E-State Modeling of HIV-1 Protease Inhibitor Binding Independent of 3D Information

Hlaing Hlaing Maw[†] and Lowell H. Hall*

Department of Chemistry, Eastern Nazarene College, Quincy, Massachusetts 02170

Received August 30, 2001

Data for HIV-1 protease inhibitors (in vitro enzyme binding) were used as a training set to develop a QSAR model based on topological descriptors, including two hydrogen E-state indices, along with a molecular connectivity χ and a κ shape index. A statistically satisfactory four-variable model was obtained for the 32 compounds in the training set, $r^2 = 0.86$, $s = 0.60$, and $q^2 = 0.79$, without the use of information from 3D geometries or detailed interaction energy calculations. The model was validated through the prediction of 15 compounds in the external test set, yielding a mean absolute error, MAE, = 0.82. Structure interpretation is given for each variable to assist in the design of new compounds. Structure features emphasized in the model include hydrogen bond donating ability, nonpolar groups, skeletal branching, and molecular globularity. On the basis of these statistical criteria, this E-state model may be considered useful for prediction of pIC_{50} values for new HIV-1 protease inhibitors.

INTRODUCTION

Over the past decade, developments in modeling the relationship between molecular structure and various drug properties have taken two distinct routes. Approaches that require significant information from 3D molecular geometries of both the enzyme active site as well as for drug molecules are often called rational or structure-based design methods. These methods require individual atom coordinates for the active site (from crystal structure data), minimum energy conformations for inhibitors, individual docking of each inhibitor into the active site, and detailed intermolecular interaction energy calculations. As a result, these methods are time-consuming and expensive but have demonstrated potential for useful drug design.[1,2]

The other line of approach is considerably less time-consuming and less expensive. These methods are based on topological representation of molecular structure and do not require 3D-based geometry information on the active site or on the drug molecules. In this approach, topological superposition of the common core skeleton is used. In addition, the methods for computation are very fast, significantly faster than those methods that require 3D geometry information. The models produced in this topological approach also yield significant structure information for the design of new compounds. E-state descriptors have been used to model binding data, providing excellent statistics.[3−13] The topological structure representation approach has also been successfully applied to heterogeneous data sets with diverse molecular structures.[7,10,11]

Significant research attention is being given to the human immunodeficiency virus (HIV), the retrovirus that is believed to cause acquired immunodeficiency syndrome (AIDS). Many research projects are attempting to find a cure for AIDS. Some researchers are focusing their attention to the HIV-1 protease. The HIV-1 protease is a proteolytic enzyme that is responsible for processing the protein precursors to the structural proteins and enzymes (reverse transcriptase, integrase, and the protease itself) of the HIV-1 virus.[9−14]

In this significant research effort there is need for sound models of the relationship between molecular structure and the HIV-1 protease inhibitor binding affinity. Such models can assist researchers to understand the structure basis of binding as well as provide a basis for development of new compounds. Part of the scientific value of these models is both their ability to lend insight into the aspects of structure that relate to biological activity and their ability to assist in the design of new compounds.

In this paper, we develop a model for HIV-1 protease inhibitor binding and demonstrate its validity by prediction of an external test set. The whole data set consists of 49 observations, split between a training and a test set, and was obtained from Holloway et al.,[15] who used intermolecular interaction energy (E_{inter}) to develop models. In three different active sites, 33 of the 49 compounds were used to create QSAR models and then to predict the remaining 16. In this paper, only topological structure descriptors are employed to create the model. No information based on 3D geometries was used. This paper shows that the topological method applied to the HIV-1 protease inhibitor binding data leads to a model that is statistically somewhat better than the one developed in the literature.[15] Further, the method is considerably less time-consuming and less expensive to apply. Finally, interpretations of significant structures features are obtained by analysis of each of the four structure descriptors in the model.

Topological Descriptors. In addition to speed and cost considerations, an important objective of modeling is to obtain useful information about the structure features that relate significantly the property being modeled. For this present case, we use the molecular structure descriptors known as electrotopological state indices[3−8,16−21] along with molecular connectivity χ indices[22−24] and κ shape indices.[23,25] The E-state indices have been used to develop models for many activities and properties in both their atom-level[3−5]

* Corresponding author phone: (617)745-3550; fax: (617)745-3509; e-mail: Halll@enc.edu.

[†] Current address: Biochemistry Laboratory, Tufts University, 136 Harrison Avenue, MV611, Boston, MA 02111.

10.1021/ci010091z CCC: $22.00 © 2002 American Chemical Society
Published on Web 02/14/2002

and atom-type forms.[3,17−21,26,27] E-state QSAR models yield structure information which reveals structure features significantly related to activity. Further, the more recent development of hydrogen E-state values (and hydrogen atom-type E-state indices[3,21,26]) has extended the capability of the E-state to provide a powerful set of structure descriptors. Several studies have investigated QSAR models of binding,[3,5,21,26] indicating their ability to represent both hydrogen-bonding groups as well as nonpolar regions of molecules. Cross-validation further supports the significance of the E-state models.[26,27] The atom-type E-state structure descriptors have also been shown to be very useful in searching a chemical database for structures similar to a desired target.[17,21,28] In this manner, the descriptors found important in an E-state QSAR model can form the basis for a similarity search of a database, experimental or virtual.

Atom-Level E-State Values. E-state indices have been defined and used in many QSAR and related studies.[3−8,16−21,26−28] Only a brief summary of the method is necessary here. In this topological approach to structure representation, information is developed for each atom (such as $>N-$, $=O$, and $-Cl$) and each hydride group (such as $-CH_3$, $-NH_2$, and $-OH$) in the molecule. For reasons of simplicity, both atoms and hydride groups are often called "atoms". *The E-state value, S(i) from eq 1, calculated for each atom in a molecule, is called the atom-level E-state index* (to distinguish it from the atom type described below). $S(i)$ is composed of an intrinsic state, I_i, plus the sum of perturbations, ΔI_{ij}, from other atoms. The E-state value for atom i in a molecule is computed as follows:

$$S(i) = I_i + \Sigma_j \Delta I_{ij} \quad \text{sum over all other atoms } j \qquad (1)$$

