Bioorganic & Medicinal Chemistry Letters 20 (2010) 6082–6087

Contents lists available at ScienceDirect

Bioorganic & Medicinal Chemistry Letters

journal homepage: www.elsevier.com/locate/bmcl

Validated predictive QSAR modeling of *N*-aryl-oxazolidinone-5-carboxamides for anti-HIV protease activity

Amit Kumar Halder, Tarun Jha *

Division of Medicinal and Pharmaceutical Chemistry, Department of Pharmaceutical Technology, PO Box 17020, Jadavpur University, Kolkata 700 032, India

ARTICLE INFO

Article history:
Received 12 January 2010
Revised 8 August 2010
Accepted 11 August 2010
Available online 13 August 2010

Keywords:
QSAR modeling
N-Aryl-oxazolidinone-5-carboxamide
Stepwise regression
RTSA index
Wang–Ford charge

ABSTRACT

Validated predictive QSAR modeling was done on some *N*-aryl-oxazolidinone-5-carboxamides for higher anti-HIV protease activities. Stepwise regression developed significant models showing importance of atom based descriptors like RTSA indices, Wang–Ford charges and different whole molecular descriptors. The true predictabilities of QSAR models were justified by challenging these against an external dataset. A representative high active compound was predicted by this modeling. It showed that internal and external validations may lead to the same conclusion.

Acquired immune deficiency syndrome (AIDS) is an end stage disease. It is manifested by gradual deterioration of the immune competence of infected patients.[1] Human immunodeficiency virus-1 (HIV-1) is the causative organism for AIDS. It belongs to the *lentiviridae* family of pathogenic retroviruses. The HIV-1 protease is a homodimeric aspartyl protease. It cleaves a 55 kDa polyprotein precursor and produces smaller functional protein fragments. These are responsible for infectivity of the budding virions.[2] Inhibition of this protease enzyme stops these processing steps giving rise to non-infectious and immature progeny virions. A number of HIV-1 protease inhibitors (indinavir, ritonavir, nelfinavir, etc.) are marketed as anti-HIV drugs. Recently, these were combined with nucleoside and/or non-nucleoside reverse transcriptase inhibitors for highly active antiretroviral therapy (HAART). It is proved to be an effective therapy against AIDS.[3] However, None of the available anti-protease drugs or treatment is completely devoid of untoward effects.[4] Phenotypic resistance and cross resistance also restricted their uses.[5] Hence, search for new more active as well as less toxic HIV-1 protease inhibitors are still in progress. To find the structural requirements for more active anti-HIV protease agents, QSAR modeling was done on some *N*-aryl-oxazolidinone-5-carboxamides.[6] The general structure of these compounds with arbitrary numbering is shown in Figure 1.

The HIV-1 protease inhibitory activity (K_i in nM) data of thirty-eight *N*-aryl-oxazolidinone-5-carboxamides were collected from the published work of Ali et al.[6] The K_i values were converted to the negative logarithmic scale (pK_i). The pK_i values were used as response variables. The activities of compounds are shown in Table 1. Atom based electrotopological state atom indices (ETSA)[7,8] and refractotopological state atom indices (RTSA)[9,10] were calculated using computer program 'MOUSE'.[11] Atomic descriptors like Wang–Ford charges, electrostatic potential charges and different whole molecular electronic descriptors were calculated by using *Chem 3D Pro* package.[12] Energy minimization of these structures was separately done under MOPAC module of AM1 (Austin Model 1) method using RHF (restricted Hartree–Fock: closed shell) wave function. Other whole molecular descriptors were calculated by *Dragon* software.[13] The total number of calculated whole molecular descriptors was about 700. Descriptors without variances were

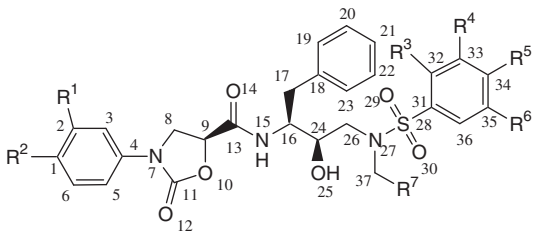


Figure 1. General structure of *N*-aryl-oxazolidinone-5-carboxamide with arbitrary numbering.

* Corresponding author. Tel.: +91 33 24146666x2495, mobile: +91 09433187443; fax: +91 3324146927.
E-mail address: tjupharm@yahoo.com (T. Jha).

0960-894X/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmcl.2010.08.050

Table 1
K_i and pK_i values of compounds in the dataset I[6]

