European Journal of Medicinal Chemistry 44 (2009) 2158–2171

Contents lists available at ScienceDirect

European Journal of Medicinal Chemistry

journal homepage: http://www.elsevier.com/locate/ejmech

Original article

QSAR models for 2-amino-6-arylsulfonylbenzonitriles and congeners HIV-1 reverse transcriptase inhibitors based on linear and nonlinear regression methods

Rongjing Hu [a,b], Jean-Pierre Doucet [b], Michel Delamar [b,*], Ruisheng Zhang [a,c,**]

[a] Department of Chemistry, Lanzhou University, Lanzhou, Gansu 730000, PR China
[b] Interfaces, Traitement, Organisation et Dynamique des Systèmes (ITODYS), Paris 7 (Paris-Diderot) University, CNRS UMR 7086, Bâtiment Lavoisier, 15 rue Jean Antoine de Baïf, 75205 Paris Cedex 13, France
[c] School of Information Science and Engineering, Lanzhou University, Lanzhou, Gansu 730000, PR China

ARTICLE INFO

Article history:
Received 1 May 2008
Received in revised form
13 October 2008
Accepted 20 October 2008
Available online 30 October 2008

Keywords:
QSAR
HIV-1 non-nucleoside reverse transcriptase inhibitors
NNRTI
SVM
PPR

ABSTRACT

A quantitative structure–activity relationship study of a series of HIV-1 reverse transcriptase inhibitors (2-amino-6-arylsulfonylbenzonitriles and their thio and sulfinyl congeners) was performed. Topological and geometrical, as well as quantum mechanical energy-related and charge distribution-related descriptors generated from CODESSA, were selected to describe the molecules. Principal component analysis (PCA) was used to select the training set. Six techniques: multiple linear regression (MLR), multivariate adaptive regression splines (MARS), radial basis function neural networks (RBFNN), general regression neural networks (GRNN), projection pursuit regression (PPR) and support vector machine (SVM) were used to establish QSAR models for two data sets: anti-HIV-1 activity and HIV-1 reverse transcriptase binding affinity. Results showed that PPR and SVM models provided powerful capacity of prediction.

1. Introduction

Acquired immune deficiency syndrome or acquired immunodeficiency syndrome (AIDS) is a collection of symptoms and infections resulting from the specific damage to the immune system caused by the human immunodeficiency virus (HIV) in humans [1]. Since it was first identified in the Western world in 1981, AIDS has developed into a worldwide pandemic of disastrous proportions. According to the latest figures published today in the UNAIDS/WHO 2006 AIDS Epidemic Update [2], an estimated 39.5 million people are living with HIV. Among them about 530,000 children less than 15 years old were infected mainly through mother-to-child transmission. In 2006, 2.9 million people died of AIDS-related illnesses [3].

There are two species of HIV which infect humans: HIV-1 and HIV-2. HIV-1 is more virulent. It is easily transmitted and is the cause of the majority of HIV infections [4]. For about 20 years, various anti-HIV-1 drugs were selected after advanced clinical trials for the treatment of patients. There are three classes [5]: (i) nucleoside reverse transcriptase inhibitors (NRTIs), such as zidovudine, and nucleotide reverse transcriptase (NtRTIs); (ii) non-nucleoside reverse transcriptase inhibitors (NNRTIs), such as nevirapine; (iii) protease inhibitors (PIs), such as saquinavir.

Reverse transcriptase (RT) provides essential enzymatic activity for HIV-1. When HIV infects a cell, reverse transcriptase copies the viral single stranded RNA genome into a double-stranded viral DNA. The viral DNA is then integrated into the host chromosomal DNA which then allows host cellular processes, such as transcription and translation to reproduce the virus. Due to its essential role in HIV-1 replication, RT is a major target for the development of antiretroviral agents [6]. Reverse transcriptase inhibitors (RTIs) block reverse transcriptase's enzymatic function and prevent completion of synthesis of the double-stranded viral DNA thus preventing HIV from multiplying. NNRTIs are one class of allosteric inhibitors which bind near the substrate binding site of RT and induce a conformational change that results in reduced enzymatic activity [7,8]. Several classes of NNRTIs were discovered, for instance: 1-(2-hydroxyethoxymethyl)-6-(phenylthio)thymine (HEPT) and 4,5,6,7-tetrahydroimidazo[4,5,1-*jk*][1,4]benzodiazepin-2(1*H*)-one (TIBO) derivatives. Recently, more types of NNRTIs were

* Corresponding author. Tel.: +33 1 57 27 54 32; fax: +33 1 57 27 72 63.
** Corresponding author. Department of Chemistry, Lanzhou University, Lanzhou, Gansu 730000, PR China. Tel.: +86 931 8914008421.
E-mail addresses: michel.delamar@univ-paris-diderot.fr (M. Delamar), zhangrs@lzu.edu.cn (R. Zhang).

0223-5234/$ – see front matter © 2008 Elsevier Masson SAS. All rights reserved.
doi:10.1016/j.ejmech.2008.10.021

designed and synthesized. Chan et al. [9] designed a class of 2-amino-6-arylsulfonylbenzonitriles and their thio and sulfinyl congeners (Fig. 1) which have anti-HIV-1 activity as NNRTIs.

Quantitative structure–activity relationships (QSAR) are based on the assumption that the biological activities of chemical compounds are quantitatively correlated with some of their physicochemical parameters such as solubility, lipophilicity, polarity and steric properties. From the early linear Hansch approach [10–12], Free-Wilson analysis [13] and molecular connectivity methods [14,15] to the late 3D-QSAR [16], comparative molecular field analysis (CoMFA) [17] and comparative molecular similarity indices analysis (CoMSIA) [18], QSAR methods have been successfully used to predict the activity of drugs and drug-likes and made large contributions to computer-aided drug design. Pattern recognition methods, such as linear learning machine (LLM), K-nearest neighbor (KNN), discriminant analysis (DA), principal component analysis (PCA), partial least squares (PLS) and cluster analysis were used to build QSAR models. The applications of artificial neural networks (ANN) in the area of QSAR were published by Aoyama, Suzuki, and Ichikawa in 1990 with the promise that "the effective application of such neural networks may bring forth a breakthrough in the current state of QSAR analysis" [19,20]. For twenty years, ANN has been a very useful tool in QSAR studies as a nonlinear regression method [21,22]. General regression neural network (GRNN), a memory-based feed forward network, has been explored for quantitative structure–property relationship (QSPR) modeling [23] as well as for developing QSAR studies [24]. Application of multivariate adaptive regression splines (MARS) to chemical studies was introduced by De Veaux et al. [25] and was successfully used in QSAR [26]. In 1996, Projection pursuit regression (PPR) as a nonparametric method was introduced into QSAR field [27]. In recent years, QSAR research was improved much because of the emergency of support vector machine (SVM), a novel learning machine. SVM was successfully used to predict the property and activity of biomolecules [28,29].

Researchers have modeled [30,31] anti-HIV activity and RT binding affinity data of 2-amino-6-arylsulfonylbenzonitriles and congeners using molecular connectivity, e-state parameter and physicochemical parameters like hydrophobicity, molar refractivity, and electronic parameter (Hammett σ). Freitas built another QSAR model using multivariate image analysis (MIA) descriptors [32]. A QSAR study of this kind of HIV-1 RT binding affinity data set was performed by Tang et al. [33] based on nonlinear PLS, back-propagation neural networks (BPNN), support vector machine (SVM) and a mixed radial basis function network-based transform for a nonlinear support vector machine (RBFN-SVM) methods with 14 descriptors generated from Cerius^2 3.5 [34]. The present study improves the QSAR model with geometrical descriptors, electrostatic descriptors and quantum-chemical descriptors generated from the CODESSA (comprehensive descriptors for structural and statistical analysis) software with 3D preference conformation of the compounds based on several linear and nonlinear methods: MLR, MARS, GRNN, RBFNN, SVM and PPR. We attempt to explore a QSAR model which has more predictive ability and compare the results of the six methods in modeling.

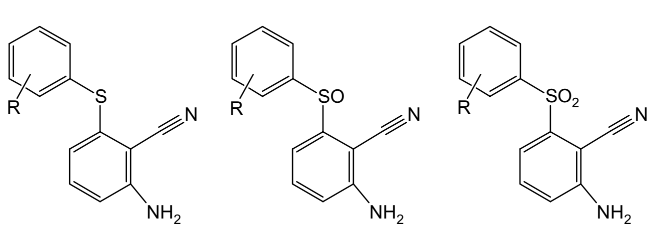


Fig. 1. 2-Amino-6-arylsulfonylbenzonitriles (compounds **1–23**) and their thio (compounds **24–38**) and sulfinyl (compounds **39–68**) congeners.

2. Results and discussion

2.1. Data sets

68 molecules (2-amino-6-arylsulfonylbenzonitriles and congeners) were selected from the literature [9]. Among them 64 compounds with precise IC50 values for anti-HIV-1 activity and 51 compounds with precise IC50 values for HIV-1 RT binding affinity were used for QSAR. Chemical structures and biological properties, (anti-HIV-1 activity and HIV-1 RT binding affinity expressed in pIC_{50} ($-\log IC_{50}$)), for the complete set of compounds (divided in the corresponding training and test sets based on principal component analysis) are presented in Tables 1 and 2.

2.2. QSAR model for an anti-HIV-1 activity data set

About 600 descriptors were calculated in CODESSA for each molecule. After forward stepwise regression and "break point" descriptor selection, the linear model for the whole set contained 6 molecular descriptors (see Section 4.2): ${}^{0}\chi_{K\&H}$, ${}^{3}\chi_{K\&H}$ and ${}^{3}K$ indices are topological; S_{Zx} is geometrical; ${}^{Max}E_{nn,CS}$ and ${}^{Max}E_{R,CH}$ are quantum-chemical descriptors, generated from the output results of the MOPAC program.

The selection of the optimum number of descriptors is shown in Fig. 2 which is a plot of R^2 and R^2_{cv} for the data set as a function of the number of descriptors. The "break point" is the "6-descriptor point" for this anti-HIV-1 activity set (Fig. 2a).

The correlation model is shown in Table 3 with a square standard error s^2 of 0.205, a square correlation coefficient R^2 of 0.805 and R^2_{CV} of 0.759. The numerical values of all of the descriptors are listed in Table 4. The linear regressions thus obtained are more satisfactory than models developed in previous studies ($R^2 < 0.8$) [30,31].

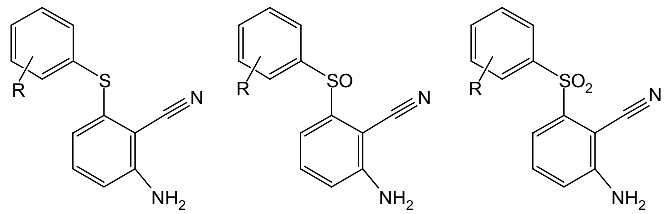
Principal components analysis (PCA) was performed with the calculated structure descriptors for the whole data set to detect the homogeneities in the data set, and further to show spatial location of samples to assist the separation of the data into training and test set.

The PCA results show that two principal components (PC1 and PC2) describe 69.2% of the overall variables, as follows: PC1 = 45.3%, PC2 = 24.4%. Since almost all variables can be accounted for by the first two PCs, their score plot is a reliable representation of the spatial distribution of the points for the data set. The plot of PC1 against PC2 (Fig. 3) displays the distribution of compounds over the first two principal components space. Several small clusters of compounds can be found in this figure, corresponding to the different S functional groups (sulfides, sulfoxides and sulfones) and types of R substituents in the aromatic ring.

