Journal of Enzyme Inhibition and Medicinal Chemistry, February 2009; 24(1): 205–223

Predictive QSAR modeling of CCR5 antagonist piperidine derivatives using chemometric tools

KUNAL ROY & ASIM SATTWA MANDAL

Division of Medicinal and Pharmaceutical Chemistry, Drug Theoretics and Cheminformatics Lab, Department of Pharmaceutical Technology, Jadavpur University, Kolkata 700 032, India

(*Received 19 December 2007; in final form 7 February 2008*)

Abstract

Quantitative structure-activity relationship (QSAR) studies have been performed on piperidine derivatives (n = 119) as CCR5 antagonists. The whole data set was divided into a training set (75% of the dataset) and a test set (remaining 25%) on the basis of *K*-means clustering technique. Models developed from the training set were used to assess the predictive potential of the models using test set compounds. Initially classical type QSAR models were developed using structural, spatial, electronic, physicochemical and/or topological parameters using statistical methods like stepwise regression, partial least squares (PLS) and factor analysis followed by multiple linear regression (FA-MLR). Using topological and structural parameters, FA-MLR provided the best equation based on internal validation ($Q^2 = 0.514$) but the best externally validated model was obtained with PLS ($R^2_{pred} = 0.565$). When structural, physicochemical, spatial and electronic descriptors were used, the best Q^2 value (0.562) was obtained from the stepwise regression derived model whereas the best R^2_{pred} value (0.571) came from the PLS model. When topological descriptors were used in combination with the structural, physicochemical, spatial and electronic descriptors, the best Q^2 and R^2_{pred} values obtained were 0.530 (stepwise regression) and 0.580 (PLS) respectively. Attempt was made to develop 3D-QSAR models using molecular shape analysis descriptors in combination with structural, physicochemical, spatial and electronic parameters. Linear models were developed using genetic function algorithm coupled with multiple linear regression. However, the results from the 3D-QSAR study were not superior to those of the classical QSAR models. Finally, artificial neural network was employed for development of nonlinear models. The ANN models showed acceptable values of squared correlation coefficient for the observed and predicted values of the test set compounds. From the view point of external predictability, selected ANN models were superior to the linear QSAR models. All reported models satisfy the criteria of external validation as recommended by Golbraikh and Tropsha (J Mol Graphics Mod 2002; 20: 269–276), whereas the majority of the models have modified r^2 (r^2_m) value of the test set for external validation more than 0.5 as suggested by Roy and Roy (QSAR Comb Sci 2008; 27: 302-313).

Keywords: *QSAR, CCR5 antagonist piperidines*

Introduction

Acquired immunodeficiency syndrome is one of the deadliest diseases in the world. This opportunistic infection (T4 cell falls below 200/μL) has no complete and successful treatment so far. Human immunodeficiency virus, a retrovirus of lentivirus family is the causative organism of this disease. About 2.9 million people including 3,80,000 children under 15 years died of AIDS in the year of 2006. In that year 4.3 million people have been newly infected with HIV virus and total numbers of HIV infected persons in the world are about 39.5 million till 2006 [1]. There are two serotypes of HIV virus that can be distinguished genetically and antigenetically. HIV-1 causes more serious and rapid infection than HIV-2. *Gag*, *pol* and *env* genes are the key elements of this viral structure. The *gag* gene is "group specific

Correspondence: K. Roy, Division of Medicinal and Pharmaceutical Chemistry, Drug Theoretics and Cheminformatics Lab, Department of Pharmaceutical Technology, Jadavpur University, Kolkata 700 032, India. E-mail: kunalroy_in@yahoo.com; http://www.geocities.com/kunalroy_in

ISSN 1475-6366 print/ISSN 1475-6374 online © 2009 Informa UK Ltd.
DOI: 10.1080/14756360802051297

antigen" composed of viral nucleocapsid. It is responsible for development of virus in the absence of *pol* and *env* genes. The *pol* gene codes for HIV enzymes like reverse transcriptase, protease and integrase. Finally the *env* gene codes for the two major glycoproteins (gp120 and gp4) of the viral envelope [2]. After entering into the blood stream this virus binds its glycoprotein (gp120) to a T4 cell's or macrophage's CD4 receptor and the coreceptor CCR5 and/or CXCR4. Binding to CD4 stimulates a conformational change to form and expose the binding site of coreceptor [3]. When virus binds to the coreceptor site, the rearrangement in that binding site occurs in such a way that fusion between the viral envelope and the cell membrane can take place. CCR or chemokine receptors are cell surface molecules. These bind peptide ligands called chemokine, thereby inducing migration of the receptor-bearing cells toward injured tissues. The injured tissues secrete chemokines into bloodstream [4]. Though *in vitro* this virus has been shown to use many coreceptors including CCR1, CCR2b, CCR3, CXCR6, CCR8, CX3CR1/V28, gpr1, gpr15, APJ, ChemR23 and RDC1 but *in vivo* the main coreceptors for infection are CCR5 and CXCR4 [3]. CCR5 permits entry of M-tropic HIV strains (R5) that predominate during early stages of the infection and are responsible for transmission of HIV-1. Individuals with a homozygous 32-bp deletion in the CCR5 gene are highly resistant to HIV-1 infection whereas heterozygous deletion may decelerate disease progression [5]. On the other hand CXCR4 is used as coreceptor by T-tropic HIV-1 strains appearing later in the disease course. This phenomenon accelerates disease progression. Genetic alterations of the CCR5 gene also control infection and disease progression. Depending on the properties of CCR5 and its interaction with HIV-1 gp120, there are two ways to inhibit the binding of HIV-1 to CCR5. Firstly sterical hindrance of gp120 binding to CCR5 can be achieved with modified or unmodified chemokines, mAbs or small molecular ligands. Secondly, internalization of CCR5 leads to the disappearance of CCR5 from the cell surface. Although CD4 is the primary receptor for HIV-1 virus, this is not efficient target for drug discovery. The fact is that binding of soluble CD4 to HIV-1 gp120 directly enables gp120 interaction with chemokine coreceptors independent of cellular CD4. CCR5 deficient individuals have no apparent immunologic defect. For these reason CCR5 constitutes very attractive target for drug development.

Predictive models have been developed using molecular modeling and multistep-docking procedure for HIV-1 entry inhibitor neomycin-arginine conjugates interaction with the CD4-gp120 binding site [6]. Liu et al used an approach combining protein structure modeling, docking and molecular dynamics simulation to build a series of structural models of the CCR5 in complexes with gp120 and CD4 [7]. Roy et al. have developed linear free energy related (LFER) model of Hansch and compared it with 3D-QSAR analyses (RSA, MSA and MFA) to find out the important molecular features of 3-(4-benzylpiperidin-1-yl)-*N*-phenylpropylamine derivatives for CCR5 binding affinity [8]. Comparative molecular field analysis and comparative molecular similarity indices studies of the derivatives of 1-(3,3-diphenylpropyl)-piperidinyl amide and urea as CCR5 receptor antagonists have been also reported [9]. Xu et al. have used an approach combining protein structure modeling, molecular dynamics simulation, automated docking and 3D QSAR analyses to investigate the detailed interactions of CCR5 with 1-amino-2-phenyl-4-(piperdin-1-yl)-butane derivatives [10]. Song et al. have compared the results obtained from CoMFA and CoMSIA on a series of piperidine-based CCR5 antagonists as an alternative approach to investigate the interaction between CCR5 antagonists and their receptor [11]. QSAR of CCR5 binding affinity of 1-(3,3-diphenylpropyl)-piperidinyl phenylacetamides using elimination selection-stepwise regression method has been reported by Afantitis et al. [12].

The present group of authors [8,13–21] has developed some anti-HIV QSAR models using compounds of different chemical classes and different types of descriptors. In continuation of such efforts, the present paper deals with predictive modeling of CCR5 binding affinity of piperidine derivatives reported by Finke et al. [22–25]. Some compounds were excluded from our study due to lack of quantitative activity data. Initially classical type QSAR models have been developed using multiple linear regression (with stepwise regression, factor analysis as variable selection technique) and partial least squares. This was followed by an attempt to develop 3D-QSAR models using molecular shape analysis descriptors along with structural, electronic, spatial and physicochemical descriptors with genetic function approximation as the statistical tool. Finally nonlinear models have also been developed using feed-forward backpropagation artificial neural network. The purpose of the present study is to develop predictive QSAR models with good validation characteristics for the CCR5 inhibitor piperidine derivatives and for this purpose different chemometric tools have been applied using different classes of descriptors for model development and comparison.

Methods and materials

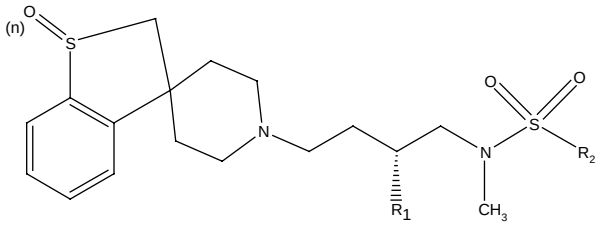
The CCR5 binding affinity data (IC_{50}) of 119 piperidine derivatives [22–25] were converted to logarithmic scale [$pIC_{50} = -\log IC_{50}$ (mM)] and then used for the QSAR study. There were total 154

piperidine derivatives in the source papers [22–25]. 35 compounds were excluded from our study due to lack of exact numerical activity values and infrequent occurrence of particular structural features. Thus, 119 compounds were selected in our study which are shown in Tables I to IV. In cases of racemic compounds (Tables I and II), only *S* configuration has been considered for modeling because the *R* isomers are less potent [22,23].

Descriptors

Three types of analyses were performed: Classical type QSAR modeling, 3D-QSAR modeling and nonlinear

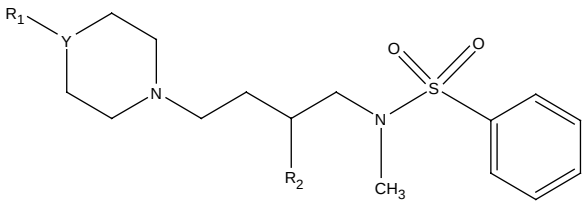
Table I. Structure and CCR5 binding affinities of sulfonyl derivatives of piperidine containing compounds.



