Additivity in the Analysis and Design of HIV Protease Inhibitors

Robert N. Jorissen,[‡,∇] G. S. Kiran Kumar Reddy,[†,§,¶] Akbar Ali,[†,§] Michael D. Altman,[||,○] Sripriya Chellappan,[‡,◆] Saima G. Anjum,[§] Bruce Tidor,[⊥] Celia A. Schiffer,[#] Tariq M. Rana,[§,+] and Michael K. Gilson*[,‡]

Center for Advanced Research in Biotechnology, UMBI, 9600 Gudelsky Drive, Rockville, Maryland 20850, Chemical Biology Program, Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School, Worcester, Massachusetts 01605, Department of Chemistry, Massachusetts Institute of Technology, Cambridge, Massachusetts, Department of Biological Engineering and Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, Massachusetts, Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School, Worcester, Massachusetts 01605

Received July 28, 2008

We explore the applicability of an additive treatment of substituent effects to the analysis and design of HIV protease inhibitors. Affinity data for a set of inhibitors with a common chemical framework were analyzed to provide estimates of the free energy contribution of each chemical substituent. These estimates were then used to design new inhibitors whose high affinities were confirmed by synthesis and experimental testing. Derivations of additive models by least-squares and ridge-regression methods were found to yield statistically similar results. The additivity approach was also compared with standard molecular descriptor-based QSAR; the latter was not found to provide superior predictions. Crystallographic studies of HIV protease−inhibitor complexes help explain the perhaps surprisingly high degree of substituent additivity in this system, and allow some of the additivity coefficients to be rationalized on a structural basis.

1. Introduction

The human immunodeficiency virus (HIV), the cause of AIDS, currently infects more 30 million people around the world,[1] and approximately 2 million people died of AIDS in 2007 alone. Current treatments include inhibiting the reverse transcriptase and protease of HIV with small molecule drugs, but their effectiveness can be diminished by the occurrence of resistance mutations in the virus.[2] New inhibitors of the HIV reverse transcriptase and protease are thus needed that will inhibit not only wild-type but also mutated forms of the virus's proteins.[3,4]

We have therefore sought to develop compounds that can inhibit both wild-type and mutated forms of the HIV protease.[5−8] The design approach is based in part upon the hypothesis that inhibitors that bind within a consensus envelope of bound substrate peptides are more likely to retain affinity for clinically relevant mutants.[9−13] Many of the HIV protease inhibitors synthesized and tested in the course of this effort possess a common chemical scaffold with three variable substituent positions (Figure 1). These compounds form an incomplete combinatorial library, and the question arose as to whether any of the unsynthesized compounds within the full library should be expected to bind HIV protease with high affinity and might thus be worth synthesizing and testing.

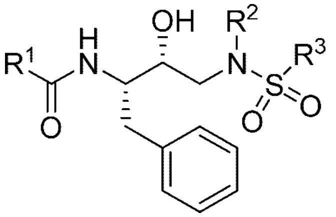


Figure 1. Framework of HIV protease inhibitors studied in this work, showing the three points of substitution.

In approaching this problem, it is natural to ask which are the best substituents and to make sure that all combinations of the best substituents have been tested. Such an approach implicitly relies upon an assumption of independence, i.e., that a substituent which appears in a high affinity compound will also tend to impart high affinity when combined with other substituents to form a different compound. The present study makes this assumption explicit through an implementation of the Free and Wilson additivity model[14] and evaluates the accuracy of the additivity model not only retrospectively but also prospectively by using it to select additional compounds from within the virtual combinatorial library for synthesis and testing. The results bear on the reliability of the additivity approximation in the present system. In addition, methodological variations are assessed, and the additivity approach is furthermore compared with a traditional quantitative structure−activity relationship (QSAR) evaluation of the same data sets. The results of the additivity analysis are also considered in the light of existing and new crystallographic complexes of bound inhibitors from the present series.

* To whom correspondence should be addressed. Phone: 240-314-6217. Fax: 240-314-6255. E-mail: gilson@umbi.umd.edu.

[†] These authors contributed equally to this study.

[‡] Center for Advanced Research in Biotechnology, UMBI.

[§] Chemical Biology Program, Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School.

[||] Department of Chemistry, Massachusetts Institute of Technology.

[⊥] Department of Biological Engineering and Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology.

[#] Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School.

[∇] Current Address: Ludwig Institute for Cancer Research, P.O Royal Melbourne Hospital, Parkville, Victoria 3050, Australia.

[○] Current address: Merck Research Laboratories, Boston, 33 Avenue Louis Pasteur, Boston, Massachusetts.

[◆] Current address: 6304 Summitview Avenue, Yakima, Washington 98908.

[¶] Current address: Prime Organics, Inc. 25-R Olympia Avenue, Woburn, Massachusetts 01801.

[+] Current address: Sanford Children's Health Research Center, Burnham Institute for Medical Research, 10901 North Torrey Pines Road, La Jolla, California 92037.

10.1021/jm8009525 CCC: $40.75 © 2009 American Chemical Society
Published on Web 01/07/2009

2. Two Methods

2.1. Data Modeling. 2.1.1. Additive Model for Affinity. The additive model assumes that the free energy contributions of the substituents of a given compound are independent and additive.[14] Two formulations are considered. In the first, one compound is chosen to be the reference compound and its substituents are the reference substituents for each of the three positions of substitution. The additivity model is then given by:

$$\mathrm{p}K_{\mathrm{i}}(a,\ b,\ c) \equiv -\log_{10} K_{\mathrm{i}}(a,\ b,\ c) \approx \mathrm{p}K_{\mathrm{i}}(1,\ 1,\ 1) + S_{1a} + S_{2b} + S_{3c} \quad (1)$$

where p$K_{\mathrm{i}}(a, b, c)$ is the predicted pK_{i} of a compound in the series with substituents *a*, *b*, and *c* at the R1, R2, and R3 positions, respectively; S_{1a} approximates the change in pK_{i} upon replacing the reference substituent at R1 ($a = 1$) with substituent a, while leaving R2 and R3 unchanged, and S_{2b} and S_{3c} have analogous interpretations for the R2 and R3 positions, respectively. The values of S_{1a}, S_{2b}, and S_{3c} are obtained by least-squares fitting, as described below. In the second formulation, similar to that of Free and Wilson,[14] the observed pK_{i} values are mean-centered and scaled, i.e., linearly transformed to yield a mean of zero and a standard deviation of one. This procedure removes the need to specify a reference molecule or reference substituents.

Two methods of parametrizing the additive models of affinity were tested, ordinary least-squares regression and ridge regression (RR[a]). The next two subsections describe these procedures.

2.1.1.1. Fitting the Additive Model by Ordinary Least-Squares Regression. Ordinary least-squares regression was used to optimize the values of the substituent parameters based on the experimental pK_{i} data. This method fits a vector $\boldsymbol{\beta}$ to the linear equation:

$$\mathbf{y}_{\mathrm{obs}} \approx \mathbf{y}_{\mathrm{fit}} = \mathbf{X}\boldsymbol{\beta} \quad (2)$$

such that the residual sum of squares deviations (RSS) is a minimum.

$$\mathrm{RSS} = \sum_{i} (\mathbf{y}_{\mathrm{obs},i} - y_{\mathrm{fit},I})^2 = (\mathbf{y}_{\mathrm{obs}} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y}_{\mathrm{obs}} - \mathbf{X}\boldsymbol{\beta}) \quad (3)$$

The values in the column vector $\boldsymbol{\beta}$ correspond to all the substituent parameters S_{1a}, S_{2b}, and S_{3c} in eq 1 other than the reference substituents; element *i* in the vector $\mathbf{y}_{\mathrm{obs}}$ equals the measured value of p$K_{\mathrm{i}}(a, b, c)$ − p$K_{\mathrm{i}}(1, 1, 1)$ for compound *i* and element *i* in the vector $\mathbf{y}_{\mathrm{fit}}$ is the corresponding fitted, and the rows of the design matrix **X** contain ones and zeroes, depending on whether or not a given molecule contains a given substituent.

Least-squares fitting was performed via singular value decomposition[15] with code implemented by the authors in C and employing an available singular value decomposition routine.[16] Additional code was written to compute predicted values of pK_{i} from eq 1.

Four of the measured K_{i} values were reported experimentally as ≥10 μM, corresponding to a p$K_{\mathrm{i}} \leq 5$. For convenience, these inequalities were converted to ranges by setting the somewhat arbitrary limit that p$K_{\mathrm{i}} \geq 2$. The same least-squares fitting procedure then was carried out for all combinations of three possibilities for each of these measurements: use of the upper limit in the fit, use of the lower limit, and exclusion of the pK_{i} value from the fit. The combination that yielded the lowest sum-squared error was chosen. Empirically, the fitted values never came close to the arbitrary lower limit of 2, so the precise value of this quantity is not an important parameter of the models. The same method was used to establish pK_{i} ranges of 11−13 for tight-binding inhibitors, as detailed in Results.

Confidence limits at the 95% level for the fitted parameters and pK_{i} predictions were obtained by the bootstrap sampling procedure.[15,17] This procedure consists of random selection of molecules from the training set with replacement, such that the total number sampled is equal to the number of molecules in the training set. For each of 500 such bootstrap samples, a set of substituent parameters was obtained from fitting and pK_{i} predictions made where applicable. Confidence limits for each parameter and pK_{i} prediction were obtained by determining the range containing the middle 95% of predictions. Note that some bootstrap samples lack data to enable fitting for some parameters and therefore do not yield parameter values needed for some of the pK_{i} predictions, so there may be fewer than 500 parameters/predictions from which to calculate the relevant confidence interval.

The statistical significance of the fit of modeled ($\mathbf{y}_{\mathrm{fit}}$) to measured ($\mathbf{y}_{\mathrm{obs}}$) p$K_{\mathrm{i}}$ values was assessed by calculating the appropriate *F*-statistic and its corresponding *p*-value. Molecules containing unique instances of their R1 substituents and their associated R1 substituent parameters were omitted from this analysis because their parameters could be trivially adjusted to give a perfect fit for that molecule's pK_{i} value. The *F*-statistic was calculated as:

$$F = \frac{(\mathrm{TSS} - \mathrm{RSS})(N - M)}{(\mathrm{RSS})(M - 1)} \quad (4)$$

where TSS is the total sum of squares $= \sum_i (\mathrm{y}_{\mathrm{obs},i} - \overline{\mathbf{y}_{\mathrm{obs}}})^2$, RSS is the residual sum of squares (eq 3), *N* is the number of observations (number of molecules included in the calculation of the *F*-statistic), and *M* is the number of parameters (number of substituent parameters included). The corresponding *p*-value was obtained from the look-up table at http://graphpad.com/quickcalcs/pvalue1.cfm with the associated numerator and denominator degrees of freedom *M* and $N - M - 1$, respectively. The calculated *p*-values were less than 0.0001 in all cases, indicating high statistical significance.

The effect of changing the reference molecule was shown to be modest. The value of any fitted parameter rarely deviated by more than 0.25 pK_{i} units from the median value of that parameter obtained from systematically changing the reference molecule to all of the training molecules in a given set. Similarly, the values of the pK_{i} predictions for each test set molecule were always within 0.25 pK_{i} units of the median predictions for that molecule (results not shown).

2.1.1.2. Fitting the Additive Model by Ridge Regression. Ridge regression [18,19] was evaluated as an alternative to ordinary least-squares for optimization of the additive substituent parameters. Ridge regression fits observed data to eq 2 but operates on mean-centered and scaled data (see above) and supplements the RSS in the target function with an additional term that penalizes fits with large substituent parameters:

$$\mathrm{RSS}_{\mathrm{RR}} = (\mathbf{y}_{\mathrm{obs}} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y}_{\mathrm{obs}} - \mathbf{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta} \quad (5)$$

where λ is the ridge parameter. The value of $\mathrm{RSS}_{\mathrm{RR}}$ is minimized by the regression coefficients $\boldsymbol{\beta}$ that solve the following linear equation:

[a] Nonstandard abbreviations: TSS, total sum of squares; RSS, residual sum of squares; GA, genetic algorithm; PLS, partial least-squares; GCV, generalized cross-validation; CC,cyclic carbamate; CI, confidence interval; add, additivity; RR, ridge regression; GARR, genetic algorithm/ridge regression; GAPLS, genetic algorithm/partial least-squares.

$$(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda I)\boldsymbol{\beta} = \mathbf{X}^{\mathrm{T}}\mathbf{y} \qquad (6)$$

The ridge parameter λ is a positive number that drives the coefficients in the β vector toward zero and drives all values of $\mathbf{y}_{\mathrm{fit}}$ toward the average value of the observed data $\mathbf{y}_{\mathrm{obs}}$. As λ goes to zero, ridge regression reduces to linear least-squares regression with mean-centered and scaled data. Shrinking the coefficients by using an appropriate value of λ can reduce the sensitivity of the fitted model to noise in the input data and thereby improve its predictivity. Even a small value of λ may be sufficient to avoid problems that can arise if the matrix product $\mathbf{X}^{\mathbf{T}}\mathbf{X}$ is ill-conditioned, i.e., if it is sensitive to numerical errors upon matrix inversion or similar operations. However, if λ is too large, then the shrinkage of the coefficients detracts from the predictivity of the model.

The code to perform ridge regression calculations was written in C by the authors. Fitting was carried out with λ set to 10^{-12}, 1, 3, 10, and 100 in order to assess the tradeoff between the increasing RSS and the decreasing size of the sum-of-squared coefficients in the β vector. The optimal value of λ will shrink the coefficients while producing only a small increase in RSS. Although methods exist to find λ through cross-validation, they were not used in this case because some of the molecules in the training set contain unique instances of some of the R1 substituents and predictions for these molecules made within the context of cross-validation (in which these molecules would be omitted from the training) would be meaningless. The λ value of 10^{-12} amounts to ordinary linear regression with mean-centered and scaled data. The present ridge regression code was not outfitted with the ability to handle data ranges, so compounds for which only inequality data are available were omitted. Also, so that the ridge regression results could be compared with the traditional QSAR results on an equal footing, compounds with substituents 5 or 6 at R2 were omitted. These substituents are not distinguished from each other by the global QSAR descriptors because they are stereoisomers of each other, and so we wished to exclude them from the QSAR analysis. This proved to remove most of the compounds containing R1 substituents 17, 18, and 19, and so the remaining four of these compounds were also removed. The resulting set of compounds used for the ridge regression and regular QSAR tests comprises compounds **1**−**55** and **71**−**106** from Table 1.

2.1.2. Modeling Affinity by QSAR with Global Molecular Descriptors. The additivity approximation was compared with traditional QSAR methods in which the molecules are represented by global molecular descriptors, which quantify aspects of the chemical structure as a whole. Parameters were fit by partial least-squares (PLS)[19,20] and ridge regression.[18,19]

Molecular descriptors were calculated with version 2.1 of the DRAGON program.[21] Here, 511 descriptors were calculated from descriptor categories 1−6 (constitutional descriptors, topological descriptors, molecular walk counts, BCUT descriptors, Galvez topological charge indices, and 2D autocorrelations). Principal components from each descriptor class were calculated with DRAGON, which generated 39 principal component-derived descriptors for the set of molecules. One set of calculations used all 39 transformed descriptors for both PLS and ridge regression. A second set of calculations used a genetic algorithm (GA) to select a subset of the 39 descriptors. The genetic algorithm used is very similar to that used by Hoffman et al.[22] for selecting molecular descriptors for PLS regression. For both GA-PLS and GA-ridge regression methods, five GA runs were performed, each with 100 generations of 100 chromosomes apiece. The descriptors chosen were those encoded by the most fit chromosome.

