



Original article

Predictive QSAR modeling of HIV reverse transcriptase inhibitor TIBO derivatives

Asim Sattwa Mandal, Kunal Roy*

Drug Theoretics and Cheminformatics Lab, Division of Medicinal and Pharmaceutical Chemistry, Department of Pharmaceutical Technology, Jadavpur University, Raja S.C. Mullick Road, Kolkata 700 032, West Bengal, India

ARTICLE INFO

Article history:

Received 4 April 2008

Received in revised form 12 July 2008

Accepted 12 July 2008

Available online 24 July 2008

Keywords:

QSAR

Reverse transcriptase inhibitor

TIBO derivatives

Validation

ABSTRACT

Comparative quantitative structure–activity relationship (QSAR) studies have been carried out on tetrahydroimidazo[4,5,1-jk][1,4]benzodiazepine (TIBO) derivatives as reverse transcriptase inhibitors ($n = 70$) using topological, structural, physicochemical, electronic and spatial descriptors. The data set was divided into training and test sets using a cluster-based method. Linear models were developed using multiple regression (with stepwise regression, factor analysis and genetic function approximation (GFA) as variable selection tools) and partial least squares (PLS) and combination of factor analysis and partial least squares (FA–PLS). Genetic function approximation (spline) and artificial neural networks (ANN) were used for the development of non-linear models. Using topological and structural descriptors, the best equation was obtained from GFA (spline) based on internal validation ($Q^2 = 0.737$), but the model with the best external validation characteristics was obtained with FA–PLS ($R_{\text{pred}}^2 = 0.707$). When structural, physicochemical, electronic and spatial descriptors were used, the best Q^2 (0.740) value was obtained from GFA (spline) whereas PLS provided the best R_{pred}^2 (0.784) value. When all descriptors were used in combination, the best R_{pred}^2 (0.760) value and the best Q^2 (0.800) value were obtained from ANN and GFA (spline), respectively. The majority of the models satisfied the criteria of external validation recommended by Golbraikh and Tropsha (2002) and the criteria of modified r^2 (r_m^2) values of the test set for external validation as suggested by Roy and Roy (2008). In order to further validate selected models, an external set of 10 TIBO derivatives, which fall within the applicability domain of the models and are not shared with the compounds of the present data set, was taken from a different source, and reverse transcriptase inhibitory activity of these compounds was predicted. Acceptable values of squared correlation coefficients between the observed and predicted values of the external set compounds were obtained from the selected models suggesting true predictive potential of the models.

© 2008 Elsevier Masson SAS. All rights reserved.

1. Introduction

Acquired immuno-deficiency syndrome (AIDS) is a leading cause of death worldwide. Globally, about 33.2 million people including 2.5 million children below 15 years are estimated to be infected with human immuno-deficiency virus (HIV), the causative organism of AIDS, in 2007 [1]. This opportunistic infection has claimed 2.1 million lives in 2007, and in this year, newly infected persons were 2.5 million. HIV primarily infects immune system of human beings such as helper T cells (specifically $CD4^+$ T cells), macrophages and dendritic cells [2]. There are three mechanisms to lower the levels of $CD4^+$ T cells. These are direct viral killing of infected cells, increased rates of apoptosis in infected cells and

killing of infected $CD4^+$ T cells by CD8 cytotoxic lymphocytes that recognize infected cells. When the number of $CD4^+$ T cells falls below a critical level, cell-mediated immunity is lost and thus the body becomes more prone to opportunistic infections. Until recently there is no successful and complete treatment invented against this retrovirus of lentivirus family. Treatment with anti-retrovirals may increase the life expectancy of the infected individuals.

Both HIV-1 and HIV-2 cause AIDS in humans. But HIV-1 is more virulent and easily transmitted [3]. The outside border of the viral structure consists of two glycoproteins, gp120 and gp41, which are breakdown products of gp160 by a protease. This glycoprotein complex helps the virus to fuse with the cells to initialize infection. After entering the blood stream, the viral particle binds the $CD4$ receptor of macrophage [4]. Binding to this receptor makes a conformational change to expose co-receptor sites like CCR5 and/or CXCR4. Out of nine genes present in this RNA virus, *gag*, *pol* and *env* are critical to viral structure and function. Reverse transcriptase

* Corresponding author. Tel.: +91 9831594140.

E-mail address: kunalroy_in@yahoo.com (K. Roy).URL: http://www.geocities.com/kunalroy_in

transcribes single stranded RNA to single stranded DNA, which is used for the synthesis of double stranded DNA. This migrates to the nucleus of the cell and integrates into host nucleus by an integrase [5]. At the last stage of the viral cycle, new HIV-1 viruses are assembled at the plasma membrane of the host cell for budding and maturation. The *gag-pol* polyproteins are translated into single protein molecule by the host ribosome. This protein molecule is again cleaved by protease into reverse transcriptase, integrase, etc. There are several targets for drug development against this disease like inhibitors of reverse transcriptase, integrase and protease enzymes and inhibitors of co-receptors (CCR5 and/or CXCR4) and fusion inhibitors.

Although there are several targets and different anti-retroviral therapies, drug resistance emerges quickly because of mutation at the transcription phase. This necessitates QSAR studies for developing good predictive models for ligands acting on different anti-HIV targets. Barreca et al. have designed, synthesized and performed structure–activity relationships and molecular modeling studies on 2,3-diaryl-1,3-thiazolidin-4-ones as reverse transcriptase inhibitors [6]. These authors have performed computational studies to delineate the ligand–RT interactions and to probe the binding of the ligands to HIV-1 RT. Rawal et al. have studied a series of 4-thiazolidines as selective inhibitors of the HIV-RT enzyme [7] and correlated the inhibitory activity with physicochemical properties using statistically significant QSAR models with good predictive ability. A structure-based design of non-nucleoside reverse transcriptase (RT) has been proposed by Mao et al. to explore the lowering of binding affinity with changes in binding pocket shape, volume and chemical properties of residue mutations [8]. Tintori et al. combined an electron–ion interaction potential (EIIP) technique with molecular modeling approaches for the identification of new HIV-1 integrase inhibitors [9]. Bhataraj and Garg investigated the effect of hydrophobicity on the design of 4-hydroxy-5,6-dihydropyran-2-ones as a new class of emerging HIV-1 protease inhibitors [10]. Kellenberger et al. have screened about 1.6 million commercially available compounds against CCR5 model by sequential filters (drug-likeness, 2-D pharmacophore, 3-D docking and scaffold clustering) and 10 compounds are detected of having binding affinity to CCR5 [11]. CoMFA and CoMSiA and docking studies have been performed by Buolamwini and Assefa on conformationally restrained cinnamoyl HIV-1 integrase inhibitors to explore binding mode at the active site [12].

The present group of authors [13–22] have performed QSAR modeling of anti-HIV compounds of different chemical series acting on different targets. In continuation of such efforts, in the present paper, we have performed QSAR modeling of tetrahydroimidazo[4,5,1-*jk*][1,4]benzodiazepine (TIBO) derivatives as reverse transcriptase inhibitors. The data set was taken from the work reported by Huuskonen [23]. In the development of QSAR models, both linear and non-linear techniques were utilized and best two models developed by us were compared to the reported models of TIBO derivatives published earlier by different researchers.

2. Methods and materials

The anti-HIV data (IC_{50}) of tetrahydroimidazo[4,5,1-*jk*][1,4]benzodiazepine (TIBO) derivatives [23] were used for our QSAR study. All compounds are shown in Table 1 serially citing their substituents and biological activities. Multiple regression (with stepwise regression, factor analysis and genetic function approximation as variable selection tools) and partial least squares (PLS) and PLS with factor analysis as preprocessing step (FA-PLS) were used as tools for the development of linear models. Non-linear models were developed using genetic function approximation (spline-based) and artificial neural networks.

2.1. Descriptors

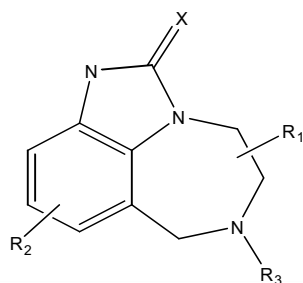
In this paper, linear and non-linear models were developed using topological, structural, physicochemical, spatial and electronic descriptors. These descriptors were divided into three categories using different combinations of descriptors. The first category was a combination of topological and structural descriptors while the second category included structural, physicochemical, spatial and electronic descriptors. All these parameters were clubbed together in the third category. All descriptors were calculated using Cerius2 version 10 software [24] running under IRIX 6.5 operating system on a Silicon Graphics workstation and are shown categorically in Table 2. The study included Balaban index (J_x), kappa shape indices, Zagreb, Wiener, connectivity indices and E-state indices as topological descriptors, molecular weight (MW), number of rotatable bonds (*Rotibonds*), number of hydrogen bond donors and acceptors and number of chiral centers as structural parameters, $A \log P$, $A \log P_{98}$, $\log P$, MR and $MolRef$ as physicochemical parameters, *RadOfGyration*, *Jurs*, *Shadow*, *Area*, *Density*, *Vm* as spatial and *Apol*, *Homo*, *Lumo* and *Sr* as electronic descriptors. The meanings of all descriptors are available at the website <http://www.accelrys.com>. All E-state parameters, which are selected in the development of models, are defined in Table 3.

2.2. Cluster analysis and validation

Cluster analysis [25] is a method of arranging objects into groups. This method divides different objects into groups in such a way that the degree of association between two objects is maximum if they possess same group and otherwise minimum. Hierarchical and non-hierarchical are two types of clustering techniques. The hierarchical technique forms relationships within clusters in subsequent steps, but in the second method, compounds are first classified into defined number of clusters based on nearest neighbor distributions in chemical space. *K*-means clustering is one of the best known non-hierarchical clustering techniques [26]. In this method, clusters are started randomly and then cluster means are calculated in the descriptor space. Molecules are reassigned to clusters whose means are closer to the position of the molecules. In this paper, clustering technique is used as a tool for selection of training set and test set compounds because in this way both training and test sets can represent all clusters and the whole data set. After development of a model from the training set, robustness and predictive power of the model should be checked on new chemical entities. The test set is used as new chemical entities for judging predictive potential of the developed models. The validation strategies check the reliability of the developed models for their possible application on a new set of data, and thus confidence of prediction can be judged [27]. Predictive capacity of a model depends on the similarity of the chemical nature between training set and test set [28–30]. If the structures of the test set are very close to the structures of the training set, the predictive potential of the models for the test set will be good. The reason is that the model has captured all features common to the training set molecules. Thus, predictive accuracy and confidence of a model for different unknown chemicals vary according to how well the training set represents the unknown chemicals and how robust the model is in extrapolating beyond the chemistry space defined by the training set. Therefore, assessing a model's predictive accuracy outside the training domain is a vital step toward defining the application domain of a model for the regulatory acceptance of QSARs.

In our study, division of the data set into training and test sets was performed based on *K*-means clustering applied on standardized topological and structural descriptors (standardized values being between 0 and 1). The whole data set was divided into 5

Table 1
Structural features and HIV-1 RT inhibitory activity values of TIBO derivatives