According the results of PCA, the whole data were divided into a training set of 48 compounds to develop the models and a test set of 16 compounds to evaluate the models based on several rules: (i) the range of the activity values of both the training set and test set should be covered from the lowest to the highest; (ii) the points corresponding to the training set in the PCA plot should not be out of the main clusters. The two sets appear in Table 1.

In order to develop the models for the anti-HIV-1 activity data set and evaluate the predictive capacity of each model, a double cross-validation was performed. Firstly, parameters were optimized to determine the best models using the leave-one-out (LOO) method, which was performed for the training set to select the optimum values of parameters. This procedure consists in removing one example from the training set, constructing the decision function on the basis of only the remaining training data and then testing on the removed example. In this fashion, one tests all examples of the training data and measures the fraction of

Table 1
Observed and predicted anti-HIV-1 activity of 2-amino-6-arylsulfonylbenzonitriles and their congeners with MLR, MARS, RBFNN, GRNN, SVM and PPR methods.



1-23 **24-38** **39-68**

No.	R	Observed	Calculated					
			MLR	MARS	RBFNN	GRNN	SVM	PPR
1	H	1.836	1.758	2.290	1.987	1.896	1.756	2.039
2	2-OCH_3	2.367	2.145	2.022	2.280	2.121	2.173	1.848
3	3-OCH_3	2.222	1.908	1.836	2.062	2.139	2.029	1.995
4	2-CH_3	1.796	2.376	2.103	2.342	2.057	1.990	2.230
5*	3-CH_3	2.215	2.060	1.997	2.187	1.985	2.028	1.745
6	4-CH_3	0.939	1.994	2.012	2.115	1.987	1.966	1.756
7	2-Cl	2.387	1.903	2.058	1.966	1.978	1.750	2.176
8	3-Cl	2.131	2.126	2.015	2.215	2.014	2.101	1.713
9	4-Cl	–	–	–	–	–	–	–
10	2-Br	1.523	2.180	2.477	2.120	2.001	1.839	2.098
11	3-Br	2.292	2.786	2.317	2.763	2.162	2.486	2.539
12	3-F	2.009	1.723	2.159	1.934	1.923	1.815	1.774
13	2-CN	–	–	–	–	–	–	–
14	3-CN	2.762	1.948	1.885	2.080	1.987	2.001	1.869
15	4-CN	1.359	2.102	1.977	2.184	1.967	2.105	1.599
16	3-CF_3	1.893	1.619	1.489	1.700	2.266	2.020	1.699
17*	3-NH_2	1.502	1.776	2.080	1.967	1.940	1.865	1.536
18	2,5-Cl_2	–	–	–	–	–	–	–
19	3,5-$(CH_3)_2$	3.367	3.862	3.470	3.832	3.596	3.629	3.507
20	3,5-Cl_2	–	–	–	–	–	–	–
21	3-Cl, 5-CH_3	2.754	2.605	2.119	2.605	2.281	2.479	2.608
22	3-OCH_3, 5-CH_3	2.699	2.676	2.094	2.664	2.370	2.524	2.535
23	3-OCH_3, 5-CF_3	2.292	2.653	2.834	2.581	2.303	2.486	2.504
24	2-OCH_3	2.319	2.202	2.304	2.213	2.183	2.125	2.579
25*	3-OCH_3	1.796	1.761	2.026	1.899	1.977	1.701	1.989
26	2-CH_3	1.032	0.983	1.513	1.139	1.388	0.838	0.905
27	3-CH_3	1.534	1.981	1.802	2.034	1.862	1.749	1.405
28	4-CH_3	1.310	1.693	1.539	1.753	1.800	1.504	1.797
29*	2-Br	1.407	1.704	2.548	1.626	2.401	1.053	1.389
30	3-Br	4.097	3.124	2.844	2.983	2.542	2.365	3.767
31*	4-Br	1.694	2.217	1.931	2.168	2.165	1.811	1.999
32	2-CN	2.409	1.607	1.668	1.666	2.081	2.007	1.885
33	3-CN	1.848	2.100	1.906	2.115	2.006	2.035	2.012
34	3-CF_3	1.398	1.770	2.226	1.727	1.956	1.488	1.632
35*	3,5-$(CH_3)_2$	3.469	2.934	2.157	2.863	3.002	3.188	3.017
36	2,5-Cl_2	2.007	2.237	2.609	2.141	2.374	2.201	2.052
37	3-Cl, 5-CH_3	3.495	3.116	2.566	3.008	3.030	3.301	3.109
38	3-OCH_3, 5-CF_3	2.684	2.458	2.353	2.406	2.689	2.877	2.643
39	H	2.699	2.144	2.841	2.155	2.757	2.605	2.735
40	2-OCH_3	3.222	3.310	3.496	3.424	3.343	3.416	3.374
41	3-OCH_3	3.046	3.116	3.132	3.101	3.116	3.239	3.468
42	4-OCH_3	1.602	1.679	1.844	1.560	1.652	1.796	1.539
43*	2-CH_3	2.638	2.459	2.907	2.419	2.853	2.862	2.267
44	3-CH_3	3.398	2.994	2.967	2.946	2.988	3.177	3.086
45	4-CH_3	2.022	2.112	2.061	2.105	2.477	2.216	2.099
46*	2-Cl	2.387	2.456	2.785	2.460	2.768	2.250	2.244
47	3-Cl	3.229	2.737	2.749	2.737	2.988	2.779	2.994
48	4-Cl	2.523	2.631	2.636	2.630	2.842	2.524	2.338
49*	2-Br	2.301	2.646	3.524	2.675	3.018	2.217	2.105
50	3-Br	3.268	3.338	3.553	3.343	3.106	3.242	3.287
51	4-Br	1.699	2.256	1.752	2.270	1.935	1.892	1.911
52	2-F	2.523	2.459	2.700	2.452	2.608	2.329	2.538
53*	3-F	2.523	1.644	1.383	1.638	2.229	1.478	1.943
54	2-CN	2.268	2.934	2.575	2.922	2.635	2.462	2.761
55	3-CN	2.62	2.531	2.440	2.540	2.901	2.490	2.627
56*	4-CN	1.097	1.202	0.658	1.215	2.097	0.692	1.510
57	3-CF_3	2.456	2.028	2.739	1.989	2.714	2.262	2.462
58*	2,5-Cl_2	3.523	2.940	3.340	3.058	3.924	2.935	2.850
59	3,5-Cl_2	4.155	4.465	4.211	4.467	4.180	3.981	4.439
60	3,5-$(CH_3)_2$	5.000	4.395	4.020	4.433	4.086	4.487	4.745
61	3-Br, 5-CH_3	4.699	4.772	4.698	4.882	4.381	4.771	4.782
62*	3-Cl, 5-CH_3	4.523	4.314	4.035	4.357	4.155	4.269	4.478
63	3-OCH_3, 5-CH_3	4.301	4.582	4.407	4.605	4.219	4.495	4.467
64	3-OCH_3, 5-CF_3	4.046	3.822	4.565	3.858	4.111	3.852	4.005

Table 1 (*continued*)

No.	R	Observed	Calculated					
			MLR	MARS	RBFNN	GRNN	SVM	PPR
65*	3-OH, 5-CH_3	3.367	2.515	1.851	2.544	2.239	2.359	2.511
66	3-OCH_2CH_3, 5-CH_3	4.222	3.822	3.888	3.877	4.096	4.095	3.826
67*	3-$O(CH_2)_2CH_3$, 5-CH_3	4.222	4.158	4.358	4.211	4.075	4.553	4.156
68*	3-$O(CH_2)_3CH_3$, 5-CH_3	3.222	3.041	3.167	2.979	2.869	2.860	3.051

Compounds labeled with "*" are the test set; other compounds are the training set.

errors over the total number of training examples. The LOO cross-validation method has the following advantages: (i) overfitting can be avoided; (ii) the model selection criteria are tractable; (iii) the computational requirements are relatively low. Then an external test set was used to evaluate the model.

For the MARS model, we used default values for parameters like "penalty" and "thresh" except for "degree". In MARS function, "penalty" means an optional value specifying the cost per degree of freedom charge and its default value is 2 while the default value of the forward stepwise stopping threshold is 0.001. "Degree" means an optional integer specifying maximum interaction degree. Its default value is 1. In this study, the optimum value of "degree" was 2.

For the RBFNN model, the "spread" and the number of the radial basis functions (the hidden layer units) are the two important parameters influencing the performances of the RBFNN. The selection of the optimal width value for RBFNN was performed by systemically changing its value in the training step. The values which gave the best LOO cross-validation result were used in the models. Each minimum error on LOO cross-validation was plotted versus the width (Fig. 4) and the minimum was chosen as the optimal condition. Finally, the number of the hidden layer units was 7 and the optimal spread was 2.5.

For the GRNN model, there is only one parameter: "spread", which is the width. As with RBFNN, leave-one-out cross-validation of the training set was performed to optimize "spread". Fig. 5 is the plot of each minimum error versus "spread" on LOO cross-validation and the minimum was chosen as the optimal condition, which is 0.25.

For the PPR model, several parameters need to be determined. These are "nterms", "span" and "optlevel". "nterms" determines the number of terms to include in the final model. "Span" describes the fraction of the observations in the span of the smoother. The levels of optimization (argument "optlevel") differ in how thoroughly the models are refitted during this process. At level 0 the existing ridge terms are not refitted. At level 1 the projection directions are not refitted, but the ridge functions and the regression coefficients are. Levels 2 and 3 refit all the terms and are equivalent for one response; level 3 is more careful to re-balance the contributions from each regressor at each step and so is a little less likely to converge to a saddle point of the sum of squares criterion. In this model, the optimum of "nterms", "span" and "optlevel" are 3, 0.3 and 2.

For the SVM model, there are three parameters to determine the performances of SVM for regression. These three parameters should be optimized in this model. They are: the capacity parameter C, ε of ε-insensitive loss function, and the parameter of the kernel type K. C is a regularization parameter that controls the trade-off between maximizing the margin and minimizing the training error. If C is too small then insufficient stress will be placed on fitting the training data. If C is too large then the algorithm will overfit the training data. Prediction error is scarcely influenced by C (if C is large enough). The optimal value for ε depends on the type of noise present in the data, which is usually unknown.

The kernel type is another important parameter. For regression tasks, the Gaussian kernel is commonly used. The form of the Gaussian function in R is as follows:

$$\exp\left(-\gamma^{*}|u-v|^{2}\right) \tag{1}$$

where γ is a constant, the parameter of the kernel; u, v are two independent variables; γ controls the amplitude of the Gaussian function and therefore, controls the generalization ability of SVM. γ should be optimized. We performed leave-one-out cross-validation to select the optimum values of the parameters. Finally the optimum values of γ, ε and C were fixed to 0.02, 0.2 and 100, respectively, and the final number of support vectors was 35.