Compd[a]	R^1	R^2	R^3	R^4	R^5	R^6	R^7	K_i[b]	pK_i
1	H	H	H	H	OCH_3	H	*i*Pr	0.100	10.000
2	F	H	H	H	OCH_3	H	*i*Pr	0.083	10.080
3	F	F	H	H	OCH_3	H	*i*Pr	0.066	10.180
4	CF_3	H	H	H	OCH_3	H	*i*Pr	0.006	11.222
5	Ac	H	H	H	OCH_3	H	*i*Pr	0.001	12.097
6	H	Ac	H	H	OCH_3	H	*i*Pr	0.004	11.398
7	OCH_3	H	H	H	OCH_3	H	*i*Pr	0.045	10.347
8	H	H	H	H	NH_2	H	*i*Pr	0.530	9.276
9	F	H	H	H	NH_2	H	*i*Pr	0.170	9.769
10	F	F	H	H	NH_2	H	*i*Pr	0.230	9.638
11	CF_3	H	H	H	NH_2	H	*i*Pr	0.042	10.377
12	Ac	H	H	H	NH_2	H	*i*Pr	0.032	10.495
13	H	Ac	H	H	NH_2	H	*i*Pr	0.184	9.735
14	F	H	H	–O–CH_2–O–		H	*i*Pr	0.107	9.971
15	F	F	H	–O–CH_2–O–		H	*i*Pr	0.085	10.071
16	CF_3	H	H	–O–CH_2–O–		H	*i*Pr	0.016	10.796
17	Ac	H	H	–O–CH_2–O–		H	*i*Pr	0.006	11.222
18	H	Ac	H	–O–CH_2–O–		H	*i*Pr	0.016	10.796
19	F	H	H	F	OCH_3	H	*i*Pr	0.070	10.155
20	F	F	H	F	OCH_3	H	*i*Pr	0.343	9.465
21	CF_3	H	H	F	OCH_3	H	*i*Pr	0.072	10.143
22	Ac	H	H	F	OCH_3	H	*i*Pr	0.133	9.876
23	H	Ac	H	F	OCH_3	H	*i*Pr	0.080	10.097
24	CF_3	H	H	H	OCF_3	H	*i*Pr	10.000	8.000
25	Ac	H	H	H	OCF_3	H	*i*Pr	2.000	8.699
26	H	H	H	OCH_3	H	H	*i*Pr	3.800	8.420
27	H	Ac	H	OCH_3	H	H	*i*Pr	0.840	9.076
28	F	H	H	H	H	H	*c*Pr	0.257	9.590
29	F	F	H	H	OCH_3	H	*c*Pr	0.580	9.237
30	H	Ac	H	H	OCH_3	H	*c*Pr	0.800	9.097
31	H	H	H	OCH_3	H	H	2-TP[c]	238.700	6.622
32	F	H	H	OCH_3	H	H	2-TP[c]	188.800	6.724
33	H	Ac	H	OCH_3	H	H	2-TP[c]	29.500	7.530
34	H	H	F	H	F	F	2-TP[c]	170.200	6.769
35	F	H	F	H	F	F	2-TP[c]	160.200	6.795
36	H	Ac	F	H	F	F	2-TP[c]	167.700	6.775
37	H	H	H	OCH_3	H	H	2THF[d]	42.000	7.377
38	F	H	H	OCH_3	H	H	2THF[d]	150.000	6.824

[a] Compound number.
[b] Taken from Ref. 6.
[c] 2-TP = 2-thiophene.
[d] 2THF = 2-tetrahydrofuran.

discarded. Highly correlated descriptors were grouped together and descriptor with the highest correlation with the pK_i was selected from that group. From the remaining 231 descriptors, selection of descriptors was done by stepwise regression analysis[14] using *F* value as stepping criteria ($F = 3.0$ for inclusion, $F = 2.9$ for exclusion). During stepwise regression, all intercorrelated independent parameters (correlation coefficient >0.6) were discarded. Calculated values of descriptors are shown in Supplementary data. The dataset[6] (henceforth, called the dataset I) was divided into the test set and the training set by Y based ranking method.[15] Molecules were ranked based on their pK_i values. Molecules in the 1st, 5th, 9th, 13th and so on rows were collected. These 10 collected compounds (almost 25% total dataset; compounds **5**, **7**, **10**, **14**, **16**, **23**, **26**, **29**, **32**, and **38**) were treated as the test set (test set I), whereas, the rest was the training set. Regression equations were justified by correlation coefficient (R), adjusted R^2 (R^2_a), variance ratio (F) at specified degrees of freedom (df), probability factor related to F ratio (p), standard error of estimate (SEE). Leave-one-out (LOO) cross validation method was used to validate models. Internal predictabilities of these equations were justified by predicted residual sum of squares (PRESS), cross-validated R^2 (R^2_{CV}), standard deviation error of prediction (SDEP) and standard error of PRESS (S_{PRESS}). External predictabilities of these models were justified by R^2_{Pred} values. Moreover, k and k' values[16,17] and root mean square (rm^2) value[18] were also considered as external predictability parameters. k and k' values should be within the range of 0.85–1.15 and R^2_{Pred} and rm^2 values should be more than 0.5.

Stepwise regression generated different equations. Initially selected equations ($R^2_{CV} > 0.75$) were considered. Two equations were developed with atom based descriptors. The first equation is:

$$\begin{aligned} pK_i = {} & 71.832(\pm 14.037) + 14.731(\pm 5.257)q_9 \\ & + 27.683(\pm 6.178)q_{10} + 14.791(\pm 2.054)q_{27} \\ & - 106.543(\pm 14.319)q_{29} \end{aligned} \quad (1)$$