The intrinsic state value, I_i, is defined as the ratio of the valence state electronegativity (given as the Kier−Hall electronegativity[3]) to a measure of the local topology (given as the number of skeletal neighbors). The perturbation term is as follows:

$$\Delta I_{ij} = (I_i - I_j)/r_{ij}^{\ 2} \qquad (2)$$

Here r_{ij} is the topological distance between atoms, given as the number of atoms in the shortest path between atoms i and j. In the manner given by eq 2, each atom's E-state value contains electronic and topological structure information from all other atoms within the structure.[3] The information encoded in the E-state value for an atom is the electron accessibility at that atom. In this sense, the E-state index encodes the potential for noncovalent intermolecular interaction.[3,21] The atoms closest to a given atom have the greatest influence on its E-state $S(i)$ value. Influence diminishes for atoms separated by a path of several bonds; the influence decreases as the square of the number of atoms in the path. *A parallel development provides the basis for hydrogen atom-level E-state indices, Hs(i).*

For this current data set of HIV-1 protease inhibitor binders, there is a common core skeleton (25 atoms in the scaffold; see Table 1 and Chart 1) among the 33 training set molecules. The E-state values for all common skeletal atoms can be used directly as variables in seeking a QSAR model. The corresponding hydrogen atom level E-state indices may also be entered into the data matrix for analysis.

Atom-Type E-State Values. For all data sets, including those with a common skeletal core as well as those with a heterogeneous group of molecules, the atom-type E-state indices provide much useful information. Each atom (or hydride group) in the molecule is classified into an atom type, such as $-OH$, $=O$, or aromatic CH. *The atom type E-state index is the sum of the individual atom level E-state values for a particular atom type.*[3,17,21] The atom-type descriptors combine three important aspects of structure information:

1, electron accessibility for the atoms of the same type;
2, presence/absence of the atom type;
3, count of the atoms in the atom type.

Hydrogen atom-type E-state descriptors form a parallel set except that accessibility refers to hydrogen atom (or proton) accessibility.

In the present HIV-1 protease inhibitor binding training data set, the inhibitor structures possess 25 atom sites in common for all molecules in the training set. The 25 atom-level E-state and hydrogen E-state indices can be used in model development along with atom-type descriptors in addition to molecular connectivity χ indices and κ shape indices.

Model Validation. An important aspect of QSAR modeling is the development of means for validation of the model. Good statistical criteria for fit to the training set are not a guarantee that the model can make accurate predictions for compounds outside the data set. In recent years, the leave-one-out (LOO) press statistic (q^2, s_{press}) has been used as a means of indicating predictive capability. Alternatively, one may set aside a selected part of the data (validation or test set) that is not used in any way to develop the model. Holloway et al. set aside 16 compounds as an external validation set that will be used in this investigation. Mean absolute error (MAE) for the prediction test set will be used as the significant criterion for assessing model quality.

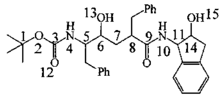
METHODS

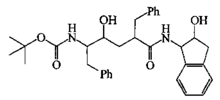
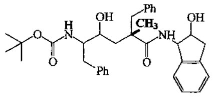
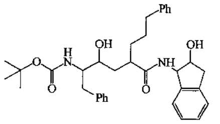
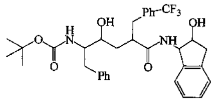
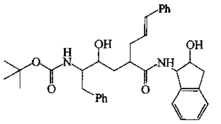
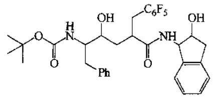
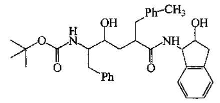
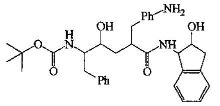
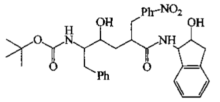
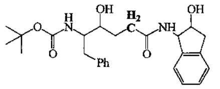
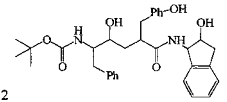
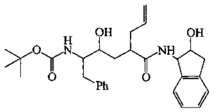
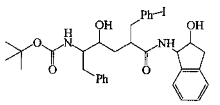
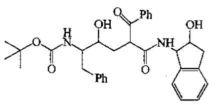
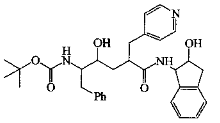
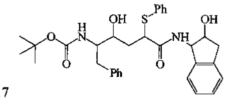
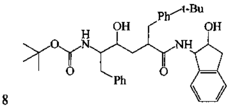
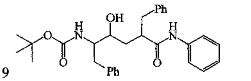
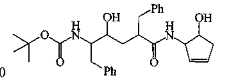
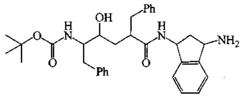
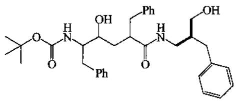
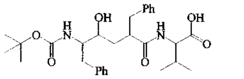
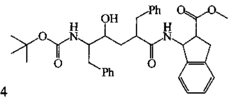
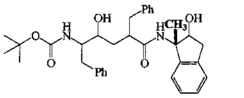
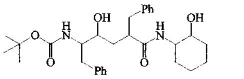
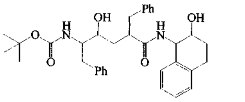
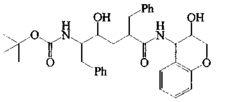
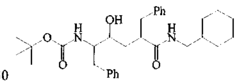
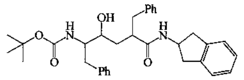
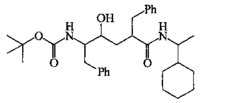
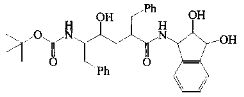
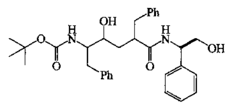
Data Entry. The inhibition data and molecular structures were taken from Holloway et al.[15] Molecular structures of 32 HIV-1 protease inhibitors are given in Table 1 along with Holloway's molecule designations under the heading "symbol". In the training set, compounds HIV-28 and HIV-29 are a chiral pair. The topological descriptors used in this investigation do not distinguish among chiral structures. For our purposes, we deleted the structure corresponding to the lower binding constant (no. 29, 7.3925) and kept the one with the higher value (no. 28, 9.7447), believing that the compound with the stronger binding more closely corresponds to the topology of the active site. As a result, there are 32 compounds in the training set.

There is one pair of chiral structures in the prediction set (compound 35−50). We deleted compound HIV-42 and kept compound HIV-43 because its pIC_{50} value (10.2676) is greater than that of compound HIV-42 (8.2676), leaving 15 compounds in the prediction set.