The calculated values for every approach are listed in Table 1. Table 5 shows the final results obtained with each model. There are large differences between them. As linear regressions, MLR yields high R^2 values of 0.793 and 0.840 for the training and test set, respectively, and low MSE values of 0.19 and 0.18. MARS gives the worst training model among all the regressions with R^2 and MSE values of 0.730 and 0.25, respectively, and the results for the test set are dissatisfactory with $R^2 = 0.478$. The results we obtained with RBFNN (MSE = 0.19 for the training set and 0.18 for the test set) were satisfactory. However, the results of GRNN with the test set (MSE = 0.32) are not as good as those obtained with RBFNN. Compared to RBFNN and GRNN, SVM offers better results. The best model is obtained with PPR with the highest R^2 value of 0.890 for the training set and the lowest MSE value (0.10). The differences between the models are apparent in Fig. 6 which show the correlation of calculated values versus observed values for each of them. The plots for SVM (e) and for PPR (f) in Fig. 6 converge along the $y = x$ line, while the other plots show appreciably more dispersion.

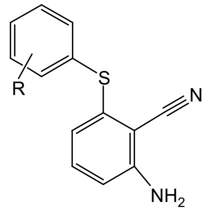
2.3. QSAR model for HIV-1 RT binding affinity data set

HIV-1 RT binding affinity is the activity of the compounds against HIV-1 reverse transcriptase. In order to avoid the intercorrelation, a correlation analysis between the values of anti-HIV-1 activity and HIV-1 RT binding affinity was performed. The correlation coefficient is 0.864. That means the two types of activity are not strictly correlated. So, different descriptors would be used for the models for HIV-1 RT binding affinity data set.

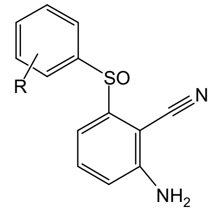
The selection of the optimum number of descriptors is shown in Fig. 2 where the "break point" is the "5-descriptor point" for this HIV-1 RT binding affinity set (Fig. 2b). These 5 descriptors were (see Section 4.2): one geometrical descriptor $S_{Zx,r}$ and four quantum-chemical ones: ${}^{\mathrm{Max}}E_{\mathrm{R,CH}}$, ${}^{\mathrm{Max}}E_{\mathrm{exc,CH}}$, ${}^{\mathrm{Max}}E_{\mathrm{ne,CN}}$, and ${}^{\mathrm{Min}}N_{\mathrm{N}}$. $S_{Zx,r}$ is not identical to S_{Zx}. This is a relative shadow area of a molecule. Among the 5 descriptors, ${}^{\mathrm{Max}}E_{\mathrm{R,CH}}$ is the only one which was used in both data sets. Table 6 shows the correlation model for the whole set with a standard deviation s^2 of 0.285, a square correlation coefficient R^2 of 0.744, and R^2_{CV} of 0.678. The values of the 5 descriptors are listed in Table 7.

As for the anti-HIV-1 activity data set, PCA was performed. It also generated two principal components. The first factor PC1′

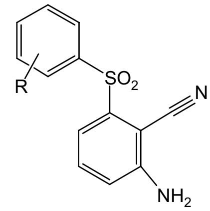
Table 2
Observed and predicted anti-HIV-1 RT binding affinity of 2-amino-6-arylsulfonylbenzonitriles and their congeners with MLR, MARS, RBFNN, GRNN, SVM and PPR methods.



1-23



24-38



39-68

No.	R	Observed	Calculated					
			MLR	MARS	RBFNN	GRNN	SVM	PPR
1	H	2.061	2.743	2.300	2.989	2.322	2.482	2.476
2	2-OCH_3	2.569	1.363	2.070	1.855	2.432	2.405	2.253
3#	3-OCH_3	2.824	2.524	2.135	2.573	2.287	2.354	2.596
4	2-CH_3	–	–	–	–	–	–	–
5	3-CH_3	3.018	3.205	2.603	2.756	2.826	2.854	3.190
6#	4-CH_3	2.244	2.590	2.591	1.937	2.659	2.434	2.614
7	2-Cl	2.143	2.269	2.133	2.460	2.246	2.307	2.202
8	3-Cl	1.796	2.349	2.030	2.420	2.407	2.171	2.182
9	4-Cl	1.921	2.212	2.030	2.237	2.454	2.069	2.314
10	2-Br	–	–	–	–	–	–	–
11#	3-Br	1.824	2.322	2.029	2.357	2.483	2.117	2.140
12	3-F	1.921	1.430	1.955	1.775	2.043	1.756	2.201
13	2-CN	2.041	1.985	2.046	2.075	2.356	2.038	1.965
14	3-CN	2.959	2.312	2.313	2.159	2.607	2.112	2.124
15	4-CN	–	–	–	–	–	–	–
16#	3-CF_3	2.149	2.451	2.369	2.386	2.635	2.209	2.265
17	3-NH_2	–	–	–	–	–	–	–
18#	2,5-Cl_2	2.456	3.107	3.160	2.992	2.729	2.701	2.897
19	3,5-$(CH_3)_2$	2.959	3.323	2.596	2.856	2.856	2.952	3.275
20	3,5-Cl_2	3.921	2.533	2.738	2.411	2.608	2.262	2.636
21	3-Cl, 5-CH_3	2.77	2.891	2.672	2.595	2.684	2.608	2.629
22#	3-OCH_3, 5-CH_3	3.854	3.421	2.624	2.815	2.902	3.039	3.327
23	3-OCH_3, 5-CF_3	1.886	1.771	2.185	1.815	2.398	1.756	1.935
24	2-OCH_3	1.921	2.386	2.925	2.373	2.680	2.216	2.234
25	3-OCH_3	1.721	1.927	1.570	1.930	1.981	2.059	1.901
26	2-CH_3	–	–	–	–	–	–	–
27	3-CH_3	2.000	1.950	1.758	1.882	2.204	2.165	1.990
28	4-CH_3	–	–	–	–	–	–	–
29	2-Br	–	–	–	–	–	–	–
30	3-Br	2.319	2.056	2.684	1.917	2.113	2.105	2.064
31	4-Br	–		–	–	–	–	–
32#	2-CN	2.004	1.543	2.720	1.491	2.493	1.688	2.113
33	3-CN	–	–	–	–	–	–	–
34	3-CF_3	–	–	–	–	–	–	–
35	3,5-$(CH_3)_2$	3.301	3.306	2.982	2.877	3.179	3.136	3.259
36	2,5-Cl_2	2.208	2.425	2.497	2.603	2.369	2.372	2.038
37	3-Cl, 5-CH_3	3.284	3.443	3.360	3.777	3.180	3.449	3.315
38#	3-OCH_3, 5-CF_3	3.046	2.686	2.425	2.237	2.734	2.664	2.528
39	H	2.161	2.513	2.474	2.519	2.540	2.325	2.544
40	2-OCH_3	2.854	2.822	2.695	2.442	2.633	2.566	2.444
41#	3-OCH_3	3.222	3.567	3.462	3.404	2.990	3.315	3.510
42#	4-OCH_3	1.886	2.656	3.412	2.434	2.610	2.386	2.555
43	2-CH_3	2.347	2.505	2.849	2.518	2.792	2.512	2.749
44	3-CH_3	3.699	3.415	4.317	3.567	3.495	3.541	3.331
45	4-CH_3	2.137	3.037	3.834	2.980	2.817	2.804	2.755
46	2-Cl	2.229	2.592	2.735	2.509	2.696	2.388	2.614
47	3-Cl	3.398	3.229	2.952	3.087	2.911	2.949	3.025
48	4-Cl	–	–	–	–	–	–	–
49#	2-Br	1.921	2.140	2.003	2.417	2.307	2.134	2.199
50	3-Br	3.699	3.090	2.944	2.927	2.874	2.791	2.857
51	4-Br	–	–	–	–	–	–	–
52	2-F	2.301	2.863	2.815	2.735	2.660	2.667	2.455
53	3-F	–	–	–	–	–	–	–
54#	2-CN	2.222	2.936	3.431	2.890	2.903	2.880	2.667
55	3-CN	2.745	3.159	3.076	3.081	2.923	2.934	3.015
56	4-CN	–	–	–	–	–	–	–
57	3-CF_3	2.276	2.765	2.894	2.706	2.813	2.496	2.329
58	2,5-Cl_2	3.523	3.539	2.931	3.180	3.198	3.358	3.491
59	3,5-Cl_2	4.523	4.499	4.410	4.941	3.769	4.687	4.714
60	3,5-$(CH_3)_2$	5.155	4.618	4.850	5.106	4.822	4.990	5.193
61	3-Br, 5-CH_3	5.523	4.568	4.660	4.949	4.849	4.955	5.028
62	3-Cl, 5-CH_3	5.301	4.648	5.004	5.197	4.907	5.130	5.291

Table 2 (*continued*)

No.	R	Observed	Calculated					
			MLR	MARS	RBFNN	GRNN	SVM	PPR
63[#]	3-OCH_3, 5-CH_3	5.000	4.652	4.581	5.152	4.716	4.989	5.291
64	3-OCH_3, 5-CF_3	4.398	4.345	3.664	4.179	3.688	4.341	4.353
65	3-OH, 5-CH_3	–	–	–	–	–	–	–
66	3-OCH_2CH_3, 5-CH_3	–	–	–	–	–	–	–
67	3-$O(CH_2)_2CH_3$, 5-CH_3	–	–	–	–	–	–	–
68	3-$O(CH_2)_3CH_3$, 5-CH_3	3.398	4.300	3.834	4.253	4.105	4.294	4.018

Compounds labeled with "#" are the test set; other compounds are the training set.

(36.6%) and the second factor PC2′ (31.0%) together account for 67.6% of the total variance in the data. The scatter was projected onto the principal components plane (Fig. 7). Fig. 7 shows again several small clusters corresponding to the different S functional groups (sulfides, sulfoxides and sulfones) and types of R substituents on the aromatic ring. With the PCA results, 51 compounds were divided into training set and test set (Table 2) and parameters were determined before building models with LOO cross-validation method.

For the MARS model, the parameters "penalty" and "thresh" were default values 2 and 0.001. The optimum value of "degree" was 1. For the RBFNN model, the "spread" parameter and the number of the hidden layer units were optimized with LOO cross-validation methods. From Fig. 8, the optimum value of "spread" is 2.25. The number of hidden layer units was 9. For the GRNN model, "spread" was optimized with leave-one-out cross-validation of the training set. The plot of RMS versus width of GRNN is shown in Fig. 9, in which the lowest point corresponds to the optimal "spread" that is 0.25. For the PPR model, the optimum of "nterms", "span" and "optlevel" were 1, 0.1 and 2. For the SVM model, the final values of γ, ε and C were 0.007, 0.2 and 100, respectively, and the final number of support vectors was 28.

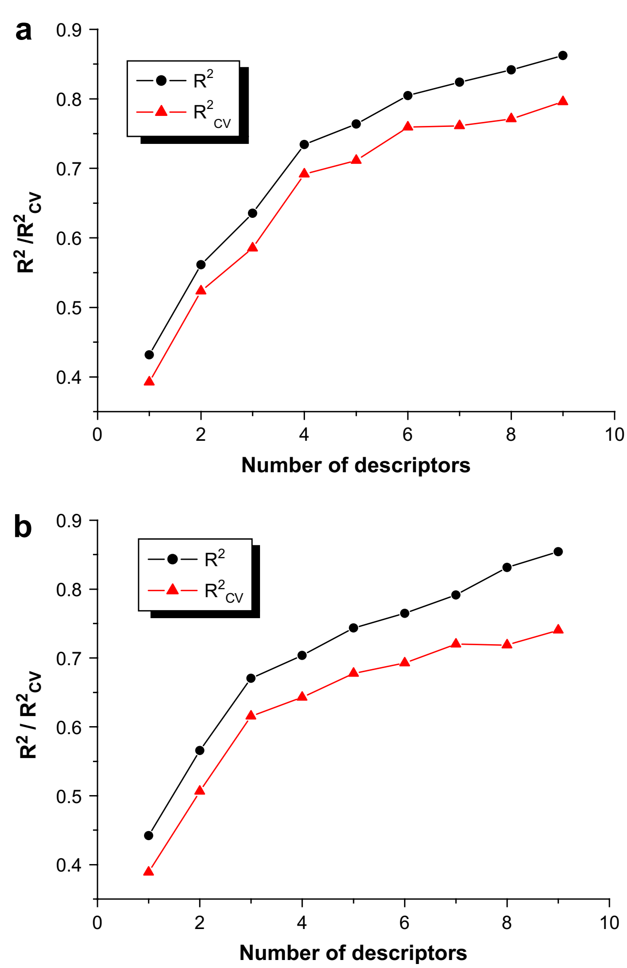


Fig. 2. Plots of R^2 and R^2_{cv} for anti-HIV-1 activity set (a) and HIV-1 RT binding affinity set (b) as a function of the number of descriptors for the two 1–9-parameter models.

