</sup>	2-Fluoro-6-methoxyphenyl	–	2.000	1.133	1.100	1.279	0.882			
19	2-Ethoxy-6-fluorophenyl	–	2.000	1.227	1.310	1.185	1.011			
20	2,6-Difluorophenyl	–	2.221	1.137	1.314	1.418	1.132			
21	2-Chloro-6-fluorophenyl	–	0.698	1.159	1.233	1.421	0.911			
22	2-Pyridyl	–	–0.279	0.903	0.758	0.342	0.528			
23	3-Pyridyl	–	0.187	0.814	0.599	0.342	0.548			
24	2-Furyl	–	1.000	0.601	0.430	0.414	0.329			
25	4-Methylthiazol-2-yl	–	0.221	0.080	0.337	0.112	0.348			
26	4-Ethylthiazol-2-yl	–	0.455	0.166	0.385	0.088	0.347			
27 <sup>b</sup>	4-Propylthiazol-2-yl	–	0.698	0.259	0.661	0.270	0.391			
28	4-Isopropylthiazol-2-yl	–	–0.398	0.339	–0.081	0.738	0.419			
29	4-Butylthiazol-2-yl	–	0.698	0.312	0.388	–0.030	0.352			
30	4-Cyanothiazol-2-yl	–	0.259	–0.663	–0.282	–0.467	0.252			
31 <sup>b</sup>	4-(Trifluoro methyl)thiazol-2-yl	–	0.698	0.019	0.703	0.682	0.588			
32	4-(Ethoxy carbonyl)thiazol-2-yl	–	–0.38	0.056	–0.003	0.068	0.267			
33	5-Chlorothiazol-2-yl	–	–0.278	–0.064	0.227	0.191	0.305			
34	5-Bromo-2-pyridyl	–	1.823	0.952	1.276	1.036	1.143			
35	5-Methyl-2-pyridyl	–	1.522	1.079	0.948	0.900	0.851			
36	2,6-Difluorophenyl	4-cyano thiazole-2-yl	0.698	0.656	0.553	1.093	0.872			
37	2,6-Difluorophenyl	5-bromo-2-pyridyl	2.259	2.312	2.218	2.027	2.240			
38 <sup>b</sup>	2,6-Difluorophenyl	5-methyl-2-pyridyl	2.222	2.403	1.867	1.772	1.880			
39	2-Ethoxy-6-fluorophenyl	5-bromo-2-pyridyl	2.522	2.340	2.288	2.073	2.172			
40	2-Pyridyl	5-bromo-2-pyridyl	2.221	2.082	1.933	1.614	1.944			
41	2,6-Difluorophenyl	4-ethylthiazol-2-yl	1.301	1.497	1.178	1.234	0.921			
<div><p>42-61</p></div>										
Compound	R <sub>2</sub>	R <sub>3</sub>	R <sub>4</sub>	R <sub>6</sub>	Ar	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>			
							MLR	PLS	FA-MLR	PCR
42	F	(CO)N(Me) <sub>2</sub>	H	F	5-Bromo-2-pyridyl	1.823	2.464	2.438	1.842	2.365
43	F	CN	H	F	5-Chloro-2-pyridyl	1.638	1.608	1.665	2.024	2.068
44 <sup>b</sup>	F	N(Me) <sub>2</sub>	H	F	5-Chloro-2-pyridyl	1.346	2.289	2.143	1.704	1.891
45	F	N(Me) <sub>2</sub>	H	F	5-Bromo-2-pyridyl	2.045	2.293	2.264	1.842	2.315
46	F	OCH <sub>3</sub>	H	F	5-Bromo-2-pyridyl	2.096	1.977	2.310	2.169	2.277
47	F	OC <sub>2</sub> H <sub>5</sub>	H	F	5-Bromo-2-pyridyl	2.154	2.115	2.304	2.169	2.255
48 <sup>b</sup>	F	CH <sub>2</sub> OCH <sub>3</sub>	H	F	5-Bromo-2-pyridyl	2.221	2.043	2.251	2.062	2.146
49	Cl	OC <sub>2</sub> H <sub>5</sub>	H	F	5-Bromo-2-pyridyl	2.397	2.099	2.186	2.169	2.208
50	Cl	OC <sub>2</sub> H <sub>5</sub>	H	F	5-Chloro-2-pyridyl	2.397	2.076	2.056	2.169	2.089
51	Cl	OC <sub>2</sub> H <sub>5</sub>	H	F	5-iodo-2-pyridyl	1.921	2.087	1.916	2.169	2.050
52	Cl	OC <sub>2</sub> H <sub>5</sub>	H	F	5-cyano-2-pyridyl	2.221	1.441	1.463	1.978	2.042
53 <sup>b</sup>	H	OCH <sub>3</sub>	H	OCH <sub>3</sub>	5-Chloro-2-pyridyl	2.096	1.848	2.094	1.903	1.761
54	H	OC <sub>2</sub> H <sub>5</sub>	H	OC <sub>2</sub> H <sub>5</sub>	5-Bromo-2-pyridyl	2.301	1.934	1.779	1.744	2.050

Table 1 (Continued)

 42-61										
Compound	R <sub>2</sub>	R <sub>3</sub>	R <sub>4</sub>	R <sub>6</sub>	Ar	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>			
							MLR	PLS	FA-MLR	PCR
55	F	H	H	OC <sub>2</sub> H <sub>5</sub>	5-Bromo-2-pyridyl	2.301	2.073	2.115	2.072	2.137
56	F	F	H	OC <sub>2</sub> H <sub>5</sub>	5-Bromo-2-pyridyl	1.745	2.106	2.358	2.169	2.224
57	F	F	H	OCH <sub>3</sub>	5-Bromo-2-pyridyl	2.221	2.026	2.396	2.169	2.301
58	F	OCH <sub>3</sub>	H	OCH <sub>3</sub>	5-Chloro-2-pyridyl	2.301	1.842	2.218	2.098	2.124
59 <sup>b</sup>	F	OC <sub>2</sub> H <sub>5</sub>	H	OCH <sub>3</sub>	5-chloro-2-pyridyl	2.522	1.917	2.249	1.965	1.894
60	F	CN	H	F	5-Bromo-2-pyridyl	1.698	1.568	1.714	2.024	2.195
61	Cl	OC <sub>2</sub> H <sub>5</sub>	H	F	5-Bromo-2-pyridyl	1.677	2.151	2.232	2.169	2.177

<sup>a</sup> pIC<sub>50</sub> = -log IC<sub>50</sub>.<sup>b</sup> Compounds used as prediction set.

dependent variable. The large number of molecular descriptors was calculated using the Hyperchem, Dragon and Gaussian packages. Some chemical parameters including molecular volume (V), molecular surface area (SA), hydrophobicity (log *p*), hydration energy (HE) and molecular polarizability (MP) were calculated using the Hyperchem Software. The Dragon software calculated different functional groups, topological, geometrical and constitutional descriptors for each molecule. The Gaussian program was employed for calculation of different quantum chemical descriptors including, dipole moment (DM), local charges, and HOMO and LUMO energies. Hardness ( $\eta$ ), softness (*S*), electronegativity ( $\chi$ ) and electrophilicity ( $\omega$ ) were calculated according to the method proposed by Thanikaivelan et al. [46]. The calculated descriptors from whole molecular structures are briefly described in Table 2.