For the GA-PLS method, the fitness was a function of the cross-validated Q^2 value, the number of compounds, n, and the number of PLS factors, c, with an additional term to penalize models using more than 6 descriptors:

$$\mathrm{Fitness} = 1 - (n-1)(1-Q^2)/(n-c) - m_6^2 \qquad (7)$$

Here m_6 equals the larger of zero and the number of descriptors selected minus six. It was added to the fitness function because we have observed that GA-PLS calculations without such a term can find a highly fit descriptor set even when the descriptors are merely random numbers arbitrarily assigned to each compound. Penalizing models with a larger number of selected descriptors ameliorates this problem (unpublished results).

For ridge regression without GA, the value of λ that corresponded to the highest value of the leave-one-out statistic, Q^2, was found using the golden section search.[15] For ridge regression with GA, the value of λ was chosen to optimize a rapidly computable approximation to Q^2, the generalized cross-validation statistic (GCV). The GCV approximates leave-one-out cross validation without actually performing repeated calculations leaving out each molecule in turn.[23] To rapidly find the GCV-optimized value of λ, the following formula was iterated to convergence,[24] starting from an initial guess of 0.1:

$$\lambda_{\mathrm{GCV}} = \frac{\mathbf{y}_{\mathrm{obs}}^{\mathrm{T}}\mathbf{P}^2\mathbf{y}_{\mathrm{obs}}\mathrm{trace}(\mathbf{A}^{-1} - \lambda_{\mathrm{GCV}}\mathbf{A}^{-2})}{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{A}^{-1}\boldsymbol{\beta}\mathrm{trace}(\mathbf{P})} \qquad (8)$$

where n is the number of data points (pK_{i} values), m is the number of variables (descriptors), $\mathbf{A} = \mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I}_m$, the projection matrix $\mathbf{P} = \mathbf{I}_n - \mathbf{X}\mathbf{A}^{-1}\mathbf{X}^{\mathrm{T}}$, and $\mathbf{I}_x$ is an identity matrix with x rows and columns. The term $\mathbf{y}^{\mathrm{T}}_{\mathrm{obs}}\ \mathbf{P}^2\ \mathbf{y}_{\mathrm{obs}}$ is the residual sum-of-squares.[24] In the rare instance when this iterative procedure failed to converge, the value of λ was left at 0.1.

The resultant GA fitness function for ridge regression was the leave-one-out cross-validation statistic Q^2 minus the quantity m_6^2 described earlier. Note that the GCV method was not used to calculate the leave-one-out cross-validation statistic Q^2 itself but only used to find reasonable values for λ during the cross-validation.

2.2. Compounds Studied as HIV-1 Protease Inhibitors. Table 1 lists all 106 compounds studied here, and Table 2 provides the chemical structures of their substituents. The R1 substituents are separated into those that contain a cyclic carbamate group and those that do not, and the inhibitors are similarly divided into those that contain a cyclic carbamate group at R1 (CC compounds) and those that do not (non-CC compounds). Some of the CC compounds achieve high affinities, but the CC substituents tend to be larger than the non-CC substituents and thus tend to protrude from the substrate envelope more than the non-CC substituents. Therefore, the substrate envelope hypothesis would suggest that inhibitors containing non-CC R1 inhibitors should better inhibit mutant forms of the HIV protease.[9−13] The design methodologies for most of the training set compounds discussed in this work have been reported elsewhere,[5−7] as have the chemical synthesis and inhibition assays.[5−7,25−27] The synthesis of new compounds is described in the next section.

2.3. Experimental Methods. 2.3.1. Chemistry. The general synthetic route applied for the preparation of the inhibitors is illustrated in Scheme 1. The Boc-protected intermediates (R)-(hydroxyethylamino)sulfonamides **110**−**112** were prepared according to the procedures described earlier.[5] Briefly, ring opening of commercially available chiral epoxide, (1S,2S)-(1-

Table 1. Substituent Indices and Observed and Fitted pK_i Values of the Molecules Used in This Study[a]

				observed		model 1		model 2		model 3	
molecule	R1	R2	R3	K_i(obs)	pK_i	fit[b]	dev	fit	dev	fit	dev
1	1	1	1	0.10	10.00	10.00	0.00	10.00	0.00	10.00	0.00
2	1	1	2	3.80	8.42	8.38	0.04	8.42	0.00	8.56	−0.14
3	1	1	4	0.530	9.28	9.21	0.07	9.26	0.02	9.61	−0.33
4	1	2	2	238.7	6.62	6.60	0.02	6.75	−0.13	7.03	−0.41
5	1	2	3	170.2	6.77	6.42	0.35	6.64	0.13	6.85	−0.09
6	1	3	2	42.0	7.38	6.93	0.45	6.97	0.41	7.08	0.29
7	2	1	1	0.083	10.08	10.35	−0.27	10.26	−0.18	10.03	0.05
8	2	1	4	0.170	9.77	9.56	0.21	9.52	0.25	9.64	0.13
9	2	1	5	0.070	10.15	9.71	0.44	9.69	0.46	9.73	0.43
10	2	1	6	0.107	9.97	10.11	−0.14	10.09	−0.12	10.23	−0.26
11	2	2	2	188.8	6.72	6.94	−0.22	7.02	−0.30	7.06	−0.34
12	2	2	3	160.2	6.80	6.77	0.03	6.91	−0.11	6.89	−0.09
13	2	3	2	150.0	6.82	7.27	−0.45	7.23	−0.41	7.12	−0.29
14	2	4	1	0.257	9.59	9.20	0.39	9.19	0.40	9.21	0.38
15	3	1	1	0.004	11.40	10.74	0.66	10.66	0.74	10.37[c]	0.63
16	3	1	2	0.84	9.08	9.12	−0.04	9.08	0.00	8.93	0.14
17	3	1	4	0.184	9.74	9.95	−0.21	9.92	−0.18	9.98	−0.25
18	3	1	5	0.080	10.10	10.10	0.00	10.08	0.02	10.06	0.03
19	3	1	6	0.016	10.80	10.51	0.29	10.48	0.32	10.57	0.22
20	3	2	2	29.5	7.53	7.34	0.19	7.41	0.12	7.40	0.13
21	3	2	3	167.7	6.78	7.16	−0.38	7.30	−0.52	7.23	−0.45
22	3	4	1	0.80	9.10	9.59	−0.49	9.58	−0.48	9.55	−0.45
23	4	1	1	0.066	10.18	10.28	−0.10	10.23	−0.05	9.98	0.20
24	4	1	4	0.230	9.64	9.49	0.15	9.49	0.15	9.59	0.05
25	4	1	5	0.343	9.46	9.64	−0.18	9.66	−0.20	9.67	−0.21
26	4	1	6	0.085	10.07	10.05	0.02	10.06	0.01	10.18	−0.11
27	4	4	1	0.58	9.24	9.13	0.11	9.16	0.08	9.16	0.08
28	5	1	1	0.006	11.22	11.05	0.17	11.01	0.21	10.70[c]	0.30
29	5	1	4	0.042	10.38	10.26	0.12	10.27	0.11	10.31	0.06
30	5	1	5	0.072	10.14	10.41	−0.27	10.43	−0.29	10.40	−0.25
31	5	1	6	0.016	10.80	10.81	−0.01	10.83	−0.03	10.90	−0.11
32	6	1	1	0.0008	12.10	11.61	0.49	11.58	0.52	10.54[c]	0.46
33	6	1	4	0.032	10.49	10.82	−0.33	10.84	−0.35	10.15	0.35
34	6	1	6	0.006	11.22	11.38	−0.16	11.40	−0.18	10.74[c]	0.26
35	6	8	1	0.232	9.63					10.52	0.89
36	6	8	9	0.019	10.72					10.90	0.18
37	7	1	1	0.117	9.93					10.30	0.37
38	7	2	3	33.0	7.48	7.48	0.00	7.48	0.00	7.16	0.32
39	7	8	1	0.046	10.34					10.29	0.05
40	8	2	3	1064.4	5.97	5.97	0.00	5.97	0.00	5.97	0.00
41	9	1	1	0.39	9.41			10.01	−0.60	9.91	−0.50
42	9	2	3	52.6	7.28	7.09	0.19	6.66	0.62	6.76	0.52
43	9	2	7	13232	4.88	5.07	−0.19	4.90	−0.02	4.90	−0.02
44	10	1	1	0.17	9.77			9.68	0.09	9.57	0.20
45	10	2	3	609.6	6.21	6.41	−0.20	6.32	−0.11	6.43	−0.21
46	10	2	7	26318	4.58	4.39	0.19	4.56	0.02	4.56	0.02
47	11	2	3	2360.7	5.63	5.63	0.00	5.63	0.00	5.63	0.00
48	12	2	2	514.6	6.29	6.29	0.00	6.29	0.00	6.29	0.00
49	13	1	1	0.093	10.03			10.34	−0.31	10.17	−0.14
50	13	2	2	40.4	7.39	7.39	0.00	7.09	0.30	7.20	0.19
51	13	8	9	0.033	10.48					10.53	−0.05
52	14	2	2	50.2	7.30	7.30	0.00	7.30	0.00	6.88	0.42
53	14	8	1	0.38	9.42					9.84	−0.42
54	15	2	2	1148.0	5.94	5.94	0.00	5.94	0.00	5.94	0.00
55	16	2	2	582.4	6.23	6.23	0.00	6.23	0.00	6.23	0.00
56	17	4	2	23.9	7.62	6.79	0.83	6.79	0.83	6.79	0.83
57	17	4	8	58.0	7.24	6.89	0.35	6.89	0.35	6.89	0.34
58	17	5	2	1169.90	5.93	6.22	−0.29	6.22	−0.29	6.22	−0.29
59	17	6	2	764.2	6.12	6.58	−0.46	6.58	−0.46	6.58	−0.47
60	17	6	8	542.9	6.27	6.69	−0.42	6.69	−0.42	6.69	−0.42
61	18	4	2	14618	4.84	5.17	−0.33	5.17	−0.33	5.17	−0.33
62	18	4	8	>10000	5−2	5.27	−0.27	5.27	−0.27	5.27	−0.27
63	18	5	2	12182.3	4.91	4.60	0.31	4.60	0.31	4.60	0.31
64	18	6	2	4763.1	5.32	4.96	0.36	4.96	0.36	4.96	0.36
65	18	6	8	>10000	5−2	5.07	−0.07	5.07	−0.07	5.07	−0.07
66	19	4	2	2024.20	5.69	5.58	0.11	5.58	0.11	5.58	0.11
67	19	4	8	>10000	5−2	5.69	−0.69	5.69	−0.69	5.69	−0.69
68	19	5	2	>10000	5−2	5.02	−0.02	5.02	−0.02	5.02	−0.02
69	19	6	2	13586.5	4.87	5.38	−0.51	5.38	−0.51	5.38	−0.51
70	19	6	8	258.8	6.59	5.48	1.11	5.48	1.11	5.48	1.10
71	20	1	1	0.24	9.62			9.40	0.22	9.54	0.08
72	20	1	9	0.12	9.92			9.80	0.12	9.92	0.01
73	20	7	1	4.17	8.38			8.92	−0.54	8.93	−0.55

Table 1. Continued

				observed		model 1		model 2		model 3	
molecule	R1	R2	R3	K_i(obs)	pK_i	fit[b]	dev	fit	dev	fit	dev
74	20	7	9	1.62	8.79			9.32	−0.53	9.31	−0.51
75	20	8	1	0.062	10.21			9.64	0.57	9.53	0.68
76	20	8	9	0.063	10.20			10.04	0.16	9.90	0.30
77	21	1	1	0.14	9.85			9.74	0.11	9.88	−0.02
78	21	1	9	0.027	10.57			10.14	0.43	10.25	0.32
79	21	7	1	1.45	8.84			9.26	−0.42	9.27	−0.43
80	21	7	9	0.309	9.51			9.66	−0.15	9.64	−0.13
81	21	8	1	0.117	9.93			9.98	−0.05	9.86	0.07
82	21	8	9	0.036	10.44			10.38	0.06	10.24	0.20
83	22	1	1	0.084	10.08			10.23	−0.15	10.37	−0.29
84	22	1	9	0.099	10.00			10.63	−0.63	10.74	−0.74
85	22	7	1	0.038	10.42			9.74	0.68	9.76	0.66
86	22	7	9	0.014	10.85			10.14	0.71	10.13	0.72
87	22	8	1	0.033	10.48			10.46	0.02	10.35	0.13
88	22	8	9	0.057	10.24			10.86	−0.62	10.73	−0.48
89	23	1	1	1.88	8.73			9.11	−0.38	9.25	−0.52
90	23	1	9	0.29	9.54			9.51	0.03	9.62	−0.09
91	23	7	1	2.48	8.61			8.63	−0.02	8.64	−0.03
92	23	7	9	0.82	9.09			9.03	0.06	9.01	0.07
93	23	8	1	0.617	9.21			9.35	−0.14	9.23	−0.02
94	23	8	9	0.063	10.20			9.75	0.45	9.61	0.59
95	24	1	1	0.173	9.76			9.56	0.20	9.69	0.07
96	24	1	9	0.115	9.94			9.96	−0.02	10.07	−0.13
97	24	7	1	0.79	9.10			9.07	0.03	9.08	0.02
98	24	7	9	0.21	9.68			9.47	0.21	9.46	0.22
99	24	8	1	0.21	9.68			9.79	−0.11	9.68	0.00
100	24	8	9	0.132	9.88			10.19	−0.31	10.05	−0.17
101	25	1	1	0.37	9.43			9.41	0.02	9.55	−0.12
102	25	1	9	0.134	9.87			9.81	0.06	9.92	−0.05
103	25	7	1	0.86	9.07			8.93	0.14	8.94	0.13
104	25	7	9	0.70	9.15			9.33	−0.18	9.31	−0.16
105	25	8	1	0.34	9.47			9.65	−0.18	9.53	−0.07
106	25	8	9	0.067	10.17			10.05	0.12	9.91	0.27
rms deviation						0.35		0.35		0.35	
R^2						0.97		0.97		0.96	
F						55		55		49	
p-value						<0.0001		<0.0001		<0.0001	

[a] The chemical structures of the substituents are shown in Table 2. The solid line separates the compounds whose R1 substituents are the larger cyclic carbamate-containing (CC) substituents from the non-CC compounds, whose R1 substituents are smaller (see Table 2). The K_i values are in nM. Statistics associated with the three least-squares regressions are also shown. [b] fit: fitted pK_i value; dev: deviation of fitted pK_i from the corresponding observed value. [c] The observed pK_i value used in the regression was changed to a range of 11−13, for reasons explained in the text.