S. no.	R1	X	R2	R3	Activity [$\log(1/IC_{50})$]		
					Obs [23]	Cal ^a	Cal ^b
1	H	S	8-Cl	DMA	7.340	7.560	7.466
2	H	S	9-Cl	DMA	6.790	7.678	7.247
3	5-Et	O	H	2-MA	4.300	4.454	4.503
4 ^c	5- <i>i</i> -Pr	O	H	2-MA	5.000	5.184	4.556
5	5- <i>i</i> -Pr	O	H	DMA	5.000	4.750	5.039
6	5,5-Di-Me	O	H	2-MA	4.640	4.604	5.041
7 ^c	4-Me	O	H	2-MA	4.490	4.599	4.632
8 ^c	4-Me	S	9-Cl	2-MA	6.170	4.062	6.708
9 ^c	4-Me	S	9-Cl	CH ₂ CH(CH ₂) ₂	5.660	4.765	6.952
10 ^c	4- <i>i</i> -Pr	O	H	<i>n</i> -Pr	4.130	5.621	4.425
11	4- <i>i</i> -Pr	O	H	2-MA	4.900	5.366	4.882
12	4- <i>n</i> -Pr	O	H	<i>n</i> -Pr	3.740	7.011	4.412
13 ^c	4- <i>n</i> -Pr	O	H	2-MA	4.320	4.345	4.998
14	7-Me	O	H	<i>n</i> -Pr	4.080	5.684	4.134
15	7-Me	O	H	DMA	4.920	5.684	4.853
16	7-Me	O	8-Cl	DMA	6.840	5.202	6.077
17 ^c	7-Me	O	9-Cl	DMA	6.790	5.017	5.850
18	7-Me	S	H	<i>n</i> -Pr	5.610	5.308	5.131
19 ^c	7-Me	S	H	DMA	7.110	6.707	6.045
20	7-Me	S	8-Cl	DMA	7.920	5.308	6.905
21 ^c	7-Me	S	9-Cl	DMA	7.640	5.682	6.899
22	4,5-Di-Me (<i>cis</i>)	O	H	DMA	4.250	6.348	4.515
23	4,5-Di-Me (<i>cis</i>)	S	H	DMA	5.650	6.027	6.021
24 ^c	4,5-Di-Me (<i>trans</i>)	S	H	CH ₂ CH(CH ₂) ₂	4.870	5.852	5.747
25	4,5-Di-Me (<i>trans</i>)	S	H	DMA	4.840	7.251	6.127
26	4-Keto-5-Me	S	9-Cl	<i>n</i> -Pr	4.300	6.545	5.764
27	4,5-Di-benzo	S	H	CH ₂ CH(CH ₂) ₂	5.000	5.503	5.368
28 ^c	5,7-Di-Me (<i>trans</i>)	S	H	DMA	7.380	5.480	6.170
29 ^c	5,7-Di-Me (<i>cis</i>)	S	H	DMA	5.940	4.299	6.416
30	5,7-Di-Me (<i>R,R; trans</i>)	O	9-Cl	DMA	6.640	6.222	5.906
31	5,7-Di-Me (<i>R,R; trans</i>)	S	9-Cl	DMA	6.320	4.888	7.052
32	5,7-Di-Me (<i>S,S; trans</i>)	O	9-Cl	DMA	5.300	5.617	5.834
33	4,7-Di-Me (<i>trans</i>)	S	H	DMA	4.590	5.304	5.705
34	5,6-CH ₂ C(=CHCH ₃)CH ₂ (<i>S</i>)	S	9-Cl	-	5.420	4.727	6.471
35	6,7-(CH ₂) ₄	S	9-Cl	-	5.700	4.582	6.299
36 ^c	5-Me (<i>S</i>)	S	8-Cl	DMA	8.300	4.856	7.530
37	5-Me (<i>S</i>)	O	9-Cl	DMA	6.740	6.222	6.395
38	5-Me (<i>S</i>)	S	9-Cl	DMA	7.370	4.331	7.303
39	5-Me (<i>S</i>)	S	9-Cl	CH ₂ CH(CH ₂) ₂	7.470	6.325	7.074
40	5-Me (<i>S</i>)	S	H	CH ₂ CH(CH ₂) ₂	7.220	6.233	6.476
41 ^c	5-Me	O	H	<i>n</i> -Pr	4.220	7.631	4.395
42	5-Me	S	H	<i>n</i> -Pr	5.780	7.317	5.549
43	5-Me	O	H	2-MA	4.460	7.187	4.785
44	5-Me	S	H	DMA	7.010	7.736	6.518
45	5-Me (<i>S</i>)	O	H	DMA	5.480	5.459	5.287
46	5-Me (<i>S</i>)	S	H	2-MA	7.580	5.920	6.181
47	H	O	H	DMA	4.900	7.129	5.138
48	H	O	H	2-MA	4.330	7.590	4.911
49	H	O	H	<i>n</i> -Pr	4.050	6.452	4.172
50	H	O	H	2-EA	4.430	6.631	4.448
51	5-Me (<i>S</i>)	S	H	DMA	7.355	7.047	6.459
52	5-Me (<i>S</i>)	O	H	Allyl	4.154	7.047	4.708
53 ^c	5-Me (<i>S</i>)	O	H	<i>n</i> -Bu	3.999	4.581	3.949
54	5-Me (<i>S</i>)	S	8-F	DMA	8.235	4.306	7.397
55	5-Me (<i>S</i>)	O	8-Br	DMA	7.324	6.655	6.894
56	5-Me (<i>S</i>)	S	8-Br	DMA	8.521	6.555	7.819
57	5-Me (<i>S</i>)	S	8-Me	DMA	7.865	4.464	7.219
58	5-Me (<i>S</i>)	S	8-OMe	DMA	7.468	4.744	6.883
59	5-Me (<i>S</i>)	S	9,10-Di-Cl	DMA	7.592	5.724	7.016
60	5-Me (<i>S</i>)	O	8-CN	DMA	5.940	6.092	6.136
61 ^c	5-Me (<i>S</i>)	S	8-CN	DMA	7.250	7.117	7.199

(continued on next page)

Table 1 (continued)

S. no.	R1	X	R2	R3	Activity [$\log(1/IC_{50})$]		
					Obs [23]	Cal ^a	Cal ^b
62	5-Me (S)	O	8-Me	DMA	6.000	4.996	5.833
63	5-Me (S)	S	10-OMe	DMA	5.330	5.673	6.083
64 ^c	5-Me (S)	O	10-OMe	DMA	5.180	5.673	4.637
65	5-Me (S)	S	10-Br	DMA	5.970	7.137	6.961
66	5-Me (S)	S	8-CHO	DMA	6.730	4.170	6.899
67	5-Me (S)	O	8-I	DMA	7.060	4.411	6.934
68	5-Me (S)	S	8-I	DMA	7.320	6.855	7.386
69 ^c	5-Me (S)	O	8-C≡CH	DMA	6.360	5.700	6.330
70	5-Me (S)	S	8-C≡CH	DMA	7.530	5.649	7.386

DMA = 3,3-Dimethylallyl; 2-MA = 2-methylallyl; 2-EA = 2-ethylallyl.

^a Obtained from the best linear model (Eq. (8)) (based on r_m^2 for the test set).

^b Obtained from the best non-linear model (ANN model 9) (based on r_m^2 for the test set).

^c Stands for a member of the test set.

clusters from each of which about 25% compounds were taken to the test set and about 75% compounds were taken to the training set. Serial numbers of compounds of different clusters are shown in Table 4.

For further validation of selected models, an external set of 10 TIBO derivatives [31], which are not shared with the compounds of the present data set but falling within the domain of applicability of the models, was selected. The prediction of a modeled response (biological activity) using QSAR is valid only if the compound being predicted is within the applicability domain of the model [32]. Not even a robust and validated QSAR model can be expected to predict reliably the modeled property for the entire universe of chemicals. The applicability domain is a theoretical region of chemical space, defined by the model descriptors and modeled response, and thus by the nature of the training set molecules.

For the development of equations different chemometric tools were utilized.

2.3. Stepwise MLR

In this method a multiple-term equation is built step by step where an initial model is recognized and then it is repeatedly altered by adding or removing a predictor variable according to the

Table 2
Categorical list of descriptors used in the development of models

Category of descriptors	Name of the descriptors
Topological	$J_x, {}^1\chi, {}^2\chi, {}^3\chi, {}^4\chi, {}^5\chi, {}^6\chi, {}^7\chi, {}^8\chi, {}^9\chi, {}^{10}\chi, {}^{11}\chi, {}^{12}\chi, {}^{13}\chi, {}^{14}\chi, {}^{15}\chi, {}^{16}\chi, {}^{17}\chi, {}^{18}\chi, {}^{19}\chi, {}^{20}\chi, {}^{21}\chi, {}^{22}\chi, {}^{23}\chi, {}^{24}\chi, {}^{25}\chi, {}^{26}\chi, {}^{27}\chi, {}^{28}\chi, {}^{29}\chi, {}^{30}\chi, {}^{31}\chi, {}^{32}\chi, {}^{33}\chi, {}^{34}\chi, {}^{35}\chi, {}^{36}\chi, {}^{37}\chi, {}^{38}\chi, {}^{39}\chi, {}^{40}\chi, {}^{41}\chi, {}^{42}\chi, {}^{43}\chi, {}^{44}\chi, {}^{45}\chi, {}^{46}\chi, {}^{47}\chi, {}^{48}\chi, {}^{49}\chi, {}^{50}\chi, {}^{51}\chi, {}^{52}\chi, {}^{53}\chi, {}^{54}\chi, {}^{55}\chi, {}^{56}\chi, {}^{57}\chi, {}^{58}\chi, {}^{59}\chi, {}^{60}\chi, {}^{61}\chi, {}^{62}\chi, {}^{63}\chi, {}^{64}\chi, {}^{65}\chi, {}^{66}\chi, {}^{67}\chi, {}^{68}\chi, {}^{69}\chi, {}^{70}\chi, {}^{71}\chi, {}^{72}\chi, {}^{73}\chi, {}^{74}\chi, {}^{75}\chi, {}^{76}\chi, {}^{77}\chi, {}^{78}\chi, {}^{79}\chi, {}^{80}\chi, {}^{81}\chi, {}^{82}\chi, {}^{83}\chi, {}^{84}\chi, {}^{85}\chi, {}^{86}\chi, {}^{87}\chi, {}^{88}\chi, {}^{89}\chi, {}^{90}\chi, {}^{91}\chi, {}^{92}\chi, {}^{93}\chi, {}^{94}\chi, {}^{95}\chi, {}^{96}\chi, {}^{97}\chi, {}^{98}\chi, {}^{99}\chi, {}^{100}\chi, \Phi, SC-0, SC-1, SC-2, SC-3_P, SC-3_C, {}^6\chi, {}^1\chi, {}^2\chi, {}^3\chi, {}^4\chi, {}^5\chi, {}^6\chi, {}^7\chi, {}^8\chi, {}^9\chi, {}^{10}\chi, {}^{11}\chi, {}^{12}\chi, {}^{13}\chi, {}^{14}\chi, {}^{15}\chi, {}^{16}\chi, {}^{17}\chi, {}^{18}\chi, {}^{19}\chi, {}^{20}\chi, {}^{21}\chi, {}^{22}\chi, {}^{23}\chi, {}^{24}\chi, {}^{25}\chi, {}^{26}\chi, {}^{27}\chi, {}^{28}\chi, {}^{29}\chi, {}^{30}\chi, {}^{31}\chi, {}^{32}\chi, {}^{33}\chi, {}^{34}\chi, {}^{35}\chi, {}^{36}\chi, {}^{37}\chi, {}^{38}\chi, {}^{39}\chi, {}^{40}\chi, {}^{41}\chi, {}^{42}\chi, {}^{43}\chi, {}^{44}\chi, {}^{45}\chi, {}^{46}\chi, {}^{47}\chi, {}^{48}\chi, {}^{49}\chi, {}^{50}\chi, {}^{51}\chi, {}^{52}\chi, {}^{53}\chi, {}^{54}\chi, {}^{55}\chi, {}^{56}\chi, {}^{57}\chi, {}^{58}\chi, {}^{59}\chi, {}^{60}\chi, {}^{61}\chi, {}^{62}\chi, {}^{63}\chi, {}^{64}\chi, {}^{65}\chi, {}^{66}\chi, {}^{67}\chi, {}^{68}\chi, {}^{69}\chi, {}^{70}\chi, {}^{71}\chi, {}^{72}\chi, {}^{73}\chi, {}^{74}\chi, {}^{75}\chi, {}^{76}\chi, {}^{77}\chi, {}^{78}\chi, {}^{79}\chi, {}^{80}\chi, {}^{81}\chi, {}^{82}\chi, {}^{83}\chi, {}^{84}\chi, {}^{85}\chi, {}^{86}\chi, {}^{87}\chi, {}^{88}\chi, {}^{89}\chi, {}^{90}\chi, {}^{91}\chi, {}^{92}\chi, {}^{93}\chi, {}^{94}\chi, {}^{95}\chi, {}^{96}\chi, {}^{97}\chi, {}^{98}\chi, {}^{99}\chi, {}^{100}\chi, Wiener, log Z, Zagreb, S_{sCH_3}, S_{dCH_2}, S_{ssCH_2}, S_{tCH}, S_{dsCH}, S_{aaCH}, S_{sssCH}, S_{tsC}, S_{dssC}, S_{aasC}, S_{aaaC}, S_{ssssC}, S_{snH_2}, S_{ssNH}, S_{tN}, S_{aaN}, S_{ssnN}, S_{dsnN}, S_{aasnN}, S_{soH}, S_{dO}, S_{ssO}, S_{sas}, S_{dsss}, S_{ddss}, S_{sF}, S_{sBr}, S_{sCl}.$
Structural	MW, Rotlbonds, Hbondacceptor, Hbonddonor, Chiralcenters.
Physicochemical	A log P, A log P98, log P, MR, MolRef.
Spatial	RadOfGyration, Jurs_SASA, Jurs_PPSA_1, Jurs_PNSA_1, Jurs_DPSA_1, Jurs_PPSA_2, Jurs_PNSA_2, Jurs_DPSA_2, Jurs_PPSA_3, Jurs_PNSA_3, Jurs_DPSA_3, Jurs_FPFA_1, Jurs_FNFA_1, Jurs_FPFA_2, Jurs_FNFA_2, Jurs_FPFA_3, Jurs_FNFA_3, Jurs_WPSA_1, Jurs_WNSA_1, Jurs_WPSA_2, Jurs_WNSA_2, Jurs_WPSA_3, Jurs_WNSA_3, Jurs_RPCG, Jurs_RNCG, Jurs_RPCS, Jurs_RNCS, Jurs_TPFA, Jurs_TASA, Jurs_RPSA, Jurs_RASA, Shadow_XY, Shadow_XZ, Shadow_YZ, Shadow_XYfrac, Shadow_XZfrac, Shadow_YZfrac, Shadow_nu, Shadow_Xlength, Shadow_Ylength, Shadow_Zlength, Area, Vm, Density, PMI_mag.
Electronic	Apol, Dipole-mag, Homo, Lumo, Sr.