### 2.3. Data screening and model building

For each class of the calculated descriptors (*i.e.* physicochemical, quantum, chemical, topological, constitutional and functional groups) separate QSAR models were constructed. For the development of QSAR equations, four different methods were used: (1) stepwise-multiple linear regression (2) MLR with factor analysis as

the data pre-processing step for variable selection (FA-MLR), (3) principal component regression analysis and (4) genetic algorithm-partial least squares (GA-PLS).

The selection of significant descriptors, which constructs a relationship between the biological activity data and the molecular structures, is an important step in QSAR modeling. Selection of significant descriptors was performed through the following steps:

- The calculated descriptors were collected in a data matrix, **D** whose number of rows and columns were the number of molecules and descriptors, respectively. First the descriptors were checked for constant or near constant values and those detected were removed from the original data matrix. The correlation of descriptors with each others and with the activity data was determined.
- The input variable in MLR must not be highly correlated. Among the collinear descriptors detected ( $r > 0.9$ ) one with the highest correlation with the activity was retained and the rest were omitted.
- The selected descriptors from each class and the experimentally anti-HIV data were analyzed by the stepwise regression SPSS (version 12.0) software.

Table 2

Brief description of some descriptors used in this study.

Descriptor type	Molecular description
Constitutional	Molecular weight, no. of atoms, no. of non-H atoms, no. of bonds, no. of heteroatoms, no. of multiple bonds (nBM), no. of aromatic bonds, no. of functional groups (hydroxyl, amine, aldehyde, carbonyl, nitro, nitroso, etc.), no. of rings, no. of circuits, no. of H-bond donors, no. of H-bond acceptors, no. of nitrogen atoms (nN), chemical composition, sum of Kier–Hall electrotopological states (Ss), mean atomic polarizability (Mp), number of rotatable bonds (RBN), mean atomic Sanderson electronegativity (Me), etc.
Topological	Molecular size index, molecular connectivity indices (X1A, X4A, X2v, X1Av, X2Av, X3Av, X4Av), information content index (IC), Kier Shape indices, total walk count, path/walk-Randic shape indices (PW3, PW4, Zagreb indices, Schultz indices, Balaban J index (such as MSD) Wiener indices, topological charge indices, Sum of topological distances between F...F (T(F...F)), Ratio of multiple path count to path counts (PCR), Mean information content vertex degree magnitude (IVDM), Eigenvalue sum of Z weighted distance matrix (SeigZ), reciprocal hyper-detour index (Rww), Eigenvalue coefficient sum from adjacency matrix (VEA1), radial centric information index, 2D Petijean shape index (PJ2), etc.
Geometrical	3D Petijean shape index (PJ3), Gravitational index, Balaban index, Wiener index, etc.
Quantum	Highest occupied molecular orbital energy (HOMO), lowest unoccupied molecular orbital energy (LUMO), most positive charge (MPC), least negative charge (LNC), sum of squares of charges (SSC), sum of square of positive charges (SSPC), sum of square of negative charges (SSNC), sum of positive charges (SUMPC), sum of negative charges (SUMNC), sum of absolute of charges (SAC), total dipole moment (DM <sub>t</sub> ), molecular dipole moment at X-direction (DM <sub>x</sub> ), molecular dipole moment at Y-direction (DM <sub>y</sub> ), molecular dipole moment at Z-direction (DM <sub>z</sub> ), electronegativity ( $\chi = -0.5$ (HOMO–LUMO)), Electrophilicity ( $\omega = (\chi^2/2)\eta$ ), hardness ( $\eta = 0.5$ (HOMO + LUMO)), softness ( $S = 1/\eta$ ).
Functional groups	Number of total tertiary carbons (nCT), number of H-bond acceptor atoms (nHAcc), number of total hydroxyl groups (nOH), number of unsubstituted aromatic C(nCaH), number of ethers (aromatic) (nRORPh), etc.
Chemical	log <i>p</i> (octanol–water partition coefficient), hydration energy (HE), polarizability (Pol), molar refractivity (MR), molecular volume (V), molecular surface area (SA).

For the development of QSAR equations, stepwise-MLR method was used.

In present study, MLR with stepwise selection and elimination of variables was applied for developing QSAR models using SPSS software (SPSS Inc., version 12.0). The resulted models were validated by leave-one out cross-validation procedure (using MATLAB software) to check their predictability and robustness. Early theoretical studies by Topliss and co-workers [47,48] and recent studies by Livingstone and Salt [49,50] indicated that by increasing the number of original descriptors with respect to the number of molecules, the probability of obtaining chance models is increased, even by using variable selection methods. To decrease the probability of getting chance models, the number of original descriptors, among which the best subset of descriptors are selected, should be kept lower than five times of the number of molecules. Therefore, for a data set with limited number of molecules, the number of original descriptors should be decreased dramatically before running variable selection. For solving this problem we first grouped descriptors into various subsets of reasonable size and then variable selection was run on each subset. The selected descriptors of each group were then collected and used as new source of descriptors for variable selection.

FA-MLR was performed on the dataset. Factor analysis was used to reduce the number of variables and to detect structure in the relationships between them. This data-processing step is applied to identify the important predictor variables and to avoid collinearities among them [51]. Principal component regression analysis, was also tried for the dataset along with FA-MLR. With PCRA collinearities among **X** variables are not a disturbing factor and the number of variables included in the analysis may exceed the number of observations [52]. In this method, factor scores, as obtained from FA, are used as the predictor variables [51]. In PCRA, all descriptors are assumed to be important while the aim of factor analysis is to identify relevant descriptors.

In this study, to model the structure–anti-HIV activity relationships better, genetic algorithm-partial least squares (GA-PLS) was employed [53,54]. Partial least squares (PLS) linear regression is a recent technique that generalizes and combines features from principal component analysis and multiple regressions. PLS is a method suitable for overcoming the problems in MLR related to multicollinear or over-abundant descriptors [55]. Application of PLS method thus allows the construction of larger QSAR equations while still avoiding over-fitting and eliminating most variables. This method is normally used in combination with cross-validation to obtain the optimum number of components [26,56]. The PLS regression method used was the NIPALS-based algorithm existed in the chemometrics toolbox of MATLAB software (version 7.1 Math Work Inc.). In order to obtain the optimum number of factors based on the Haaland and Thomas *F*-ratio criterion, leave-one out cross-validation procedure was used [57].

### 3. Results and discussion

Firstly, separate stepwise selection-based MLR analyses were performed using different types of descriptors, and then, an MLR equation was obtained utilizing the pool of all calculated descriptors. The results are summarized in Tables 3–5 for different classes of molecules and in Table 6 for the whole studied molecules. After this, the results obtained by FA-MLR, PCR and GA-PLS analysis of data will be discussed.

#### 3.1. MLR models for subset of molecules

Table 3 provides the derived equations for the molecules 1–24. In this series the quantum parameters did not represent a significant impact on the biological activity. The first equation of Table 3 ( $E_1$ ), obtained from the pool of chemical descriptors, explains that hydration energy (HE) has a positive effect whereas

**Table 3**  
The result of MLR analysis with different type of descriptors for molecules 1–24 ( $n = 24$ ).

No.	Descriptor source	MLR equations	$r^2$	S.E. <sup>a</sup>	RMS <sub>CV</sub> <sup>b</sup>	$q^2$	$F^c$
E <sub>1</sub>	Chemical	$pIC_{50} = 4.534 (\pm 0.963) + 0.358 (\pm 0.095)HE - 0.283 (\pm 0.125) \log p$	0.41	0.53	0.63	0.20	7.50
E <sub>2</sub>	Constitutional	$pIC_{50} = -2.270 (\pm 0.904) + 0.978 (\pm 0.172)nF + 15.328 (\pm 4.062)RBF$	0.63	0.42	0.44	0.55	18.08
E <sub>3</sub>	Topological	$pIC_{50} = -14.137 (\pm 4.438) - 0.915 (\pm 0.166)SEigv + 35.87 (\pm 11.600)X1Av$	0.60	0.44	0.48	0.50	15.86
E <sub>4</sub>	Geometrical	$pIC_{50} = -0.338 (\pm 0.338) + 0.081 (\pm 0.022)DELS$	0.38	0.54	0.57	0.27	13.68
E <sub>5</sub>	Functional groups	$pIC_{50} = 0.350 (\pm 0.840) - 0.517 (\pm 0.194)nHDon + 0.392 (\pm 0.163)nHAcc$	0.40	0.54	0.61	0.23	7.08
E <sub>6</sub>	Molecular descriptor	$pIC_{50} = 1.908 (\pm 0.748) - 0.506 (\pm 0.114)SEigv + 0.237 (\pm 0.070)HE$	0.62	0.43	0.47	0.51	17.47

<sup>a</sup> S.E. = Standard error of regression.

