Receptor-Independent 4D-QSAR Analysis of a Set of Norstatine Derived Inhibitors of HIV-1 Protease

Craig L. Senese and A. J. Hopfinger*

Laboratory of Molecular Modeling and Design, Department of Medicinal Chemistry and Pharmacognosy,
College of Pharmacy - University of Illinois at Chicago, 833 South Wood Street,
Chicago, Illinois 60612-7231

Received March 4, 2003

A training set of 27 norstatine derived inhibitors of HIV-1 protease, based on the 3(S)-amino-2(S)-hydroxyl-4-phenylbutanoic acid core (AHPBA), for which the -log IC_{50} values were measured, was used to construct 4D-QSAR models. Five unique RI-4D-QSAR models, from two different alignments, were identified ($q^2 = 0.86-0.95$). These five models were used to map the atom type morphology of the lining of the inhibitor binding site at the HIV protease receptor site as well as predict the inhibition potencies of seven test set compounds for model validation. The five models, overall, predict the -log IC_{50} activity values for the test set compounds in a manner consistent with their q^2 values. The models also correctly identify the hydrophobic nature of the HIV protease receptor site, and inferences are made as to further structural modifications to improve the potency of the AHPBA inhibitors of HIV protease. The finding of five unique, and nearly statistically equivalent, RI-4D-QSAR models for the training set demonstrates that there can be more than one way to fit structure−activity data even within a given QSAR methodology. This set of unique, equally good individual models is referred to as the manifold model.

INTRODUCTION

Since the publication of its crystal structure in 1989,[1,2] the aspartyl protease from the human immunodeficiency virus type-1 (HIV-1) has been a major target to break the retrovirus replication cycle[3−8] and thereby stunt the progression from HIV carrier to acquired immune deficiency syndrome (AIDS). A common feature of the retrovirus family, to which HIV-1 belongs, is the translation of viral proteins and enzymes as polyproteins that require further processing by viral enzymes to become functional.[9]

HIV-1 protease (HIVPR) is one such viral enzyme which cleaves specific peptide bonds found in the polyprotein products of the *gag* and *gag-pol* genes. As postulated, and hoped, the inhibition of HIVPR results in the cessation of processing of the *gag* and *gag-pol* gene products and subsequent production of noninfectious virions.[10−12] These findings, coupled with previous work done on other aspartyl proteases, make HIVPR an attractive drug target for anti-retroviral therapy.

Early work on HIVPR inhibitors focused on derivatives of the endogenous ligands,[13] specifically Tyr-Pro and Phe-Pro templates which were converted to peptidomimetic analogues. Peptide based drugs generally suffer from several drawbacks, including high production costs and typically poor pharmacokinetic profiles. The search for drug candidates with good potencies, half-lives, and favorable toxicity profiles resulted in the discovery of two general classes (cyclic and acyclic) of non-peptide derived inhibitors of HIVPR.

A new lead compound, 3(S)-amino-2(S)-hydroxyl-4-phenylbutanoic acid (AHPBA), belonging to the nonpeptidic acyclic class of HIVPR inhibitors, is currently being explored. AHPBAs are norstatine derived transition state mimics that show high potency and a good pharmacokinetic profile.[14−16] A constant influx of new drug candidates appears necessary to match the rapid rate of drug resistance exhibited by HIV-1. Therefore, the full potential of the AHPBA derived class of HIVPR inhibitors needs to be explored. Previous SAR studies[17] of these inhibitors provide information to permit the application of molecular modeling and QSAR analysis in the search for the optimum drug candidates in this class.

The 4D-QSAR paradigm[18] combines molecular modeling and QSAR analysis to permit construction of, in this application, quantitative models of enzyme inhibition as a function of the conformation, alignment, and putative binding pharmacophore of the AHPBA inhibitors. The AHPBAs and its derivatives are flexible molecules. Therefore, the analysis of these inhibitors with a QSAR method that incorporates conformational flexibility in QSAR model construction may provide unique insight from the structure−activity information contained in a given data set. The purpose of this study is to utilize the 4D-QSAR method to assess novel possible structural modifications of the AHPBA class of HIVPR inhibitors with the goal of expanding the pool of AHPBA derived inhibitors of HIVPR.

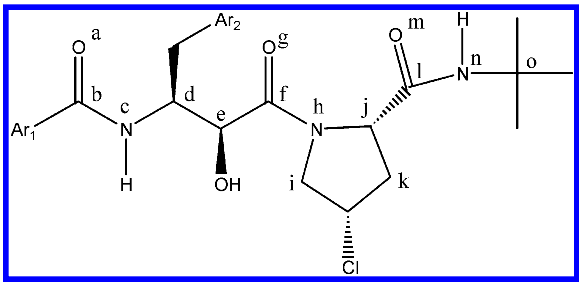
METHOD

Training Set of AHPBAs. The training set of AHPBAs used in this work is the same as that of a previous 3D-QSAR study performed by Huang et al.[17] with a few notable methodological exceptions. The AHPBA core structure as well as the atom lettering used for alignment identification is given in Table 1. The 27 training set compounds are listed along with their corresponding -log IC_{50} values in Table 2. The inhibitory measures of the AHPBA compounds against the HIVPR enzyme were determined via a sodium dodecyl

* Corresponding author e-mail: hopfingr@uic.edu.

10.1021/ci0340456 CCC: $25.00
Published on Web 05/31/2003

Table 1. The Three Ordered-Atom Alignments Used in the 4D-QSAR Analysis of the AHPBA Derived Inhibitors of HIV-1



alignment	atom 1	atom 2	atom 3
1	f	h	j
2	f	h	i
3	e	f	h
4	e	f	g
5	b	c	d
6	c	b	a
7	d	e	f
8	k	j	l
9	j	l	n
10	j	l	m
11	l	n	o
12	b	d	f
13	a	e	g
14	a	d	f

sulfate polyacrylamide gel electrophoresis assay (SDS-PAGE) described in a previous publication.[4] Briefly, the inhibitors were evaluated for their ability to prevent cleavage at any of four sites of a recombinant *gag* substrate in the presence of HIVPR. The range of -log IC_{50} values for the training set spans less than 2 orders of magnitude (7.30 to 9.10). This is a relatively small range in inhibition activity, and the inhibitory measures are skewed toward potent inhibition. Therefore, significant 4D-QSAR models should reflect subtle differences in the training set compounds leading to differentiation in high activity. Table 3 contains a set of seven AHPBA test compounds and their inhibitory activity values which has been used to assess the predictive power of the 4D-QSAR models built from the training set. Two test compounds employed in the original 3D-QSAR study of ref 17 were not considered in this work because their activity measures (-log $IC_{50} = 6.00$ for both) are nearly one-and-a-half orders of magnitude outside the range of the training set.

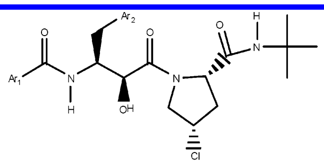
4D-QSAR Analysis. The 4D-QSAR methodology has been presented previously in detail[18] and will only be summarized here. The commercial version (V3.0) of the 4D-QSAR package was employed in this study.[19] We are currently expanding the 4D-QSAR paradigm to include the receptor geometry when it is available. Thus, we now distinguish between receptor-independent (RI) and receptor-dependent (RD) 4D-QSAR analyses. This study is an RI-4D-QSAR analysis. The first step in the analysis is to generate a reference grid cell lattice in which to place the 3D structure of each training set compound. In this study the grid cell lattice is composed of a set of one angstrom cubes. The 3D structures of the training set compounds are constructed and optimized in Hyperchem 6.0.[20] The preferred compound geometry is determined via molecular mechanics with an MM+ force field, and the partial charges are assigned using a semiempirical AM1 method. In a major point of departure from the reported 3D-QSAR study,[17] each of the training set compounds generated for the 4D-QSAR analysis is *not* restricted to a *single conformation* based on a predicted bound conformation for the inhibitors.