All the results are gathered in Table 8 and the calculated values versus observed values correlations are plotted in Fig. 10. As with the anti-HIV-1 activity data set, MARS yielded the worst model for the HIV-1 RT binding affinity data set with a very low R^2 value of 0.345 and a very high MSE value of 0.59 for the test set. The predictive capacity of this model is poor. The nonlinear regressions of RBFNN and GRNN are also not very satisfactory with poor test results. SVM test set results are much better than those of RBFNN and GRNN. R^2 of the SVM model is 0.811 for the training set and 0.802 for the test set. Finally, PPR yielded the best model with the highest R^2 value of 0.843 for the training set and of 0.843 for the test set.

For the same data set, Tang et al. built QSAR models [33]. With SVM, the squared correlation coefficient (R^2) was 0.846 for the training set and 0.753 for the test set. However, the descriptors used by these authors were very different and more numerous than ours. They used 14 descriptors such as total charge, subgraph count indices (SC-3 cluster), a spatial descriptor (Shadow-Zlength) and the Jurs descriptor PPSA2 (total charge weighted partial positively charged molecular surface area).

2.4. Comparison of the six models

For the two data sets, the above results show a similar trend: linear models are not very satisfactory; the MARS models yield the poorest results whereas SVM and PPR are the best regression approaches to build QSAR models.

Table 3
Descriptors, coefficients, standard error, and *t* values for the linear model of anti-HIV-1 activity data set.

Physico-chemical meaning	Descriptors	Coefficient	Standard error	*t* value
Kier & Hall index (order 0)	${}^0\chi_{K\&H}$	1.665	0.244	6.816
Kier & Hall index (order 3)	${}^3\chi_{K\&H}$	−1.873	0.336	−5.579
Kier shape index (order 3)	${}^3K$	−1.590	0.408	−3.896
ZX shadow	S_{ZX}	-7.429×10^{-2}	1.155×10^{-2}	−6.433
Max n–n repulsion for a C–S bond	${}^{Max}E_{nn,CS}$	0.277	0.048	5.808
Max resonance energy for a C–H bond	${}^{Max}E_{R,CH}$	−29.844	7.206	−4.142
Intercept	Constant	2.839×10^2	80.94	3.508
F value	39.12			

Table 4
The values of the 6 descriptors of anti-HIV-1 activity set.

Compound	$^0\chi_{K\&H}$	$^3\chi_{K\&H}$	^3K	S_{ZX}	$^{Max}E_{nn,CS}$	$^{Max}E_{R,CH}$
1	9.3681	3.2093	2.4457	51.5608	171.2395	11.1037
2	10.699	3.5766	2.7207	55.7409	171.1171	11.1165
3	10.699	3.4890	2.9183	59.0610	171.1629	11.1111
4	10.2908	3.7219	2.4953	55.5209	171.0864	11.1096
5	10.2908	3.4374	2.6905	56.8409	171.2547	11.1038
6	10.2908	3.4870	2.6905	56.6409	171.2701	11.1039
7	10.4247	3.8045	2.6160	54.5209	170.9948	11.103
8	10.4247	3.4790	2.8180	56.0409	171.1553	11.1016
10	11.2547	4.3168	2.6964	54.3809	171.1782	11.1073
11	11.2547	3.7372	2.9029	57.1609	171.1629	11.1017
12	9.6687	3.2439	2.6600	53.2408	171.1018	11.1032
14	10.238	3.4110	2.7151	58.1209	171.1094	11.1000
15	10.238	3.4495	2.7151	56.8209	171.1782	11.0964
16	10.9247	3.6092	3.3350	61.0210	171.0788	11.0997
17	9.8681	3.3059	2.6731	55.1009	171.2012	11.1036
19	11.2134	3.6248	2.9359	61.0610	171.2701	11.1039
21	11.3473	3.6634	3.0646	60.1610	171.1639	11.1012
22	11.6217	3.6897	3.1642	62.1610	171.1818	11.1020
23	12.2556	3.8727	3.8099	61.5810	170.9833	11.0958
24	11.1073	4.1152	2.5109	48.5007	169.27	11.1202
25	11.1073	4.0470	2.6734	52.0608	169.5008	11.1008
26	10.699	4.2605	2.2944	58.9610	169.3375	11.1124
27	10.699	3.9954	2.4532	53.8409	169.7284	11.1027
28	10.699	4.0451	2.4532	58.5010	169.9018	11.0977
29	11.663	4.8555	2.4750	55.0009	168.6426	11.1010
30	11.663	4.2952	2.6429	49.7608	169.4120	11.0975
31	11.663	4.3664	2.6429	61.3410	169.4394	11.0918
32	10.6462	4.0724	2.3368	56.2209	168.3881	11.0949
33	10.6462	3.9690	2.4908	52.0008	169.2697	11.0962
34	11.3329	4.1672	3.0807	54.2809	169.2091	11.0991
35	11.6217	4.1828	2.6892	59.7010	169.7055	11.0790
36	11.8894	4.6166	2.7488	58.3410	168.449	11.0832
37	11.7556	4.2215	2.8049	56.4809	169.3825	11.0786
38	12.6638	4.4307	3.5430	65.4411	169.0305	11.0754
39	10.1846	4.3253	2.1023	49.3407	172.8206	11.1061
40	11.5155	4.6539	2.3991	51.6608	172.1076	11.1011
41	11.5155	4.6051	2.5385	54.2209	172.5112	11.0901
42	11.5155	4.6415	2.5385	66.6811	173.4627	11.1127
43	11.1073	4.7992	2.1912	50.1408	172.4901	11.1087
44	11.1073	4.5534	2.3264	50.6608	172.8221	11.0986
45	11.1073	4.6031	2.3264	61.981	173.0502	11.096
46	11.2411	4.8818	2.2907	49.7408	171.405	11.0975
47	11.2411	4.5950	2.4306	54.6009	172.2887	11.0911
48	11.2411	4.6477	2.4306	66.1611	172.55	11.0959
49	12.0712	5.3942	2.357	50.9808	171.0202	11.0962
50	12.0712	4.8532	2.4999	56.6209	172.3124	11.0916
51	12.0712	4.9244	2.4999	69.1212	172.3484	11.0898
52	10.4852	4.4153	2.1675	46.3207	171.6656	11.1014
53	10.4852	4.3599	2.3016	49.4408	172.0793	11.0858
54	11.0545	4.6111	2.2385	47.8807	170.9951	11.0900
55	11.0545	4.5270	2.3708	55.5009	172.0931	11.0906
56	11.0545	4.5655	2.3708	68.4012	171.9927	11.0977
57	11.7411	4.7252	2.934	60.3410	171.9706	11.0929
58	12.2977	5.1553	2.6185	57.9209	170.8647	11.0773
59	12.2977	4.8181	2.7674	46.8207	171.7744	11.0761
60	12.0299	4.7408	2.5530	50.0008	172.8274	11.0788
61	12.9939	5.0191	2.7299	54.3809	172.3269	11.0767
62	12.1638	4.7795	2.6593	49.8808	172.3096	11.0774
63	12.4382	4.8057	2.7670	49.6408	172.5209	11.0791
64	13.0720	4.9887	3.3827	55.0809	171.6603	11.0775
65	11.4771	4.5812	2.5385	46.9607	172.4026	11.0825
66	13.1453	4.8810	3.1367	60.8210	172.5476	11.0791
67	13.8524	5.0429	3.5471	68.3612	172.5469	11.0792
68	14.5595	5.3366	3.9572	75.4413	172.5500	11.0791

2.5. External predictive ability of the models

Some authors have attracted attention to the predictive ability of models. For instance, Golbraikh and Tropsha [35] showed that high values for the LOO cross-validated q^2 is a necessary but not sufficient condition to ensure high predictive ability. More recently, Roy and Roy [36] defined the R^2_m parameter:

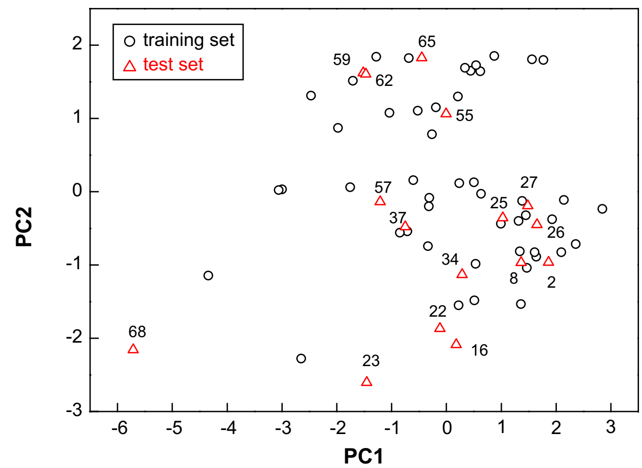


Fig. 3. Scatter plot of HIV-1 RT binding affinity set compounds on principal components' plane.

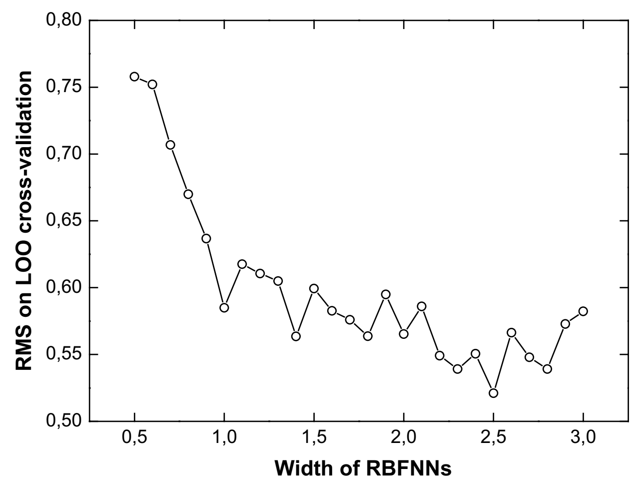


Fig. 4. RMS versus width of RBFNN on LOO cross-validation.