$n = 28;\ R = 0.934;\ R^2_A = 0.851;\ F(4,23) = 39.594;$
$p < 0.00001;\ SEE = 0.558;\ R^2_{CV} = 0.820;\ PRESS = 10.196;$
$SDEP = 0.603;\ S_{PRESS} = 0.665,\ R^2_{Pred} = 0.782;\ k = 1.009;$
$k' = 0.985;\ rm^2 = 0.726.$

where n is the number of compounds in the training set. The q_9, the q_{10}, the q_{27} and the q_{29} are Wang–Ford charges of atom numbers 9, 10, 27, and 29 of the general structure (Fig. 1). For atom based descriptors, the best equation was obtained with the RTSA index of the atom number 34 (R_{34}), Wang–Ford charges of atom numbers 9 and 32 (the q_9 and the q_{32}, respectively) and the existence of amino group at R^5 position of the general structure (I_1). The equation is:

$$\begin{aligned} pK_i = {} & 15.647(\pm 0.616) - 1.801(\pm 0.155)R_{34} \\ & + 14.902(\pm 5.138)q_9 - 5.779(\pm 0.670)q_{32} \\ & - 1.256(\pm 0.345)I_1 \end{aligned} \quad (2)$$

$n = 28;\ R = 0.941;\ R^2_A = 0.865;\ F(4,23) = 44.340;$
$p < 0.00001;\ SEE = 0.531;\ R^2_{CV} = 0.834;\ PRESS = 9.369;$
$SDEP = 0.578;\ S_{PRESS} = 0.638;\ R^2_{Pred} = 0.754;\ k = 0.979;$
$k' = 1.016;\ rm^2 = 0.689.$

When only whole molecular descriptors were subjected to stepwise regression analysis, two equations showed R^2_{CV} value more than 0.75. The first equation is:

$$\begin{aligned} pK_i = {} & 20.166(\pm 3.305) - 0.009(\pm 0.002)QXXm \\ & - 0.349(\pm 0.115)n\mathrm{CaH} - 6.456(\pm 2.591)PJI2 \\ & + 4.916(\pm 1.065)Lop \end{aligned} \quad (3)$$

$n = 28;\ R = 0.917;\ R^2_A = 0.814;\ F(4,23) = 30.549;$
$p < 0.00001;\ SEE = 0.624;\ R^2_{CV} = 0.776;\ PRESS = 12.670;$
$SDEP = 0.673;\ S_{PRESS} = 0.742;\ R^2_{Pred} = 0.536;\ k = 0.988;$
$k' = 0.999;\ rm^2 = 0.501.$

Where geometric descriptor *QXXm*[19] stands for *Qxx* COMMA2 value (weighted by atomic masses). The *nCaH* is the number of unsubstituted aromatic carbon with sp^2 orbital. Topological descriptors like *Lop*[20] and the *PJI2*[21] are lopping centric index and 2D Petitjean shape index, respectively. This equation explains 81.4% variance and predicts 77.6% variance of biological activity. The second equation is:

$$\begin{aligned} pK_i = {} & 48.354(\pm 6.225) - 3.940(\pm 1.508)Dispp \\ & + 1.420(\pm 0.106)nCp - 92.402(\pm 21.527)X2Av \\ & - 1.695(\pm 0.232)S2K \end{aligned} \quad (4)$$

$n = 28;\ R = 0.945;\ R^2_A = 0.874;\ F(4,23) = 47.643;$
$p < 0.00001;\ SEE = 0.515;\ R^2_{CV} = 0.843;\ PRESS = 8.892;$
$SDEP = 0.563;\ S_{PRESS} = 0.622;\ R^2_{Pred} = 0.782;\ k = 1.011;$
$k' = 0.983;\ rm^2 = 0.677.$

The *Dispp*[22] or the d COMMA2 value (weighted by atomic polarizability) is a geometric descriptor. The *X2Av*[23] and the *S2K*[24] are topological descriptors representing the average valence connectivity index (χ-2) and the 2-path Kier alpha modified shape index, respectively. Functional descriptor *nCp* is the number of primary carbons having sp^3 orbital. When atom based and whole molecular descriptors were considered together, the best equation was found. It is:

$$\begin{aligned} pK_i = {} & 16.701(\pm 2.072) - 6.799(\pm 0.670)q_{35} \\ & + 0.454(\pm 0.157)L/Bw + 0.631(\pm 0.108)nCp \\ & - 2.259(\pm 0.385)VEA1 \end{aligned} \quad (5)$$

$n = 28;\ R = 0.966;\ R^2_A = 0.922;\ F(4,23) = 80.669;$
$p < 0.00001;\ \mathrm{SEE} = 0.404;\ R^2_{CV} = 0.902;\ \mathrm{PRESS} = 5.567;$
$\mathrm{SDEP} = 0.446;\ S_{\mathrm{PRESS}} = 0.492;\ R^2_{\mathrm{Pred}} = 0.825;\ k = 0.989;$
$k' = 1.006;\ \mathrm{rm}^2 = 0.790.$

In Eq. 5, the q_{35} is the Wang–Ford charge of the atom number 35. The *L/Bw*[25] is length to breadth ratio and the *VEA1*[26] is eigenvector coefficient sum of adjacency matrix. Sometimes stepwise regression produces equations with biased intercorrelated descriptors with low effects on the dependent parameter. To overcome this, variation inflation factors (VIF)[27,28] of each model were determined. The VIF was calculated from $1/1 - r^2$ where r^2 is the multiple correlation coefficient of one descriptor's effect regressed on the remaining molecular descriptors. For VIF value larger than 10, information of the descriptor may be hidden by other descriptors. The VIF factors of different descriptors of Eqs. 1–5 are provided in Supplementary data. The highest value obtained was 2.834 for q_{27} in Eq. 1. To check whether the built models are biased to a particular splitting method or not, k means cluster analysis (k-MCA)[29] was adopted. The k-MCA divided the dataset in four clusters containing 9, 5, 14, and 10 members each depending on Euclidian distance. Nine compounds (25% compounds of the total dataset, compounds **5**, **9**, **15**, **19**, **21**, **23**, **27**, **31**, and **38**) were selected randomly for designing the test set (test set II). The remaining compounds were treated as the training set (Training set II). The QSAR Eqs. of 1–5 were redeveloped. These are:

$$\begin{aligned} pK_i = {} & -71.832(\pm 14.037) + 14.731(\pm 5.257)q_9 \\ & + 27.683(\pm 6.178)q_{10} + 14.791(\pm 2.054)q_{27} \\ & - 106.543(\pm 14.319)q_{29} \end{aligned} \quad (6)$$

