ORIGINAL PAPER

Molecular docking and 3D-QSAR studies of HIV-1 protease inhibitors

**Vijay M. Khedkar · Premlata K. Ambre ·
Jitender Verma · Mushtaque S. Shaikh ·
Raghuvir R. S. Pissurlenkar · Evans C. Coutinho**

Abstract HIV-1 protease is an obligatory enzyme in the replication process of the HIV virus. The abundance of structural information on HIV-1PR has made the enzyme an attractive target for computer-aided drug design strategies. The daunting ability of the virus to rapidly generate resistant mutants suggests that there is an ongoing need for new HIV-1PR inhibitors with better efficacy profiles and reduced toxicity. In the present investigation, molecular modeling studies were performed on a series of 54 cyclic urea analogs with symmetric P2/P2′ substituents. The binding modes of these inhibitors were determined by docking. The docking results also provided a reliable conformational superimposition scheme for the 3D-QSAR studies. To gain insight into the steric, electrostatic, hydrophobic and hydrogen-bonding properties of these molecules and their influence on the inhibitory activity, comparative molecular field analysis (CoMFA) and comparative molecular similarity indices analysis (CoMSIA) were performed. Two different alignment schemes *viz. receptor-based* and *atom-fit alignment*, were used in this study to build the QSAR models. The derived 3D-QSAR models were found to be robust with statistically significant r^2 and r^2_{pred} values and have led to the identification of regions important for steric, hydrophobic and electronic interactions. The predictive ability of the models was assessed on a set of molecules that were not included in the training set. Superimposition of the 3D-contour maps generated from these models onto the active site of enzyme provided additional insight into the structural requirements of these inhibitors. The CoMFA and CoMSIA models were used to design some new inhibitors with improved binding affinity. Pharmacokinetic and toxicity predictions were also carried out for these molecules to gauge their ADME and safety profile. The computational results may open up new avenues for synthesis of potent HIV-1 protease inhibitors.

Introduction

HIV-1 protease (HIV-1PR) is probably the most extensively investigated enzyme for therapeutic intervention in the short history of structure-based drug design. The development of HIV protease inhibitors is regarded as one of the most successful examples of structure-based drug design efforts to date. This enzyme is a critical component in the replicative cycle of the human immunodeficiency virus (HIV) that cleaves the polyproteins transcribed from the *gag* and *pol* genes into enzymes and structural proteins essential for the assembly and maturation of infectious virions [1, 2]. The finding that inactivation of this viral encoded aspartyl protease produces a progeny of virions that are immature and noninfectious, elicited intense efforts in the development of specific and potent inhibitors targeted against this enzyme, as a novel therapy for AIDS.

V. M. Khedkar · P. K. Ambre · J. Verma · M. S. Shaikh ·
R. R. S. Pissurlenkar · E. C. Coutinho (✉)
Department of Pharmaceutical Chemistry,
Bombay College of Pharmacy,
Kalina, Santacruz (E),
Mumbai 400098, India
e-mail: evans@bcpindia.org

Since then, HIV-1PR has become a prime target for therapeutic intervention in this disease. Moreover, availability of several X-ray crystal structures of HIV-1 protease has guided structure-based searches for specific inhibitors.

The ability of the virus to circumvent the inhibitors *via* mutation clearly signifies the need for improved therapeutic agents. To facilitate the design of more specific and potent HIV-1 protease inhibitors, we must improve our understanding of the principles of molecular recognition for this enzyme. Two distinct characteristics of HIV-1 protease have been identified to date that distinguish it from the human aspartyl proteases - renin and pepsin: first, the active form of the enzyme is a homodimer, where each monomer contributes equally to the active site; and second, the presence of a structural water molecule that bridges the inhibitor molecule to the flap of the protein *via* hydrogen bonds. While hydroxyethylene isosteres and phosphinates are known to be the first C2 symmetric molecules able to recognize the HIV-1PR [3, 4], C2 symmetric cyclic urea (CU) based inhibitors are one of the first molecules that displace the structural water molecule involved in ligand-receptor interaction [5]. Since then, the CU scaffold has been extensively explored by medicinal chemists. Although some molecular modification and structure-activity relationships have been reported [6, 7], more quantitative insight into structure-activity relationship is needed for rational development of improved CU inhibitors in the future. A CoMFA study to understand the structure activity relationship for the CU class of HIV-1 protease inhibitors has been reported [8]. The authors used an approach where the ligands were built in the active pocket and further subjected to constrained minimization to identify the bioactive conformation for structural alignment in CoMFA. Due to such structural constraints, it is unlikely that a complete exploration of the active site would have been possible. This limitation can be addressed using molecular docking by which it is possible to search the active site more thoroughly. In an effort to gain deeper insight into the ligand-receptor interaction for the CU scaffold, we present here a 3D-QSAR study using both CoMFA and CoMSIA techniques on the series of CU analogs containing symmetric P2/P2′ substituents with HIV-1PR inhibitory activity [6]. Molecular docking was carried out to identify the probable binding conformation of these CU analogs; this also served as a guideline for the structural alignment. The advantage of using *docking-based alignment* over constrained minimization within the active site is that docking explores the binding pocket more extensively while constrained minimization may limit the conformational search to local minima. In addition, *atom-fit alignment* was also carried out to compare against the results obtained from *docking-based alignment*. The objective of using the CoMSIA approach was to understand the role of H-bonding and hydrophobic interactions which cannot be envisaged by

CoMFA alone. These additional fields provide better visualization and interpretation of the derived correlation in terms of field contribution to the activity of the compounds. The results obtained from 3D-QSAR analysis were further utilized to design new molecules whose activities were predicted to be manifold better than existing CU analogs. In silico prediction of pharmacokinetics and toxicity was carried out for the designed molecules to assess their metabolic stability and therapeutic safety.
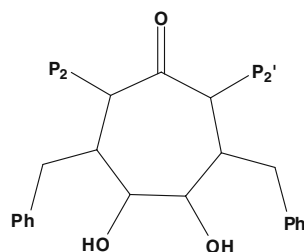
Computational details

The molecular modeling packages Sybyl (v7.1, Tripos Inc., USA) [9] running on a Pentium IV computer under the Linux RedHat Enterprise WS 4 and the GOLD Suite (CCDC, UK) [10] running on a Pentium IV computer under the Windows OS were used in this modeling study.

Ligand preparation

The set of HIV-1PR inhibitors used in the molecular modeling study was taken from literature [6]. The chemical structures of the 54 cyclic urea analogs and their experimental pK_i values are given in Table 1. The molecules span a 5 log unit activity range and molecular field descriptors have a good distribution across the range of values. Thus, the dataset was found to be appropriate for the purpose of 3D-QSAR analysis. In this study, the negative log of K_i (pK_i) was used, as it gives numerically a larger value for the active compound than that for the inactive compound. These pK_i values were used as the dependent variable in the 3D-QSAR calculations.

The 3D-structures of the molecules were sketched with the *Builder* module of Sybyl. Generally, the global minimum energy structure, assumed to be the bioactive conformation, is used in 3D-QSAR studies. Since in the present case, a crystal structure of HIV-1PR in complex with one of the inhibitors in the dataset (DMP323) is available, the 3D structures of the remaining molecules in the dataset were built using DMP323 as the template. The ligand geometries were then optimized by energy minimization using the Powel gradient method, the Tripos force field, Gasteiger Hückel charges and a distance dependent dielectric, till a gradient of 0.01 kcal mol^{-1} Å^{-1} was reached.

To derive statistically significant QSAR models, the dataset was split into training and test sets on the basis of structural, chemical and biological diversity using similarity search techniques [11] *viz.* D-Optimal design, Tanimoto similarity coefficient and the Euclidian distance matrix criteria defined in Cerius2 (Accelrys Inc., USA) [12]. The selection technique searches for diverse molecules by comparing their chemical nature in 2D space.

Table 1 Symmetric cyclic urea inhibitors of HIV-1PR

Mol ID	P2/P2′	pK$_i$	Mol ID	P2/P2′	pK$_i$
1	Methyl	5.24	28	2-picolyl	6.84
2	Ethyl	7.00	29	3-picolyl	8.01
3	*n*-propyl	8.10	30	4-picolyl	7.05
4	*n*-butyl	8.85	31	α-naphthylmethyl	7.07
5	*n*-pentyl	8.80	32	β-naphthylmethyl	9.51
6	*n*-hexyl	8.34	33	*o*-fluorobenzyl	7.47
7	*n*-heptyl	6.59	34	*m*-fluorobenzyl	8.52
8	CH$_2$CH$_2$OCH$_3$	6.10	35	*p*-fluorobenzyl	8.85
9	CH$_2$CH$_2$OCH$_2$CH$_3$	5.96	36	*o*-chlorobenzyl	6.62
10	CH$_2$CH$_2$OCH$_2$CH$_2$OCH$_3$	5.11	37	*m*-chlorobenzyl	9.05
11	*i*-butyl	7.31	38	*p*-chlorobenzyl	8.28
12	*i*-pentyl	7.92	39	*m*-bromobenzyl	8.85
13	*i*-hexyl	8.15	40	*p*-bromobenzyl	7.57
14	*i*-heptyl	7.52	41	*m*-methylbenzyl	8.15
15	*i*-octyl	6.96	42	*p*-methylbenzyl	8.24
16	Neohexyl	7.44	43	*m*-(trifluoromethyl)benzyl	7.66
17	Allyl	8.28	44	*p*-(trifluoromethyl)benzyl	7.29
18	2-methylpropen-3-yl	8.14	45	*o*-methoxybenzyl	5.73
19	Isoprenyl	8.74	46	*m*-methoxybenzyl	8.80
20	CH$_2$CH$_2$OCHCH$_2$	7.22	47	*p*-methoxybenzyl	6.80
21	3-propynyl	7.66	48	*m*-nitrobenzyl	8.55
22	Cyclopropylmethyl	8.68	49	*m*-iodobenzyl	9.38
23	Cyclobutylmethyl	8.89	50	*p*-(hydroxymethyl)benzyl (DMP323)	9.47
24	Cyclopentylmethyl	8.37	51	*m*-(hydroxymethyl)benzyl	9.85
25	Cyclohexylmethyl	7.43	52	*p*-hydroxybenzyl	9.92
26	N-morpholino-2-ethyl	5.40	53	*m*-hydroxybenzyl	9.92
27	Benzyl	8.52	54	*p*-(HOCH$_2$)benzyl (enantiomer of DMP323)	5.78

Receptor preparation

The crystal structure of HIV-1 protease in complex with the inhibitor DMP323 was taken from the Protein Data Bank (PDB entry 1 qbs). The catalytically active protease is a homodimer of two identical subunits with 99 amino acid residues, where each monomer contributes one of the two catalytic aspartic acid residues in the active site. The active site exists in a cleft at the interface of the two monomers. The formation of the active site of aspartyl proteases by the union of two identical subunits is unique to retroviruses. The water molecules present in the protein-inhibitor complex were deleted during docking since no water molecule was found to be conserved in the different crystal structures. The atom types were corrected and hydrogen atoms then added to the protein.