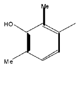
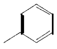
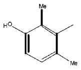
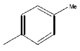
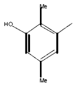
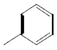
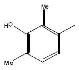
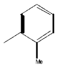
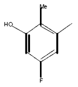
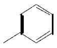
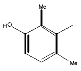
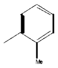
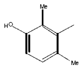
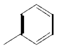
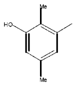
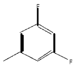
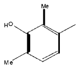
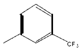
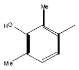
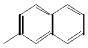
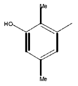
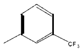
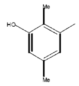
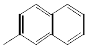
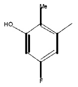
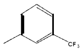
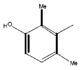
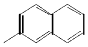
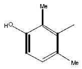
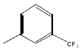
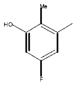
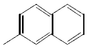
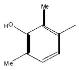
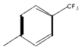
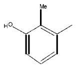
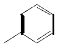
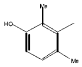
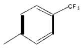
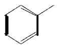
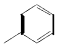
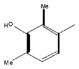
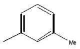
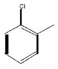
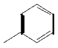
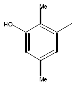
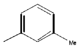
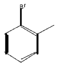
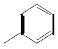
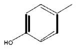
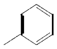
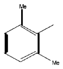
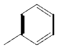
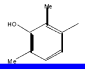
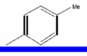
The *interaction pharmacophore elements*, or IPEs, defined in Table 4, were assigned to the intramolecular energy minimized 3D structure of each inhibitor and the *conformational ensemble profile*, or CEP, was then generated for each training set inhibitor. A molecular dynamics simulation (MDS) is used to create the CEP. The MOLSIM software package with the extended MM2 force field is currently utilized to perform the MDS.[21] The molecular dielectric was set to 3.5, and the simulation temperature was fixed at 300 K. A sampling time of 100 ps was employed, over which a total of 1000 conformations of each inhibitor were recorded. The CEP was created by recording the atomic coordinates and conformational energy every 0.1 ps throughout the simulation, resulting in 1000 “snapshots” of the inhibitor as it traverses through the set of thermodynamically available conformer states.

Following generation of the CEP of each inhibitor, the molecular alignments are chosen for the training set. Three-ordered atom alignment rules were used in this study. In general, the alignments are chosen to span the common framework (core) of the molecules in the training set. Alignments using atoms from the right, left, and middle of the common framework and alignments that use atoms that span the common framework should be used to ensure a complete alignment analysis. In this way information relating to the substituent properties of the compounds is obtained from the resulting models. This alignment strategy is reflected in the alignments chosen for this study (Table 1). Alignments 1 and 2 focus on the middle-right portion of the framework. Alignments 3, 4, and 7 focus on the middle of the inhibitors, around the norstatine like region, while alignments 5 and 6 focus on the left-hand portion of the molecule. The right-hand portion of the training set compounds, near the proline ring, is captured in alignments 8 through 11, while alignments 12 through 14 span several regions of the inhibitors.

All conformations from the CEP of every compound are placed in the grid cell lattice space according to a selected trial alignment. The occupancy of the grid cells by each IPE type is recorded over the CEP and forms the set of grid cell occupancy descriptors, or GCODs, to be utilized as the pool of trial descriptors in the model building and optimization process. The 4D-QSAR method also offers the option of including additional non-GCOD descriptors, such as log P, to the pool of trial descriptors. Addition of non-GCOD descriptors may help to improve the quality of the models, or to provide additional information regarding the data set. The genetic function approximation (GFA)[22] is used to optimize the 4D-QSAR models. The results of the GFA analysis are a set of RI-4D-QSAR models describing the activity of the compounds based on the occupancy of certain grid cells by specific IPE types. The location of these grid cells as well as the sign and magnitude of their regression coefficients can provide direct information regarding the pharmacophore/binding features of the training set compounds and indirect information about the features of the target receptor.

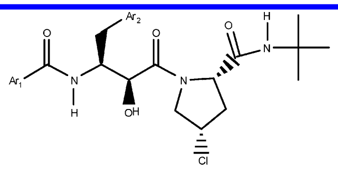
Table 2. The 27 Training Set Compounds and Their Inhibition Potency Measures



Compound	Ar_1	Ar_2	$-\log IC_{50}$	Compound	Ar_1	Ar_2	$-\log IC_{50}$
1			8.10	15			8.44
2			8.89	16			7.30
3			8.77	17			7.46
4			8.60	18			8.92
5			8.46	19			8.49
6			8.72	20			9.10
7			8.68	21			8.55
8			8.34	22			8.92
9			8.20	23			9.10
10			8.41	24			7.59
11			8.60	25			7.96
12			8.89	26			7.92
13			7.50	27			8.44
14			8.10				

The *best* models in a 4D-QSAR study can be chosen based on a number of different criteria. Typically, the first model quality parameter considered is the statistical measure of fit of a model. The leave-one-out cross-validated correlation coefficient, or q^2, is currently used as the preferred parameter of model fitness. As the models will vary in their number of model terms, it is important to determine how many to include in the best model. The typical way to determine the optimal number of descriptor terms to include in the best model is to plot the number of model term versus the cross-validated correlation coefficient. The point of the plot at which the q^2 does not significantly increase with the addition of an additional model term is chosen as the optimal number of model terms. A test set, that is analogue compounds not included in the training set, is also used to evaluate the predictive power of a QSAR model. Other parameters of model quality are the regression coefficients and the constant term of a model. Ideally, a model with an equal, or nearly equal, number of positive and negative regression coefficients is preferred and reflects model stability. Correspondingly, the value of the constant term should be midway in value across the range of activity values of the training set. Finally, if binding properties, such as locations of key residues and/or hydrophobic regions, of the training set compounds with the target receptor or the receptor topology are known a priori, the GCODs of the best models may be directly evaluated by comparison with the IPE profile of the receptor, i.e., amino acid residue composition of the inhibitor binding

Table 3. The 9 Test Set Compounds and Their Inhibition Potency Measures



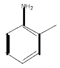
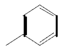
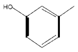
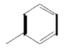
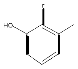
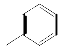
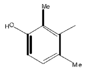
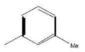
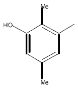
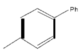
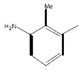
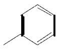
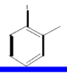
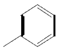
Compound	Ar_1	Ar_2	-log IC_{50}
1			7.41
2			8.05
3			8.82
4			8.47
5			8.89
6			8.27
7			7.85

Table 4. Interaction Pharmacophore Elements, IPEs, Used in 4D-QSAR Analyses

IPE description (abbreviation)	IPE code
all atoms in the molecule (any)	0
nonpolar atoms (np)	1
polar (+) atoms (p+)	2
polar (−) atoms (p-)	3
hydrogen bond acceptor atoms (hba)	4
hydrogen bond donor atoms (hbd)	5
aromatic atoms (a)	6
non-hydrogen atoms (hs)	7

region of the enzyme. Finally, there are other statistical measures such as r^2, standard error (SE), lack-of-fit (LOF), and *F*-values that are indicators of model fitness.[22,25] It is usually the case that a combination of (or, all of) these criteria, as is the case in the current study, are considered in order to propose the *best* model from a 4D-QSAR analysis.