Sl. No.	Number of oxygen atom (n)	Structural Features R1	Structural Features R2	Anti-HIV Activity ($-\log IC_{50}$(mM)) obs	cal[a]	cal[b]	cal[c]	cal[d]
1	0	(S)-3,4 diCl-phenyl	Phenyl	3.000	3.829	3.659	3.834	3.752
2	1	(S)-3,4 diCl-phenyl	Phenyl	4.456	3.984	3.919	3.890	3.928
3	2	(S)-3,4 diCl-phenyl	Phenyl	4.000	4.072	4.063	3.580	3.920
4	1	(S)-3,4 diCl-phenyl	2-Thienyl	4.222	3.886	3.762	3.652	3.697
5	2	(S)-3,4 diCl-phenyl	2-Thienyl	3.921	3.867	4.007	3.368	3.683
6	1	(S)-3,4 diCl-phenyl	N-dimthyl	3.469	3.750	3.898	3.325	3.590
7*	1	(S)-3,4 diCl-phenyl	Benzyl	3.229	4.319	4.263	4.021	4.119
8	1	(S)-3,4 diCl-phenyl	Methyl	3.071	3.375	3.581	2.948	3.314
9	1	(S)-3,4 diCl-phenyl	n-Octyl	2.854	4.462	4.548	4.240	4.795
10	1	(S)-3,4 diCl-phenyl	Cyclopentyl	4.000	3.876	3.857	3.482	3.815
11*	1	(S)-3,4 diCl-phenyl	Cyclohexyl	4.000	4.044	3.717	3.793	4.014
12	1	(S)-3,4 diCl-phenyl	2-Cl-phenyl	4.097	4.044	3.693	3.814	4.025
13	1	(S)-3,4 diCl-phenyl	3-Cl-phenyl	4.155	4.099	4.099	3.804	4.047
14	1	(S)-3,4 diCl-phenyl	4-Cl-phenyl	4.398	4.131	4.122	3.876	4.052
15	2	(S)-3,4 diCl-phenyl	3-NO_2-phenyl	3.824	4.554	4.809	3.989	4.169
16*	2	(S)-3,4 diCl-phenyl	4-NO_2-phenyl	4.222	4.595	4.969	3.968	4.254
17*	1	(S)-3,4 diCl-phenyl	4-MeO-phenyl	4.398	4.294	4.212	4.093	4.259
18	1	(S)-3,4 diCl-phenyl	4-Phenyl-phenyl	4.398	4.579	4.572	4.638	4.591
19	1	(S)-3,4 diCl-phenyl	Naphth-1-yl	3.444	3.855	3.386	4.342	3.901
20	1	(S)-3,4 diCl-phenyl	Naphth-2-yl	4.222	4.098	4.264	4.222	4.073
21*	1	(S)-3,4 diCl-phenyl	Indan-5-yl	4.155	4.302	3.998	4.015	4.225
22	1	(S)-3,4 diCl-phenyl	Pyridin-3-yl	4.000	3.879	3.805	3.882	3.826
23	1	(S)-3,4 diCl-phenyl	Quinolin-8-yl	4.046	3.965	3.585	4.473	4.096
24	1	(S)-3,4 diCl-phenyl	Quinolin-3-yl	3.921	4.152	4.140	4.336	4.035
25*	1	(S)-3,4 diCl-phenyl	1-Me-Imidazol-4-yl	3.469	4.060	3.813	3.760	3.866
26	0	(R/S)-phenyl	Phenyl	3.347	3.572	3.599	3.759	3.594
27	1	(R/S)-phenyl	Phenyl	4.456	3.872	3.875	3.889	3.606
28	2	(R/S)-phenyl	Phenyl	4.523	3.883	3.690	3.853	3.693
29	1	(R/S)-2-Cl-phenyl	Phenyl	2.699	3.857	3.583	3.861	3.692
30	2	(R/S)-2-Cl-phenyl	Phenyl	2.886	3.886	3.521	3.634	3.690
31	0	(S)-3-Cl-phenyl	Phenyl	3.569	3.674	3.497	3.910	3.638
32	1	(S)-3-Cl-phenyl	Phenyl	5.000	4.029	4.106	3.949	3.808
33	2	(S)-3-Cl-phenyl	Phenyl	4.824	4.209	3.984	3.845	3.798
34*	1	(S)-4-Cl-phenyl	Phenyl	3.569	4.043	3.972	3.928	3.714
35	1	(S)-4-F-phenyl	Phenyl	3.244	3.831	3.538	3.920	3.738
36*	1	(R/S)-3,5 diCl-phenyl	Phenyl	4.046	4.253	4.238	3.817	3.910
37	2	(R/S)-3,5 diCl-phenyl	Phenyl	3.959	4.251	4.243	3.817	3.915

* stands for a member of the test set; cal[a] obtained from the best r_m^2 result of ANN (Model 6); cal[b] obtained from the best r_m^2 result of GFA-MLR (Equation 11); cal[c] obtained from the best r_m^2 result of stepwise regression (Equation 7); cal[d] obtained from the best r_m^2 result of PLS (Equation 8)

Table II. Structure and CCR5 binding affinities of non-spiro piperidine derivatives.

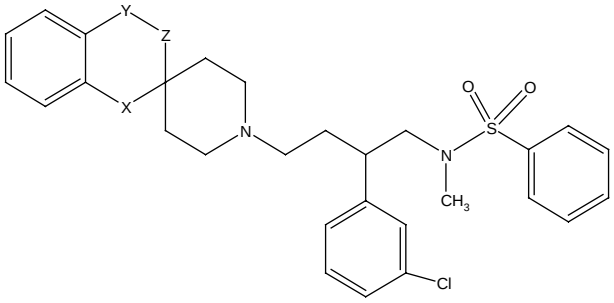


	Structural Features			Anti-HIV Activity (-logIC$_{50}$(mM))				
Sl. No.	R1	R2	Y	obs	cal[a]	cal[b]	cal[c]	cal[d]
38	Phenyl	(R/S)-Phenyl	—CH—	3.921	3.552	3.732	3.931	3.737
39	Phenyl	(R/S)-2-Cl-phenyl	—CH—	2.523	3.518	3.322	3.895	3.839
40*	Phenyl	(S)-3-Cl-phenyl	—CH—	4.523	3.659	3.709	3.887	3.864
41	Phenyl	(S)-4-F-phenyl	—CH—	3.000	3.522	3.505	3.828	3.767
42*	Phenyl	(R/S)-3,5-diCl-phenyl	—CH—	3.523	3.632	3.798	3.740	3.973
43	Phenyl	(R/S)-3-F-phenyl	—CH—	4.000	3.505	3.344	3.895	3.762
44*	Phenyl	(R/S)-3-Me-phenyl	—CH—	4.097	3.717	3.832	3.824	3.885
45*	Phenyl	(R/S)-3-Et-phenyl	—CH—	3.959	3.815	3.895	4.141	4.098
46	Phenyl	(R/S)-3-CF_3-phenyl	—CH—	3.301	3.828	3.736	3.949	3.935
47	Phenyl	(R/S)-4-Me-phenyl	—CH—	3.699	3.764	3.932	3.910	3.894
48*	Phenyl	(R/S)-3,5-di-Me-phenyl	—CH—	3.796	3.802	3.954	3.939	4.021
49*	Phenyl	(R/S)-3,4-di-F-phenyl	—CH—	3.244	3.161	3.207	4.026	3.724
50	Phenyl	(R/S)-3,4-di-Me-phenyl	—CH—	4.222	3.761	3.974	3.890	4.015
51	Phenyl	(R/S)-3-Me-4-F-phenyl	—CH—	3.745	3.806	3.571	4.010	3.902
52	Phenyl	(R/S)-3-F-4-Me-phenyl	—CH—	3.959	3.853	3.867	3.760	3.923
53	Phenyl	3-Cl-phenyl	—N—	3.155	3.720	3.724	4.054	3.836
54*	2-Methyl-phenyl	3-Cl-phenyl	—N—	2.620	3.704	3.840	3.896	3.946
55	2-Methyl-phenyl	3-Cl-phenyl	—CH—	3.398	3.868	3.924	3.883	3.984
56	2-MeO-phenyl	3-Cl-phenyl	—CH—	4.155	4.166	3.819	4.335	4.211
57	3-CF_3-phenyl	3-Cl-phenyl	—CH—	3.921	4.041	4.216	4.057	4.115
58*	4-Cl-phenyl	3-Cl-phenyl	—CH—	3.699	3.846	4.026	3.850	3.996
59	4-F-phenyl	3-Cl-phenyl	—CH—	4.602	3.774	3.877	3.917	3.924
60	Benzyl	3-Cl-phenyl	—CH—	3.602	3.839	3.822	3.825	4.080
61	$C_6H_5CH_2CH_2$	3-Cl-phenyl	—CH—	4.187	3.937	4.237	3.993	4.318
62	$C_6H_5CH_2CH_2CH_2$	3-Cl-phenyl	—CH—	5.301	4.209	4.553	4.338	4.573

* stands for a member of the test set; cal[a] obtained from the best r^2_m result of ANN (Model 6); cal[b] obtained from the best r^2_m result of GFA-MLR (Equation 11); cal[c] obtained from the best r^2_m result of stepwise regression (Equation 7); cal[d] obtained from the best r^2_m result of PLS (Equation 8)

modeling using artificial neural network. For the development of classical type QSAR models, topological, structural, physicochemical, spatial and electronic descriptors were used. At first topological and structural descriptors were utilized to develop 2D models using multiple linear regressions (with stepwise regression and factor analysis as variable selection techniques) and partial least squares. Then structural, physicochemical, spatial and electronic descriptors were combined to build models using the same techniques. Finally, in search of better predictive models, topological parameters were combined with structural, physicochemical, spatial and electronic descriptors and models were developed. For the development of 3D models molecular shape analysis descriptors were combined with structural, physicochemical, spatial and electronic descriptors. All descriptors were calculated using Cerius2 version 10 [26] running under IRIX 6.5 operating system on a Silicon Graphics computer and are shown categorically in Table V. In this study, topological descriptors considered were Balaban index (Jx), Kappa shape indices, Zagreb, Wiener, connectivity indices and E-state indices. Molecular weight (MW), numbers of rotatable bonds (Rotlbonds), number of hydrogen bond donors and acceptors and number of chiral centers were used as structural descriptors. Physicochemical descriptors used in the study include AlogP, AlogP98, LogP, MR and MolRef. Spatial descriptors like RadOfGyration, Jurs, Shadow, Area, Density, Vm and electronic parameters like charge, Fcharge, Apol, HOMO, LUMO and Sr were used in the study. DIFFV, COSV, Fo, NCOSV and ShapeRMS were employed as molecular shape analysis descriptors to develop 3D QSAR models. A full list of descriptors is given in Table V and their definitions can be found

Table III. Structure and CCR5 binding affinities of spiro piperidine derivatives.