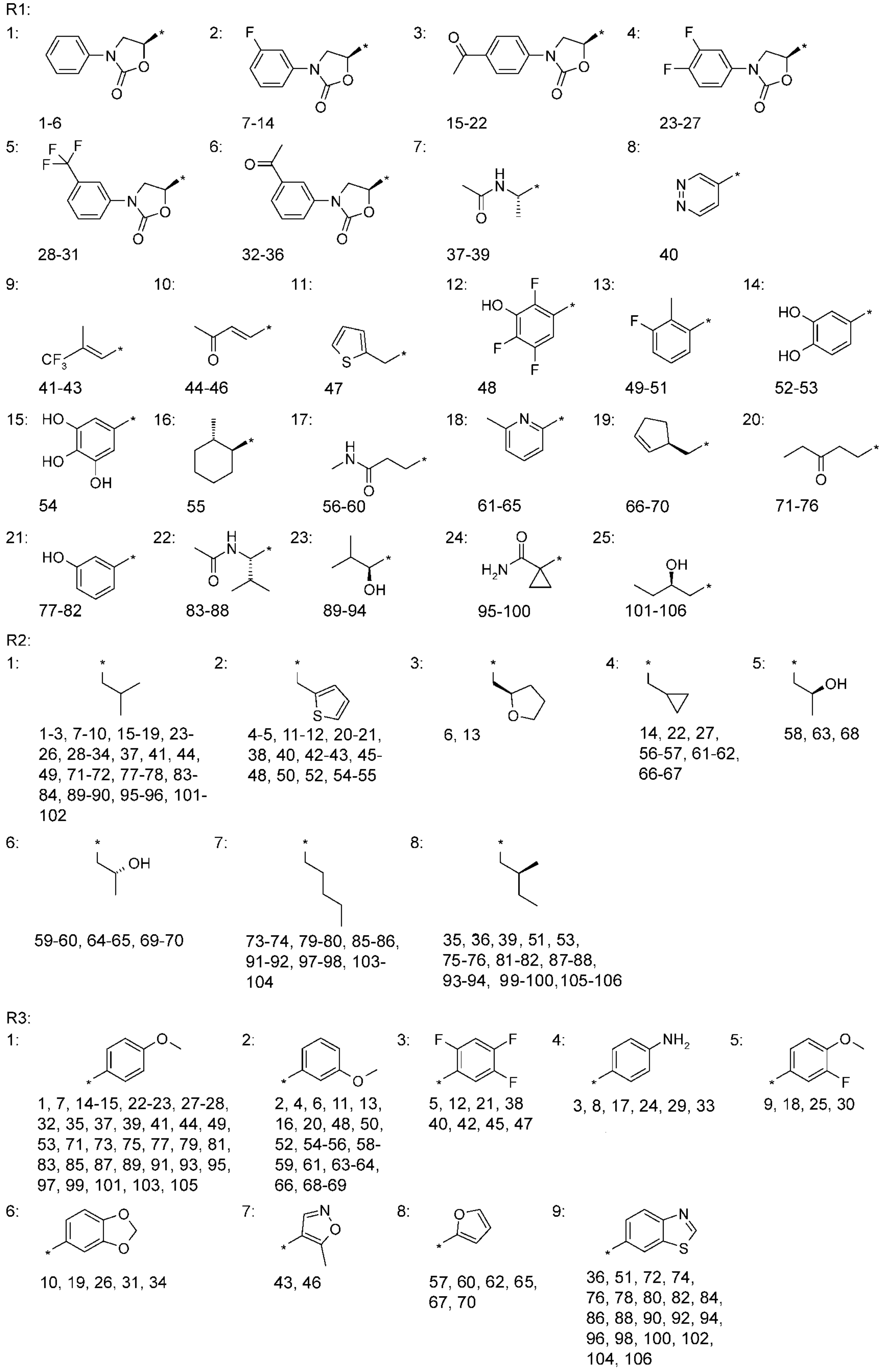
oxiranyl-2-phenylethyl)carbamic acid *tert*-butyl ester **107** with selected R^2 primary amines provided the amino alcohols **108** and **109**. Reactions of selected R^3 sulfonyl chlorides with **108** and/or **109** gave the sulfonamide intermediates, (*R*)-(hydroxyethylamino)sulfonamides **110**−**112**. After removing the Boc protection, the free amine fragments were coupled with selected R^1 carboxylic acids using two different coupling methods: The cyclic carbamate-based acid fragment was first converted to the corresponding acyl chloride and then reacted with amine to provide the target compounds **35** and **36** (method A);[5] the selected carboxylic acids were reacted with free amines using EDCI/HOBt in H_2O-CH_2Cl_2 (1:1) mixture to afford the designed inhibitors **37**, **39**, **41**, **44**, **49**, **51**, and **53** (method B).[25]

2.3.2. HIV-1 Protease Inhibition Assays. The HIV-1 protease inhibitory activities of all newly designed inhibitors were determined by a fluorescence resonance energy transfer (FRET) method.[5,26] Protease substrate, (Arg-Glu(EDANS)-Ser-Gln-Asn-Tyr-Pro-Ile-Val-Gln-Lys(DABCYL)-Arg) was purchased from Molecular Probes. The energy transfer donor (EDANS) and acceptor (DABCYL) dyes were labeled at two ends of the peptide, respectively, to perform FRET. Fluorescence measurements were carried out on a fluorescence spectrophotometer (Photon Technology International) at 30 °C. Excitation and emission wavelengths were set at 340 and 490 nm, respectively. Each reaction was recorded for about 10 min. Wild-type HIV-1 protease (Q7K) was desalted through PD-10 columns (Amersham Biosciences). Sodium acetate (20 mM, pH 5) was used as elution buffer. Apparent protease concentrations were around 50 nM estimated by UV spectrophotometry at 280 nm. All inhibitors were dissolved in dimethylsulfoxide (DMSO) and diluted to appropriate concentrations. Protease (2 μL) and inhibitor (2 μL) or DMSO were mixed and incubated for 20−30 min at room temperature before initializing substrate cleavage reaction. For all experiments, 150 μL of 1 μM substrate were used in substrate buffer [0.1 M sodium acetate, 1 M sodium chloride, 1 mM ethylenediaminetetraacetic acid (EDTA), 1 mM dithiothreitol (DTT), 2% DMSO and 1 mg/mL bovine serum albumin (BSA) with an adjusted pH 4.7]. Inhibitor binding dissociation constant (K_i) values were obtained by nonlinear regression fitting (GraFit 5, Erithacus software) to the plot of initial velocity as a function of inhibitor concentrations based on the Morrison equation.[27] The initial velocities were derived from the linear range of reaction curves.

3. Results

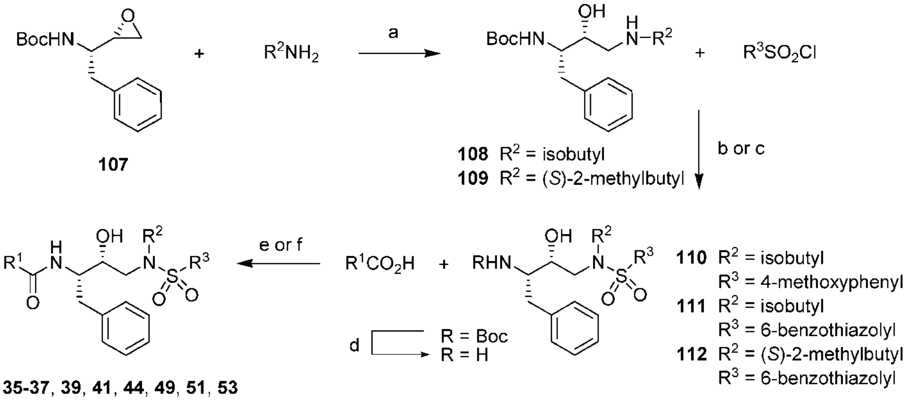
Three rounds of modeling and two rounds of synthesis were carried out. The first additivity model (model 1) was generated with the least-squares regression methodology using the pK_i values of 61 molecules that had been synthesized to date. On

Table 2. The R1, R2, and R3 Substituents of the Compounds Listed in Table 1[a]



[a] The attachment points to the scaffold are indicated by an asterisk. The indices of the compounds in which each substituent are found in are listed below each substituent's chemical structure.

Scheme 1. Reaction Scheme for the Synthesis of Designed Protease Inhibitors[a]



[a] Reagents and conditions: (a) EtOH, 80 °C, 3 h; (b) aq Na_2CO_3, CH_2Cl_2, 0 °C to rt, 4−8 h; (c) Et_3N, CH_2Cl_2, 0 °C to rt, 4−8 h; (d) TFA, CH_2Cl_2, rt, 1 h; (e) EDCI, HOBt, H_2O-CH_2Cl_2 (1:1), 0 °C, 24 h; (f) acid activation: $(COCl)_2$, rt, overnight; coupling: Et_3N, THF, 0 °C to rt, 6 h.

the basis of this model, several further compounds were proposed, synthesized, and tested. The second additivity model (model 2) was generated after the pK_i values of 39 more compounds, including some of the proposed compounds, became available. The second model was used to propose 7 more compounds, and these were designed, synthesized, and tested. In addition, an evaluation of errors in the models led to an adjustment in the treatment of the experimental data for the highest affinity compounds. Finally, the third additivity model (model 3) was constructed based upon the measured pK_i values of the 61 + 39 + 6 = 106 compounds that had been studied experimentally. This third model provides the current best estimates of the affinity contributions of the various substituents.

Additional calculations were carried out to compare the various regression methods with each other and to compare the additivity model with traditional descriptor-based QSAR.

3.1. Additivity Models. 3.1.1. First Additivity Model and Cycle of Inhibitor Design.

The first additivity model was constructed by using ordinary least-squares regression to fit 61 molecules' pK_i values to eq 1, with molecule **1** (K_i = 0.1 nM)[5] as the reference compound. Parameters for 19, 6, and 8 substituents at the R1, R2, and R3 positions, respectively, were fitted. A plot of the fitted vs observed pK_i values (Figure 2) shows that there is a good fit of the data to eq 1. Some of the data points have zero residual because their corresponding molecules contain the only instances of their R1 substituents (Table 1), and so the value of their corresponding R1 substituent parameters take on values that yield zero error.

The fitted parameters listed in Table 3 approximately quantify the change in pK_i from 10 (K_i = 0.1 nM) upon replacing a substituent in the reference compound by the listed substituent. These values range from 1.61 (R1 substituent 6) to −3.82 (R3 substituent 7). The bootstrap-derived 95% confidence limits' ranges vary from 0.78 to 3.22 with median 1.16 and indicate that comparison of the various substituent parameters must be approached with caution due to the large uncertainties in these values.

The parameters for all of the CC substituents (R1 substituents 1−6) are greater than or equal to zero. This observation reinforces the general observation of the potency of compounds that incorporate this moiety.[5] The parameters from this least-squares fit indicate that R1 substituent 6 is the most potent of these moieties. In contrast, only five of the 13 non-CC R1 substituents have fitted parameters close to or greater than zero.

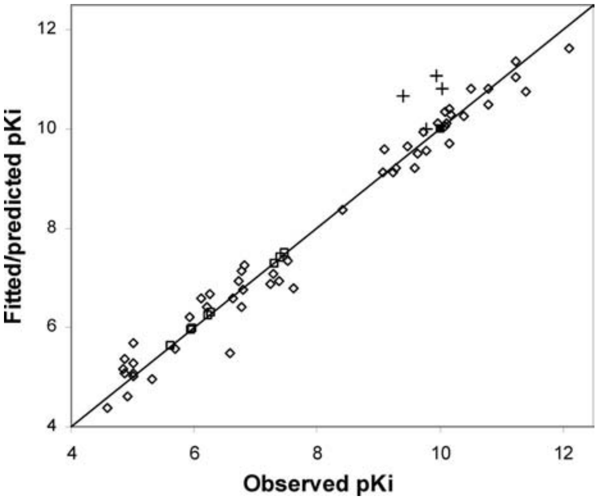


Figure 2. Comparison with experiment of calculated pK_i values for 61 (fitted) plus 4 (predicted) HIV protease inhibitors (model 1). The filled square corresponds to the reference compound, the unfilled squares indicate compounds with only one example of a given R1 substituent, and the unfilled diamonds represent the remaining 52 compounds in the training set. The crosses represent four test set compounds whose pK_i values were predicted from the parameters obtained from this training set (also see Table 4).

Of these, three (7, 13, and 14) are found in only one molecule each, and the other two (9 and 10) are found in only two molecules.

The fitted parameters in Table 3 show that there is one clearly preferred moiety for the R2 position, substituent 1. This group's substituent parameter is more than one log unit greater than those of the other R2 substituents, indicating that replacement of R2 substituent 1 with any of the other R2 groups would be expected to reduce binding affinity by more than 1 order of magnitude.

The most potent contributor to binding affinity at the R3 position, substituent 1, again is the reference substituent for this position. Three other R3 substituents decrease binding by less than an order of magnitude relative to the corresponding reference substituent, i.e., their fitted parameter substituents are greater than −1. Two of these three moieties contain an oxygen atom at the 4-position of the phenyl ring, as does the reference substituent. The other is the 4-anilino group, which is also found in the potent HIV protease inhibitors amprenavir[28] and darunavir.[29]

Table 3. Substituent Parameters Obtained from Least-Squares Fitting of pK_i Values for 61 Molecules (model 1)[a]

		parameter	n[b]	95% CI
R1:	1	0.00	5	0.00−0.00
	2	0.35	8	0.03−0.94
	3	0.74	8	0.24−1.21
	4	0.28	5	−0.03−0.77
	5	1.05	4	0.56−1.44
	6	1.61	3	1.08−2.10
	7	1.07	1	0.64−1.84
	8	−0.44	1	−1.12−0.30
	9	0.67	2	−1.76−1.46
	10	−0.01	2	−1.87−0.70
	11	−0.79	1	−1.59 to −0.08
	12	−0.31	1	−0.69−0.35
	13	0.80	1	0.33−1.45
	14	0.70	1	0.26−1.23
	15	−0.66	1	−1.05−0.00
	16	−0.36	1	−0.76−0.29
	17	−0.44	5	−1.89−0.98
	18	−2.06	5	−3.18 to −0.41
	19	−1.65	5	−2.92 to −0.22
R2:	1	0.00	22	0.00−0.00
	2	−1.78	18	−2.34 to −1.40
	3	−1.45	2	−2.36 to −0.70
	4	−1.15	9	−1.89 to −0.63
	5	−1.71	3	−3.11 to −0.63
	6	−1.35	6	−2.83 to −0.15
R3:	1	0.00	8	0.00−0.00
	2	−1.62	21	−2.06 to −1.25
	3	−1.80	8	−2.40 to −0.93
	4	−0.79	6	−1.25 to −0.46
	5	−0.64	4	−1.13 to −0.06
	6	−0.24	5	−0.72−0.20
	7	−3.82	2	−4.55 to −1.54
	8	−1.52	6	−2.63 to −0.08

[a] The cyclic carbamate R1 substituents are those above the solid line. [b] n: number of training set molecules containing the specified substituent; 95% CI: 95% confidence interval.

Compounds containing four of the five most potent non-CC R1 substituents (groups 7, 9, 10, and 13), plus the relatively potent reference R2 and R3 substituents, were proposed and their pK_i values predicted from additivity considerations (Table 4). These compounds were subsequently synthesized and tested and were found to have subnanomolar affinity (Table 4). This result contrasts with the corresponding training set molecules for which the most potent non-CC inhibitor is molecule **56** (CARB-AD37[8]) which has a K_i value of 23.9 nM (pK_i 7.62). This represents a successful application of the additivity method to design molecules with increased affinity. On the other hand, the measured binding constants of two of these compounds are more than an order of magnitude less than their corresponding predictions. Also, only two of the compounds' pK_i values lie within their respective predictions' 95% confidence limit.

3.1.2. Second Additivity Model and Cycle of Inhibitor Design. In the course of this research, 36 additional molecules from a separate inhibitor design project were synthesized and tested, the MIT-2 library from ref 7. Some of the substituent moieties in these molecules were not present in the 61-molecule training set used in the first additivity model, so another fit of parameters was performed. Additionally, the observed values of three of the molecules in Table 4 were included in this new training set. (Compound **37** had not yet been synthesized and was therefore not included at this stage.) The fit of this set of 100 molecules is shown in Figure 3, and the substituent parameters are presented in Table 5.

The parameter values for the R1 substituents 9, 10, and 13 each noticeably decrease from their respective values in the first additivity model, presumably reflecting the inclusion of pK_i values lower than the values predicted from the first set of parameters. The additivity analysis also identified three of the new substituents as potent contributors to inhibitor binding: R1 substituent 22 (parameter value 0.23); R2 substituent 8 (parameter value 0.23), and R3 substituent 9 (parameter value 0.40).