It is also possible, however, that several models and/or alignments equally satisfy the criteria stated above. In this situation, a set of models, or a *manifold* model, may be necessary to fully encompass all of the information consistent with the training set. Models that are statistically significant, yet not highly correlated to one another, provide unique information relative to the training set. By determining the set of models in this situation, one can be assured that the QSAR analysis is fully extracting the information available in the training set. In the current study, multiple, and largely independent, 4D-QSAR models have been used as a manifold model in determining the important binding characteristics of the inhibitors to the enzyme.

Once a *best* model, or set of models, is chosen from the 4D-QSAR analysis, the active conformation of each of the compounds in the training set can be postulated *relative to that model*. This important operation is accomplished by first determining the conformations of the CEP that are within a threshold energy limit, i.e., only thermodynamically accessible conformations are considered, and then determining which of these possible conformations has the highest activity as predicted by the model. The resulting postulated active conformation provides a reasonable working hypothesis regarding the 3D binding mode of each of the compounds of the training set. The postulated active conformation of a potent member of the training set can be used as a structural template for rational drug design[23] or as a starting point for other QSAR analyses that require a user postulated active conformation, such as CoMFA.[24]

Table 5. The Statistical Quality, as Measured by r^2 and q^2, of the Five-Term Models for the 14 Alignments

alignment[a]	r^2 range	q^2 range
1	0.78−0.83	0.61−0.76
2	0.83−0.85	0.73−0.76
3	0.84−0.86	0.75−0.77
4	0.83−0.85	0.75−0.79
5	0.93−0.95	0.89−0.92
6	0.94−0.98	0.91−0.96
7	0.90−0.91	0.79−0.83
8	0.87−0.88	0.75−0.79
9	0.90−0.91	0.82−0.84
10	0.88−0.91	0.80−0.82
11	0.90−0.92	0.85−0.86
12	0.90−0.92	0.84−0.89
13	0.86−0.94	0.80−0.92
14	0.90−0.93	0.86−0.90

[a] For alignment definitions, see Table 1.

RESULTS

As previously discussed, the 14 alignments given in Table 1 were chosen to cover the conserved structural elements of the core AHPBA compound for the training set. The statistical measures of fit for the best RI-4D-QSAR models for each of the trial alignments are given in Table 5. Five-term models were chosen as the basis of comparison between the alignments because this is the optimum model size for the majority of alignments. It is readily apparent from Table 5 that alignments 1−4 and alignment 8 are statistically inferior to the remaining alignments, all of which exhibit higher r^2 and q^2 values. The best alignments, as determined by statistical quality of the models, focus on the region of the molecule near the Ar_1 substituent, while alignments involving atoms near the proline typically have lower statistical significance.

Evaluating statistical quality in terms of q^2 for the five term models suggests that alignment 6 is the best alignment. From a plot of q^2 versus the number of model terms, Figure 1(a), it appears that for alignment 6 the optimum model size is four terms, as this is the point at which q^2 does not significantly increase with the addition of another GCOD term. However, the plot of the predicted residual sum of squares for the test set compounds versus the number of model terms, Figure 1(b), indicates that the five-term model is more predictive in extrapolating to compounds outside the training set. Therefore, the optimum model size for alignment 6 is considered to be five terms. The cross-correlation matrix of the residuals of fit for the top-ten models, from the GFA

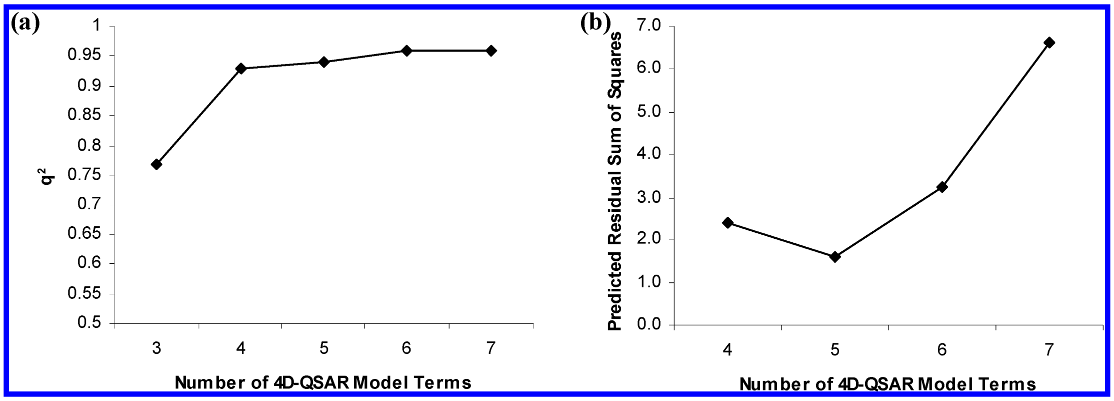


Figure 1. Plot of q^2 value versus number of 4D-QSAR model terms for alignment 6 (a). Also shown is the plot of the predicted residual sum of squares for the test compounds versus the number of 4D-QSAR model terms (b).

Table 6. The Cross-Correlation Matrix of the Residuals of Fit for the Top-Ten Models for Alignment 6

	model 1	model 2	model 3	model 4	model 5	model 6	model 7	model 8	model 9	model 10
model 1	1									
model 2	0.81	1								
model 3	0.95	0.80	1							
model 4	0.78	0.95	0.84	1						
model 5	0.94	0.72	0.96	0.80	1					
model 6	0.77	0.74	0.79	0.76	0.77	1				
model 7	0.59	0.76	0.64	0.78	0.61	0.54	1			
model 8	0.71	0.54	0.73	0.60	0.77	0.54	0.85	1		
model 9	0.63	0.56	0.62	0.60	0.67	0.82	0.75	0.81	1	
model 10	0.61	0.53	0.65	0.60	0.69	0.67	0.75	0.82	0.84	1

optimization trajectory in which the best five-term model was found, is given in Table 6. An r-value of greater than 0.7 indicates that two models are highly correlated and can be considered to represent the training set in the same manner. Highly correlated models typically have similar GCOD representations in grid lattice space and vary only by small differences in the GCOD coefficients found in the equation. The matrix in Table 6 indicates there are two unique models, model 1 and model 7, present among the top-ten models for alignment 6. These 4D-QSAR models are given as eq 1, model 1, and eq 2, model 7.

$$\text{-log IC}_{50} = 8.52 + 4.09\text{*GC1 (np)} - 14.15\text{*GC2 (np)} - 3.02\text{*GC3 (np)} + 11.73\text{*GC4 (any)} - 5.32\text{*GC5 (np)} \quad (1)$$

$$N = 27,\ r^2 = 0.98,\ q^2 = 0.95$$

$$-\log \text{IC}_{50} = 8.59 - 15.72\text{*GC1 (np)} - 2.44\text{*GC2 (np)} + 13.05\text{*GC3 (any)} - 6.52\text{*GC4 (any)} \quad (2)$$

$$N = 27,\ r^2 = 0.95,\ q^2 = 0.93$$

The cross-correlation matrix for the GCODs represented in eqs 1 and 2 is given in Table 7. Any pair of GCODs that are highly correlated in a 4D-QSAR model, i.e., absolute r value greater than 0.7, may reflect possible positive or negative cooperative effects between the regions identified by these GCODs in the receptor. There are no such correlations in these two models. Therefore, each of the GCODs represented in the models is independent. One common feature of eqs 1 and 2 is the grid cell types of the model indicate that nonpolar interactions are the dominant source of inhibitor-enzyme binding which has been identified as characteristic of AHPBA binding to HIVPR in previous studies.[4,16,17] The unusually high values for the cross-validated correlation coefficients suggest that these models may be overfitting the training set data. However, the seven compound test set utilized in this study indicates that this is not the case, as the models do an acceptable job at extrapolating