Docking protocol

The program GOLD was used to explore the probable binding conformation of the cyclic urea analogs in the active site of the target protein. The program uses a genetic algorithm to search the configuration space in the active site. The docking protocol was validated by reproducing the

pose of the CU analog DMP323 in its crystal structure (PDB entry 1 qbs) [13]. GOLD parameters optimized for the docking protocol were: (a) the dihedral angles of the rotatable bonds in the ligand; (b) geometries of the ligand ring (flipping ring corners); (c) the dihedral angles of protein OH and NH_3^+ groups; and (d) mapping of the H-bond fitting points. All these variables were randomized at the start of a docking run.

The docking protocol was then implemented on the remaining molecules in the dataset. The study was carried for 20 GA runs, which was found to be optimum to reproduce the pose of DMP323 in its crystal. The other GA parameters *viz.* the population size and the genetic operators were kept at their default values. All molecules in the dataset were docked into the receptor active site shaped by residues in a 7.0 Å vicinity of the ligand. This active site comprised of Arg8, Leu23, Asp25, Thr26, Gly27, Ala28, Asp29, Asp30, Thr31, Val32, Lys45, Met46, Ile47, Gly48, Gly49, Ile50, Gly51, Phe53, Leu76, Thr80, Pro81, Val82, Asn83, Ile84, and Gly86 residues from chains A and B thus forming a symmetric binding site.

CoMFA and CoMSIA studies

In the present study, two 3D-QSAR methodologies, CoMFA and CoMSIA, were adopted to understand the local physicochemical properties involved in the ligand-receptor interaction. Since its advent in 1988, comparative molecular field analysis (CoMFA) [14] has been cherished as one of the most valuable 3D-QSAR methods. A 3D cubic lattice with a grid spacing of 1 Å and extending 4 Å beyond the aligned molecules in all directions was created to encompass the aligned molecules. CoMFA fields were generated using the sp^3 carbon probe with a +1 charge to calculate steric (Lennard–Jones 6–12 potential) field energies and electrostatic (Coulombic potential) fields with a distance-dependent dielectric at each lattice point. The fields were scaled by the CoMFA-Standard method available in Sybyl. The default cut-off for energy greater than 30 kcal mol^{-1} was applied; the electrostatic values were also dropped for each row where the steric cut-off was reached. No smoothing functions were employed.

Comparative Molecular similarity indices analysis (CoMSIA) [15] calculates similarity indices in the space surrounding each of the molecules in the dataset, that have previously been aligned by some methodology. The CoMSIA approach circumvents some of the inherent deficiencies arising from the functional form of the Lennard-Jones and Coulombic potentials used in CoMFA. CoMSIA employs a Gaussian type distance dependence function between the probe and the atoms of the molecules; this avoids singularities at the atomic positions and the dramatic changes in the potential energy for threshold grids

in the proximity of the surface. This functional form requires no arbitrary definition of cut-off limits, and the similarity indices can be calculated at all grid points inside and outside the molecule. The CoMSIA technique calculates five different similarity index fields (steric, electrostatic, hydrophobic, hydrogen bond donor, and hydrogen bond acceptor) with a similar probe atom as CoMFA. The objective of using these additional fields is not to improve the predictive power of the 3D-QSAR models but to partition the various properties into spatial locations where they play a vital role in determining the binding affinity. The major advantage of CoMSIA lies in better visualization and interpretation of the correlations that have been obtained in terms of field contributions. Strictly speaking, the contours obtained from regression analysis indicate those lattice points in space where a particular property contributes significantly to the variance in binding affinity.

The contour maps derived from both the CoMFA and CoMSIA models permit an understanding of the steric, electrostatic, hydrophobic, and hydrogen bonding requirements for ligand binding. Some information about the binding domain of the receptor can also be gleaned from the contour map.

Structural alignment

The most critical step in the CoMFA and CoMSIA procedures is the relative alignment of all ligands, such that they have comparable conformations and orientations in space. Both *receptor-based* and *atom-fit molecular alignment* strategies were applied to derive the 3D-QSAR models.

Receptor-based alignment

An underlying postulate in 3D-QSAR analyses is that all molecules in the dataset bind to the receptor in a similar way. The docking solutions with the best scores, and conformations similar to the ligand in the crystal structure were overlaid to derive the *receptor-based molecular alignment*. This approach is also termed as the *docking-based alignment* [16, 17]. The alignment based on this strategy is shown in Fig. 1(a).

Atom-fit alignment

The *atom-fit alignment* procedure is based on the best matching of preselected atoms. The cyclic urea ring atoms were used as the basis for the alignment of all compounds in the series. The molecules in the dataset were aligned using the "Database Align" option available in Sybyl. The alignment based on *atom-fit* strategy is shown in Fig. 1(b).
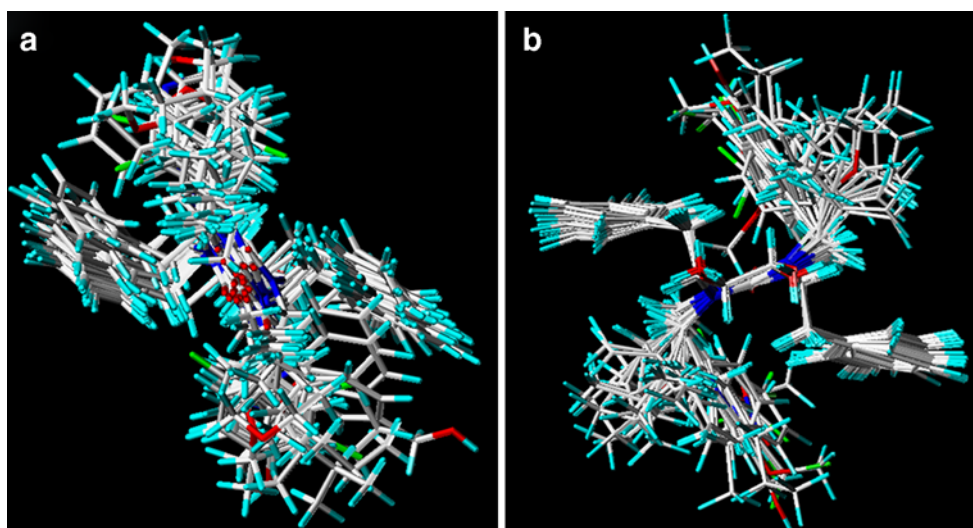
Both the alignment strategies produced highly reliable CoMFA and CoMSIA models. In contrast to CoMFA, CoMSIA is believed to be less sensitive to changes in molecular alignment and provides smooth and interpretable contour maps as a result of the Gaussian type function used to calculate the molecular similarity indices.

Statistical analysis

The CoMFA and CoMSIA field energies were used as independent variables while the pK_i values formed the dependent variables. A partial least squares (PLS) [18] regression was then run to derive the 3D-QSAR models. The predictive ability of the derived statistical models was evaluated using the "leave-one-out" (LOO) cross-validation procedure [19, 20]. In this method, each compound is systematically excluded from the data set and its activity predicted by a model that is deduced from the remaining compounds. In order to minimize the possibility of over-fitting data, the number of components corresponding to the lowest PRESS value was used to derive the final PLS regression models. SAMPLS [21] with leave-one-out (LOO) validation and no column filtering were used for cross-validation calculations to determine the q^2 (r^2_{cv}) and standard error of prediction (SEP). The PLS analysis was repeated without cross-validation with the optimum number of components to determine the conventional correlation coefficient r^2, the standard errors (SE) and the F-value. The cross-validated correlation coefficient (q^2), was calculated using Eq. 1

$$r^2_{cv} = (\text{PRESS}_0 - \text{PRESS})/(\text{PRESS}_0) \qquad (1)$$

where PRESS_0 is the average of observed biological activity over the dataset; PRESS [22] is the sum of squares

of the differences between the actual and the predicted activity and is calculated as

$$\text{PRESS} = \sum (Y_{obs} - Y_{pred})^2 \qquad (2)$$

where, Y_{obs} and Y_{pred} are actual and predicted values of the biological activity respectively.

The robustness and statistical confidence of the 3D-QSAR models were evaluated by cross-validation with the bootstrap formalism. Bootstrap analysis [23] involves simulating a large number of datasets that are of the same size as the original. These datasets are obtained by randomly choosing samples from the original dataset. The statistical calculation is performed on each of these bootstrap samplings. The difference in the parameters calculated from the original data and the average of the parameters obtained from the N bootstrap runs is a measure of the bias of the original calculation.