“stepping criteria” (in this study $F = 4$ for inclusion and $F = 3.9$ for exclusion for the forward selection method) [33]. When a specified maximum number of steps has been reached or stepping is not possible, the process was terminated. At each step all variables are assessed and evaluated to determine which one will contribute the most to the equation. The method selected for stepwise regression is forward selection and backward elimination. The criteria “F to Enter” and “F to Remove” determine how significant or insignificant is the contribution of a variable in the regression equation, respectively, for adding to the equation and removing from the equation. A limitation of the stepwise regression search approach is that it is a very biased procedure, because it assumes that there is only one best subset of X variables and seeks to identify it. There

Table 3
Definition of E-state parameters

Name of descriptors	Meaning of descriptors	
S_{dssC}	E-state value of carbon atom in the fragment	
S_{aaCH}	E-state value of carbon atom in the fragment	
S_{aasC}	E-state value of carbon atom in the fragment	
S_{dO}	E-state value of carbon atom in the fragment	$=O$
S_{ds}	E-state value of carbon atom in the fragment	$=S$
S_{dsCH}	E-state value of carbon atom in the fragment	$=CH-$
S_{ssNH}	E-state value of nitrogen atom in the fragment	
S_{sssCH}	E-state value of carbon atom in the fragment	$-CH$
S_{tsC}	E-state value of carbon atom in the fragment	$\equiv C-$
S_{ssnN}	E-state value of nitrogen atom in the fragment	

Table 4
Serial numbers of compounds under different clusters

Cluster no.	Serial numbers of compounds
1	9, 24, 27, 34, 35, 39, 40
2	1, 2, 8, 19, 44, 46, 51
3	7, 14, 18, 41, 42, 43, 47, 48, 49, 50, 52, 53
4	3, 4, 6, 10, 11, 12, 13, 15, 16, 17, 22, 26, 37, 45, 62
5	5, 20, 21, 23, 25, 28, 29, 30, 31, 32, 33, 36, 38, 54, 55, 56, 57, 58, 59, 60, 61, 63, 64, 65, 66, 67, 68, 69, 70

is often no unique 'best' subset, and all possible regression models with a similar number of X variables as in the stepwise regression solution should be fitted subsequently to study whether some other subsets of X variables might be better.

2.4. PLS

When the number of factors is large and remarkable collinearity persists among them, partial least squares (PLS) is an ideal technique for that case. This procedure generalizes and combines features from principal component and multiple regression. To avoid overfitting, a strict test for the significance of each consecutive PLS component is necessary and then stopping when the components are non-significant. Leave-one out method was employed to select optimum number of components [34,35]. This ensures that the QSAR equations are selected based on their ability to predict the data rather than to fit the data. In our study, variables with smaller coefficients were removed from PLS regression analysis until there was no further improvement in leave-one-out (LOO) crossvalidation R^2 (Q^2) value irrespective of the components. In case of leave-one-out crossvalidation, one compound is deleted from the data set and a model is developed from the reduced set. The deleted compound is predicted from the model, which was developed excluding the compound. The process is repeated until all compounds are deleted once. The outcome from the cross-validation procedure is crossvalidated R^2 ($LOO-Q^2$) which is used as a criterion of both robustness and predictive ability of the model. Crossvalidated squared correlation coefficient R^2 ($LOO-Q^2$) is calculated according to Eq. (1)

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

In the above equation, \bar{Y} means average activity value of the training data set while Y_{obs} and Y_{pred} represent observed and predicted activity values. Often, a high Q^2 value ($Q^2 > 0.5$) is considered as a proof of high predictive ability of the model.

2.5. FA-PLS

This is the combination of factor analysis (FA) and partial least squares (PLS), where FA is used for initial selection of descriptors following which PLS is performed. Factor analysis [36,37] is a tool to find out the relationship among variables. It reduces variables into few latent factors from which important variables are selected for PLS regression. Here also leave-one-out method is used as a tool for selection of optimum number of components for PLS.

2.6. GFA

A genetic algorithm (GA) is a search technique [38–40] used in computing to find exact or approximate solutions to optimization and search problems. Genetic function approximation was conceived from (i) genetic algorithm originally developed by Fraser and others, and later popularized by Holland, (ii) Friedman's

multivariate adaptive regression splines (MARS) algorithm. GFA is done as follows: (i) an initial population of equations is generated by random choice of descriptors; (ii) pairs from the population of equations are chosen at random and "crossovers" are performed and progeny equations are generated; (iii) it is better at discovering combinations of features that take advantage of correlations between multiple features; (iv) the fitness of each progeny equation is assessed by lack-of-fit (LOF) measure; (v) it can use a larger variety of equation term types in construction of its models; (vi) if the fitness of new progeny equation is better, then it is preserved. The models with proper balance of all statistical terms are used to explain variance in the biological activity. A distinctive feature of GFA is that it produces a population of models (e.g., 100), instead of generating a single model, as do most other statistical methods. The range of variations in this population gives added information on the quality of fit and importance of the descriptors. Because of additional information about the models, this algorithm produces superior models in comparison to traditional stepwise or similar techniques. The fitness function, i.e., lack-of-fit (LOF) is calculated by the following formula:

$$\text{LOF} = \frac{\text{LSE}}{\left(1 - \frac{c+dp}{M}\right)^2} \quad (2)$$

In the above equation, c is the number of basis functions (other than constant term), d is smoothing parameter (adjustable by the user), M is the number of samples in the training set, LSE is least square error, and p is the total number of features contained in all basis functions. The best model contains the best fitness score. The genetic crossover operation is repeatedly performed on the basis of LOF scores of the models. GFA can build not only linear models but also higher-order polynomials, splines and Gaussians.

In our study we have used GFA as a tool for the development of both linear and non-linear models. In the first case GFA was utilized for selection of variables, which were subsequently used in making of models using MLR technique. GFA (spline) was used as a tool for non-linear models. Attempt was made to develop GFA models using different levels of iterations. Based on the LOF and/or Q^2 values of the best models of a population, the final number of iterations for the development of the models was set. The GFA model development process at the selected level of iterations was subjected to randomization test to judge absence of over-training.

2.7. G/PLS

The combination of (i) genetic function approximation and (ii) partial least squares is referred to as G/PLS. Both of these methods are valuable analytical tools for QSAR modeling where number of descriptors is more than samples. The variables which are selected by GFA are submitted for PLS regression to weigh the relative contributions of the selected variables in the final model. Attempt was made to develop G/PLS models using different levels of iterations. Based on the Q^2 and/or LSE values of the best models of a population, the final number of iterations for the development of the models was set. In general, number of iterations required for the G/PLS models was less than that for the GFA models as higher number of iterations in case of the G/PLS technique leads to poorer models. The G/PLS model development process at the selected level of iterations was subjected to randomization test to judge absence of over-training.

2.8. Artificial neural networks

Artificial neural networks [41] are inspired from the information-processing pattern of the biological nervous system. Input, hidden and output layers are the main components of most neural

networks. The input layer takes information directly from input files, and the output layer sends information directly to the outside world through computer or any other mechanical control system. There may be many hidden layers between input and output layers. These hidden layers are interconnected with each other. Depending on the modes of function, there are different types of neural networks like feedforward backpropagation, probabilistic neural network, counter propagation, self-organizing map, etc. In our study, feedforward backpropagation was selected as a tool for optimization of neural networks. Multilayer perceptron (MLP) method under “Custom Network Designer” has been selected to design the network. In the first phase, backpropagation method was selected for formation of the network using training set. The error term, i.e., difference between output of the network and the desired output is backpropagated to the transfer function (sigmoid function) for adjustment of weight. The output [42] can be represented by the following equation.

$$O_j = f(i_j) = \frac{1}{1 + \exp(-\beta i_j)} \quad (3)$$

In Eq. (3), O_j is the output of node j and β is a gain, being able to adjust the form of the function. Usually β is taken as 1. Using the error signal to adjust the connected weights, the following adjusted weights are obtained for the output layer.

$$W_{ij}(\text{new}) = W_{ij}(\text{old}) + \eta \delta_i O_j + \alpha [\Delta W_{ij}(\text{old})] \quad (4)$$

In backpropagation method, the learning of the network followed the Delta Rule, which starts with the calculated difference between the actual outputs and the desired outputs. This difference or error is backpropagated to the transfer function for adjustment of connection weights, which can minimize the error. The complex part of this learning mechanism is to determine which input contributed the most to an incorrect output and how does that element get changed to correct the error. During the learning process, a forward sweep is made through the network, and the output of each element is computed layer by layer. The difference between the output of the final layer and the desired output is back propagated to the previous layer until the input layer is reached. In second phase conjugate gradient descent was used. This method is a good secondary and advanced method of training multilayer perceptron. It is generally used for the network of large number of weights and/or multiple output units. It is a batch update algorithm whereas backpropagation adjusts the weights of the network. Learning rate and momentum of each epoch are adjusted and weight decay is regularized.

Most work on assessing performance in neural modeling concentrates on approaches to resampling. When a particular number of resampling is selected, the number of available cases is divided into 3 subsets (training, selection and test sets). A neural network is optimized using a training subset. Often, a separate subset (the selection subset) is used to halt training to mitigate over-learning, or to select from a number of models trained with different parameters. Then, a third subset (the test subset) is used to perform an unbiased estimation of the network's likely performance. Although the use of a test subset allows us to generate unbiased performance estimates, these estimates may exhibit high variance. Ideally, one would like to repeat the training procedure a number of different times, each time using new training, selection and test cases drawn from the population – then, one could average the performance prediction over the different test subsets, to get a more reliable indicator of generalization performance. In reality, one seldom has enough data to perform a number of training runs with entirely separate training, selection and test subsets. Cross-validation is the simplest resampling technique. We have

crossvalidated the networks using 15 or 20 resampling. The sum of total number of compounds in training, selection and test sets may not be equal to the total number of compounds used for development of the network. During 20 resampling, number of cases selected for training, selection and test compounds were 25, 13 and 2 while in case of 15 resampling, these values were 28, 14 and 3 respectively.

2.9. Model quality

The statistical quality of the multiple regression equations [43] was examined by different parameters like *square of correlation coefficient* (R^2), *explained variance* (R_a^2), *standard error of estimate* (s) and *variance ratio* (F) at specified *degrees of freedom* (df). All accepted MLR equations have regression coefficients and F ratios significant at 95% and 99% levels, respectively, if not stated otherwise. The generated QSAR equations were validated by *leave-one-out* or *LOO statistics* [44,45] and *crossvalidation* R^2 (Q^2) and *predicted residual sum of squares* (PRESS) values were reported. In case of external validation, predictive capacity of a model was judged by its application for prediction of test set activity values and calculation of predictive R^2 (R_{pred}^2) value as shown below:

$$R_{\text{pred}}^2 = 1 - \frac{\sum (Y_{\text{pred}(\text{Test})} - Y_{(\text{Test})})^2}{\sum (Y_{(\text{Test})} - \bar{Y}_{\text{training}})^2} \quad (5)$$

In the above equation, $Y_{\text{pred}(\text{Test})}$ and $Y_{(\text{Test})}$ indicate predicted and observed activity values, respectively, of the test set compounds and $\bar{Y}_{\text{training}}$ indicates mean activity value of the training set. Furthermore, squared correlation coefficient (r^2) between observed and predicted values of the test set compounds was also noted.

2.10. Software

MINITAB [46] was used for cluster analysis, stepwise regression and PLS. SPSS [47] and STATISTICA [48] were used for factor analysis and ANN, respectively. Cerius2 version 4.10 [24] was used for GFA (linear and non-linear) and G/PLS.

3. Results and discussion

3.1. QSAR using topological and structural descriptors

3.1.1. Stepwise regression


In the final model, 7 variables were selected. The equation of this model was developed using F criterion ($F = 4$ for inclusion; $F = 3.9$ for exclusion).

$$\begin{aligned} pIC_{50} = & 10.18 + 0.99(\pm 0.797)S_{\text{dssC}} - 0.89(\pm 0.687)S_{\text{aaCH}} \\ & + 11.1(\pm 11.113)^3 \chi_{\text{CH}} - 0.333(\pm 0.266)SC - 3.P \\ & + 1.61(\pm 1.923)^3 \chi_p^v + 4.3(\pm 7.611)Jx \\ & - 0.76(\pm 1.419)Hbondacceptor \end{aligned} \quad (6)$$

$$\begin{aligned} n_{\text{Training}} = & 52, R^2 = 0.765, R_{\text{adj}}^2 = 0.728, F = 20.50, s \\ = & 0.697, Q^2 = 0.685, \text{PRESS} = 28.685, n_{\text{Test}} \\ = & 18, R_{\text{pred}}^2 = 0.536. \end{aligned}$$

The 95% confidence intervals of the regression coefficients are mentioned within parentheses. This equation could explain and

predict 72.8% and 68.5%, respectively. The predicted R^2 (R^2_{pred}) value for the test set is 0.536. The positive coefficients of two molecular connectivity indices (${}^3\chi_{\text{CH}}$, ${}^3\chi_{\text{p}}^{\text{v}}$) indicate that the number of branched vertices and flexibility are the determining factors of the biological activity. The coefficients of E-state indices have shown that both the electronic character and topological environment of

the fragment $>\text{C}=\text{C}$ increase while those of the fragment 

decrease the activity. Kier and Hall subgraph count index ($SC\text{-}3\text{-}P$) and hydrogen bond acceptor have negative impact on the activity while Balaban index (J_{x}) has positive impact on the activity.