Table 7. The Cross-Correlation Matrix of Grid Cell Occupancy Descriptors (GCODs) for (a) Eq 1 and (b) Eq 2

(a) Eq 1

	GC1	GC2	GC3	GC4	GC5
GC1	1				
GC2	0.01	1			
GC3	0.43	−0.09	1		
GC4	−0.08	−0.04	−0.32	1	
GC5	−0.22	0.33	−0.1	0.49	1

(b) Eq 2

	GC1	GC2	GC3	GC4
GC1	1			
GC2	−0.09	1		
GC3	−0.04	−0.32	1	
GC4	0.06	−0.07	0.53	1

Table 8. Predicted and Observed -log IC_{50} Values for the Test Set Compounds Using the Two Unique Models, Eqs 1 and 2, from Alignment 6

test	predicted activity		obsd	residual activity	
compd	eq 1	eq 2	activity	eq 1	eq 2
1	8.42	8.26	7.41	1.01	0.85
2	8.15	7.63	8.05	0.1	−0.42
3	8.67	8.74	8.82	−0.15	−0.08
4	8.52	8.48	8.47	0.05	0.01
5	8.73	8.00	8.89	−0.16	−0.89
6	8.98	9.10	8.27	0.71	0.83
7	7.89	7.79	7.85	0.04	−0.06

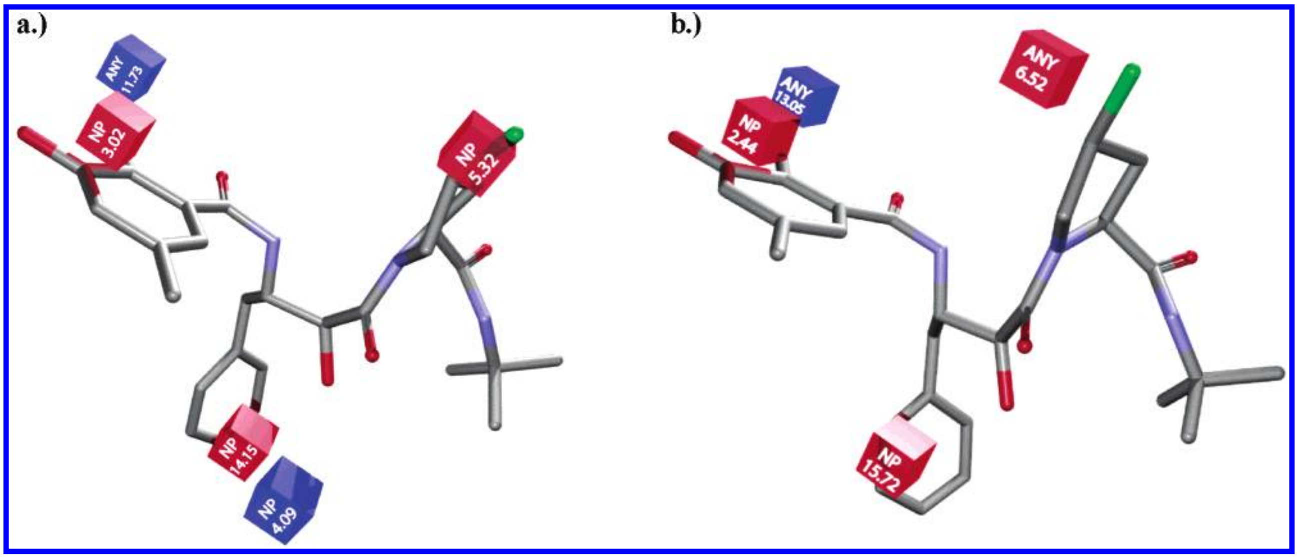


Figure 2. Predicted active conformation for compound 2 using the two unique models from alignment 6, eq 1 (a) and eq 2 (b). The grid cells from the model are the colored 1 angstrom cubes. Blue grid cells are those that contribute positively to inhibition potency, while the red grid cells reduce potency. The five grid cells shown are labeled with their corresponding IPE type as well as the magnitude of the regression coefficient.

to predict the -log IC_{50} for compounds not used in the model building process.

Table 8 lists the observed and predicted inhibition potencies of the test set compounds as calculated using eqs 1 and 2. Equation 1 produces accurate predictions for five of the seven test set compounds, and eq 2 is accurate for four of the seven compounds. Both equations share difficulty in the prediction of two of the compounds. A possible explanation for the less accurate predictions for test compounds 1 and 6 for eqs 1 and 2, and compound 5 for eq 2, is that these molecules contain functional groups that are not present in any of the training set compounds. Test compounds 1 and 6 contain an amine group on Ar_1, and compound 5 contains an extra phenyl ring bound by a carbon−carbon bond to the Ar_2 moiety. The other test set compounds all contain functional groups that are represented in the training set.

Although the correlation of the residuals of fit is less than 0.7, the two distinct models of alignment 6 have similar pharmacophore representations. Figure 2 is a graphical representation (or 3D-pharmacophore) of the RI-4D-QSAR models, eqs 1 and 2, using compound 2 in its postulated active conformation as a reference structure. The corresponding grid cells for the two models appear to represent the same regions in ligand−receptor space, with the only difference being an extra grid cell of type “nonpolar” (regression coefficient of 4.09) in the Ar_2 region for eq 1. Another indication that these two models have some common features is they predict an identical active conformation for compound 2.

The 4D-QSAR models appear to define three inhibitor-enzyme interaction regions relative to the AHPBA analogues. The first region involves the Ar_1 substituent and is represented by a positive GCOD of IPE type “any” with a magnitude, i.e., regression coefficient from the model, of 11.73 for eq 1 and 13.05 for eq 2, and a negative GCOD of IPE type nonpolar with a magnitude of 3.02 for eq 1 and 2.44 for eq 2. The second region is located around the Ar_2 substituent and, like the first region, is again represented by both a positive and negative GCOD for eq 1 but just a negative GCOD for eq 2. All GCODs are of IPE type “nonpolar” with absolute magnitudes of 4.09 and 14.15, respectively, for eq 1, and 15.72 for eq 2. The third inhibitor-enzyme interaction region is located near the proline ring and is characterized by a negative GCOD of type “nonpolar”, with a magnitude of 5.32 for eq 1 and a negative GCOD of type “any” with a magnitude of 6.52 for eq 2. These three regions identified by the 4D-QSAR models may be sites on the inhibitor which differentiate the subtle changes in binding potency among these tight binding AHPBA analogues to HIVPR. A general feel for the binding properties of the inhibitors can be inferred from this model. The negative GCODs of IPE type “nonpolar” near the hydroxyl of Ar_1 suggests this is an electrostatic or hydrogen-bonding region of the receptor, and replacement of the hydroxyl with a nonpolar substituent or removal of the hydroxyl would result in a loss in activity. This interpretation is supported when compound 4 is compared to compound 27. Compound 4 contains a hydroxyl group in the meta position of Ar_1, which is the correct position to interact with the electrostatic region of the receptor identified by the 4D-QSAR model. The sole difference between compound 4 and compound 27 is the presence/absence of the meta hydroxyl, in compound 27 the hydroxyl is absent and this inhibitor cannot take advantage of the electrostatic region and, consequently, has a lower inhibition potency than compound 4.

