

Comparative Binding Energy Analysis of HIV-1 Protease Inhibitors: Incorporation of Solvent Effects and Validation as a Powerful Tool in Receptor-Based Drug Design

Carlos Pérez, Manuel Pastor, Angel R. Ortiz,[†] and Federico Gago*

Departamento de Farmacología, Universidad de Alcalá, E-28871 Alcalá de Henares, Madrid, Spain

Received August 11, 1997

A comparative binding energy (COMBINE) analysis (Ortiz et al. *J. Med. Chem.* **1995**, *38*, 2681–2691) has been performed on a training set of 33 HIV-1 protease inhibitors, and the resulting regression models have been validated using an additional external set of 16 inhibitors. This data set was originally reported by Holloway et al. (*J. Med. Chem.* **1995**, *38*, 305–317), who showed the usefulness of molecular mechanics interaction energies for predicting the activity of novel HIV-1 protease inhibitors within the framework of the MM2X force field and linear regression techniques. We first used the AMBER force field on the same set of three-dimensional structures to check up on any possible force-field dependencies. In agreement with the previous findings, the calculated raw ligand–receptor interaction energies were highly correlated with the inhibitory activities ($r^2 = 0.81$), and the linear regression model relating both magnitudes had an acceptable predictive ability both in internal validation tests ($q^2 = 0.79$, $SDEP_{cv} = 0.61$) and when applied to the external set of 16 different inhibitors ($SDEP_{ex} = 1.08$). When the interaction energies were further analyzed using the COMBINE formalism, the resulting PLS model showed improved fitting properties ($r^2 = 0.89$) and provided better estimations for the activity of the compounds in the external data set ($SDEP_{ex} = 0.83$). Computation of the electrostatic part of the ligand–receptor interactions by numerically solving the Poisson–Boltzmann equation did not improve the quality of the linear regression model. On the contrary, incorporation of the solvent-screened residue-based electrostatic interactions and two additional descriptors representing the electrostatic energy contributions to the partial desolvation of both the ligands and the receptor resulted in a COMBINE model that achieved a remarkable predictive ability, as assessed by both internal ($q^2 = 0.73$, $SDEP_{cv} = 0.69$) and external validation tests ($SDEP_{ex} = 0.59$). Finally, when all the inhibitors studied were merged into a single expanded set, a new model was obtained that explained 91% of the variance in biological activity ($r^2 = 0.91$), with very high predictive ability ($q^2 = 0.81$, $SDEP_{cv} = 0.66$). In addition, the COMBINE analysis provided valuable information about the relative importance of the contributions to the activity of individual residues that can be fruitfully used to design better inhibitors. All in all, COMBINE analysis is validated as a powerful methodology for predicting binding affinities and pharmacological activities of congeneric ligands that bind to a common receptor.

Introduction

Structural information of biological macromolecules and their complexes with ligands is increasingly being used in modern medicinal chemistry research programs. As a consequence, there is a pressing need for novel computational methods that can utilize this structural information about ligand–receptor complexes in a more quantitative way, both to improve existing leads and to design de novo compounds with accurately predicted binding affinities. This information is often crucial, and for practical use, it should be obtained at reasonable computational expense. During the drug design cycle, experimental structures for a target protein complexed to one or more ligands can allow researchers to examine the details of the binding site and to design new

analogues.¹ Once the novel ligands have been docked into the binding site, attempts are made to calculate their binding affinities in advance of experiment, to prioritize synthetic efforts.

A wide variety of methods exist for the calculation of such binding affinities.^{2,3} A first category of methods partition the free energy of binding and assume the additivity of the different components. The way the free energy is partitioned varies among the different methods, as does the formalism that is used to evaluate each term. Hydrophobic, electrostatic, and entropic contributions are usually considered, but there are uncertainties associated with the particular choice of parameters employed.⁴ These methods are computationally fast and are suitable at the beginning of a drug development project in the search for new lead compounds.

A second class of commonly used methods rely on molecular mechanics calculations.^{2,5–8} The binding energy between ligand and receptor is approximated to

[†] Present address: Department of Molecular Biology, TPC-5, The Scripps Research Institute, La Jolla, CA 92037.

* Corresponding author. Telephone: +34-1-885 45 14. Fax: +34-1-885 45 91. E-mail: fgago@fisfar.alcala.es.

the interaction enthalpy calculated by means of empirical energy functions that represent van der Waals repulsion and dispersion interactions by a Lennard-Jones term and electrostatic interactions in the form of a Coulomb term that uses atom-centered point charges.² In most cases, these molecular mechanics calculations are performed on a structure that is taken to represent the ensemble average of each complex. Entropy contributions are usually ignored although solvation terms are sometimes added to the scoring function by calculating changes in buried nonpolar surface area⁹ or differences in the ease of desolvation of both the ligands and the binding site upon complex formation.¹⁰ In favorable cases, most often involving the complexes of a given target with a congeneric series of ligands, experimental measures of activity (e.g. K_i , IC_{50}) or binding affinity (e.g. K_d) can be correlated straightforwardly with these plain molecular mechanics interaction energies making use of simple linear regression techniques.⁵⁻⁸ But even though the results of such quantitative structure-activity relationship (QSAR) studies can be satisfactory for the training set of compounds, the resulting regression models are seldom accurate enough for predicting the potency of yet unsynthesized compounds that were not included in the derivation of the model.

A third category of methods is in principle the most rigorous but also the most computationally demanding one: binding free energy differences can be obtained in the context of molecular dynamics or Monte Carlo simulations making use of free energy perturbation (FEP) or thermodynamic integration (TI) methods¹¹ that involve nonphysical stepwise conversions between pairs of rather similar ligands in the free and bound states. These procedures are very time-consuming, and it is not always clear that the results they yield are fully converged.¹¹ Recently, a more tractable variant of this approach has been presented that makes use of linear response theory¹² and requires only simulations at the endpoints of the mutations; the free energy of binding is then determined from a linear combination of the differences in the average inhibitor-environment interaction energies between the bound and unbound states.¹³ Overall, however, this last category of methods is, with present computational resources, impractical for the comparative study of even small series of receptor-ligand complexes, although the methods have the potential of providing great physical insight about the reasons for the differences in binding affinities between very similar ligands.¹⁴

We have shown that the predictive ability of a molecular mechanics receptor-based QSAR model can be considerably improved if the total interaction energy for each complex is partitioned into a number of fragment-based contributions and a subset of those energy variables that account for most of the variance within the series is selected and properly weighted. The method, termed COMBINE (comparative binding energy) analysis,¹⁵ is formally similar to CoMFA (comparative molecular field analysis) in that a data matrix is produced containing a large number of energy descriptors. Both methodologies then make use of partial least squares (PLS) analysis¹⁶ to generate a predictive model, but whereas CoMFA only provides information about the interaction properties of the ligands, COM-

BINE analysis produces a set of "weights" that represent the relative importance for activity of each residue-ligand interaction. Implicit in COMBINE analysis is the realization that experimental quantities directly related to binding free energies only seldom exhibit a good correlation with calculated binding energies. The method also assumes that energy calculations are usually "noisy" due to limited conformational sampling, inaccuracies in the modeled structures, and/or inadequacies in current potential energy functions. In such situations, the PLS method¹⁶ employed is especially useful as it can balance the different interaction energy contributions in such a way that the "signal" (characterized by showing a correlation with the dependent variable) is increased whereas the "noise" (not showing any more significant correlation with the biological property than random numbers) is reduced. The resulting PLS regression model then establishes a relation between the *differences* in activity or binding affinity among the members of the series and *differences* in some of the calculated energy contributions. The constant term in the regression equation, which corresponds to the average value of activity or binding affinity in the series, contains all those contributions that are common to all the compounds studied together with those that are neglected (such as entropy).¹⁵

The first application of COMBINE analysis was on a set of human phospholipase A₂ inhibitors for which linear regression techniques and plain molecular mechanics global interaction energies produced unacceptable QSAR models.¹⁵ The ability of PLS to increase the signal-to-noise ratio was demonstrated by producing a regression model with considerable predictive ability. As a new test case for the COMBINE methodology, we have now chosen a set of inhibitors of human immunodeficiency virus type I (HIV-1) protease developed at Merck Research Laboratories (Table 1) for which a good correlation between calculated intermolecular interaction energies (using the molecular mechanics program OPTIMOL and the MM2X force field) and enzyme inhibition was initially obtained;⁶ the resulting linear regression equation was subsequently used in a structure-based design program to predict the activity of new candidate inhibitors in advance of their synthesis and pharmacological evaluation. When the experimental pIC_{50} values were determined, it was seen that, for a set of 16 compounds, the average absolute error in the predicted values was 1.01 log units across a range of 5.10 log units, with the most significant outlier being a pseudo-symmetric inhibitor (**36** = L-700,417) whose activity was overpredicted by 3 orders of magnitude. The Merck study reported that no improvements in the quality of their model could be achieved upon incorporation of solvation effects, via a continuum solvation method,¹⁷ or use of the CHARMM force field.⁶

With this same set of enzyme-inhibitor complexes, we have used the AMBER force field¹⁸ and ESP charges¹⁹ to further investigate any possible force field dependencies in the reported correlation and have attempted to derive a more accurate QSAR model by introducing a number of successive refinements which can be summarized as follows:

(i) replacement in the linear regression model of the electrostatic energies given by the Coulombic expression

Table 1. HIV-1 Protease Inhibitors Included in both the Training Set (1–34) and the Prediction Set (35–50)⁶