<sup>b</sup> RMS<sub>CV</sub> = Root mean square of cross-validation.

<sup>c</sup> *F* = Fisher ratio.

**Table 4**  
The result of MLR analysis with different type of descriptors for molecules 25–35 ( $n = 11$ ).

No.	Descriptor source	MLR equations	$r^2$	S.E.	RMS <sub>CV</sub>	$q^2$	<i>F</i>
E <sub>7</sub>	Constitutional	$pIC_{50} = 1.672 (\pm 0.314) - 1.453 (\pm 0.347)nBnz$	0.66	0.44	0.48	0.55	17.50
E <sub>8</sub>	Topological	$pIC_{50} = -7.78 (\pm 2.111) + 2.281 (\pm 0.581)VEA1$	0.63	0.46	0.50	0.51	15.41
E <sub>9</sub>	Geometrical	$pIC_{50} = -0.069 (\pm 0.277) + 0.067 (\pm 0.026)G(N \cdots N)$	0.42	0.57	0.70	0.13	6.58
E <sub>10</sub>	Functional groups	$pIC_{50} = -1.234 (\pm 0.432) + 1.453 (\pm 0.347)nCaH$	0.66	0.44	0.48	0.55	17.50
E <sub>11</sub>	Molecular descriptor	$pIC_{50} = -5.42 (\pm 2.329) - 2.020 (\pm 0.302)nBnz + 3.131 (\pm 0.979)J3D$	0.85	0.31	0.38	0.73	22.83

**Table 5**  
The result of MLR analysis with different type of descriptors for molecules 42–61 ( $n = 20$ ).

No.	Descriptor source	MLR equations	$r^2$	S.E.	RMS <sub>CV</sub>	$q^2$	<i>F</i>
E <sub>12</sub>	Constitutional	$pIC_{50} = 1.744 (\pm 0.102) + 0.312 (\pm 0.086)nO$	0.42	0.24	0.26	0.30	13.10
E <sub>13</sub>	Topological	$pIC_{50} = 1.730 (\pm 0.103) + 0.014 (\pm 0.004)T(N \cdots O)$	0.43	0.24	0.25	0.30	13.10
E <sub>14</sub>	Geometrical	$pIC_{50} = 1.734 (\pm 0.100) + 0.039 (\pm 0.010)G(O \cdots S)$	0.44	0.23	0.25	0.32	14.35
E <sub>15</sub>	Functional groups	$pIC_{50} = 1.767 (\pm 0.092) + 0.305 (\pm 0.079)nROR$	0.45	0.23	0.25	0.32	14.74
E <sub>16</sub>	Molecular descriptor	$pIC_{50} = 1.767 (\pm 0.092) + 0.305 (\pm 0.079)nROR$	0.45	0.23	0.25	0.32	14.74



**Table 6**

The result of MLR analysis with different type of descriptors for molecules 1–61 (51 molecules as calibration and 10 molecules as prediction sets).

No.	Descriptor source	MLR equations	$r^2$	S.E.	RMS <sub>CV</sub>	$q^2$	F
E <sub>17</sub>	Chemical	$\text{pIC}_{50} = -0.021 (\pm 0.849) + 0.008 (\pm 0.002)\text{Mass} - 0.421 (\pm 0.123) \log p + 0.111 (\pm 0.049)\text{HE}$	0.60	0.58	0.61	0.52	22.48
E <sub>18</sub>	Quantum	$\text{pIC}_{50} = 1.884 (\pm 2.692) - 35.373 (\pm 6.599)\text{STD} - 1.588 (\pm 0.449)\text{softness} - 2.153 (\pm 0.901)\text{SQNC}$	0.60	0.58	0.65	0.45	22.10
E <sub>19</sub>	Constitutional	$\text{pIC}_{50} = 5.952 (\pm 1.161) - 1.512 (\pm 0.236)\text{nS} - 0.369 (\pm 0.136)\text{nBM} + 0.238 (\pm 0.116)\text{nF}$	0.66	0.53	0.55	0.60	30.07
E <sub>20</sub>	Topological	$\text{pIC}_{50} = -13.734 (\pm 4.720) - 0.012 (\pm 0.003)\text{D/DR05} + 2.731 (\pm 0.776)\text{IVDE} + 11.668 (\pm 4.464)\text{BIC5}$	0.60	0.56	0.55	0.60	29.81
E <sub>21</sub>	Geometrical	$\text{pIC}_{50} = -0.744 (\pm 1.003) - 0.302 (\pm 0.057)\text{G}(\text{S} \cdots \text{S}) + 0.923 (\pm 0.363)\text{J3D} + 0.037 (\pm 0.018)\text{G}(\text{S} \cdots \text{Br})$	0.70	0.50	0.52	0.63	34.31
E <sub>22</sub>	Functional groups	$\text{pIC}_{50} = 0.767 (\pm 0.592) - 0.538 (\pm 0.208)\text{nCaR} + 0.668 (\pm 0.315)\text{nNHR}$	0.64	0.54	0.56	0.58	41.31
E <sub>23</sub>	Molecular descriptor	$\text{pIC}_{50} = 3.312 (\pm 0.371) - 0.420 (\pm 0.044)\text{G}(\text{S} \cdots \text{S}) + 0.144 (\pm 0.039)\text{HE}$	0.70	0.50	0.25	0.65	54.49

$\log p$  shows a negative effect on the HIV-1 reverse transcriptase inhibitory activity of the compounds. Decreasing lipophilicity facilitates both the transport of molecule from the hydrophobic membrane and hydrophobic interaction with receptor. The second equation was obtained from the constitutional descriptors (E<sub>2</sub>), which explains the positive effect of the number of fluorine atoms (nF) and rotatable bond fraction (RBF) on HIV-1 reverse transcriptase inhibitory activity of compounds 1–24. It has a high statistical quality and can explain more than 60% of variances in the studied biological activity.

The equation E<sub>3</sub> shows that among the topological descriptors Eigenvalue sum from van der Waals weighted distance matrix (SEigv) has a negative effect and average valence connectivity index chi-1 (X1Av) has a positive effect on HIV-1 reverse transcriptase inhibitory activity. The usefulness of the topological indices in a wide variety of QSAR studies has been indicated in many literatures [58].