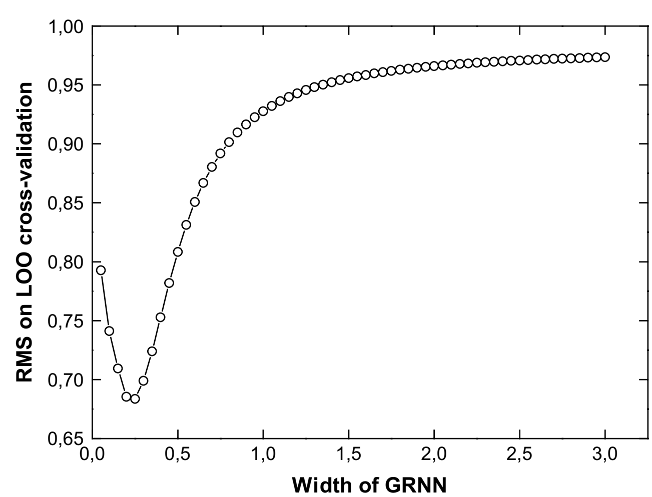


Fig. 5. RMS versus width of GRNN on LOO cross-validation.

Table 5
Results of the QSAR models for anti-HIV-1 activity set based on MLR, MARS, RBFNN, GRNN, PPR and SVM.

Data set	R^2						MSE					
	MLR	MARS	RBFNN	GRNN	PPR	SVM	MLR	MARS	RBFNN	GRNN	PPR	SVM
Training set	0.793	0.730	0.791	0.814	0.890	0.831	0.19	0.25	0.19	0.18	0.10	0.16
Test set	0.840	0.478	0.833	0.686	0.882	0.850	0.18	0.48	0.18	0.32	0.15	0.21

$$R_m^2 = R^2\left(1 - \sqrt{\left|R^2 - R_0^2\right|}\right) \quad (2)$$

where R^2 and R_0^2 between the observed and predicted values are calculated from the test set with and without intercept, respectively. This parameter highlights the fact that R^2 and R_0^2 should not be significantly different, and an R_m^2 value greater than 0.5 is an indicator of good external predictability.

In order to test our best correlations with this new approach, we calculated R^2, R_0^2, and R_m^2 for the PPR and SVM models. Results are presented in Table 9.

They confirm the good quality of these models and moreover indicate good external predictability, especially for the anti-HIV activity PPR model.

3. Conclusions

Two data sets with anti-HIV-1 activity and HIV-1 RT binding affinity of a series of 2-amino-6-arylsulfonylbenzonitriles and congeners were analyzed by QSAR studies. The descriptors, which were calculated and selected by CODESSA, involved topological descriptors, geometrical descriptors and quantum-chemical descriptors. Regression models were built with six linear and

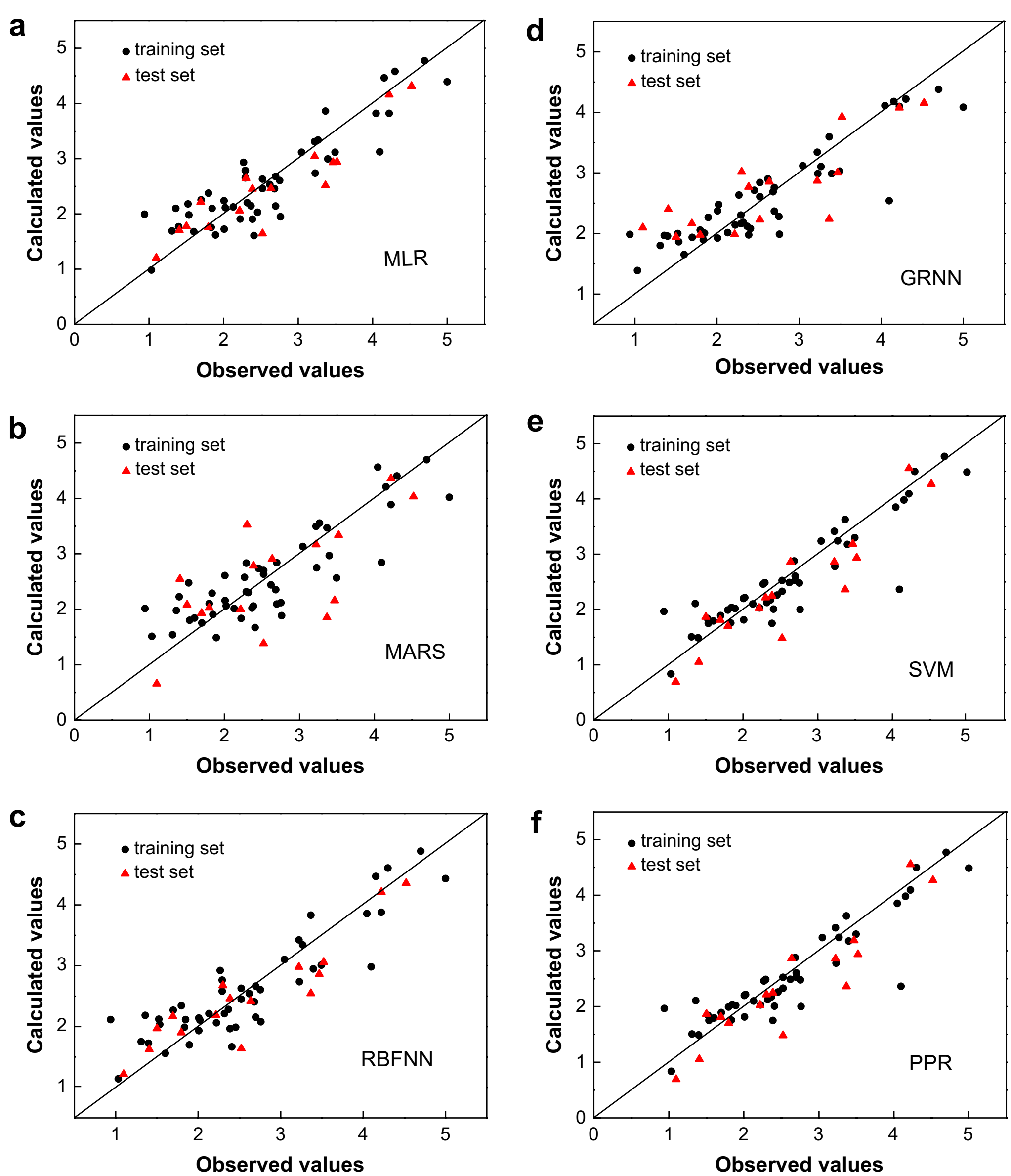


Fig. 6. Calculated values versus observed values of activity using MLR (a), MARS (b), RBFNN (c), GRNN (d), SVM (e) and PPR (f) modeling for anti-HIV-1 activity data set. The diagonal in the six plots is the $y = x$ line.

Table 6
Descriptors, coefficients, standard error, and *t*-values for the linear model of HIV-1 RT binding affinity set.

Physico-chemical meaning	Descriptors	Coefficient	Standard error	*t* value
Max resonance energy for a C–H bond	$^{Max}E_{R,CH}$	−62.67	8.345	−7.510
Max exchange energy for a C–H bond	$^{Max}E_{exc,CH}$	7.596	4.141	1.834
Max e–n attraction for a C–N bond	$^{Max}E_{ne,CN}$	−4.098	0.860	−4.767
ZX shadow/*ZX* rectangle	$S_{ZX,r}$	−5.246	1.658	−3.164
Min electroph. react. index for a N atom	$^{Min}N_N$	3.329×10^3	1.091×10^3	3.053
Intercept	Constant	2.102×10^3	3.650×10^2	5.760
F value	26.11			

nonlinear approaches: MLR, MARS, RBFNN, GRNN, PPR and SVM. PPR and SVM models showed their powerful capacity to predict both anti-HIV-1 activity and HIV-1 RT binding affinity of 2-amino-6-arylsulfonylbenzonitriles and congeners.

Most of the descriptors used herein relate to global properties of molecules, and cannot be directly used as a guide to the synthesis of new molecules. However, these descriptors, with the PPR and SVM techniques, due to their good predictive ability, could be used to assess the activity and affinity towards HIV-1 RT of new molecules in the three series of compounds we studied.

Table 7
The values of the 5 descriptors of HIV-1 RT binding affinity set.

Compound	$^{Max}E_{R,CH}$	$^{Max}E_{exc,CH}$	$^{Max}E_{ne,CN}$	$S_{ZX,r}$	$^{Min}N_N$
1	11.1037	5.3206	351.7096	0.6773	2.92E−04
2	11.1165	5.3186	351.7582	0.7706	3.28E−04
3	11.1111	5.3163	351.6990	0.6472	3.11E−04
5	11.1038	5.3498	351.7145	0.6312	3.19E−04
6	11.1039	5.3506	351.7205	0.6936	2.36E−04
7	11.1030	5.3145	351.6395	0.7451	1.56E−04
8	11.1016	5.3105	351.6828	0.6940	1.36E−04
9	11.1016	5.3105	351.6872	0.6914	9.46E−05
11	11.1017	5.3105	351.6856	0.6760	1.04E−04
12	11.1032	5.3099	351.6816	0.7223	1.47E−04
13	11.1031	5.3114	351.6225	0.7350	3.47E−05
14	11.1000	5.3094	351.6708	0.6616	2.61E−05
16	11.0997	5.3090	351.6731	0.6431	3.77E−05
18	11.0957	5.3077	351.6272	0.6284	8.27E−05
19	11.1039	5.3507	351.7177	0.6264	3.53E−04
20	11.0980	5.3086	351.6753	0.6639	6.82E−05
21	11.1012	5.3479	351.6865	0.6462	1.60E−04
22	11.1020	5.3622	351.7008	0.6491	3.41E−04
23	11.0958	5.3167	351.6698	0.6735	5.81E−05
24	11.1202	5.3195	351.8844	0.5919	2.98E−04
25	11.1008	5.3161	352.1721	0.5902	4.69E−04
27	11.1027	5.3520	352.1390	0.6153	4.46E−04
30	11.0975	5.3075	352.1378	0.5674	3.77E−04
32	11.0949	5.3098	351.9581	0.6578	6.61E−05
35	11.0790	5.3533	352.1343	0.6419	4.47E−04
36	11.0832	5.3050	352.0021	0.6521	1.75E−04
37	11.0786	5.3509	352.1332	0.5819	3.87E−04
38	11.0754	5.3167	352.2012	0.6695	3.82E−04
39	11.1061	5.3145	351.6475	0.5826	3.85E−05
40	11.1011	5.3318	351.6196	0.6237	3.75E−05
41	11.0901	5.3160	351.6164	0.5946	3.41E−05
42	11.1127	5.3184	351.6624	0.6448	4.82E−05
43	11.1087	5.3575	351.6723	0.5865	5.07E−05
44	11.0986	5.3633	351.6450	0.5533	3.92E−05
45	11.0960	5.3448	351.6399	0.6409	3.88E−05
46	11.0975	5.3054	351.7221	0.6015	4.39E−05
47	11.0911	5.3003	351.6533	0.5975	2.95E−05
49	11.0962	5.3050	351.7664	0.6229	4.93E−05
50	11.0916	5.3000	351.6596	0.6126	2.91E−05
52	11.1014	5.3041	351.6447	0.5575	3.10E−05
54	11.0900	5.304	351.7529	0.5783	1.26E−05
55	11.0906	5.2987	351.6770	0.5905	2.11E−05
57	11.0929	5.2982	351.6831	0.6325	1.96E−05
58	11.0773	5.3000	351.7159	0.6469	2.47E−05
59	11.0761	5.2994	351.6371	0.5359	1.69E−05
60	11.0788	5.3556	351.6387	0.5553	4.15E−05
61	11.0767	5.3519	351.6591	0.5632	3.20E−05
62	11.0774	5.3551	351.6522	0.5489	3.23E−05
63	11.0791	5.3639	351.6201	0.5665	3.66E−05
64	11.0775	5.3166	351.6107	0.5869	1.28E−05
68	11.0791	5.3686	351.6285	0.6327	3.71E−05

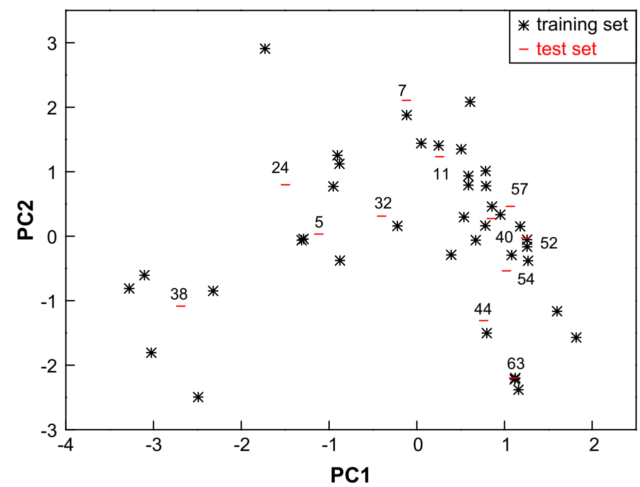


Fig. 7. Scatter plot of HIV-1 RT binding affinity set compounds on principal components' plane.