No	Chemical Structure	$\Delta H_{\text{Lamber}}^a$	ΔG_{ele}^b	ΔG_{ele}^c	$\Delta G_{\text{desolv}}^L^d$	$\Delta G_{\text{desolv}}^R^e$	$\Delta G_{\text{solv}}^L^f$	pIC_{50}^g	pIC_{50}^h
1		-80.56	10.75	10.87	5.69	16.96	-10.13	9.60	8.86
3		-76.15	12.02	11.95	5.24	17.33	-9.26	8.11	7.43
4		-84.12	12.95	13.28	6.19	18.60	-11.52	9.72	9.72
5		-82.76	11.71	11.93	5.81	17.64	-10.56	9.59	9.68
6		-82.74	13.86	14.04	6.93	18.22	-11.90	9.64	9.83
7		-79.56	12.55	12.81	6.36	18.50	-10.46	9.22	9.20
8		-81.92	10.94	11.06	5.70	17.12	-9.98	9.54	9.21
9		-81.36	11.31	11.41	6.04	17.14	-12.57	9.51	9.21
10		-84.51	11.23	11.37	5.80	17.48	-11.97	9.57	9.98
11		-67.78	11.28	11.41	5.55	17.19	-9.73	5.53	6.08
12		-81.53	10.86	11.07	6.37	16.65	-11.77	9.80	9.25
13		-74.17	10.22	10.31	5.24	16.92	-9.25	7.56	7.18
14		-83.14	12.06	12.28	5.87	17.75	-10.37	9.14	9.68
15		-81.17	11.32	11.32	5.70	17.41	-10.20	8.27	8.84

Table 1 (Continued)

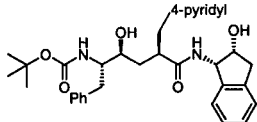
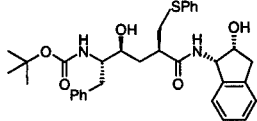
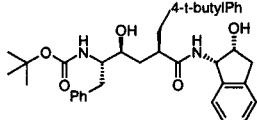
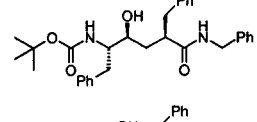
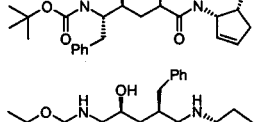
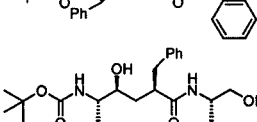
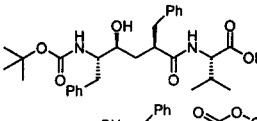
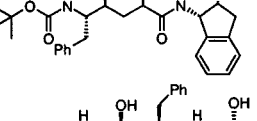
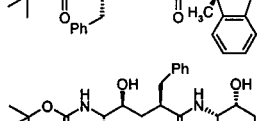
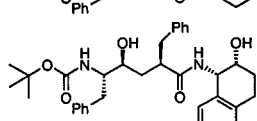
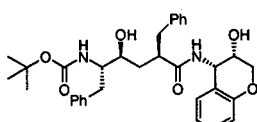
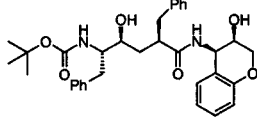
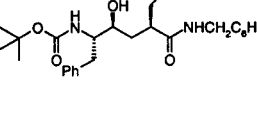

No	Chemical Structure	$\Delta H_{\text{Lamber}}^a$	ΔG_{ele}^b	ΔG_{ele}^c	$\Delta G_{\text{desolv}}^L^d$	$\Delta G_{\text{desolv}}^R^e$	$\Delta G_{\text{solv}}^L^f$	pIC_{50}^g	pIC_{50}^h
16		-81.85	12.74	12.86	6.25	17.39	-11.26	9.28	8.88
17		-80.40	11.14	11.11	5.79	17.52	-10.34	9.60	9.03
18		-85.76	11.42	11.56	5.83	17.53	-10.02	9.77	10.14
19		-73.56	11.69	11.83	5.52	16.62	-9.90	6.94	6.88
20		-75.20	12.52	12.73	5.77	17.00	-10.03	8.02	7.39
21		-77.68	12.21	12.28	5.55	17.09	-9.79	7.47	8.17
22		-70.79	11.59	11.70	5.32	18.00	-10.08	6.16	5.97
23		-69.82	11.93	11.93	4.93	16.96	-9.39	6.79	6.37
24		-75.61	12.55	12.79	5.59	18.92	-10.30	7.18	7.30
25		-78.84	12.64	12.85	6.42	17.17	-10.86	6.67	7.31
26		-74.83	11.90	11.94	5.21	17.31	-9.14	6.91	7.16
27		-81.09	13.58	13.74	6.97	18.61	-11.51	9.16	9.28
28		-82.53	11.80	11.97	7.13	17.49	-12.36	9.75	9.84
29		-76.09	16.41	16.39	6.54	18.84	-11.32	7.39	7.82
30		-72.19	13.75	13.90	4.95	18.13	-8.51	4.52	-i

Table 1 (Continued)

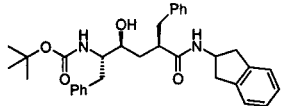
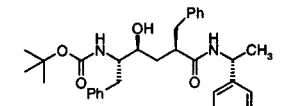
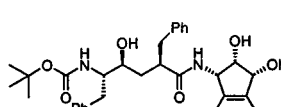
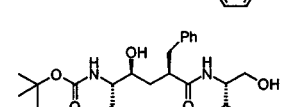
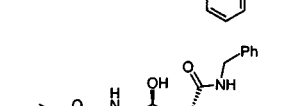
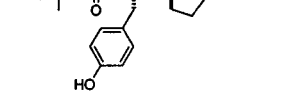
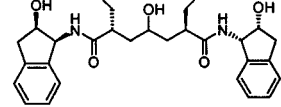
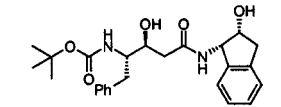
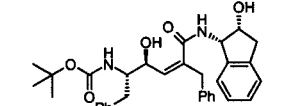
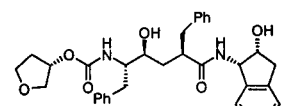
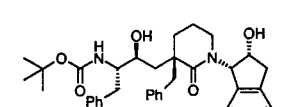
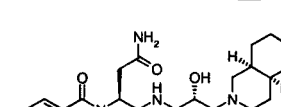
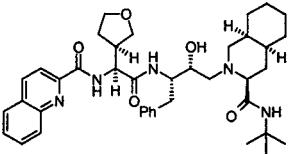
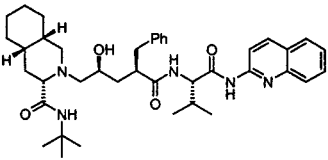
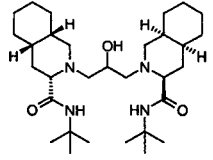
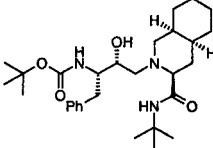
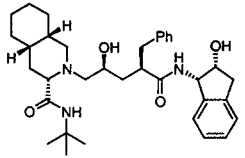
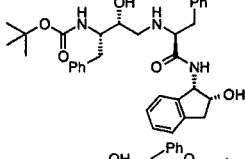
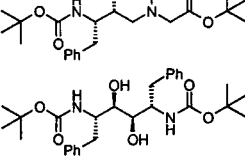
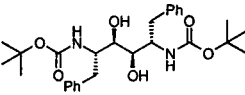
No	Chemical Structure	$\Delta H_{\text{Lamber}}^a$	ΔG_{ele}^b	ΔG_{ele}^c	$\Delta G_{\text{desolv}}^L^d$	$\Delta G_{\text{desolv}}^R^e$	$\Delta G_{\text{solv}}^L^f$	pIC_{50}^g	pIC_{50}^h
31		-76.80	11.88	12.06	5.53	17.09	-9.78	6.89	7.01
32		-75.58	13.20	13.24	5.62	17.45	-9.62	6.84	7.24
33		-82.20	11.81	12.03	6.33	17.57	-10.95	10.00	9.45
34		-74.16	11.62	11.62	5.95	16.62	-10.56	7.41	7.64
35		-65.12	11.81	12.10	6.22	15.81	-11.31	6.23	5.06
36		-88.28	11.14	11.07	6.56	18.34	-11.66	9.16	9.29
37		-61.83	8.40	8.53	6.27	15.77	-10.86	6.25	6.23
38		-79.81	12.13	12.36	6.01	16.87	-10.65	8.89	8.15
39		-83.26	10.58	10.84	6.89	16.90	-11.86	10.22	9.31
40		-66.18	12.40	12.35	6.38	16.61	-11.57	5.90	5.19
41		-86.00	14.63	14.65	7.69	17.94	-16.79	9.64	8.66
42		-81.48	11.13	11.16	6.54	17.01	-13.08	8.27	7.83

Table 1 (Continued)

No	Chemical Structure	$\Delta H_{\text{Lamber}}^a$	ΔG_{ele}^b	ΔG_{ele}^c	$\Delta G_{\text{desolv}}^L^d$	$\Delta G_{\text{desolv}}^R^e$	$\Delta G_{\text{sol}}^L^f$	pIC_{50}^g	pIC_{50}^h
43		-91.73	13.83	14.03	6.16	18.20	-12.74	10.27	9.91
44		-80.34	14.35	14.31	5.35	18.88	-10.59	7.28	7.33
45		-73.94	11.03	11.05	3.06	16.00	-5.02	5.17	5.79
46		-70.77	10.49	10.49	4.95	15.25	-7.78	5.52	5.76
47		-80.71	11.57	11.45	5.43	17.60	-9.40	8.12	7.88
48		-72.88	14.16	14.45	7.91	17.05	-13.86	6.64	7.32
49		-68.08	10.81	10.90	4.98	16.90	-9.04	5.33	5.06
50		-66.90	12.82	13.02	6.64	16.38	-10.99	5.86	6.10

^a Total interaction energy calculated with AMBER; ^b electrostatic binding free energy calculated by means of equation 1; ^c electrostatic binding free energy calculated by means of equation 5; ^d electrostatic contribution to the free energy of desolvation of the ligand upon complex formation; ^e electrostatic contribution to the free energy of desolvation of the receptor upon complex formation; ^f electrostatic contribution to the free energy of solvation of the ligand; ^g experimental; ^h predicted with model C_{delphi} ; ⁱ compound excluded from the analysis.

in AMBER with those derived by solving the Poisson–Boltzmann equation according to classical continuum electrostatic theory.