Compounds were entered as drawings with ChemDraw[29] and structure data saved as MDL mol files. The atoms in the common core of all molecules were numbered identically as shown in Table 1. (Only first 15 numbers shown.) All structure indices used in this investigation were computed from Molconn-Z, ver 3.50.[30] Structure input was validated

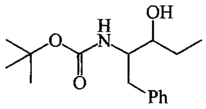
Table 1. HIV-1 Protease Inhibitor Structures and Binding Values (pIC50): Training Set



Id / Symbol[a]	Structure Drawing	pIC50 Obs[b]	pIC50 Calc[c]	pIC50 Res[d]
1- HIV-1		9.6021	9.38	0.22
2- HIV-3		8.1135	7.76	0.45
3- HIV-4		9.7212	9.62	0.10
4- HIV-5		9.5850	9.57	0.01
5- HIV-6		9.6383	9.74	-0.11
6- HIV-7		9.2218	8.16	1.39
7- HIV-8		9.5376	9.32	0.22
8- HIV-9		9.5086	9.16	0.35
9- HIV-10		9.5686	9.29	0.28
10- HIV-11		5.5325	5.36	0.17
11- HIV-12		9.7959	9.22	0.58
12- HIV-13		7.5607	8.04	-0.48
13- HIV-14		9.1427	9.25	-0.11
14- HIV-15		8.2660	9.49	-1.22
15- HIV-16		9.2757	9.16	0.11
16- HIV-17		9.6021	9.36	0.24
17- HIV-18		9.7696	10.04	-0.27
18- HIV-19		6.9431	6.34	0.60
19- HIV-20		8.0209	8.13	-0.11
20- HIV-21		7.4653	7.15	0.31
21- HIV-22		6.1612	6.64	-0.48
22- HIV-23		6.7932	7.40	-0.61
23- HIV-24		7.1785	7.59	-0.41
24- HIV-25		6.6728	7.70	-1.03
25- HIV-26		6.9144	7.70	-0.79
26- HIV-27		9.1549	9.42	-0.27
27- HIV-28		9.7447	9.40	0.33
28- HIV-30		4.5229	5.46	-0.94
29- HIV-31		6.8861	7.15	-0.27
30- HIV-32		6.8356	5.77	1.06
31- HIV-33		10.0000	9.69	0.31
32- HIV-34		7.4134	6.68	0.73

Note: Topological differences in the common core skeleton are indicated with bold bonds and/or substituents for structures HIV-3, HIV-11, HIV-22, HIV-25, and HIV-34. See text for discussion. [a] Symbol: designation used by Holloway.[15] [b] Experimental value for $-\log(IC_{50})$ = pIC50 for inhibitor binding. [c] pIC50 value calculated from the QSAR model, eq 3. [d] pIC50 − calc.

Chart 1



by visual inspection of the ChemDraw drawings as well as the structure analysis provided by Molconn-Z output.

For QSAR analysis, we selected 19 atom-level E-state and 14 (nonzero) hydrogen E-state indices and 25 nonzero atom-type E-state and hydrogen E-state indices, along with molecular connectivity χ indices and κ shape indices that have nonzero variance. The pairwise correlation matrix was examined for correlation coefficients greater than 0.80. For each such occurrence, one of the pair of correlated variables was eliminated. Selection of a variable to be retained is primarily experience-based. Preference for retention is given to variables thought to be more easily interpreted in terms of molecular structure. For example, the valence molecular connectivity χ path indices are correlated with the first-order χ valence index ${}^1\chi^v$ ($r > 0.80$). We selected the ${}^1\chi^v$ index (deleting the others) to include in the modeling because it may be more easily interpretable. (See interpretation in a later section). However, since ${}^1\chi^v$ is found in the best model, we also tested ${}^2\chi^v$. In fact, the ${}^2\chi^v$ index is selected in the final model because it leads to the best MAE for the prediction test set. Many high intercorrelations occur in the data matrix because of the large (constant) common core structure. Training set compounds exhibit little structure variation in the region of atoms numbered 1−7 (See Table 1). One significant part of the structures containing 20 atoms (see Chart 1) is present in all training set structures. After data matrix reduction, 29 variables remained for statistical analysis in model development.

Statistical Analysis. The data matrix was submitted for statistical analysis using the SAS system.[31] The RSQUARE selection method in proc REG was used to examine every QSAR equation from 1 to 5 variables, listing the top 10 most statistically significant. The number of variables was limited to 5 to preserve a reasonable ratio of the number of observations to number of variables. RSQUARE is not a stepwise procedure; all possible sets of variables are considered, and those with the largest r^2 values are listed. Eight variables appear important because they occur several times in the equations:

HS^T(HBd), sum of hydrogen E-state values for hydrogen bond donors;

HS^T(other), sum of hydrogen E-state values for nonpolar CH bonds;

Hs[CH(11)], Hs[NH(10)], Hs[CH(8)], hydrogen E-state value for individual atoms;

S[C(2)], E-state value for individual atom;

S^T(OH), atom-type E-state for −OH group;

${}^1\chi^v$, first-order molecular connectivity χ index;

${}^2\kappa_\alpha$, ${}^3\kappa_\alpha$, second- and third-order κ shape indices.

Most prominent among these descriptors are E-state indices, both atom type and atom level. These structure descriptors emphasize hydrogen bond donating ability, HS^T(HBd), Hs(NH_{10}), and S^T(−OH), nonpolar structure features, HS^T(other), Hs[CH(11)], and s[C(2)], hydrogen accessibility at a few atomic sites, Hs[CH(11)], Hs[(NH(10)], and Hs[(CH(8)], and skeletal structure (χ and κ indices).

The best single-variable equation involves HS^T(HBd); the single best two-variable equation includes HS^T(HBd) and HS^T(other). The HS^T(HBd) descriptor is the sum of hydrogen E-state values (Hs(*i*)) for hydrogen bond donor groups. HS^T(other) is the sum of hydrogen E-state values for nonpolar CH groups. Examination of the training set structures reveals that four groups in the common core skeleton are capable of forming hydrogen bonds: NH(4); OH(13); NH(10); OH(15). (See Table 1 for atom numbering scheme and individual structures.) The structure environments of NH(4) and OH(13) are nearly constant in the training set, and their Hs values do not vary significantly. Consequently, the HS^T(HBd) descriptor was redefined to include only Hs values from NH(10) and OH(15):

$$HS^T(HBd) = Hs[NH(10)] + Hs[OH(15)]$$

(Model equations that included Hs[NH(4)] and Hs[OH(13)] yielded statistical tests that indicate their Hs values do not make significant contribution to the equation; that is, they have *t* values considerably less than 1.0.)