	Structural Features		Anti-HIV Activity ($-\log IC_{50}$(mM))				
Sl. No.	X	Y-Z	obs	cal[a]	cal[b]	cal[c]	cal[d]
63	-[e]	—CH_2CH_2—	3.745	3.834	3.611	3.956	3.686
64⋆	-[e]	—$NHCH_2$—	4.301	3.852	3.813	3.877	3.776
65	-[e]	—$C(O)CH_2$—	5.301	3.809	3.745	4.085	3.671
66	-[e]	—C(O)NH—	4.347	4.027	4.073	4.009	3.726
67	-[e]	—C(O)N(Me)	4.000	4.016	3.943	3.854	3.796
68	-[e]	—$C(O)NHCH_2$—	4.456	4.060	4.204	4.113	3.862
69	-[e]	—$NHC(O)CH_2$—	4.456	4.145	4.401	3.890	3.886
70	-[e]	—$CH(OH)CH_2$—	4.000	4.029	4.280	3.952	3.899
71	—CH_2—	—O—	3.585	3.632	3.586	3.427	3.703

⋆ stands for a member of the test set; cal[a] obtained from the best r^2_m result of ANN (Model 6); cal[b] obtained from the best r^2_m result of GFA-MLR (Equation 11); cal[c] obtained from the best r^2_m result of stepwise regression (Equation 7); cal[d] obtained from the best r^2_m result of PLS (Equation 8); [e]The X in these structures is a single bond

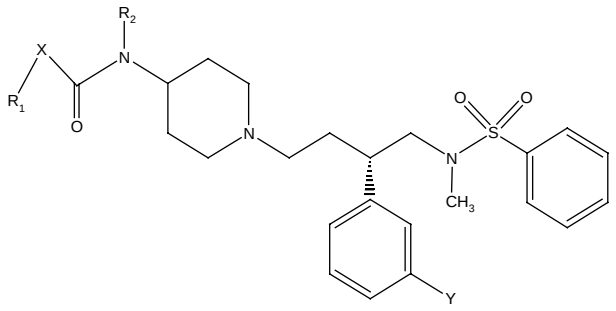
at the Cerius2 tutorial available at the website http://www.accelrys.com.

Cluster analysis and validation

The main target of any QSAR modeling is that the developed model should be robust enough to be capable of making accurate and reliable predictions of biological activities of new compounds. So, QSAR models which are developed from training set should be validated using new chemical entities for checking the predictive capacity 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 confidence of prediction can thus be judged [27]. For maximum cases, appropriate external data set is not available for prediction purpose. That is why the original data set is divided into training and test sets. A model's predictive accuracy and confidence for different unknown chemicals varies 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. So, the selection of the training set is significantly important in QSAR analysis. Predictive potential of a model on the new data set is influenced by the similarity of chemical nature between training set and test set [28–30]. The test set molecules will be predicted well when these molecules are very similar to the training set compounds. The reason is that the model has represented all features common to the training set molecules. There are different techniques available for division of the data set into training and test sets like statistical molecular design, self-organizing map, clustering, Kennard-Stone selection, sphere exclusion, etc. [31]. In the present case we have used clustering technique as the method for training set selection. Cluster analysis [32] is a technique to arrange the 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. There are two types of clustering: i) hierarchical clustering and ii) non-hierarchical clustering. One of the important non-hierarchical techniques is *K*-means clustering [33] which has been used in the present study. In this method clusters are started randomly and then cluster means are calculated in descriptor space. Molecules are reassigned to clusters whose means are closer to the position of molecules. After clustering, the test set compounds are selected from each cluster because both test set and training set can represent all clusters and characteristics of the whole dataset.

In our study the whole data set was divided into training and test sets based on *K*-means clustering and the models developed the training set were externally validated using test set. During internal validation the models were crossvalidated using leave-one-out

Table IV. Structure and CCR5 binding affinities of piperidine derivatives.



	Structural Features				Anti-HIV Activity ($-\log IC_{50}$(mM))				
Sl. No.	R1	R2	X	Y	obs	cal[a]	cal[b]	cal[c]	cal[d]
72	Me	H	O	H	3.000	3.612	3.720	4.017	3.823
73	t-Bu	H	O	H	3.000	3.902	4.063	3.332	4.225
74	t-Bu	Et	O	H	4.523	4.640	4.580	4.137	4.562
75⋆	Me	Me	O	H	3.824	3.957	3.586	4.071	3.944
76	Me	Et	O	H	4.398	4.108	4.121	4.216	4.143
77	Me	n-Pr	O	H	4.699	4.199	4.257	4.492	4.405
78	Me	n-Bu	O	H	4.824	4.634	4.449	4.598	4.628
79	Me	n-C_6H_{13}	O	H	5.000	4.792	4.776	5.170	5.123
80	Me	c-C_6H_{11}—CH_2	O	H	5.222	4.841	4.363	4.694	4.840
81⋆	Me	Bn	O	H	4.000	4.869	4.476	4.862	4.731
82	Et	c-C_6H_{11}—CH_2	O	H	4.456	4.705	4.634	4.728	5.109
83	Bn	c-C_6H_{11}—CH_2	O	H	3.097	5.026	5.010	5.258	5.446
84⋆	Et	Et	O	H	4.398	4.370	4.221	4.516	4.408
85	t-Bu	Et	O	H	4.602	4.609	4.760	3.877	4.564
86	c-C_6H_{11}—CH_2	Et	O	H	4.824	5.060	4.791	4.796	4.957
87⋆	Ph	Et	O	H	5.000	4.756	4.608	4.955	4.752
88	Bn	Et	O	H	5.699	5.017	4.879	5.070	4.834
89	Bn	Et	O	Cl	5.699	5.324	4.924	5.283	5.011
90	Bn	Me	O	H	5.301	4.917	4.834	4.856	4.653
91⋆	Bn	n-Pr	O	H	5.699	5.134	5.101	5.179	5.190
92	Bn	n-Pr	O	Cl	5.398	5.501	5.269	5.445	5.219
93	Bn	n-Bu	O	H	5.301	5.329	5.116	5.237	5.341
94	Bn	Allyl	O	H	5.824	5.264	5.410	5.220	5.181
95⋆	2-Me-C_6H_4—CH_2	n-Pr	O	H	5.398	5.191	4.952	5.141	5.156
96	3-Me-C_6H_4—CH_2	n-Pr	O	H	5.523	5.239	4.943	4.968	5.191
97	4-Me-C_6H_4—CH_2	n-Pr	O	H	5.523	5.503	5.518	5.402	5.358
98	4-CF_3—C_6H_4—CH_2	n-Pr	O	H	5.222	5.049	4.887	4.914	5.202
99	4-NO_2—C_6H_4—CH_2	n-Pr	O	H	5.824	5.330	5.185	5.288	5.262
100	4-NO_2—C_6H_4—CH_2	Allyl	O	H	5.699	5.740	5.684	5.605	5.486
101⋆	4-NO_2—C_6H_4-CH_2	Allyl	O	Cl	5.699	5.967	6.184	5.669	5.579
102	3-$NH_2COC_6H_4$—CH_2	n-Pr	O	H	6.097	5.951	6.041	5.667	5.653
103	4-$NH_2COC_6H_4$—CH_2	n-Pr	O	H	5.699	5.545	5.718	5.596	5.572
104	4-$NH_2COC_6H_4$—CH_2	n-Pr	O	Cl	5.523	5.670	5.835	5.627	5.680
105	Bn	n-Pr	O	H	5.699	5.165	5.080	5.300	5.142
106	Me	H	NH	H	3.000	3.768	3.894	3.938	3.856
107	Me	Et	NH	H	3.921	4.159	4.097	4.257	4.180
108⋆	Bn	H	NH	H	4.000	4.351	4.798	4.672	4.535
109	Bn	n-Pr	NH	H	5.602	5.344	5.649	5.695	5.193
110	Ph	n-Pr	NH	H	5.398	4.969	5.150	5.027	5.001
111	Bn	n-Pr	N-Me	H	4.699	5.166	5.627	5.222	5.267
112⋆	(S)-α-Me-Bn	n-Pr	NH	H	4.125	5.026	5.245	5.360	5.147
113⋆	4-NO_2-Bn	Allyl	NH	H	6.125	5.731	6.041	6.011	5.461
114⋆	Me	Et	–	H	3.921	4.188	4.255	4.276	4.167
115	Ph	n-Pr	–	H	4.000	4.852	5.214	4.958	4.975
116	Bn	n-Pr	–	H	5.523	5.180	5.201	5.258	5.186
117	$PhOCH_2$	n-Pr	–	H	5.398	5.397	5.349	5.410	5.486
118⋆	$PhCH_2CH_2$	n-Pr	–	H	4.699	5.104	5.419	5.489	5.468
119	4-NO_2-Bn	Allyl	–	H	5.699	5.471	5.672	5.803	5.240

⋆ stands for a member of the test set; cal[a] obtained from the best r^2_m result of ANN (Model 6); cal[b] obtained from the best r^2_m result of GFA-MLR (Equation 11); cal[c] obtained from the best r^2_m result of stepwise regression (Equation 7); cal[d] obtained from the best r^2_m result of PLS (Equation 8)

Table V. Categorical list of descriptors used in the development of models.

Category of Descriptors	Name of the Descriptors
Topological	Jx, $^{1}\kappa$, $^{2}\kappa$, $^{3}\kappa$, $^{1}\kappa_{am}$, $^{2}\kappa_{am}$, $^{3}\kappa_{am}$, Φ, SC-0, SC-1, SC-2, SC-3_P, SC-3_C, $^{0}\chi$, $^{1}\chi$, $^{2}\chi$, $^{3}\chi_{p}$, $^{3}\chi_{c}$, $^{0}\chi^{v}$, $^{1}\chi^{v}$, $^{2}\chi^{v}$, $^{3}\chi_{p}^{v}$, $^{3}\chi_{c}^{v}$, Wiener, Zagreb, S_sCH3, S_dCH2, S_ssCH2, S_dsCH, S_aaCH, S_sssCH, S_dssC, S_aasC, S_aaaC, S_ssssC, S_sNH2, S_ssNH, S_aaN, S_sssN, S_ddsN, S_aasN, S_sOH, S_dO, S_ssO, S_ssS, S_aaS, S_dssS, S_ddssS, S_sF, S_sCl.
Structural	MW, Rotlbonds, Hbond acceptor, Hbond donor, Chiral centers.
Physicochemical	AlogP, AlogP98, LogP, 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_FPSA_1, Jurs_FNSA_1, Jurs_FPSA_2, Jurs_FNSA_2, Jurs_FPSA_3, Jurs_FNSA_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_TPSA, 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	Charge, Fcharge, Apol, Dipole-mag, HOMO, LUMO, Sr.
Molecular Shape	DIFFV, COSV, Fo, NCOSV, ShapeRMS.

method. At first all independent variables were standardized between 0 and 1. All molecules with standardized descriptors were classified into six clusters based on *K*-means clustering. Serial numbers of compounds under different clusters were shown in Table VI. From these six clusters 75% of the total compounds were selected as training set and remaining 25% were selected as test set.