To evaluate the confidence limits of the generated PLS models to small systematic perturbations of the response variable, the technique of y-scrambling [24] was used. It is a non-parametric approach that helps determine the sensitivity of the QSAR model to chance correlations, without disturbing the underlying covariance structure of the data. The test involves scrambling the biological data (150 trials) and deriving the model once again.

Predictive r^2 (r^2_{pred})

The predictive ability of the CoMFA and CoMSIA models is expressed by predictive r^2 value [25], which is related to the cross-validated r^2 (q^2) and is computed using the formula

$$r^2_{pred} = (\text{SD} - \text{PRESS})/\text{SD} \qquad (3)$$

where SD is the sum of squared deviations between the biological activities of the test set and the mean activity of

the training set; while PRESS is the sum of squared deviations between the actual and the predicted activities of the test set molecules. The r^2_{pred} is based solely on molecules included in the test set.

Results and discussion

Docking analysis

The outcome of the docking studies was found to be in harmony with report for the cyclic urea analog DMP323 co-crystallized with the HIV protein (PDB code: 1 qbs) [6]. The root-mean-square deviation (RMSD) of the conformation of DMP323 determined by docking with the experimental conformation was found to be 0.49 Å (Fig. 2). This means that the docking procedure will reliably reproduce the receptor bound conformation of other molecules in the dataset. The best docked conformation of the most active molecule 52 in the HIV-1PR active site is shown in Fig. 3. In contrast to the flipped conformation adopted by the cyclic sulfamide HIV-1PR inhibitors in the binding site, a reasonably symmetric orientation was observed for the cyclic urea class of compounds. The P1/P1′ side chain occupied the S1/S1′ pockets while the P2/P2′ side chain occupied the S2/S2′ pockets. When these molecules were aligned over one another, their functional elements superimposed quite well. All the crucial interactions observed between HIV-1PR and DMP323 in the crystal structure were also reproduced in the most active compound 52, suggesting a fairly consistent binding mode for all the CU analogs (Fig. 4). The contacts between the backbone atoms of all the inhibitors and the enzyme were consistent; in particular, the hydrogen bonding interactions between the catalytic residues (Asp 25/25′, Asp 29/29′, Asp 30/30′ and Ile 50/50′) and the inhibitor were conserved. The poses of
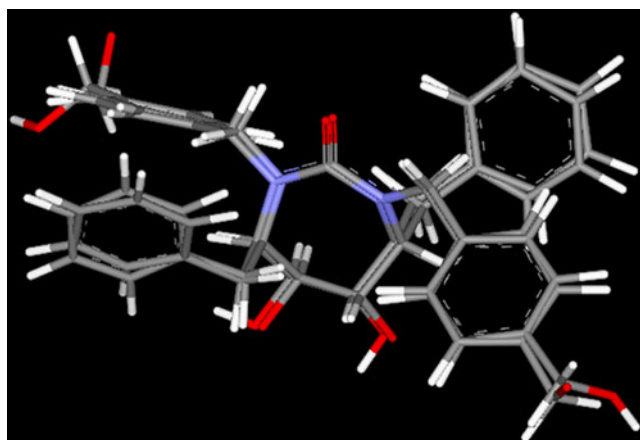

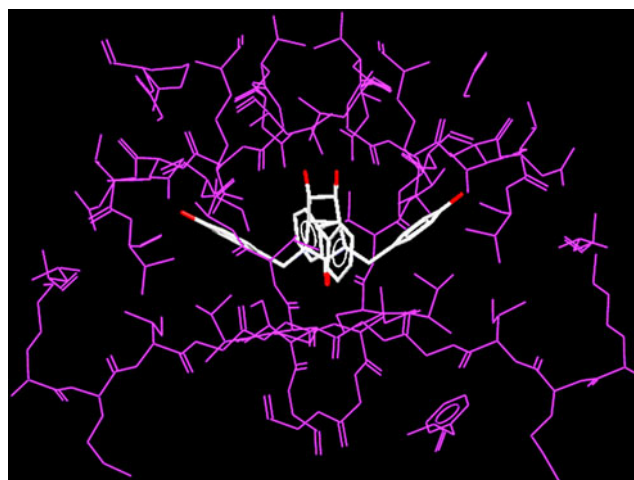Fig. 2 Overlay of the X-ray pose of DMP323 over its best docked conformation

Fig. 3 Binding mode of molecule 52 in the HIV-1PR active site

the ligands with good docking scores were analyzed, and those that matched the crystal structure were used to generate the molecular alignment.

3D-QSAR analyses

Various 3D-QSAR models were generated and the best one was selected based on its statistical parameters. The statistical parameters for the 3D-QSAR models generated by *receptor-based* and *atom-fit alignment* strategies are given in Table 2.

CoMFA analysis

Thirty one of the total 54 CU analogs constituted the training set and the remaining 23 compounds formed the test set. These two sets of molecules were used to derive and validate the CoMFA model based on *receptor-based alignment*. The corresponding number of molecules in the training and test sets was 37 and 17 respectively for the *atom-fit aligned* CoMFA model. The training and test set compounds were selected such that the structural diversity and spread of activity in the dataset was maintained.

For the CoMFA model generated using *receptor-based alignment*, partial least squares (PLS) regression produced a cross-validated correlation coefficient q^2 of 0.68 with five components. The non cross-validated PLS analysis produced a correlation coefficient (r^2) of 0.96, F value of 192.23 and an estimated standard error (SE) of 0.21. The steric field descriptors explain 90% of the variance, while the electrostatic descriptors explain 10% of the variance, suggesting that the contribution of the steric field predominates. Robustness of the model was judged from the bootstrap analysis. The bootstrap analysis produced a correlation coefficient (r^2_{bs}) of 0.98 which upholds the statistical validity of the CoMFA model. Y-scrambling was

Fig. 4 Schematic representation of the interactions of molecule 52 in the HIV-1PR active site
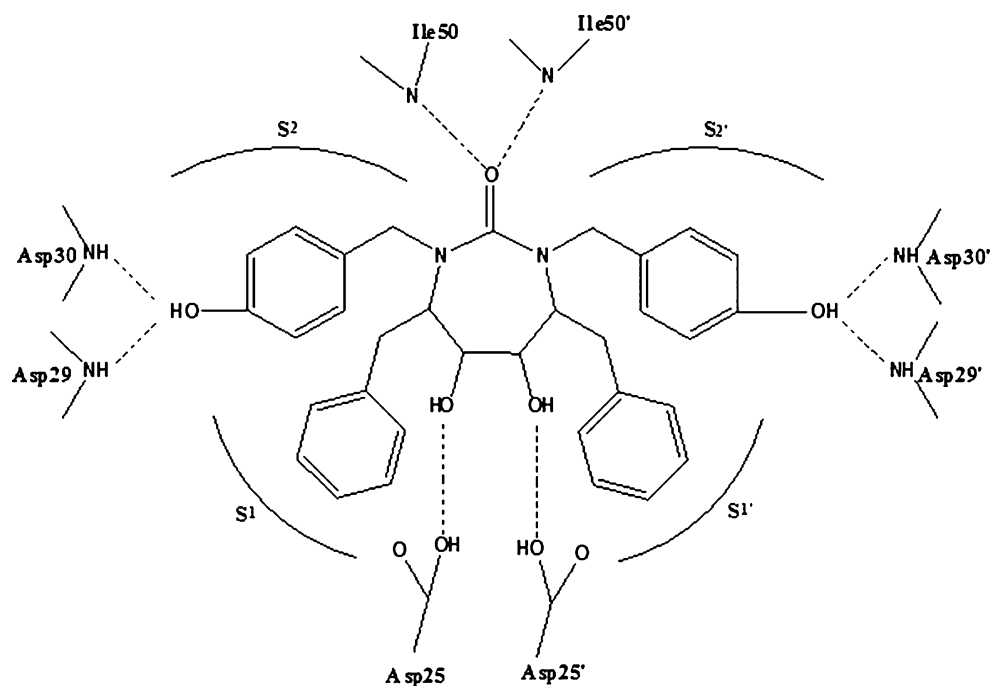
carried out to eliminate the possibility of chance correlation. In all cases, negative q^2 values with an average of -0.27 and an average correlation coefficient r^2 of 0.39 was obtained. Thus the 3D-QSAR models were found to be significantly better than the y-scrambled models.

For the CoMFA model derived using *atom-fit alignment*, the cross-validated correlation coefficient (q^2) was found to be 0.63, while the non cross-validated PLS analysis with six components produced a conventional r^2 of 0.93, F value of 166.64, an estimated standard error (SE) of 0.31, bootstrap r^2 (r^2_{bs}) 0.95 and y-scrambled r^2 0.19. These

values signify a good statistical correlation and that the CoMFA model has satisfactory linearity. The respective relative contributions of steric and electrostatic fields were 87% and 13%. The statistics of the model generated by the *atom-fit alignment* was not better than the *receptor-based alignment*.

CoMSIA analysis

The CoMSIA analysis gives an additional structural insight into the putative binding sites of the ligand-receptor

Table 2 Statistical results of the CoMFA and CoMSIA analyses

PLS statistics	Receptor-based alignment		Atom-fit alignment	
	CoMFA	CoMSIA	CoMFA	CoMSIA
N	54	54	54	54
q^2	0.68	0.65	0.63	0.67
r^2	0.96	0.94	0.93	0.90
r^2_{pred}	0.69	0.73	0.66	0.54
r^2_{bs}	0.98	0.97	0.95	0.92
r^2 y-scrambling	0.39	0.28	0.19	0.18
F	192.23	174.28	166.64	146.46
SE	0.21	0.30	0.31	0.24
PLS components	5	6	6	6
Field contribution				
Steric	0.90	0.23	0.87	0.74
Electrostatic	0.10	0.17	0.13	0.08
Hydrophobic				
H-bond donor		0.28		0.18
H-bond acceptor		0.32		

complex. The CoMSIA analysis was performed using the same structural alignment and same training and test sets as defined in the CoMFA studies. The CoMSIA isopleths represent the most relevant molecular interaction fields, namely - steric, electrostatic, hydrophobic, H-bond donor and H-bond acceptor. The value of the attenuation factor (α) was set to 0.3.