A new model was formed using the redefined HS^T(HBd) descriptor along with HS^T(other). Calculated pIC50 values based on this equation indicated three compounds that are significantly overpredicted: nos. 3, 11, and 25. Examination of the structures for these compounds indicates that the topology of their core skeletons is significantly different from that of the remaining 29 structures. For structures HIV-3 and HIV-11, topology differs at atom no. 8. For compound HIV-25, the difference occurs at atoms no. 11. These variations in core skeleton topology are indicated as bold bonds in Table 1. For compounds HIV-22 and HIV-34, a difference in core skeletal topology occurs at atom nos. 11 and 15; their residuals are not large like those for compounds 3, 11, and 25. No other deviations in topology occur in the common core skeleton structures.

We formulate the hypothesis that because of the variations in the common backbone in each of these cases, the OH group at atom site no. 15 does not form an effective hydrogen bond. Other structure variations do not have this effect on the hydrogen-bonding scheme. To implement the effects of this hypothesis, the HS^T(HBd) descriptor was modified for these five cases; the hydrogen E-state value for atom no. 15 [Hs(15)] was removed from HS^T(HBd). To test the effects of this uniformly applied hypothesis, we examined the regression statistics when the modified (HBd) descriptor is used in the model. Statistics for the resulting equation improved significantly. The SAS RSQUARE selection was rerun for a number of variables from one to five, using the redefined and modified variable HS^T(HBd). We proceeded to examine the best equations for each number of variables from one to four.

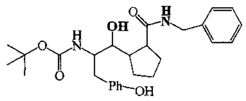
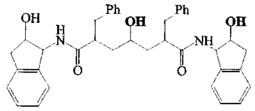
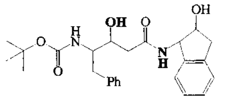
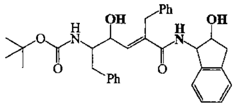
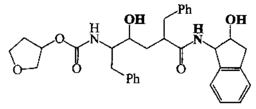
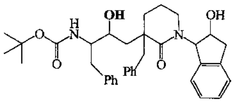
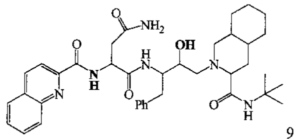
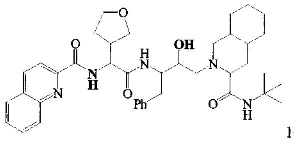
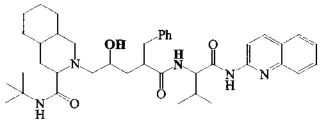
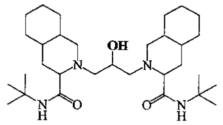
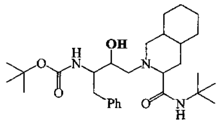
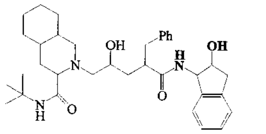
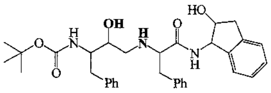
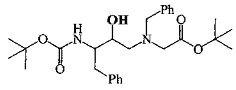
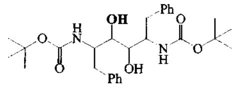
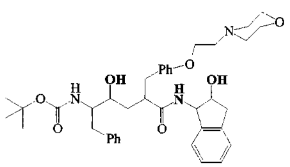
The best one-variable equation is based on the redefined HS^T(HBd) ($r^2 = 0.62$, $q^2 = 0.56$).

The best two-variable equation uses HS^T(HBd) and HS^T(other) ($r^2 = 0.81$, $q^2 = 0.76$).

The best three-variable equation includes HS^T(HBd), HS^T(other), and ${}^1\chi^v$ ($r^2 = 0.82$, $q^2 = 0.74$).

The best four-variable equation includes HS^T(HBd), HS^T(other), ${}^1\chi^v$, and ${}^2\kappa_\alpha$ ($r^2 = 0.86$, $q^2 =0.79$). Because ${}^1\chi^v$ is intercorrelated with higher order χ path indices, statistics were obtained for the four-variable equation by substituting

Table 2. HIV-1 Protease Inhibitor Structures and Binding Values: Test Set

Id / Symbol[a]	Structure Drawing	pIC50		
		Obs[b]	Calc[c]	Res[d]
1- HIV-35		6.2299	5.36	0.87
2- HIV-36		9.1612	11.37	-2.21
3- HIV-37		6.2457	5.41	0.84
4- HIV-38		8.8861	9.34	-0.46
5- HIV-39		10.2220	9.48	0.74
6- HIV-40		5.8965	7.11	-1.21
7- HIV-41		9.6383	9.05	0.59
8- HIV-43		10.2676	10.00	0.27
9- HIV-44		7.2774	8.07	-0.79
10- HIV-45		5.1675	4.50	0.67
11- HIV-46		5.5229	4.27	1.25
12- HIV-47		8.1163	8.13	-0.01
13- HIV-48		6.6402	5.73	0.91
14- HIV-49		5.3279	3.98	1.35
15- HIV-50		5.8617	5.74	0.13
16- HIV-2		9.35[e]	9.88	-0.53

Note: Groups that are hypothesized to engage in hydrogen bonding in the active site are indicated in bold as −**NH**− or −**OH**. [a]Symbol: designation used by Holloway.[15] [b] Experimental value for $-\log(\text{IC}_{50})$ = pIC50 for inhibitor binding. [c] pIC50 value calculated from eq 3. [d] pIC50 − calc. [e] See ref 32.

higher order χ indices for ${}^{1}\chi^{v}$. The equation using ${}^{2}\chi^{v}$ yielded direct statistics very similar to those with ${}^{1}\chi^{v}$. However, as indicated in the next section, the best MAE value for test set prediction was also obtained using ${}^{2}\chi^{v}$. The final model, using ${}^{2}\chi^{v}$, is given as eq 3 in the Discussion. The observed, calculated, and residual values are given in Table 1.

The implementation of the hydrogen-bonding hypothesis [redefining and modifying HS^{T}(HBd)] has greatly improved the model because the observations with the large residuals no longer have such large influence on the regression, thus allowing the other descriptors to better represent their encoded structure information. Such improvement lends support to the hypothesis; however, the results of the two validation studies give even greater support to the hypothesis.