Alignment 14 is an alignment using nonadjacent atoms that incorporates one atom from alignment 6 and provides excellent model statistics (Table 5). The plot of q^2 versus the number of 4D-QSAR model terms, Figure 3(a), indicates that a six-term model is the optimum size. However, as with alignment 6, the plot of the predicted residual sum of squares for the test set compounds versus the number of GCOD descriptor terms for the best (as determined by q^2) four-, five-, six-, and seven-term models, Figure 3(b), shows that a five-term model is the best model. The cross-correlation matrix of the residuals of fit for the top-ten models, from the GFA optimization trajectory in which the best five-term model was found, is given in Table 9. Cross-correlations greater than 0.7 are highlighted in bold, indicating that the two models are similar to one another. When a column of

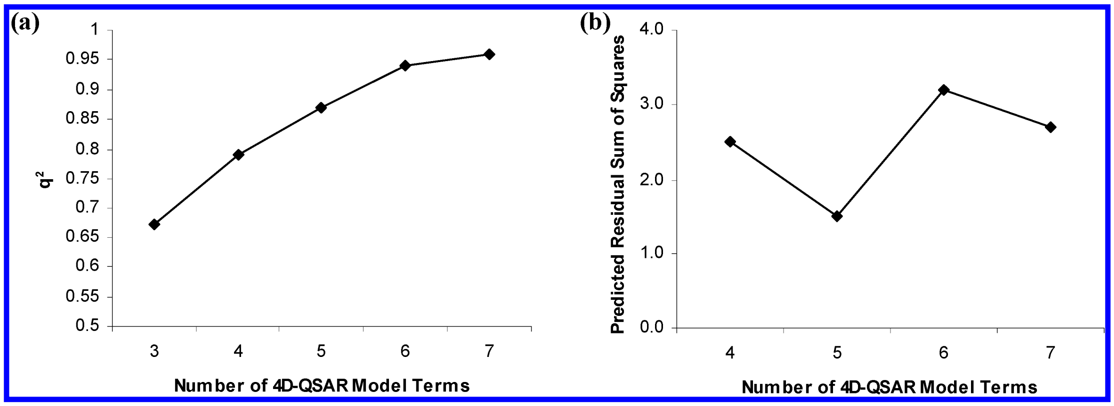


Figure 3. Plot of q^2 value versus number of 4D-QSAR model terms for alignment 14 (a). Also shown is the plot of the predicted residual sum of squares for the test compounds versus the number of 4D-QSAR model terms (b).

Table 9. The Cross-Correlation Matrix of the Residuals of Fit for the Top-Ten Models for Alignment 14

	model 1	model 2	model 3	model 4	model 5	model 6	model 7	model 8	model 9	model 10
model 1	**1**									
model 2	**0.91**	**1**								
model 3	**0.89**	**0.91**	**1**							
model 4	0.54	0.46	0.52	**1**						
model 5	**0.91**	**0.93**	**0.76**	0.5	**1**					
model 6	0.47	0.53	0.49	0.3	0.49	**1**				
model 7	0.51	0.43	0.41	0.51	0.52	**0.72**	**1**			
model 8	0.63	0.63	0.63	0.54	0.62	**0.89**	**0.87**	**1**		
model 9	0.65	0.63	0.63	0.62	0.64	**0.75**	**0.89**	**0.95**	**1**	
model 10	0.54	0.58	0.54	0.37	0.55	**0.99**	**0.75**	**0.91**	**0.79**	**1**

the cross-correlation matrix contains several r-values of less than 0.7, this indicates the presence of several distinct models. From the cross-correlation of residuals matrix there appear to be three distinct models, numbered 1, 4, and 6. These models are given below as eqs 3−5, respectively.

$$\text{-log IC}_{50} = 8.05 + 3.99\text{*GC1 (any)} + 3.41\text{*GC2 (np)} - 9.29\text{*GC3 (any)} + 9.38\text{*GC4 (any)} - 10.84\text{*GC5 (any)} \quad (3)$$

$$N = 27,\ r^2 = 0.93,\ q^2 = 0.90$$

$$-\text{log IC}_{50} = 7.80 + 5.41\text{*GC1 (np)} - 0.83\text{*GC2 (any)} - 9.01\text{*GC3 (any)} - 5.22\text{*GC4 (any)} + 2.01\text{*GC5 (np)} \quad (4)$$

$$N = 27,\ r^2 = 0.91,\ q^2 = 0.87$$

$$-\text{log IC}_{50} = 7.95 + 5.07\text{*GC1 (np)} - 10.27\text{*GC2 (any)} + 4.72\text{*GC3 (np)} - 7.86\text{*GC4 (any)} \quad (5)$$

$$N = 27,\ r^2 = 0.89,\ q^2 = 0.86$$

The cross-correlation matrices for the GCODs in eqs 3−5 are given in Table 10. As with the models from alignment 6, there is no correlation between pairs of GCODs, indicating that they are all independent. All three of the models, eqs 3−5, satisfy several of the parameters discussed as indicators of a good model, high q^2 value, both positive and negative regression coefficients as well as regression constants that are in the middle of the activity range for the data set (7.30−9.10). When considered together, these three equations indicate that nonpolar interactions as well as spatial considerations are important pharmacophore features of the AH-PBA data set. In fact, of the 15 unique grid cells that appear in the top-ten models for alignment 14, five are of IPE type "nonpolar" and 10 are IPE type "any". These three equations, like eqs 1 and 2, are 4D-QSAR models dominated by nonpolar and, likely, steric ("any") interactions.

Table 10. The Cross-Correlation Matrix of Grid Cell Occupancy Descriptors (GCODs) for (a) Eq 3, (b) Eq 4, and (c) Eq 5

(a) Eq 3

	GC1	GC2	GC3	GC4	GC5
GC1	1				
GC2	−0.41	1			
GC3	−0.17	0.43	1		
GC4	−0.26	0.19	−0.04	1	
GC5	−0.05	−0.15	0.04	0.56	1

(b) Eq 4

	GC1	GC2	GC3	GC4	GC5
GC1	1				
GC2	−0.01	1			
GC3	0.43	0.15	1		
GC4	0.37	0.35	0.55	1	
GC5	−0.5	0.35	−0.22	−0.25	1

(c) Eq 5

	GC1	GC2	GC3	GC4
GC1	1			
GC2	0.43	1		
GC3	−0.55	−0.14	1	
GC4	0.18	0.31	−0.33	1

To further evaluate the three models from alignment 14, each was used to predict the inhibition potencies of the seven test set compounds. The results are given in Table 11. It is immediately apparent that eqs 4 and 5 have nearly identical predictive features, and both are superior to eqs 1−3 in terms of predicting the activities of the test set compounds. Also evident from Table 11 is that the three 4D-QSAR models of alignment 14 do not have the same difficulty accurately predicting the -log IC_{50} of the two amine containing test set compounds, 1 and 6, as do the 4D-QSAR models from alignment 6. This better prediction performance of eqs 4 and 5 may be attributed to the different GCODs found for the two alignments in the Ar_1 region of the molecules. Thus, incorporating more compounds in the training set containing the amine moiety may provide further information relating to the binding preferences of the HIVPR inhibitors.