Equation E<sub>4</sub> of Table 3 demonstrates the effect of the geometrical descriptors. It includes the positive effect of molecular electrotopological variation (DELS) on HIV-1 reverse transcriptase inhibitory activity. However, this equation is not so significant and can only explain 38% of variances in the activity data. The MLR equation of Table 3, E<sub>5</sub>, was obtained from the pool of the functional groups descriptors and explained the negative effect of number of donor atoms for H-bonds (with N and O) (nHDon) and positive effect of number of acceptor atoms for H-bonds (N, O, F) (nHAcc) on the HIV-1 reverse transcriptase inhibitory activity of the compounds. The last Equation E<sub>6</sub> was obtained from the all types of calculated descriptors. Stepwise selection and elimination of variables produced a two-parametric QSAR equation. This equation shows that the topological (SEigv) and the chemical (HE) parameters are major factors that affect the biological activity of compounds. It can explain and predict more than 60% of variances in the biological activity data.

In Table 4 the resulted equations for the molecules 25–35 are provided. In this series the chemical and quantum parameters did not represent significant impact on the biological activity. The univariate QSAR model obtained from the constitutional descriptors, E<sub>7</sub>, has number of benzene-like rings (nBnz) as input variable. It has the negative effect of inhibitory activity of compounds. The second equation of Table 4 was found by using the topological descriptors (E<sub>8</sub>). This equation explained the positive effect of eigenvector coefficient sum from adjacency matrix (VEA1) index on the HIV-1 reverse transcriptase inhibitory activity of compounds 25–35. The equation E<sub>9</sub> of Table 4 was obtained from the pool of the geometrical descriptors. It includes the positive effect of sum of geometrical distances between N  $\cdots$  N (G (N  $\cdots$  N)) on the inhibitory activity. The effect of the functional groups on the HIV-1 reverse transcriptase inhibitory activity of the studied compounds has been described by equation E<sub>10</sub> of Table 4. The positive sign of the coefficient of the nCaH proposed that an increase in the

number of unsubstituted aromatic carbons (sp<sup>2</sup>) resulted in an enhanced activity. The equation E<sub>11</sub> was derived from the pool of all calculated descriptors. It shows the negative effect of the nBnz and positive effect of 3D-Balaban index (J3D) on the inhibitory activity. Looking at the pIC<sub>50</sub> values, it is obvious that compounds with pyridyl moiety (compounds 34, 35) were less potent than those with thiazolyl substituent which is not a benzene-like heteroaryl ring. J3D encodes the compactness of the molecules. Comparison of the pIC<sub>50</sub> data shows that the smaller substituents are much better than bigger ones, e.g. 4-methylthiazole-2-yl (R<sub>1</sub> in compound 25) gives a more potent compound than 4-propylthiazole-2-yl (R<sub>1</sub> in compound 27). Maybe this is due to a smaller room at this part of reverse transcriptase binding site. This two-parametric model can explain and predict 85% and 73% of variances in the biological activity of molecules 25–35.

Since the number of molecules in series 36–41 is small, chance correlation would happen thus QSAR models resulted for these compounds derived by using different sets of descriptors would not be reliable.

For the compounds 42–61 (Table 5) no significant QSAR models were obtained from the source of the chemical and quantum descriptors. Among the constitutional descriptors, the number of Oxygen atoms (nO) appeared in the resulted model (E<sub>12</sub>) and it has a positive effect on the inhibitory activity of this compounds. The equation obtained from the pool of topological parameter, E<sub>13</sub>, shows a positive sign of the coefficient of T (N  $\cdots$  O) (sum of topological distances between N  $\cdots$  O) descriptor. This univariate equation can explain and reproduce 43% and 30% of the variances in the inhibitory activity data. The effect of geometrical descriptors on the HIV-1 reverse transcriptase inhibitory activity of the studied compounds has been described by equation E<sub>14</sub> of Table 5. It explains the positive effect of sum of geometrical distances between O  $\cdots$  S (G (O  $\cdots$  S)) on the HIV-1 reverse transcriptase inhibitory activity. The one-parametric equation E<sub>15</sub> was found by using functional groups descriptors indicating nROR (number of ethers (aliphatic)) as input variable, and could explain only about 45% of variance in the biological activity data. There was a poor relationship between the number of aliphatic ethers and the biological activity of the compounds which was reflected in the pIC<sub>50</sub> of compounds 48 and 49. The last equation, E<sub>16</sub>, was derived from all calculated descriptors. Stepwise selection and elimination of variables produced a one-parametric equation which is similar to what was obtained for E<sub>15</sub>.

### 3.2. MLR models for all molecules

As it was explained in the previous section, for each subset of molecules separate MLR-based QSAR models were obtained from the sources of different types of descriptors. Interestingly, for each type of descriptors, similar QSAR models were obtained for different subsets of molecules. This suggests that it is possible to

generate a single QSAR model for all studied molecules from each group of descriptors. The data set was classified into calibration and prediction set by homogenous sampling of the 10 prediction molecules from the factor spaces of the calculated descriptors. Attention was made to have prediction molecules from all substructures.

The resulted QSAR models from different types of descriptors are listed in Table 6. As it is observed, among the QSAR models obtained from the sources of descriptors calculated from the whole molecular structure, those created from the constitutional and geometrical ( $E_{19}$  and  $E_{21}$ ) represented the highest statistical quality. As well as  $\log p$  and HE that were discussed previously, Mass was also introduced in the QSAR model of chemical indices ( $E_{17}$ ). Among the selected quantum descriptors ( $E_{18}$ ), SQNC was already selected by the equations used for subsets of molecules whereas Softness and STD are new parameters. It reveals that standard deviation of dipole moments in the X, Y and Z directions is a controlling factor for binding of compounds to HIV-1 RT.

The effect of the constitutional descriptors on the HIV-1 reverse transcriptase inhibitory activity of the studied compounds has been described by equation  $E_{19}$  of Table 6. It explained the negative effect of number of multiple bonds (nBM) and number of sulfur atoms (nS) and positive effect of number of fluorine atoms (nF) on the biological activity. The negative sign of the coefficient of the nBM and nS proposed that decreasing the number of multiple bonds of compounds and the number of sulfur atoms, respectively, resulted in activity enhancement. The positive sign of the coefficient of the nF proposed that an increase in the number of fluorine atoms of the compounds, resulted in activity enhancement. It has a good statistical quality for predicting the activity of the inhibitors (i.e.  $r^2 = 0.66$  and  $q^2 = 0.60$ ). The MLR equation of Table 6 found from the pool of the topological descriptors ( $E_{20}$ ) explained the positive effect of mean information content vertex degree equality (IVDE) and bond information content (neighborhood symmetry of 5-order) (BIC5) and the negative effect of distance/detour ring index of order 5 (D/Dr05) on HIV-1 reverse transcriptase inhibitory activity.

Among the groups of descriptors obtained from the whole molecular structures, the geometrical indices resulted in the most significant QSAR model,  $E_{21}$ , for predicting the anti-HIV-1 RT activity of the studied molecules. This model explains the positive effects of 3D-Balaban index (J3D) and sum of geometrical distances between S...Br (G (S...Br)) and negative effect of sum of geometrical distances between S...S (G (S...S)) on the inhibitory activity. It could explain and predict 70% and 63% of the variance in  $pIC_{50}$  data, respectively.

The equation obtained from the effect of the functional groups parameter on the inhibitory activity of the studied compounds ( $E_{22}$ ) shows a negative sign of the effect of number of substituted aromatic  $C(sp^2)$  (nCaR) and positive effect of number of secondary amines (aliphatic) (nNHR) on HIV-1 reverse transcriptase inhibitory activity. This negative sign of nCaR proposed that decreasing the number of substituted aromatic carbons ( $sp^2$ ), resulted in an enhancement in the activity. The positive coefficient of nNHR shows that increasing the number of secondary amines (aliphatic) resulted in activity enhancement. As it was shown in the last row of Table 6, the resulted QSAR model ( $E_{23}$ ) represents high ability (about 70%) to explain and predict the activity of the studied compounds. This model is a combination of chemical (HE) and geometrical (G (S...S)) descriptors. These descriptors are major factors that affect HIV-1 reverse transcriptase inhibitory activity of the compounds.