Validation Study. The QSAR model (eq 3) was employed to predict the pIC50 values in the external validation test set. The mean absolute error for the prediction test set is MAE = 0.82. The correlation between the predicted values and the observed pIC50 values is $r^2 = 0.84$. The primary challenge in performing predictions occurred in assigning which potential hydrogen-bonding groups are hypothesized to form hydrogen bonds. This approach is presented in the Discussion. The 15 predicted pIC_{50} values are shown in Table 2 along with the observed, calculated, and residual values and also in Figure 1.

Compound HIV-2 was not included in the Holloway study. We have used that compound as a second external validation study. Its experimental value was obtained from Thompson et al..[32]

RESULTS AND DISCUSSION

QSAR Model. The model based on the selected four variables yielded statistical information as follows:

$$\text{pIC50} = 0.731(\pm 0.086)\text{HS}^{\text{T}}(\text{HBd}) + 0.283(\pm 0.058)\text{HS}^{\text{T}}(\text{other}) + 0.312(\pm 0.164)\,{}^{2}\chi^{v} - 0.420(\pm 0.156)\,{}^{2}\kappa_{\alpha} - 0.518(\pm 1.970) \quad (3)$$

$$r^2 = 0.86 \quad s = 0.60 \quad n = 32 \quad q^2 = 0.79 \quad s_{\text{press}} = 0.74$$

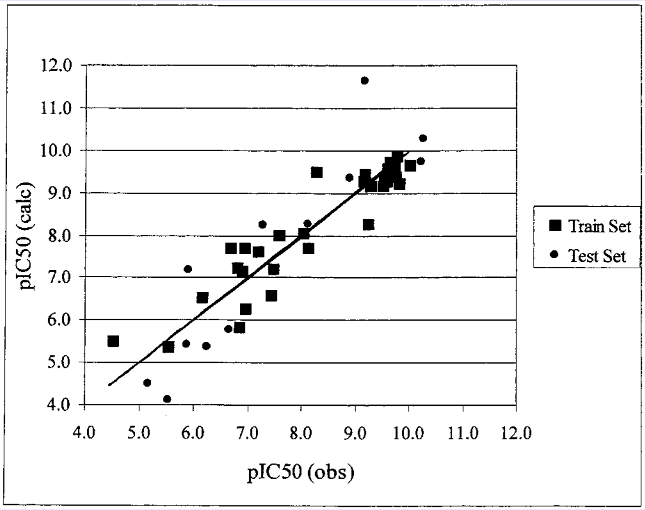


Figure 1. Plot of calculated pIC50 values (QSAR model, eq 3) for HIV-1 protease inhibitor data set versus observed pIC50 values for both the training set (marked by squares) and the test set (marked by dots).

Quantities in parentheses are the standard deviations of the coefficients; q^2 and s_{press} are based on the leave-one-out (LOO) method. A plot of the calculated pIC$_{50}$ versus observed pIC$_{50}$ is given in Figure 1. An examination of the plot of residuals versus observed pIC$_{50}$ (not shown) revealed no trends and appears randomly distributed. The observed pIC50 values, calculated (calc), residuals (res), and predicted residuals (pres from LOO), are given in Table 1. The variables in the model (eq 3) are essentially independent; the largest intercorrelation is $r^2 = 0.48$ for HS^{T}(other) and $^{2}\kappa_{\alpha}$; the next largest is $r^2 = 0.31$ for $^{2}\chi^{v}$ and $^{2}\kappa_{\alpha}$.

The QSAR model for the HIV-1 protease binding data was selected on the basis of three statistical criteria: r^2 and q^2 for the direct model and the mean absolute error (MAE) for the test set and the predicted value for compound HIV-2. The MAE value was considered the most important statistic because an equation should not be considered a QSAR model unless it has demonstrated predictive value. For this present data set, the MAE is at a minimum value for the four-variable model, as follows: for one variable, MAE = 1.06; for two variables, 1.07; for three variables, 0.91; for four variables, 0.82; for five variables, 1.04.

Interpretation of the Model. The four-variable model (eq 3) successfully represents the pIC50 data for both training and test sets. Each of the variables is a descriptor of molecular structure and will be discussed at this point to indicate the specific structure information encoded in each.

HS^{T}(HBd) Descriptor. The variable with the single best correlation to pIC50 is HS^{T}(HBd), a representation of hydrogen bond donation ability.[3,21,26] The descriptor, HS^{T}(HBd), is the sum of the hydrogen E-state values for groups that act as hydrogen bond donors. In this present data set, these groups include NH(10) and OH(15). The group NH(10) is present is all the training set structures. The group OH(15) is present in all but eight members of the training set.

An indication of the significance of the HS^{T}(HBd) descriptor is gained by examining the list of the training set structures that is rank-ordered on HS^{T}(HBd). Ranking the data set on each of the descriptors is a useful way to explore the structure implications of the model. Fourteen of the 16 largest pIC50 values are ranked with the 16 largest HS^{T}(HBd) values. The 5 most strongly binding compounds are ranked in the top 11 HS^{T}(HBd) values. Also 14 of the 16 smallest pIC50 values are ranked with the 16 smallest HS^{T}(HBd) values. Clearly, the hydrogen-bonding abilities of the designated groups [NH(10) and OH(15)] are important in this data set. As a single variable, the correlation between pIC$_{50}$ and HS^{T}(HBd) is $r^2 = 0.62$ and $q^2 = 0.56$. The HS^{T}(HBd) variable contributes on average 13.3% to the calculated pIC50 value and ranges from 7.2% to 17.7%. Because of its positive coefficient, the larger the value of HS^{T}(HBd), the greater the value of calculated pIC$_{50}$.

An important consideration in the structure interpretation from this model is the ability to estimate the contribution of hydrogen bonding groups to the calculated pIC50 value. The actual contribution to pCI50 is the product of the hydrogen E-state value (Hs) of the group times the coefficient of HS^{T}(HBd) in the model, 0.731. For the training set, the average Hs value for OH(15) is 2.64, typical of OH groups bonded to saturated carbons. The average contribution of the OH(15) group to pIC50 is 1.93. The influence of substituents near the OH group in this data set causes its contribution to vary from 1.86 to 2.00. The encoding of the influence of molecular environment on hydrogen bond donating ability results from the nature of the hydrogen E-state formalism. This encoding is not a mere count of donors, but it is a variable measure of hydrogen bond donating ability arising from hydrogen atom accessibility for intermolecular interactions.[3,17] For the NH(10) group in the core skeleton, the average Hs(10) value is 1.48, making its typical contribution 1.35. The influence of structure variation causes the NH(10) contribution to vary from 1.35 to 1.60. This structure information on these two groups may be of assistance in the design of new inhibitors.