4. Methods

4.1. Descriptors calculation and selection

To encode the features of molecules with molecular descriptors is an important step to obtain a QSAR model. The descriptors used as independent variables in QSAR modeling were calculated with the CODESSA software, on the basis of the minimum energy

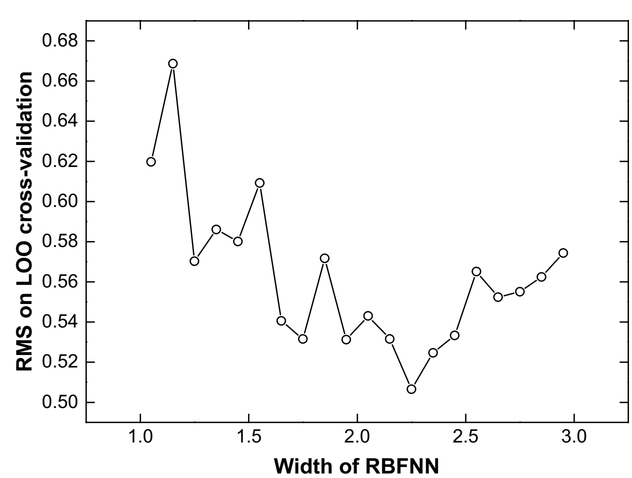


Fig. 8. RMS versus width of RBFNN on LOO cross-validation.

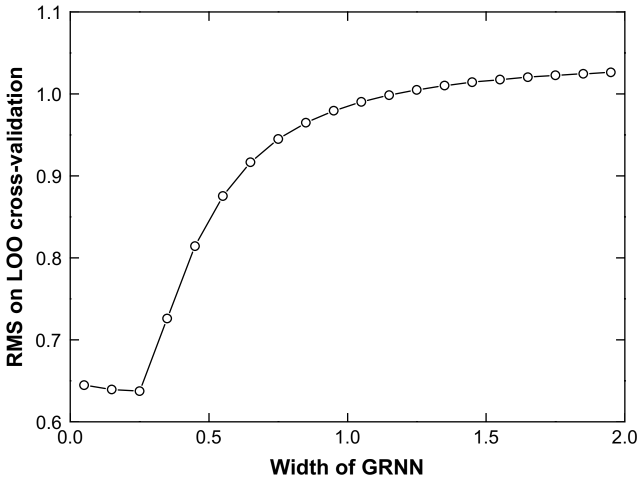


Fig. 9. RMS versus width of GRNN on LOO cross-validation.

molecular geometries optimized by the Hyperchem Program (Version 7.0) [37] and MOPAC software [38] based on AM1 [39] semi-empirical method.

CODESSA [40] is a comprehensive program for developing quantitative structure–activity/property relationships (QSAR/QSPR) by integrating all necessary mathematical and computational tools. Its function is to (i) calculate a large variety of molecular descriptors on the basis of the 3D geometrical structure and/or quantum-chemical wave functions of chemical compounds; (ii) develop (multi)linear and nonlinear QSPR models on the chemical and physical properties or biological activity of chemical compounds; (iii) carry out cluster analysis of the experimental data and molecular descriptors; (iv) give tools for interpreting the developed models; (v) predict property values for any chemical compound with known molecular structure. The CODESSA software produces more than 500 constitutional, topological, geometrical, electrostatic, quantum-chemical and thermodynamical molecular descriptors [41] and performs the statistical analyses in the descriptor space. It provides various methods for statistical analysis of experimental data such as linear (e.g. Multiple Linear Regression, Best Multiple Linear Regression, and Heuristic Method) and nonlinear regression (e.g. nonlinear iterative partial least squares (NIPALS)).

After the calculation of the molecular descriptors, we used the heuristic method (HM) in CODESSA to accomplish the pre-selection of the descriptors and build the linear model. Its advantages are the high speed and no software restrictions on the size of the data set. The heuristic method can either quickly give a good estimation about what quality of correlation to expect from the data, or derive several best regression models. Besides, it shows which descriptors have bad or missing values, which ones are insignificant (from the standpoint of a single-parameter correlation), and which ones are highly intercorrelated. This information is helpful in reducing the number of descriptors involved in the search for the best QSAR/QSPR model.

First of all, all descriptors are checked to ensure: (a) that values of each descriptor are available for each structure and (b) that there is a variation in these values. Descriptors for which values are not available for every structure in the data in question are discarded. Descriptors having a constant value for all structures in the data set are also discarded. Thereafter all possible one-parameter regression models are tested and insignificant descriptors removed. As a next step, the program calculates the pair correlation matrix of descriptors and further reduces the descriptor pool by eliminating highly correlated descriptors. All two-parameter regression models with remaining descriptors are subsequently developed and ranked by the regression correlation coefficient R^2. A stepwise addition of further descriptor scales is performed to find the best multi-parameter regression models with the optimum values of statistical criteria (highest values of R^2, the cross-validated R^2_{cv}, and the F values).

A simple "break point" technique was used to control the model expansion in the improvement of the statistical quality of the model, by analyzing the plot of the number of descriptors involved in the obtained models versus squared correlation coefficient values corresponding to those models (see Fig. 2). Frequently, the statistical improvement of the regression model is less significant ($\Delta R^2 < 0.02$–0.04) beyond a certain number of independent variables in the model ('breaking point'). Consequently, the model corresponding to the breaking point is considered the best/optimum model.

4.2. Descriptors and their physicochemical meanings

Although five different types of descriptors are provided by the CODESSA software, only three types have been selected to build our models. Our results show that they have strong relationship with bioactivity in the series of molecules we studied. These are topological, geometrical and quantum-chemical descriptors. Precise definitions can be found in Ref. [40].

Kier and Hall descriptors [15,42] are related to molecular connectivity. These indices quantify molecular structure, encoding structural features such as size, branching, unsaturation, heteroatom content and cyclicity. There are four orders of indices (0, 1, 2, and 3) related to atomic valence connectivity, one bond path valence connectivity, two bond fragment valence connectivity and three contiguous bond fragment valence connectivity, respectively. Molecular shape descriptors are also called Kappa shape indices, they are derived from the number of paths in the molecular skeleton and the number of atoms. We used ${}^0\chi_{K\&H}$, ${}^3\chi_{K\&H}$ and ${}^3K$ (Table 3).

Geometrical descriptors reflect the size and geometrical shape of the molecule. Different from topological descriptors, they are related to three-dimensional (3D) molecular structures and 3D coordinates are required to calculate them. In QSAR, geometrical descriptors, especially those related to molecular volume and molecular surface area are important. Shadow indices [43] are a series of molecular surface area projections. Molecular surfaces are projected on the XY, YZ and XZ planes to obtain shadow areas and relative shadow areas of a molecule. The shape of molecules is an important factor in their interaction with proteins. S_{Zx} revealed relevant in this work (Tables 3 and 6).

Quantum-chemical descriptors [44] are related to atomic charges, the highest occupied molecular orbital (HOMO), the lowest unoccupied molecular orbital (LUMO), orbital electron density and molecular polarizability. Therefore, quantum-chemical

Table 8
Results of the QSAR models for HIV-1 RT binding affinity set based on MLR, MARS, RBFNN, GRNN, PPR and SVM.

Data set	R^2						MSE					
	MLR	MARS	RBFNN	GRNN	PPR	SVM	MLR	MARS	RBFNN	GRNN	PPR	SVM
Training set	0.738	0.707	0.825	0.808	0.843	0.811	0.27	0.30	0.18	0.23	0.16	0.20
Test set	0.750	0.345	0.651	0.694	0.843	0.802	0.22	0.59	0.29	0.29	0.15	0.16

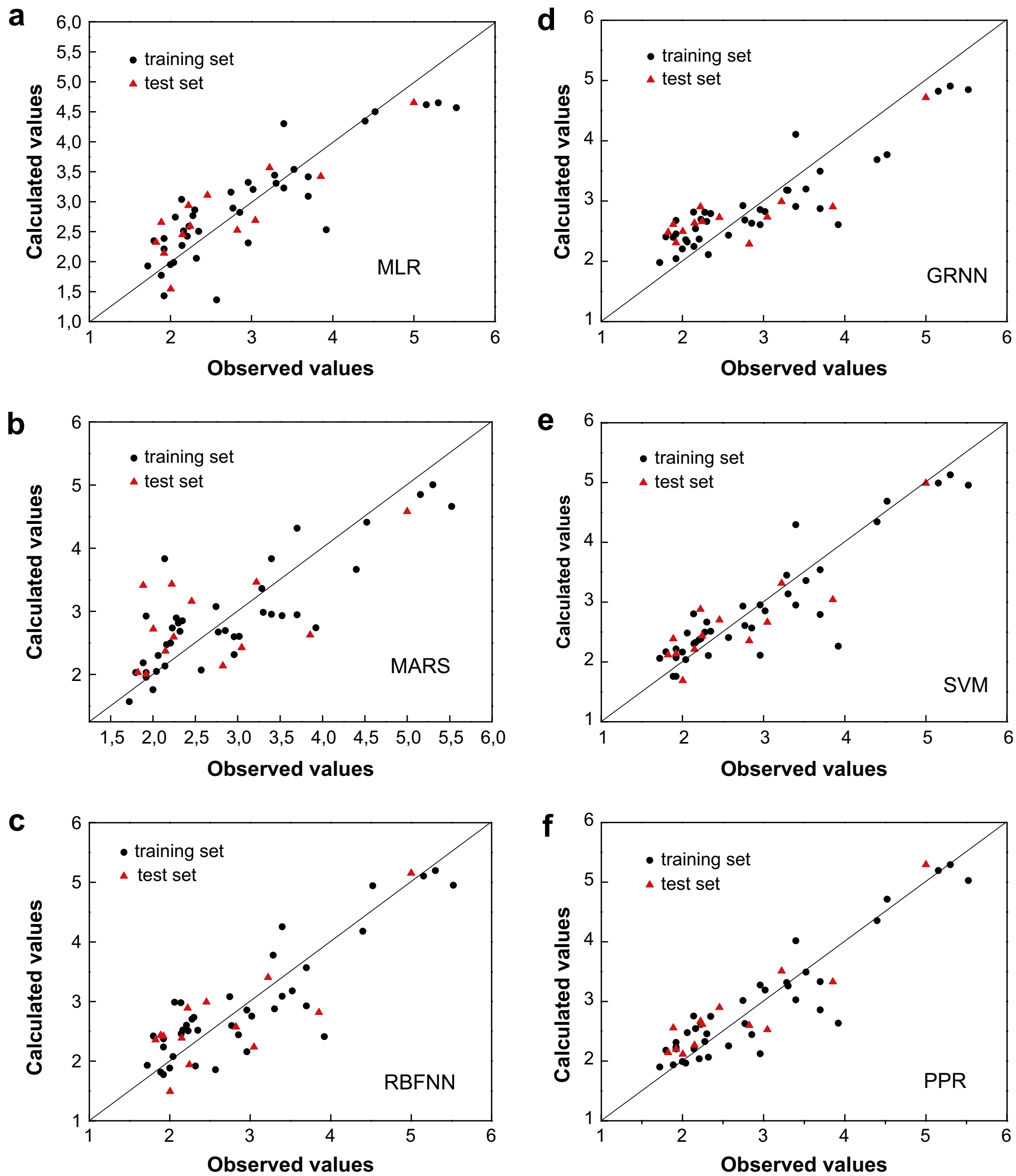


Fig. 10. Calculated values versus observed values of affinity using MLR (a), MARS (b), RBFNN (c), GRNN (d), SVM (e) and PPR (f) modeling for HIV-1 RT binding affinity data set. The diagonal in the six plots is the $y = x$ line.