Table 11. (a) Predicted and Observed -log IC_{50} Values and (b) the r^2 Values of Prediction of the Test Set Compounds for the Three Unique Models from Alignment 14

(a) Predicted and Observed -log IC_{50} Values

test	predicted activity			obsd	residual activity		
compd	eq 3	eq 4	eq 5	activity	eq 3	eq 4	eq 5
1	7.91	6.67	7.59	7.41	0.50	0.26	0.18
2	7.70	7.41	7.99	8.05	−0.35	−0.64	−0.06
3	7.90	8.73	8.56	8.82	−0.92	−0.09	−0.26
4	8.51	8.43	8.50	8.47	0.04	−0.04	0.03
5	8.23	8.59	8.31	8.89	−0.66	−0.30	−0.58
6	8.57	8.60	8.69	8.27	0.30	0.33	0.42
7	8.19	7.87	7.85	7.85	0.34	0.02	0.02

(b) r^2 Values of Prediction

	eq 3	eq 4	eq 5
r^2 of test set prediction[a]	−0.05	0.61	0.64

[a] Intercept set to 0.

DISCUSSION

Several pairs of compounds in the training set differ only by the position of a single substituent, yet show marked differences in inhibition potency. Figure 4 presents two such training set molecules, 4 and 17, relative to eqs 4 and 5. The -log IC_{50} difference of compounds 4 and 17 is more than one log unit (an order of magnitude in IC_{50}), yet the only structural difference is the presence of an ortho methyl group on Ar_2 of compound 17. The intersection of this methyl substituent with a activity reducing GCOD term, present in

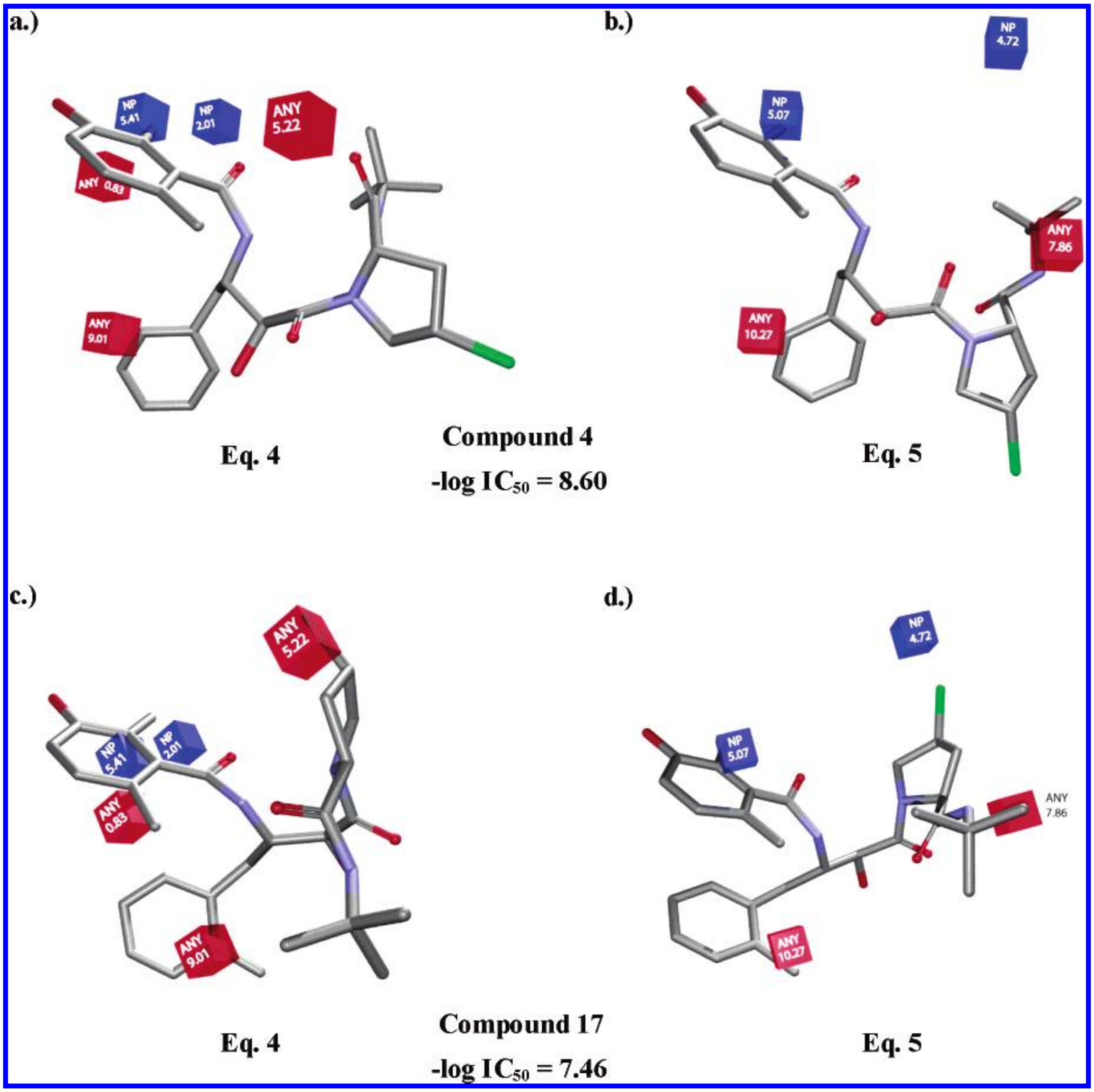


Figure 4. The predicted active conformations of compounds 4 (a,b) and 17 (c,d) by the 4D-QSAR models from alignment 14, eq 4 (a,c) and eq 5 (b,d). The IPE type and magnitude of the regression coefficient is given for each grid cell. Blue grid cells contribute positively to inhibition potency, while red grid cells reduce potency.