HS^{T}(other) Descriptor. The second variable in the best two-variable equation we discuss next, HS^{T}(other), the sum of the hydrogen atom-level E-state indices for all nonpolar hydrogen atoms in the molecule.[3,21,26] When the training set is rank-ordered on HS^{T}(other), 10 of the 16 largest pIC50 values are ranked with the 16 largest HS^{T}(other) values. Likewise smaller pIC50 values are ranked with smaller HS^{T}(other) values. For the two-variable equation, the correlation coefficient is $r^2 = 0.81$ and $q^2 = 0.76$. When the training set is rank-ordered on the pIC50 value calculated from the two-variable model, 15 of the largest pIC50 values are ranked with the 16 largest calculated values. Such information gives strong confirmation of the significance of these two structure descriptors for this training set. The HS^{T}(other) descriptor contributes the largest percentage of the calculated pIC50 values: 40.2% on the average and ranges from 35.3% to 43.8%. Because of the positive coefficient on HS^{T}(other), larger values are related to larger pIC$_{50}$ values. In the training set, two benzyl groups and the *tert*-butyl group contribute significantly to HS^{T}(other), along with nonpolar portions of substituents.

The contribution of nonpolar groups to the calculated pCI50 value may be computed from the product of its HS^{T}(other) contribution and its coefficient in the model, 0.283. The typical contribution of the benzyl group at position no. 5 to pIC50 (based on the sum of hydrogen E-state values) is computed to be 1.8. A phenyl group contributes typically 1.5, and a *tert*-butyl group contributes 0.4. The contribution of any individual nonpolar CH group may be computed from its Hs value times the coefficient. (Molconn-Z provides Hs values for each atom in the structure.) This information may also be of assistance in the design of new inhibitors.

Because the HS^T(other) descriptor is the sum for all nonpolar hydrogen atoms, it is possible to partition HS^T(other) among four atom-type hydrogen E-state indices:

HS^T(other) = HS^T(Csats) + HS^T(Csatu) + HS^T(arom) + HS^T(rest).[3]

These descriptors represent hydrogen atoms in different molecular environments:

HS^T(Csats), hydrogens on saturated carbon atoms bonded to saturated carbon atoms; HS^T(Csatu), hydrogens on saturated carbon atoms bonded to unsaturated carbon atoms; HS^T(arom), hydrogens on aromatic carbon atoms;

HS^T(rest), hydrogens on other carbon atoms, including vinyl and acetylenic.

When substituted for the HS^T(other) descriptors, none of these descriptors gives the same quality statistics of the HS^T(other) descriptor. This analysis indicates that no one of these types of nonpolar regions of the carbon skeleton dominates the HS^T(other) descriptor.

This discussion indicates the combined significance of specific hydrogen bond donating groups and of nonpolar regions of inhibitors toward their binding. A good general account of the ranking of compounds is obtained from the two-variable model. The other two structure descriptors provide detailed molecular architecture or skeletal information.

$^2\chi^v$ Descriptor. The third variable in the model is the second-order valence molecular connectivity χ index, $^2\chi^v$. This variable increases with increased branching in the structure; it is a measure of overall skeletal variation.[22,24] The $^2\chi^v$ index also increases with molecular size, including atom count. However, the $^2\chi^v$ descriptor contains much more structure information, including skeletal branching and heteroatom content. The largest $^2\chi^v$ value is recorded for HIV-18, the largest structure in the training set with a *para-tert*-butyl group on the benzyl group at position no. 8. The *tert*-butyl group is also the most branched substituent in the training set, adding further to the $^2\chi^v$ value. The smallest $^2\chi^v$ value occurs for HIV-11, the smallest structure, the only one with no benzyl group at atom no. 8. (See Table 1.) When added as the third variable in the equation, the correlation improves to $r^2 = 0.82$ and $q^2 = 0.75$; for the prediction test set, MAE = 0.91. The $^2\chi^v$ contributes 17.9% on the average to the calculated pIC_{50} and ranges from 16.5% to 19.9%. Because of the positive coefficient on $^2\chi^v$, larger values are related to larger pIC_{50} values.

The structure information encoded in $^2\chi^v$ does not describe the major influence on binding; the $^2\chi^v$ index should be understood as providing fine-tuning of structure information. For compounds with strong hydrogen bond donation and adequate nonpolar groups, binding is improved with greater size and increased branching.

$^2\kappa_\alpha$ Descriptor. The fourth variable in the model, $^2\kappa_\alpha$, is the second-order κ shape index. The $^2\kappa_\alpha$ index encodes molecular globularity; it decreases with increasing degree of globularity.[23,25] None of the structures in the data set is highly globular; however, HIV-11 has the smallest $^2\kappa_\alpha$ value (no benzyl group at position no. 8), making it the structure with the highest degree of globularity in the data set. The $^2\kappa_\alpha$ descriptor contributes 28.7% on the average to the calculated pIC_{50} and ranges from 26.4% to 33.6%. Because of the negative coefficient on $^2\kappa_\alpha$, smaller values (greater tendency to globularity) are related to larger pIC_{50} values.

The addition of the whole molecule indices, $^2\kappa_\alpha$ and $^2\chi^v$, to the more structure-specific descriptors, HS^T(HBd) and HS^T(other), improves the statistics from $r^2 = 0.81$, $q^2 = 0.76$, and MAE = 1.07 to $r^2 = 0.86$, $q^2 = 0.79$, and MAE = 0.82. These two whole molecule descriptors are understood as providing molecular architecture information to fine-tune the model for prediction.

In summary, the model given as eq 3 indicates that new candidate structures for increased binding should possess strong hydrogen-bonding groups, appropriately placed in the skeleton, should incorporate substantial nonpolar substituents, should have significant skeletal branching, and should possess more globularity. Candidate structures can be evaluated values by using eq 3 for prediction for pIC50.