For the CoMSIA model generated using *receptor-based alignment*, statistical parameters showed that steric, electrostatic, H-bond donor and H-bond acceptor features significantly influence the activity of the compounds. The CoMSIA model generated from a combination of these fields had a q^2 of 0.65 with six components, non cross-validated r^2 0.94, standard error of prediction 0.30, F value 174.28 and bootstrap r^2 0.94. Y-scrambling yielded an average q^2 of -0.25 and an average r^2 of 0.28 indicating a very low probability of chance correlation. The contributions of the steric, electrostatic, hydrogen bond donor and acceptor fields to this model were 22.6, 16.7, 28.6 and 32.1%, respectively.

The CoMSIA model based on *atom-fit alignment* was derived using steric, electrostatic and hydrogen bond donor fields. Partial least square (PLS) regression produced a model with cross-validated correlation coefficient (q^2) of 0.67, conventional correlation coefficient (r^2) of 0.90 with optimum number of components as 6, standard error of prediction 0.24 and F value 146.46. The contribution of steric, electrostatic and hydrogen bond donor fields to the HIV-1PR inhibitory activity was found to be 74.5, 7.5 and 18.0% respectively. Bootstrap analysis for 100 runs produced an r^2_{bs} value of 0.92 while Y-scrambling produced an average q^2 value of -0.17 and r^2 of 0.18 confirming the robustness of the derived CoMSIA model. It can be seen that the CoMSIA model generated with *receptor-based* alignment could explain the variance in activity better than the model generated with *atom-fit alignment*.

Predictivity of CoMFA and CoMSIA models

The 3D-QSAR models were assessed for their predictive abilities by the test set molecules not included in the development of the models. External validation is considered to be the most acceptable validation method for a predictive QSAR, since the test set molecules are completely neglected during the training of the model. The predicted pK$_i$ values were found to be in good agreement with the experimental values within the statistically tolerable error range. The predictive correlation coefficient r^2_{pred} for CoMFA and CoMSIA models derived from *receptor-based alignment* was 0.69 and 0.73 respectively. The value is 0.66 and 0.54 respectively for the CoMFA and CoMSIA models derived using *atom-fit alignment*. The good agreement between experimental and predicted values

for the test set molecules suggests that these CoMFA and CoMSIA models are reliable and can be used in the design of novel HIV-1PR inhibitors within this structural class. No outliers were detected in the analysis. The predicted and experimental pK$_i$ values for the test set molecules are shown in Tables 3 and 4 and the plots for the same are shown in Fig. 5(a)–(d).

The higher values of the predictive r^2 (r^2_{pred}) obtained for models constructed using *receptor-based alignment* is an indication that the *receptor-based alignment* can effectively account for the interactions these molecules make with the enzyme active site; this may not always be demonstrated by the *atom-fit alignment* strategy.

Contour analysis

The results of the CoMFA and CoMSIA models were visualized through contour maps. These maps show regions in space where variation in specific molecular properties increase or decrease potency. The default values of 80% contribution for favored and 20% for disfavored regions were set for visualization of the contour maps. The molecular fields around the most active compound 52

Table 3 The experimental and predicted pK$_i$ values for the test set molecules using *receptor-based alignment*

Molecule ID	Experimental pK$_i$	Predicted pK$_i$ (CoMFA)	Predicted pK$_i$ (CoMSIA)
3	5.24	6.34	6.13
9	8.10	8.10	8.04
17	7.31	6.86	7.12
18	7.92	7.78	7.03
19	8.15	6.99	7.35
20	7.52	7.15	7.43
21	6.96	7.04	6.99
23	8.28	8.34	8.46
25	8.74	8.38	8.23
27	7.66	7.09	8.48
32	5.40	6.67	4.98
35	8.01	7.64	7.41
37	7.07	7.19	7.19
38	9.51	8.91	9.10
40	8.52	8.58	8.04
43	9.05	8.92	8.67
45	8.85	8.11	8.54
48	8.24	8.46	8.08
49	7.66	7.17	7.06
54	8.55	8.37	8.18
56	9.47	8.87	8.70
57	9.85	9.41	8.94
61	5.78	7.62	7.42

Table 4 The experimental and predicted pK$_i$ values for the test set molecules using *atom-fit alignment*

Molecule ID	Experimental pK$_i$	Predicted pK$_i$ (CoMFA)	Predicted pK$_i$ (CoMSIA)
17	7.31	8.24	7.88
18	7.92	7.98	7.23
20	7.52	7.56	6.58
23	8.28	8.00	6.63
26	7.22	6.41	7.83
27	7.66	7.17	8.58
33	8.52	7.99	6.75
37	7.07	7.61	7.04
38	9.51	8.23	7.96
40	8.52	8.44	8.02
42	6.62	7.15	8.14
43	9.05	8.95	7.47
45	8.85	8.41	8.78
48	8.24	7.61	9.15
49	7.66	7.61	7.98
56	9.47	8.71	8.52
61	5.78	6.23	5.99

(Table 1) is displayed in Figs. 6 and 7. The contour maps have been shown only for the CoMFA and CoMSIA models derived with *receptor-based alignment* for the sake of brevity. These contour maps are important tools in drug design, as they show regions in 3D space where modifications of steric and electrostatic fields strongly correlate with concomitant changes in biological activity.

CoMFA contour maps

The steric contours observed for the CoMFA model generated by *receptor-based alignment* (Fig. 6a) are more or less similar to the steric contours of the model derived with *atom-fit alignment*. The steric interactions are represented by green and yellow isopleths, where increase in activity is associated with more bulk near the green and less bulk near the yellow contours. Large green contours were observed around the P2/P2′ side chain, indicating that steric bulk at these positions is important for HIV-1PR inhibitory activity. Published X-ray structures [6] have also shown that the S2/S2′ pockets are essentially large. Evidence for this comes from molecules 49, 50, 51, 52, 53 (Table 1) which have bulky aromatic groups at the P2/P2′ site. Their
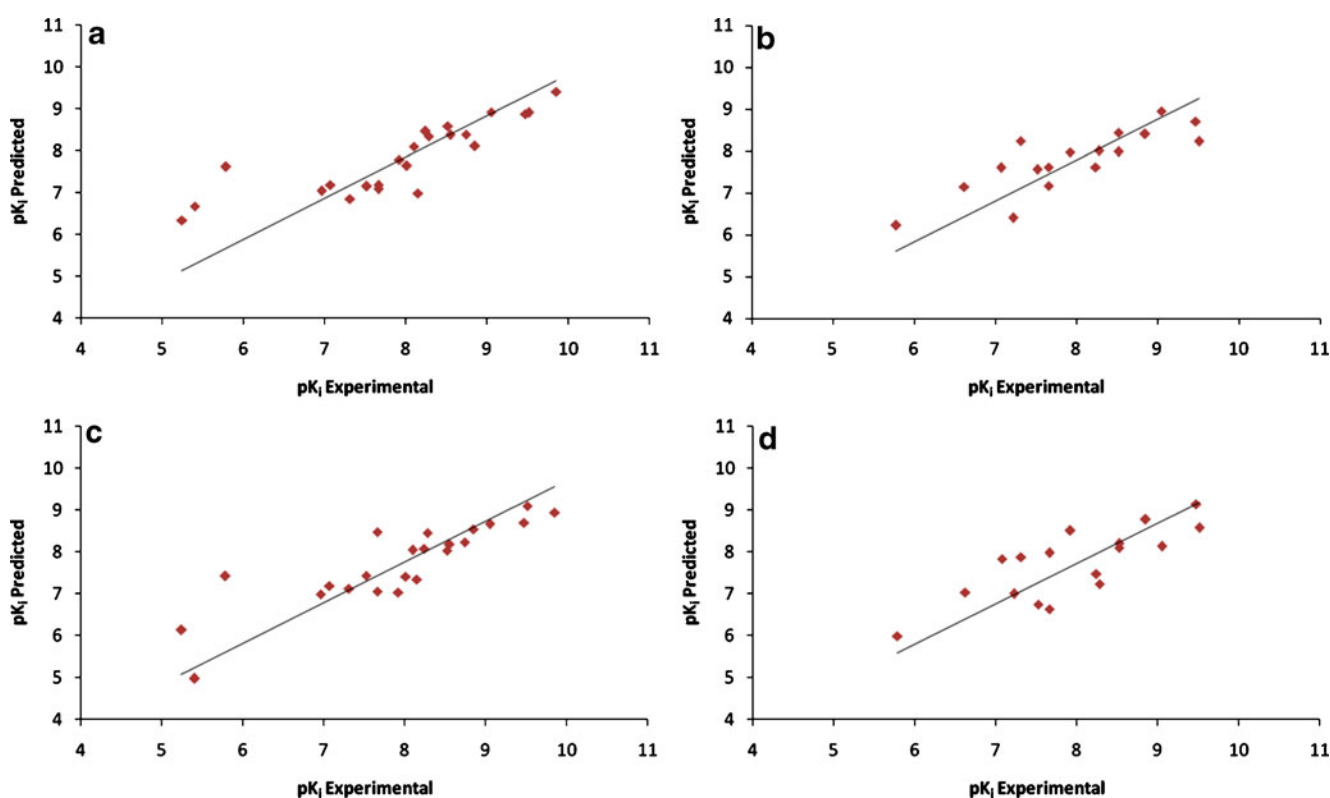