3.1.2. PLS

The equation of the best model was obtained from 7 variables with 2 components selected by crossvalidation.

$$p\text{IC}_{50} = 9.471 - 0.129SC - 3.P + 1.036{}^3\chi_{\text{c}}^{\text{v}} + 0.006MW - 0.369S_{\text{aaCH}} + 0.377S_{\text{aasC}} - 0.060S_{\text{dO}} + 0.088S_{\text{dS}} \quad (7)$$

$$n_{\text{Training}} = 52, R^2 = 0.639, R^2_{\text{adj}} = 0.624, F = 43.33, Q^2 = 0.499, \text{PRESS} = 45.600, n_{\text{Test}} = 18, R^2_{\text{pred}} = 0.619.$$

The explained variance and predicted variance of Eq. (7) were 62.4% and 49.9%, respectively, of the total variance of reverse transcriptase inhibitory activity. Here the result of crossvalidation (internal validation) is not encouraging ($Q^2 < 0.5$) but the result of external validation is good ($R^2_{\text{pred}} = 0.619$). The effects of Kier and Hall subgraph count index and E-state index (S_{aaCH}) on the activity in the model are similar to those of Eq. (6). In Eq. (7), electronic and topological environment of the fragments $\text{O}=\text{C}$ and $\text{S}=\text{C}$ have opposite effects on the activity. Molecular weight, valence molecular connectivity index of 3rd order and E-state index (S_{aasC}) have positively influenced the biological activity.

3.1.3. FA-PLS

In factor analysis 10 factors could explain the data matrix to the extent of 95.8%. The reverse transcriptase inhibitory activity was moderately loaded with factor 1 (loaded in ${}^1\kappa$, ${}^2\kappa$, ${}^3\kappa$, ${}^1\kappa_{\text{am}}$, ${}^2\kappa_{\text{am}}$, ${}^3\kappa_{\text{am}}$, Φ , SC_0 , SC_1 , SC_2 , $SC_3\text{-}P$, $SC_3\text{-}C$, ${}^0\chi$, ${}^1\chi$, ${}^2\chi$, ${}^3\chi_{\text{p}}$, ${}^3\chi_{\text{c}}$, ${}^0\chi^{\text{v}}$, ${}^1\chi^{\text{v}}$, ${}^2\chi^{\text{v}}$, ${}^3\chi_{\text{p}}^{\text{v}}$, ${}^3\chi_{\text{c}}^{\text{v}}$, Wiener, $\log Z$, Zagreb, MW, S_{dsCH}), factor 3 (S_{aasC} , S_{ssNH} , S_{sssN} , S_{dO} , S_{dS}) and factor 8 (S_{dsCH}) and poorly loaded with factor 4 (*Hbondacceptor*, S_{scI}), factor 6 (S_{tCH} , S_{tsC}), factor 7 (*Rotlbond*), factor 9 (S_{sBr}), factor 10 (S_{st}), factor 11 (S_{ssO}) and factor 12 (S_{sf}). Based on the variables found important in factor analysis, PLS regression was carried out and the following equation composed of 9 variables was obtained using 2 crossvalidated components.

$$p\text{IC}_{50} = 1.709 + 0.705{}^3\kappa_{\text{am}} - 0.046SC - 3.P + 0.005MW - 0.436\text{Chiralcenters} + 0.139S_{\text{dsCH}} - 0.251S_{\text{aaCH}} + 0.381S_{\text{aasC}} + 1.080S_{\text{ssNH}} - 0.051S_{\text{dO}} \quad (8)$$

$$n_{\text{Training}} = 52, R^2 = 0.636, R^2_{\text{adj}} = 0.621, F = 42.72, Q^2 = 0.526, \text{PRESS} = 43.158, n_{\text{Test}} = 18, R^2_{\text{pred}} = 0.707.$$

Kier and Hall subgraph count index, molecular weight, E-state indices (S_{aasC} , S_{dO} , S_{aaCH}) have effects on the inhibitory activity similar to Eqs. (6) and (7). Besides this, the kappa alpha-modified shape index of 3rd order, electronic and topological environment of E-state indices

(S_{dsCH} , S_{ssNH}) have positively influenced the activity whereas number of chiral centers decreases the activity. Eq. (8) could explain 62.1% of the variance of the RT inhibitory activity of TIBO derivatives while the leave-one-out predicted variance was 52.6%. The predicted R^2 value for the test set is higher than those of Eqs. (6) and (7).

3.1.4. GFA-MLR (50,000 iterations)

Eq. (9) was obtained using 50,000 iterations with 4 variables.

$$p\text{IC}_{50} = 20.145 + 2.047(\pm 1.583)S_{\text{sssCH}} - 0.911(\pm 0.494)S_{\text{aaCH}} + 1.046(\pm 0.691)S_{\text{dssC}} - 0.265(\pm 0.220)SC - 3.P \quad (9)$$

$$n_{\text{Training}} = 52, R^2 = 0.707, R^2_{\text{adj}} = 0.682, F = 28.29, s = 0.754, \text{LOF} = 0.717, Q^2 = 0.649, \text{PRESS} = 31.965, n_{\text{Test}} = 18, R^2_{\text{pred}} = 0.533.$$

The activity increases with increase of E-state indices (S_{sssCH} and S_{dssC}) and decrease of Kier and Hall subgraph count index and E-state index (S_{aaCH}). This equation could explain and predict 68.2% and 64.9% of the variance of the activity, respectively.

3.1.5. G/PLS (1000 iterations)

The following equation was developed with 4 variables using 3 components.

$$p\text{IC}_{50} = 12.073 + 0.044\text{Zagreb} - 0.679S_{\text{aaCH}} - 0.154SC - 3.P - 0.101S_{\text{dO}} \quad (10)$$

$$n_{\text{Training}} = 52, R^2 = 0.628, R^2_{\text{adj}} = 0.605, F = 27.78, Q^2 = 0.561, \text{LSE} = 0.520, \text{PRESS} = 32.071, n_{\text{Test}} = 18, R^2_{\text{pred}} = 0.568.$$

Eq. (10) could explain and predict 67.8% and 64.7%, respectively, of the variance of anti-HIV activity. The predicted R^2 value for the test set is 0.568. In Eq. (10), except Zagreb, other 3 independent variables (S_{aaCH} , $SC\text{-}3\text{-}P$ and S_{dO}) have negative impact on the biological activity. The Zagreb index is defined as the sum of the squares of vertex valencies.

3.1.6. GFA (spline) (75,000 iterations)

Eq. (6) was obtained with 5 variables using 75,000 iterations.

$$p\text{IC}_{50} = 6.807 - 1.464(\text{Chiralcenters} - 1) - 5.117(2.58 - {}^3\kappa_{\text{am}}) - 0.833(\text{Rotlbonds} - 2) + 0.073(11.909 - S_{\text{dO}}) + 0.828(14.917 - {}^1\kappa) \quad (11)$$

$$n_{\text{Training}} = 52, R^2 = 0.791, R^2_{\text{adj}} = 0.768, F = 34.72, s = 0.644, \text{LOF} = 0.722, Q^2 = 0.737, \text{PRESS} = 23.886, n_{\text{Test}} = 18, R^2_{\text{pred}} = 0.689.$$

Eq. (11) shows a negative coefficient of $(\text{Chiralcenters}-1)$, which indicates that the number of chiral centers above 1 is detrimental to the activity. The value of kappa alpha-modified shape index of 3rd order above 2.58 is conducive for the RT inhibitory activity as evident from the negative coefficient of $(2.58 - {}^3\kappa_{\text{am}})$. The negative coefficient of $(\text{Rotlbonds}-2)$ implies that the number of rotatable bonds should

be less than 2. The negative coefficient of $(11.909 - S_{do})$ implies that the value of S_{do} should be below 11.909 for optimum activity. A value of kappa shape index of 1st order below 14.197 is necessary for positive contribution to the inhibitory activity.

3.2. QSAR using structural, physicochemical, spatial and electronic descriptors

3.2.1. Stepwise regression

Using structural, physicochemical, spatial and electronic descriptors, the following equation was obtained with 5 independent variables ($F = 4$ for inclusion; $F = 3.9$ for exclusion).

$$\begin{aligned} pIC_{50} = & -4.623 + 0.038(\pm 0.036)Jurs_WNSA_1 \\ & - 0.185(\pm 0.233)Dipole_mag + 0.98(\pm 0.934)A \log P \\ & + 1.80(\pm 2.411)Jurs_RPCS + 31(\pm 42.593)Jurs_RNCG \end{aligned} \quad (12)$$

$$\begin{aligned} n_{\text{Training}} &= 52, R^2 = 0.709, R^2_{\text{adj}} = 0.677, F = 22.41, s \\ &= 0.759, Q^2 = 0.606, \text{PRESS} = 35.837, n_{\text{Test}} \\ &= 18, R^2_{\text{pred}} = 0.154. \end{aligned}$$

All regression coefficients are significant at 95% confidence level and the corresponding confidence intervals are mentioned within parentheses. This model could explain 67.7% and predict 60.6% of the variance of the RT inhibitory activity. However, the external validation statistics of Eq. (12) are very poor. Probably, the combination of variables selected by the stepwise method is good only for fitting data and not for prediction. In this equation all three Jurs descriptors have positive effects on the pIC_{50} . Jurs descriptors combine shape and electronic information to characterize molecules and are developed by mapping atomic partial charges on solvent-accessible surface areas of individual atoms. *WNSA-1* is the surface weighted negatively charged partial surface area, which is calculated from multiplication of total charge weighted negative surface area by the total molecular solvent-accessible surface area and dividing by 1000. *RPCS* is the relative positive charge surface area, which is calculated from solvent-accessible surface area of the most positive atom divided by relative positive charge, and *RNCG* is relative negative charge, which is calculated from charge of most negative atom divided by the total negative charge. The positive coefficient of $A \log P$ indicates that the increase in lipophilicity of these compounds increases the activity. However, the electronic descriptor *Dipole_mag* has negative influence on the inhibitory activity.

3.2.2. PLS

In case of PLS regression, Eq. (10) with 7 independent variables and 1 component (optimized with crossvalidation) was obtained.

$$\begin{aligned} pIC_{50} = & 4.257 + 0.139 A \log P + 0.206 A \log P98 \\ & + 0.018MolRef - 0.164Lumo + 0.005Jurs_PNSA_1 \\ & - 16.900Jurs_FPSA_3 + 0.008Jurs_WNSA_1 \end{aligned} \quad (13)$$

$$\begin{aligned} n_{\text{Training}} &= 52, R^2 = 0.536, R^2_{\text{adj}} = 0.527, F = 57.80, Q^2 \\ &= 0.500, \text{PRESS} = 45.497, n_{\text{Test}} = 18, R^2_{\text{pred}} \\ &= 0.784. \end{aligned}$$

This equation could explain 52.7% of the variance of the activity. Though the internal predictive capacity was not so good

($Q^2 = 0.5$), external validation characteristics are encouraging ($R^2_{\text{pred}} = 0.784$). Here both types of partition coefficient ($A \log P$ and $A \log P98$) and molar refractivity (*MolRef*) appear in the equation. The positive coefficients of these terms indicate that the activity increases with increase in lipophilicity and size of the molecules. *Jurs_PNSA_1* and *Jurs_WNSA_1* are responsible for the improvement of activity while *Jurs_FPSA_3* decreases it. *PNSA_1* is the acronym of partial negative surface area, which is the summation of solvent-accessible surface areas of all negatively charged atoms. *FPSA_1* is the fractional charged partial surface area, which is obtained by dividing partial positive surface area by the total molecular solvent-accessible surface area. *Lumo* is an acronym of lowest unoccupied molecular orbital energy and indicator of electrophilicity. The negative coefficient of *Lumo* indicates that the activity decreases with increase in electrophilicity of molecules.

3.2.3. FA-PLS

From the factor analysis on the data matrix using structural, electronic, spatial and physicochemical descriptors, it was observed that 7 factors could explain the data matrix to the extent of 95.5%. The anti-HIV activity was moderately loaded with factor 1 (loaded in *Hbondacceptor*, *MR*, *Dipole_mag*, *Jurs_PNSA_1*, *Jurs_DPSA_1*, *Jurs_PNSA_2*, *Jurs_DPSA_3*, *Jurs_FPSA_1*, *Jurs_FNSA_1*, *Jurs_FNSA_2*, *Jurs_FNSA_3*, *Jurs_WNSA_1*, *Jurs_WNSA_2*, *Jurs_WNSA_3*, *Jurs_TPSA*, *Jurs_TASA*, *Jurs_RPSA*, *Jurs_RASA*, *PMI_mag*), factor 3 ($A \log P$, $A \log P98$, *MolRef*, *Jurs_RPCG*, *Jurs_RPCS*), factor 10 (*Jurs_RNCS*) and weakly loaded with factor 4 (*Shadow_xzfrac*, *Shadow_nu*, *Shadow_zlength*), factor 5 (*Shadow_xyfrac*, *Shadow_ylength*), factor 6 (*Homo*) and factor 9 (*Jurs_FPSA_3*). The following FA-PLS model was developed with 7 variables and 3 crossvalidated components.

$$\begin{aligned} pIC_{50} = & 4.161 + 0.401 A \log P + 0.021MolRef \\ & - 0.254Dipole_mag - 18.170Jurs_FPSA_3 \\ & + 0.014Jurs_WNSA_1 - 0.069Jurs_WNSA_3 \\ & - 0.607Chiralcenters \end{aligned} \quad (14)$$

$$\begin{aligned} n_{\text{Training}} &= 52, R^2 = 0.685, R^2_{\text{adj}} = 0.665, F = 34.80, Q^2 \\ &= 0.628, \text{PRESS} = 33.827, n_{\text{Test}} = 18, R^2_{\text{pred}} \\ &= 0.698. \end{aligned}$$

Better explained variance ($R^2_{\text{adj}} = 66.5\%$) and predictivity ($Q^2 = 62.8\%$) were shown by this FA-PLS model in comparison to Eq. (13). However, the predicted R^2 for the test set was comparatively less (0.698 versus 0.784). Like Eqs. (12) and (13), $A \log P$, *MolRef*, *Dipole_mag*, *Jurs_FPSA_3* and *Jurs_WNSA_1* show positive effects on the activity. *Jurs_WNSA_3* has negative effects on the activity. *Jurs_WNSA_3* is surface weighted charged partial surface area, which is obtained by multiplying atomic charge weighted negative surface area by the total molecular solvent-accessible surface area and dividing by 1000. *Jurs_FPSA_3* is fractional charged partial surface area, which is calculated by dividing total charge weighted positive surface area with the total molecular solvent-accessible surface area. Higher number of chiral centers is detrimental to the anti-HIV activity of the TIBO derivatives.