$n = 29;\ R = 0.924;\ R^2_A = 0.830;\ F(4,24) = 35.206;$
$p < 0.00001;\ \mathrm{SEE} = 0.595;\ R^2_{CV} = 0.807;\ \mathrm{PRESS} = 11.276;$
$\mathrm{SDEP} = 0.624;\ S_{\mathrm{PRESS}} = 0.684.$

Similarly:

$$\begin{aligned} pK_i = {} & 16.197(\pm 0.617) - 1.863(\pm 0.171)R_{34} \\ & + 8.389(\pm 4.603)q_9 - 5.273(\pm 0.681)q_{32} \\ & - 1.194(\pm 0.358)I_1 \end{aligned} \quad (7)$$

$n = 29;\ R = 0.931;\ R^2_A = 0.845;\ F(4,24) = 9.281;$
$p < 0.00001;\ \mathrm{SEE} = 0.568;\ R^2_{CV} = 0.809;\ \mathrm{PRESS} = 11.123;$
$\mathrm{SDEP} = 0.619;\ S_{\mathrm{PRESS}} = 0.681.$

another model:

$$\begin{aligned} pK_i = {} & 20.166(\pm 3.305) - 0.009(\pm 0.002)QXXm \\ & - 0.349(\pm 0.115)nCaH - 6.456(\pm 2.591)PJI2 \\ & + 4.916(\pm 1.065)Lop \end{aligned} \quad (8)$$

$n = 29;\ R = 0.900;\ R^2_A = 0.814;\ F(4,23) = 30.549;$
$p < 0.00001;\ \mathrm{SEE} = 0.681;\ R^2_{CV} = 0.721, \mathrm{PRESS} = 16.297;$
$\mathrm{SDEP} = 0.750;\ S_{\mathrm{PRESS}} = 0.824.$

the other model:

$$\begin{aligned} pK_i = {} & 48.608(\pm 5.393) - 4.952(\pm 1.274)Dispp \\ & + 1.444(\pm 0.103)nCp - 98.825(\pm 19.710)X2Av \\ & - 1.603(\pm 0.200)S2K \end{aligned} \quad (9)$$

$n = 29;\ R = 0.949;\ R^2_A = 0.885;\ F(4,24) = 54.718;$
$p < 0.00001;\ \mathrm{SEE} = 0.490;\ R^2_{CV} = 0.809;\ \mathrm{PRESS} = 8.918;$
$\mathrm{SDEP} = 0.554;\ S_{\mathrm{PRESS}} = 0.610.$

and the final one:

$$\begin{aligned} pK_i = {} & 16.701(\pm 2.072) - 6.799(\pm 0.670)q_{35} \\ & + 0.454(\pm 0.157)L/Bw + 0.631(\pm 0.108)nCp \\ & - 2.259(\pm 0.385)VEA1 \end{aligned} \quad (10)$$

$n = 29;\ R = 0.960;\ R^2_A = 0.922;\ F(4,23) = 80.669;$
$p < 0.00001;\ SEE = 0.681;\ R^2_{CV} = 0.890;\ \mathrm{PRESS} = 6.448;$
$\mathrm{SDEP} = 0.471;\ S_{\mathrm{PRESS}} = 0.518.$

Comparing earlier Eqs. 1–5, Eqs. 6–10 show similar statistical qualities. Therefore, QSAR models are not dependent on any particular splitting method. Stepwise regression procedure is highly susceptible to chance correlation. Randomization test was performed to circumvent this. Values of pK_is were permuted randomly and repeatedly to generate QSAR models. Resulting scrambled regression coefficients (R_{SCR}) are compared with regression coefficients of the original QSAR models. For statistically significant models, original regression coefficients (R) should be significantly greater than R_{SCR}.[30] For 95% confidence interval, 19 trials were made. The R_{SCR} values are shown in Supplementary data. As R_{SCR} values for different models range from 0.477 to 0.565 (much less than non-randomized R_s), robustness of the original models was confirmed. Determination of the applicability domain[31] is a theoretical approach for ensuring the predictability of the model property for the entire set of chemicals. Applicability domain is a theoretical region in a chemical space defined by the model descriptor and model response. Predictions for only those chemicals that fall into that space are considered reliable. *Extent of extrapolation*[31] that calculates applicability domain by leverage for each chemical was adopted and the details of it are avoided. Prediction is considered unreliable for compounds of leverage value greater than $3p/n$ where p is the number of model variables plus one and n is the number of datapoints. For Eqs. 1–5, the warning limit for these compounds was 0.54. Leverage values for compounds of Eqs. 1–5 are provided in Supplementary data. All compounds, except compound **13** for Eq. 1, lie within the warning limit.