The predicted values of the activity for calibration set (by cross-validation) and prediction set for MLR analysis are listed in Table 1 and are plotted against the corresponding experimental values in Fig. 1. The statistical parameters of prediction set are listed in

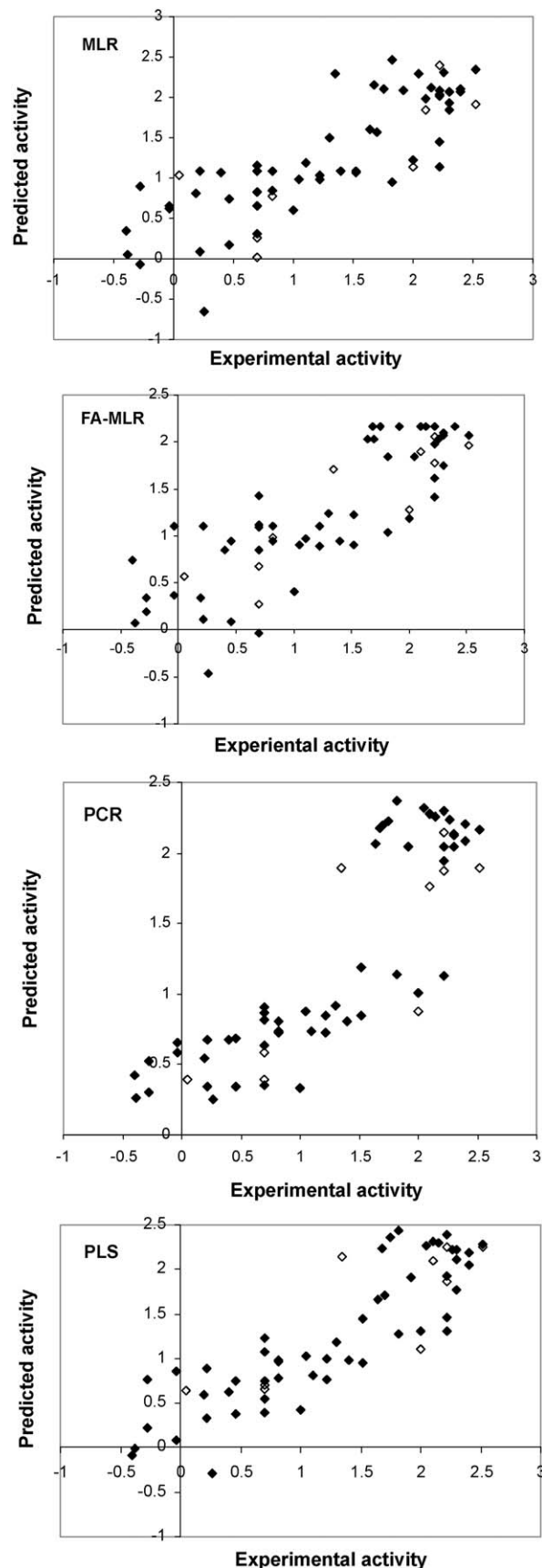


Fig. 1. Plots of the cross-validated predicted activity against the experimental activity for the QSAR models obtained by MLR, GA-PLS, FA-MLR, PCR methods.

**Table 7**

Statistical parameters for testing prediction ability of the MLR, FA-MLR, PCR and GA-PLS models.

Model	$q^2$	RMSE <sub>CV</sub>	$r_p^2$	RMSE <sub>P</sub>
MLR	0.65	0.25	0.61	0.49
FA-MLR	0.65	0.51	0.65	0.53
PCR	0.74	0.45	0.74	0.40
GA-PLS	0.69	0.48	0.69	0.51

**Table 8**

Numerical values of factor loading numbers 1–4 for descriptors after VARIMAX rotation.

	1	2	3	4	Commonality
HE	0.176	−0.195	−0.195	0.041	0.766
log <i>p</i>	−0.337	−0.324	−0.324	−0.087	0.546
MASS	0.717	0.526	0.526	0.329	0.901
STD	−0.234	−0.921	−0.921	−0.166	0.943
SQNC	−0.048	0.175	0.175	0.950	0.936
SOFTNESS	−0.107	0.942	0.942	0.048	0.927
nBM	0.530	0.452	0.452	0.207	0.635
nS	−0.932	−0.238	−0.238	−0.037	0.929
nF	0.646	0.206	0.206	−0.276	0.615
IVDE	0.668	−0.427	−0.427	−0.010	0.773
BIC5	0.047	0.830	0.830	−0.115	0.782
D/DR05	−0.933	−0.184	−0.184	0.009	0.905
J3D	0.380	0.751	0.751	0.410	0.877
G (S··S)	−0.913	−0.076	−0.076	−0.025	0.907
G (S··Br)	0.578	−0.103	−0.103	0.035	0.362
nCaR	−0.908	−0.095	−0.095	−0.064	0.884
nNHR	0.956	0.082	0.082	0.001	0.921
pIC <sub>50</sub>	0.846	0.136	0.136	−0.003	0.740
%Variance	40.980	22.270	8.900	7.540	79.690

**Table 7.** The correlation coefficient of prediction is 0.61, which means that the resulted QSAR model could predict 61% of variances in the anti-HIV activity data. It has a root mean square error of 0.49.

### 3.3. FA-MLR and PCRA

**Table 8** shows the four factor loadings of the variables (after VARIMAX rotation) for the compounds tested against HIV-1 reverse transcriptase. As it is observed, about 79% of variances in the original data matrix could be explained by the selected four factors.

Based on the procedure explained in the experimental section, the following two-parametric equation was derived.

$$pIC_{50} = 3.312(\pm 0.371) - 0.673(\pm 0.970)G(S\cdots S) + 0.259(\pm 0.070)HE; \\ r^2 = 0.70, S.E. = 0.49, F = 54.49 \quad q^2 = 0.65 \quad RMS_{cv} = 0.51, N = 51 \quad (1)$$

Eq. (1) could explain 70% of the variance and predict 65% of the variance in pIC<sub>50</sub> data. This equation describes the effect of geometrical (G (S··S)) and chemical (HE) indices on anti-HIV activity of the studied molecules.

When factor scores were used as the predictor parameters in a multiple regression equation using forward selection method (PCRA), the following equation was obtained:

$$pIC_{50} = 1.240(\pm 0.059) + 0.676(\pm 0.060)f_1 - 0.276(\pm 0.060)f_4 \\ + 0.270(\pm 0.060)f_3; \quad r^2 = 0.79, S.E. = 0.42, \\ F = 56.48, q^2 = 0.74, RMS_{cv} = 0.45, N = 51 \quad (2)$$

Eq. (2) shows also high equation statistics (79% explained variance and 74% predict variance in pIC<sub>50</sub> data). Since factor scores are used instead of selected descriptors, and any factor score contains information from different descriptors, loss of informa-

tion is thus avoided and the quality of PCRA equation is better than those derived from FA-MLR.