For the development of equations different chemometric tools were utilized.

Stepwise regression

In this method an initial model is identified and then it is repeatedly altered by adding or removing a predictor variable according to the "stepping criteria" (in this study F = 4 for inclusion and F = 3.9 for exclusion for the forward selection method) [34]. The search is terminated when stepping is no longer possible or when a specified maximum number of steps has been reached. Specifically, at each step all variables are evaluated to determine the most contributing predictor 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 respectively the contribution of a variable in the regression equation for adding the term to the equation and removing from the equation.

PLS

For PLS, "leave-one-out" method was used for crossvalidation to obtain the optimum number of components. PLS is a useful technique for constructing predictive models when the factors are many (e.g., greater than the number of observations) and they are highly collinear. This technique [35] generalizes and combines features from principal component and multiple regression. In the development of models there are many factors which contribute to the model. But some of them have capability to change the response largely and others have very low contribution to the response. So, the primary target of PLS regression is to find out those latent factors which are responsible for large variation in the response. In this present data set, the variables with smaller coefficients based on standardized regression coefficients were removed from the PLS regression, until there was no further improvement in Q^2 value, irrespective of the components. To avoid overfitting, the significance of each consecutive PLS component is examined and it is stopped when the components are non-significant.

FA-MLR

Factor analysis [36,37] is a statistical procedure used to disclose relationships among many variables. It allows large numbers of intercorrelated variables to be condensed into fewer dimensions, called factors. It is a data processing step to identify the variables contributing to the response variable. In our study biological activity data of the training set and all descriptors were extracted by principle component method and rotated by VARIMAX rotation to obtain Thurston's simple structure. The effective variables were selected from rotated component matrix obtained from the previous operation. Linear regression was performed using these variables.

Table VI. Serial numbers of compounds under different clusters.

Cluster No.	Serial Numbers of Compounds
1	1; 26; 31; 38; 39; 40; 41; 42; 43; 44; 45; 46; 47; 48; 49; 50; 51; 52; 53; 54; 55; 56; 57; 58; 59; 60; 61; 62; 63; 64; 65; 66; 67; 68; 69; 70; 71.
2	79; 80; 81; 82; 83; 86; 87; 88; 89; 90; 91; 92; 93; 94; 95; 96; 97; 105; 108; 109; 110; 111; 112; 115; 116; 117; 118.
3	98; 99; 100; 101; 102; 103; 104; 113; 119.
4	2; 4; 6; 7; 8; 9; 10; 11; 12; 13; 14; 17; 18; 19; 20; 21; 22; 23; 24; 25; 27; 29; 32; 34; 35.
5	3; 5; 15; 16; 28; 30; 33; 36; 37.
6	72; 73; 74; 75; 76; 77; 78; 84; 85; 106; 107; 114.

Molecular shape analysis [38]

Molecular shape analysis (MSA) was used as a 3D-QSAR technique. In our study the steps to perform MSA were i) generation of conformers and energy minimization, ii) hypothesizing an active conformer (global minimum of the most active compound), iii) selecting a shape reference compound based on active conformation, iv) performing pair-wise molecular shape superimposition using maximum common subgroup (MCSG) method, v) measurement of molecular shape commonality using MSA descriptors, vi) determination of other molecular features by calculating structural, spatial, physicochemical and electronic parameters, vii) selection of conformers and viii) generation of QSAR equations by linear genetic function approximation (GFA) followed by multiple linear regression. Attempt was also made to develop nonlinear models using Artificial Neural Network (ANN). Multiple conformations of every molecule were generated using optimal search as a conformational search method. Conformers of each molecule were subjected to energy minimization procedure to generate a low energy conformation for each structure. Energy minimization had been performed using a smart minimizer under open force field (OFF). Maximum common subgroup (MCSG) method was used for alignment of molecules. This method searches the largest subset of atoms in the atoms in the shape reference compound that is shared by all the structures in the study table and uses this subset for alignment. A rigid fit of atom pairings was performed to superimpose each structure so that it overlays the shape reference compound.

Genetic function approximation-multiple linear regression

Genetic algorithms [39] are derived from an analogy with the mutation of DNA. This algorithm was initially imagined from i) Holland's genetic algorithm and ii) Friedman's multivariate adaptive regression splines (MARS) algorithm. In this algorithm an individual or model is represented as a linear string in which information about DNA (the series of basis functions) of that individual or model is stored. Based on this information the activity model is reconstructed using least-squares regression to regenerate the coefficients. Genetic algorithm makes superior models to those developed using stepwise regression techniques because genetic algorithm contains additional information about the models. A "fitness function or lack of fit (LOF)" is used to estimate the quality of an individual, so that best individual receives the best fitness score. The error measurement term LOF is determined by the following equation:

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

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 number of samples in the training set; LSE is least squares error and p is total numbers of features contained in all basis functions.

Once models in the population have been rated using the LOF score, the genetic cross over operation is repeatedly performed. Individual (or model) with best fitness score is considered as potential member to transmit its genetic material for mutation, in which some parts of genetic material are taken from each parent and recombined to create the child. After many mating steps average fitness of individuals (models) in the population increases as good combinations of genes are discovered and spread through the population. It can build not only linear models but also higher-order polynomials, splines and Gaussians. But in our present work, splines were not used. Descriptors, which were selected by this algorithm, were subjected to multiple linear regression for generation of models.

Artificial neural network [40]

Artificial Neural Network (ANN) is an information-processing pattern that is inspired by the way biological nervous systems, such as the brain, process information. Maximum networks contain at least three layers - input, hidden and output. The layers of input neurons receive the data either from input files or directly from electronic sensors in real-time applications. The output layer sends information directly to the outside world, to a secondary computer process or to other devices such as a mechanical control system. Between input and output layers there may be many hidden layers. These internal layers contain many of the neurons in various interconnected structures. Based on the function there are different types of neural networks like feed-forward back-propagation, counter propagation, probabilistic neural network, self-organizing map etc. But here in the present study for the development of our nonlinear models, feed-forward backpropagation method was used. Multilayer perceptron (MLP) method under "Custom Network Designer" had 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 back propagated to the transfer function (sigmoid function) for adjustment of weight. The output [41] can be represented as by the following equation.

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

where 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.

$$W_{i_j}(new) = W_{i_j}(old) + \eta\delta_i O_j + \alpha[\Delta W_{i_j}(old)]$$

In backpropagation method the learning of the network has followed the Delta Rule, which starts with the calculated difference between the actual outputs and the desired outputs. Using this error, connection weights are increased in proportion to the error times a scaling factor for global accuracy. The complex part of this learning mechanism is for the system 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 2nd 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 numbers of weights and/or multiple output units. It is a batch update algorithm whereas back propagation adjusts the weights of the network. Learning rate and momentum of each epoch are adjusted and weight decay is regularized. Crossvalidated resampling of advanced technique was used as sampling procedure during formation of network. When a particular number of resampling is selected, the numbers of available cases are divided into 3 subsets (training, selection and test sets). Training subset is used to optimize the network. The second subset, i.e., selection set is used to prevent the training from becoming over learned. Finally, a test subset is used to estimate the performance of that network.

Although the use of a test subset set 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.

Model quality

The statistical performances of the multiple regression equations [42] were evaluated 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* [43,44] and *cross-validation* 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^2_{pred}) value.

Softwares

MINTAB [45] was used for cluster analysis, stepwise regression and PLS. SPSS [46] was utilized in the operation of FA-MLR and STATISTICA [47] was used for ANN. Cerius2 version 4.10 [26] was used for MSA and GFA analyses.

Results and discussion

Classical type QSAR

QSAR using topological and structural descriptors

Stepwise regression. The following equation was obtained using F criterion (F = 4 for inclusion; F = 3.9 for exclusion).

$$\begin{aligned} pIC_{50} = {} & 3.278 + 4.090(\pm 3.072)^2\kappa \\ & - 1.570(\pm 2.777)^3\kappa_{am} \\ & - 0.700(\pm 1.337)S_{\_ssCH_2} \end{aligned} \quad (1)$$

$$n_{Training} = 90, R^2 = 0.566, R^2_{adj} = 0.551,$$

$$F = 37.370, F_{\max} = 15.990, s = 0.600,$$

$$Q^2 = 0.510, PRESS = 34.981, n_{Test} = 29,$$

$$R^2_{\text{pred}} = 0.504.$$

In the above equation, three variables were selected for development of the model. All regression coefficients are significant at 95% confidence level and the corresponding confidence intervals are mentioned within parentheses. The above equation could explain 55.1% of the variance of the CCR5 binding affinity while the leave-one-out predicted variance was 51.0%. The positive coefficient of the kappa shape index of 2nd order indicates that the CCR5 binding affinity increases with increment of branching whereas kappa alpha-modified shape index of 3rd has negative impact on the affinity. Contribution of the covalent radii and hybridization states are considered in the kappa alpha-modified shape index. The negative coefficient of the E-state index ($S_{\_ssCH_2}$) shows that both the electronic

character and topological environment of carbon atom in the fragment $-CH_2$ are responsible for lowering the CCR5 binding affinity. Equation (1) contains 3 independent variables whereas total numbers of observations are 90. According to Eriksson et al. [28] number of compounds should be at least 5 times higher than the number of selected independent variables. So, this model maintains the recommended ratio. When a multiple linear regression model has been developed from a large pool of variables then critical F test can be used to judge its significance [48,49]. The reason is that an effect known as "selection bias" makes the resulting model more significant than they really are. According to Livingstone and Salt, a critical F 5% value should be used to judge the significance of MLR models constructed by best subset selection and the critical value ($F_{\max}$) is calculated as follows [49]:

$$F_{\max} = \frac{29.96n^{3.18}N^{0.21}}{p^{0.82}} e^{\ln(v2)[1.06\ln(v2)-0.97\ln(n)-3.97]}$$

In the above equation, p is the number of predictor variables used in a MLR equation, k is the total number of variables from which the p variables have been chosen and n is the number of compounds. For Equation (1), the values of p, k and n are 3, 56 and 90 respectively. N is defined as k!/(p!(k − p)!) and v2 is the second degree of freedom of the F-statistics, i.e., n-p-1. For Equation (1), $F_{\max}$ is calculated to be 15.990 whereas the F value of the equation is 37.370. Thus, Equation (1) passes the critical F test. When Equation (1) was used for prediction of the CCR5 binding affinity of the compounds that were not used for model development, the predictive R^2 (R^2_{pred}) value was found to be 0.504.