Fig. 5 Plot of the experimental vs. predicted pK$_i$ values for the test set molecules (**a**) based on the CoMFA model derived by *receptor-based alignment* (**b**) based on the CoMFA model derived by *atom-fit alignment* (**c**) based on the CoMSIA model derived by *receptor-based alignment* (**d**) based on the CoMSIA model derived by *atom-fit alignment*

Fig. 6 The CoMFA molecular interaction fields (*receptor-based alignment*) around molecule 52 (a) steric contours - favored (green); disfavored (yellow) (b) electrostatic contours - electropositive (blue); electronegative (red)
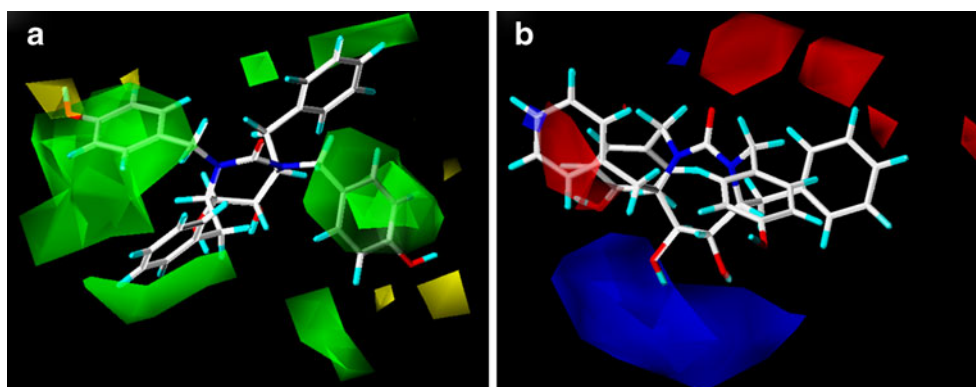

activities are higher compared to molecules 1, 2, 3, 7, 8, 9, 10, 11 and 12 which have smaller substituents (Table 1) at the P2/P2′ site. However, small yellow contours were also observed beyond the green isopleths at the *para*-position of the P2/P2′ phenyl ring. This indicates that although steric substitution is favored around P2/P2′ side chain, extending the side chain beyond a certain length may be detrimental for activity. It is noteworthy that green contours associated with steric substitution were also observed around the P1/P1′ side chains. This means that there is a possibility for further structural modification to improve the inhibitory activity. These green isopleths around P1/P1′ side chains were not observed in the steric map of the CoMFA model generated by *atom-fit alignment*.

The CoMFA electrostatic contour plots are represented by blue and red isopleths. A blue contours suggests that substituents should be electron deficient while red contours indicate that substituents should be electron rich to improve

activity. The electropositive contours were found to dominate the CoMFA model over its electronegative counterpart. The electrostatic contour maps for the CoMFA model generated by *receptor-based alignment* (Fig. 6b) as well as the model generated by *atom-fit alignment* showed prominent blue contours in the vicinity of the ethylene diol moiety suggesting that electropositive substituents in this region would favor potency. Additional blue contours were also observed around the *para* positions of P1/P1′ side chains in case of the *atom-fit* aligned model. Red contours favoring electronegative substitutions were seen around the carbonyl oxygen of the cyclic urea scaffold in the case of the *receptor-based* aligned model. This suggests that this carbonyl oxygen could be replaced with another electronegative substituent like sulfur. Red contours were also observed around the *ortho* and *meta* positions of the P1/P1′ phenyl rings in the case of both *atom-fit* as well as *receptor-based* aligned models.