As it is observed from **Table 8**, in the case of each factor, the loading values for some descriptors are much higher than those of the others. These high values for each factor indicate that this factor contains higher information about which descriptors. It should be noted that all factors have information from all descriptors but the contribution of descriptor in different factors are not equal. For example, factors 1 and 2 have higher loadings for the chemical, quantum, topological, constitutional geometrical and functional groups indices, whereas information about the quantum, geometrical and topological descriptors is highly incorporated in factors 3 and 4. Therefore, from the factor scores used by equation E<sub>2</sub>, significance of the original variables for modeling the activity can be obtained. Factor score 1 indicates importance of Mass, nS, D/DR05, G (S··S), nCaR and nNHR (the chemical, constitutional, topological, geometrical and functional groups indices). Factor score 2 indicates importance of STD, J3D and BIC5 (the quantum, geometrical and topological descriptors), Factor scores 3 and 4 signify the importance of STD, Softness, SQNC, BIC5 and J3D (the quantum topological and geometrical descriptors).

The predicted values of the activity for calibration set (by cross-validation) and prediction set for FA-MLR and PCRA are listed in **Table 1** and are plotted against the corresponding experimental values in **Fig. 1**. The statistical parameters of prediction set are listed in **Table 7**. The correlation coefficient of prediction for FA-MLR analysis is 0.65, which means that the obtained QSAR model could predict 65% of variances in the anti-HIV-1 RT activity data. It has a root mean square error of 0.53. The correlation coefficient of prediction for PCRA analysis is 0.74. This means that the derived QSAR model could predict 74% of variances in the inhibitory activity data. The root mean square error of PCRA analysis was 0.40. Whilst the data of this analysis show acceptable prediction, we see that the predicted values of some molecules are near to each other. It indicates the suitability of the proposed QSAR model based on PCRA analysis.

### 3.4. GA-PLS model

In PLS analysis, the descriptors data matrix is decomposed to orthogonal matrices with an inner relationship between the dependent and independent variables. Unlike MLR analysis, the multicollinearity problem in the descriptors is omitted by PLS analysis. Because a minimal number of latent variables are used for modeling in PLS; this modeling method coincides with noisy data better than MLR. Since redundant variables degrade the performance of PLS analysis, similar to other regression methods, a variable selection method must be employed to find the more convenient set of descriptors. Here, GA was used as variable selection method. The data set (*n* = 61) was divided into two group: calibration set (*n* = 51) and prediction set (*n* = 10). Given 51 calibration samples; cross-validation procedure was used to find the optimum number of latent variables for each PLS model. GA produces a population of acceptable models in each run. In this work, many different GA-PLS runs were conducted using different initial set of populations (50–250) and therefore a large number of acceptable models were created.

The most convenient GA-PLS model that resulted in the best fitness contained nine descriptors including two chemical parameters (HE and log *p*), one quantum descriptor (SQNC), one constitutional (nS), two topological indices (IVDE and BIC5) and three geometrical (J3D, G (S··S) and G (S··Br)) parameters. These descriptors were already used by different MLR-based QSAR models. The PLS estimate of the regression coefficients are shown in **Fig. 2**. Since these constants were calculated based on the



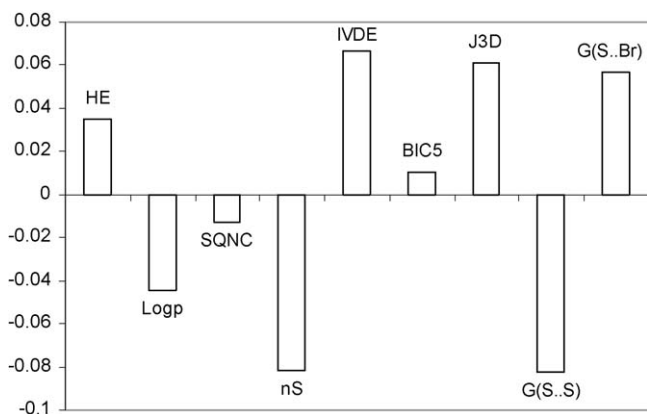


Fig. 2. PLS regression coefficients for the variables used in GA-PLS model.

normalized descriptor values, they can be used as a measure of the importance of the corresponding descriptor. As it is observed, the constitutional (nS) and geometrical indices (G (S..S)) represent the most significant contribution in the obtained QSAR model followed by the geometrical and topological parameters (J3D and IVDE). Interestingly, in the same manner as MLR-based QSAR model, the parameters HE, IVDE, J3D and G (S..Br) have positive and log *p*, SQNC, nS, BIC5, G (S..S) have negative effects.

The statistical parameters of the resulted PLS-based QSAR model are given in Table 7. It could explain and predict about 69% of variances in the anti-HIV-1 RT inhibitory activity of the studied molecules. The correlation coefficient of prediction for GA-PLS analysis is 0.69, which means that the resulted QSAR model could predict 69% of variances in the inhibitory activity data. It has a root mean square error of 0.51. The predicted activities are represented in Table 1 and are plotted against the corresponding experimental values in Fig. 1.

#### 4. Conclusions

Quantitative relationships between molecular structure and HIV-1 reverse transcriptase inhibitory activity of PETT derivatives were discovered by different chemometrics methods. MLR, FA-MLR, PCRA and GA-PLS were employed to make connections between structural parameters and enzyme inhibition. Different QSAR models for molecules 1–24 revealed that constitutional parameters (nF and RBF) had significant impact on HIV-1 reverse transcriptase inhibitory activity of the compounds. In molecules 25–35 significant role of constitutional (nBnz) and geometrical (J3D) parameters on the inhibitory activity was observed. Compounds with the general structure of molecules 25–35 (Table 1) with heteroaryl substituents other than benzene-like ones (such as pyridine) and small compact structures are more favored as potent anti-HIV compounds. For compounds 42–61 no significant QSAR models were obtained from the source of the different descriptors. Using the pool of all types of calculated descriptors a new QSAR model was derived for selected calibration set compounds. In this model the importance of geometrical (G (S..S)) and chemical (HE) parameters on HIV-1 reverse transcriptase inhibitory activity was indicated. The results of FA-MLR analysis showed the effects of geometrical (G (S..S)) and chemical (HE) indices on the inhibitory activity. It was similar to those obtained from MLR analysis for selected calibration set compounds. GA-PLS analysis indicated that the constitutional (nS) and geometrical indices (G (S..S)) were the most significant parameters on HIV-1 RT inhibitory activity. A comparison between the different statistical methods employed revealed that PCRA represented superior results and it could

explain and predict 74% and 79% of variances in the pIC<sub>50</sub> data, respectively.

#### Acknowledgment

This work was supported by Isfahan Pharmaceutical Sciences Research Center.