PLS. In case of PLS, the following equation was developed from seven independent variables with one component selected by crossvalidation.

$$\begin{aligned} pIC_{50} =& 3.366 + 0.371\,{}^1\kappa + 0.356\,{}^2\kappa + 0.286\,{}^3\kappa \\ &+ 0.329\,{}^2\kappa_{am} + 0.306\phi + 0.367\,Wiener \\ &+ 0.280\,Rotlbonds \end{aligned} \quad (2)$$

$$n_{Training} = 90, R^2 = 0.511, R^2_{adj} = 0.506,$$

$$F = 92.060, s = 0.349,$$

$$Q^2 = 0.488, PRESS = 36.552,$$

$$n_{Test} = 29, R^2_{pred} = 0.565.$$

Equation (2) could explain and predict 50.6% and 48.8% respectively of the variance of the CCR5 binding affinity. Here, the results of crossvalidation (internal validation) are not encouraging (Q^2 less than 0.5), but external predictive capability of the model on the test data set is good (R^2_{pred} being 0.565). In this model, kappa shape indices of 1st, 2nd, 3rd order and kappa alpha-modified shape index of 2nd order have positive impact on the CCR5 binding affinity. Besides these, flexibility index, Wiener index and number of rotatable bonds have positively influenced the biological activity.

FA-MLR. From the factor analysis on the data matrix consisting of the CCR5 binding affinity with topological and structural descriptors, it was observed that 10 factors could explain the data matrix to the extent of 95.135%. The anti-HIV activity was moderately loaded with factor 2 (loaded in ${}^2\kappa$, ${}^3\kappa$, ${}^2\kappa_{am}$, ${}^3\kappa_{am}$, Φ, S_{sCH3}, S_{aasC}, Rotlbond) and weakly loaded with factor 1 (loaded in Jx, SC_1, SC_3_P, SC_3_C, ${}^2\chi$, ${}^3\chi_c$, ${}^0\chi^v$, ${}^1\chi^v$, ${}^2\chi^v$, ${}^3\chi^v_p$, Zagreb, S_{sssCH}, S_{do}, MW), factor 3 (loaded in S_{aaCH}), factor 4 (loaded in S_{ssssC}), factor 5 (loaded in S_{dCH2}), factor 9 (loaded in S_{ssCH2}), factor 11 (loaded in S_{aaaC}, S_{aaN}), factor 13 (loaded in S_{ssS}), factor 15 (loaded in Sr) and factor 16 (loaded in S_{sCl}). Based on the factor analysis, the following variables were selected for multiple linear regression. The best equation evolved was as follows:

$$\begin{aligned} pIC_{50} =& 3.489 + 2.473(\pm 0.952)^2\kappa \\ &- 0.682(\pm 1.354)S\_{sF} \\ &- 0.942(\pm 1.367)S\_{ssCH_2} \end{aligned} \quad (3)$$

$$n_{Training} = 90, R^2 = 0.561, R^2_{adj} = 0.546,$$

$$F = 36.630, F_{\max} = 15.990, s = 0.604,$$

$$Q^2 = 0.514, PRESS = 34.708, n_{Test} = 29,$$

$$R^2_{\text{pred}} = 0.470.$$

Equation (3) involved three descriptors explaining and predicting 54.6% and 51.4% respectively of the variance of the CCR5 binding affinity. But the predictive capacity of the model on the test data set was not satisfactory (R^2_{pred} being less than 0.5). According to the Pearson Correlation method there was no significant intercorrelation within these variables [Intercorrelation table not shown]. The critical $F_{\max}$ value for Equation (3) calculated according to Livingstone and Salt [49] is 15.990. The F value of Equation (3) being 36.630, this equation passes the critical F value test.

QSAR using structural, physicochemical, spatial and electronic descriptors

Stepwise regression. Using structural, physicochemical, spatial and electronic descriptors, the following

equation was obtained with six independent variables (F = 4 for inclusion; F = 3.9 for exclusion).

$$pIC_{50} = 4.455 + 1.700(\pm 1.916)MR - 1.780(\pm 1.623)A\log P98 + 1.680(\pm 2.001)Jurs\_SASA - 0.950(\pm 1.212)Hbonddonor - 2.280(\pm 2.554)Shadow\_YZfrac + 1.730(\pm 2.474)Shadow\_XZfrac \quad (4)$$

$$n_{Training} = 90, R^2 = 0.628, R^2_{adj} = 0.602,$$

$$F = 23.390, F_{\max} = 42.108, s = 0.565,$$

$$Q^2 = 0.562, PRESS = 31.287, n_{Test} = 29,$$

$$R^2_{pred} = 0.438.$$

Like Equation (1), all regression coefficients were significant at 95% confidence level and the corresponding confidence intervals were mentioned within parentheses. This model could explain 60.2% and predict 56.2% of the variance of the CCR5 binding affinity. The external prediction ability of Equation (4) is not encouraging (R^2_{pred} being 43.8%). According to this model, the CCR5 binding affinity increases with increase in molar refractivity and decrease in partition coefficient values. The value of Jurs_SASA is calculated by mapping atomic partial charges on total solvent accessible surface areas of individual atoms. This descriptor has positively influenced the CCR5 binding affinity of the piperidine derived compounds. Again, with increase in the number of the hydrogen-bond donors, the binding affinity decreases as evidenced from the negative coefficient of the parameter Hbonddonor. Fraction of the area of molecular shadow in the XZ plane has positive impact on the activity whereas the effect of fraction of area of molecular shadow in the YZ plane is detrimental. Though the model has maintained the ratio of 1:5 [28] between the numbers of descriptors and the numbers of observations but it is unable to fulfill the criterion of the critical *F* test [49].

PLS. In case of PLS regression, Equation (5) with seven independent variables and one component (optimized with crossvalidation) was obtained.

$$pIC_{50} = 3.197 + 0.410MR + 0.419Jurs\_SASA + 0.374Jurs\_PPSA\_2 + 0.372Jurs\_WPSA\_2 + 0.364Area + 0.368V_m + 0.292Rotlbonds \quad (5)$$

$$n_{Training} = 90, R^2 = 0.508, R^2_{adj} = 0.503,$$

$$F = 91.010, s = 0.351,$$

$$Q^2 = 0.480, PRESS = 37.138, n_{Test} = 29,$$

$$R^2_{pred} = 0.571.$$

Like Equation (4), molar refractivity (MR) and Jurs_SASA show positive coefficients in this model. Besides this, Jurs_PPSA_2 (total charge weighted positive surface area: partial positive solvent accessible surface area multiplied by the total positive charge) and Jurs_WPSA_2 (surface-weighted charged partial surface area) show positive coefficients in the model. Increase in the number of rotatable bonds also improves the binding affinity. The positive coefficient of Area (van der Waals area of a molecule) indicates that exposing capacity of molecules to external environment is conducive for the CCR5 binding affinity. This descriptor is related to binding, transport and solubility. Similarly, molecular volume (V_m) has positive effect to the response variable. This model could explain 50.3% of variance of the affinity. It also could predict 48% of variance of the affinity (internal validation). But predictive potential on the test set is significant (R^2_{pred} being 0.571). The quality of Equation (5) is also better than that of Equation (2).

FA-MLR. In this case, three factors could explain the data matrix to the extent of 95.590%. The CCR5 binding affinity was highly loaded with factor 2 (loaded in MR, MolRef, Jurs_SASA, Jurs_PPSA_2, Jurs_DPSA_2, Jurs_FPSA_2, Jurs_WPSA_1, Jurs_WPSA_2, Jurs_RPCG, Jurs_RNCG, Jurs_TASA, Shadow_XY, Shadow_YZ, Area, V_m, Rotlbond), moderately loaded with factor 12 (loaded in HOMO) and poorly loaded with factor 5 (loaded in Shadow_XZfrac, Shadow_YZfrac, Shadow_nu, Shadow_Zlength). Using structural, physicochemical, spatial and electronic descriptors, FA-MLR led to an equation which was inferior in statistical quality to that of stepwise regression and PLS derived equations. All regression coefficients were significant at 95% confidence level and the corresponding confidence intervals were mentioned within parentheses.

$$pIC_{50} = 3.796 + 2.293(\pm 1.147)MR - 0.924(\pm 1.542)Shadow\_YZfrac \quad (6)$$

$$n_{Training} = 90, R^2 = 0.498, R^2 = 0.486, F = 43.080,$$

$$F_{\max} = 12.865, s = 0.642,$$

$$Q^2 = 0.466, PRESS = 38.124, n_{Test} = 29, R^2_{pred} = 0.434.$$

This equation could explain and predict 48.6% and 46.6% respectively of variance of the CCR5 binding affinity. Like Equations (4) and (5), here molar refractivity (MR) has showed positive influence to the CCR5 binding affinity. The fraction of area of molecular shadow in the YZ plane is unfavorable for the CCR5 binding affinity. According to Pearson Correlation method molar refractivity is weakly correlated with Shadow_YZfrac [r = −0.280]. This model passes the critical F test recommended by Livingstone and Salt [49] as the value of variance ratio crosses the critical $F_{\max}$ value.

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

Stepwise regression. Equation (7) consisting of 3 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.

$$pIC_{50} = 3.674 + 4.130(\pm 3.066)^2\kappa - 1.920(\pm 2.767)^3\kappa_{am} - 0.800(\pm 1.477)Shadow\_YZfrac \quad (7)$$

$$n_{Training} = 90, R^2 = 0.567, R^2_{adj} = 0.552, F = 37.600, F_{\max} = 20.801, s = 0.599, Q^2 = 0.530, PRESS = 33.591, n_{Test} = 29, R^2_{pred} = 0.549.$$

The 95% confidence intervals of independent variables were mentioned within parentheses. The positive coefficient of kappa shape index of 2nd order and negative coefficient of alpha-modified kappa shape index of 3rd order are obtained similar to Equation (1). This model showed 55.2% explained and 53% predicted variances which were inferior to corresponding results of Equation (4) obtained from stepwise regression excluding topological descriptors. However, predictive potential of Equation (7) on the test set was superior to that of Equation (4) (R^2_{pred} value of Equation (7) being 0.549 compared to corresponding value of 0.438 for Equation (4)]. This model also passes the critical *F* test recommended by Livingstone and Salt [49].