Fig. 7 The CoMSIA molecular interaction fields (*receptor-based alignment*) around molecule 52 (a) steric contours - favored (green); disfavored (yellow) (b) electrostatic contours - electropositive (blue); electronegative (red) (c) hydrogen bond donor contours – favored (cyan); disfavored (purple) (d) hydrogen bond acceptor contours – favored (magenta); disfavored (red)
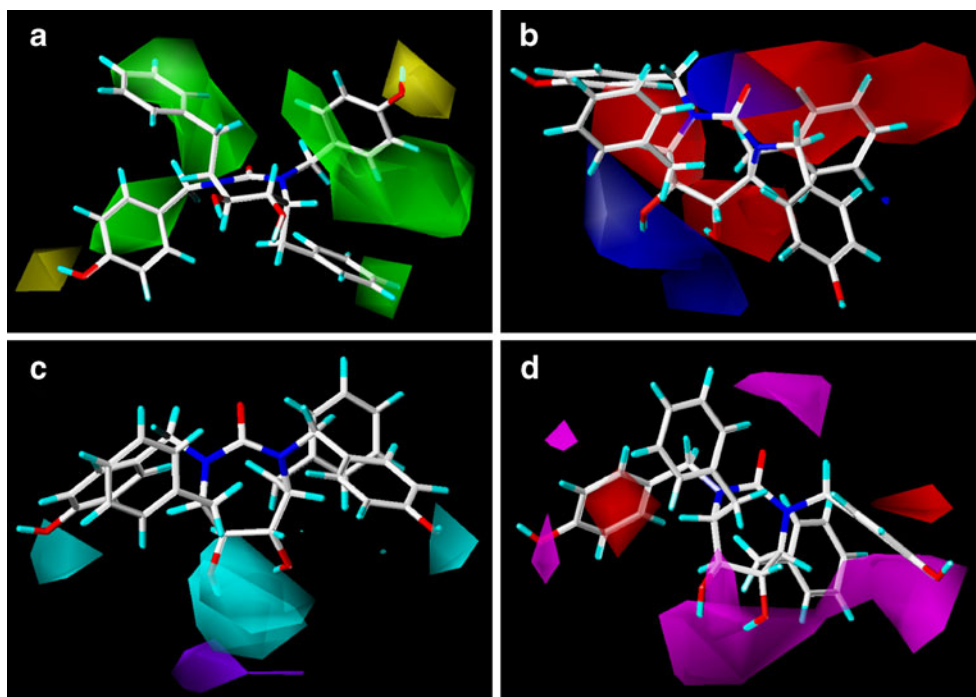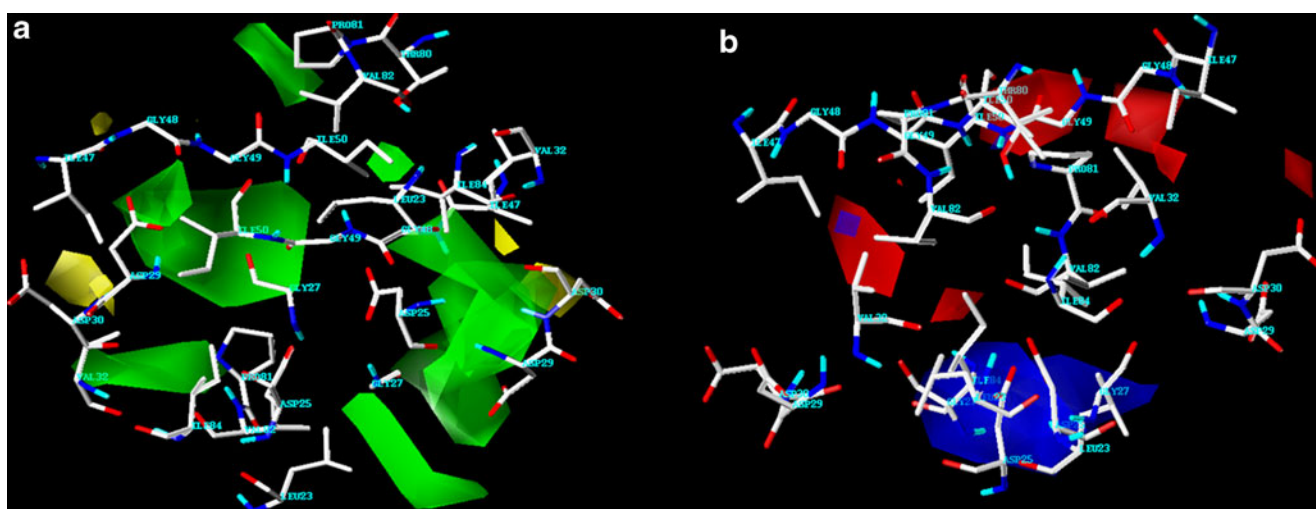
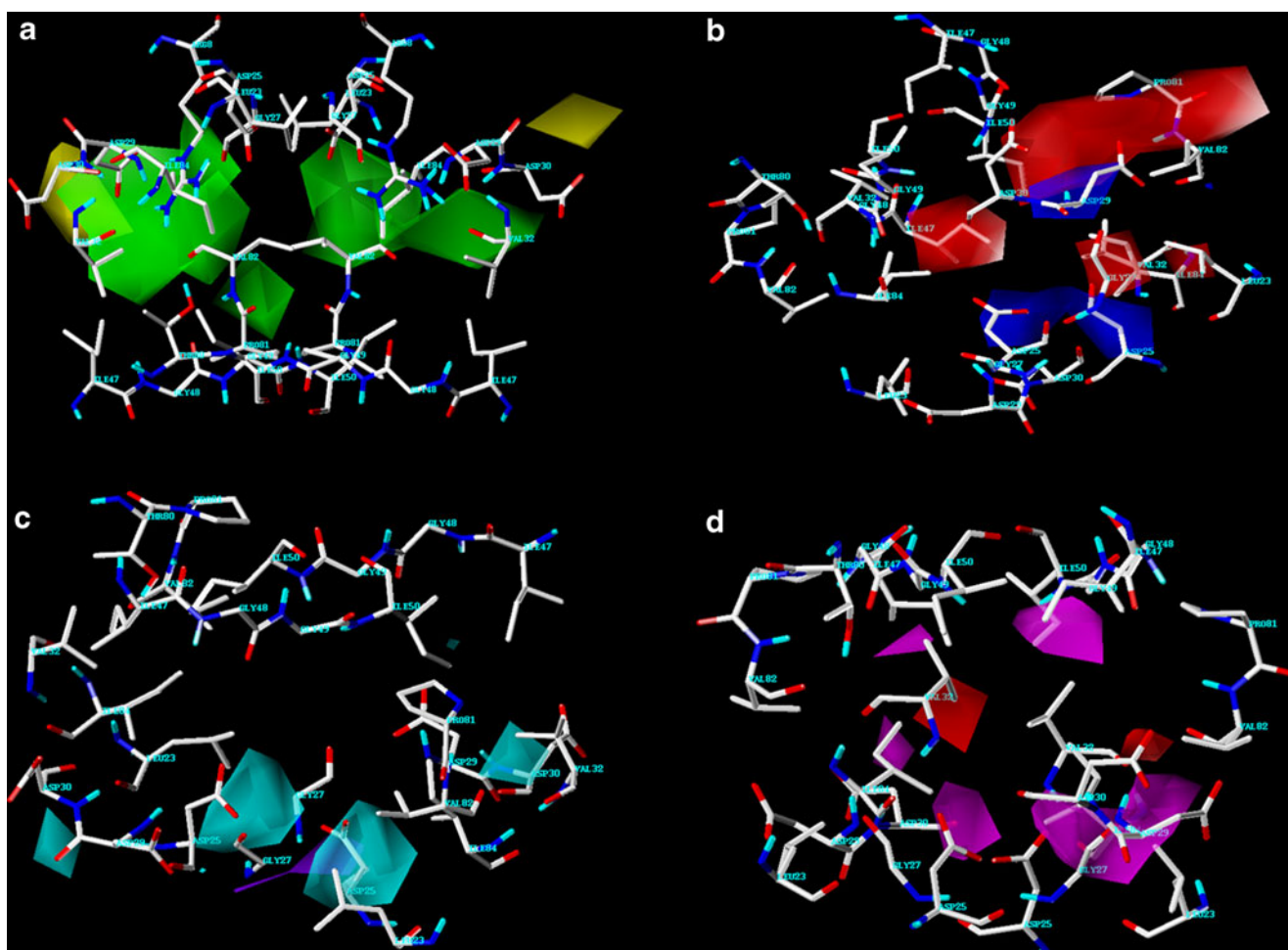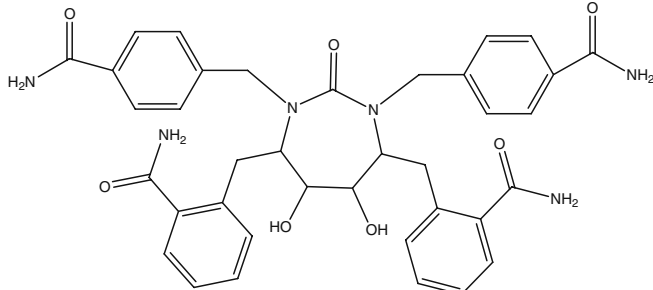**Fig. 8** CoMFA steric (green/yellow), electrostatic (red/blue), fields projected in the active site of HIV-1PR
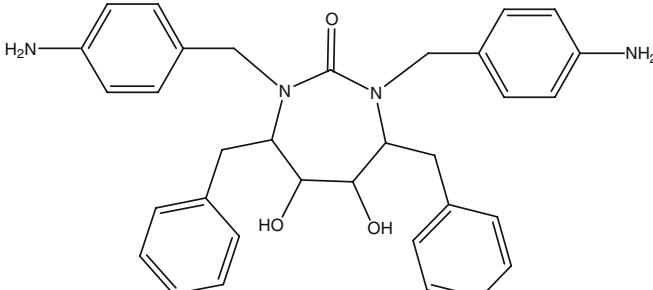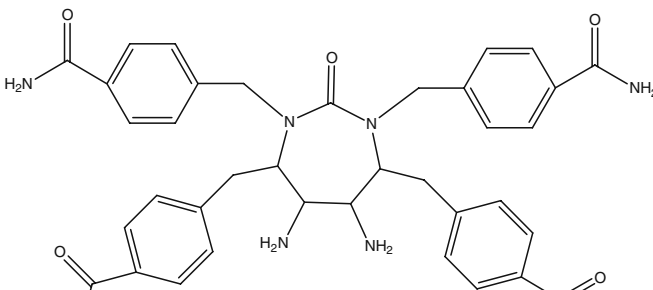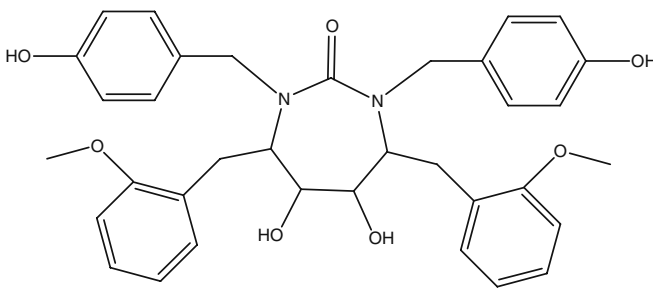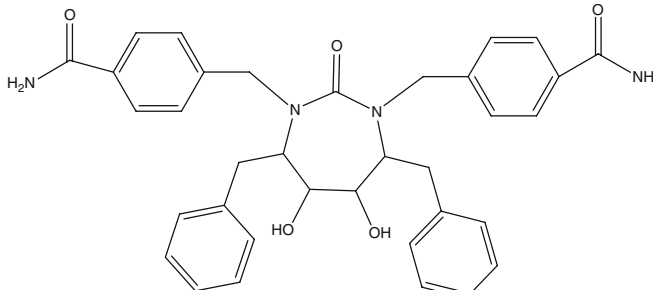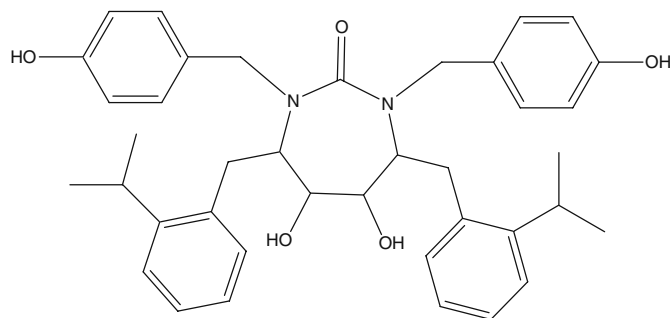

Fig. 9 CoMSIA steric (green/yellow), electrostatic (blue/red), H-bond donor (cyan/ purple) and H-bond acceptor (magenta/ red) fields projected in the active site of HIV-1PR

Table 5 HIV-1PR inhibitors that have been designed based on the QSAR results and their predicted activities
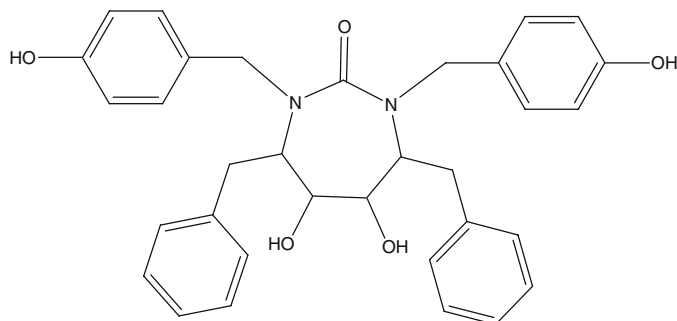
Molecule ID	Structure	Predicted pK$_i$
01		12.32
02		11.39
03		11.10
04		11.13
05		11.55