3.2.4. GFA-MLR (50,000 iterations)

This GFA-MLR model was developed using 50,000 iterations with 5 variables. The coefficient of each variable is significant at 95% confidence interval, which is mentioned within parenthesis with respective variable.

$$pIC_{50} = 4.139 - 0.188(\pm 0.137)Jurs_WNSA_3 \\ - 0.837(\pm 0.841)Chiralcenters \\ + 0.659(\pm 0.440)A \log P - 0.271(\pm 0.246)Dipole_mag \\ - 1.174(\pm 1.670)Hbondacceptor \quad (15)$$

$$n_{Training} = 52, R^2 = 0.730, R_{adj}^2 = 0.700, F = 24.83, s \\ = 0.731, LOF = 0.725, Q^2 = 0.622, PRESS \\ = 34.425, n_{Test} = 18, R_{pred}^2 = 0.650.$$

Presence of hydrogen bond acceptor groups, number of chiral centers, *Jurs_WNSA_3* and *Dipole_mag* reduce the activity while the thermodynamic descriptor (*A log P*) is responsible for improvement of activity according to the coefficients of respective variables. The model could explain and predict 70% and 62.2%, respectively, of the variance of the activity. The predicted R^2 value for the test set was 0.650.

3.2.5. G/PLS (5000 iterations)

Eq. (16) was obtained with 8 variables and 3 components (crossvalidated by internal validation).

$$pIC_{50} = 7.643 - 0.300Lumo + 0.649 A \log P \\ + 0.043Jurs_DPSA_3 - 0.257Dipole_mag - 0.158Homo \\ - 65.209Jurs_FPSA_3 - 0.581Chiralcenters \\ + 0.662Jurs_RPCS \quad (16)$$

$$n_{Training} = 52, R^2 = 0.727, R_{adj}^2 = 0.710, F = 42.54, Q^2 \\ = 0.625, LSE = 0.454, PRESS = 34.110, n_{Test} \\ = 18, R_{pred}^2 = 0.492.$$

Although this equation could explain and predict 71% and 62.5%, respectively, of the variance external predictivity is not encouraging ($R_{pred}^2 = 0.492$). *Lumo*, *Dipole_mag*, *Homo*, *Jurs_FPSA_3* and *Chiralcenters* are unfavorable but *A log P*, *Jurs_DPSA_3* and *Jurs_RPCS* are favorable for the anti-HIV activity. *Homo* is an acronym for highest occupied molecular orbital energy and represents nucleophilicity. *DPSA_3* is the difference between atomic charge weighted positive solvent-accessible surface area and atomic charge weighted negative solvent-accessible surface area.

3.2.6. GFA (spline) (50,000 iterations)

Eq. (17) was developed with 4 variables using 50,000 iterations.

$$pIC_{50} = 7.234 + 7.786(Jurs_FNFA_2 + 1.184) \\ - 7.910(0.457 - Jurs_RPCS) \\ - 1.360(10.544 - Shadow_ylength) \\ - 0.110(141.294 - Jurs_PNSA_1) \quad (17)$$

$$n_{Training} = 52, R^2 = 0.786, R_{adj}^2 = 0.767, F = 43.05, s \\ = 0.644, LOF = 0.634, Q^2 = 0.740, PRESS \\ = 23.699, n_{Test} = 18, R_{pred}^2 = 0.587.$$

The coefficient of (*Jurs_FNFA_2 + 1.184*) indicates that the activity value is facilitated when the *Jurs_FNFA_2* values are less negative (i.e., higher) than -1.184 . *Jurs_FNFA_2* is fractional negatively charged partial surface area, which is obtained from dividing total

charge weighted negative surface area by the total molecular solvent-accessible surface area. The values of other 3 coefficients show that the value of *Jurs_RPCS* below 0.457, the value of *Jurs_PNSA_1* below 141.294 and the value of *Shadow_ylength* below 10.544 are detrimental for optimum activity. Shadow descriptors are used to characterize the shape of the molecules. These are calculated by projecting the molecular surface on 3 mutually perpendicular planes *XY*, *YZ*, *ZX*. *Shadow_ylength* is the length of molecule in the *Y* dimension. This equation could explain and predict 76.7% and 74%, respectively, of the variance of the anti-HIV activity.

3.3. QSAR using combined (topological, structural, physicochemical, spatial and electronic) set of descriptors

3.3.1. Stepwise regression

Eq. (18) consisting of 5 independent variables was developed from stepwise regression. Here, the combined pool of descriptors was subjected to *F* criterion ($F = 4$ for inclusion; $F = 3.9$ for exclusion) to get an equation in a stepwise manner.

$$pIC_{50} = 2.146 + 0.070(\pm 0.036)Jurs_WNSA_1 \\ - 0.141(\pm 0.117)S_{dO} - 0.312(\pm 0.262)Dipole_mag \\ + 1.61(\pm 1.981)Jurs_RPCS - 0.51(\pm 0.817)S_{tsc} \quad (18)$$

$$n_{Training} = 52, R^2 = 0.718, R_{adj}^2 = 0.687, F = 23.40, s \\ = 0.747, Q^2 = 0.629, PRESS = 33.755, n_{Test} \\ = 18, R_{pred}^2 = 0.686.$$

The ability of the model to explain ($R_{adj}^2 = 0.687$) and predict ($Q^2 = 0.629$) was better than Eq. (12) but inferior to Eq. (6). However, the predicted R^2 value for Eq. (18) was better than those of both Eqs. (6) and (12). The activity increases with larger values of *Jurs_WNSA_1*, *Jurs_RPCS* and lower values of *S_dO*, *Dipole_mag* and electronic and topological environment of E-state index (S_{tsc}).

3.3.2. PLS

Eq. (19) composed of 7 variables was obtained using 1 component.

$$pIC_{50} = 2.911 + 0.614^3 k_{am} + 0.198 A \log P - 0.233Lumo \\ + 0.201Shadow_ylength - 0.164S_{aaCH} + 0.313S_{dssc} \\ - 0.032S_{dO} \quad (19)$$

$$n_{Training} = 52, R^2 = 0.621, R_{adj}^2 = 0.613, F = 81.94, Q^2 \\ = 0.577, PRESS = 38.489, n_{Test} = 18, R_{pred}^2 \\ = 0.741.$$

Although this equation could explain 61.3%, which was inferior to Eq. (7) the predictive potential ($Q^2 = 0.577$) was superior to both Eqs. (7) and (13). The predicted R^2 value for the test set was more than that of Eq. (7) and less than that of Eq. (13). Increase of kappa alpha-modified shape index, *A log P*, *Shadow_ylength*, S_{dssc} and decrease of *Lumo*, S_{aaCH} and S_{dO} are helpful for the optimum activity.

3.3.3. FA-PLS

From factor analysis it was observed that 7 factors could explain the data matrix to the extent of 95.8%. Anti-HIV activity of

Table 5
Comparative study of statistical quality of QSAR models obtained using different statistical tools

Type of descriptors	Statistical method	R ² (training set)	R _a ² (training set)	Q ² (training set)	F	PRESS	R _{pred} ² (test set)
Topological + structural	Stepwise	0.765	0.728	0.685	20.50	28.685	0.536
	PLS	0.639	0.624	0.499	43.33	45.600	0.619
	FA-PLS	0.636	0.621	0.526	42.72	43.158	0.707
	GFA-MLR	0.707	0.682	0.649	28.29	31.965	0.533
	G/PLS	0.628	0.605	0.561	27.78	32.071	0.568
	GFA (spline)	0.791	0.768	0.737	34.72	23.886	0.689
Structural + physicochemical + spatial + electronic	Stepwise	0.709	0.677	0.606	22.41	35.837	0.154
	PLS	0.536	0.527	0.500	57.80	45.497	0.784
	FA-PLS	0.685	0.665	0.628	34.80	33.827	0.698
	GFA-MLR	0.730	0.700	0.622	24.83	34.425	0.650
	G/PLS	0.727	0.710	0.625	42.54	34.110	0.492
	GFA (spline)	0.786	0.767	0.740	43.05	23.699	0.587
Topological + structural + physicochemical + spatial + electronic	Stepwise	0.718	0.687	0.629	23.40	33.755	0.686
	PLS	0.621	0.613	0.577	81.94	38.489	0.741
	FA-PLS	0.656	0.642	0.575	46.68	38.647	0.674
	GFA-MLR	0.816	0.792	0.759	33.29	21.956	0.529
	G/PLS	0.773	0.754	0.704	40.10	26.878	0.517
	GFA (spline)	0.835	0.821	0.800	59.44	18.164	0.612

TIBO derivatives was moderately loaded with factor 2 (loaded in *Jurs_FNSA_2*, *Jurs_FPSA_3*, *Jurs_FNSA_3*, *Jurs_WNSA_1*, *Jurs_WNSA_2*, *Jurs_WNSA_3*, *Jurs_RNCS*, *Jurs_TPSA*, *Jurs_TASA*, *Jurs_RPSA*, *Jurs_RASA*, *S_aaCH*, *S_sCl*), factor 3 (*A log P*) and factor 10 (*Jurs_RNCG*) and poorly loaded with factor 1 (*¹k*, *²k*, *³k*, *¹k_{am}*, *²k_{am}*, *³k_{am}*, *Φ*, *SC_0*, *SC_1*, *SC_2*, *SC_3_P*, *SC_3_C*, *⁰χ*, *¹χ*, *²χ*, *³χ*, *³χ_p*, *⁰χ_v*, *¹χ_v*, *²χ_v*, *³χ_v*, *³χ_v_C*, *Wiener*, *log Z*, *Zagreb*, *MW*, *log P*, *Jurs_SASA*, *Jurs_PPSA_2*, *Jurs_PNSA_2*, *Jurs_PPSA_3*, *Jurs_WPSA_1*, *Jurs_WPSA_2*, *Jurs_WPSA_3*, *Jurs_RNCG*, *Shadow_{XY}*, *Shadow_{XZ}*, *Area*, *V_m*, *S_{sCH3}*, *S_{dsCH}*), factor 5 (*Homo*, *S_{sBr}*), factor 6 (*Shadow_{XYfrac}*, *S_{tCH}*, *S_{tsC}*) and factor 9 (*Shadow_{Ylength}*). The following FA-PLS equation comprising 9 variables was obtained with 2 components (optimized by crossvalidation).

$$\begin{aligned}
 pIC_{50} = & 5.747 + 0.578^3 k_{am} + 0.232^2 \chi^v - 0.038 Jurs - PPSA_3 \\
 & + 25.091 Jurs_{RNCG} - 26.744 Jurs - FPSA_3 \\
 & - 0.164 S_{aaCH} + 0.236 A \log P - 3.941 Shadow_{XYfrac} \\
 & + 0.194 Shadow_{Ylength}
 \end{aligned}
 \quad (20)$$

$$\begin{aligned}
 n_{Training} = & 52, R^2 = 0.656, R_{adj}^2 = 0.642, F = 46.68, Q^2 \\
 = & 0.575, PRESS = 38.647, n_{Test} = 18, R_{pred}^2 \\
 = & 0.674.
 \end{aligned}$$

This equation could explain and predict 64.2% and 57.5%, respectively, of variance of the activity. Here kappa alpha-modified shape index of 3rd order, molecular valence connectivity index of 2nd order, *Jurs_RNCG*, *A log P* and *Shadow_{Ylength}* have positive influence and *Jurs_PPSA_3*, *Jurs_FPSA_3*, *S_{aaCH}* and *Shadow_{XYfrac}* have negative influence on the biological activity. *Jurs_PPSA_3* is the summation of the product of solvent-accessible surface area and partial charge for all positively charged atoms.