PLS. The following PLS equation consisting of seven independent variables was obtained with one component.

$$pIC_{50} = 3.283 + 0.355^2\kappa + 0.328^2\kappa_{am} + 0.305\phi + 0.391MR + 0.279Rotlbonds + 0.400Jurs\_SASA + 0.355Jurs\_WPSA\_2 \quad (8)$$

$$n_{Training} = 90, R^2 = 0.515, R^2_{adj} = 0.509, F = 93.520, s = 0.346, Q^2 = 0.490, PRESS = 36.443, n_{Test} = 29, R^2_{pred} = 0.580.$$

Like previous Equations (Equations (2) and (5)), this equation contains positive coefficients of kappa shape index of 2nd order, kappa alpha-modified shape index of 2nd order, flexibility index ( ), molar refractivity (MR), number of rotatable bonds (Rotlbonds), Jurs_SASA and Jurs_WPSA_2. This model could explain and predict 50.9% and 49% of the variance of the CCR5 binding affinity. Both of these statistics are better than those of two previous PLS results (Equations (2) and (5)). In fact, predictive potential of this Equation (R^2_{pred}) on the test chemical entities is also superior to those of Equations (2) and (5).

FA-MLR. The following FA-MLR equation was obtained with only one variable. In this model only kappa shape index of 2nd order has been selected based on factor loading pattern.

$$pIC_{50} = 3.272 + 2.256(\pm 0.933)^2\kappa \quad (9)$$

$$n_{Training} = 90, R^2 = 0.512, R^2_{adj} = 0.506, F = 92.340, F_{\max} = 12.186, s = 0.629, Q^2 = 0.490, PRESS = 36.408, n_{Test} = 29, R^2_{pred} = 0.526.$$

Though the results of explained variance and predicted variance of Equation (9) were inferior to those of Equation (3), predictive capacity (R^2_{pred}) on the test data set was better than those of both Equations (3) and (6). It passes the F test as the F value is greater than the critical $F_{\max}$ value [49].

A Comparative study of statistical parameters of classical QSAR models using different descriptors is given in Table VII.

Table VII. Comparative study of statistical parameters of classical QSAR models using different descriptors.

Type of Descriptors	Statistical Method	R^2 (Training Set)	Ra^2 (Training Set)	Q^2 (Training Set)	F	S	R^2_{pred} (Test Set)
Topological + structural	Stepwise	**0.566**	0.551	0.510	37.370	0.600	0.504
	PLS	0.511	0.506	0.488	92.060	0.349	**0.565**
	FA-MLR	0.561	0.546	**0.514**	36.630	0.604	0.470
Structural + physicochemical + spatial + electronic	Stepwise	**0.628**	0.602	**0.562**	23.390	0.565	0.438
	PLS	0.508	0.503	0.480	91.010	0.351	**0.571**
	FA-MLR	0.498	0.486	0.466	43.080	0.642	0.434
Topological + structural + physicochemical + spatial + electronic	Stepwise	**0.567**	0.552	**0.530**	37.600	0.599	0.549
	PLS	0.515	0.509	0.490	93.520	0.346	**0.580**
	FA-MLR	0.512	0.506	0.490	92.340	0.629	0.526

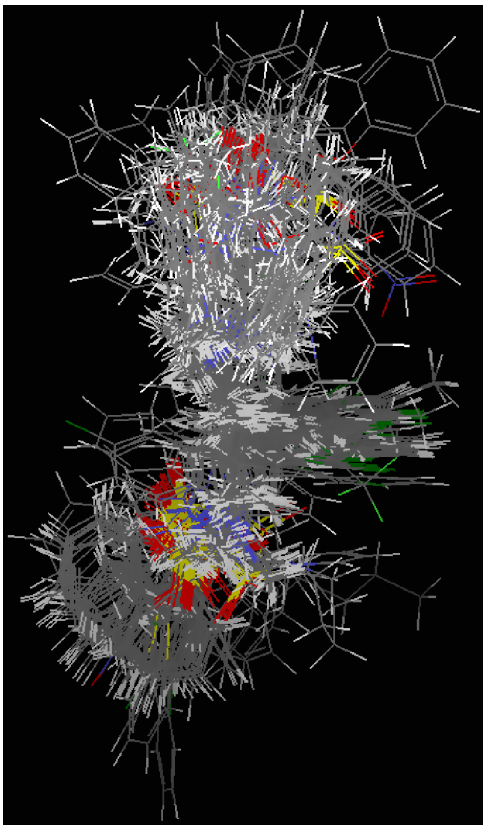


Figure 1. Aligned view of the training set molecules.

3D QSAR

In order to gain further insight into the structure-activity relationships, additional study was made using molecular shape analysis. This study was conducted using MSA descriptors along with additional descriptors like structural, physicochemical, spatial and electronic parameters. We have developed two types of models: i) linear (using genetic unction approximation combined with multiple linear regression) and ii) nonlinear (artificial neural network). To develop 3D QSAR models, the training data set compounds were aligned (shown in Figure 1) to the shape reference compound (compound 102) as detailed in the Materials and Methods section.

The following two Equations (Equations 10-11) were among the best ones based on LOF score obtained from genetic function approximation (50000 iterations) combined with multiple linear regression; however, none of these equations contain any MSA descriptor.

$$pIC_{50} = 3.764 - 4.184MW + 3.538Jurs\_PNSA\_1 + 5.809Area - 1.815A\log P98 - 0.641Hbonddonor \quad (10)$$

$$n_{Training} = 90, R^2 = 0.614, R^2_{adj} = 0.591,$$

$$F = 26.760, F_{\max} = 34.463,$$

$$s = 0.573, LOF = 0.387,$$

$$Q^2 = 0.550, PRESS = 32.152, n_{Test} = 29,$$

$$R^2_{pred} = 0.505.$$

Equation (10) suggests that the CCR5 binding affinity decreases with increase in molecular weight, partition coefficient and number of hydrogen bond donor groups. Again positive impacts of Area and Jurs descriptor (Jurs_PNSA_1) are observed. Jurs_PNSA_1 is partial negative surface area which is sum of the solvent-accessible surface areas of all negatively charged atoms. The explained variance and predicted variance are 59.1% and 55%. The predictive R^2 value is found to be 0.505.

$$pIC_{50} = 3.554 - 3.598MW + 2.959Jurs\_PNSA\_1 + 5.159Area - 1.260A\log P98 \quad (11)$$

$$n_{Training} = 90, R^2 = 0.593, R^2_{adj} = 0.574,$$

$$F = 30.990, F_{\max} = 24.798,$$

$$s = 0.585, LOF = 0.389,$$

$$Q^2 = 0.537, PRESS = 33.075, n_{Test} = 29,$$

$$R^2_{pred} = 0.520.$$

Equation (11) is similar to Equation (10), only lacking the term *Hbonddonor*. The removal of this term decreases R^2 and Q^2 values, but the R^2_{pred} value increases. Although both of the models (Equations 10 and 11) has maintained the ratio of 1:5 [28] between the number of descriptors and the number of observations, Equation 10 does not satisfy the criteria of the critical *F* value due to large pool of selected independent variables [49]. The absence of MSA descriptors in the models indicate that 3D QSAR could not provide better models over Classical type QSAR for this data set.

A Comparative study of statistical parameters of the GFA-MLR models is given in Table VIII.

Nonlinear modeling

For the development of better predictive models, nonlinear modeling with artificial neural network was also tried. We have formed the network 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 iterations of backpropagation and conjugate gradient descent, varying numbers of hidden layers and units per layer, a number of models were developed. In this study certain numbers of iterations, hidden layers, elements per layer etc. were selected. Then the number of a particular parameter was changed by fixing the other parameters. Here we have presented 6 best networks using different iterations and different hidden layers in Table IX. In the best network (bold faced model based on the squared correlation coefficient between the observed and predicted values of the test set compounds), 3 hidden layers of 43, 39, 36 elements respectively were used. Numbers of iterations selected for backpropagation and conjugate gradient descent were 700 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. During 20 resampling, numbers of cases selected for training, selection and test were 56, 26 and 4 respectively.

Further tests on external predictability

To know performance of the prediction, squared correlation coefficient values between the observed and predicted values of the test set compounds with intercept (r^2) and without intercept (r^2_0) were calculated. These values of all models have been represented in Tables X, XI and XII. All the models (except first two

Table VIII. Comparative study of statistical parameters of GFA-MLR models using different descriptors.

Type of Descriptors	Eq. No.	R^2 (Training Set)	Ra^2 (Training Set)	Q^2 (Training Set)	LOF	F	s	R^2_{pred} (Test Set)
MSA + Structural + physicochemical + spatial + electronic	10	0.614	0.591	0.550	0.387	26.760	0.573	0.505
	11	0.593	0.574	0.537	0.389	30.990	0.585	0.520

Table IX. Comparative study of best two network using different hidden layers.

Model No.	No. of Hidden Layer	No. of units in 3 different layers	No. of cross validated resampling	No. of epoch in backpropagation followed by conjugate gradient descent	Absolute Error Mean	Correlation Coefficient (r^2) between Obs. and Pred. values of the test set
1	3	40 38 36	10	500, 200	0.482	0.634
2	3	**43 39 36**	**20**	**700, 300**	**0.403**	**0.683**
3	2	43 39	10	500, 200	0.369	0.659
4	2	43 39	15	800, 300	0.373	0.638
5	1	43	20	700, 400	0.387	0.641
6	1	43	25	700, 400	0.377	0.635

ANN models) 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 [50]. According to Golbraikh and Tropsha [49], 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 XIII.

Moreover, R^2_{pred} value is mainly controlled by the value of $(Y_{test} - \bar{Y}_{training})^2$, i.e., the difference between observed value of test set and mean of training data set. Thus, it may not truly reflect the predictive capability on new dataset. Besides 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 been defined in the following manner [51]

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

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_o^2 value. In the best case r_m^2 will be equal to r^2 whereas in the worst case r_m^2 value will be zero. Here, the r_m^2 values of Equations 3, 10 and 11 and first two models of ANN are less than the recommended value (0.5). The best r_m^2 value is obtained from the ANN model 6 (Table XIV).

Overview

Different statistical methods like stepwise regression, PLS and FA-MLR have been applied to model CCR5 binding affinity of piperidine derivatives using different combinations of topological, structural, physicochemical, spatial and electronic descriptors to develop classical type QSAR models. Using topological and structural parameters the best equation based on internal validation was obtained

Table X. Comparison of external predictability characteristics of different models obtained from the training set using classical QSAR.