(ii) decomposition of the calculated intermolecular interaction energies into a data matrix of *per residue* van der Waals and electrostatic contributions to be used as regressor variables in a PLS analysis following the COMBINE strategy.

(iii) implementation of a new formalism that allows the solvent-corrected electrostatic interaction energy between ligand and receptor to be described in terms of individual residue contributions plus two additional regressor variables representing the cost in electrostatic energy of desolvating the ligand and the receptor binding site upon complex formation.

Methodology

(a) Model Building of the Complexes and Parametrization of the Inhibitors. The training set of inhibitors contained a hydroxyethylene moiety as an isosteric replacement of the tetrahedral transition state of the hydrolyzing amide bond of the substrate polypeptides, whereas the set of candidate inhibitors (“prediction set”) contained hydroxyethylene, hydroxyethylamine, statine, and symmetrical diol isosteres. Both sets of compounds included modifications in the P_1' and P_2' substituents, that is, those that interact with binding pockets S_1' and S_2' on one side of the enzyme (Table 1). Coordinates for molecules **1**, **3–34** in their complexes with the enzyme have been previously reported⁶ and

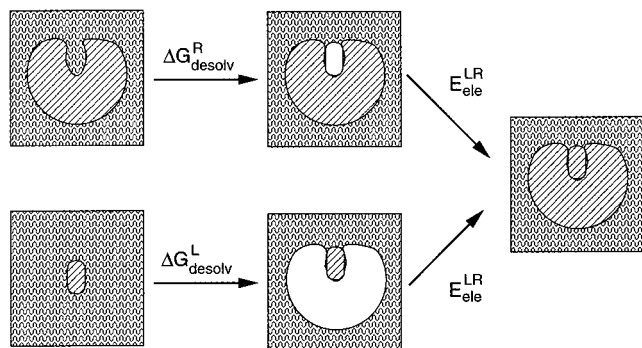


Figure 1. Schematic representation for calculating the electrostatic contribution to the free energy of binding (ΔG_{ele}). The wavy pattern represents the high dielectric solvent ($\epsilon_s = 80$) that surrounds the low dielectric solutes ($\epsilon_m = 4$). Hatched or blank areas in the solutes denote the presence or absence, respectively, of atomic charges in the calculation. ΔG_{desolv} stands for the loss of electrostatic interaction between the solvent and either the receptor ($\Delta G_{\text{desolv}}^R$) or the ligand ($\Delta G_{\text{desolv}}^L$) whereas $E_{\text{ele}}^{\text{LR}}$ represents the electrostatic interaction between ligand and receptor in the presence of surrounding solvent.

were kindly provided to us by the Merck researchers. They were modeled on the crystallographic structure of the complex of HIV-1 protease with L-689,502 solved at 2.25 Å resolution.²⁰ All the complexes included the water molecule that stabilizes the closed conformation of the dimeric enzyme by bridging a β -hairpin from each monomer to the inhibitors (WAT 199).²¹ In agreement with the experimental evidence,²² the carboxyl of AspA25 was protonated whereas that of AspB25 was not. Molecules **35–50** were model-built in Insight using standard bond lengths and angles based on the structure of similar inhibitors, and they were manually docked into the active site using the interactive graphics program Insight II.²³ Atom-centered charges for all the inhibitors were derived by fitting the molecular electrostatic potential calculated²⁴ with the AM1 Hamiltonian²⁵ to a monopole–monopole expression.¹⁹ The same sets of charges and radii were used in programs AMBER²⁶ and DelPhi,²⁷ described below. Covalent and nonbonded parameters for the inhibitors were derived, by analogy or through interpolation,²⁸ from those already present in the AMBER database.¹⁸

(b) Molecular Mechanics Calculations. All-atom AMBER force field parameters¹⁸ were used for the inhibitors, the water molecule, and the protein. For strict comparison of our results with those from the Merck study, the protein atoms should have been kept rigid, as they were in the original work.⁶ On the other hand, we felt that the complexes refined with the MM2X force field should be allowed to adapt to the AMBER force field, especially in view of the different sets of van der Waals radii used in the two force fields. A compromise between these two alternatives was reached by performing a mild and progressive refinement of the docked conformation of each inhibitor in the bound complex. In a first stage, only the hydrogen atoms of each complex were allowed to reorient. Then the geometry of the inhibitor was optimized while the atoms of both the protein and the water molecule were held fixed. Finally, the whole complex was energy minimized but the protein atoms were restrained to their crystallographic positions by means of a harmonic potential

with a force constant of 32 kcal mol⁻¹ Å⁻². Each of the energy minimizations proceeded using the steepest descent algorithm for 100 steps and then switching to the conjugate gradient minimizer until the root-mean-square value of the potential energy gradient was below 0.0001 kcal mol⁻¹ Å⁻¹. A cutoff of 10.0 Å and a distance-dependent dielectric constant ($\epsilon = r_{ij}$) were used throughout.

(c) Continuum Electrostatics Calculations. It is not trivial to include the electrostatic free energy of binding calculated using continuum methods in a way appropriate for use in COMBINE analysis. If ligand–receptor interactions, calculated in the presence of solvent, are decomposed into residue-based contributions, and each pair of interactions is considered independently, the approximations introduced can be considerable. This is due to the fact that the electrostatic energy, when expressed as a function of the electric field, does not obey the superposition principle, the reason being that the electrostatic energy is quadratic in the fields. In linear dielectric media, the total work (W), i.e. the electrostatic energy, of assembling a set of charges is

$$W = \frac{1}{2} \int \mathbf{D} \cdot \mathbf{E} \, d\tau = \frac{1}{2} \int \epsilon E^2 \, d\tau$$

where \mathbf{D} is the electrostatic displacement ($\mathbf{D} = \epsilon \mathbf{E}$ in linear dielectric media), and E is the electric field produced by this set of charges. If the total set of charges is divided into two subsystems, 1 and 2, and the electrostatic energy of each subsystem in the dielectric medium is calculated independently (W_1 and W_2), we will find that $W \neq W_1 + W_2$ because $E_2 = (E_1 + E_2)^2 \neq E_1^2 + E_2^2$. Thus, the sum of the reaction fields created by each residue is not the same as the reaction field created by all residues when taken together, i.e., there are cross-terms that must be taken into account when a partitioning scheme is devised.

Following the classical approach, the overall electrostatic free energy change upon binding (ΔG_{ele}) can be calculated from the total electrostatic energy of the system by running three consecutive calculations on the same grid:²⁹ one for all the atoms in the complex ($G_{\text{ele}}^{\text{LR}}$), one for the ligand atoms alone ($G_{\text{ele}}^{\text{L}}$), and a third one for the receptor atoms alone ($G_{\text{ele}}^{\text{R}}$). Since the grid definition is the same in the three calculations, the artifactual grid energy cancels out when the electrostatic contribution to the binding free energy is expressed as the difference in energy between the product and the reactants:

$$\Delta G_{\text{ele}} = G_{\text{ele}}^{\text{LR}} - (G_{\text{ele}}^{\text{L}} + G_{\text{ele}}^{\text{R}}) \quad (1)$$

This method, however, is unsuitable for our purposes because it does not allow partitioning at the residue level. An alternative is to consider a different description of the binding process (Figure 1) consisting of first desolvating the apposing surfaces of both ligand and receptor and then letting the charges of the two molecules interact. It is then possible to separate the change in electrostatic free energy on molecular association (ΔG_{ele}) into three components:^{29,30} (i) the ligand–receptor interaction energy in the presence of the surrounding solvent ($E_{\text{ele}}^{\text{LR}}$), (ii) the change in solvation

energy of the ligand upon binding ($\Delta G_{\text{desolv}}^{\text{L}}$), and (iii) the change in solvation energy of the receptor upon binding ($\Delta G_{\text{desolv}}^{\text{R}}$):

$$\Delta G_{\text{ele}} = E_{\text{ele}}^{\text{LR}} + (\Delta G_{\text{desolv}}^{\text{L}} + \Delta G_{\text{desolv}}^{\text{R}}) \quad (2)$$

This decomposition is exact, contains all cross-terms, and can be profitably used in COMBINE analysis. The first term in eq 2, that is, the electrostatic energy of interaction between the group of R atoms in the receptor and the group of L atoms in the ligand, can be described (in kcal mol⁻¹) by

$$E_{\text{ele}}^{\text{LR}} = \sum_{i=1}^R q_i \phi_i \quad (3)$$

where q represents the atomic point charges of the receptor, and ϕ_i is the electrostatic potential at each of the R atoms of the receptor created by the L atoms of the ligand. Equation 3 can be recasted in the form of a sum of N residue-based contributions (e_n), where N is the number of residues in the receptor, each of these containing K atoms:

$$E_{\text{ele}}^{\text{LR}} = \sum_{n=1}^N \sum_{k=1}^K q_{nk} \phi_{nk} = \sum_{n=1}^N e_n \quad (4)$$

This allows the total electrostatic free energy of ligand binding to be expressed as a sum of residue-based contributions plus two additional terms corresponding to the electrostatic component of the desolvation free energy of both the ligand and the receptor:

$$\Delta G_{\text{ele}} = \sum_{n=1}^N e_n + \Delta G_{\text{desolv}}^{\text{L}} + \Delta G_{\text{desolv}}^{\text{R}} \quad (5)$$

The electrostatic potentials used in eqs 3 and 4 can be calculated either as

$$\phi_i = 332 \sum_{j=1}^L \frac{q_j}{\epsilon r_{ij}} \quad (6)$$

where ϵ is the relative permittivity of the homogeneous dielectric medium and r_{ij} is the separation between every pair of atoms, or to include the potential created by the response of the surrounding solvent to the ligand charges, by solving the linear form of the Poisson–Boltzmann equation

$$\nabla \cdot [\epsilon(\vec{r}) \nabla \phi(\vec{r})] = -4\pi\rho(\vec{r}) + \bar{\kappa}^2 \phi(\vec{r}) \quad (7)$$

where ρ is the fixed solute charge distribution, $\bar{\kappa}$ is the modified Debye–Hückel constant that accounts for a Boltzmann distribution of the ions in solution, and $\epsilon(\vec{r})$ and $\phi(\vec{r})$ are the dielectric constant and the electrostatic potential, respectively, as a function of position. The solvent-corrected potential calculated by means of eq 7 can be either that generated by the charges on the receptor at the positions of the uncharged ligand atoms or, alternatively, that created by the charges on the ligand at the location of each of the uncharged atoms of the receptor (Figure 1). It is the latter that we have computed for the purpose of calculating the residue-based contributions to $E_{\text{ele}}^{\text{LR}}$ described in eq 4. Thus, the

only effect that is missed by this approach is the solvent polarization created by the receptor charges and its corresponding cross-terms. However, the consequence of ignoring this effect is almost negligible when the complexes of a common receptor with a series of congeneric ligands are considered,¹⁰ as is demonstrated below by the similar $E_{\text{ele}}^{\text{LR}}$ values computed by means of eqs 1 and 5.