A QSAR modeling faces a serious drawback when a model fails to predict activities of compounds outside the dataset. For a small dataset with limited extent of structural variations this problem is further manifested. Success of a model is not truly justified until it predicts activities of structurally diverse compounds. Considering this, sixteen structurally related anti-HIV I protease compounds (compounds **39–51**) were collected from the work of Reddy et al.[32] and this dataset is henceforth called Dataset II. These compounds were treated as the test set (test set III) for the Dataset I (training set III). General structure of compounds **39–51** is shown in Figure 2. The activities of these compounds are provided in Table 2. Structures, activities of compounds **52–54** are given in Table 3.

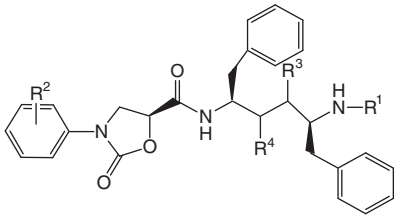


Figure 2. General structure of compounds **39**–**51**.

Calculated values of whole molecular descriptors for the dataset II are shown in Supplementary data. As positions of these atoms are changed for the dataset II, it was not possible to check predictabilities of atom based descriptors. Hence, Eqs. 1, 2, and 5 were excluded from external validation process. The predictabilities of other equations were challenged with respect to the test set III. Predictabilities of these models were justified by R^2_{Pred}, k, k' and rm^2 values. After the new QSAR modeling, Eqs. 3 and 4 produced Eqs. 11 and 12, respectively and these are:

$$\begin{aligned} \text{p}K_\text{i} = {} & 17.244(\pm 3.528) - 0.008(\pm 0.002)QXXm \\ & - 0.368(\pm 0.119)nCaH - 4.379(\pm 2.866)PJI2 \\ & + 5.696(\pm 1.113)Lop \end{aligned} \tag{11}$$

$n = 38;\ R = 0.869;\ R^2_\text{A} = 0.725;\ F(4,33) = 25.441;$
$p < 0.00001;\ \text{SEE} = 0.782;\ R^2_\text{CV} = 0.679;\ \text{PRESS} = 26.413;$
$\text{SDEP} = 0.834;\ S_\text{PRESS} = 0.895;\ R^2_\text{Pred} = -3.483;\ k = 0.700;$
$k' = 1.567;\ \text{rm}^2 = 0.226.$

$$\begin{aligned} \text{p}K_\text{i} = {} & 49.228(\pm 5.506) - 4.470(\pm 1.423)Dispp \\ & + 1.490(\pm 0.105)nCp - 100.870(\pm 20.166)X2Av \\ & - 1.642(\pm 0.210)S2K \end{aligned} \tag{12}$$

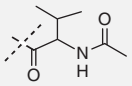
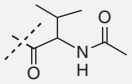
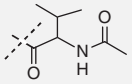
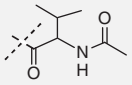
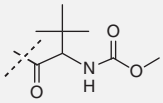
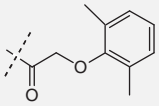
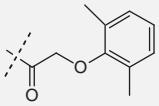
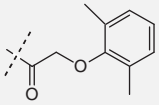
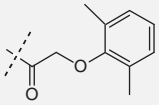
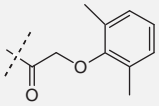
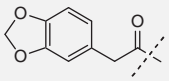
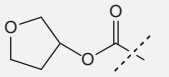
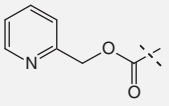
$DC = \mathbf{43}, \mathbf{51}, \mathbf{54};\ n = 38;\ R = 0.929;\ R^2_\text{A} = 0.847;$
$F(4,33) = 52.135;\ p < 0.00001;\ \text{SEE} = 0.584;$
$R^2_\text{CV} = 0.820;\ \text{PRESS} = 14.852;\ \text{SDEP} = 0.625;$
$S_\text{PRESS} = 0.671;\ R^2_\text{Pred} = 0.588;\ k = 1.114;\ k' = 0.924;\ \text{rm}^2 = 0.184.$

Evidently, predictability of Eq. 11 for the test set II was poor whereas the same of Eq. 12 was moderate. For Eq. 12, three compounds (**43, 51**, and **54**) were deleted as these showed extreme varied predictive results. These compounds may act through different mechanism(s) of action(s). For Eq. 11, no deletion could be made as most of the predictive values were out of range. Now, for dataset I, several QSAR models were developed but equations with R^2_CV less than 0.75 were discarded. After obtaining unsatisfactory predictability results from the developed models, other models were tested. One QSAR model produced Eq. 13 on Y based ranking for the dataset I. This equation is:

$$\begin{aligned} \text{p}K_\text{i} = {} & -9.838(\pm 5.030) + 14.912(\pm 3.475)IVDE \\ & - 1.581(\pm 0.979)J3D - 1.638(\pm 0.405)SEigm \\ & - 0.394(\pm 0.120)G(S\ldots S) \end{aligned} \tag{13}$$

$n = 28;\ R = 0.894;\ R^2_\text{A} = 0.763;\ F(4,23) = 22.791;$
$p < 0.00001;\ \text{SEE} = 0.704;\ R^2_\text{CV} = 0.700;\ \text{PRESS} = 16.989;$
$\text{SDEP} = 0.779;\ S_\text{PRESS} = 0.859;\ R^2_\text{Pred} = 0.755;\ k = 1.034;$
$k' = 0.971;\ \text{rm}^2 = 0.715.$