3.3.4. GFA-MLR (50,000 iterations)

The following equation was built up with 6 variables using 50,000 iterations.

$$\begin{aligned}
 pIC_{50} = & 28.739 - 1.155 (\pm 0.491) S_{aaCH} \\
 & + 2.576 (\pm 1.345) S_{sssCH} + 0.900 (\pm 0.713) S_{dssC} \\
 & + 2.483 (\pm 2.191)^3 k_{am} - 0.251 (\pm 0.181) Dipole_{mag} \\
 & - 2.430 (\pm 1.353) \log Z
 \end{aligned}
 \quad (21)$$

$$\begin{aligned}
 n_{Training} = & 52, R^2 = 0.816, R_{adj}^2 = 0.792, F = 33.29, s \\
 = & 0.610, LOF = 0.544, Q^2 = 0.759, PRESS \\
 = & 21.956, n_{Test} = 18, R_{pred}^2 = 0.529.
 \end{aligned}$$

Table 6
Lists of selected variables in the different equations

Equation number	Selected variables
Eq. (6)	<i>S_{dssC}</i> , <i>S_{aaCH}</i> , <i>³χ_{ch}</i> , <i>SC-3_P</i> , <i>³χ_v</i> , <i>Jx</i> , <i>Hbondacceptor</i>
Eq. (7)	<i>SC-3_P</i> , <i>³χ_v</i> , <i>MW</i> , <i>S_{aaCH}</i> , <i>S_{aasC}</i> , <i>S_{dO}</i> , <i>S_{ds}</i>
Eq. (8)	<i>³k_{am}</i> , <i>SC-3_P</i> , <i>MW</i> , <i>Chiralcenters</i> , <i>S_{dsCH}</i> , <i>S_{aaCH}</i> , <i>S_{ssNH}</i> , <i>S_{dO}</i>
Eq. (9)	<i>S_{sssCH}</i> , <i>S_{aaCH}</i> , <i>S_{dssC}</i> , <i>SC-3_P</i>
Eq. (10)	<i>Zagreb</i> , <i>S_{aaCH}</i> , <i>SC-3_P</i> , <i>S_{dO}</i>
Eq. (11)	(<i>Chiralcenters-1</i>), (<i>2.58⁻³k_{am}</i>), (<i>Rotlbonds-2</i>), (<i>11.909-S_{dO}</i>), (<i>14.917^{-k}</i>)
Eq. (12)	<i>Jurs_WNSA_1</i> , <i>Dipole_{mag}</i> , <i>A log P</i> , <i>Jurs_RPCS</i> , <i>Jurs_RNCG</i>
Eq. (13)	<i>A log P</i> , <i>A log P98</i> , <i>MolRef</i> , <i>Lumo</i> , <i>Jurs_PNSA_1</i> , <i>Jurs_FPSA_3</i> , <i>Jurs_WNSA_1</i>
Eq. (14)	<i>A log P</i> , <i>MolRef</i> , <i>Dipole_{mag}</i> , <i>Jurs_FPSA_3</i> , <i>Jurs_WNSA_1</i> , <i>Jurs_WNSA_3</i> , <i>Chiralcenters</i>
Eq. (15)	<i>Jurs_WNSA_3</i> , <i>Chiralcenters</i> , <i>A log P</i> , <i>Dipole_{mag}</i> , <i>Hbondacceptor</i>
Eq. (16)	<i>Lumo</i> , <i>A log P</i> , <i>Jurs_DPSA_3</i> , <i>Dipole_{mag}</i> , <i>Homo</i> , <i>Jurs_FPSA_3</i> , <i>Chiralcenters</i> , <i>Jurs_RPCS</i>
Eq. (17)	(<i>Jurs_FNSA_2 + 1.184</i>), (<i>0.457-Jurs_RPCS</i>), (<i>10.544-Shadow_{Ylength}</i>), (<i>141.294-Jurs_PNSA_1</i>)
Eq. (18)	<i>Jurs_WNSA_1</i> , <i>S_{dO}</i> , <i>Dipole_{mag}</i> , <i>Jurs_RPCS</i> , <i>S_{tsC}</i>
Eq. (19)	<i>³k_{am}</i> , <i>A log P</i> , <i>Lumo</i> , <i>Shadow_{Ylength}</i> , <i>S_{aaCH}</i> , <i>S_{dssC}</i> , <i>S_{dO}</i>
Eq. (20)	<i>³k_{am}</i> , <i>²χ_v</i> , <i>Jurs_PPSA_3</i> , <i>Jurs_RNCG</i> , <i>Jurs_FPSA_3</i> , <i>S_{aaCH}</i> , <i>A log P</i> , <i>Shadow_{XYfrac}</i> , <i>Shadow_{Ylength}</i>
Eq. (21)	<i>S_{aaCH}</i> , <i>S_{sssCH}</i> , <i>S_{dssC}</i> , <i>³k_{am}</i> , <i>Dipole_{mag}</i> , <i>log Z</i>
Eq. (22)	<i>S_{ssNH}</i> , <i>³k_{am}</i> , <i>³χ_v</i> , <i>Dipole_{mag}</i> , <i>S_{aaCH}</i>
Eq. (23)	(<i>10.631-Shadow_{Ylength}</i>), (<i>0.434-Jurs_RPCS</i>), (<i>SC-3_P-44</i>), (<i>138.17-Jurs_PNSA_1</i>)

Table 7
Comparative study of developed networks with different architecture

Model no.	No. of hidden layers	No. of units in different layers	No. of crossvalidated resampling	No. of epochs in backpropagation followed by conjugate gradient descent	Absolute error mean	R^2_{pred} (test set)	Correlation coefficient (r^2) between obs. and pred. values of the test set
1	3	48 54 15		500,200	0.547	0.760	0.841
2	3	48 54 20		800,300	0.660	0.690	0.765
3	2	49 15		700,200	0.617	0.736	0.799
4	2	58 15		500,200	0.568	0.739	0.780
5	1	43 15		800,300	0.654	0.673	0.771
6	1	46 15		500,200	0.581	0.707	0.712
7	3	6 5 15		500,200	0.601	0.701	0.706
8	3	4 6 15		500,200	0.607	0.711	0.720
9	2	6 20		800,300	0.573	0.731	0.731
10	2	6 20		700, 200	0.586	0.696	0.693
11	1	4 15		500, 200	0.628	0.653	0.648
12	1	5 20		800, 300	0.602	0.652	0.650

Predictivity ($Q^2 = 0.759$) and explained variance ($R^2_{\text{adj}} = 0.792$) of this model were better than both Eqs. (9) and (15). However, the predicted R^2 value for the test set was inferior to those of both Eqs. (9) and (15). The 95% confidence interval of each independent variable is mentioned within parenthesis. Here S_{aaCH} , $\log Z$ and $Dipole_{mag}$ are responsible for lowering of activity whereas S_{sssCH} , S_{dssC} and kappa alpha-modified shape index of 3rd order decline inhibitory activity of compounds. $\log Z$ is the logarithm of Hosoya index. It is related to the number of edges in the graph.

3.3.5. G/PLS (5000 iterations)

This G/PLS model included 6 variables for the development of model with 4 components.

$$pIC_{50} = 4.930 + 2.400S_{ssN} + 5.458^3\kappa_{am} - 1.924^2\kappa - 0.876^2\chi^v - 0.259Dipole_{mag} - 0.815S_{aaCH} \quad (22)$$

$$n_{\text{Training}} = 52, R^2 = 0.773, R^2_{\text{adj}} = 0.754, F = 40.10, Q^2 = 0.704, \text{LSE} = 0.396, \text{PRESS} = 26.878, n_{\text{Test}} = 18, R^2_{\text{pred}} = 0.517.$$

In Eq. (22) values of R^2_{adj} (0.754) and Q^2 (0.704) were better than both Eqs. (10) and (16). However, the predicted R^2 value for the test set was less than that of Eq. (10) and more than that of Eq. (16). Here, kappa shape index of 2nd order, valence molecular connectivity index of 2nd order, $Dipole_{mag}$, E-state index (S_{aaCH}) are responsible for lowering of anti-HIV activity whereas E-state index (S_{sssN}), kappa alpha-modified shape index of 3rd order have positive impact on the activity.

3.3.6. GFA (spline) (50,000 iterations)

Eq. (23) was developed with 4 variables using 50,000 iterations.

Table 8
Comparison of external predictability characteristics (test set) of different models developed from the training set

Type of descriptors	Statistical method	r^2	r^2_0	$(r^2 - r^2_0)/r^2$	r^2_m
Topological + structural	Stepwise	0.565	0.536	0.052	0.468
	PLS	0.615	0.615	2.6×10^{-5}	0.613
	FA-PLS	0.753	0.743	0.013	0.677
	GFA-MLR	0.540	0.528	0.022	0.481
	G/PLS	0.574	0.566	0.014	0.523
	GFA (spline)	0.687	0.685	0.003	0.657
Structural + physicochemical + spatial + electronic	Stepwise	0.455	0.190	0.582	0.221
	PLS	0.847	0.794	0.060	0.655
	FA-PLS	0.744	0.725	0.025	0.642
	GFA-MLR	0.696	0.657	0.055	0.560
	G/PLS	0.719	0.555	0.229	0.427
	GFA (spline)	0.645	0.582	0.099	0.483
Topological (E-state) + structural + physicochemical + spatial + electronic	Stepwise	0.734	0.685	0.066	0.572
	PLS	0.756	0.737	0.025	0.653
	FA-PLS	0.680	0.679	0.002	0.655
	GFA-MLR	0.545	0.525	0.036	0.468
	G/PLS	0.525	0.510	0.029	0.460
	GFA (spline)	0.667	0.612	0.083	0.510

Table 9
Comparison of external predictability characteristics (test set) of different ANN models

Model no.	r^2	r_0^2	$(r^2 - r_0^2)/r^2$	r_m^2
1	0.841	0.778	0.074	0.630
2	0.765	0.692	0.096	0.558
3	0.799	0.732	0.084	0.592
4	0.780	0.735	0.058	0.615
5	0.771	0.669	0.132	0.525
6	0.712	0.711	0.001	0.698
7	0.706	0.697	0.014	0.637
8	0.720	0.707	0.018	0.639
9^a	0.731	0.730	0.001	0.714
10	0.693	0.692	0.001	0.673
11	0.648	0.647	0.001	0.634
12	0.650	0.648	0.004	0.618

^a The best model based on r_m^2 value is shown in bold face.

$$\begin{aligned}
 pIC_{50} = & 7.967 - 1.219(10.631 - Shadow_{Ylength}) \\
 & - 8.111(0.434 - Jurs_RPCS) - 0.224(SC - 3_P - 44) \\
 & - 0.079(138.17 - Jurs_PNSA_1)
 \end{aligned}
 \quad (23)$$

$$\begin{aligned}
 n_{Training} = & 52, R^2 = 0.835, R_{adj}^2 = 0.821, F = 59.44, s \\
 & = 0.565, LOF = 0.488, Q^2 = 0.800, PRESS \\
 & = 18.164, n_{Test} = 18, R_{pred}^2 = 0.612.
 \end{aligned}$$

All independent variables in this model have negative coefficients. So, the difference between a variable and corresponding constant term should be negative for getting optimum activity. It could explain and predict 82.1% and 80.0%, respectively, which were better than previous 2 GFA (spline) models. However, the predicted R^2 value for the test set was inferior to that of Eq. (11) and better than that of Eq. (17).

A comparative study of statistical quality of different equations has been shown in Table 5. A list of descriptors appearing in different equations is given in Table 6. Among 18 reported equations, 9 equations contain the descriptor S_{aaCH} (the most frequently occurring descriptor). The number of occurrence of each of the descriptors $Dipole_mag$ and $A \log P$ is 7, while each of the descriptors $SC-3_P$ and S_{dO} occurs 6 times. Each of the descriptors $Jurs_RPCS$, $Chiralcenters$ and $^3k_{am}$ appear 5 times in 18 equations. Some of the descriptors, which never appeared in the reported 18 equations are $Apol$, $Area$, Vm and $Density$.

3.3.7. Artificial neural networks

Apart from GFA (spline), another non-linear modeling (ANN) has been performed to search for better predictive models. Initially, all descriptors were used for this non-linear modeling. Selected descriptors based on the equations of linear modeling were then tried for ANN model development. The network has been formed with the training set using backpropagation in the 1st phase and conjugate gradient descent in the 2nd phase. The developed network was used to estimate the biological activity of the test set compounds. Using different numbers of iterations in backpropagation and conjugate gradient descent methods, different numbers of hidden layers and units per layer, a number of models were developed. In this study we have changed a particular parameter fixing others to get an optimum network architecture. We have presented 12 selected networks using different iterations and different hidden layers in Table 7. The first 6 models used higher number of hidden nodes while the remaining six models used comparatively lower number of hidden nodes. Although best r^2 value (squared correlation coefficient between observed and

predicted values of the test set compounds) [49] was obtained from 1st ANN model 9th ANN model provided the best r_m^2 value (modified r^2 as defined in Ref. [50]; also see Section 3.4). In general, higher r^2 values were obtained from the ANN models with comparatively higher number of hidden nodes while higher r_m^2 values were obtained from those with comparatively lower number of hidden nodes. We have selected the best model on the basis of the best r_m^2 value. In that network (ANN model 9), two hidden layers of 6 and 5 elements were used. The number of iterations selected for backpropagation and conjugate gradient descent were 800 and 300, respectively. Initialization method selected for network was "random uniform". Weight decay was regularized in both phases (decay factor = 0.01, scale factor = 1). Learning rate and momentum of each epoch were adjusted to 0.01 and 0.3, respectively. The number of crossvalidated resampling was set to 20 for this model. ANN models developed from reduced set of descriptors (those appearing in the equations) showed poor predictive ability in comparison to those developed from the whole pool of descriptors.