06

11.48

07

11.29

08

11.33

09

12.68

10   12.403

11   11.84

12   12.12

13   10.95
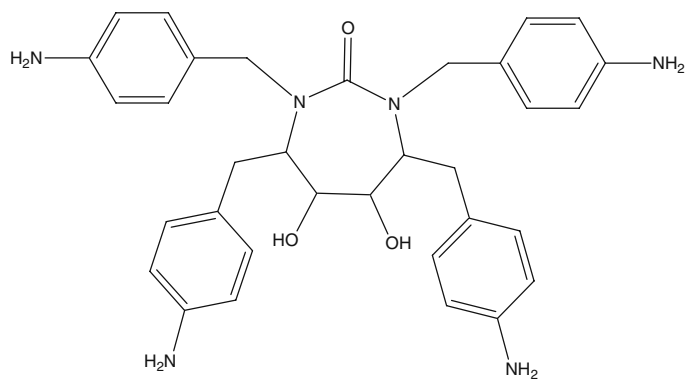
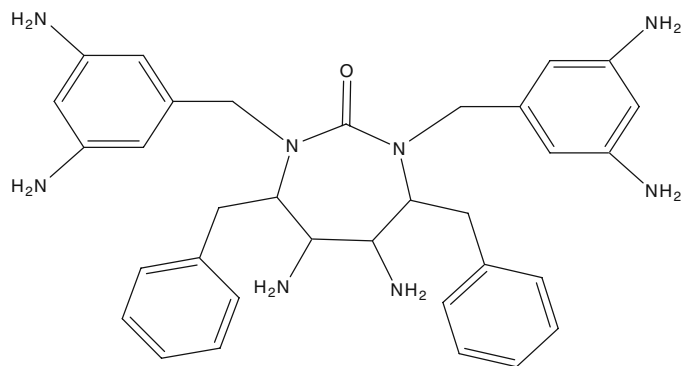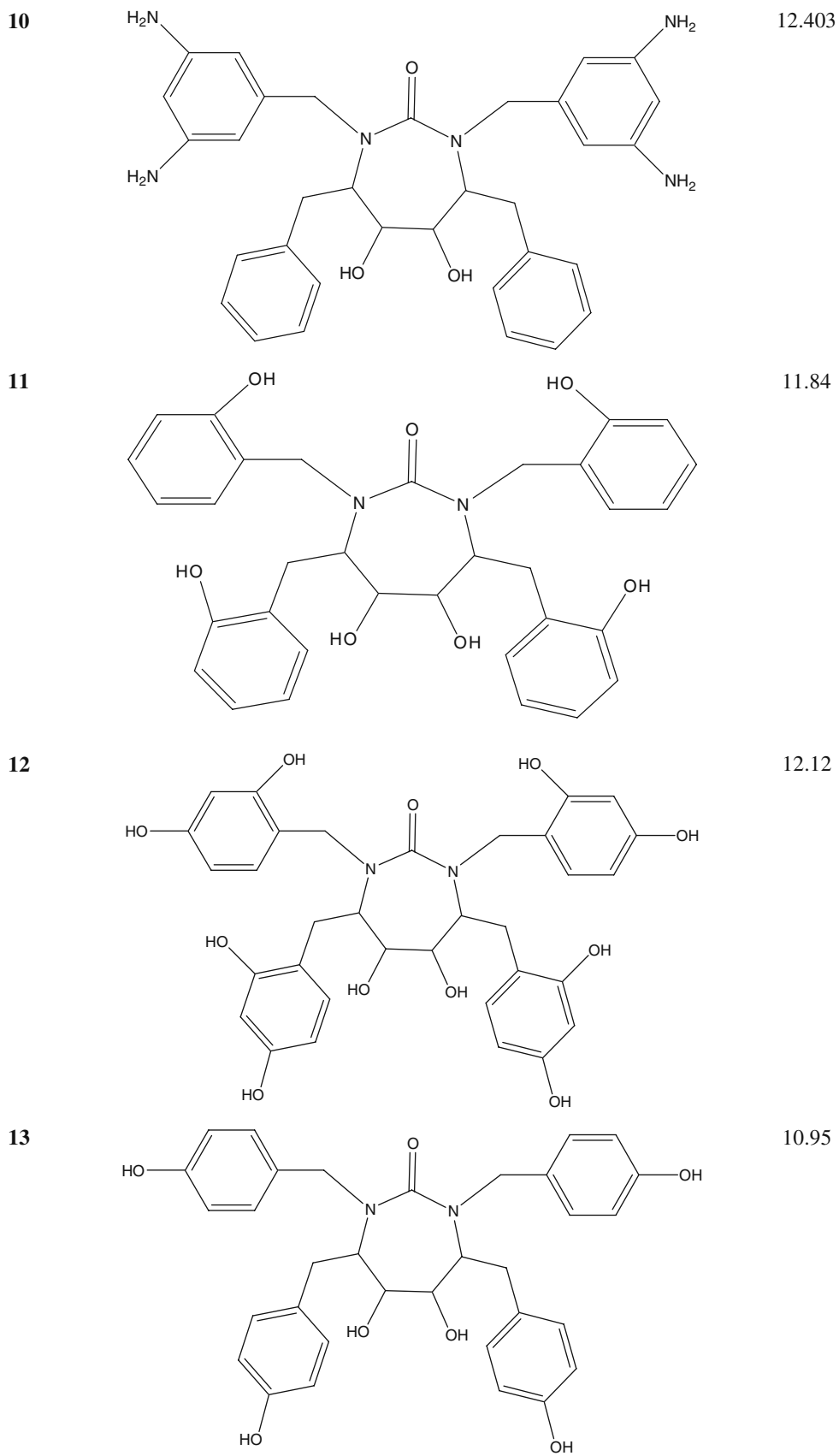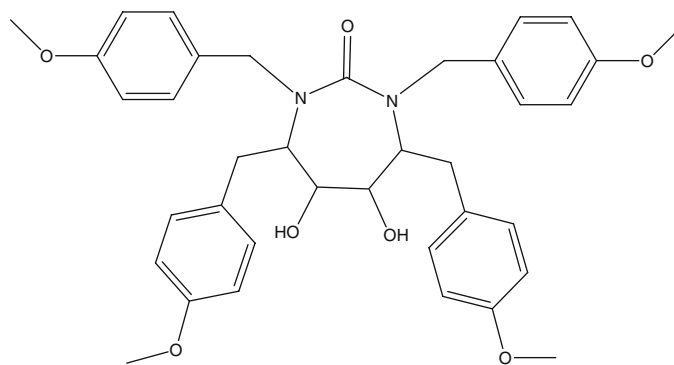**14**   10.355

CoMSIA contour maps

The CoMSIA model gives additional structural insights into the ligand-receptor interactions. It is similar to CoMFA, except that in addition to the steric and electrostatic fields, CoMSIA defines explicit hydrophobic and hydrogen bond donor–acceptor fields, which are not available with standard CoMFA analysis. The most active compound 52 (Table 1) is displayed surrounded by the CoMSIA fields, as an example.

Four fields, *i.e.*, steric, electrostatic, H-bond donor and acceptor (Fig. 7) were found to contribute significantly to the CoMSIA model derived using *receptor-based alignment*. On the other hand, the best statistical results for the CoMSIA model derived with *atom-fit alignment* were obtained with steric, electrostatic, and H-bond donor fields. There is a great correspondence between the CoMSIA contour maps and the CoMFA maps. In harmony with the CoMFA steric contours, large green isopleths were observed over the P2/P2′ side chains (Fig. 7a). Additional yellow contours were observed around the *para*-position of the P2/P2′ phenyl rings in the case of the *receptor-based* aligned model but were not seen in the *atom-fit aligned* model.

The CoMSIA electrostatic contours (Fig 7b) showed prominent red contours that favor electronegative substitution around the phenyl rings of the P1/P1′ side chains. Blue colored contours favoring electropositive substitutions were observed around the ethylene diol moiety of the cyclic urea scaffold. Blue contours were also observed around the *ortho* position of the phenyl rings of the P1/P1′ side chains.

Hydrogen bond donor isopleths are represented by cyan and purple contours (Fig. 7c). Cyan isopleths were observed around the vicinal diol and around the *para* positions of the P2/P2′ side chain. This means that incorporating hydrogen bond donor functionalities around these positions would favor activity. This is in accord with

the fact that aspartic acid residues (Asp30/30′, Asp29/29′ and Asp25/25′) present in the receptor adjacent to the P2/P2′ side chain and near the vicinal diol are involved in hydrogen bond interaction. Hydrogen bond acceptor contours are represented by magenta and red isopleths, where magenta indicates regions that favor hydrogen bond acceptor substituents, while red isopleths represent regions in the molecule that disfavor hydrogen bond acceptor substituents. Magenta contours were observed around the vicinal diol and the *para*-position of P2/P2′ side chain of the inhibitor. This implies that these regions could be substituted with functionalities capable of accepting hydrogen bonds (Fig. 7d). Small magenta isopleths were also observed around the carbonyl group of the cyclic urea scaffold suggesting a hydrogen bond acceptor group at this position would favor activity. These contours are complementary to the crystal structure where it was observed that this carbonyl moiety is involved in hydrogen bonding with isoleucine (Ile50/50′) residues in the enzyme. In addition, red contours were observed around the *ortho* and *meta* position of the P2/P2′ phenyl rings indicating that the presence of hydrogen bond acceptor functionalities at these positions is not favorable for activity. Therefore, it can be concluded that substituents capable of H-bonding interactions would be preferred at the vicinal diol and around the P2/P2′ positions to improve the activity.

Mapping of 3D-QSAR models onto the HIV-1PR binding site

For the 3D-QSAR models derived using *receptor-based alignment* it is reasonable to assume that the CoMFA and CoMSIA features would be in harmony with the active site of the enzyme. Therefore the 3D contour maps were superimposed onto the active site of HIV-1PR to understand the relationship between the CoMFA/CoMSIA fields and the surroundings of the ligand binding site of the

enzyme (Figs. 8 and 9). The sterically favorable regions lie in the S1/S1′ and S2/S2′ catalytic binding pocket, which harbor the hydrophobic residues Val 32/32′, Val82/82′, Gly 48/48′, Ile 50/50′, and Ile 84/84′. The diol functionality associated with cyclic urea scaffold is involved in hydrogen bonding with Asp 25/25′. The P2/P2′ substituents also take part in hydrogen bonding interaction with Asp 30/30′ and Asp 29/29′. The urea oxygen accepts two hydrogen bonds from the Ile 50/50′. These observations are consistent with the electrostatic and the hydrogen bond donor/acceptor isopleths. Thus the field contour maps correlate well with the environmental characteristics of the active site of HIV-PR.

Design of new CU analogs

The structural requirements identified in the present study can be utilized strategically in the design of novel, potent and selective molecules with improved binding affinity. For the inactivation of HIV-1PR by cyclic urea analogs, the CoMFA and CoMSIA studies suggested that increasing the bulk at the P1/P1′ and P2/P2′ side chains of the cyclic urea scaffold would favor affinity and efficacy. The results also suggested that electropositive substituents capable of hydrogen bonding with key aspartic acid residues in the receptor are favored around the diol , the P1/P1′ and P2/P2′ side chains. Therefore, based on 3D-QSAR analysis presented here, we have designed some potential candidates (Table 5).