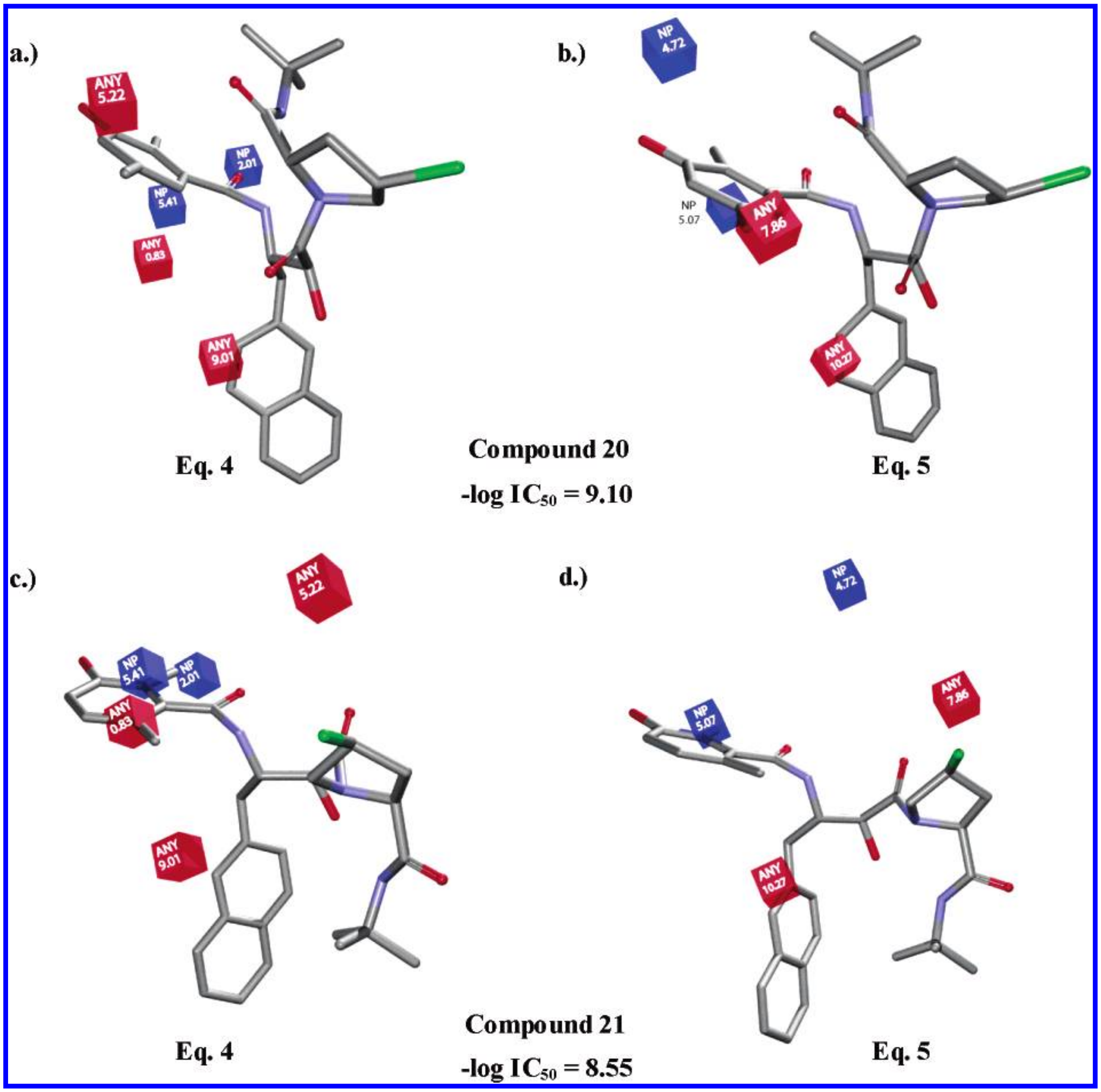


Figure 5. The predicted active conformations of compounds 20 (a,b) and 21 (c,d) by the 4D-QSAR models from alignment 14, eq 4 (a,c) and eq 5 (b,d). The IPE type and magnitude of the regression coefficient is given for each grid cell. Blue grid cells contribute positively to inhibition potency, while red grid cells reduce potency.

both models, [−9.01*GC3(any) for eq 3, and −10.27*GC2(any) for eq 4], in the region of the Ar_2 group has a large effect on inhibition potency. This finding may suggest that there is a spatial restriction/limitation in this region of the receptor, and any substituent on Ar_2 must be either in the meta or para position. Of course, this relationship may also be illustrating a classic ortho effect, limiting energetically favorable active conformations due to the steric bulk of the ortho substituent on Ar_2. Another glimpse of the binding site geometry is realized from the inhibition potency enhancing

GCOD of IPE type “nonpolar” located near the ortho position of Ar_1 for both eqs 4 and 5. This GCOD suggests that there must be a hydrophobic region of the receptor near the Ar_1 ortho position, and potential inhibitors would benefit from exploiting this region with various types of nonpolar ortho substituents in addition to the methyl group. This pharmacophore site of the inhibitor/receptor interaction can be further explored by two other training set compounds, 20 and 21, that differ only in the position of a methyl group on Ar_1, yet their activity difference is nearly a half order in magnitude.

Figure 5 displays the active conformations of compounds 20 and 21 positioned in grid cell lattice space as predicted by eqs 4 and 5. The GCODs located in the Ar_1 and Ar_2 regions are similar in location/IPE type for eqs 4 and 5, while they differ in the proline region, attributing to the distinct nature of the models. The two 4D-QSAR models suggest very similar predicted active conformations for compound 20 and nearly similar conformations for compound 21. However, the predicted conformations for compound 20 compared to compound 21 are significantly different. This may be attributed to the fact that the more active conformation depicted in compound 20 is energetically unfavorable for compound 21, and, therefore, a conformation with a lower energy, and correspondingly a lower activity, is chosen as the active conformation. Another explanation for the difference in activity can be seen with the GCOD term [−0.83*GC2(any)] from eq 4. This activity decreasing

GCOD term is directly occupied by the second ortho methyl group of compound 21. This loss in potency with methyl occupancy suggests that this region of the receptor may not have sufficient space to tolerate more than one ortho group on Ar_1. Comparing the activity of compound 4 with compounds 2 and 23 further supports this observation. Compound 4 contains two ortho methyl substituents and thereby has a lower -log IC_{50} than compound 2 that contains one ortho and one meta methyl substituent and compound 23 that contains a single ortho methyl substituent. Combining the information obtained from these two models and four inhibitors yields a message for this region that varying the nonpolar properties of a single ortho substituent may help to increase potency, but adding a second ortho substituent would most certainly reduce potency.

The sensitivity of model quality to the 14 different alignments may indirectly reflect the mode of action for this set of inhibitor compounds. The alignments that constrain the region of the molecule located near the Ar_1 and Ar_2 substituents, such as alignments 5, 6, 12, 13, and 14, produce models with the highest r^2 and xv-r^2 values. This may mirror the need for that region of the molecule to be less mobile in the active site, while the proline region is allotted some conformational freedom. This possible mode of binding behavior is also indicated in the predicted active conformations of compounds 20 and 21 in Figure 5. The two compounds show similar conformations for the Ar_1 and Ar_2 regions, while the conformation of the proline region differs. Therefore, future efforts at improved inhibitors with increased potency may focus on creating a rigid framework for the Ar_1 and Ar_2 region in the interest of reducing the entropic penalty paid by the molecules in the process of assuming the proper conformation.

Overall, the 4D-QSAR models presented in this study effectively discriminate small potency differences among very potent analogue inhibitors. A training set of potent inhibitors with a small range in activity, as presented in this paper, would not likely capture the baseline interactions leading to the high -log IC_{50} values but rather characterize the interactions leading to the perturbations from baseline high potency. From this work it is apparent that the interactions driving high inhibition potency are nonpolar/hydrophobic in nature, although it is also possible that steric crowding involving the ortho position of the Ar_1 region of the training set compounds is crucial to high potency. Focusing on these interactions (or regions of the inhibitors' structures) may be beneficial in the search for additional potent HIVPR inhibitors having the AHPBA core.

ACKNOWLEDGMENT

We gratefully acknowledge support from the Laboratory of Molecular Modeling and Design at the University of Illinois at Chicago and from The ChemBats21 Group, Incorporated.

REFERENCES AND NOTES

(1) Wlodawer, A.; Miller, M.; Jaskolski, M.; Sathyanarayana, B. K.; Baldwin, E.; Weber, I. T.; Selk, L. M.; Clawson, L.; Schenider, J.; Kent, S. B. H. Structure of complex of synthetic HIV-1 Protease with a substrate-based inhibitor at 2.3 Å resolution. *Science* **1989**, *246*, 1149−1152.

(2) Navia, M. A.; Fitzgerald, P. M. D.; McKeever, B. M.; Leu, C.-T.; Heimbach, J. C.; Herber, W. K.; Sigal, I. S.; Darke, P. L.; Springer, J. P. Three-dimensional structure of aspartyl protease from human immunodeficiency virus HIV-1. *Nature* **1989**, *337*, 615−620.

(3) Roberts, N. A.; Craig, J. C.; Duncan, I. B. HIV proteinase inhibitors. *Bio. Soc. Trans.* **1992**, *20*, 513−516.