Type of Descriptors	Statistical Method	r^2	r_o^2	$(r^2 - r_o^2)/r^2$	r_m^2
Topological + structural	Stepwise	0.553	0.544	0.016	0.502
	PLS	0.578	0.577	0.001	0.562
	FA-MLR	0.535	0.526	0.018	0.483
Structural + physicochemical + spatial + electronic	Stepwise	0.578	0.563	0.027	0.506
	PLS	0.591	0.589	0.003	0.564
	FA-MLR	0.517	0.516	0.002	0.500
Topological + structural + physicochemical + spatial + electronic	Stepwise	0.592	0.591	0.003	0.568
	PLS	0.609	0.604	0.008	0.566
	FA-MLR	0.5473	0.5472	6.94E-06	0.546

Table XI. Comparison of external predictability characteristics of different GFA-MLR models.

Type of Descriptors	Model No.	r^2	r_o^2	$(r^2 - r_o^2)/r^2$	r_m^2
MSA + Structural + physicochemical + spatial + electronic	1	0.581	0.551	0.052	0.480
	2	0.572	0.555	0.029	0.480

Table XII. Comparison of external predictability characteristics of different ANN models.

Model No.	r^2	r_o^2	$(r^2 - r_o^2)/r^2$	r_m^2
1	0.634	0.560	0.117	0.461
2	0.683	0.548	0.198	0.432
3	0.659	0.637	0.034	0.560
4	0.638	0.632	0.011	0.586
5	0.641	0.634	0.011	0.586
6	0.6352	0.6350	0.0003	0.627

Table XIV. Comparison of best r_m^2 between observed and predicted values of the test set compounds using different techniques.

Statistical Methods	r_m^2 value
Stepwise Regression (Equation 7)	0.568
PLS (Equation 8)	0.566
FA-MLR (Equation 9)	0.546
GFA-MLR (Equation 11)	0.480
ANN (Model 6)	0.627

with FA-MLR ($Q^2 = 0.514$). But predictive potential of this model on the test chemicals was not satisfactory ($R^2_{pred} = 0.470$). According to the external validation statistics, the best model have been reported using PLS ($R^2_{pred} = 0.565$). However, this model produced insignificant Q^2 value (0.488). Only the stepwise regression derived model has given both acceptable Q^2 (0.510) and R^2_{pred} (0.504) values. When structural, physicochemical, spatial and electronic descriptors were used in combination, the best Q^2 value (0.562) was obtained from the stepwise regression derived model. But here also the external validation parameter (R^2_{pred}) is not satisfactory. The only significant R^2_{pred} value came from PLS ($R^2_{pred} = 0.571$). Next, topological descriptors were combined with other (structural, physicochemical, spatial and electronic) descriptors in search of better predictive models. In this case, the best Q^2 and R^2_{pred} values obtained were 0.530 (stepwise regression) and 0.580 (PLS regression) respectively. On using topological descriptors along with other descriptors, predictive R^2 value increased marginally (PLS model). In case of the 3D-QSAR study, linear models have been tried to develop from genetic function approximation using MSA descriptors in combination with structural, electronic, physicochemical and spatial descriptors. Although Equation (10) with five descriptors gave the best explained ($R_a^2 = 0.591$) and predicted variance ($Q^2 = 0.550$) of the CCR5 binding affinity along with lowest LOF score among the GFA models, higher predictive R^2 ($R^2_{pred} = 0.520$) was obtained in case of Equation (11) with four descriptors. None of these two equations contain MSA descriptors and their quality (explained variance, crossvalidated R^2 and predicted R^2) is not better than those of the best models obtained from classical QSAR approach. In search of better predictive models, nonlinear modeling was performed with artificial neural network. The models showed acceptable value of squared correlation coefficient for the observed and predicted values of the test set compounds. Further statistical validation was performed as recommended by Golbraikh and Tropsha [50] and Roy and Roy [51]. All models except first two models of ANN have satisfied the criteria of $(r^2 - r_0^2)/r^2$ value being less than 0.1. When r_m^2 test was been performed, Equations 3, 10 and 11 and first two models of ANN did not pass the test. The best r_m^2 value is obtained from the ANN model 6 (Table XIV). The scatter plots of observed versus predicted values of the test set compounds for five selected models using different techniques based on best r_m^2 values are shown in Figure 2.

Table XIII. Calculated values of k and k′ for different models as defined by Golbraikh and Tropsha [49].

	k′	k
Stepwise Regression (Equation 1)	1.0268	0.9593
PLS (Equation 2)	1.0144	0.9721
FA-MLR (Equation 3)	1.0323	0.9536
Stepwise Regression (Equation 4)	1.0529	0.9361
PLS (Equation 5)	1.0176	0.9695
FA-MLR (Equation 6)	1.0404	0.9459
Stepwise Regression (Equation 7)	1.0286	0.9592
PLS (Equation 8)	1.0213	0.9664
FA-MLR (Equation 9)	1.0188	0.967
GFA-MLR (Equation 10)	1.0295	0.9571
GFA-MLR (Equation 11)	1.025	0.9614
ANN (Model 1)	1.0425	0.9454
ANN (Model 2)	0.9746	1.0109
ANN (Model 3)	1.0097	0.9786
ANN (Model 4)	1.0156	0.9727
ANN (Model 5)	1.0164	0.9721
ANN (Model 6)	1.0209	0.9678

Conclusion

Among the classical QSAR models, the best model was obtained with stepwise regression using combination of structural, physicochemical, electronic and spatial descriptors based on internal validation while the best model based on external validation was obtained from PLS using combination of topological and other (structural, physicochemical, electronic and spatial) descriptors. The 3D-QSAR linear models did not provide any better result over the classical QSAR models with respect to both internal and external validations. However, when nonlinear mapping technique was applied to the set of 3D-QSAR descriptors, the best model based on modified r^2 (r_m^2) value was developed using one hidden layer.

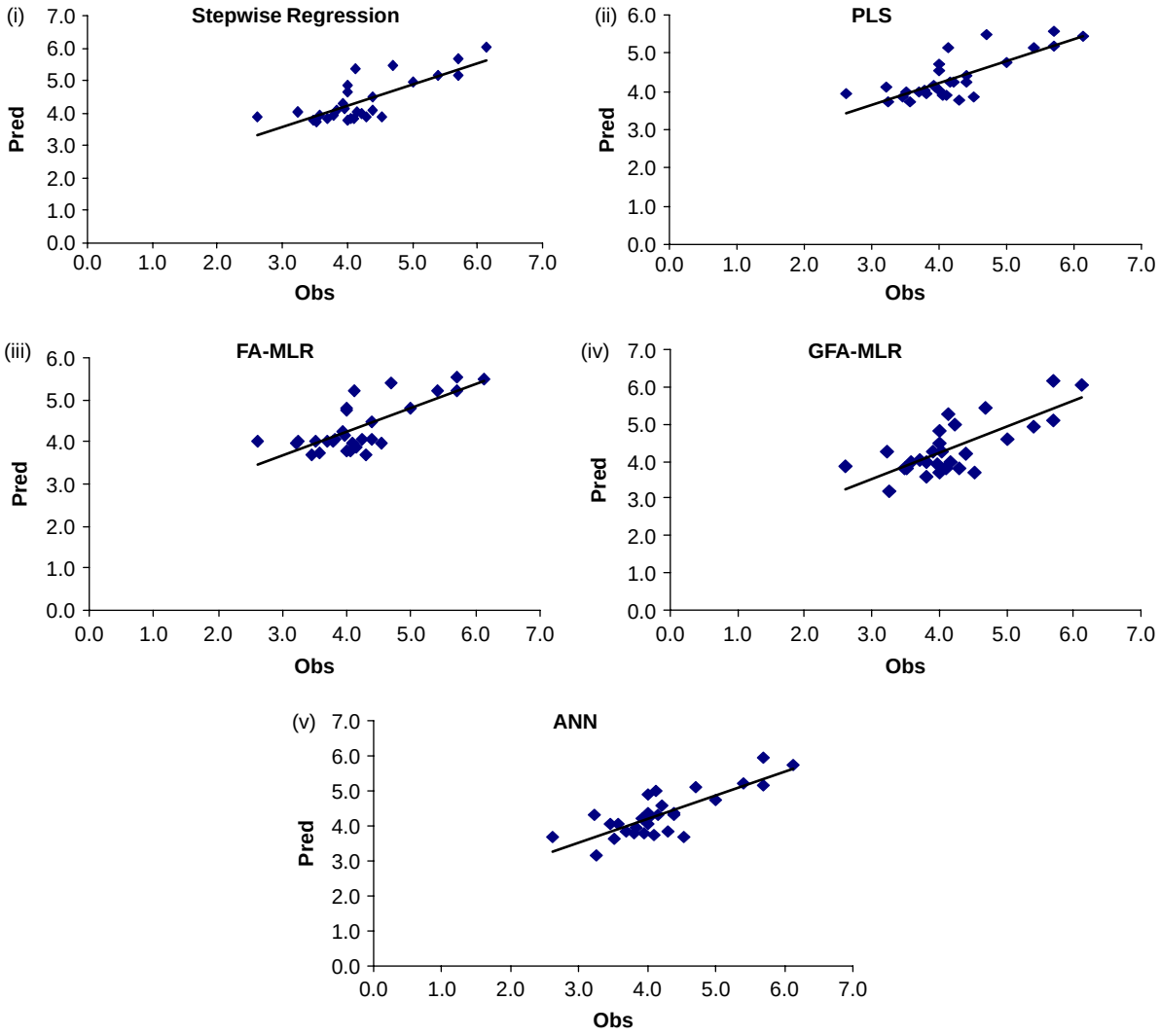


Figure 2. Scatter plots of observed versus predicted values of the test set compounds obtained from the best models (based on r^2_m value for the test set compounds) using (i) stepwise regression (Equation 7), (ii) PLS (Equation 8), (iii) FA-MLR (Equation 9), (iv) GFA-MLR (Equation 11) and (v) ANN (6^{th} model).

This confirms that nonlinear modeling outperforms the external predictability of linear models for this data set.