In silico pharmacokinetic/ toxicity predictions

In silico ADME and toxicity predictions were carried out for the designed molecules with an objective to understand their pharmacokinetic and safety profiles. This was done with the *ADME/Tox web-box v3.5* [26, 27] tool available online. The in silico ADME/toxicity prediction data is shown in Table 6. Most of the molecules showed a pK_a for the basic group around 8.0 indicating that the amino moiety is strongly basic and for the acidic group around 10.0 suggesting that the hydroxyl group is weakly acidic. The average plasma protein binding was found to be 94–99%. Inspection of the volume of distribution and plasma protein binding shows that most of these compounds are hydrophobic in nature. These drugs would predominantly bind to *alpha1*-acid glycoprotein, albumin and lipoproteins. Poor oral bioavailability was predicted for the designed molecules. Since *ADME/Tox web-box v3.5* does not predict clearance values, the bioavailability predictions may have been derived only from the solubility and permeability of the molecules. The logD or logP data suggest that these molecules have adequate aqueous solubility. Thus the low predicted bioavailability may be due to reduced permeability of the compounds. Further, it is possible that extensive

Table 6 ADME and toxicity prediction data of the designed molecules. The results have been compared with the most active molecule 52 in the dataset

Sr. No.	Solubility		Physico-chemical determinants				% oral bioavailability	Distribution			Genotoxicity	Acute toxicity LD$_{50}$ Mouse (mg/kg)	Acute toxicity LD$_{50}$ Rat (mg/kg)
	Water (logSw)	Buffer (pH 7.4) (logS)	logD (pH 7.4)	LogP	pk$_a$ (acid)	pk$_a$ (base)		Plasma Protein Binding	V$_d$ (L/kg)		Ames test	Oral	Oral
01	-4.03	-4.03	-0.62	-0.62	–	–	<30%	33.90%	0.99		0.004	5200	12000
02	-4.22	-4.22	3.71	3.71	–	4.90±0.50	<30%	99.46%	2.63		0.513	1100	2000
03	-3.90	-3.28	-1.39	-0.77	10.30±0.50	7.90±0.50	<30%	63.65%	1.86		0.000	4900	11000
04	-4.23	-4.23	4.52	4.53	–	–	<30%	99.86%	3.03		0.002	1500	1400
05	-3.82	-3.82	2.61	2.61	10.30±0.50	–	<30%	94.39%	2.39		0.005	3000	5100
06	-5.13	-5.13	6.90	6.90	–	–	<30%	99.94%	5.72		0.001	1300	860
07	-4.10	-4.10	4.60	4.61	10.30±0.50	–	<30%	99.76%	3.10		0.001	1400	2500
08	-4.17	-4.16	1.57	1.57	–	4.90±0.50	<30%	98.61%	1.72		0.992	1300	1900
09	-3.34	-2.72	-0.56	0.06	–	7.90±0.50	<30%	82.07%	2.41		0.970	1700	2700
10	-3.92	-3.92	0.21	0.21	–	5.50±0.50	<30%	92.50%	1.42		0.996	1800	2900
11	-4.32	-4.32	3.36	3.36	10.30±0.50	–	<30%	99.79%	3.10		0.000	2000	2900
12	-4.39	-4.39	0.36	0.36	9.60±0.50	–	<30%	99.82%	1.74		0.000	2300	4000
13	-3.95	-3.95	3.36	3.36	10.30±0.50	–	<30%	99.79%	2.48		0.000	2000	2900
14	-4.34	-4.34	5.69	5.69	–	–	<30%	99.86%	4.33		0.037	1400	760
52	-4.10	-4.10	4.60	4.61	10.30±0.50	–	<30%	99.76%	3.10		0.001	1400	2500

first pass metabolism may also be the reason for low bioavailability. High oral as well as *i.v.* LD_{50} dose means that these molecules are reasonably safe. Also the designed molecules showed a low potential for genotoxicity as depicted by the Ames test. To conclude, the predicted pharmacokinetic and toxicity profile for the designed molecules appeared to be favorable for hit/lead optimization.

Conclusions

The 3D-QSAR models presented in this investigation establishes the molecular basis for inhibition of HIV-1PR. The 3D-QSAR models described herein possesses good internal and external consistency. The predictive ability of these models is manifested in the good correlation between experimental and predicted pK_i values for the test molecules. The CoMFA and CoMSIA contour plots provided a fruitful insight into the different field contributions toward the overall activity. The contour maps as well as the docking studies were used for designing new molecules whose predicted activities were found to be better than the parent molecules. The ADMET/toxicity predictions for the designed molecules were also found to be within tolerance limits. Overall, the results described in this paper provide a better understanding of ligand HIV-1PR interactions and thus offer guidelines for ligand design plus a predictive model for scoring novel synthetic candidates against HIV-1PR.

References

1. Henderson LE, Benveniste RE, Sowder R, Copeland TD, Schultz AM, Oroszlan S (1988) Molecular characterization of gag proteins from simian immunodeficiency virus (SIVMne). J Virol 62:2587–2595

2. Loeb DD, Hutchison CA 3rd, Edgell MH, Farmerie WG, Swanstrom R (1989) Mutational analysis of human immunodeficiency virus type 1 protease suggests functional homology with aspartic proteinases. J Virol 63:111–121

3. Erickson J, Neidhart DJ, VanDrie J, Kempf DJ, Wang XC, Norbeck DW, Plattner JJ, Rittenhouse JW, Turon M, Wideburg N et al. (1990) Design, activity, and 2.8 A crystal structure of a C2 symmetric inhibitor complexed to HIV-1 protease. Science 249:527–533

4. Dreyer GB, Boehm JC, Chenera B, DesJarlais RL, Hassell AM, Meek TD, Tomaszek TA, Lewis M (1993) A symmetric inhibitor binds HIV-1 protease asymmetrically. Biochemistry 32:937–947

5. Lam PY, Jadhav PK, Eyermann CJ, Hodge CN, Ru Y, Bacheler LT, Meek JL, Otto MJ, Rayner MM, Wong YN et al. (1994) Rational design of potent, bioavailable, nonpeptide cyclic ureas as HIV protease inhibitors. Science 263:380–384

6. Lam PY, Ru Y, Jadhav PK, Aldrich PE, DeLucca GV, Eyermann CJ, Chang CH, Emmett G, Holler ER, Daneker WF, Li L, Confalone PN, McHugh RJ, Han Q, Li R, Markwalder JA, Seitz SP, Sharpe TR, Bacheler LT, Rayner MM, Klabe RM, Shum L, Winslow DL, Kornhauser DM, Hodge CN et al. (1996) Cyclic HIV protease inhibitors: synthesis, conformational analysis, P2/P2′ structure-activity relationship, and molecular recognition of cyclic ureas. J Med Chem 39:3514–3525

7. Schaal W, Karlsson A, Ahlsen G, Lindberg J, Andersson HO, Danielson UH, Classon B, Unge T, Samuelsson B, Hulten J, Hallberg A, Karlen A (2001) Synthesis and comparative molecular field analysis (CoMFA) of symmetric and nonsymmetric cyclic sulfamide HIV-1 protease inhibitors. J Med Chem 44:155–169

8. Debnath AK (1999) Three-dimensional quantitative structure-activity relationship study on cyclic urea derivatives as HIV-1 protease inhibitors: application of comparative molecular field analysis. J Med Chem 42:249–259

9. Sybyl. *Sybyl*, version 7.1; Tripos Associates Inc: 1699S Hanley Rd, St. Louis, MO 631444, USA, 2005

10. Gold. *GOLD*, version 3.1; Cambridge Crystallographic Data Centre (CCDC), UK, 2005

11. Nayyar A, Malde A, Jain R, Coutinho E (2006) 3D-QSAR study of ring-substituted quinoline class of anti-tuberculosis agents. Bioorg Med Chem 14:847–856

12. Accelrys I. *Cerius2*, version 4.8; San Diego, CA, USA, 1998

13. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The protein data bank. Nucleic Acids Res 28:235–242

14. Cramer RD, Patterson DE, Bunce JD (1988) Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. J Am Chem Soc 110:5959–5967

15. Klebe G, Abraham U, Mietzner T (1994) Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. J Med Chem 37:4130–4146

16. Datar PA, Coutinho EC (2004) A CoMFA study of COX-2 inhibitors with receptor based alignment. J Mol Graph Model 23:239–251

17. Pissurlenkar RR, Shaikh MS, Coutinho EC (2007) 3D-QSAR studies of Dipeptidyl peptidase IV inhibitors using a docking based alignment. J Mol Model 13:1047–1071

18. Wold S, Johansson E, Cocchi M (1993) PLS: partial least squares projections to latent structures. In: Kubinyi H (ed) 3D QSAR in drug design: theory, methods and applications. ESCOM Science, Leiden, pp 523–550

19. Stone M (1974) Cross-validatory choice and assessment of statistical predictions. J Roy Stat Soc B 36:111–147

20. Richard D, Cramer RD III, Bunce JD, Patterson DE, Frank IE (1988) Crossvalidation, bootstrapping, and partial least squares compared with multiple regression in conventional QSAR studies. Quant Struct-Act Relat 7:18–25

21. Bush BL, Nachbar RB (1993) Sample-distance partial least squares: PLS optimized for many variables, with application to CoMFA. J Comput Aided Mol Des 7:587–619

22. Deep R (2006) Regression. In: Deep R (ed) Probability and statistics. Academic, London, UK, pp 455–515

23. Shao J (1996) Bootstrap model selection. J Am Stat Assoc 91:655–665
24. Rucker C, Rucker G, Meringer M (2007) y-Randomization and its variants in QSPR/QSAR. J Chem Inf Model 47:2345–2357
25. Marshall GR (1998) Binding-site modeling of unknown receptors. In: Kubinyi H, Martin YC, Folkers G (eds) 3D QSAR in drug design: Theory methods and applications. Springer, London, pp 80–116
26. Didziapetris R, Reynolds DP, Japertas P, Zmuidinavicius D, Petrauskas A (2006) In silico technology for identification of potentially toxic compounds in drug discovery. Curr Comput-Aided Drug Des 2:95–103
27. Zmuidinavicius D, Japertas P, Petrauskas A, Didziapetris R (2003) Progress in toxinformatics: the challenge of predicting acute toxicity. Curr Top Med Chem 3:1301–1314