Validation Study. To assess the potential for predictive ability of the E-state model (eq 3), the pIC50 value for each of the structures in the test set was predicted. The primary challenge in this effort was determining whether a potential hydrogen bond donor group is actually engaged in hydrogen bonding. We note that, in the training set, certain topological features were hypothesized to preclude OH(15) from forming an effective hydrogen bond. The same hypothesis was applied for the structures in the test set, through examination of the topology of the molecular skeleton of each member of the test set. For example, HIV-35 does not possess an OH group at position no. 15; hence, no value is included for Hs(15) in HS^T(HBd). On the other hand, HIV-38 does possess an OH at position no. 15 and its skeleton is very similar to those in the training set; Hs[OH(15)] is retained in HS^T(HBd). By contrast, HIV-40 possesses a six-membered ring involving position nos. 8 and 10, a topology very different from the training set, suggesting that the OH at position no. 15 does not engage in a hydrogen bond; Hs[OH(15)] is removed from HS^T(HBd). Likewise, HIV-48 possesses a topologically different arrangement for OH(15), suggesting that the OH at position no. 15 does not engage in a hydrogen bond; Hs[OH(15)] is removed from HS^T(HBd). In this manner, assignments were made as indicated by bold bonds and groups in Table 2. See the notes at the bottom of Table 2.

On the basis of the calculated values from eq 3, the mean absolute error of the average predictions is computed as MAE = 0.82. The corresponding root-mean-squared error is found to be rms = 0.95. An examination of these statistical values along with the residual values indicates reasonable predictive quality for the model. One compound (HIV-36) has a large residual, 2.21, which is 3.7 times the standard deviation of the regression (eq 3) (See Table 2). We note that the 3D-based model found a somewhat larger prediction residual for HIV-36: 2.85.[15] Furthermore, the MAE for the prediction test set reported in the literature is MAE = 0.93, a higher value than found with the present model, 0.82.

In a further validation of the QSAR model, we predicted compound number HIV-2 from Holloway's paper. A larger *para* substituent occurs on the benzyl group at position no. 8: $-OCH_2CH_2$-morpholino (HIV-2 is shown as entry 16 in

Table 2). Our model (eq 3) predicts the binding activity of that compound as 9.878. Experimental data (IC_{50} = 0.45 nM) for HIV-2 were obtained from Thompson et al. as entry no. 29 in Table 1.[32] The experimental pIC50 value, 9.35, is in very good agreement with the value predicted from eq 3; the residual, −0.53, is less than the standard error of our QSAR model, 0.60. This prediction is further confirmation of the quality of the topological model developed in this paper.

To predict binding for compounds not in the original data set compounds, we suggest the following approach. First, the structure information described above can be used to design new compounds. Second, one can use the model based on the full data set, eq 3, to calculate pIC50 for each new compound. We note that the press statistic standard error, $s_{press} = 0.74$, is only somewhat larger than for the model standard error, $s = 0.60$, suggesting that this equation may be useful for prediction.

CONCLUSIONS

For the HIV-1 protease inhibitor binding data set, an excellent QSAR model is developed with two E-state structure descriptors along with one molecular connectivity χ index and one κ shape index. Furthermore, external validation of the QSAR model is obtained by prediction of the test set, producing an MAE = 0.82, better than that obtained by the 3D-based methods reported in the literature.[15] The structure information encoded in the four descriptors is described in detail, indicating structurally significant features that may be useful for new compound design. The excellent results were obtained independent of information from 3D geometry based considerations.

Scientific value for this modeling is indicated by two aspects of this investigation. First, detailed structure information is available from the four structure descriptors used in the model. The structure information derived from the modeling also includes a hypothesis involving hydrogen bonding for OH groups attached to the common core skeleton. The analysis given permits specific directions to be given for synthesis of compounds with improved binding. Second, the modeling process does not require time-consuming and expensive methods based on obtaining and using detailed 3D structure information. This modeling process can be applied in a very straightforward manner, leading to a useful model, which illuminates structure aspects significantly related to inhibitor binding.

The reason for success of this topologically based method is the nature of the source structure information used in the topological representation of molecular structure. Topological descriptors such as the E-state and χ indices draw their information from fundamental aspects of molecular structure. Topological structure descriptors are a representation of molecular structure that arises from the chemical identity of each atom, including valence state, and the nature of the set of connections in the molecular skeleton, the chemical bonding pattern.[3,17,21] Through appropriate mathematical processes described in the literature, this encoded information is transformed into such significant structure features as valence state electronegativity, whose atom-by-atom differences are strongly related to electron distribution.[3,17,21] Further, significant relations among structure features are obtained from analysis of the network of chemical bonds,[22,23] leading to chi indices of skeletal branching and κ indices of molecular shape. The combination of valence state electronegativity and skeletal characterization leads to electron accessibility for the E-state formalism and potential for noncovalent intermolecular interaction, intermolecular accessibility, for the molecular connectivity formalism.[33]

ACKNOWLEDGMENT

L.H.H. wishes to acknowledge the DuPont Chemical Corporation for a grant which provided partial support for this research investigation.

REFERENCES AND NOTES

(1) Greer, J.; Erickson, J. W.; Baldwin, J. J.; Varney, M. D. Application of the Three-Dimensional Structures of Protein Target Molecules in Structure-Based Design. *J. Med. Chem.* **1994**, *37*, 1035−1054.

(2) Bugg, C. E.; Carson, W. M.; Montgomery, J. A. Drugs by Design. *Sci. Am.* **1993**, Dec, 92−98.

(3) Kier, L. B.; Hall, L. H. *Molecular Structure Description: The Electrotopological State*; Academic Press: San Diego, CA, 1999.

(4) Kier, L. B.; Hall, L. H. Inhibition of Salicylamide Binding: An Electrotopological State Analysis. *Med. Chem. Res.* **1992**, *2*, 497−502.

(5) Gough, J. D.; Hall, L. H. QSAR Models of the Antileukemic Potency of Carboquinones: Electrotopological State and Chi Indices. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 356−361.

(6) (a) Huuskonen, J. Estimation of Aqueous Solubility for a Diverse Set of Organic Compounds Based on Molecular Topology. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 773−777. (b) Huuskonen, J. QSAR Modeling with the Electrotopological State: TIBO Derivatives. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 425−429.

(7) Pantakar, S. J.; Jurs, P. C. Prediction of IC_{50} Values for ACAT Inhibitors from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 706−723.

(8) Gozolbes, R.; Galvez, J.; Garcia-Domenech, R.; Derouin, F. Molecular Search of New Active Drugs Against *Toxoplasma Gondii*. *SAR QSAR Environ. Res.* **1999**, *10*, 47−60.