descriptors can be classified into three main categories: (i) charge distribution-related descriptors; (ii) valency-related descriptors; (iii) quantum mechanical energy-related descriptors.

Most of the quantum descriptors that were used in this work: ${}^{Max}E_{nn,CS}$, ${}^{Max}E_{R,CH}$, ${}^{Max}E_{exc,CH}$ and ${}^{Max}E_{ne,CN}$ (Tables 3 and 6) belong to quantum mechanical energy-related descriptors, which characterize the total energy of the molecule and the intramolecular energy distribution using different partitioning schemes. Maximum nuclear repulsion energy between two given atoms describes the nuclear repulsion in the molecule and may be related to the conformational (rotational, inversional) changes or atomic reactivity in the molecule. Maximum electronic exchange energy between two given atoms reflects the change in the Fermi correlation energy between two electrons localized on atoms A and B, respectively. It is important in determining the conformational changes of the molecule and its spin properties. Maximum nuclear–electron attraction energy between two given atoms (in our case, ${}^{Max}E_{ne,CN}$, Table 6) and maximum resonance energy between two given atoms (${}^{Max}E_{R,CH}$, Tables 3 and 6) are also energy-related. ${}^{Min}N_N$ (Table 6) related to LUMO energy and electrophilic reactivity was also a relevant parameter.

Table 9
External predictability of the SVM and PPR models.

	R^2	R_0^2	R_m^2
Anti-HIV-1 activity/PPR	0.882	0.946	0.660
Anti-HIV-1 activity/SVM	0.850	0.969	0.555
HIV-1 RT binding affinity/PPR	0.843	0.967	0.548
HIV-1 RT binding affinity/SVM	0.802	0.999	0.446

4.3. Multiple linear regression (MLR)

Multiple linear regression fits a linear model of the form:

$$Y = b_0 + b_1X_1 + b_2X_2 + \ldots + b_kX_k + e \quad (3)$$

where Y is the dependent variable (response) and $X_1, X_2, \ldots, X_k$ are the independent variables (predictors) and e is random error. $b_0, b_1, b_2, \ldots, b_k$ are known as the regression coefficients, which have to be estimated from the data. The MLR algorithm chooses regression coefficients so as to minimize the squared sum of the difference between predicted values and measured values. MLR is performed either to study the relationship between the response variable and predictor variables or to predict the response variable based on the predictor variables.

4.4. Multivariate adaptive regression splines (MARS)

Multivariate adaptive regression splines was explored by Friedman and other workers [45,26]. This is an adaptive regression procedure well suited to problems with a large number of predictor variables. MARS is a generalization of stepwise linear regression, but the regression is fitted using a series of basis functions. The basis functions consist of one single spline function or two (or more) functions, for example, $b^-{}_q(x-t)$ and $b^+{}_q(x-t)$ with

$$b_q^-(x-t) = [-(x-t)]_q^+ = \begin{cases} (t-x)^q, & \text{if } x < t \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

$$b_q^+(x-t) = [+(x-t)]_q^+ = \begin{cases} (x-t)^q, & \text{if } x < t \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

where $b^-{}_q(x-t)$ and $b^+{}_q(x-t)$ are describing, respectively, the regions right and left of the knot location t and q the power to which the spline is raised. The superscript "+" indicates a value of "0" for negative values of the argument. Then the different base functions are combined in one multidimensional model, which describes the response as a function of the explanatory variables. The result is a complex nonlinear model of the form:

$$\hat{y} = a_0 + \sum_{m=1}^{M} a_m B_m(x) \tag{6}$$

where $\hat{y}$ is the predicted value for the response variable; a_0 is the coefficient of the constant base function; M is the number of base functions and B_m and a_m is the mth base function and its coefficient.

Generally, three steps are used in a MARS analysis. The first one consists of a stepwise procedure in which a global model is built that usually overfits the training data. During each iteration the best pair of basis functions is selected in order to improve the model. All possible predictors and knot locations (for each predictor) are evaluated. At the end of each iteration, so-called interactions may also be introduced if this improves the model. This building process continues until a user-defined maximum number of basis functions ($M_{\max}$) is reached.

4.5. Radial basis function neural networks (RBFNN)

A radial basis function neural network is an artificial neural network which uses radial basis functions as activation functions. It consists of an input layer, a hidden layer and an output layer. Each layer is fully connected to the following one and the hidden layer is composed of a number of nodes with radial activation functions called radial basis functions. The input layer does not process the information; it only distributes the input vectors to the hidden layer. The hidden layer of RBFNN consists of a number of RBF units (n_h) and bias (b_k). Each hidden layer unit represents a single radial basis function, with associated center position and width. Each neuron on the hidden layer employs a radial basis function as a nonlinear transfer function to operate on the input data. The most often used RBF is a Gaussian function that is characterized by a center (c_j) and a width (r_j). In this study, the Gaussian was selected as a radial basis function. RBF operates by measuring the Euclidian distance between input vector (x_i) of pattern I and the radial basis function centre (c_j), and performs a nonlinear transformation according to the formula $h(x_i) = \exp[-(x_i - c_j)^2/r_j^2]$, where h_j is the output of hidden unit j. The operation of the output layer is linear:

$$y_k(x) = \sum_{j=1}^{n_k} w_{kj} h_j(x) + b_k \tag{7}$$

where y_k is the kth output unit for the input vector x, w_{kj} is the weight connection between the kth output unit and the jth hidden layer unit, and b_k is the bias. The training procedure when using RBF involves selecting centers, width and weights. In this paper, the forward subset selection routine was used to select the centers from training set samples. The adjustment of the connection weight between the hidden layer and the output layer was performed using a least-squares solution after the selection of centers and width of radial basis functions. The final layer is only a linear weighted output.

4.6. General regression neural networks (GRNN)

A general regression neural network which is designed for regression, was selected instead of back-propagation (BP) neural networks. It was introduced by Specht in 1991 [46]. GRNN is a nonparametric estimator that calculates a weighted average of the target values of training patterns by the probability density function using Parzen's nonparametric estimator. For GRNN, the predicted value is the most probable value $E(y|x)$:

$$E(y|x) = \hat{y}(x) = \frac{\int_{-\infty}^{+\infty} y f(x,y) \mathrm{d}y}{\int_{-\infty}^{+\infty} f(x,y) \mathrm{d}y} \tag{8}$$

where $f(x,y)$ is the probability density function. This can be estimated from the training set by using the Parzen's nonparametric estimator [47]:

$$f(x,y) = \frac{1}{n\sigma} \sum_{i=1}^{n} W \frac{(x - x_i)}{\sigma} \tag{9}$$

where n is the sample size, σ is a scaling parameter that defines the width of the bell curve that surrounds each sample point, $W(d)$ is a weighting function that has its largest value at $d = 0$, and $(x - x_i)$ is the distance between the unknown sample and a data point. The Gaussian function is frequently used as the weighting function because it is well behaved, easily calculated, and satisfies the conditions required by Parzen's estimator. Substituting Parzen's nonparametric estimator for $f(x,y)$ and performing the integrations leads to the fundamental equation of GRNN.

$$\hat{y}(x) = \frac{\sum_{i=1}^{n} y_i \exp(-D(x, x_i))}{\sum_{i=1}^{n} \exp(-D(x, x_i))} \tag{10}$$

where

$$D(x, x_i) = \sum_{j=1}^{p} \left(\frac{x_j - x_{ij}}{\sigma_j} \right)^2 \tag{11}$$

GRNN consists of 4 layers: input, hidden, summation, and output layers. The input layer provides input values to all neurons in the hidden layer and has as many neurons as the number of descriptors in the training set. The number of hidden neurons is determined by the total number of compounds in the training set. The summation layer has two neurons, which, respectively, calculate the numerator and the denominator of Eq. (10). The output layer has the single neuron performing a division of the outputs of the two summation neurons to obtain the predicted response.

4.7. Projection pursuit regression (PPR)

Projection pursuit regression is one of nonparametric methods. It was developed in 1981 by Friedman and Stuezle [48]. PPR models the response variable by a linear combination of predictor functions g_i and reduces a p-dimensional problem to at most p one-dimensional nonparametric subproblems. For many practical problems, the data is usually high dimensional. It has been a common practice to use lower dimensional linear projections of the data for visual inspection. The lower dimension is usually 1 or 2 (or may be 3). More precisely, if $X1,\ldots, Xn$, $X \in IR^p$ are p-dimensional data, then a k-dimensional ($k < p$) linear projection is $Z_1,\ldots, Z_n$, $Z \in IR^k$ where $Z_i = \alpha^T X_i$ for some $p \times k$ matrix α such that $\alpha^T \alpha = I_k$, the k-dimensional identity matrix. Such a matrix α is called orthonormal. α^T is the transpose matrix of α. Since there are infinitely many projections from a higher dimension to a lower dimension, it is important to have a technique to pursue a finite sequence of projections that can reveal the most interesting structures of the data. Friedman and Turkey successfully implemented the idea combining both projection and pursuit, which is called projection pursuit.

In a typical regression problem, (X, Y) is an observable pair of random variables from a distribution F, where $X \in IR^p$ is a p-dimensional variable (called predictor) and $Y \in IR$ is a response; and the goal is to estimate the regression function.

$$f(x) = E(Y|X = x) \tag{12}$$

i.e. the conditional expectation of Y given $X = x$, using a random sample $(X_1, Y_1),\ldots, (X_n, Y_n)$ from F. PPR approximates the regression function $f(x)$ by a finite sum of ridge functions

$$g^{(m)}(x) = \sum_{i=1}^{m} g_i\left(\alpha_i^T x\right) \tag{13}$$

where α_i are $p \times k$ orthonormal matrices, m is the number of ridge functions. PPR model can be used to approximate a large class of function by suitable choices of α_i and g_i.