The substituents identified as contributing to high affinity were incorporated into a new set of additivity-designed compounds (Table 6). Some of these compounds contained the most potent non-CC substituents in the R1 position (substituents 7, 13, and 14). Others contained R1 substituent 6, the most potent of the CC substituents according to this second additivity model (Table 5). Comparison of the predicted pK_i values with those obtained from experiment are shown in Table 6 and Figure 3. The pK_i value of compound **37**, which had not been synthesized at this point, also was predicted using the updated substituent parameter values.

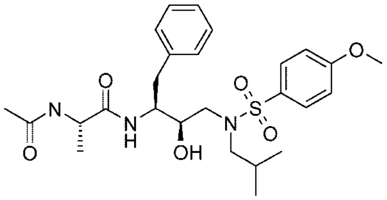
The observed pK_i values of the newly synthesized compounds are all less than predicted (Table 6), much as observed for the previous set of predictions (Table 4). The predictions for the non-CC compounds deviate from the observed values by −0.49 to −1.36. Two of these four compounds, **39** and **51**, are more potent than nearly all of the non-CC training set inhibitors (Table 1). Interestingly, these two compounds are the best predicted of the new compounds and their observed pK_i values are inside the 95% confidence intervals of the corresponding predictions. The observed pK_i values of the three CC compounds are less than predicted by more than one unit; i.e., the observed K_i values are greater than predicted by more than an order of magnitude. This unexpected result is considered in more detail later in the next section.

3.1.3. Analysis of First and Second Additivity Models. 3.1.3.1. Overestimation of Affinities. As noted above, the pK_i values predicted with the additivity model consistently exceeded the observed affinities, yielding inaccurate results in particular for compounds containing R1 substituent 6. Additional calculations were performed to address this issue.

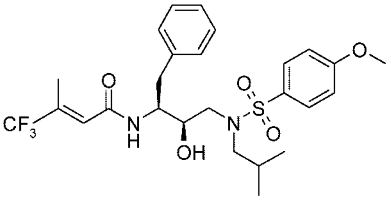
The training set compounds with R1 substituent 6 are all of high affinity (Table 1) and this contributes to the high value of the fitted parameter for this substituent (Table 5). This in turn appears to lead to excessively high pK_i predictions for the three designed compounds that contain this substituent (Table 6). It has been reported that the binding constants of inhibitors whose K_i values are below approximately 10 pM (pK_i > 11) cannot be reliably measured using the standard fluorometric assay.[30] To crudely account for this uncertainty in the measured pK_i values, a new additivity model was constructed in which training set pK_i values greater than 11 were replaced with the range 11−13. Note, however, that no new compounds were added to the data set. New predictions were made for the compounds in Table 6; Table 7 shows the results. The predictions made with the new model are closer to the experimentally determined pK_i values than those from the original model, and the improvements are pronounced for the three compounds that contain R1 substituent 6. Of these, the errors in two of the three predictions are within the range observed for the non-CC compounds. This analysis suggests that one reason for overprediction of affinities in the first and second additivity models is experimental uncertainty in the highest measured pK_i values. Allowing for this uncertainty, as done here, improves the predictions.

Even after allowing for these uncertainties, however, the predictions for two of the four non-CC compounds remained

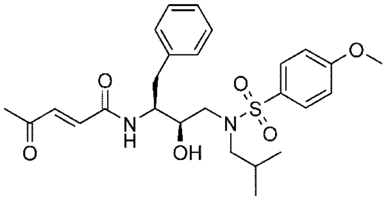
Table 4. Comparison of Predicted and Observed pK_i Values for Four Molecules Not Present in the 61 Molecule Training Set (model 1)[a]



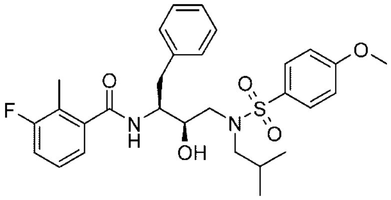
37



41



41



49

				observed		predicted		
molecule	R1	R2	R3	K_i (nM)	pK_i	pK_i	error	95% CI
37	7	1	1	0.117	9.93	11.07	−1.13	10.64−11.84
41	9	1	1	0.39	9.41	10.67	−1.26	8.24−11.46
44	10	1	1	0.17	9.77	9.99	−0.22	8.13−10.70
49	13	1	1	0.093	10.03	10.80	−0.77	10.33−11.45

[a] The predictions, including the bootstrap estimates of the 95% confidence intervals, were made prior to synthesis and testing of the compounds.

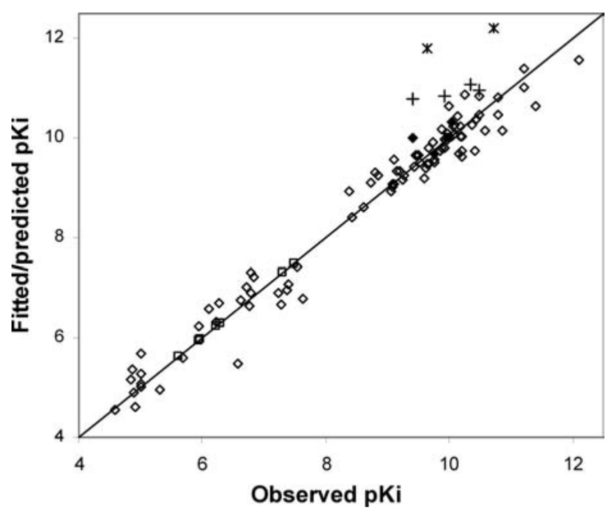


Figure 3. Comparison with experiment of calculated pK_i values for 100 (fitted) plus 6 (predicted) HIV protease inhibitors (model 2). The filled square corresponds to the reference compound, the unfilled squares indicate compounds with only one example of a given R1 substituent, the filled diamonds indicate the three compounds from Table 4 that were included in the 100 molecule training set, and the unfilled diamonds represent the remaining 89 compounds in the training set. The crosses and stars represent 6 test set compounds whose pK_i values were predicted with model 2 (Table 6): two contain the cyclic carbamate moiety (stars) and four do not (crosses).

outside the calculated 95% confidence intervals. To investigate these persistent overestimates, we constructed two artificial training sets and corresponding test sets of molecules. The first training and test sets (labeled A) consist of the compounds with R1 substituents 1−5, with half of the compounds containing R1 substituent 3 placed in the test set (Table 8). The second training and test set (labeled B) were generated by exchanging the first sets' training and test compounds that contain R1 substituent 3. The pK_i values of the molecules in the first test set are all overpredicted, similar to what was observed for the additivity-designed molecules (Tables 4, 6, and 7), but the pK_i values of three of the four molecules in the second test set were underpredicted. This change in the general direction of prediction was accompanied by a decrease in the fitted parameter value for the common R1 substituent 3. These results indicate that overprediction of affinity is not a problem intrinsic to the additivity method, but varies with the compounds in the training set (Table 9).

3.1.3.2. Ridge Regression versus Ordinary Least-Squares Regression. Least-squares fitting can overfit the training set data and so the subsequent predictions can vary significantly with small changes in the data such as changes due to measurement error or differences in training set composition. The technique of ridge regression attenuates this sensitivity by reducing the magnitudes of the fitted parameters and decreases the risk of overfitting the training data. Accordingly, this method was tested for the various data sets (compounds whose pK_i values were predicted using models 1 and 2, and the test sets from models A and B) with different values of the ridge parameter, λ, in order to determine its effect on the quality of the predictions. Note that the lowest value used, $\lambda = 10^{-12}$, reduces the method to ordinary least-squares regression with mean-centered and scaled data, and the predictions from this approach are similar to those from ordinary least-squares regression with a reference compound (Table 10).

Assessment of the sum of squared coefficients from the trained ridge regression models versus the training set sum-squared residuals suggests optimal values of the ridge parameter, λ, to be either 1 or 3, as these choices reduce the sum of squared coefficients while only slightly increasing the sum-squared residual for the training set. The test set rms prediction error for $\lambda = 1$ is approximately the same as our original, reference molecule-based additivity-based method, and it is generally slightly smaller for $\lambda = 3$ (Table 10). As λ is increased to 10 and 100, however, the errors rise. These results suggest that using ridge regression instead of ordinary least-squares regression may lead to a slight improvement in the pK_i predictions, given an appropriate choice of λ.

3.1.4. Third and Final Additivity Model. A third additivity model was constructed based on the full set of 106 compounds and with all pK_i values greater than 11 treated as ranges of

Table 5. Substituent Parameters Obtained from Least-Squares Fitting of pK_i Values for 100 Molecules (model 2)[a]

		parameter (1)	parameter (2)	*n* (2)	95% CI (2)
R1:	1	0.00	0.00	5	0.00−0.00
	2	0.35	0.26	8	−0.07−0.71
	3	0.74	0.66	8	0.16−1.18
	4	0.28	0.23	5	−0.15−0.57
	5	1.05	1.01	4	0.56−1.30
	6	1.61	1.58	3	0.92−2.10
	7	1.07	0.84	1	0.20−1.50
	8	−0.44	−0.67	1	−1.26 to −0.06
	9	0.67	*0.01*	3	−1.66−1.07
	10	−0.01	*−0.32*	3	−1.58−0.19
	11	−0.79	−1.02	1	−1.61 to −0.34
	12	−0.31	−0.46	1	−0.93−0.26
	13	0.80	*0.34*	2	0.03−1.04
	14	0.70	0.55	1	0.08−1.14
	15	−0.66	−0.81	1	−1.42 to −0.17
	16	−0.36	−0.52	1	−1.01−0.08
	17	−0.44	−0.56	5	−1.99−0.81
	18	−2.06	−2.18	5	−3.51 to −0.72
	19	−1.65	−1.76	5	−3.09 to −0.25
	20		*−0.60*	6	−1.07 to −0.23
	21		*−0.26*	6	−0.65−0.13
	22		*0.23*	6	−0.26−0.80
	23		*−0.89*	6	−1.16 to −0.44
	24		*−0.44*	6	−0.80 to −0.19
	25		*−0.59*	6	−0.87 to −0.25
R2:	1	0.00	0.00	37	0.00−0.00
	2	−1.78	−1.67	18	−2.28 to −1.08
	3	−1.45	−1.45	2	−2.21 to −0.76
	4	−1.15	−1.08	9	−1.84 to −0.50
	5	−1.71	−1.64	3	−3.09 to −0.28
	6	−1.35	−1.28	6	−2.61−0.09
	7		−0.49	12	−0.86 to −0.13
	8		0.23	12	−0.13−0.56
R3:	1	0.00	0.00	29	0.00−0.00
	2	−1.62	−1.58	21	−1.92 to −1.21
	3	−1.80	−1.69	8	−2.40 to −0.94
	4	−0.79	−0.74	6	−1.11 to −0.36
	5	−0.64	−0.58	4	−1.05 to −0.05
	6	−0.24	−0.18	5	−0.56−0.31
	7	−3.82	−3.45	2	−4.49 to −1.75
	8	−1.52	−1.48	6	−2.55 to −0.37
	9		0.40	18	0.13−0.64

[a] Cyclic carbamate R1 subsituents are above the solid line. Parameters (1) from model 1 (Table 3) are included for comparison; (2) indicates data for model 2.

11−13 (see above). This is arguably the best model because it uses all the available data and includes the improved treatment of experimental uncertainty. As summarized in Table 11, the parameters for R1 substituents 6, 7, and 14 decrease relative to the second model because of the inclusion of pK_i values below those predicted with the second additivity model. The parameter value for R1 substituent 13 falls less than the other three, presumably because the prediction for the compounds containing this substituent was more accurate than the predictions for the other six compounds (Table 6). The best of the non-CC R1 substituents are those numbered 7, 13, and 22, and the best of the CC R1 substituents are 3, 5, and 6.

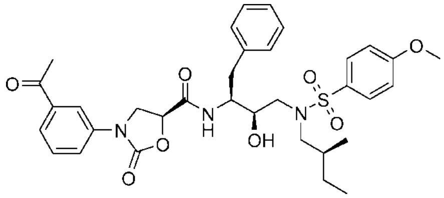
The parameter value of R2 substituent 8 fall from 0.23 to −0.01, relative to the second model, suggesting that this substituent's contribution to binding is approximately the same as that of R2 substituent 1. These two moieties are the most potent contributors to binding in the R2 position. The best of the R3 substituents are 1, 6, and 9. There are no major changes for R3 substituent parameters between the second and third additivity models.

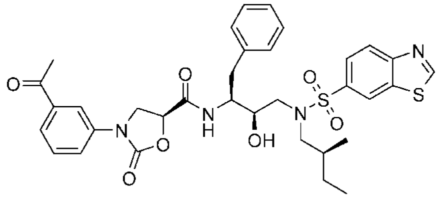
3.2. Additivity Approach versus Descriptor-Based QSAR. As an alternative to the substituent additivity paradigm, QSAR calculations were performed using the ridge regression and partial least-squares techniques, with the molecules represented via global molecular descriptors based on chemical structure. Two sets of partial least-squares and ridge regression calculations were performed: one pair of calculations using a genetic algorithm to find a small number of descriptor principal components (usually six) that maximize a cross-validation-based fitness function, and a second pair of calculations using all 39 descriptor principal components. While there are noticeable differences in these pK_i predictions, relative to those of the additivity method, there does not seem to be a uniform improvement in prediction accuracy (Table 12). Thus, these traditional QSAR models do not appear to provide any advantage over the additivity approach for this system.

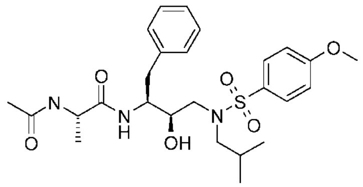
3.3. Crystallographic Correlations. 3.3.1. Structural Basis of Additivity. Crystallographic data on a number of HIV protease ligand complexes (molecules **15**, **28**, **31**, **32**, **34**, **38**, **42**, **50**, **52**, **56**, **57**, **75**, **82**, **86**, and **94**)[5,7,8,unpublished] enable the additivity approximation be rationalized on a structural basis. Thus, as shown in Figure 4a, substituents at the three positions of the scaffold occupy separate regions of the protease and do not contact each other. Furthermore, the common chemical scaffold of the various inhibitors (Figure 1) remains in essentially the same pose for all of the inhibitors. The lack of substituent-substituent contacts and of substituent-induced shifts of the scaffold provides a structural rationale for the reliability of the additivity model. It is likely that additivity would have been less reliable if the phenyl group of the scaffold had represented a fourth variable substituent because some of the R1 substituents, notably the cyclic carbamates, interact with the phenyl group (Figure 4A). Interestingly, the main-chain of Gly 48A in the protease adopts a somewhat different conformation for the CC versus the non-CC inhibitors, possibly to avoid a clash with the cyclic carbamate group (Figure 4B). If this shift in the conformation of the protease in response to the CC substituents had extended to other protease subsites, it might have generated large deviations from independent substituent additivity. In fact, however, the conformational shift is local to the R1 group, so it can be accounted for by the fitted affinity contributions of the R1 substituents. Thus, the substituent parameters can incorporate more effects than simply the direct interactions of ligand substituents with the protease.