#### References

- [1] S. Fauci, The human immunodeficiency virus: infectivity and mechanisms of pathogenesis, *Science* 239 (1998) 617–622.
- [2] E. De Clercq, Toward improved anti-HIV chemotherapy: therapeutic strategies for intervention with HIV infections, *J. Med. Chem.* 38 (1995) 2491–2517.
- [3] C. Ahgren, K. Backro, F.W. Bell, A.S. Cantrell, M. Clemens, J.M. Colacino, J.B. Deeter, J.A. Engelhardt, M. Hogberg, S.R. Jaskunas, N.G. Johanssons, The PETT series, a new class of potent nonnucleoside inhibitors of human immunodeficiency virus type 1 reverse transcriptase, *Antimicrob. Agents Chemother.* 39 (1995) 1329–1335.
- [4] R. Pauwels, K. Andries, J. Desmyter, D. Schols, M.J. Kukla, H.J. Breslin, A. Raeymaeckers, J. Van Gelder, R. Woestenborghs, J. Heykants, Potent and selective inhibition of HIV-1 replication in vitro by a novel series of TIBO derivatives, *Nature* 343 (1990) 470–474.
- [5] M. Baba, H. Tanakas, E. De Clercq, R. Pauwels, J. Balzarini, D. Schols, H. Nakashima, C.F. Perno, R.T. Walker, T. Miyasaka, Highly specific inhibition of human immunodeficiency virus type 1 by a novel 6-substituted acycloauridine derivative, *Biochem. Biophys. Res. Commun.* 165 (1989) 1375–1381.
- [6] J.P. Kleim, R. Bender, U.M. Billhardt, C. Meichsner, G. Riess, M. Rosner, I. Winkler, A. Paessens, Activity of a novel quinoxaline derivative against human immunodeficiency virus type 1 reverse transcriptase and viral replication, *Antimicrob. Agents Chemother.* 37 (1993) 1659–1664.
- [7] R. Pauwels, K. Andries, Z. Debyser, P. Van Daele, D. Schols, A.M. Vandamme, C.G.M. Janssen, J. Anne, G. Cauwenbergh, J. Desmyter, J. Heykants, M.A.C. Janssen, E. De Clercq, P.A.J. Janssen, Potent and highly selective human immunodeficiency virus type 1 (HIV-1) inhibition by a series of alpha-anilino-phenylacetamide derivatives targeted at HIV-1 reverse transcriptase, *Proc. Natl. Acad. Sci. U.S.A.* 90 (1993) 1711–1715.
- [8] J. Balzarini, M.J. Pérez-Pérez, A. San-Felix, D. Schols, D.C.F. Perno, A. Vandamme, M.J. Camarasa, E. De Clercq, 2',5'-Bis-O-(tert-butylidimethylsilyl)-3'-spiro-5''-(4''-amino-1'',2''-oxathiole-2'', 2'-dioxide) pyrimidine (TSAO) nucleoside analogues: highly selective inhibitors of human immunodeficiency virus type 1 that are targeted at the viral reverse transcriptase, *Proc. Natl. Acad. Sci. U.S.A.* 89 (1992) 4392–4396.
- [9] D.L. Romero, M. Busso, C.K. Tan, F. Reusser, J.R. Palmer, S.M. Poppe, P.F. Aristoff, K.M. Downey, A.G. So, L. Resnick, W.G. Tarpley, Nonnucleoside reverse transcriptase inhibitors that potently and specifically block human immunodeficiency virus type 1 replication, *Proc. Nat. Acad. Sci.* 88 (1991) 8806–8810.
- [10] V.J. Merluzzi, K.D. Hargrave, M. Labadia, K. Grozinger, M. Skoog, J.C. Wu, C.K. Shih, K. Eckner, S. Hattox, J. Adams, Inhibition of HIV-1 replication by a nonnucleoside reverse transcriptase inhibitor, *Science* 250 (1990) 1411–1413.
- [11] J.P. Moore, M. Stevenson, New targets for inhibitors of HIV-1 replication, *Nat. Rev. Mol. Cell Biol.* 1 (2000) 40–49.
- [12] C.M. Tarby, Recent advances in the development of next generation non-nucleoside reverse transcriptase inhibitors, *Curr. Top. Med. Chem.* 4 (2004) 1045–1057.
- [13] P.R. Clapham, A. McKnight, Cell surface receptors, virus entry and tropism of primate lentiviruses, *J. Gen. Virol.* 83 (2002) 1809–1829.
- [14] J.A. Hoxie, C.C. LaBranche, M.J. Endres, J.D. Turner, J.F. Berson, R.W. Doms, T.J. Matthews, CD-4 independent utilization of the CXCR4 chemokine receptor by HIV-1 and HIV-2, *J. Reprod. Immunol.* 41 (1998) 197–211.
- [15] J. Hurwitz, J.P. Leis, RNA-dependent DNA polymerase activity of RNA tumor viruses I. Directing influence of DNA in the reaction, *J. Virol.* 9 (1972) 116–129.
- [16] J.P. Leis, J. Hurwitz, RNA-dependent DNA polymerase activity of RNA tumor viruses II. Directing influence of RNA in the reaction, *J. Virol.* 9 (1972) 130–142.
- [17] D.J. Hazuda, P. Felock, M. Witmer, A. Wolfe, K. Stillmock, J.A. Grobler, A. Espeseth, L. Gabryelski, W. Schleif, C. Blau, M.D. Miller, Inhibitors of strand transfer that prevent integration and inhibit HIV-1 replication in cells, *Science* 287 (2000) 646–650.
- [18] C. Duda-Seiman, M.V. Putz, D. Ciubotariu, QSAR modeling of anti-HIV activity with HEPT derivatives, *Dig. J. Nanomater. Biostruct.* 2 (2007) 207–219.
- [19] H.C. Castro, N.I. Loureiro, M. Pujol-Luz, A.M. Souza, M.G. Albuquerque, D.O. Santos, L.M. Cabral, I.C. Frugulhetti, C.R. Rodrigues, HIV-1 reverse transcriptase: a therapeutic target in the spotlight, *Curr. Med. Chem.* 13 (2006) 313–324.
- [20] E. De Clercq, The role of non-nucleoside reverse transcriptase inhibitors (NNRTIs) in the therapy of HIV infection, *Antiviral. Res.* 38 (1998) 153–179.
- [21] P.A. Janssen, P.J. Lewi, E. Arnold, F. Daeyaert, M. De Jonge, J. Heeres, L. Koymans, M. Vinkers, J. Guillemont, E. Pasquier, M. Kukla, D. Ludovici, K. Andries, M.P. De Bethune, R. Pauwels, K. Das, A.D. Clark Jr., Y.V. Frenkel, S.H. Hughes, B. Medaer, F. De Knaep, H. Bohets, F. De Clerck, A. Lampo, P. Williams, P. Stoffels, In search of a novel anti-HIV drug: multidisciplinary coordination in the discovery of 4-[[4-[(1E)-2-cyanoethenyl]-2,6-dimethylphenyl]amino]-2-pyrimidinyl]amino benzonitrile (R278474, Rilpivirine), *J. Med. Chem.* 48 (2005) 1901–1909.
- [22] B. Hemmateejad, R. Miri, M. Akhond, M. Shamsipur, QSAR study of the calcium channel antagonist activity of some recently synthesized dihydropyridine