With regard to the latter two terms of eqs 2 and 5, corresponding to the differences in electrostatic free energies of solvation of ligand and receptor upon complex formation, they were calculated by considering the effects on the respective electrostatic free energies of replacing the high dielectric medium of the solvent with the low dielectric medium of the other molecule in those regions that are occupied by the binding partner in the complex (Figure 1).

Each of the components of eq 5, which describes the electrostatic effects of ligand binding, enters the energy matrix for COMBINE analysis as a different variable. From the previous derivation, and as demonstrated below in the Results section, it is clear that the total electrostatic binding free energy is partitioned taking into account the cross-terms in the reaction field and no double counting is done, i.e. the sum of electrostatic terms in the energy matrix yields the electrostatic free energy of binding for that particular compound. This is an important feature that allows an easier interpretation of the regression models, avoiding convoluted effects in the energetic description of the variables. It is also worth noting that the formalism presented creates an electrostatic block of similar variance to the van der Waals block, making the direct use of PLS analysis possible without invocation of scaling procedures, which may produce spurious results in 3D QSAR.³¹

The Poisson–Boltzmann equation was solved using a finite difference method, as implemented in the DelPhi²⁷ module of Insight II. The atomic coordinates employed were those of the AMBER-optimized complexes. The interior of the protein, the ligands, and the complexes was considered a low dielectric medium ($\epsilon = 4$) whereas the surrounding solvent was treated as a high dielectric medium ($\epsilon = 80$) with ionic strength of 0.145 M. Cubic grids with a resolution of 0.5 Å were centered on the molecular systems considered, and the charges were distributed onto the grid points.²⁹ Solvent-accessible surfaces,³² calculated with a spherical probe of 1.4 Å radius,³³ defined the solute boundaries, and a minimum separation of 11 Å was left between any solute atom and the borders of the box. The potentials at the grid points delimiting the box were calculated analytically by treating each charge atom as a Debye–Hückel sphere.²⁹ Further definition of a finer grid (0.25 Å spacing) and use of new boundary potentials linearly interpolated from those calculated in the previous run (focusing)³⁴ were not justified as test calculations showed that the electrostatic energies did not change significantly.

For comparison purposes, the electrostatic contribution to the solvation free energy of each inhibitor ($\Delta G_{\text{ele}}^{\text{solv}}$) was also calculated. It was obtained by subtracting the total electrostatic free energy in vacuo ($G_{\text{ele}}^{\text{vac}}$) from the total electrostatic free energy in water

($G_{\text{ele}}^{\text{wat}}$).²⁹ To this end, two calculations with identical grid mappings and the same interior dielectric ($\epsilon = 4$) were run for each inhibitor. The exterior dielectric was set to either 1, to reproduce vacuum conditions, or to 80, to simulate the aqueous environment.

(d) Partitioning of the Intermolecular Interaction Energy and Pretreatment of the Resulting Energy Matrix. The calculated ligand–receptor interaction energies in the refined complexes were partitioned on a per residue basis using the ANAL module in AMBER. A dielectric constant of 4 was used, and all atom pairs were included in the calculation. Each inhibitor was regarded as a single fragment and no intramolecular energy terms were considered. Since there are 2 protein subunits of 99 amino acids each plus 1 water molecule, and 2 energy contributions (van der Waals and electrostatic) are considered for each residue, 398 variables were used to characterize each complex. These energy descriptors were used as input for the GOLPE program.³⁵

Any positive energy values were truncated to 0.0 kcal mol⁻¹ as they were thought to arise from differences in force field parameters between MM2X and AMBER. Noise was reduced by zeroing those interaction energies with absolute values lower than 0.1 kcal mol⁻¹ and by removing any variables with a standard deviation below 0.1 kcal mol⁻¹. This pretreatment reduced the number of variables that entered the PLS analysis from 398 to around 50, making variable selection unnecessary. No further scaling was carried out. Inspection of the activity data, on the other hand, revealed that inclusion of compound **30** in the analysis was not warranted as this is a molecule with very low activity ($\text{IC}_{50} = \sim 30 \mu\text{M}$) and high leverage.

(e) Partial Least Squares (PLS) Analysis. The PLS analysis of the pretreated data was carried out as implemented in the GOLPE 3.0 program.³⁶ The models were validated using both internal and external validation tests. For internal validation, cross-validation was used, assigning the compounds randomly to one of five groups of approximately the same size, excluding from the analysis each group in turn, and repeating the whole procedure 20 times. This cross-validation procedure provides a safer alternative to the more widely preferred leave-one-out method and gives more conservative results: a smaller cross-validated correlation coefficient (q^2) and a higher cross-validated standard deviation error of predictions (SDEP_{cv}).³⁷ For external validation, the PLS models obtained were used to predict the biological activity of 16 compounds (prediction set) not included in the training set. The quality of the predictions was assessed by the value of the external SDEP (SDEP_{ex}).

The optimal dimensionality of each model was determined from the evolution of the cross-validation indexes as a function of the number of latent variables extracted. All of the reported models contain two latent variables (LV); further increase of the number of model dimensions was not justified as it did not lead to any improvement or the improvement achieved was not significant.

Results

(a) Correlation of Inhibitory Potencies with Calculated Binding Energies. The quality of the

correlation between the intermolecular interaction energies calculated with AMBER and the biological activities of these inhibitors ($r^2 = 0.81$) was found to be slightly better than that reported for the same series using the MM2X force field ($r^2 = 0.74$; $r^2 = 0.78$ if compound **30** is not excluded).⁶ In cross-validation, the predictive indices of our equivalent model (L_{amber}) were likewise comparable ($q^2 = 0.79$, SDEP_{cv} = 0.61 vs $q^2 = 0.75$), which is indicative that the good correlation first reported by the Merck researchers⁶ is not exclusive to the MM2X force field, at least according to these criteria. When the regression equation obtained for the training set was used to predict the activity of the external set, the SDEP_{ex} was 1.08 log units. Figure 2a shows that the model behaves remarkably well for the most active compounds but predicts the less active molecules poorly, some with an error in pIC₅₀ of about 2 log units (e.g. **37**, observed 6.25, predicted 3.92; **45**, observed 5.17, predicted 7.12).

(b) Calculation of Electrostatic Energies with DelPhi and Replacement of the Coulombic Term.

A measure of the uncertainties introduced in the continuum calculations when mapping the charges and the dielectric constants onto a grid of finite size was obtained by comparing the AMBER electrostatic interaction energies ($\epsilon = 4$) with those calculated by DelPhi using a uniform dielectric of 4 for both the complex interior and the surrounding medium. The average difference amounted to about 0.2 kcal mol⁻¹, and this was also the degree of discrepancy in electrostatic interaction energies calculated by means of eqs 1 or 5 (Table 1). These average errors were considered sufficiently low to make us feel comfortable with the conditions used in the continuum electrostatics calculations and reassured us about the correctness of the partitioning scheme employed.

When the electrostatic term calculated with AMBER was replaced with the overall electrostatic energy change upon binding computed with DelPhi (ΔG_{ele}), a new set of global interaction energies was obtained. The quality of the correlation with the inhibitory activities of this new model (L_{delphi}), however, deteriorated slightly (Figure 2b) with respect to the previous model (Table 2). This result is analogous to that reported by the Merck researchers and corroborates their finding that incorporation of solvation effects was not accompanied by any improvement of the linear regression model.⁶ This lack of improvement may be somewhat surprising at first sight as the DelPhi calculation now includes desolvation effects and takes into account the reaction field of the surrounding solvent around the complex. Nevertheless, it should be borne in mind that continuum calculations are very sensitive to small changes in atomic coordinates, and for this reason the information contained in the ΔG_{ele} values is probably not devoid of “noise”.