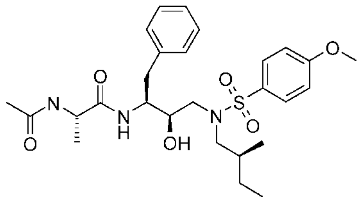
3.3.2. Structural Analysis of Substituents and Affinities. 3.3.2.1. R1 Substituent. No single chemical feature is common to the most potent R1 substituents, but a number of them contain a carbonyl group that forms a hydrogen bond to the main chain nitrogen of Asp 29A. These include the CC substituents (substituent parameters −0.02−0.70) whose other contacts are nonpolar in nature. The non-CC substituents 7 and 22 (parameter values 0.30 and 0.37, respectively) also make hydrogen bonds to the main chain nitrogen of Asp 29A and the main chain oxygen of Gly 48A similar to those made by the bound substrate.[9] The less efficacious substituent 20 (parameter value −0.46) lacks a group to make the second of these two hydrogen bonds although its atoms occupy similar positions to those of R1 substituents 7 and 22. The R1 substituent 17 makes hydrogen bonds to the same two protease atoms as do substituents 7 and 22 but with a different geometry, a consequence of its amide group being in the opposite orientation to those in R1 substituents 7 and 22. (Table 2). This suggests that the precise geometry of these hydrogen bonds may be important for binding.

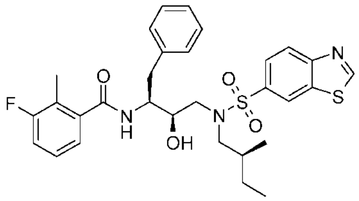
Table 6. Comparison of Predicted (model 2) and Observed pK_i Values for Seven Molecules Not Present in the Second Training Set

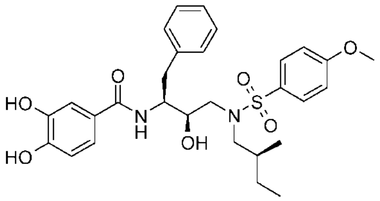

35


36


37


39


51


53

				observed		predicted		
molecule	R1	R2	R3	K_i (nM)	pK_i	pK_i	error	95% CI
35	6	8	1	0.232	9.63	11.81	−2.18	11.01−12.45
36	6	8	9	0.019	10.72	12.21	−1.49	11.39−12.87
37	7	1	1	0.117	9.93	10.84	−0.91	10.20−11.50
39	7	8	1	0.046	10.34	11.07	−0.73	10.30−11.83
51	13	8	9	0.033	10.48	10.97	−0.49	10.34−11.76
53	14	8	1	0.38	9.42	10.78	−1.36	10.21−11.43

Table 7. Comparison of Original Model 2 Predictions with New Model 2 Predictions Where Training Data Account for the Imprecision of High pK_i Data by Using a Range

					model 2			modified model 2		
molecule	R1	R2	R3	obs	pred[a]	95% CI	error	pred	95% CI	error
35	6	8	1	9.63	11.81	11.01−12.45	−2.18	11.29	10.91−13.48	−1.65
36	6	8	9	10.72	12.21	11.39−12.87	−1.49	11.69	11.18 −13.98	−0.97
37	7	1	1	9.93	10.84	10.20−11.50	−0.91	10.77	10.18 −11.31	−0.84
39	7	8	1	10.34	11.07	10.30−11.83	−0.73	11.01	10.30 −11.68	−0.67
51	13	8	9	10.48	10.97	10.34−11.76	−0.49	10.93	10.32−11.70	−0.45
53	14	8	1	9.42	10.78	10.21−11.43	−1.36	10.70	10.15−11.37	−1.28

[a] Pred: predicted pK_i value using the substituent parameters from the corresponding least-squares fit.

R1 substituents 13, 14, and 21 (substituent parameter values 0.71, −0.15, and −0.12, respectively), which are all substituted phenyls, also tend to generate potent inhibitors. Crystal structures of representative bound ligands containing these substituents, compounds **50** (MIT-1-KK80), **52** (MIT-1-KK81), and **82** (MIT-2-AD93),[7] show that they make mainly hydrophobic contacts with the protease, although the hydroxyl groups of the latter two also make hydrogen bonds with the protease. The substituent parameters of several less potent substituents can also be understood from an examination of these three crystal structures. The low affinity parameters of R1 substituents 12 and 15 (substituent parameters −0.74 and −1.09), which are chemically similar to 13, 14, and 21, can be rationalized in terms of expected steric clashes of substituents on their respective benzene rings with the protease. The low values of the substituent parameters of R1 substituents 8 and 18 (−0.88 and −2.57, respectively) can be rationalized by the lack of nearby hydrogen bond donor moieties from the protease in appropriate geometries (unpublished model building). The R1 substituent 9, which is nearly as potent as the larger R1 substituent 1, interacts with the protease through hydrophobic contacts.

3.3.2.2. R2 Substituent. The most potent of the R2 substituents studied in this work, substituents 1, 7, and 8, contain no rings or heteroatoms. Examination of the crystal structures of relevant HIV protease-inhibitor complexes (compounds **15**, **28**, **31**, **32**, **34**, **75**, **82**, **86**, and **94)** shows that the corresponding atoms of these substituents closely superimpose. In contrast, substituent 4, which is a cyclized version of R2 substituent 1, occupies a similar region of space as substituent 1 but has a slightly smaller volume. Replacement of substituent 1 with substituent 4 leads to a reduction in affinity of approximately 1 order of magnitude, suggesting that the precise shape complementarity of this group is important for inhibitor binding.

3.3.2.3. R3 Substituent. The crystal structures of complexes with compounds **15**, **28**, **31**, **32**, **34**, **75**, **82**, **86**, and **94** show that the three R3 substituents with the largest parameter values, substituents 1, 5, 6, and 9, each have an atom at the 4-position

Table 8. Least-Squares Fitting and Subsequent pK_i Prediction for Two Different Permutations of Training and Test Set Molecules

training set A							training set B						
molecule	R1	R2	R3	obs	fit	dev	molecule	R1	R2	R3	obs	fit	dev
1	1	1	1	10.00	10.00	0.00	**1**	1	1	1	10.00	10.00	0.00
2	1	1	2	8.42	8.42	0.00	**2**	1	1	2	8.42	8.60	−0.17
3	1	1	4	9.28	9.50	−0.22	**3**	1	1	4	9.28	9.47	−0.20
4	1	2	2	6.62	6.64	−0.02	**4**	1	2	2	6.62	6.58	0.05
5	1	2	3	6.77	6.46	0.30	**5**	1	2	3	6.77	6.68	0.08
6	1	3	2	7.38	7.01	0.37	**6**	1	3	2	7.38	7.00	0.37
7	2	1	1	10.08	10.19	−0.11	**7**	2	1	1	10.08	10.19	−0.11
8	2	1	4	9.77	9.68	0.09	**8**	2	1	4	9.77	9.67	0.10
9	2	1	5	10.15	9.75	0.40	**9**	2	1	5	10.15	9.80	0.35
10	2	1	6	9.97	10.14	−0.17	**10**	2	1	6	9.97	10.10	−0.13
11	2	2	2	6.72	6.83	−0.11	**11**	2	2	2	6.72	6.77	−0.05
12	2	2	3	6.80	6.65	0.14	**12**	2	2	3	6.80	6.88	−0.08
13	2	3	2	6.82	7.19	−0.37	**13**	2	3	2	6.82	7.20	−0.37
14	2	4	1	9.59	9.48	0.11	**14**	2	4	1	9.59	9.30	0.29
15	3	1	1	>11[a]	10.76	0.24[b]	**16**	3	1	2	9.08	8.90	0.17
19	3	1	6	10.80	10.71	0.08	**17**	3	1	4	9.74	9.78	−0.04
20	3	2	2	7.53	7.40	0.13	**18**	3	1	5	10.10	9.91	0.18
21	3	2	3	6.78	7.23	−0.45	**22**	3	4	1	9.10	9.41	−0.32
23	4	1	1	10.18	10.06	0.12	**23**	4	1	1	10.18	10.10	0.08
24	4	1	4	9.64	9.56	0.08	**24**	4	1	4	9.64	9.57	0.07
25	4	1	5	9.46	9.62	−0.16	**25**	4	1	5	9.46	9.71	−0.24
26	4	1	6	10.07	10.01	0.06	**26**	4	1	6	10.07	10.00	0.07
27	4	4	1	9.24	9.35	−0.11	**27**	4	4	1	9.24	9.21	0.03
28	5	1	1	>11[a]	10.83	0.17[b]	**28**	5	1	1	>11[a]	10.83	0.17[b]
29	5	1	4	10.38	10.32	0.05	**29**	5	1	4	10.38	10.31	0.07
30	5	1	5	10.14	10.39	−0.25	**30**	5	1	5	10.14	10.44	−0.30
31	5	1	6	10.80	10.78	0.02	**31**	5	1	6	10.80	10.74	0.06

rms deviation	0.21	rms deviation	0.19
R^2	0.98	R^2	0.98
F	68	*F*	61
P-value	<0.0001	*P*-value	<0.0001

test set A							test set B						
molecule	R1	R2	R3	obs	pred	error	molecule	R1	R2	R3	obs	pred	error
16	3	1	2	9.08	9.18	−0.11	**15**	3	1	1	>11[a]	10.31	0.69[b]
17	3	1	4	9.74	10.26	−0.52	**19**	3	1	6	10.80	10.21	0.59
18	3	1	5	10.10	10.33	−0.23	**20**	3	2	2	7.53	6.88	0.65
22	3	4	1	9.10	10.05	−0.96	**21**	3	2	3	6.78	6.99	−0.22

[a] pK_i range specified as 11−13 for the calculation. [b] Calculation performed assuming an observed pK_i of 11.

Table 9. Comparison of Fitted Parameters Obtained from the Two Training Sets of Table *8*

		model A			model B		
		parameter	*n*	95% CI	parameter	*n*	95% CI	difference
R1:	1	0.00	5	0.00−0.00	0.00	5	0.00−0.00	0.00
	2	0.19	8	−0.11−0.59	0.19	8	−0.15−0.59	−0.01
	3	0.76	4	0.10−3.00	0.31	4	−0.27−0.75	0.45
	4	0.06	5	−0.38−0.36	0.10	5	−0.30−0.44	−0.04
	5	0.83	4	0.35−3.00	0.83	4	0.35−3.00	−0.01
R2:	1	0.00	16	0.00−0.00	0.00	17	0.00−0.00	0.00
	2	−1.78	6	−2.31 to −1.36	−2.02	4	−2.59 to −1.54	0.24
	3	−1.41	2	−2.07 to −1.04	−1.59	2	−2.35 to −1.04	0.18
	4	−0.71	2	−1.13 to −0.33	−0.89	3	−1.43 to −0.42	0.18
R3:	1	0.00	6	0.00−0.00	0.00	6	0.00−0.00	0.00
	2	−1.58	6	−1.78 to −1.12	−1.40	6	−1.73 to −0.86	−0.17
	3	−1.76	3	−2.35 to −1.20	−1.30	2	−1.79 to −0.71	−0.46
	4	−0.50	4	−0.73 to −0.09	−0.53	5	−0.75 to −0.14	0.03
	5	−0.44	3	−0.90−0.17	−0.39	4	−0.80−0.19	−0.04
	5	−0.05	4	−0.43−0.33	−0.10	3	−0.45−0.22	0.05

of a six-membered ring, which can accept a hydrogen bond from Asp 30B. (This should also occur for substituent 5, which is not represented in any of the crystal structures). The next best R3 substituent is the 4-anilino group (substituent 4), which is found in the HIV protease inhibitors amprenavir[28] and darunavir.[29] Interestingly, the 4-anilino group is associated with superior pharmacokinetics relative to substituents 1 and 6,[29] and this advantage apparently outweighs a small sacrifice in affinity. This reminds us that binding affinity is not the sole criterion for selection of substituents in drug candidates.

4. Discussion

The approximation of substituent additivity is found to be informative and useful for a series of HIV protease inhibitors having a common chemical scaffold. Statistical analysis yields significant least-squares fits of measured and modeled pK_i values, and the fitted substituent parameters enabled the design of new inhibitors with subnanomolar binding affinities.

A major virtue of the additivity method is its obvious interpretability and its consequent utility for designing molecules that consist of new combinations of substituents already tested in a partial combinatorial set of inhibitors. The non-CC compounds that were designed from additivity considerations used R1 groups that were originally found in inhibitors with only moderate potency, yet these new compounds bind the protease target with better than 1 nM affinity, representing a successful application of the methodology. Application of additivity to design CC compounds with increased affinity was less successful, but it should be noted that two of the three

Table 10. Comparison of pK_i Predictions Made by Least-Squares Regression and Ridge Regression Approaches to Fitting the Additivity Model[a]

molecule	R1	R2	R3	obs[e]	add	LS	RR1	RR3	RR10	RR100
				Model 1[c]						
37	7	1	1	9.93	10.97	11.11	11.06	10.91	10.57	9.42
41	9	1	1	9.41	10.58	10.72	10.62	10.41	9.99	9.05
44	10	1	1	9.77	9.90	10.04	9.95	9.77	9.43	8.83
49	13	1	1	10.03	10.71	10.85	10.84	10.76	10.49	9.38
training set sum-squared residuals					1.83	1.73	1.81	2.08	3.45	31.51
training set sum-squared coefficient						0.45	0.40	0.36	0.29	0.10
training set rms error					0.20	0.19	0.20	0.21	0.27	0.83
test set rms error					0.85	0.98	0.92	0.79	0.52	0.65
				Model 2[c]						
35	6	8	1	9.63	11.29	11.24	11.20	11.10	10.85	10.07
36	6	8	9	10.72	11.69	11.64	11.58	11.43	11.09	10.13
37	7	1	1	9.93	10.77	10.76	10.75	10.64	10.36	9.56
39	7	8	1	10.34	11.01	11.00	10.95	10.79	10.41	9.46
51	13	8	9	10.48	10.93	10.93	10.84	10.70	10.39	9.61
53	14	8	1	9.42	10.70	10.69	10.61	10.47	10.15	9.35
training set sum-squared residuals					6.47	6.46	6.70	7.14	8.74	31.28
training set sum-squared coefficient						0.93	0.60	0.51	0.39	0.16
training set rms error					0.28	0.28	0.28	0.29	0.32	0.61
test set rms error					1.06	1.03	0.98	0.87	0.63	0.61
				Model A						
16	3	1	2	9.08	9.18	9.08	9.14	9.18	9.16	9.04
17	3	1	4	9.74	10.26	10.24	10.21	10.15	10.01	9.47
18	3	1	5	10.10	10.33	10.33	10.29	10.22	10.06	9.49
22	3	4	1	9.10	10.05	10.05	10.02	9.95	9.77	9.28
training set sum-squared residuals					1.11	1.04	1.06	1.18	2.14	21.05
training set sum-squared coefficient						0.38	0.36	0.34	0.27	0.06
training set rms error					0.20	0.20	0.20	0.21	0.28	0.88
test set rms error					0.56	0.55	0.53	0.48	0.37	0.34
				Model B						
15	3	1	1	>11[d]	10.31	10.34	10.31	10.26	10.11	9.57
19	3	1	6	10.80	10.21	10.22	10.20	10.16	10.06	9.60
20	3	2	2	7.53	6.88	6.87	6.92	7.02	7.28	8.33
21	3	2	3	6.78	6.99	6.98	7.01	7.08	7.27	8.28
training set sum-squared residuals					0.98	0.97	0.99	1.08	1.85	17.00
training set sum-squared coefficient						0.38	0.37	0.34	0.27	0.06
Training set rms error					0.19	0.19	0.19	0.20	0.26	0.79
Test set rms error					0.57[e]	0.56[e]	0.56[e]	0.57[e]	0.64[e]	1.26[e]

[a] The test sets are the two sets of molecules for which predictions from models 1 and 2 were made, and also test sets A and B. [b] obs: observed pK_i values; add: pK_i predictions from the original additivity method; LS: least-squares using ridge regression code and with $\lambda = 10^{-12}$; RR1: ridge regression with $\lambda = 1$; RR3: ridge regression with $\lambda = 3$; RR10: ridge regression with $\lambda = 10$; RR100: ridge regression with $\lambda = 100$. [c] Compounds **56**−**70** were omitted from training set, as discussed in the Methods section. [d] pK_i range specified as 11−13 for the calculation. [e] Calculation performed assuming an observed pK_i of 11 for compounds with pK_i > 11.

designed compounds have affinities close to the limit of measurement accuracy, 10 pM (pK_i = 11).