3.4. Further tests on external predictability

Although R_{pred}^2 is used to test predictive ability on new chemical entities this parameter is influenced by the value of $\sum(Y_{test} - \bar{Y}_{training})^2$, i.e., the sum of differences between observed values of test set compounds and mean of training data set. Thus, this parameter may not truly reflect the predictive capability of the model on a new data set, as it is dependent on selection of training set compounds. So, squared correlation coefficient values between the observed and predicted values of the test set compounds with intercept (r^2) and without intercept (r_0^2) were calculated to know performance of the prediction. These values of all models have been represented in Tables 8 and 9. All the models (except Eqs. (7) and

Table 10
Calculated values of k and k' for different models as defined by Golbraikh and Tropsha [49]

Type of descriptors	Statistical method	k	k'
Topological + structural	Stepwise	1.019	0.959
	PLS	1.010	0.972
	FA-PLS	1.047	0.943
	GFA-MLR	0.990	0.986
	G/PLS	0.993	0.985
	GFA (spline)	0.993	0.991
Structural + physicochemical + spatial + electronic	Stepwise	0.952	1.008
	PLS	1.027	0.964
	FA-PLS	1.041	0.947
	GFA-MLR	1.026	0.958
	G/PLS	1.063	0.920
	GFA (spline)	1.006	0.973
Topological (E-state) + structural + physicochemical + spatial + electronic	Stepwise	1.016	0.970
	PLS	1.002	0.986
	FA-PLS	1.023	0.962
	GFA-MLR	0.987	0.989
	G/PLS	1.003	0.973
	GFA (spline)	0.983	0.998
ANN model no. 1		1.035	0.956
ANN model no. 2		1.019	0.967
ANN model no. 3		1.005	0.982
ANN model no. 4		1.000	0.987
ANN model no. 5		1.008	0.976
ANN model no. 6		1.022	0.965
ANN model no. 7		1.005	0.981
ANN model no. 8		1.001	0.985
ANN model no. 9		1.014	0.973
ANN model no. 10		1.004	0.981
ANN model no. 11		1.002	0.981
ANN model no. 12		1.008	0.975

Table 11

List of best five r_m^2 values [50] between observed and predicted values of the test set compounds obtained from models using different techniques

Statistical methods	r_m^2 value
ANN (model no. 9)	0.714
ANN (model no. 6)	0.698
FA-PLS (Eq. (8))	0.677
ANN (model no. 10)	0.673
GFA (spline) (Eq. (11))	0.657

(11) and 5th ANN model) have satisfied the requirement of the value of $(r^2 - r_0^2)/r^2$ being less than 0.1 as recommended by Golbraikh and Tropsha [49]. According to them models are considered acceptable, if they satisfy all of the following conditions: (i) $Q^2 > 0.5$, (ii) $r^2 > 0.6$, (iii) r_0^2 or $r_0'^2$ is close to r^2 , such that $[(r^2 - r_0^2)/r^2]$ or $[(r^2 - r_0'^2)/r^2] < 0.1$ and $0.85 \leq k \leq 1.15$ or $0.85 \leq k' \leq 1.15$. When the observed values of the test set compounds (Y axis) are plotted against the predicted values of the compounds (X axis) setting intercept to zero, slope of the fitted line gives the value of k . Interchange of the axes gives the value of k' . A list of values of k' and k for different models is given in Table 10.

Besides this, a high value of squared regression coefficient (r^2) between observed and predicted values of the test set compounds does not necessarily mean that the predicted values are very near to observed activity (there may be considerable numerical difference between the values though maintaining an overall good intercorrelation). To better indicate external predictive capacity of a model a modified r^2 term (r_m^2) was defined in the following manner [50].

$$r_m^2 = r^2 \left(1 - \sqrt{r^2 - r_0^2} \right)$$

In case of good external prediction, predicted values will be very close to observed activity values. So, r^2 value will be very near to r_0^2 value. In the best case r_m^2 will be equal to r^2 . Here, the r_m^2 values of Eqs. (6), (9), (12), (16), (17), (21) and (22) are less than the recommended value (0.5). The best r_m^2 value was obtained from 9th ANN model (Table 11). A comparison of the present models with those published earlier [51–54] is given in Table 12. It may be mentioned here that Huuskonen [23] developed QSAR models based on E-state indices of the common atoms of the TIBO derivatives (in addition to log P and MR values) while we have used E-state indices based on atom types (defined in Table 3) in this paper.

3.5. Randomization test for the genetic models

Genetic analysis is a flexible modeling technique, which may lead to over-trained models. To check absence of over-training, we

have randomized the model development process at the selected levels of iterations. The results of such tests are given in Table 13, which show that mean values of the correlation coefficients (R) of the random models are significantly lower than those of the corresponding nonrandom models. This proves that the models are not over-trained.

3.6. True external validation of the selected models

In order to further validate selected models, a data set of reverse transcriptase inhibitor TIBO derivatives from a different source [31] was considered and it was found that 14 compounds reported in that communication [31] were not included in the present data set. To check the domain of applicability, a cluster analysis was performed on the training set combined with the external set of 14 compounds. Hierarchical clustering was done using centroid as the linkage method and Euclidean as the distance measure. Out of 5 clusters generated, 3 clusters did not contain any training set compounds. Due to the absence of training set compounds in these clusters, external set compounds (four in number) present in these 3 clusters are not represented by the training set compounds. The prediction of the response using models is valid only if the compound being predicted is within the applicability domain of the model. Thus, these four compounds (S. nos. 15c, 18a, 18b and 18c of the original paper [31]) were excluded from the external set and the remaining 10 compounds were predicted using selected models (Table 14). The squared correlation coefficients (r^2) between observed and predicted activity values of the external set compounds and root mean square error of prediction (RMSEP) were reported for selected models (Table 15). The RMSEP values were calculated according to following equation:

$$\text{RMSEP} = \sqrt{\frac{(Y_{\text{obs}} - Y_{\text{pred}})^2}{N_{\text{Test}}}} \quad (24)$$

In the above equation, N_{Test} indicates the number of test set compounds. Acceptable r^2 values as reported in Table 15 suggest the predictive potential of the models. It is interesting to note that the best r^2 value (external set) was given by the model (ANN model 9) showing the best r_m^2 value (test set).

4. Overview

Different statistical methods like stepwise regression, partial least squares (PLS), combination of factor analysis and PLS (FA-PLS), genetic function approximation (GFA) coupled with multiple linear regression have been applied to model tetrahydroimidazo[4,5,1-*jk*][1,4]benzodiazepine or TIBO derivatives as reverse transcriptase

Table 12

Comparative study of different models on TIBO derivatives reported by different researchers

Reference	Modeling technique	Descriptors	No. of compounds	R^2	Q^2	F	R_{pred}^2	r^2 (test)
Zhou and Madura [51]	CoMFA	Electrostatic and steric field descriptors	50 (no division into training and test sets)	0.972	0.704	N/A	N/A	N/A
	CoMSiA	Similarity (electrostatic, steric, hydrophobic, hydrogen bond donor and acceptor) descriptors	50 (no division into training and test sets)	0.944	0.776	N/A	N/A	N/A
Toropov et al. [52]	OCWLG1	Descriptors from graph of atomic orbitals	37 (training), 20 (test)	0.888	N/A	279.0	0.885	N/A
Hannongbua et al. [53]	CoMFA	Electrostatic and steric field descriptors	46 (training), 24 (test)	0.941	0.771	195.0	0.870	N/A
Huuskonen [23]	MLR	E-state indices + log P + MR	46 (training), 24 (test)	0.737	0.663	22.4	0.800	N/A
Solov'ev and Varnek [54]	SMF	Topological	64 (training), 7 (test)	0.920	0.736	7.1	0.859	N/A
Present study	GFA (Spline) [Eq. (23)]	Topological, structural, electronic, spatial and physicochemical	52 (training), 18 (test)	0.835	0.800	59.44	0.612	0.667
	ANN (model 9)	Topological, structural, electronic, spatial and physicochemical	52 (training), 18 (test)	N/A	N/A	N/A	0.731	0.731

N/A = not available.

Table 13
Results of randomization test applied on genetic model development process

Eq. no.	Modeling technique	R from nonrandom model	Confidence level (%)	Mean value of R from random trials \pm standard deviation
(9)	GFA-MLR	0.841	90	0.371 \pm 0.170
(10)	G/PLS	0.792	90	0.499 \pm 0.045
(11)	GFA (spline)	0.889	90	0.436 \pm 0.155
(15)	GFA-MLR	0.854	90	0.337 \pm 0.132
(16)	G/PLS	0.853	90	0.582 \pm 0.054
(17)	GFA (spline)	0.887	90	0.559 \pm 0.170
(21)	GFA-MLR	0.903	90	0.391 \pm 0.167
(22)	G/PLS	0.879	90	0.601 \pm 0.077
(23)	GFA (spline)	0.914	90	0.562 \pm 0.174

inhibitors ($n = 70$) using different combinations of topological, structural, physicochemical, electronic and spatial descriptors. Non-linear models have also been developed from GFA (spline) and artificial neural networks (ANN). The best equation using topological and structural descriptors was developed from genetic function approximation (spline) ($Q^2 = 0.737$) based on internal validation. When the developed equations were applied on the test set, the FA-PLS method provided the best results ($R^2_{\text{pred}} = 0.707$). The second set of descriptors was combination of structural, electronic, spatial and physicochemical ones. Here external predictive capacity (test set) of models developed from stepwise regression and G/PLS methods has unacceptable values though the leave-one-out predicted variance (Q^2) values are moderately good. This reconfirms the absence of any correlation between the internal and external validation statistics [30]. Probably, the combinations of variables selected in these cases are good only for fitting data and not for prediction. In this set, GFA (spline) provided the best internally crossvalidated squared correlation coefficient ($Q^2 = 0.740$), whereas the best externally (test set) validated R^2 (0.784) result was obtained from the PLS method. When all descriptors were combined there was improvement in best Q^2 value (0.800 from GFA (spline)), but the best R^2_{pred} (0.760 from ANN) value deteriorated. Based on internal validation (Q^2 value), Eq. (23) [GFA (spline) model developed from the combined pool of descriptors] was the best model. All equations, except Eq. (12), showed acceptable values of squared correlation coefficient with intercept and without intercept for the observed and predicted values of the test set compounds. Further statistical validation was performed as recommended by Golbraikh and Tropsha [49] and Roy and Roy [50]. Except Eqs. (7) and (11) and 5th ANN model, all models have satisfied the criteria of $(r^2 - r_0^2)/r^2$ value being less than 0.1. When modified squared correlation coefficient (externally

Table 14
Structural features, observed and predicted reverse transcriptase inhibitory activities of external set compounds [31] using selected models

S. no.	R1	X	R2	R3	Activity [$\log(1/IC_{50})$]		
					Obs [31]	Pred ^a	Pred ^b
T1	5-Me (S)	S	8-SCH ₃	DMA	8.301	7.606	7.483
T2	5-Me (S)	S	8-OEt	DMA	7.018	7.486	7.176
T3	5-Me (S)	O	8-CONH ₂	DMA	5.197	4.840	5.625
T4	5-Me (S)	O	9-NO ₂	CH ₂ CH(CH ₂) ₂	4.476	3.492	5.144
T5	5-Me (S)	O	8-NH ₂	CH ₂ CH(CH ₂) ₂	3.071	4.899	5.966
T6	5-Me (S)	O	8-N(CH ₃) ₂	CH ₂ CH(CH ₂) ₂	5.177	5.415	5.799
T7	5-Me (S)	O	9-NH ₂	CH ₂ CH(CH ₂) ₂	4.218	4.981	5.857
T8	5-Me (S)	O	9-N(CH ₃) ₂	CH ₂ CH(CH ₂) ₂	5.177	5.482	5.768
T9	5-Me (S)	S	9-NO ₂	CH ₂ CH(CH ₂) ₂	5.611	4.869	6.471
T10	5-Me (S)	S	9-F	DMA	7.602	6.506	7.588

^a Obtained from the best linear model (Eq. (18)) [based on r_m^2 for the test set].

^b Obtained from the best non-linear model (ANN model 9) [based on r_m^2 for the test set].

Table 15
Comparative study of r^2 values and RMSEP values for prediction of external set compounds using selected models

Statistical methods	r^2 value	RMSEP value
ANN (model no. 9)	0.724	1.178
ANN (model no. 6)	0.622	1.120
FA-PLS (Eq. (8))	0.675	0.873
ANN (model no. 10)	0.718	1.133
GFA (spline) (Eq. (11))	0.667	1.012

validated) or r_m^2 values are calculated majority of models (except Eqs. (6), (9), (12), (16), (17), (21) and (22)) have exceeded recommended value (0.5). The best r_m^2 value was obtained from 9th ANN model. Further, higher r^2 values were obtained from the ANN models with comparatively higher number of hidden nodes, while higher r_m^2 values were obtained from those with comparatively lower number of hidden nodes. A comparative study of the present models with previous models [51–54] on TIBO derivatives published by different researchers has been represented in Table 11. Scatter plots of observed versus calculated or predicted values of the training and test set compounds obtained from the best two models (based on r_m^2 value) are given in Figs. 1 and 2.