(4) Sakurai, M.; Sugano, M.; Handa, H.; Komai, T.; Yagi, R.; Nishigaki, T.; Yabe, Y. Studies of HIV-1 protease inhibitors. I. Incorporation of a reduced peptide, simple amino alcohol, and statine analogue at the scissile site of substrate sequences. *Chem. Pharm. Bull.* **1993**, *41*, 1369−1377.

(5) Kaldor, S. W.; Kalish, V. J.; Davies, J. F.; Shetty, B. V.; Fritz, J. E.; Appelt, K.; Burgess, J. A.; Campanale, K. M.; Chirgadze, N. Y.; Clawson, D. K.; Dressman, B. A.; Hatch, S. D.; Khalil, D. A.; Kosa, M. B.; Lubbehusen, P. P.; Muesing, M. A.; Patrick, A. K.; Reich, S. H.; Su, K. S.; Tatlock, J. H. Viracept (nelfinavir mesylate, AG1343): a potent, orally bioavailable inhibitor of HIV-1 protease. *J. Med. Chem.* **1997**, *40*, 3979−3985.

(6) De Lucca, G. V.; Liang, J.; Aldrich, P. E.; Calabrese, J.; Cordova, B.; Klabe, R. M.; Rayner, M. M.; Chang, C.-H. Design, synthesis, and evaluation of tetrahydropyrimidines as an example of a general approach to nonpeptide HIV protease inhibitors. *J. Med. Chem.* **1997**, *40*, 1707−1719.

(7) Gupta, S. P.; Babu, M. S. Quantitative structure−activity relationship studies on cyclic cyanoguanidines acting as HIV-1 protease inhibitors. *Bioorg. Med. Chem.* **1999**, *7*, 2549−2553.

(8) Hagen, S. E.; Domagala, J.; Gajda, C.; Lovdahl, M.; Tait, B. D.; Wise, E.; Holler, T.; Hupe, D.; Nouhan, C.; Urumov, A.; Zeikus, G.; Zeikus, E.; Lunney, E. A.; Pavlovsky, A.; Gracheck, S. J.; Saunders: J.; VanderRoest, S.; Brodfuehrer, J. 4-Hydroxy-5,6-dihydropyrones as inhibitors of HIV protease: the effect of heterocyclic substituents at C-6 on antiviral potency and pharmacokinetic parameters. *J. Med. Chem.* **2001**, *44*, 2319−2332.

(9) Ratner, L.; Haseltine, W.; Patarca, R.; Llvak, K. J.; Starcich, B.; Josephs, S. F.; Doran, E. R.; Ratalski, J. A.; Whitehorn, E. A.; Baumeister, K.; Ivanoff, L.; Petteway, S. R.; Pearson, M. L.; Lautenberger, J. A.; Papas, T. S.; Ghrayeb, J.; Chang, N. T.; Gallo, R. C.; Wong-Staal, F. Complete nucleotide sequence of the AIDS virus, HLTV-III. *Nature (London)* **1985**, *313*, 277−284.

(10) Kohl, N. E.; Emini, E. A.; Schleif, W. A.; Davis, L. J.; Meimbach, J. C.; Dixon, R. A. F.; Scolnick, E. M.; Sigal, I. S. Active human immunodeficiency virus protease is required for viral infectivity. *Proc. Natl. Acad. Sci. U.S.A.* **1988**, *85*, 4686−4690.

(11) Ashorn, P.; McQuade, T. J.; Thaisrivongs, S.; Tomasselli, A. G.; Tarpley, W. G.; Moss, B. An inhibitor of the protease blocks maturation of human and simian immunodeficiency viruses and spread of infection. *Proc. Natl. Acad. Sci. U.S.A.* **1990**, *87*, 7472−7476.

(12) McQuade, T. J.; Tomasselli, A. G.; Liu, L.; Karacostas, V.; Moss, B.; Sawyer, T. K.; Heinrikson, R. L.; Ratpley, W. G. A synthetic HIV-1 protease inhibitor with antiviral activity arrests HIV-like particle maturation. *Science* **1990**, *247*, 454−456.

(13) Roberts, N. A.; Martin, J. A.; Kinchington, D.; Broadhurst, A. V.; Craig, J. C.; Duncan, I. B.; Galpin, S. A.; Handa, B. K.; Kay, J.; Krohn, A.; Lambert, R. W.; Merrett, J. H.; Mills, J. S.; Parkes, K. E. B.; Redshaw, S.; Ritchie, A. J.; Taylor, D. L.; Thomas, G. J.; Machin, P. J. Rational design of peptide-based HIV proteinase inhibitors. *Science* **1990**, *248*, 358−361.

(14) Sakurai, M.; Higashida, S.; Sugano, M.; Komai, T.; Yagi, R.; Ozawa, Y.; Handa, H.; Nishigaki, T.; Yabe, Y. Structure−activity relationships of HIV-1 PR inhibitors containing AHPBA. *Bioorg. Med. Chem.* **1994**, *2*, 807−825.

(15) Komai, T.; Higashida, S.; Sakurai, M.; Nitta, T.; Kasuya, A.; Miyamaoto, S.; Yagi, R.; Ozawa, Y.; Handa, H.; Mohri, H.; Yasuoka, A.; Oka, S.; Nishigaki, T.; Kimura, S.; Shimada, K.; Yabe, Y. Structure−activity relationships of HIV-1 PR inhibitors containing AHPBA II. Modification of pyrrolidine ring at P1′ proline. *Bioorg. Med. Chem.* **1996**, *4*, 1365−1377.

(16) Takashiro, E.; Wantanabe, T.; Nitta, T.; Kasuya, A.; Miyamoto, S.; Ozawa, Y.; Yagi, R.; Nishigaki, T.; Shibayama, T.; Nakagawa, A.; Iwamoto, A.; Yabe, Y. Structure−activity relationships of HIV-1 protease inhibitors containing AHPBA III. Modification of P2 site. *Bioorg. Med. Chem.* **1998**, *6*, 595−604.

(17) Huang, X.; Xu, L.; Luo, X.; Fan, K.; Ji, R.; Pei, G.; Chen, K.; Jiang, H. Elucidating the inhibiting mode of AHPBA derivatives against HIV-1 protease and building predictive 3D-QSAR models. *J. Med. Chem.* **2002**, *45*, 333−343.

(18) Hopfinger, A. J.; Wang, S.; Tokarski, J. S.; Jin, B.; Albuquerque, M.; Madhav, P. J.; Duraiswami, G. Construction of 3D-QSAR models using the 4D-QSAR analysis formalism. *J. Am. Chem. Soc.* **1997**, *119*, 10509−10524.

(19) 4D-QSAR, Version 3.0; The Chem21 Group, Inc., 1780 Wilson Drive, Lake Forest, IL 60045, 2002.

(20) *HyperChem Program Release 6.01 for Windows;* Hypercube, Inc., 2000.

(21) *MOLSIM V3.0;* D. C. Doherty and The Chem21 Group, Inc., 1780 Wilson Drive, Lake Forest IL 60045, 1998.

(22) Rogers, D.; Hopfinger, A. J. Application of genetic function approximation to quantitative structure−activity relationships and quantitative structure−property relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854−866.

(23) Terfloth, L.; Gasteiger, S. Neural networks and genetic algorithms in drug design. *Drug Discovery Today 6 (15 suppl*) **2001**, s102−s108.

(24) Cramer, R. D.; Patterson, D. E.; Bunce, J. D. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.

(25) Kleinbaum, D. G.; Lawrence, L. K.; Muller, K. E.; Nizam, A. *Applied regression analysis and other multivariate methods*, 3rd ed.; Brooks/Cole Publishing Co.: CA, 1998.

CI0340456