The prediction errors of the CC compounds were generally larger than those of the non-CC compounds. However, the magnitude of these errors decreased when the highest training set pK_i values were replaced by ranges that more realistically expressed the uncertainty of these experimental data. This approach effectively capped the pK_i values at 11 and reduced the predicted affinities to values closer to those observed experimentally. The prediction error of compound **36** with observed pK_i = 9.63, was still relatively large, however. This compound differs only in its R2 substituent from that of the more potent compound **32** (pK_i > 11);[5] the R2 substituents are 8 and 1, respectively, which only differ by a methyl group. From additivity considerations, one would expect their pK_i difference to be very small, as the value of the additivity parameter for the replacement of the R2 substituent 1 with substituent 8 is −0.01 (Table 11). The differing consequences of the methyl group for these two pairs of compounds point to a breakdown of additivity or perhaps experimental uncertainties greater than supposed.

The generality of the present conclusions regarding additivity are necessarily limited by the number and accuracy of the available data. For example, some of the models described here were generated using a training set that includes only one or a few instances of a given substituent. The affinity contributions assigned to such substituents are therefore not optimally tested and, to the extent that the system is nonadditive, the apparent accuracy of the additive model will depend upon which specific compounds appear in the training and test sets, as reported in Section 3.1.3.1. On the other hand, the tests reported here are strengthened by the fact that they involve blind predictions of

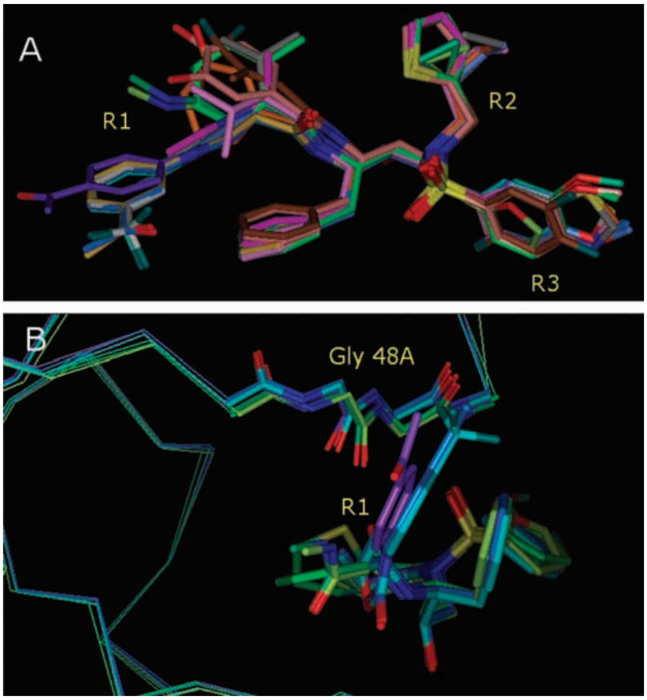


Figure 4. Superimposed HIV protease inhibitors. (A) Molecules **38** (magenta), **42** (brown), **56** (green-cyan), **28** (cyan), **32** (blue), **75** (cobalt blue), **82** (orange), **94** (pink), **15** (violet), **57** (green-yellow), **31** (off-white), **34** (gold), **86** (gray), **50** (lime green), **52** (salmon pink). (B) Superimposed CC HIV protease inhibitors **28** (cyan), **32** (blue) and **15** (violet), and the non-CC inhibitors **56** (green-cyan), **57** (green-yellow) and **50** (lime green). The orientations of the Gly 48A carbonyl groups in the HIV protease structures with CC inhibitors bound are seen to differ from those in the non-CC inhibitor-bound crystal structures. For simplicity, only one protease monomer is shown for each structure.

the affinities of new compounds. It is also worth emphasizing that, although a more redundant data set would make the statistics more informative and reliable, it would not necessarily improve the additivity of the models, which derives not from the statistics but from the physics of the protein−ligand system (see below). In fact, if the system were perfectly additive, then a training set with only $M_1 + M_2 + M_3$ compounds, where M_i is the number of different substituents at site R_i, would allow us to predict the affinities of all $M_1 \times M_2 \times M_3 - (M_1 + M_2 + M_3)$ other compounds in the combinatorial library with perfect accuracy. Finally, it is worth emphasizing that experimental uncertainty can lead even an additive system to appear nonadditive, and it is worth accounting explicitly for least the larger known uncertainties when constructing and testing an additive model (Section 3.1.3.1).

As just discussed, the accuracy of the pK_i predictions can depend upon the molecules in the training set. Thus, interchanging four molecules between test and training sets with a total of 27 + 4 = 31 compounds led to noticeably different accuracies. In one case, the predicted pK_is were consistently too high, and in the other, the pK_is were too low. In addition, the parameter of the R1 substituent common to the interchanged molecules was assigned values differing by 0.45 pK_i units. These artificial training and test sets exhibit behavior similar to that of the previous sets of molecules, which displayed consistent overprediction of pK_i values and changes in parameter values upon changes to the training set. Thus, the model generated by a given set of pK_i values should necessarily be seen as provisional and should be updated upon obtaining relevant new data. We sought to reduce the sensitivity of the predictions to the choice of training set data by using ridge regression instead of ordinary least-squares. The apparently optimal value of the ridge parameter, $\lambda = 3$, the largest value tried that did not increase the sum-squared residuals by an excessive amount, generally led to only a minor improvement in prediction accuracy. This suggests that the use of ordinary least-squares regression is not a major reason for errors in the original predictions. Still, for three of the four test sets, increasing the value of λ did somewhat improve the pK_i predictions (Table 10).

Table 11. Third Additivity Model (model 3), with Substituent Parameters Obtained from Least-Squares Fitting of pK_i Values from 107 Molecules and with pK_i Values above 11 Replaced by the Range 11−13[a]

		least-squares fit				
		parameter (1)[b]	parameter (2)	parameter (3)	n (3)	95% CI (3)
R1:	1	0.00	0.00	0.00	5	0.00−0.00
	2	0.24	0.16	0.03	8	−0.30−0.56
	3	0.58	0.50	0.37	8	−0.11−0.80
	4	0.15	0.10	−0.02	5	−0.44−0.40
	5	0.87	0.82	0.70	4	0.18−3.00
	6	1.10	1.05	0.54	5	−0.21−2.96
	7	0.97	0.77	0.30	3	−0.07−0.79
	8	−0.54	−0.74	−0.88	1	−1.51 to −0.33
	9	0.58	*−0.02*	−0.09	3	−0.59−0.80
	10	−0.10	*−0.35*	−0.43	3	−1.84 to −0.08
	11	−0.88	−1.08	−1.23	1	−1.74 to −0.82
	12	−0.40	−0.54	−0.74	1	−1.31 to −0.27
	13	0.71	*0.30*	0.17	3	−0.09−0.75
	14	0.61	0.47	−0.15	2	−0.81−0.68
	15	−0.75	−0.89	−1.09	1	−1.56 to −0.62
	16	−0.45	−0.59	−0.79	1	−1.39 to −0.32
	17	−0.66	−0.77	−0.95	5	−2.40−0.47
	18	−2.28	−2.39	−2.57	5	−3.75 to −1.09
	19	−1.86	−1.97	−2.16	5	−3.62 to −0.62
	20		*−0.60*	−0.46	6	−1.00 to −0.03
	21		*−0.26*	−0.12	6	−0.48−0.20
	22		*0.23*	0.37	6	−0.17−0.94
	23		*−0.89*	−0.75	6	−1.09 to −0.29
	24		*−0.44*	−0.31	6	−0.61 to −0.03
	25		*−0.59*	−0.45	6	−0.71 to −0.15
R2:	1	0.00	*0.00*	0.00	38	0.00−0.00
	2	−1.77	−1.67	−1.53	18	−2.13 to −0.90
	3	−1.48	−1.48	−1.48	2	−2.13 to −0.91
	4	−1.01	−0.95	−0.82	9	−1.64 to −0.22
	5	−1.58	−1.51	−1.38	3	−3.03 to −0.30
	6	−1.22	−1.15	−1.03	6	−2.43−0.23
	7		*−0.49*	−0.61	12	−0.96 to −0.23
	8		*0.23*	−0.01	17	−0.35−0.33
R3:	1	0.00	*0.00*	0.00	33	0.00− 0.00
	2	−1.54	−1.50	−1.44	21	−1.79 to −0.94
	3	−1.72	−1.62	−1.61	8	−2.30 to −0.75
	4	−0.61	−0.56	−0.39	6	−0.76−0.08
	5	−0.49	−0.43	−0.31	4	−0.80−0.27
	6	−0.05	0.01	0.20	5	−0.20−0.97
	7	−3.74	−3.41	−3.48	2	−4.40 to −1.89
	8	−1.43	−1.40	−1.33	6	−2.69 to −0.27
	9		*0.40*	0.37	20	0.11−0.60

[a] Parameters that are italicized are for substituents not used to train models 1 or 2. [b] Number in parentheses indicates refers to the relevant additivity model. Cyclic carbamate compounds are above the solid line.

The accuracy of the additivity approach was furthermore compared to that of standard descriptor-based QSAR methods. The results were mixed, with each method yielding slightly better accuracy in some cases but not in others. However, the additivity method is arguably superior in the present application because it provides clear guidance on the contributions of individual substituents to affinity. Thus, it is trivial to use additivity information to propose new molecules for synthesis,

Table 12. Comparison of pK_i Predictions via Additivity with Predictions from Whole Molecule Descriptor-Based QSAR Methods[a]

						QSAR			
molecule	R1	R2	R3	obs[d]	add	RR	GARR	PLS	GAPLS
				Model 1					
37	7	1	1	9.93	10.97	9.88	10.05	9.74	10.57
41	9	1	1	9.41	10.58	9.82	10.99	9.68	11.02
44	10	1	1	9.77	9.90	9.64	10.24	9.56	10.27
49	13	1	1	10.03	10.71	10.42	10.65	10.46	11.12
rms prediction error					0.85	0.29	0.88	0.29	1.05
				Model 2					
35	6	8	1	9.63	11.29	10.55	10.48	10.48	10.78
36	6	8	9	10.72	11.69	11.10	10.70	11.11	11.22
37	7	1	1	9.93	10.77	9.66	10.15	9.72	10.18
39	7	8	1	10.34	11.01	9.48	9.48	9.45	9.60
51	13	8	9	10.48	10.93	10.60	10.14	10.70	10.34
53	14	8	1	9.42	10.70	9.56	9.21	9.61	9.44
rms prediction error					1.06	0.55	0.53	0.55	0.61
				Model A					
16	3	1	2	9.08	9.18	9.42	9.96	9.61	9.89
17	3	1	4	9.74	10.26	9.98	10.16	9.99	10.13
18	3	1	5	10.10	10.33	10.49	10.72	10.29	10.65
22	3	4	1	9.10	10.05	10.06	10.11	10.20	10.11
rms prediction error					0.56	0.56	0.77	0.63	0.74
				Model B					
15	3	1	1	>11[c]	10.31	10.34	10.15	10.23	10.17
19	3	1	6	10.80	10.21	10.37	10.25	9.90	10.13
20	3	2	2	7.53	6.88	6.96	7.12	7.01	7.09
21	3	2	3	6.78	6.99	7.12	7.61	7.49	7.60
rms prediction error					0.57[d]	0.52[d]	0.69[d]	0.74[d]	0.71[d]

[a] The tests set are the same as in Table 10. [b] obs: observed pK_i values; add: pK_i predictions from the original additivity method; RR: ridge regression; GARR: ridge regression with genetic algorithm used for descriptor selection; PLS: partial least-squares; GAPLS: partial least-squares with genetic algorithm used for descriptor selection. [c] pK_i range specified as 11−13 for the calculation. [d] Calculation performed assuming an observed pK_i of 11.

albeit with the restriction that the substituents must have been included in at least one training set molecule. Furthermore, the additivity-based method can distinguish between chemically similar substituents, such as R1 substituents 14 and 15, whose contributions to binding affinity differ by approximately 1 order of magnitude. In contrast, a QSAR method, which represents molecules by whole-molecule descriptors, is less well suited to make such fine distinctions.

Why does the additivity approximation work so well for the present ligand−protein series, and will it be equally applicable to other systems? These question may be addressed by considering some of the physical requirements that must be met in order for the simple additivity model to work, as follows:

1. The substituents should not contact each other in either the bound or free state, or else changing R1, for example, would change the relative affinities of ligands with different R2 substituents. This requirement is most easily met by a large ligand with small, widely separated substituents.

2. Changing one substituent should not shift the position of the other substituents in the binding site. For example, changing from a small R1 substituent to a bulky one could generate nonadditivity by shifting the entire ligand and thereby forcing R2 to move away from its original position. This requirement is perhaps most easily met if all the substituents are similar in size, the ligand has a flexible scaffold, and small changes in the conformation of the scaffold do not lead to significant changes in its interaction with the protein (for example, the flexible part of the scaffold might be solvent-exposed).

3. Changing R1 must not cause a change in protein conformation that propagates to the interaction site of R2, because this, too, would alter the interaction of R2 with the protein. Meeting this requirement may be facilitated if all the substituents at a given site make similar interactions with the protein and if the protein is very rigid or, perhaps, so soft that conformational shifts remain local.

There may also be nonadditive effects on conformational fluctuations, and hence the entropy, of the ligand and the protein, but it is not clear what additional requirements avoiding these nonadditivities places on the system.

The protein−ligand series considered here does, arguably, go a long way to meeting these requirements, given the flexibility of the scaffold (Figure 1) and the fact that the substituents bind in discrete regions of the protein without contacting each other (Figure 4). In addition, the substituents at each site are fairly similar in size, and the largest variations in size occur at R1 and R3, which lie at the two entrances of the active site tunnel where substituents of varied size can be accommodated without steric clashes or protein reorganization. The available crystal structures bear out the expectation that changes at one ligand site produce minimal perturbation at the other sites. The additivity approximation may well be useful for other protein−ligand systems that meet the requirements laid out above.

5. Summary

The concept of substituent additivity proves to be a pleasingly simple and practical way to use existing binding affinity data for the design of new HIV protease inhibitors with high affinity and a good fit to the substrate envelope. The present study also has broader significance, because there are few other articles in which an additivity analysis was followed up and tested by synthesis of compounds containing the most favorable substituents (but see ref 31). The accuracy of initial predictions with the additive model motivated a careful look at this methodology, highlighting the provisional nature of the mathematical models obtained and the importance of accounting for measurement imprecision for high affinity compounds. Correlations with crystallographic data rationalize the observed additivity and the contributions of the various substituents to binding affinity. Analysis of the physical requirements for substituent independence provide guidance to the identification of other systems where additivity analysis is likely to be useful.