Table 2
K_i and pK_i values of compounds in the dataset II[36]

Compd[a]	R^1	R^2	R^3	R^4	K_i[b]	pK_i
39		H	H	OH	131.400	6.881
40		3,4-diF	H	OH	2.13.000	8.672
41		$3CF_3$	H	OH	2.620	8.582
42		3-Ac	H	OH	0.980	9.009
43		H	H	OH	3.510	8.455
44		H	H	OH	8.050	8.094
45		3Ac	H	OH	35.110	7.455
46		4Ac	H	OH	39.870	7.399
47		H	OH	H	4731.000	5.325
48		$3CF_3$	OH	H	21.630	7.665
49		H	H	H	657.000	6.182
50		H	OH	H	1800.000	5.745
51		$3CF_3$	OH	H	41.090	7.386

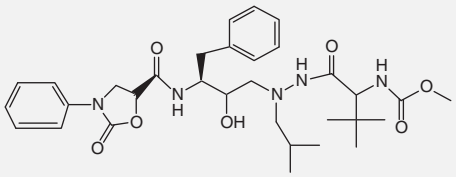
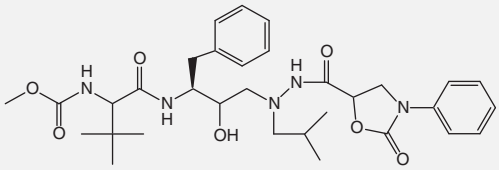
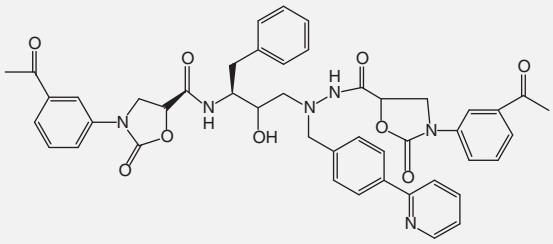
[a] Compd: compound number.
[b] Taken from Ref. 36.

Eq. 13 predicts 76.3% and explains 70.0% variances. VIF values, average scrambled regression coefficient, leverage values and correlation matrix of Eq. 13 are provided in Supplementary data. Upon cluster based splitting technique, these descriptors produced the following equation:

$$\begin{aligned} \text{p}K_\text{i} = {} & -13.928(\pm 5.338) + 18.100(\pm 3.679)IVDE \\ & - 2.177(\pm 0.908)J3D - 1.845(\pm 0.416)SEigm \\ & - .398(\pm 0.111)G(S\ldots S) \end{aligned} \tag{14}$$

$n = 29;\ R = 0.894;\ R^2_\text{A} = 0.767;\ F(4,24) = 24.027;$
$p < 0.00001;\ \text{SEE} = 0.697;\ R^2_\text{CV} = 0.725;\ \text{PRESS} = 16.080;$
$\text{SDEP} = 0.745;\ S_\text{PRESS} = 0.818.$

Table 3
K_i and pK_i values of the compounds in the dataset II[36]

Compd[a]	Structures	K_i[b]/pK_i
52		7.000/8.155
53		10.600/7.975
54		1090.000/5.963

[a] Compd: compound number.
[b] Taken from Ref. 36.

External validation for the test set III produced Eq. 15.

$$\begin{aligned} pK_i = &-15.794(\pm5.005) - 1.783(\pm0.927)J3D \\ &- 0.304(\pm0.115)G(S\ldots S) + 18.7052(\pm3.535)IVDE \\ &- 1.934(\pm0.425)SEigm \end{aligned} \quad (15)$$

$n = 38;\ R = 0.867;\ R_A^2 = 0.721;\ F(4,33) = 24.876;$
$p < 0.00001;\ \mathrm{SEE} = 0.789;\ R_{CV}^2 = 0.679;\ \mathrm{PRESS} = 26.462;$
$\mathrm{SDEP} = 0.834;\ S_{\mathrm{PRESS}} = 0.802;\ R_{\mathrm{Pred}}^2 = 0.878;\ k = 0.936;$
$k' = 1.074;\ \mathrm{rm}^2 = 0.610.$

For the dataset I, inherent statistical qualities of Eqs. 13 and 14 are inferior to earlier developed models [R_{CV}^2 values <0.75] but the predictability of the model for external data is higher than Eqs. 11 and 12 as suggested by R_{Pred}^2, k, k' and rm^2 values of Eq. 15.