Acknowledgements

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] www.rhodes.edu/biology/glindquester/viruses/pagespass/hiv/hiv.html
[3] www.retrovirology.com/content/4/1/50
[4] Carrington M, Dean M, Martin MP, O'Brien SJ. Genetics of HIV-1 infection: Chemokine receptor CCR5 polymorphism and its consequences. Hum Mol Genet 1990;8(10): 1939–1945.
[5] Mack M, Pfirstinger J, Haas J, Nelson PJ, Kufer P, Riethmuller G, Schlondorff D. Preferential targeting of CD4-CCR5 complexes with bifunctional inhibitors: A novel approach to block HIV-1 infection. Immunology 2005;175: 7586–7593.
[6] Berchanski A, Lapidot A. Prediction of HIV-1 entry inhibitors neomycin-arginine conjugates interaction with the CD4-gp120 binding site by molecular modeling and multi-step docking procedure. Biochim Biophys Acta 2007;1768 (9):2107–2119.
[7] Liu S, Fan S, Sun Z. Structural and functional characterization of the human CCR5 receptor in complex with HIV gp120 envelope glycoprotein and CD4 receptor by molecular modeling studies. J Mol Model 2003;9(5):329–336.
[8] Roy K, Leonard JT. QSAR analyses of 3-(4-Benzylpiperidin-1-yl)-N-phenylpropylamine derivatives as potent CCR5 antagonists. J Chem Inf Model 2005;45:1352–1368.
[9] Aher YD, Agrawal A, Bharatam PV, Garg P. 3D-QSAR studies of substituted 1-(3,3-diphenylpropyl)-piperidinyl amides and

ureas as CCR5 receptor antagonists. J Mol Model 2007;13(4): 519–529.

[10] Xu Y, Liu H, Niu C, Luo C, Luo X, Shen J, Chen K, Jiang H. Molecular docking and 3D QSAR studies on 1-amino-2-phenyl-4-(piperidin-1-yl)-butanes based on the structural modeling of human CCR5 receptor. Bioorg Med Chem 2004;12(3):6193–6208.

[11] Song M, Breneman CM, Sukumar N. Three-dimensional quantitative structure-activity relationship analyses of piperidine-based CCR5 receptor antagonists. Bioorg Med Chem 2004;12(2):489–499.

[12] Afantitis A, Melagraki G, Sarimveis H, Koutentis PA, Markopoulos J, Igglessi-Markopoulou O. Investigation of substituent effect of 1-(3,3-diphenylpropyl)-piperidinyl phenylacetamides on CCR5 binding affinity using QSAR and virtual screening techniques. J Comput Aided Mol Des 2006; 20:83–95.

[13] Roy K, Leonard JT. QSAR modeling of HIV-1 reverse transcriptase inhibitor 2-amino-6- arylsulfonylbenzonitriles and congeners using molecular connectivity and E-state parameters. Bioorg Med Chem 2004;12:745–754.

[14] Leonard JT, Roy K. Classical QSAR modeling of HIV-1 reverse transcriptase inhibitor 2-amino-6-arylsulfonylbenzonitriles and congener. QSAR Comb Sci 2004;23:23–35.

[15] Leonard JT, Roy K. QSAR modeling of anti-HIV activities of alkenyldiarylmethanes using topological and physicochemical descriptors. Drug Des Discov 2003;18:165–180.

[16] Leonard JT, Roy K. Classical QSAR modeling of CCR5 receptor binding affinity of substituted benzylpyrazoles. QSAR Comb Sci 2004;23:387–398.

[17] Roy K, Leonard JT. Classical QSAR modeling of anti-HIV 2,3-diaryl-1,3-thiazolidin-4-ones. QSAR Comb Sci 2005;24: 579–592.

[18] Roy K, Leonard JT. QSAR by LFER model of cytotoxicity data of anti-HIV 5-phenyl-1-phenylamino-1H-imidazole derivatives using principal component factor analysis and genetic function approximation. Bioorg Med Chem 2005;13: 2967–2973.

[19] Roy K, Leonard JT. Topological QSAR modeling of cytotoxicity data of anti-HIV5-phenyl-1-phenylamino-imidazole derivatives using GFA, G/PLS, FA and PCRA techniques. Indian J Chem 2006;45A:126–137.

[20] Leonard JT, Roy K. QSAR by LFER model of HIV protease inhibitor mannitol derivatives using FA-MLR, PCRA, and PLS techniques. Bioorg Med Chem 2006;14:1039–1046.

[21] Leonard JT, Roy K. Comparative QSAR modeling of CCR5 receptor binding affinities of substituted 1-(3,3-diphenylpropyl)-piperidinyl amides and ureas. Bioorg Med Chem Lett 2006;16:4467–4474.

[22] Dorn CP, Finke PE, Oates B, Budhu RJ, Mills SG, MacCoss M, Malkowitz L, Springer MS, Daugherty BL, Gould SL, DeMartino JA, Siciliano SJ, Carella A, Carver G, Holmes K, Danzeisen R, Hazuda D, Kessler J, Lineberger J, Miller M, Schleif WA, Emini EA. Antagonists of the human CCR5 receptor as anti-HIV-1 agents. Part 1: Discovery and initial structure-activity relationships for 1-amino-2-phenyl-4-(piperidin-1-yl) butanes. Bioorg Med Chem Lett 2001;11: 259–264.

[23] Finke PE, Meurer LC, Oates B, Mills SG, MacCoss M, Malkowitz L, Springer MS, Daugherty BL, Gould SL, DeMartino JA, Sicilino SJ, Carella A, Carver G, Holmes K, Danzeisen R, Hazuda D, Kessler J, Lineberger J, Miller M, Schleif WA, Emini EA. Antagonists of the human CCR5 receptor as anti-HIV-1 agents. Part 2: Structure-activity relationships for substituted 2-aryl-1-[*N*-(methyl)-*N*-(phenylsulfonyl) amino]-4-(piperidin-1-yl) butanes. Bioorg Med Chem Lett 2001;11:265–270.

[24] Finke PE, Meurer LC, Oates B, Shah SK, Loebach JL, Mills SG, MacCoss M, Castonguay L, Malkowitz L, Springer MS, Gould SL, DeMartino JL. Antagonists of the human CCR5 receptor as anti-HIV-1 agents. Part 3: A proposed pharmacophore model for 1-[*N*-(methyl)-*N*-(phenylsulfonyl) amino]-2-(phenyl)-4-[4-(substituted)piperidin-1-yl] butanes. Bioorg Med Chem Lett 2001;11:2469–2473.

[25] Finke PE, Oates B, Mills SG, MacCoss M, Malkowitz L, Springer MS, Gould SL, DeMartino JA, Carella A, Carver G, Holmes K, Danzeisen R, Hazuda D, Kessle J, Lineberger J, Miller M, Schleif WA, Emini EA. Antagonists of the human CCR5 receptor as anti-HIV-1 agents. Part 4: Synthesis and structure-activity relationships for 1-[*N*-(methyl)-*N*-(phenylsulfonyl)amino]-2-(phenyl)-4-(4-(*N*-(alkyl)-*N*-(benzyloxycarbonyl)amino)piperidin-1-yl)butanes. Bioorg Med Chem Lett 2001;11:2475–2479.

[26] Cerius2 Version 4.10 is a product of Accelrys Inc., San Diego, CA.

[27] Roy K, Mandal AS. Development of linear and nonlinear predictive QSAR models and their external validation using molecular similarity principle for anti-HIV indolyl aryl sulfones. J Enz Inhib Med Chem 2007. DOI: 10.1080/14756360701811379.

[28] Eriksson L, Jaworska J, Worth AP, Cronin MTD, McDowell RM, Gramatica P. Methods for reliability and uncertainty assessment and for applicability evaluations of classification and regression-based QSARs. Environ Health Perspect 2003;111:1361–1375.

[29] Guha R, Jurs PC. Determining the validity of a QSAR model - A classification approach. J Chem Inf Model 2005;45:65–73.

[30] Leonard JT, Roy K. On selection of training and test sets for the development of predictive QSAR models. QSAR Comb Sci 2006;25(3):235–251.

[31] Roy K. On some aspects of validations of predictive QSAR models. Expert Opin Drug Discov 2007;2:1567–1577.

[32] Everitt B, Landau S, Leese M. Cluster analysis. London: Arnold; 2001.

[33] Dougherty ER, Barrera J, Brun M, Kim S, Cesar RM, Chen Y, Bittner M, Trent JM. Inference from clustering with application to gene-expression microarrays. J Comput Biol 2002;9:105–126.

[34] Darlington RB. Regression and linear models. New York: McGraw-Hill; 1990.

[35] Wold S. PLS for multivariate linear modeling. In: van de Waterbeemd H, editor. Chemometric methods in molecular design. Weinheim: VCH; 1995. p 195–218.

[36] Franke R. Theoretical drug design methods. Amsterdam: Elsevier; 1984.

[37] Franke R, Gruska A. Principal component and factor analysis. In: van de Waterbeemd H, editor. Chemometric methods in molecular design. Weinheim: VCH; 1995. p 113–163.

[38] Hopfinger AJ, Tokarsi JS. Three dimensional quantitative structure-activity relationship analysis. In: Charifson PS, editor. Practical applications of computer-aided drug design. New York: Marcel Dekker; 1997. p 105–164.

[39] Rogers D, Hopfinger AJ. Application of genetic function approximation to quantitative structure-activity relationships and quantitative structure-property relationships. J Chem Inf Comput Sci 1994;34:854–866.

[40] Zupan J, Gasteiger J. Neural networks in chemistry and drug design. Weinheim: Wiley-VCH; 1999.

[41] Tang Y, Jiang HL, Chen KX, Ji RY. QSAR study of artemisinin (Qinghaosu) derivatives using neural network method. Indian J Chem 1996;35B:325–332.

[42] Snedecor GW, Cochran WG. In: van de Waterbeemd H, editor. Statistical methods. New Delhi: Oxford and IBH; 1967. p 381.

[43] Wold S, Eriksson L. In: van de Waterbeemd H, editor. Chemometric methods in molecular design. Weinheim: VCH; 1995. p 312.

[44] Debnath AK. In: Ghose AK, Viswanadhan VN, editors. Combinatorial library design and evaluation. New York: Marcel Dekker Inc.; 2001. p 73.
[45] MINITAB is a statistical software of Minitab Inc., USA.
[46] SPSS is a statistical software of SPSS Inc., USA.
[47] STATISTICA is a statistical software of STATSOFT Inc., USA.
[48] Livingstone DJ, Salt DW. Judging the significance of multiple linear regression models. J Med Chem 2005;48:661–663.
[49] http://www.port.ac.uk/research/cmd/research/selectionbiasinmultipleregression
[50] Golbraikh A, Tropsha A. Beware of q^2!, J Mol Graphics Mod 2002;20:269–276.
[51] Roy P, Roy K. On some aspects of variable selection for partial least squares regression models. QSAR Comb Sci 2008;27: 302–313.