6. Experimental Section (Chemistry)

6.1. General. Proton nuclear magnetic resonance (^1H NMR) and carbon nuclear magnetic resonance (^{13}C NMR) spectra were recorded with a Varian Mercury 400 MHz NMR spectrometer operating at 400 MHz for ^1H and 100 MHz for ^{13}C NMR. Chemical shifts are reported in ppm (δ scale) relative to the solvent signal, and coupling constant (J) values are reported in Hertz. Data are represented as follows: chemical shift, multiplicity (s = singlet, d = doublet, t = triplet, q = quartet, m = multiplet, dd = doublet of doublets, br = broad), coupling constant in Hz, and integration. High resolution mass spectra (HRMS) were recorded on Waters Q-TOF Premier mass spectrometer by direct infusion of solutions of each compound using electrospray ionization (ESI) in positive mode. All reactions were performed in oven-dried round-bottomed or modified Schlenk flasks fitted with rubber septa under N_2 atmosphere unless otherwise noted. All final coupling reactions were carried out at 0.5 mmol scale unless stated otherwise. Air- and moisture-sensitive liquids, and solutions were transferred via syringe or stainless steel cannula. Organic solutions were concentrated under reduced pressure by rotary evaporation at 35−40 °C. Flash and column chromatography was performed using silica gel (230−400 mesh, Merck KGA). Analytical thin-layer chromatography (TLC) was performed using silica gel (60 F-254) coated aluminum plates (Merck KGA) and spots were visualized by exposure to ultraviolet light (UV) and/or exposure to an acidic solution of *p*-anisaldehyde (anisaldehyde) followed by brief heating. Dichloromethane was dried over P_2O_5 and distilled, tetrahydrofuran (THF) was distilled from sodium/benzophenone, and anhydrous *N*,*N*-dimethylformamide was purchased from Aldrich and used as received. All other reagents and solvents were purchased from commercial sources and used as received.

6.2. Typical Procedure for the Coupling Reactions (Method A). 6.2.1. (5*S*)-3-(3-Acetylphenyl)-*N*-[(1*S*,2*R*)-2-hydroxy-3-[[(4-methoxyphenyl)sulfonyl][(2*S*)-2-methylbutyl]amino]-1-(phenylmethyl)propyl]-2-oxo-oxazolidine-5-carboxamide (35). Excess oxalyl chloride was added to solid (*S*)-3-(3-acetylphenyl)-2-oxo-oxazolidine-5-carboxylic acid (0.125 g, 0.5 mmol), and the resulting mixture was stirred at room temperature overnight. The oxalyl chloride was removed by distillation under reduced pressure and residue dried under high vacuum for 30 min. A solution of the resulting acid chloride in dry THF (5 mL) was used in the coupling reaction. To an ice-cooled mixture of the Boc deprotected amine (0.5 mmol) in dry THF (5 mL) was added Et_3N (0.15 mL, 1.1 mmol), followed by the slow addition of the acid chloride solution. After 15 min, the reaction mixture was warmed to room temperature and stirred until reaction was complete (monitored by TLC). Small amounts of water and ethyl acetate were added, and layers were separated. The organic extract was washed with saturated aqueous NaCl solution, dried (Na_2SO_4), filtered, and evaporated. The residue was purified by flash chromatography on silica gel using ethyl acetate−hexanes (3:1) mixture as eluent to provide the target compound (0.30 g, 92%) as white solid. ^{1}H (400 MHz, $CDCl_3$) δ 7.89 (t, J = 2.0 Hz, 1H), 7.83 (m, 1H), 7.77 (m, 1H), 7.76−7.72 (m, 2H), 7.52 (t, J = 8.4 Hz, 1H), 7.13 (dd, J = 8.4, 1.6 Hz, 2H), 7.03−6.98 (m, 4H), 6.86 (dt, J = 8.4, 1.2 Hz, 1H), 6.75 (d, J = 10.0 Hz, 1H), 4.80 (dd, J = 9.6, 5.6 Hz, 1H), 4.25 (m, 1H), 4.08 (t, J = 9.6 Hz, 1H), 3.92 (m, 1H), 3.87 (s, 3H), 3.65 (d, J = 2.4 Hz, 1H), 3.41 (dd, J = 9.6, 6.0 Hz, 1H), 3.20 (dd, J = 15.6, 9.6 Hz, 1H), 3.12−3.04 (m, 2H), 2.98 (dd, J = 15.2, 2.8 Hz, 1H), 2.82 (dd, J = 13.2, 7.2 Hz, 1H), 2.76 (dd, J = 13.6, 10.4 Hz, 1H), 2.65 (s, 3H), 1.62 (m, 1H), 1.52 (m, 1H), 1.11 (m, 1H), 0.90−0.86 (m, 6H). ^{13}C NMR (100 MHz, $CDCl_3$) δ 197.64, 168.60, 163.37, 153.04, 138.12, 138.08, 137.42, 129.79 (2C), 129.76, 129.72, 129.60 (2C), 128.63 (2C), 126.73, 124.77, 123.04, 117.56, 114.65 (2C), 72.40, 69.91, 57.61, 55.89, 53.87, 53.39, 48.25, 35.66, 33.74, 26.98, 26.65, 17.16, 11.26. HRMS (ESI) *m*/*z*: calcd for $C_{34}H_{42}N_3O_8S$ [M + H]$^+$ 652.2693; found, 652.2714.

6.2.2. (5*S*)-3-(3-Acetylphenyl)-*N*-[(1*S*,2*R*)-3-[(6-benzothiazolylsulfonyl)[(2*S*)-2-methylbutyl]amino]-2-hydroxy-1-(phenylmethyl)propyl]-2-oxo-oxazolidine-5-carboxamide (36). Coupling method A; solvent for flash chromatography: EtOAc−hexanes (4:1); yield: 0.310 g, 91%; white solid. ^{1}H NMR (400 MHz, $CDCl_3$) δ 9.22 (s, 1H), 8.49 (d, J = 1.6 Hz, 1H), 8.27 (d, J = 8.4 Hz, 1H), 7.93−7.90 (m, 2H), 7.82−7.76 (m, 2H), 7.52 (t, J = 8.4 Hz, 1H), 7.13 (dd, J = 8.4, 1.6 Hz, 2H), 7.02 (t, J = 8.0 Hz, 2H), 6.89−6.82 (m, 2H), 4.80 (dd, J = 9.6, 5.6 Hz, 1H), 4.26 (m, 1H), 4.08 (t, J = 9.6 Hz, 1H), 3.98 (m, 1H), 3.63 (d, J = 3.6 Hz, 1H), 3.43 (dd, J = 9.2, 5.6 Hz, 1H), 3.25 (dd, J = 15.2, 9.2 Hz, 1H), 3.16−3.05 (m, 3H), 2.93 (dd, J = 13.2, 6.8 Hz, 1H), 2.78 (dd, J = 13.6, 10.8 Hz, 1H), 2.65 (s, 3H), 1.66 (m, 1H), 1.51 (m, 1H), 1.12 (m, 1H), 0.90−0.86 (m, 6H). ^{13}C NMR (100 MHz, CDCl3) δ 197.66, 168.69, 158.32, 155.89, 153.07, 138.09, 138.07, 137.37, 135.63, 134.66, 129.73, 129.58 (2C), 128.67 (2C), 126.79, 125.09, 124.84, 124.72, 123.04, 122.60, 117.58, 72.46, 69.94, 57.62, 53.88, 53.52, 48.26, 35.70, 33.73, 26.98, 26.67, 17.15, 11.28. HRMS (ESI) *m*/*z*: calcd for $C_{34}H_{39}N_4O_7S_2$ [M + H]$^+$ 679.2260; found, 679.2287.

6.3. Typical Procedure for the Coupling Reactions (Method B). 6.3.1. (2*S*)-2-(Acetylamino)-*N*-[(1*S*,2*R*)-2-hydroxy-3-[[(4-methoxyphenyl)sulfonyl](2-methylpropyl)amino]-1-(phenylmethyl)propyl]-propanamide (37). To a solution of the *N*-[(1*S*,2*R*)-2-Hydroxy-3-[[(4-methoxyphenyl)sulfonyl](2-methylpropyl)amino]-1-(phenylmethyl)propyl]-carbamic acid *tert*-butyl ester[5] (0.254 g, 0.5 mmol) in CH_2Cl_2 (15 mL) was added TFA (5 mL) and the mixture was stirred at room temperature for 1 h. Solvents were removed under reduced pressure, and the residue was dissolved in CH_2Cl_2, washed with 10% aqueous $NaHCO_3$ solution, dried (Na_2SO_4), filtered, and evaporated under reduced pressure to provide the free amine as white solid. To an ice-cooled solution of this amine in a mixture of $H_2O-CH_2Cl_2$ (1:1) (12 mL) were added *N*-Ac-Ala-OH (0.079 g, 0.6 mmol) followed by HOBt (0.092 g, 0.6 mmol) and EDCI (0.115 g, 0.6 mmol) under N_2 atmosphere. The reaction mixture was stirred at 0−4 °C until the reaction was complete (monitored by TLC). A small amount of CH_2Cl_2 was added and layers were separated. The organic extract was washed with saturated aqueous NaCl solution, dried (Na_2SO_4), filtered, and evaporated under reduced pressure. The residue was purified by flash chromatography on silica gel using $CHCl_3$−MeOH (19:1) mixture as eluent to provide the target compound (0.235 g, 90%) as white solid. ^{1}H (400 MHz, $CDCl_3$) δ. 7.74−7.70 (m, 2H), 7.27−7.16 (m, 5H), 6.99−6.95 (m, 2H), 6.65 (d, J = 8.8 Hz, 1H), 5.87 (d, J = 7.6 Hz, 1H), 4.32 (m, 1H), 4.15 (m, 1H), 4.07 (d, J = 3.6 Hz, 1H), 3.86 (s, 3H), 3.85 (m, 1H, overlapping signal), 3.14−3.02 (m, 3H), 2.92−2.84 (m, 3H), 1.88 (s, 3H), 1.84 (m, 1H), 1.19 (d, J = 6.8 Hz, 3H), 0.87 (d, J = 6.8 Hz, 3H, overlapping signal), 0.86 (d, J = 6.4 Hz, 3H, overlapping signal). ^{13}C NMR (100 MHz, $CDCl_3$) δ 172.82, 170.29, 163.24, 138.05, 130.19, 127.72 (2C), 129.58 (2C), 128.65 (2C), 126.71, 114.57 (2C), 72.79, 58.93, 55.86, 54.20, 53.55, 49.13, 35.44, 27.41, 23.38, 20.34, 20.18, 18.27. HRMS (ESI) *m*/*z*: calcd for $C_{26}H_{38}N_3O_6S$ [M + H]$^+$ 520.2481; found, 520.2461.

6.3.2. (2*S*)-2-(Acetylamino)-*N*-[(1*S*,2*R*)-2-hydroxy-3-[[(4-methoxyphenyl)sulfonyl][(2*S*)-2-methylbutyl]amino]-1-(phenylmethyl)propyl]-propanamide (39). Coupling method B; solvent for flash chromatography: EtOAc-hexanes (3:2); yield: 0.235 g, 88%; white solid. ^{1}H (400 MHz, $CDCl_3$) δ 7.74−7.70 (m, 2H), 7.27−7.16 (m, 5H), 6.99−6.95 (m, 2H), 6.61 (d, J = 8.8 Hz, 1H), 5.86 (d, J = 7.6 Hz, 1H), 4.33 (m, 1H), 4.16 (m, 1H), 4.02 (d, J = 3.6 Hz, 1H), 3.86 (s, 3H), 3.83 (m, 1H), 3.12−3.0 (m, 3H), 2.97−2.88 (m, 2H), 2.82 (dd, J = 13.2, 7.6 Hz, 1H), 1.89 (s, 3H), 1.60 (m, 1H), 1.44 (m, 1H), 1.20 (d, J = 6.8 Hz, 3H), 1.05 (m, 1H), 0.86−0.82 (m, 6H). ^{13}C NMR (100 MHz, CDCl3) δ 172.81, 170.27, 163.24, 138.0, 130.11, 129.73 (2C), 129.60 (2C), 128.65 (2C), 126.72, 114.57 (2C), 72.73, 57.58, 55.86, 54.11, 53.54, 49.13, 35.45, 33.63, 26.79, 23.37, 18.35, 17.12, 11.33. HRMS (ESI) *m*/*z*: calcd for $C_{27}H_{40}N_3O_6S$ [M + H]$^+$ 534.2638; found, 534.2630.

6.3.3. *N*-[(1*S*,2*R*)-2-Hydroxy-3-[[(4-methoxyphenyl)sulfonyl](2-methylpropyl)amino]-1-(phenylmethyl)propyl]-3-methyl-4,4,4-trifluoro-2-butenamide (41). Coupling method B; solvent for flash chromatography: EtOAc−hexanes (1:1); yield: 0.240 g, 88%; gummy solid. ^{1}H (400 MHz, $CDCl_3$) δ 7.70−7.66 (m, 2H), 7.32−7.20 (m, 5H), 6.99−6.95 (m, 2H), 6.14 (m, 1H), 5.95 (d, J = 8.8 Hz, 1H), 4.26 (m, 1H), 3.99 (d, J = 2.8 Hz, 1H), 3.87 (s, 3H), 3.11 (dd, J = 15.2, 8.8 Hz, 1H), 3.03 (dd, J = 14.0, 5.6 Hz, 1H), 2.99−2.90 (m, 3H), 2.79 (dd, J = 13.6, 6.8 Hz, 1H), 2.04 (d, J = 1.6 Hz, 3H), 1.83 (m, 1H), 0.90 (d, J = 6.4 Hz, 3H), 0.87 (d, J = 6.4 Hz, 3H). ^{13}C NMR (100 MHz, CDCl3) δ 164.74, 163.34, 138.64 (q, J = 30.0 Hz), 137.69, 129.90, 129.67 (2C), 129.52 (2C), 128.90 (2C), 126.95, 124.82 (t, J = 272.6 Hz), 123.50 (q, J = 5.2 Hz), 114.60 (2C), 72.74, 59.09, 55.87, 53.92, 53.85, 34.90, 27.55, 20.34, 20.13, 12.17. HRMS (ESI) *m*/*z*: calcd for $C_{26}H_{34}F_3N_2O_5S$ [M + H]$^+$ 543.2141; found, 543.2100.

6.3.4. *N*-[(1*S*,2*R*)-2-Hydroxy-3-[[(4-methoxyphenyl)sulfonyl](2-methylpropyl)amino]-1-(phenylmethyl)propyl]-4-oxo-2-pentenamide (44). Coupling method B; solvent for flash chromatography: EtOAc−hexanes (3:2); yield: 0.225 g, 89%; white foamy solid. ^{1}H (400 MHz, CDCl$_3$) δ 7.70−7.66 (m, 2H), 7.31−7.19 (m, 5H), 7.99−6.93 (m, 2H), 6.91 (d, $J = 15.6$ Hz, 1H), 6.59 (d, $J = 15.2$ Hz, 1H), 6.22 (d, $J = 8.8$ Hz), 4.28 (m, 1H), 4.06 (d, $J = 2.8$ Hz, 1H), 3.93 (m, 1H), 3.87 (s, 3H), 3.10 (dd, $J = 15.2$, 8.4 Hz, 1H), 3.03−2.96 (m, 3H), 2.91 (dd, $J = 14.0$, 8.4 Hz, 1H), 2.79 (dd, $J = 13.6$, 6.8 Hz, 1H), 2.31 (s, 3H), 1.82 (m, 1H), 0.88 (d, $J = 6.8$ Hz, 3H, overlapping signal), 0.86 (d, $J = 6.8$ Hz, 3H, overlapping signal). ^{13}C NMR (100 MHz, CDCl$_3$) δ 197.81, 164.51, 163.33, 137.61, 137.21, 133.64, 129.89, 129.67 (2C), 129.52 (2C), 128.91 (2C), 126.98, 114.62 (2C), 72.65, 59.07, 55.88, 54.46, 53.75, 34.84, 29.11, 27.52, 20.35, 20.14. HRMS (ESI) *m*/*z*: calcd for $C_{26}H_{35}N_2O_6S$ $[M + H]^+$ 503.2216; found, 503.2221.