In 2D QSAR analyses, a number of equations are developed, but equations of higher statistical qualities (evidenced by R, R_{CV}^2, etc.) are mainly highlighted. To obtain statistically robust models, an arbitrary cut off value of R_{CV}^2 (0.75) was initially considered. After including more number of compounds from other dataset (the Dataset II), two types of equations were found. One type of equations [Eqs. 3 and 4] has good statistical qualities for inherent validation but have moderate [e.g., Eq. 4] or poor [e.g., Eq. 3] statistical predictions for Dataset II. The other type of equation [Eq. 13] may have lesser overall statistical quality for dataset I, as compared to the first kind of equations [Eqs. 3 and 4] but has satisfactory, if not excellent, predictability for a set of compounds (of Dataset II) that not only lie outside the data but also have much more structural variations as well as complexities. The first kind of equations highlights the structural details of a set of compounds. Their predictabilities are limited to the dataset. The second kind of equations [Eq. 13] is more versatile. Their scope of predictability may be safely extrapolated for the design of new compounds. If any equation, that is, Eq. 3, gave anomalous predictions for external dataset, it is better to ignore the equation, and thus, detail of this model is avoided. The inconsistency between inherent and external validation for small dataset may be due to the reason that for inherent validation, some descriptors were present in dormant stages due to limited number of datapoints and complexities of molecules. The influences of these descriptors become prominent when a different dataset is challenged against these. Hence, equations like Eq. 13 are difficult to find by conventional QSAR methods devoted to inherent validation. The only way to find these models is to expose different equations with external datasets containing compounds with huge structural variations.

Eq. 1 shows importance of the q_9, the q_{10}, the q_{27} and the q_{29} for the dataset I. The positive coefficients of the q_9, the q_{10} and the q_{27} indicate that the positive charge of the atom number 9 and the negative charges of atom numbers 10 and 27 may be increased. The negative charge of the atom number 29 may be decreased. The charge distributions of atom numbers 27 and 29 may be directly correlated with the type of moiety associated with the atom number 27, that is, the nitrogen atom. The charge distributions on these atoms suggest that comparing small groups like isopropyl or cyclopropyl, bulky groups like thiophene and/or tetrahydrofuran may have an influence on the decrease of the negative charge of the atom number 27 and on the increase in the negative charge of the atom number 29 (the O atom). Substitutions on nitrogen was earlier reported as a contributing factor for interaction with the hydrophobic pockets of the HIV-I protease enzyme[33] and this information complies with current QSAR study. The negative coefficient with Wang–Ford charge of the atom number 32 in Eq. 2 suggests that Wang–Ford charge of this atom may be decreased. The *R*-state index is assumed to be related to the dispersive or van der Waals interactions with the enzyme. From the negative coefficient of R_{34}, it is evident that the increase in the value of this parameter may be unfavorable. The negative coefficient of I_1 defies the presence of amino group at the atom number 34. In Eq. 4, negative coefficients of the *Dispp*, the *S2K* and the *X2Av* indicate that low values of these descriptors may be favorable. The positive coefficient of the *nCp* suggests that the total number of primary sp^3 car-

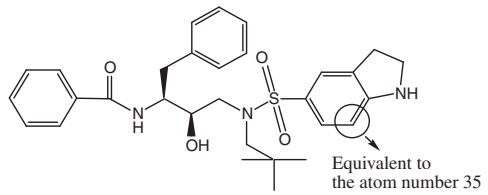


Figure 3. Structure of the designed proposed compound.

bon may increase the activity. In Eq. 5, the negative coefficient of Wang–Ford charge at the atom number 35 suggests that the charge of the atom may be decreased. This descriptor also indicates that the presence of electron donating group in R^5 position and the absence of electron withdrawing group in R^6 position may be conducive to the activity. The positive coefficient of *L/Bw* suggests that the required ratio of length and breadth may be higher. The negative coefficient of *VEA1* signifies that the value of this descriptor may be lowered. Lastly, Eq. 13 shows importance of topological descriptors Balaban 3D index (*J3D*), mean information content index vertex degree equality (*IVDE*), eigenvalue sum for mass weighed distance matrix (*SEigm*) and geometric descriptor sum of geometric distance between two sulfur atoms [*G*(*S*–*S*)]. Regression coefficients show that increased values of *IVDE* and decreased values of the *J3D*, the *SEigm* and the *G*(*S*...*S*) may be favorable. As values of *G*(*S*...*S*) are positive for compounds **31–36** and zero for all other compounds, the presence of bulky thiophene group at R^7 position of the dataset I may be unfavorable.

Based on the QSAR predictions, one new compound was designed. This one was proposed for higher activity. For this design, descriptors of Eq. 5, undoubtedly the best internal validated model, as well as that of Eq. 13, proved to be the best external validated model, were considered. The structure of the proposed compound is given in Figure 3. The predicted pK_i value of this compound from Eq. 5 and 13 were 12.095 and 12.099, respectively. These activities are almost equivalent to that of compound **5** which showed the highest activity (pK_i = 12.097). Significance of this designed compound lies in the fact that its predicted activities are similar for the best internal and external validated models. Moreover, the molecule (Fig. 3) retained its central hydroxyl group which may form hydrogen bonds with the carboxylate groups of the catalytic aspartic acids of the protease enzyme as observed in the co-crystal structures. However, this is just one of the many possible structures that may be designed from the current QSAR modeling. Thus, this work shows that both internal and external validations were equally important and may lead to the same prediction and conclusion. Regression coefficients of Eqs. 1–5 and 13 are significant at 95% confidence level as shown by their *t*- and *p*-values (provided as Supplementary data). Observed (Obs), calculated (Calcd), residual (Res), predicted residual (Pres) and LOO predicted activities of Eqs.1–5 and 14, correlation matrices of dependent and independent variables and Plots of observed versus predicted values of important equations like Eqs. 5, 13, and 15 are shown in Supplementary data. The recommended ratio of the number of predictor parameters to number of data point of 1:5 was maintained.[34–36]

Acknowledgements

The authors are grateful to All India Council for Technical Education (AICTE), New Delhi for awarding a research project.

Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.bmcl.2010.08.050.

References and notes

1. Alberts, B.; Johnson, A.; Lewis, J.; Raff, M.; Roberts, K.; Walter, P. *Molecular Biology of the Cell*; Garland Fransis-Taylor and Fransis group: New York, 2002. Chapter 23.
2. Chakravarty, A. K. *Immunology and Immunotechnology*; Oxford University Press: New Delhi, 2006. Chapter 16.
3. Barbaro, G.; Scozzafava, A.; Mastrolorenzo, A.; Supuran, C. T. *Curr. Pharm. Des.* **2005**, *11*, 1805.
4. Safrin, S. In *Basic and Clinical Pharmacology*; Katzung, B. G., Ed.; McGraw Hill: New Delhi, 2004; pp 801–857.
5. Tripathi, K. D. *Essentials of Medical Pharmacology*; Jaypee Brothers Medical Publishers Ltd: New Delhi, 2003. Chapter 60.
6. Ali, A.; Reddy, G. S. K. K.; Cao, H.; Anjum, S. G.; Nalam, M. N. L.; Schiffer, C. A.; Rana, T. M. *J. Med. Chem.* **2006**, *49*, 7342.
7. Hall, H.; Mohney, B.; Kier, L. B. *Quant. Struct.-Act. Relat.* **1991**, *10*, 43.
8. de Gregorio, C.; Kier, L. B.; Hall, L. H. *J. Comput. Aided Mol. Des.* **1998**, *12*, 557.
9. Ghose, A. K.; Pritchett, A.; Crippen, G. M. *J. Comput. Chem.* **1988**, *9*, 80.
10. Carrasco, R.; Padron, A. J.; Galvez, J. *J. Pharm. Pharm. Sci.* **2004**, *7*, 19.
11. MOUSE, a computer program developed by Jadavpur University.
12. Chem 3D Pro Version 5.0 and Chem Draw Ultra Version 5.0 are programs Cambridge Soft Corporation, USA.
13. DRAGON web version 2.1 is a QSAR software developed by Milano Chemometrics and QSAR Research Group, Dipartimento di Scienze dell'Ambiente e del Territorio Universitàdegli Studi di Milano, Bicocca.
14. Bhattacharya, P.; Roy, K. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 3737.
15. Hemmateenejad, B. *J. Chem.* **2004**, *18*, 475.
16. Tropsha, A.; Gramatica, P.; Gomber, V. K. *QSAR Comb. Sci.* **2003**, *22*, 69.
17. Golbraikh, A.; Tropsha, A. *J. Mol. Graphics Modell.* **2002**, *20*, 269.
18. Roy, P. P.; Roy, K. *QSAR Comb. Sci.* **2008**, *27*, 302.
19. Balaban, A. T. *Theor. Chim. Acta* **1979**, *53*, 355.
20. Silvermann, B. D. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1470.
21. Petitjean, M. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 331.
22. Randic, M. *Int. J. Quantum Chem. Quantum Biol. Symp.* **1998**, *15*, 201.
23. Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure Activity Analysis*; RSP-Wiley: Chichetser, UK, 1986.
24. Kier, L. B. *Quant. Struct.-Act. Relat.* **1985**, *4*, 109.
25. Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: Weinheim (Ger.), 2000. Chapter 3.
26. Balaban, A. T.; Ciubotariu, D.; Medeleanu, M. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 517.
27. Jaiswal, M.; Khadikar, P. V.; Scozzafava, A.; Supuran, C. T. *Bioorg. Med. Chem. Lett.* **2004**, *14*, 3283.
28. Shapiro, S.; Guggenheim, B. *Quant. Struct.-Act. Relat.* **1998**, *17*, 327.
29. Tropsha, A.. In *Burger's Medicinal Chemistry and Drug Discovery*; Abraham, D. J., Ed.; John Wiley and Sons: New Jersey, 2003; Vol. 1, pp 49–75.
30. Deswal, S.; Roy, N. *Eur. J. Med. Chem.* **2006**, *41*, 1339.
31. Gramatica, P. *QSAR Comb. Sci.* **2007**, *26*, 694.
32. Reddy, G. S. K. K.; Ali, A.; Nalam, M. N. L.; Anjum, S. G.; Cao, H.; Nathans, R. S.; Schiffer, C. A.; Rana, T. M. *J. Med. Chem.* **2007**, *50*, 4316.
33. Lam, P. Y. S.; Jadhav, P. K.; Eyermann, C. E.; Hodge, C. N.; Ru, Y.; Bachelor, L. T.; Meek, J. L.; Otto, M. J.; Rayner, M. M.; Wong, Y. N.; Chang, C. H.; Weber, P. C.; Jackson, D. A.; Sharpe, T. R.; Erickson-Viitanen, S. *Science* **1994**, *263*, 380.
34. Topliss, J. G.; Edwards, R. P. *J. Med. Chem.* **1979**, *22*, 1238.
35. Walker, J. D.; Jaworska, J.; Comber, M. H.; Schultz, T. W.; Dearden, J. C. *Environ. Toxicol. Chem.* **2003**, *22*, 1653.
36. Eriksson, L.; Jaworska, J.; Worth, A. P.; Cronin, M. T. D.; McDowell, R. M.; Gramatica, P. *Environ. Health Perspect.* **2003**, *111*, 1361.