(c) Breakdown of the Ligand–Receptor Interaction Energies and COMBINE Analysis of the Energy Matrix. The partitioning scheme used in COMBINE analysis allows the dissection of the ligand–receptor interaction energies into many individual van der Waals and electrostatic contributions and provides a wealth of information that can be easily comprehended

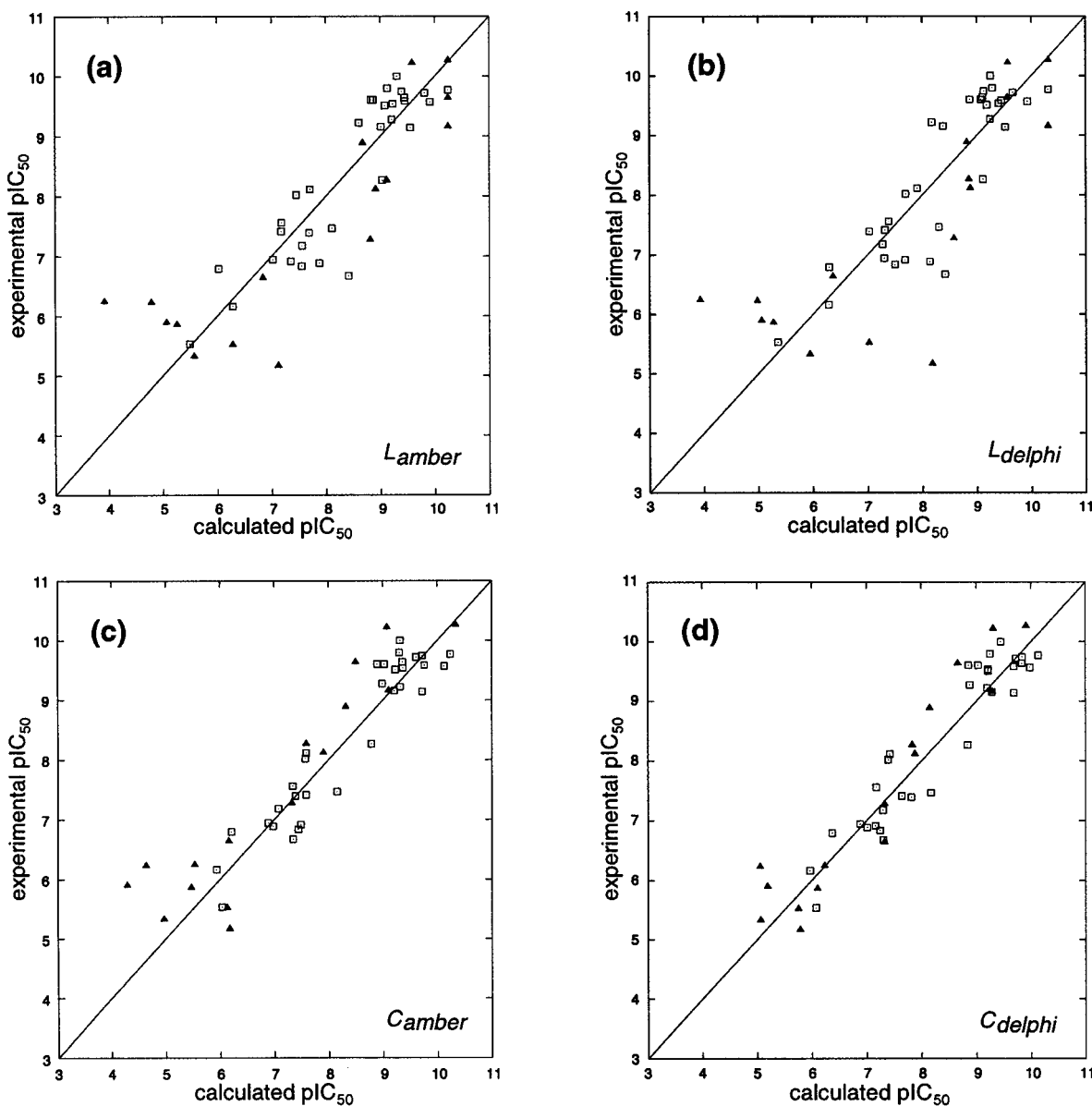


Figure 2. Correlations between experimental and calculated activities (pIC_{50} values) according to the different regression models for compounds in the training set (open squares) and the prediction set (filled triangles). (a) linear regression model (AMBER van der Waals and electrostatic: L_{amber}), (b) linear regression model (AMBER van der Waals and DelPhi electrostatic: L_{delphi}), (c) COMBINE model (AMBER van der Waals and electrostatic: C_{amber}), (d) COMBINE model (AMBER van der Waals and DelPhi electrostatic: C_{delphi}). See Methodology for details.

Table 2. Comparison of the Different Regression Models

model	no. of objects	no. of variables	no. of LVs	r^2	q^2	SDEP _{cv}	SDEP _{ex}
L_{amber}	32	1	1	0.81	0.79	0.61	1.08
L_{delphi}	32	1	1	0.77	0.75	0.65	1.23
C_{amber}	32	48	2	0.89	0.70	0.72	0.83
C_{delphi}	32	47	2	0.90	0.73	0.69	0.59
C_{expanded}	48	54	2	0.91	0.81	0.66	

from a simple plot. Thus, the visual inspection of the energy values emanating from each protein residue and the water molecule (Figure 3) highlights not only the amino acids whose interaction with the ligand varies more from complex to complex but also those parts of the enzyme that give rise to unfavorable interaction energies. The presence of these disturbing positive energy values passed unnoticed in the previous analyses that considered just a global interaction energy term (L_{amber} and L_{delphi}), and is most likely a consequence of

differences in the van der Waals radii between the two force fields used for modeling and refining the complexes (MM2X and AMBER). The suggestion that the rather mild energy minimization performed using AMBER on the MM2X geometries has not been able to refine the complexes completely is supported by other work from our laboratory on a different set of fully optimized enzyme–inhibitor complexes for which no positive energy values were observed.³⁸ Further refinement of the HIV-1 protease–inhibitor complexes with AMBER was ruled out because the structural changes this procedure was likely to introduce would have precluded direct comparison of our results with the results of the Merck study (the average rmsd for all protein non-hydrogen atoms between initial and refined complexes was 0.06 Å). Instead, and in order to avoid the undesirable effect of all those variables, a cutoff of 0.0 kcal mol⁻¹ was applied (see Methodology).

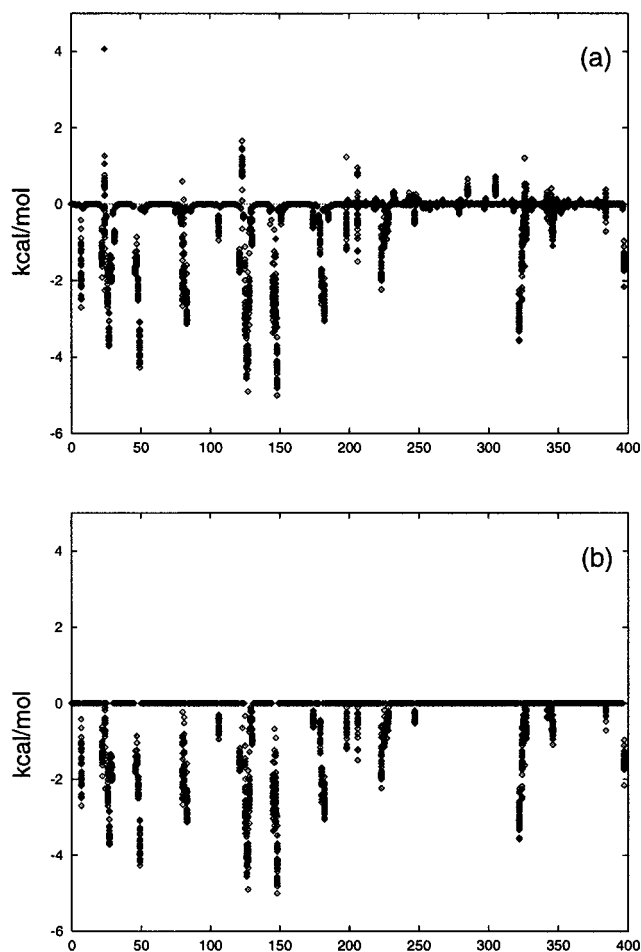


Figure 3. Interaction energies (a) before and (b) after pretreatment. These are the variables that enter the PLS analysis leading to the C_{amber} model. Each diamond represents a value present in the energy matrix. On the horizontal axis, the variables are ordered sequentially and correspond to van der Waals (1–199) and electrostatic (200–398) interactions between the ligands and either the protein residues (1–198 and 200–397) or the water molecule (199 and 398). See Methodology for details.

The residue-based van der Waals and electrostatic interactions calculated in AMBER, pretreated as described in the Methodology section, were analyzed using PLS. The experimental data fit to the resulting model (C_{amber}) better than to the linear regression model (L_{amber}), as manifested by the higher correlation coefficient ($r^2 = 0.89$) and, graphically, by the plot of calculated vs experimental activities (Figure 2c). The COMBINE model clearly outperforms the previous model in fitting some of the former outliers and the overall fit is slightly improved. The predictive ability of the model is also increased, as assessed by the results of the external validation ($\text{SDEP}_{\text{ex}} = 0.83$). Figure 2c clearly shows that the values of activity predicted for the compounds in the external set are much closer to experiment than those predicted by the previous model. However, it is also clear from comparison of the SDEP values that the standard error in prediction for the compounds in the external set is somewhat larger than that for the molecules belonging to the training set. This can be taken as evidence that there is still room for improvement in the model. The C_{amber} model still predicts the less active compounds poorly, and some

compounds are underpredicted by more than 1 log unit (e.g. **41**, observed 9.64, predicted 8.50; **40**, observed 5.90, predicted 4.29). Moreover, there appears to be a general tendency to underpredict the activity of the compounds in the external dataset: the predicted minus observed residual is negative for 12 out of the 16 products, with an average value of -0.45 .