- derivatives. An application of genetic algorithm for variable selection in MLR and PLS methods, *Chemom. Intell. Lab. Syst.* 64 (2002) 91–99.
- [23] B. Hemmateenejad, R. Miri, M. Akhond, M. Shamsipur, Quantitative structure activity relationship study of recently synthesized 1,4-dihydropyridine calcium channel antagonists. Application of Hansch analysis methods, *Arch. Pharm. Pharm. Med. Chem.* 10 (2002) 472–480.
- [24] C. Hansch, D. Hoekman, H. Gao, Comparative QSAR; toward a deeper understanding of chemo-biological interaction, *Chem. Rev.* 96 (1996) 1045–1075.
- [25] A. Fassihi, R. Sabet, QSAR study of p56<sup>lck</sup> protein tyrosine kinase inhibitory activity of flavonoid derivatives using MLR and GA-PLS, *Int. J. Mol. Sci.* 9 (2008) 1876–1892.
- [26] A. Fassihi, D. Abedi, L. Saghaie, R. Sabet, H. Fazeli, Gh. Bostaki, O. Deilami, H. Sadihpour, Synthesis, antimicrobial evaluation and QSAR study of some 3-hydroxypyridine-4-one and 3-hydroxypyran-4-one derivatives, *Eur. J. Med. Chem.* 44 (2009) 2145–2157.
- [27] C. Hansch, T. Fujita, A method for the correlation of biological activity and chemical structure, *J. Am. Chem. Soc.* 86 (1964) 1616–1626.
- [28] J. Wang, L. Zhang, G. Yang, C.G. Zhan, Quantitative structure–activity relationship for cyclic imide derivatives of protoporphyrinogen oxidase inhibitors: a study of quantum chemical descriptors from density functional theory, *J. Chem. Inf. Comput. Sci.* 44 (2004) 2099–2105.
- [29] L.P. Hammett, The effect of structure upon the reactions of organic compounds. Benzene derivatives, *J. Am. Chem. Soc.* 59 (1937) 96–103.
- [30] B. Hemmateenejad, M. Sanchooli, Substituent electronic descriptors for fast QSAR/QSPR, *J. Chemom.* 21 (2007) 96–107.
- [31] R. Todeschini, V. Consonni, *Handbook of Molecular Descriptors*, Wiley-VCH, Weinheim, 2000.
- [32] H. Schmidt, Multivariate prediction for QSAR, *Chemom. Intell. Lab. Syst.* 37 (1997) 125–134.
- [33] C. Hansch, D. Hoekman, H. Gao, Chem-bioinformatics and QSAR: a review of QSAR lacking positive hydrophobic terms, *Chem. Rev.* 101 (2001) 619–627.
- [34] S. Wold, J. Trygg, A. Berglund, H. Antti, Some recent developments in PLS modeling, *Chemom. Intell. Lab. Syst.* 58 (2001) 131–150.
- [35] F.A. Pasha, H.K. Srivastava, H.K. Singh, P.P. Semiempirical, QSAR study and ligand receptor interaction of estrogens, *Mol. Div.* 9 (2005) 215–220.
- [36] O.J. D'Cruz, F.M. Uckun, Dawn of non-nucleoside inhibitor-based anti-HIV microbicides, *J. Antimicrob. Chemother.* 57 (2006) 411–423.
- [37] S. Gayen, B. Debnath, S. Samanta, T. Jha, QSAR study on some anti-HIV HEPT analogues using physicochemical and topological parameters, *Bioorg. Med. Chem.* 12 (2004) 1493–1503.
- [38] S. Agatonovic-Kustrin, I.G. Tucker, M. Zecevic, L.J. Ziva-novic, Prediction of drug transfer into human milk from theoretically derived descriptors, *Anal. Chem. Acta* 418 (2000) 181–195.
- [39] Milano Chemometrics and QSPR Group, Dragon, version 2.1, Milano, Italy, 2002.
- [40] V. Ravichandran, R.K. Agrawal, Predicting anti-HIV activity of PETT derivatives: CoMFA approach, *Bioorg. Med. Chem. Lett.* 17 (2007) 2197–2202.
- [41] V. Ravichandran, B.R.P. Kumar, S. Sankar, R.K. Agrawal, Comparative molecular similarity indices analysis for predicting anti-HIV activity of phenylethylthiourea (PET) derivatives, *Med. Chem. Res.* 17 (2008) 1–11.
- [42] V. Ravichandran, V.K. Mourya, R.K. Agrawal, QSAR modeling of HIV-1 reverse transcriptase inhibitory activity with pett derivatives, *Dig. J. Nanomater. Biostruct.* 3 (2008) 9–17.
- [43] Gaussian, version 98, Revision A.7, Gaussian Inc., Pittsburgh, PA, 1998.
- [44] F.W. Bell, A.S. Cantrell, M. Hogberg, S.R. Jaskunas, N.G. Johansson, C.L. Jordan, M.D. Kinnic, P. Lind, J.M. Morin, R. Noreen, B. Oberg, J.A. Palkowitz, C.A. Parrish, P. Pranc, C. Sahlberg, R.J. Ternansky, R.T. Vasileff, L. Vrang, S.J. West, H. Zhang, X.X. Zhou, Phenethylthiazolethiourea (PETT) compounds, a new class of HIV-1 reverse transcriptase inhibitors. 1. Synthesis and basic structure–activity relationship studies of PETT analogs, *J. Med. Chem.* 38 (1995) 4929–4936.
- [45] A.S. Cantrell, P. Engelhardt, M. Hogberg, S.R. Jaskunas, N.G. Johansson, C.L. Jordan, M.D. Kinnic, P. Lind, J.M. Morin, R. Noreen, B. Oberg, J.A. Palkowitz, C.A. Parrish, P. Pranc, C. Sahlberg, R.J. Ternansky, R.T. Vasileff, L. Vrang, S.J. West, H. Zhang, X.X. Zhou, Phenethylthiazolethiourea (PETT) compounds as a new class of HIV-1 reverse transcriptase inhibitors. 2. Synthesis and further structure–activity relationship studies of PETT analogs, *J. Med. Chem.* 39 (1996) 4261–4274.
- [46] P. Thanikaivelan, V. Subramanian, J.R. Rao, B.U. Nair, Application of quantum chemical descriptor in quantitative structure activity and structure property relationship, *Chem. Phys. Lett.* 323 (2000) 59–70.
- [47] J.G. Topliss, R.J. Costello, Chance correlation structure–activity studies using multiple regression analysis, *J. Med. Chem.* 15 (1972) 1066–1078.
- [48] J.G. Topliss, R.P. Edwards, Chance factors in studies of quantitative structure–activity relationships, *J. Med. Chem.* 22 (1979) 1238–1244.
- [49] D.J. Livingstone, D.W. Salt, Judging the significance of multiple linear regression models, *J. Med. Chem.* 48 (2005) 661–663.
- [50] D.W. Salt, S. Ajmani, R. Crichton, D.J. Livingstone, An improved approximation to the estimation of the critical *F* values in best subset regression, *J. Chem. Inf. Model.* 47 (2007) 143–149.
- [51] R. Franke, A. Gruska, Chemometrics methods in molecular design, in: H. van Waterbeemd (Ed.), *Methods and Principles in Medicinal Chemistry*, vol. 2, VCH Publishers, Weinheim, 1995, pp. 113–119.
- [52] H. Kubinyi, The quantitative analysis of structure–activity relationships, in: M.E. Wolff (Ed.), 5th ed., *Burger's medicinal chemistry and drug discovery*, vol. 1, Wiley, New York, 1995, pp. 506–509.
- [53] R. Leardi, Application of genetic algorithm-PLS for feature selection in spectral data sets, *J. Chemom.* 14 (2000) 643–655.
- [54] R. Leardi, A.L. Gonzalez, Genetic algorithm applied to feature selection in PLS regression: how and when to use them, *Chemom. Intell. Lab. Syst.* 41 (1998) 195–207.
- [55] O. Deeb, B. Hemmateenejad, A. Jaber, R. Garduno-Juarez, R. Miri, Effects of the electronic and physicochemical parameters on the carcinogenesis activity of some sulfa drug using QSAR analysis based on genetic-MLR & genetic-PLS, *Chemosphere* 67 (2007) 2122–2130.
- [56] R. Leardi, Genetic algorithms in chemometrics and chemistry: a review, *J. Chemom.* 15 (2001) 559–569.
- [57] B. Hemmateenejad, Optimal QSAR analysis of the carcinogenic activity of drugs by correlation ranking and genetic algorithm-based, *J. Chemom.* 108 (2004) 475–485.
- [58] A.R. Katritzky, E.V. Gordeeva, Traditional topological indices versus electronic, geometrical and molecular descriptors in QSAR/QSPR research, *J. Chem. Inf. Comput. Sci.* 33 (1993) 835–857.