(9) Tantillo, C.; Ding, J.; Jacobo-Molina, A.; Nanni, R. G.; Boyer, P. L.; Hughes, S. H.; Pauwels, R.; Andires, K.; Jansen, P. A.; Arnold, E. Locations of Anti_AIDS drug binding sites and resistance mutations in the three-dimensional structure of HIV-1 reverse transcriptase. Implications for mechanisms of drug inhibition and resistance. *J. Mol. Biol.* **1994**, *243*, 369−387.

(10) Karn, J.; Graeble, M. New Insights into the Mechanism of Hiv-1 Tran-Activation. *Trends Genet.* **1992**, *8*, 365−368.

(11) Antoni, B.; Stein, S. B.; Rabson, A. B. Regulation of human immunodeficiency virus infection: Implications for pathogenesis. *Adv. Virus Res.* **1994**, *43*, 53−145.

(12) Huff, J. R. HIV Protease: A Novel Chemotherapeutic Target for AIDS. *J. Med. Chem.* **1991**, *34*, 2305−2314.

(13) Kohl, N. E.; Emini, E. A.; Schleif, W. A.; Davis, L. K. J.; Heimbach, J. C.; Dixon, R. A.; Scolnick, E. M.; Sigal, I. Active human immunodeficiency virus is required for viral infection. *Proc. Nat. Acad. Sci. U.S.A.* **1998**, *85*, 4686.

(14) Farmergie, W. G.; Loeb, D. D.; Casavant, N. C.; Hutchinson, C. A., III; Edgel, M. H.; Swanstrom, R. Expression and processing of the AIDS virus reverse transcriptase in *Escherichia coli*. *Science* **1987**, *236*, 305−308.

(15) Holloway, M. K.; Wai, J. M.; Halgren, T. A.; Fitzgerald, P. M. D.; Vacca, J. P.; Dorsey, B. D.; Levin, R. B.; Thompsom, W. J.; Chen, J.; deSolms, S. J.; Gaffin, N.; Ghosh, A. K.; Giuliani, E. A.; Graham, S. L.; Guare, J. P.; Hungate, R. W.; Lyle, T. A.; Sanders, w. M.; Tucker, T. J.; Wiggins, M.; Wiscount, C. M.; Woltersdorf, O. W.; Young, S. D.; Darke, P. L.; Zugay, J. A. Priori Prediction of Activity for HIV-1 Protease Inhibitors Employing Energy Minimization in the Active Site. *J. Med. Chem.* **1995**, *38*, 305−317.

(16) Hall, L. H.; Mohney, B. K.; Kier, L. B. Comparison of electrotopological state indexes with molecular orbital parameters: Inhibition of MAO by hydrazides. *Quant. Struct.-Act. Relat.* **1993**, *12*, 44−48.

(17) Hall, L. H.; Kier, L. B. Electrotopological State Indices for Atom Types: A Novel Combination Of Electronic, Topological and Valence State Information. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1039−1045.

(18) Gough, J. D.; Hall, L. H. QSAR Models of the Antileukemic Potency of Carboquinones: Electrotopological State and χ Indices. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 356−361.

(19) Gough, J. D.; Hall, L. H. Modeling the toxicity of amide herbicides using the electrotopological state. *Environ. Toxicol. Chem.* **1999**, *18*, 1069−1075.

(20) Kier, L. B.; Hall, L. H. Database Organization and Similarity Searching with E-State Indices. In *Symposium on Computer Methods for Structure Representation*; Kluwer Academic Publishing Co.: New York, Dordrecht, The Netherlands, 2001; pp 33−49.

(21) Hall, L. H.; Kier, L. B. The Electrotopological State: Structure Modeling for QSAR and Database Analysis. In *Topological Indices and Related Descriptors in QSAR and QSPR*; Devillers, J., Balaban, A. T., Eds.; Gordon and Breach: Reading, U.K., 1999; pp 491−562.

(22) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure-Activity Analysis*; Research Studies Press, John Wiley and Sons: Chichester, U.K., 1986.

(23) Hall, L. H.; Kier, L. B. The Molecular Connectivity Chi Indexes and Kappa Shape Indexes in Structure-Property Relations. In *Reviews of Computational Chemistry*; Boyd, D., Lipkowitz, K., Eds.; VCH Publishers: Weinheim, Germany, 1991; Chapter 9, pp 367−422.

(24) Hall, L. H.; Kier, L. B. Molecular Connectivity Chi Indices for Database Analysis and Structure-Property Modeling. In *Topological Indices and Related Descriptors in QSAR and QSPR*; Devillers, J., Balaban, A. T., Eds.; Gordon and Breach: Reading, U.K., 1999; pp 307−360.

(25) Kier, L. B.; Hall, L. H. The Kappa Indices for Modeling Molecular Shape and Flexibility. In *Topological Indices and Related Descriptors in QSAR and QSPR*; Devillers, J., Balaban, A. T., Eds.; Gordon and Breach: Reading, U.K., 1999; pp 455−490.

(26) Maw, H. H.; Hall, L. H. E-State Modeling of Dopamine Transporter Binding. Validation of Model for a Small Data Set. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1270−1275.

(27) Maw, H. H.; Hall, L. H. E-State Modeling of Corticosteroid Binding Affinity. Validation of Model for a Small Data Set. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1248−1254.

(28) Hall, L. H.; Kier, L. B. The E-State as the Basis for Molecular Structure Space Definition and Structure Similarity. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 784−791.

(29) ChemDraw, ver 4.5; CambridgeSoft: Cambridge, MA 02139.

(30) Molconn-Z, ver 3.50, available from Hall Associates Consulting, 2 Davis Street, Quincy, MA, 02170; also from EduSoft, LC, PO Box 1811, Ashland, VA 23005, and SciVision, Inc, 200 Wheeler Road, Burlington, MA 01803.

(31) SAS, ver 8.0, SAS Institute, Cary, NC 27513.

(32) Thompson, W. J.; Fitzgerals, P. M. D.; Holloway, M. K.; Emini, E.; Darke, P. L.; McKeever, B. M.; Schleif, W. A.; Quintero, J. C.; Zugay, J. A.; Tucker, T. J.; Schwering, J. E.; Homnick, C. F.; Nunberg, J.; Springer, J. P.; Huff, J. R. Synthesis and Antiviral Activity of a Series of HIV-1 Protease Inhibitors with Functionality Tethered to the P1 or P1′ Phenyl Substituents: X-ray Crystal Structure Assisted Design. *J. Med. Chem.* **1992**, *35*, 1685−1701.

(33) Kier, L. B.; Hall, L. H. Intermolecular Accessibility: The Meaning of Molecular Connectivity. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 792−795.

CI010091Z