(d) Replacement of the Coulombic Term in the Interaction Energy Matrix. The residue-based electrostatic interaction energies computed with DelPhi, as depicted in eq 4, were used to replace the corresponding AMBER values in the COMBINE energy matrix. Besides, to account for the change in the electrostatic energy of solvation of the ligand and the enzyme binding site upon complex formation (Figure 1), two new variables, $\Delta G_{\text{desolv}}^{\text{R}}$ and $\Delta G_{\text{desolv}}^{\text{L}}$, were incorporated into the analysis as additional regressors. Pretreatment of this new data matrix as in the previous example reduced the initial 400 variables per complex to only 47, which were analyzed by PLS as before. The new COMBINE model (C_{delphi}) neatly outperforms the previous C_{amber} model (Table 2). The quality of the fit is marginally better ($r^2 = 0.90$) than in the case of the C_{amber} model but, more importantly, the predictive ability is significantly improved, as assessed by a higher cross-validated correlation coefficient in internal validation ($q^2 = 0.73$) and a smaller error in prediction ($\text{SDEP}_{\text{cv}} = 0.69$). The external predictions are even better, and for the first time the external SDEP falls below the cross-validated SDEP, reaching the remarkable value of $\text{SDEP}_{\text{ex}} = 0.59$. Comparison between calculated and experimental activities (Figure 2d) clearly shows that the accuracy of this model is greatly improved, especially for the weaker inhibitors. It can be seen that the residuals of the predictions are distributed more evenly although there persists a certain tendency to underestimate the potency of the compounds in the external set (the average predicted-minus-observed residual is -0.24). The most significant outlier is compound **41** but its activity is underpredicted by less than 1 order of magnitude (Table 1).

To obtain further insight into the correlation, the PLS model C_{delphi} was transformed into a pseudo-MLR (multiple linear regression) model of the form

$$\text{pIC}_{50} = \sum_i a_i X_i + b \quad (8)$$

where each pseudo-coefficient a_i summarizes the contribution to the activity of variable X_i in each of the two latent variables extracted. Since these pseudocoeficients are used mainly to compare the relative importance for activity of each variable, they are weighted so as to make them independent of the scale and the variance of the original variables. Figures 4 and 5 provide a visual representation of these weighted PLS pseudo-coefficients and graphically highlight the protein residues whose van der Waals and/or electrostatic interactions with the ligands are more important for describing the variation in activity within the series. This should allow the design of further structural modifications within the series to be guided by focusing synthetic efforts on the regions highlighted by the analysis.³⁹ It must be borne in mind that structural variation within the series is primarily concentrated on the P1' and P2' substituents. Interactions with around 30 amino acids

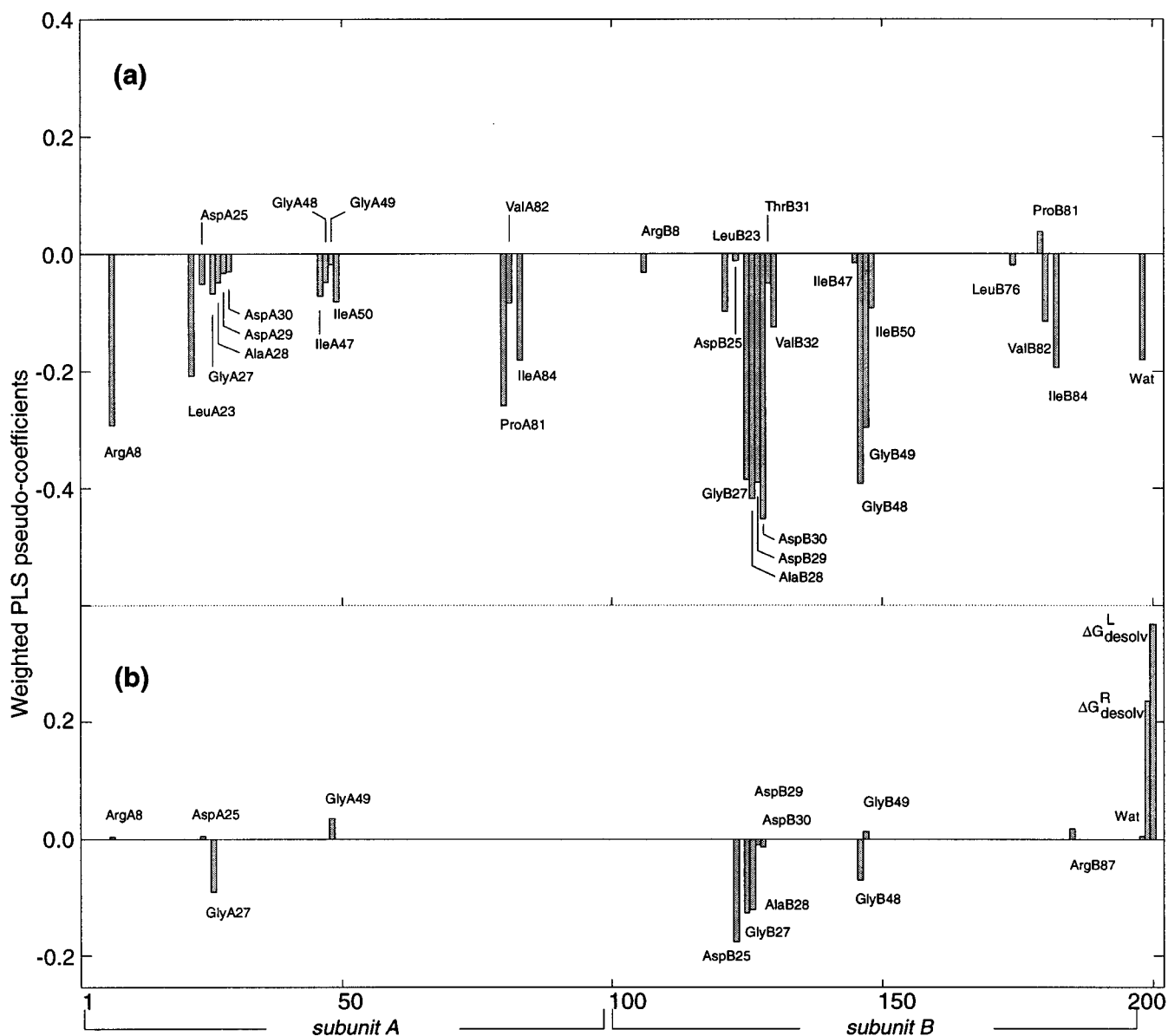


Figure 4. Weighted PLS pseudocoefficients of model C_{delphi} for each of the (a) van der Waals and (b) electrostatic interaction energies studied. The last two variables in (b) represent the electrostatic contributions to the desolvation of the ligand ($\Delta G_{\text{desolv}}^{\text{L}}$) and the receptor ($\Delta G_{\text{desolv}}^{\text{R}}$). Relevant residues have been labeled.

account for the differences in activity, and van der Waals interactions outnumber electrostatic interactions. With regard to the van der Waals coefficients, the more significant fragments selected include the water molecule and a number of amino acids that are clustered around the region containing the unprotonated catalytic aspartic acid (GlyB27, AlaB28, AspB29, AspB30), the flap region (GlyB48, GlyB49), the strand between residues 80–86 (ProA81, ValA82, IleA84), and the “bridge” residues between binding pockets S1’ and S2’ at the interface between subunits A and B (ArgA8, LeuA23). The electrostatic coefficients, on the other hand, are mostly contributed by the unprotonated AspB25, as the most important protein residue, followed by GlyB27, AlaB28, GlyA27, and finally GlyB48; the largest contributions, by far, correspond to the electrostatic energy of desolvation of both ligand and receptor. The hydrogen-bonding potential of the ligands is fulfilled in a very similar way with the only exception of an additional polar group present in some inhibitors which is located

at the boundaries of the S1’ binding pocket. For this reason many electrostatic interactions do not contribute significantly to the differences in activity, and the PLS coefficients arise from few residues. On the other hand, the bulk of the P1’ and P2’ substituents is more efficiently explored (Table 1).

The improvement observed in model C_{delphi} with respect to C_{amber} can be seen as a direct consequence of using a better method for describing the residue-based intermolecular electrostatic energies ($E_{\text{ele}}^{\text{LR}}$) but also as a consequence of the new information added by the two variables that represent the changes in the electrostatic energy of solvation upon complex formation ($\Delta G_{\text{desolv}}^{\text{R}}$ and $\Delta G_{\text{desolv}}^{\text{L}}$). The influence of considering the solvent reaction field in the calculation of $E_{\text{ele}}^{\text{LR}}$ using DelPhi appears to be small since these interaction energy values and those calculated using AMBER and a uniform dielectric of 4 are very similar (data not shown). This probably reflects the fact that the inhibitors are