The descriptor S_{aaCH} was found to be the most frequently occurring (9 times) descriptor in 18 reported equations. Some of the other frequently occurring descriptors in 18 equations are $Dipole_mag$ and $A \log P$ (7 out of 18), $SC-3_P$ and S_dO (6 out of 18), $Jurs_RPCS$, $Chiralcenters$ and $^3k_{am}$ (5 out of 18). An attempt was also made to develop ANN models using descriptors appearing in the equations. ANN models obtained from such reduced set of descriptors showed poor predictive ability in comparison to those developed from the whole pool of descriptors.

A set of 10 compounds falling within the applicability domain of the models was taken from a different source [31] and reverse

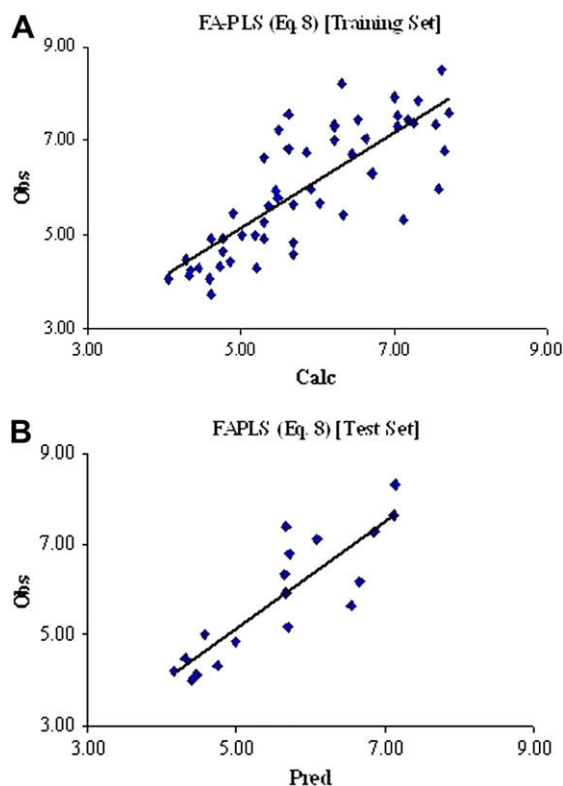


Fig. 1. Scatter plots of observed versus calculated/predicted values of (A) training set compounds and (B) test set compounds using Eq. (8).

transcriptase inhibitory activity of these compounds was predicted using selected models. The squared correlation coefficients between the observed and predicted values were found acceptable for the models. It is interesting to note that the best r^2 value (external set) was given by the model (ANN model 9) showing the best r_m^2 value (test set).

5. Conclusion

Among all the linear and non-linear models, the best model based on internal validation was obtained with non-linear modeling technique GFA (spline) using combination of topological, structural, physicochemical, electronic and spatial descriptors, while the best model based on external validation (R_{pred}^2 value from test set) was obtained from PLS method using structural, physicochemical, electronic and spatial descriptors. Predictive quality of the models was also further tested using modified r^2 (r_m^2) value. On the basis of this value (r_m^2), ANN provided the best model using combined set of topological parameters, structural, physicochemical, electronic and spatial descriptors. In general, higher r^2 (squared correlation coefficient between observed and predicted values of the test set compounds) values were obtained from the ANN models with comparatively higher number of hidden nodes, while higher r_m^2 values were obtained from those with comparatively lower number of hidden nodes. S_{aCH} , $SC-3_P$, S_{dO} , $Jurs_RPCS$, $Chiralcenters$ and $^3k_{am}$ were among the most frequently occurring descriptors in the equations. ANN models developed from reduced set of descriptors (those appearing in the equations)

showed poor predictive ability in comparison to those developed from the whole pool of descriptors. Selected models could also satisfactorily predict external set of TIBO derivatives falling within the applicability domain of the models. This indicates true predictive potential of the developed models.

Acknowledgement

Financial support under the DST Fast Track Scheme for Young Scientists (DST, Govt. of India, New Delhi) is thankfully acknowledged.

References

- [1] <www.unaids.org>.
- [2] <<http://en.wikipedia.org/wiki/HIV>>.
- [3] <www.rhodes.edu/biology/giindquester/viruses/pagespass/hiv/hiv.html>.
- [4] <www.retrovirology.com/content/4/1/50>.
- [5] <uhavax.hartford.edu/bugl/hiv.htm>.
- [6] M.L. Barreca, J. Balzarini, A. Chimirri, H.D. Hölting, M. Hölting, A.M. Monforte, P. Monforte, C. Pannecouque, A. Rao, M. Zappala, J. Med. Chem. 45 (2002) 5410–5413.
- [7] R.K. Rawal, Y.S. Prabhakar, S.B. Katti, E. De Clercq, Bioorg. Med. Chem. 13 (2005) 6771–6776.
- [8] C. Mao, E.A. Sudbeck, T.K. Venkatachalam, F.M. Uckun, Antivir. Chem. Chemother. 10 (1999) 233–240.
- [9] C. Tintori, F. Manetti, N. Veljkovic, V. Perovic, J. Vercammen, S. Hayes, S. Massa, M. Witvrouw, Z. Debyser, V. Veljkovic, M. Botta, J. Chem. Inf. Model. 47 (2007) 1536–1544.
- [10] B. Bhhataraj, R. Garg, Bioorg. Med. Chem. 13 (2005) 4078–4084.
- [11] E. Kellenberger, J.Y. Springael, M. Parmentier, M.H. Haas, J.L. Galzi, D. Rognan, J. Med. Chem. 50 (2007) 1294–1303.
- [12] J.K. Buolamwini, H. Assefa, J. Med. Chem. 45 (2002) 841–852.
- [13] K. Roy, J.T. Leonard, Bioorg. Med. Chem. 12 (2004) 745–754.
- [14] J.T. Leonard, K. Roy, QSAR Comb. Sci. 23 (2004) 23–35.
- [15] J.T. Leonard, K. Roy, Drug Des. Discov. 18 (2003) 165–180.
- [16] J.T. Leonard, K. Roy, QSAR Comb. Sci. 23 (2004) 387–398.
- [17] K. Roy, J.T. Leonard, QSAR Comb. Sci. 24 (2005) 579–592.
- [18] K. Roy, J.T. Leonard, Bioorg. Med. Chem. 13 (2005) 2967–2973.
- [19] K. Roy, J.T. Leonard, Indian J. Chem. 45A (2006) 126–137.
- [20] K. Roy, J.T. Leonard, J. Chem. Inf. Model. 45 (2005) 1352–1368.
- [21] J.T. Leonard, K. Roy, Bioorg. Med. Chem. 14 (2006) 1039–1046.
- [22] J.T. Leonard, K. Roy, Bioorg. Med. Chem. Lett. 16 (2006) 4467–4474.
- [23] J. Huuskonen, J. Chem. Inf. Comput. Sci. 41 (2001) 425–429.
- [24] Cerius2 Version 4.10 is a product of Accelrys Inc., San Diego, CA.
- [25] B. Everitt, S. Landau, M. Leese, Cluster Analysis, Arnold, London, 2001.
- [26] E.R. Dougherty, J. Barrera, M. Brun, S. Kim, R.M. Cesar, Y. Chen, M. Bittner, J.M. Trent, J. Comput. Biol. 9 (2002) 105–126.
- [27] K. Roy, A.S. Mandal, J. Enzyme Inhib. Med. Chem. (2008). <<http://dx.doi.org/10.1080/14756360701811379>>.
- [28] L. Eriksson, J. Jaworska, A.P. Worth, M.T.D. Cronin, R.M. McDowell, P. Gramatica, Environ. Health Perspect. 111 (2003) 1361–1375.
- [29] R. Guha, P.C. Jurs, J. Chem. Inf. Model. 45 (2005) 65–73.
- [30] J.T. Leonard, K. Roy, QSAR Comb. Sci. 25 (2006) 235–251.
- [31] W. Ho, M.J. Kukla, H.J. Breslin, D.W. Ludovici, P.P. Grous, C.J. Diamond, M. Miranda, J.D. Rodgers, C.Y. Ho, E.D. Clercq, R. Pauwels, K. Andries, M.A.C. Janssen, P.A.J. Janssen, J. Med. Chem. 38 (1995) 794–802.
- [32] K. Roy, Exp. Opin. Drug Discov. 2 (2007) 1567–1577.
- [33] R.B. Darlington, Regression and Linear Models, McGraw-Hill, New York, 1990.
- [34] S. Wold, in: H. van de Waterbeemd (Ed.), Chemometric Methods in Molecular Design, VCH, Weinheim, 1995, p. 195.
- [35] Y. Fan, L.M. Shi, K.W. Kohn, Y. Pommier, J.N. Weinstein, J. Med. Chem. 44 (2001) 3254.
- [36] R. Franke, Theoretical Drug Design Methods, Elsevier, Amsterdam, 1984, p. 184.
- [37] R. Franke, A. Gruska, in: H. van de Waterbeemd (Ed.), Chemometric Methods in Molecular Design, VCH, Weinheim, 1995, p. 113.
- [38] A. Fraser, Aust. J. Biol. Sci. 10 (1957) 484–491.
- [39] A. Fraser, D. Burnell, Computer Models in Genetics, McGraw-Hill, New York, 1970.
- [40] D. Rogers, A.J. Hopfinger, J. Chem. Inf. Comput. Sci. 34 (1994) 854–866.
- [41] J. Zupan, J. Gasteiger, Neural Networks in Chemistry and Drug Design, Wiley-VCH, Weinheim, 1999.
- [42] Y. Tang, H.L. Jiang, K.X. Chen, R.Y. Ji, Indian J. Chem. 35B (1996) 325–332.
- [43] G.W. Snedecor, W.G. Cochran, in: H. van de Waterbeemd (Ed.), Statistical Methods, Oxford and IBH, New Delhi, 1967, p. 381.
- [44] S. Wold, L. Eriksson, in: H. van de Waterbeemd (Ed.), Chemometric Methods in Molecular Design, VCH, Weinheim, 1995, p. 312.
- [45] A.K. Debnath, in: A.K. Ghose, V.N. Viswanadhan (Eds.), Combinatorial Library Design and Evaluation, Marcel Dekker, Inc., New York, 2001, p. 73.
- [46] MINITAB is a statistical software of Minitab Inc.; USA, <<http://www.minitab.com>>.
- [47] SPSS is a statistical software of SPSS Inc.; USA, <<http://www.spss.com>>.

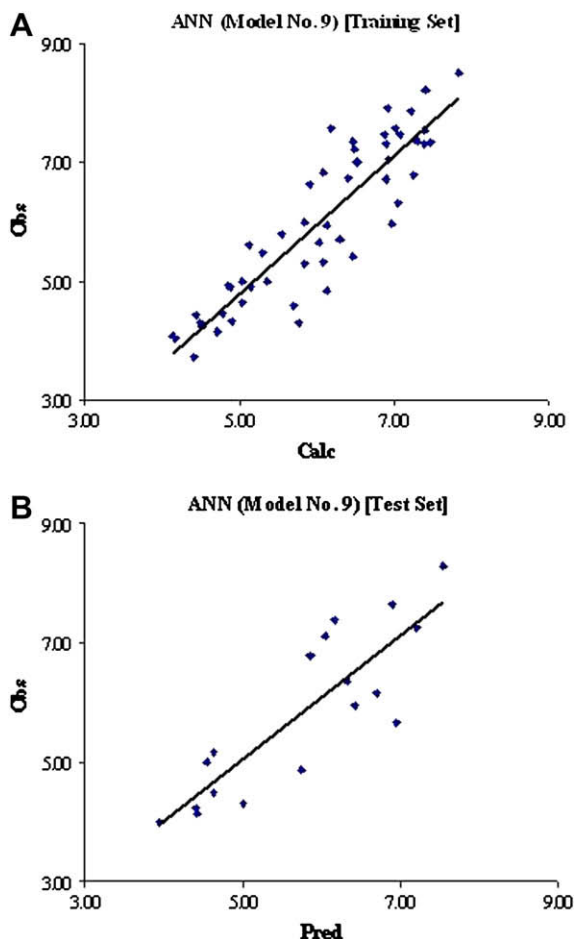


Fig. 2. Scatter plots of observed versus calculated/predicted values of (A) training set compounds and (B) test set compounds using ANN model 9.

- [48] STATISTICA is a statistical software of STATSOFT Inc.; USA, <<http://www.statsoft.com/>>.
- [49] A. Golbraikh, A. Tropsha, *J. Mol. Graph. Model.* 20 (2002) 269–276.
- [50] P. Roy, K. Roy, *QSAR Comb. Sci.* 27 (2008) 302–313.
- [51] Z. Zhou, J.D. Madura, *J. Chem. Inf. Comput. Sci.* 44 (2004) 2167–2178.
- [52] A.A. Toropov, A.P. Toropova, I.V. Nesterov, O.M. Nabiev, *J. Mol. Struct. (Theor. Chem)* 640 (2003) 175–181.
- [53] S. Hannongbua, P. Pungpo, J. Limtrakul, P. Wolschann, *J. Comput. Aided Mol. Des.* 13 (1999) 563–577.
- [54] V.P. Solov'ev, A. Varnek, *J. Chem. Inf. Comput.* 43 (2003) 1703–1719.