6.3.5. 3-Fluoro-*N*-[(1*S*,2*R*)-2-hydroxy-3-[[(4-methoxyphenyl)sulfonyl](2-methylpropyl)amino]-1-(phenylmethyl)propyl]-2-methyl-benzamide (49). Coupling method B; solvent for flash chromatography: EtOAc-hexanes (1:1); yield: 0.220 g, 81%; white foamy solid. ^{1}H (400 MHz, CDCl$_3$) δ 7.71 (m, 2H), 7.33−7.21 (m, 5H), 7.10−7.0 (m, 2H), 6.98 (m, 2H), 6.78 (dd, $J = 7.6$, 1.2 Hz, 1H), 5.98 (d, $J = 8.8$ Hz, 1H), 4.38 (m, 1H), 3.99 (m, 1H), 3.87 (s, 3H), 3.21−3.12 (m, 3H), 3.01−2.93 (m, 2H), 2.86 (dd, $J = 13.2$, 6.8 Hz, 1H), 2.07 (d, $J = 2.0$ Hz, 3H), 1.89 (m, 1H), 0.93 (d, $J = 6.8$ Hz, 3H), 0.89 (d, $J = 6.8$ Hz, 3H). ^{13}C NMR (100 MHz, CDCl$_3$) δ 169.43 (d, $J = 2.9$ Hz), 163.33, 161.54 (d, $J = 243.9$ Hz), 138.40 (d, $J = 4.4$ Hz), 137.89, 129.99, 129.68 (2C), 129.57 (2C), 128.91 (2C), 127.24 (d, $J = 8.1$ Hz), 126.98, 123.73 (d, $J = 18.3$ Hz), 122.21 (d, $J = 3.7$ Hz), 117.0 (d, $J = 22.7$ Hz), 114.62 (2C), 73.19, 59.19, 55.88, 54.40, 53.94, 35.14, 27.58, 20.35, 20.16, 11.32 (d, = 4.4 Hz). HRMS (ESI) *m*/*z*: calcd for $C_{29}H_{36}FN_2O_5S$ $[M + H]^+$ 543.2329; found, 543.2319.

6.3.6. 3-Fluoro-*N*-[(1*S*,2*R*)-2-hydroxy-3-[[(4-methoxyphenyl)sulfonyl][(2*S*)-2-methylbutyl]amino]-1-(phenylmethyl)propyl]-2-methyl-benzamide (51). Coupling method B; solvent for flash chromatography: EtOAc-hexanes (3:1); yield: 0.240 g, 82%; white solid. ^{1}H (400 MHz, CDCl$_3$) δ 9.21 (s, 1H), 8.44 (d, $J = 1.6$ Hz, 1H), 8.25 (d, $J = 8.8$ Hz, 1H), 7.88 (dd, $J = 8.8$, 1.6 Hz, 1H), 7.33−7.21 (m, 5H), 7.11−7.0 (m, 2H), 6.79 (dd, $J = 8.8$, 1.2 Hz, 1H), 6.03 (d, $J = 8.4$ Hz, 1H), 4.39 (m, 1H), 4.02 (m, 1H), 3.22 (m, 2H), 3.16 (dd, $J = 14.0$, 5.2 Hz, 1H), 3.10 (dd, $J = 13.6$, 7.6 Hz, 1H), 3.0 (dd, $J = 14.0$, 9.6 Hz, 1H), 2.94 (dd, $J = 13.6$, 8.0 Hz, 1H), 2.08 (d, $J = 2.0$ Hz, 3H), 1.68 (m, 1H), 1.49 (m, 1H), 1.09 (m, 1H), 0.88−0.82 (m, 6H). ^{13}C NMR (100 MHz, CDCl$_3$) δ 169.49 (d, $J = 3.0$ Hz), 161.55 (d, $J = 243.9$ Hz), 158.27, 155.83, 138.29 (d, $J = 4.4$ Hz), 137.75, 135.81, 134.63, 129.56 (2C), 128.97 (2C), 127.29 (d, $J = 8.1$ Hz), 127.09, 125.03, 124.68, 123.76 (d, $J = 18.3$ Hz), 122.47, 122.20 (d, $J = 3.7$ Hz), 117.0 (d, $J = 22.8$ Hz), 73.09, 57.67, 54.54, 53.76, 35.16, 33.77, 26.74, 17.12, 11.35 (d, $J = 3.0$ Hz), 11.33. HRMS (ESI) *m*/*z*: calcd for $C_{30}H_{35}FN_3O_4S_2$ $[M + H]^+$ 584.2053; found, 584.2068.

6.3.7. 3,4-Dihydroxy-*N*-[(1*S*,2*R*)-2-hydroxy-3-[[(4-methoxyphenyl)sulfonyl][(2*S*)-2-methylbutyl]amino]-1-(phenylmethyl)propyl]-benzamide (53). Coupling method B; solvent for flash chromatography: EtOAc; yield: 0.160 g, 57%; white solid. ^{1}H (400 MHz, CDCl$_3$ + 2 drops CD$_3$OD) δ 7.63−7.59 (m, 2H), 7.25 (m, 5H), 7.18 (m, 1H), 7.14 (d, $J = 2.0$ Hz, 1H), 7.0 (dd, $J = 8.8$, 2.4 Hz, 1H), 6.91−6.87 (m, 2H), 6.80 (d, $J = 8.0$ Hz, 1H), 6.68 (d, $J = 8.4$ Hz, 1H), 4.28 (m, 1H), 3.92 (m, 1H), 3.82 (s, 3H), 3.16 (dd, $J = 14.8$, 4.0 Hz, 1H), 3.04 (d, $J = 6.8$ Hz, 1H), 2.96 (dd, $J = 14.8$, 7.6 Hz, 1H), 2.85 (d, $J = 7.2$ Hz, 1H), 2.16 (m, 2H), 1.60 (m, 1H), 1.36 (m, 1H), 1.01 (m, 1H), 0.83−0.78 (m, 6H). ^{13}C NMR (100 MHz, CDCl$_3$ + 2 drops CD$_3$OD) δ 168.42, 168.34, 163.26, 148.64, 144.40, 138.03, 129.64 (4C), 128.80 (2C), 126.81, 125.70, 119.52, 114.78, 114.54 (2C), 72.95, 57.64, 55.83, 54.45, 53.66, 35.32, 33.61, 26.92, 17.05, 11.28. HRMS (ESI) *m*/*z*: calcd for $C_{29}H_{37}N_2O_7S$ $[M + H]^+$ 557.2321; found, 557.2341.

Acknowledgment. We thank Dr. Madhavi Nalam for sharing crystal structures prior to publication and Dr. Sune Peterson for pointing out that our additivity analysis “reinvented the wheel” first devised by Free and Wilson. A generous gift of chiral epoxides from Kaneka USA is gratefully acknowledged. This publication was made possible by grants nos. GM066524 and GM061300 from the National Institute of General Medical Sciences (NIGMS) of the NIH. The contents of this publication are solely the responsibility of the authors and do not necessarily represent the official views of the NIGMS.

References

(1) Joint United Nations Programme on HIV/AIDS (UNAIDS) and World Health Organization (WHO). *AIDS Epidemic Update: December, 2007*; **2007**; http://www.unaids.org.

(2) Richman, D. D. HIV chemotherapy. *Nature* **2001**, *410*, 995–1001.

(3) Pauwels, R. New non-nucleoside reverse transcriptase inhibitors (NNRTIs) in development for the treatment of HIV infections. *Curr. Opin. Pharmacol.* **2004**, *4*, 437–446.

(4) Turner, S. R. HIV protease inhibitors—the next generation. *Curr. Med. Chem.: Anti-Inf. Agents* **2002**, *1*, 141–162.

(5) Ali, A.; Reddy, G. S. K. K.; Cao, H.; Anjum, S. G.; Nalam, M. N. L.; Schiffer, C. A.; Rana, T. M. Discovery of HIV-1 protease inhibitors with picomolar affinities incorporating *N*-aryl-oxazolidinone-5-carboxamides as novel P2 ligands. *J. Med. Chem.* **2006**, *49*, 7342–7356.

(6) Reddy, G. S. K. K.; Ali, A.; Nalam, M. N. L.; Anjum, S. G.; Cao, H.; Nathans, R. S.; Schiffer, C. A.; Rana, T. M. Design and synthesis of HIV-1 protease inhibitors incorporating oxazolidinones as P2/P2′ ligands in pseudosymmetric dipeptide isosteres. *J. Med. Chem.* **2007**, *50*, 4316–4328.

(7) Altman, M. D.; Ali, A.; Reddy, G. S. K. K.; Nalam, M. N.; Ghafoor, S.; Cao, H.; Chellappan, S.; Kairys, V.; Fernandes, M. X.; Gilson, M. K.; Schiffer, C. A.; Rana, T. M.; Tidor, B. HIV-1 protease inhibitors from inverse design in the substrate envelope exhibit subnanomolar binding to drug-resistant variants. *J. Am. Chem. Soc.* **2008**, *30*, 6099–6113.

(8) Chellappan, S.; Reddy, G. S. K. K.; Ali, A.; Nalam, M. N. L.; Anjum, S. G.; Cao, H.; Kairys, V.; Fernandes, M. X.; Altman, M. D.; Tidor, B.; Rana, T. M.; Schiffer, C. A.; Gilson, M. K. Design of mutation-resistant HIV protease inhibitors with the substrate envelope hypothesis. *Chem. Biol. Drug Des.* **2007**, *69*, 298–313.

(9) Prabu-Jeyabalan, M.; Nalivaika, E.; Schiffer, C. A. Substrate shape determines specificity of recognition for HIV-1 protease: analysis of crystal structures of six substrate complexes. *Structure.* **2002**, *10*, 369–381.

(10) King, N. M.; Prabu Jeyabalan, M.; Nalivaika, E. A.; Schiffer., C. A. Combating susceptibility to drug resistance: lessons from HIV-1 protease. *Chem. Biol* **2004**, *11*, 1333–1338.

(11) King, N. M.; Prabu-Jeyabalan, M.; Nalivaika, E. A.; Wigerinck, P.; de Bethune, M.-P.; Schiffer., C. A. Structural and thermodynamic basis for the binding of TMC114, a next-generation human immunodeficiency virus type 1 protease inhibitor. *J. Virol.* **2004**, *78*, 12012–12021.

(12) Prabu-Jeyabalan, M.; King, N. M.; Nalivaika, E. A.; Heilek-Snyder, G.; Cammack, N.; Schiffer, C. A. Substrate envelope and drug resistance: crystal structure of RO1 in complex with wild-type human immunodeficiency virus type 1 protease. *Antimicrob. Agents Chemother.* **2006**, *50*, 1518–1521.

(13) Chellappan, S.; Kairys, V.; Fernandes, M. X.; Schiffer, C.; Gilson, M. K. Evaluation of the substrate envelope hypothesis for inhibitors of HIV-1 protease. *Proteins.* **2007**, *68*, 561–567.

(14) Free, S. M., Jr.; Wilson, J. W. A mathematical contribution to structure−activity studies. *J. Med. Chem.* **1964**, *7*, 395–399.

(15) Press, W. H.; Flannery, B. P.; Teukolsky, S. A.; Vetterling, W. T. *Numerical Recipes in C: The Art of Scientific Computing*: Cambridge University Press: New York, 1992.

(16) Cook, D. C Code for Computing: A Grand Tour; Department of Statistics, Iowa State University: Ames, IA, 1997; http://www.public.iastate.edu/∼dicook/JSS/paper/code.html.

(17) Efron, B.; Tibshirani, R. J. *An Introduction to the Bootstrap*; Chapman & Hall/CRC: Boca Raton, FL, 1994.

(18) Hoerl, A. E.; Kennard, R. W. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* **1970**, *12*, 55–67.

(19) Frank, I. E.; Friedman, J. H. A Statistical View of Some Chemometrics Regression Tools. *Technometric.* **1993**, *35*, 109–135.

(20) Geladi, P.; Kowalski, B. R. Partial least squares regression- a tutorial. *Anal. Chim. Acta* **1986**, *185*, l–17.

(21) Todeschini, R.; Consonni, V., Pavan, M., *DRAGON 2.1*; Milano Chemometrics and QSAR Research Group: Milan, Italy, 2002.

(22) Hoffman, B.; Cho, S. J.; Zheng, W.; Wyrick, S.; Nichols, D. E.; Mailman, R. B.; Tropsha, A. Quantitative structure−activity relationship modeling of dopamine D(1) antagonists using comparative molecular field analysis, genetic algorithms-partial least-squares, and K nearest neighbor methods. *J. Med. Chem.* **1999**, *42*, 3217–3226.

(23) Golub, G. H.; Heath, M.; Wahba, G. Generalised cross-validation as a method for choosing a good ridge parameter. *Technometrics* **1979**, *21*, 215–223.

(24) Orr, M. J. L. Introduction to Radial Basis Function Networks; http://anc.ed.ac.uk/rbf/rbf.html.

(25) Ho, G.-J.; Emerson, K. M.; Mathre, D. J.; Shuman, R. F.; Grabowski, E. J. J. Carbodiimide-mediated amide formation in a two-phase system. A high-yield and low-racemization procedure for peptide synthesis. *J. Org. Chem.* **1995**, *60*, 3569–3570.

(26) Matayoshi, E. D.; Wang, G. T.; Krafft, G. A.; Erickson, J. Novel fluorogenic substrates for assaying retroviral proteases by resonance energy transfer. *Science* **1990**, *247*, 954–958.

(27) Greco, W. R.; Hakala, M. T. Evaluation of methods for estimating the dissociation constant of tight binding enzyme inhibitors. *J. Biol. Chem.* **1979**, *254*, 12104–12109.

(28) Kim, E. E.; Baker, C. T.; Dwyer, M. D.; Murcko, M. A.; Rao, B. G.; Tung, R. D.; Navia, M. A. Crystal structure of HIV-1 protease in complex with VX-478, a potent and orally bioavailable inhibitor of the enzyme. *J. Am. Chem. Soc.* **1995**, *117*, 1181–1182.

(29) Surleraux, D. L.; Tahri, A.; Verschueren, W. G.; Pille, G. M.; de Kock, H. A.; Jonckers, T. H.; Peeters, A.; De Meyer, S.; Azijn, H.; Pauwels, R.; de Bethune, M. P.; King, N. M.; Prabu-Jeyabalan, M.; Schiffer, C. A.; Wigerinck, P. B. Discovery and selection of TMC114, a next generation HIV-1 protease inhibitor. *J. Med. Chem.* **2005**, *48*, 1813–1822.

(30) Miller, J. F.; Andrews, C. W.; Brieger, M.; Furfine, E. S.; Hale, M. R.; Hanlon, M. H.; Hazen, R. J.; Kaldor, I.; McLean, E. W.; Reynolds, D.; Sammond, D. M.; Spaltenstein, A.; Tung, R.; Turner, E. M.; Xu, R. X.; Sherrill, R. G. Ultra-potent P1 modified arylsulfonamide HIV protease inhibitors: the discovery of GW0385. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 1788–1794.

(31) Dirlam, J. P.; Czuba, L. J.; Dominy, B. W.; James, R. B.; Pezzullo, R. M.; Presslitz, J. E.; Windisch, W. W. Synthesis and antibacterial activity of 1-hydroxy-1-methyl-1,3-dihydrofuro[3,4-*b*]quinoxaline 4,9-dioxide and related compounds. *J. Med. Chem.* **1979**, *22*, 1118–1121.

JM8009525