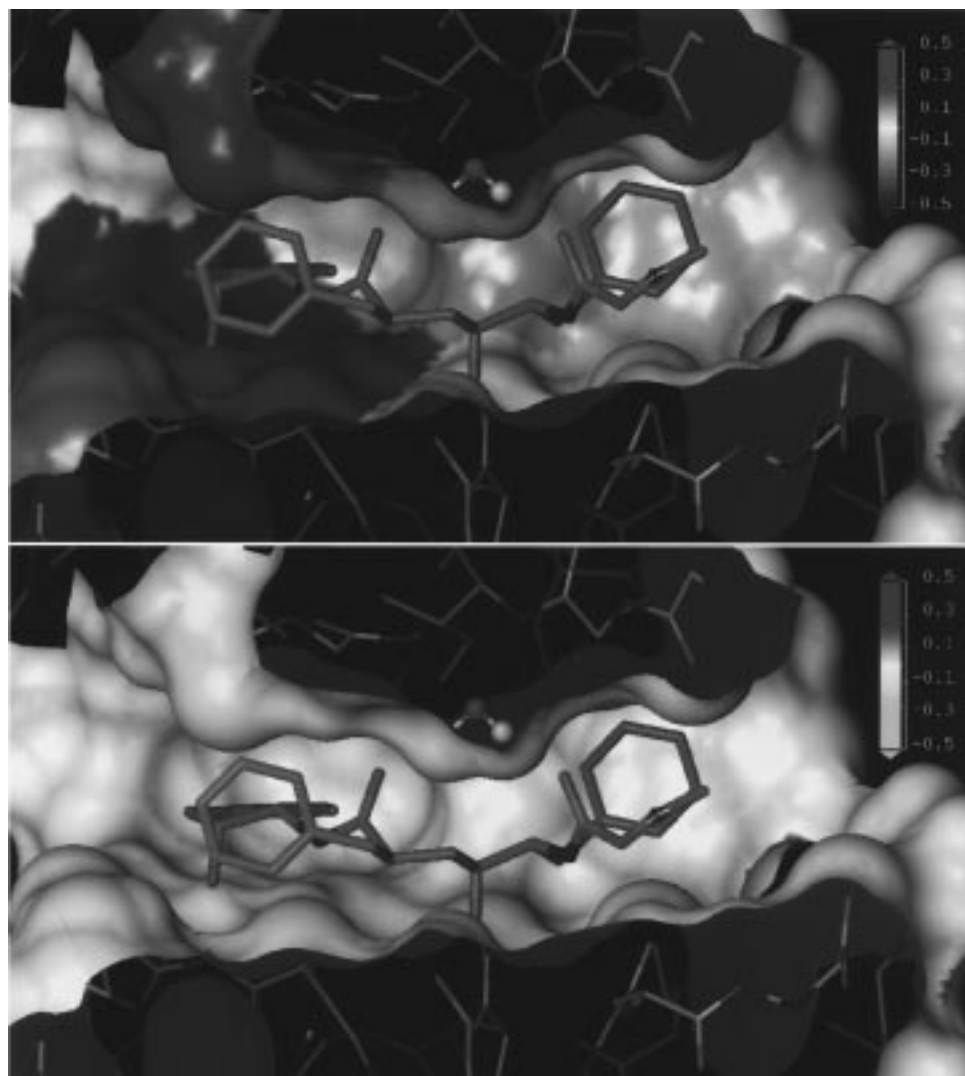


Figure 5. View of a section of the active site of HIV-1 protease in a complex with compound **1** (C atoms in green). The Connolly surface³³ around the protein (C atoms in gray) and water (represented as ball and stick) atoms has been color-coded according to the (top) van der Waals and (bottom) electrostatic weighted PLS pseudocoeficients reported in Figure 4. A scale is provided on the right-hand corner of each picture.

mostly buried in the active site of the protease. To assess the effect of including $\Delta G_{\text{desolv}}^{\text{R}}$ and $\Delta G_{\text{desolv}}^{\text{L}}$ on the quality of the model, we calculated the changes that take place in the coefficients of the C_{delphi} model when these energy descriptors are omitted from the analysis (Figure 6). When the information provided by these two variables is not present, the model does in fact resemble the C_{amber} model (virtually identical r^2 , q^2 , and SDEP_{cv} ; slightly improved $\text{SDEP}_{\text{ex}} = 0.79$). The change in coefficients reported in Figure 6 reveals that the incorporation of these two new variables replaces some information already contained in the previous model. In its search for the best possible correlation with activity, and in the absence of variables that account for solvation effects, the PLS algorithm presumably finds a solution in which the coefficients implicitly include those effects. These coefficients encompass both the effect on the activity of the measured binding enthalpies and some other contributions (such as solvation effects) that are collinear with the calculated energy variables. Such collinearities, even if they could be valid for the training set, are hardly general enough to give accurate predictions for the external

compounds. As a consequence, upon incorporation of the desolvation energies, the model becomes more generally applicable and the quality of the external predictions is improved.

(e) Expanded Data Set. True validation of COMBINE models, as of any other QSAR models, comes when they show their value in predicting accurately the activities of new compounds. We have already shown how C_{delphi} performs remarkably well in this respect. However, elaboration of an existing drug lead is a cyclic process involving repeated rounds of compound design, synthesis and testing.¹ QSAR models derived at one stage need to be continually updated incorporating as much new information as possible. In the present example, once the molecules in the prediction set have been synthesized and evaluated, it makes sense to merge them into the training set to produce an expanded set of inhibitors on which to test the methodology with a view to bringing the model up to date (Figure 7).

With this aim, a new training set encompassing the 48 inhibitors studied was built and analyzed following the same strategy that produced the C_{delphi} model. The

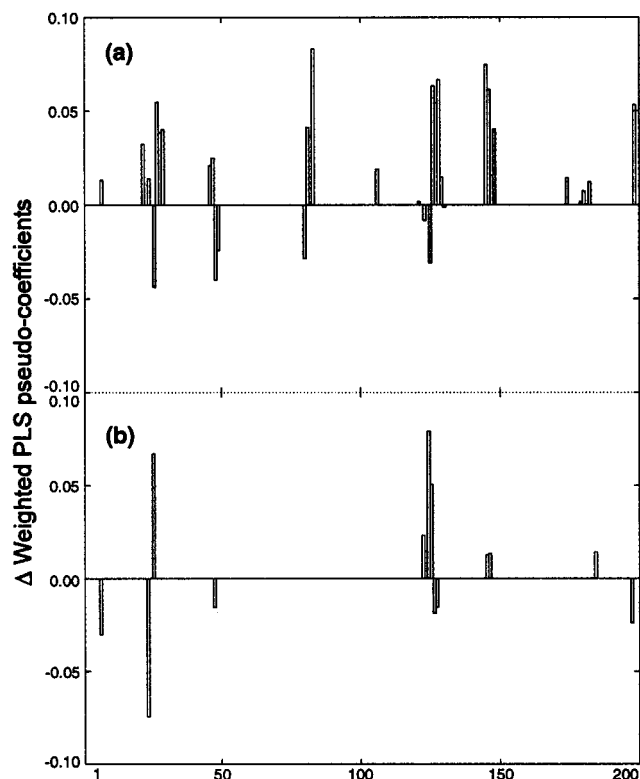


Figure 6. Differential plot showing the changes on the weighted PLS pseudo-coefficients of the (a) van der Waals and (b) remaining electrostatic variables upon removal from model C_{delphi} of $\Delta C_{\text{desolv}}^L$ and $\Delta C_{\text{desolv}}^R$. These values were calculated by subtracting the coefficients provided by a model analogous to C_{delphi} but which does not include the desolvation energy terms from those provided by C_{delphi} . Note that the scale has been changed with respect to that in Figure 4.

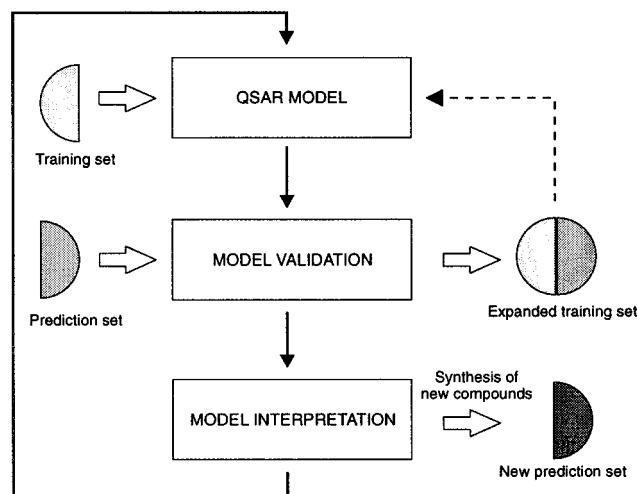


Figure 7. QSAR model building and updating in an iterative cycle of drug design.

resulting COMBINE model (C_{expanded}) proved excellent both in fitting ($r^2 = 0.91$) and in predictive ability ($q^2 = 0.81$, $\text{SDEP}_{\text{cv}} = 0.66$). Figure 8 shows how this updated model provides more accurate estimations of activity for the products in the former prediction set (standard deviation error of calculation for these 16 compounds = 0.42) than any of the previous models (e.g. $\text{SDEP}_{\text{ex}} = 0.59$ in C_{delphi}). This must be seen as a direct consequence of the incorporation into the expanded training set of unique information contained in some of the

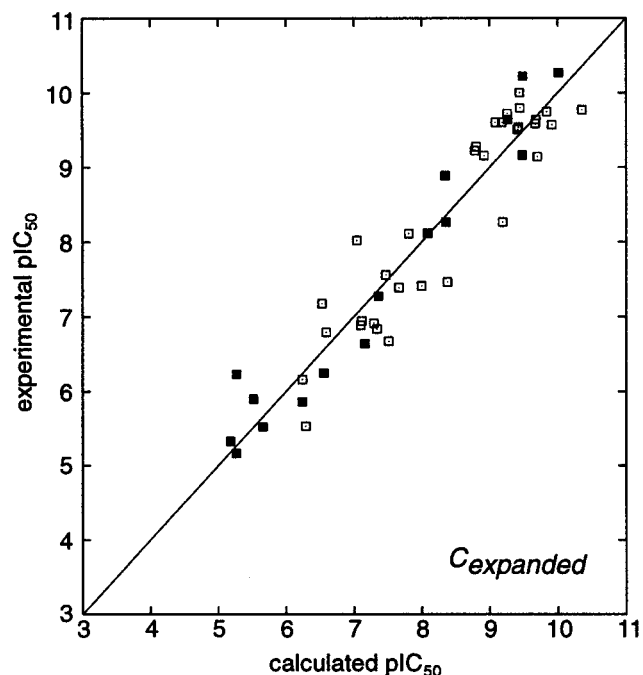


Figure 8. Plot of experimental vs calculated inhibitory activities (pIC_{50} values) for a single expanded set of compounds comprising those belonging to the former training set (open squares) and the former prediction test (filled squares).

compounds of the former prediction set. The lack of this information in the preceding models necessarily limited the quality of the predictions that could be attained.