In 1985 Friedman [49] presented a more efficient algorithm, suitable for multiple response regression and classification. This study is using this method to construct a PPR model. In this algorithm, g_i are found by smoothing operation that entails a backfitting. Specially, given $g^{(0)} = 0$, for $i \geq 1$, it iteratively estimates α_i by maximum of an index and g_i by a low dimensional nonparametric regression estimate based on the projected data (z_j, r_j), where $r_j = Y_j - g^{(i-1)}(X_j)$ are the residuals at the ith step and $z_i = \alpha^T{}_i X_j$, $j = 1,\ldots, n$. The procedure is repeated forward (and perhaps a backward fitting is allowed to adjust for the previous fitted pair) until the residual sum of squares $\sum \gamma_j^2$ is less than a predetermined value. A different smoother for g_i, or index, or fitting order may be used and hence yields a different PPR algorithm. In this work, we smoothed the (x, y) values with Friedman's "super smoother". Details of Friedman's "super smoother" are found in R Documentation [50].

4.8. Support vector machines (SVM)

Support vector machines were developed by Vapnik [51]. A version of a SVM for regression was proposed in 1997 by Vapnik et al. [52]. The theory of support vector regression (SVR) has been extensively described [53,54]. Thus, a brief description is given here. SVM is based on the structure risk minimization (SRM) principle, which has been shown to be superior to the traditional empirical risk minimization (ERM) principle [55]. In SVM, the basic idea is to map the data x into a higher dimensional feature space with a kernel function, $K(x_i, x_j)$. Then linear regression is conducted in this space. The prediction or approximation function is:

$$f(x) = \sum_{i-1}^{l}\left(\alpha_i^* - \alpha_i\right)K(x, x_i) + b \tag{14}$$

where l is the number of support vectors, b is bias, α_i and $\alpha^*{}_i$ are the introduced Lagrange multipliers that are determined by maximizing the following form of function:

$$\Phi\left(\alpha_i, \alpha_i^*\right) = \sum_{i=1}^{l} y_i\left(\alpha_i + \alpha_i^*\right) - \varepsilon \sum_{i=1}^{l}\left(\alpha_i + \alpha_i^*\right) - \frac{1}{2}\sum_{i,j=1}^{l}\left(\alpha_i + \alpha_i^*\right) \times \left(\alpha_j + \alpha_j^*\right) \times \left(\alpha_i - \alpha_j\right) \tag{15}$$

subject to:

$$\sum_{i=1}^{l}\left(\alpha_i + \alpha_i^*\right) = 0 \tag{16}$$

$$0 \leq \alpha_i \leq C \tag{17}$$

$$0 \leq \alpha_j \leq C \tag{18}$$

where l is the training set size and C is a penalty for training errors.

In Eq. (14), the value of the kernel function $K(x, x_i)$ is equal to the inner product of two vectors x and x_i in the feature space $\Phi(x)$ and $\Phi(x_i)$. That is, $K(x, x_i) = \Phi(x)\Phi(x_i)$. The elegance of using kernel function lies in the fact that one can deal with feature spaces of arbitrary dimensionality without having to compute the map $\Phi(x)$ explicitly. Any function that satisfies Mercer's condition can be used as the kernel function. In SVR, the Gaussian kernel $K(u,v) = \exp(-|u - v|^2/\delta^2)$ is commonly used.

MLR, MARS, PPR, and SVM were performed using the R 2.1.1 statistical software [50]. RBFNN and GRNN were implemented by M-file based on MATLAB script for RBFNN [56,57].

Acknowledgement

We are indebted to the National Natural Science Foundation of China for support through grant 90612016 and 60773108, and to the Ministry of Science and Technology of China for support through grant 2005DKA64001. We are also indebted to the China Research Council for financial support to Miss HU.

References

[1] J.L. Marx, Science 217 (1982) 618–621.
[2] <http://www.unaids.org/en/HIV_data>.
[3] <http://www.who.int/hiv/mediacentre/news64/>.
[4] J.D. Reeves, R.W. Doms, J. Gen. Virol. 83 (2002) 1253–1265.
[5] E. De Clercq, Med. Res. Rev. 22 (2002) 531–565.
[6] R.A. Katz, M.A. Skalka, Annu. Rev. Biochem. 63 (1994) 133–173.
[7] R. Esnouf, J. Ren, C. Ross, Y. Jones, D. Stammers, D. Stuart, Nat. Struct. Biol. 2 (1995) 303–308.
[8] E. De Clercq, Antiviral Res. 38 (1998) 153–179.
[9] J.H. Chan, J.S. Hong, R.N. Hunter III, G.F. Orr, J.R. Cowan, D.B. Sherman, S.M. Sparks, B.E. Reitter, C.W. Andrews III, R.J. Hazen, M.St. Clair, L.R. Boone, R.G. Ferris, K.L. Creech, G.B. Roberts, S.A. Short, K. Weaver, R.J. Ott, J. Ren, A.H.D. Stuart, D.K. Stammers, J. Med. Chem. 44 (2001) 1866–1882.
[10] C. Hansch, T. Fujita, J. Am. Chem. Soc. 86 (1963) 1616–1626.
[11] H. Kubinyi (Ed.), QSAR: Hansch Analysis and Related Approaches, VCH, Weinheim, 1993.
[12] C. Hansch, A. Leo (Eds.), Exploring QSAR, American Chemical Society, Washington DC, 1995.
[13] S.M. Free, J.W. Wilson, J. Med. Chem. 7 (1964) 395–399.

[14] M. Randic, J. Am. Chem. Soc. 97 (1975) 6609–6615.
[15] L.B. Kier, L.H. Hall (Eds.), Molecular Connectivity in Structure–Activity Analysis Research, Letchworth, England, 1986.
[16] G.M. Crippen, J. Med. Chem. 22 (1979) 988–997.
[17] R.D. Cramer, D.E. Patterson, J.D. Bunce, J. Am. Chem. Soc. 110 (1988) 5959–5967.
[18] G. Klebe, U. Abraham, T. Mietzner, J. Med. Chem. 37 (1994) 4130–4146.
[19] T. Aoyama, Y. Suzuki, H. Ichikawa, J. Med. Chem. 33 (1990) 905–908.
[20] T. Aoyama, Y. Suzuki, H. Ichikawa, J. Med. Chem. 33 (1990) 2583–2590.
[21] X.J. Yao, A. Panaye, J.P. Doucet, R.S. Zhang, H.F. Chen, M.C. Liu, Z.D. Hu, B.T. Fan, J. Chem. Inf. Comput. Sci. 44 (2004) 1257–1266.
[22] A. Szczurek, M. Maciejewska, Talanta 64 (2004) 609–617.
[23] T. Niwa, J. Chem. Inf. Comput. Sci. 43 (2003) 113–119.
[24] P.D. Mosier, P.C. Jurs, J. Chem. Inf. Comput. Sci. 42 (2002) 1460–1470.
[25] R.D. De Veaux, D.C. Psichogios, L.H. Ungar, Comput. Chem. Eng. 17 (1993) 819–837.
[26] Q.S. Xu, M. Daszykowskia, B. Walczaka, F. Daeyaert, M.R. De Jonge, J. Heeres, L.M.H. Koymans, P.J. Lewi, H.M. Vinkers, P.A. Janssen, D.L. Massart, Chemom. Intell. Lab. Syst. 72 (2004) 27–34.
[27] V. Nguyen-Cong, B.M. Rode, Eur. J. Med. Chem. 31 (1996) 479–484.
[28] H.X. Liu, R.S. Zhang, X.J. Yao, M.C. Liu, Z.D. Hu, B.T. Fan, J. Chem. Inf. Comput. Sci. 43 (2003) 1288–1296.
[29] H.X. Liu, R.J. Hu, R.S. Zhang, X.J. Yao, M.C. Liu, Z.D. Hu, B.T. Fan, J. Comput. Aided Mol. Des. 19 (2005) 33–46.
[30] K. Roy, J.T. Leonard, Bioorg. Med. Chem. 12 (2004) 745–754.
[31] J.T. Leonard, K. Roy, QSAR Comb. Sci. 23 (2004) 23–35.
[32] M.P. Freitas, Org. Biomol. Chem. 4 (2006) 1154–1159.
[33] L.J. Tang, Y.P. Zhou, J.H. Jiang, H.Y. Zou, H.L. Wu, G.L. Shen, R.Q. Yu, J. Chem. Inf. Model. 47 (2007) 1438–1445.
[34] Cerius 2, Version 3.5, Molecular Simulations Inc., San Diego, USA, 1999.
[35] A. Golbraikh, A. Tropsha, J. Mol. Graph. Modell. 20 (2002) 269–276.
[36] P.P. Roy, K. Roy, QSAR Comb. Sci. 27 (2008) 302–313.
[37] HyperChem, Release 7.0 for Windows, Hypercube, Inc., 2002.
[38] J.P.P. Stewart, MOPAC 6.0, Quantum Chemistry Program Exchange, QCPE, No. 455, Indiana University, Bloomington, IN, 1989.
[39] M.J.S. Dewar, E.G. Zoebisch, E.F. Healy, J.J.P. Stewart, J. Am. Chem. Soc. 107 (1985) 3898–3902.
[40] <http://www.codessa-pro.com>.
[41] M. Karelson, Molecular Descriptors in QSAR/QSPR, Wiley-Interscience, New York, 2000.
[42] L.B. Kier, L.H. Hall, Eur. J. Med. Chem. 12 (1977) 307–312.
[43] R.H. Rohrbaugh, P.C. Jurs, Anal. Chim. Acta 199 (1987) 99–109.
[44] A.R. Katritzky, V.S. Lobanov, M. Karelson, Comprehensive Descriptors for Structural and Statistical Analysis, Reference Manual, Version 2.0 (1994).
[45] J.H. Friedman, Ann. Stat. 19 (1991) 1–67.
[46] D.F. Specht, IEEE Trans Neural Networks 2 (1991) 568–576.
[47] E. Parzen, Ann. Math. Stat. 33 (1962) 1065–1076.
[48] J.H. Friedman, W. Stuetzle, J. Am. Stat. Assoc. 76 (1981) 817–823.
[49] J.H. Friedman, Classification and Multiple Regression through Projection Pursuit. Technical Report LCS 12, Stanford University, Laboratory for Computational Statistics, Stanford, CA, 1985.
[50] <http://www.r-project.org>.
[51] V. Vapnik (Ed.), The Nature of Statistical Learning Theory, Springer, New York, 1995.
[52] V. Vapnik, S. Golowich, A. Smola, Support Vector Method for Function Approximation, Regression Estimation, and Signal Processing, in: M.C. Mozer, M.I. Jordan, T. Petsche (Eds.), Advances in Neural Information Processing Systems 9, MIT Press, Cambridge, MA, 1997, pp. 281–287.
[53] V. Vapnik (Ed.), Statistical Learning Theory, Wiley, New York, 1998.
[54] B. Scholkopf, C. Burges, A. Smola (Eds.), Advances in Kernel Methods – Support Vector Learning, MIT Press, Cambridge, MA, 1999.
[55] S.R. Gunn, M. Brown, K.M. Bossley, Intel. Data Anal. 1208 (1997) 313–323.
[56] M.J.L. Orr (Ed.), Introduction to Radial Basis Function Networks, Centre for Cognitive Science, Edinburgh University, 1996.
[57] M.J.L. Orr (Ed.), MATLAB Routines for Subset Selection and Ridge Regression in Linear Neural Networks, Centre for Cognitive Science, Edinburgh University, 1996.