Discussion

Accurate prediction of ligand–receptor affinities is handicapped by a limited understanding of the relative importance and inter-relationships of the variables which define the 3D QSAR.⁴⁰ For binding to take place, the interaction energy between the binding partners in the complex must overcome their interaction energy with the solvent, but the overall effect of these opposing tendencies on binding affinity is hard to predict: the desolvation penalty can increase by incorporation of polar groups but may also improve binding if good hydrogen bonds or favorable electrostatic interactions are established with the protein target.¹⁴ On the other hand, when congeneric series of ligands are studied, the structural variations on the substituents associated with one or more of the receptor's binding pockets are likely to cooperatively affect the binding of the rest of the molecule. This makes comparisons of ligand components meaningless unless they are restricted to variations in a single position of an otherwise identical structure.⁴⁰ The problem in medicinal chemistry is that there can be many structural changes in a series, and recognizing the relation between the change in structure and the change in activity or binding affinity is difficult due to the many variables involved. The rationale behind a COMBINE analysis is to accurately translate the distinct ligand–receptor interactions in a set of complexes into a large number of informative variables in order to find a correlation between these descriptors and the biological activity. It is hoped that the data matrix can capture the essence of all the structural changes in the compounds studied (training set), and that the resulting quantitative model will highlight

those variables that are more important for improved activity. Once this information is gained, it can be used to advantage in the design of new structural changes. The method of choice for extracting the relevant information is PLS. PLS projects the original matrix of energy terms on to a small number of orthogonal latent variables and gives weights to the original energy contributions according to their importance in the model.

If the ratio between the really informative variables and those that do not contribute to explaining the differences in activity (noisy variables) is too low, the PLS method may fail to obtain a valid model. In a previous example with a set of human phospholipase A₂ inhibitors,¹⁵ intramolecular energy terms were included and the rather large inhibitors were split into a number of fragments. As a result, the number of variables that entered the PLS analysis was very high, and the use of variable selection methods³⁷ was justified. In the present case, our calculations have been limited to a conformational substate of the protein, as found in the crystal structure of HIV-1 protease complexed with L-689,502,²⁰ and a pretreatment of the data consisting of setting very small values to zero and removing those variables that take nearly constant values in the matrix was enough to reduce the number of energy descriptors down to around 50 (Table 2). This means that the variation in activity along the series of HIV-1 protease inhibitors studied can be explained by considering interactions with no more than 30 amino acid residues (Figure 4), which is about the number of unique residues involved in all subsite interactions with a wide range of inhibitors.²¹

The improvement in the accuracy of the external predictions of the COMBINE models relative to the linear regression models (Table 2) is remarkable, all the more if we consider that C_{amber} and C_{delphi} are derived, respectively, using the same data as L_{amber} and L_{delphi} , the only difference being in the tools used for the analysis. Comparison of the cross-validation results, however, should be interpreted with caution as the regression method used to derive each type of model is rather different. COMBINE models contain many more variables and the method itself is more "flexible". As a consequence, when some of the objects are taken out during the cross-validation procedure, the weights given to the different variables change in order to improve the fit, and a slightly poorer estimation is obtained compared to the more "rigid" linear regression models.

Also indisputable is the beneficial effect in a COMBINE model (C_{delphi}) of considering the reaction field originated by the response of the solvent molecules and ions to the field generated by the solute charges, as well as the desolvation of the binding interface upon complex formation. In this respect, it is worth noting that in some published studies the electrostatic contributions to the solvation free energies of the ligands ($\Delta G_{\text{ele}}^{\text{solv}}$) are included in the regression equation that attempts to correlate biological activities with calculated binding energies.^{6,8} This electrostatic free energy represents the change in electrostatic energy in going from vacuum to solvent but the really meaningful quantity that may complement the calculated binding enthalpies is the cost in electrostatic free energy of desolvating both the ligand ($\Delta G_{\text{desolv}}^{\text{L}}$) and the binding site on the receptor

($\Delta G_{\text{desolv}}^{\text{R}}$) (Figure 1). Nevertheless, using data from Table 1, it can be seen that in this particular case the $\Delta G_{\text{ele}}^{\text{solv}}$ values calculated for the inhibitors show a clear correlation ($r^2 = 0.80$) with the $\Delta G_{\text{desolv}}^{\text{L}}$ values, probably due to the fact that most of the variation in polarity occurs on groups that contact the protein and the ligands are almost completely buried within the receptor in the complex. Also noteworthy is the fact that the sign of the weighted PLS pseudo-coefficients assigned by the C_{delphi} model to $\Delta G_{\text{desolv}}^{\text{L}}$ and $\Delta G_{\text{desolv}}^{\text{R}}$ is inverted with respect to that of the majority of the remaining contributions (Figure 4). Since the sign of these terms is positive (Table 1), this result implies that the larger the desolvation energy the better the binding. There is no discrepancy between this finding and the notion that desolvation of both the ligands and the receptor site opposes binding.²⁹ Desolvation and binding are inextricably linked (Figure 1), and the analysis simply highlights the fact that the substituents present in the best inhibitors lead to improved binding but also incur a greater penalty in terms of desolvation energies.

The weights given by the PLS analysis to the original variables used in a COMBINE study are valuable per se because they can be used to highlight those residues whose interactions with the ligand are more correlated with the biological activity in the series studied. These weights can be studied "dimensionwise" for each latent variable considered, or "modelwise", condensed in the weighted PLS pseudocoefficients (Figures 4 and 5). In either case, they can be used similarly to the grid plots of PLS weights or pseudocoefficients provided by a CoMFA study to show the key regions for the ligand-receptor interaction. One obvious advantage of their use is that they pinpoint not only those regions of interest in three-dimensional space but also the actual ligand-receptor interactions involved. Independently of the quality of the model obtained, the simpler linear correlation models lack all this information. Interestingly, the structural mapping of the different contributions to the activity resulting from our analysis is very similar to that reported for the binding energetics of a series of 13 HIV-1 protease inhibitors using a very different approach.⁴¹

The target functions used to estimate the properties related to the binding free energies are very sensitive to inevitable errors in the atomic coordinates.⁴² These inaccuracies are probably exacerbated in modeled complexes and require careful analysis. Unpaired buried polar groups in the protein ligand interface, for example, are strongly adverse to binding but can pass unnoticed when only global intermolecular energies are considered. The COMBINE methodology can be a useful tool to unveil deficiencies such as these. A simple graphical overview of the energy matrix, such as that shown in Figure 3, can easily highlight sources of docking errors.

The lack of improvement in going from the L_{amber} model to L_{delphi} probably reflects the fact that electrostatic free energy contributions calculated using continuum methods are especially sensitive to small changes in atomic coordinates as well as to the way the molecular surfaces are calculated, so that a certain amount of noise is introduced in the global electrostatic interaction energy term. On the other hand, the performance of C_{delphi} is notably improved over that of C_{amber} because

COMBINE analysis filters out this noise by changing the weights given to those residues whose interaction energies with the ligands fluctuate but do not contribute to explaining differences in activity. At the same time, C_{delphi} benefits from the better description of the binding electrostatics, which include solvation effects, and a highly predictive regression model is produced.

Mutagenesis experiments suggest that ligand–receptor interactions can be analyzed in terms of a limited set of discrete interactions rather than as the sum of a larger set of independent weak interactions.⁴³ For different HIV-1 protease inhibitors, distinctly different patterns of resistance arise, with mutations at different sites.⁴⁰ Given that the number of structures of mutant enzymes with different inhibitors is continually increasing,⁴⁴ the methodology presented in this paper can make it possible to study the interrelationships among the different binding energy contributions and address the issue of resistance in molecular design.

Conclusions

The good correlation found between calculated molecular mechanics interaction energies and experimental inhibitory potencies for a series of HIV-1 protease inhibitors using the MM2X force field has been corroborated using the AMBER force field. A more accurate QSAR model was derived when the ligand–receptor interaction energies were decomposed into residue-based van der Waals and electrostatic contributions, and a PLS analysis was done to fit the activity data (COMBINE analysis). The use of a continuum method to calculate the electrostatic contributions to the binding energy taking into account the desolvation of the binding interface provided an alternate set of electrostatic energy descriptors. When they were incorporated into the analysis in place of the Coulombic energy interactions, the model improved even further.

On the basis of our results from a study involving trypsin and a short series of flavonoid inhibitors,¹⁰ we concluded that the good correlations obtained by different authors for congeneric series using only force field calculations^{5–8} could be accounted for, at least in part, by the fact that solvation effects tend to be of similar magnitude within the series. Thus, when these effects are ignored, reasonable correlations can be derived although a closer agreement between calculated and experimental free energy differences can be obtained when desolvation is explicitly included.¹⁰ Our observations from the considerably larger series of congeneric HIV-1 protease inhibitors studied here confirm our previous findings and, at the same time, demonstrate that molecular mechanics-based QSAR studies on ligand–receptor complexes can benefit greatly from proper incorporation of solvation effects into a COMBINE framework based on residue-based interaction energy decomposition.

Note. Coordinates of all the complexes, atomic charges, AMBER parameters, and interaction energy matrixes are available from the authors upon request (e-mail: fgago@fisfar.alcala.es).

Acknowledgment. We thank Dr. Kate Holloway (Merck Research Laboratories) for providing the coordinates of HIV-1 protease in complex with molecules **1** and **3–34** and both referees for extremely useful com-

ments and suggestions. Fellowships to C. Pérez (Spanish Ministerio de Educación y Cultura) and A. R. Ortiz (Comunidad Autónoma de Madrid) are gratefully acknowledged. Biosym/Molecular Simulations, Inc. contributed a software license and the University of Alcalá Computing Centre granted a generous allowance of computer time on a SGI Power Challenge server. This research has been financed by the Spanish CICYT (Projects SAF94-630 and SAF96-231).

References

- (1) (a) Greer, J.; Erickson, J. W.; Baldwin, J. J.; Varney, M. D. Application of the Three-Dimensional Structures of Protein Target Molecules in Structure-Based Drug Design. *J. Med. Chem.* **1994**, *37*, 1035–1054. (b) Whittle, P. J.; Blundell, T. L. Protein Structure-Based Drug Design. *Annu. Rev. Biophys. Biomol. Struct.* **1994**, *23*, 349–375. (c) Babine, R. E.; Bender, S. L. Molecular Recognition of Protein–Ligand Complexes: Applications to Drug Design. *Chem. Rev.* **1997**, *97*, 1359–1472.
- (2) Ajay; Murcko, M. A. Computational Methods to Predict Binding Free Energy in Ligand–Receptor Complexes. *J. Med. Chem.* **1995**, *38*, 4953–4967.
- (3) Gilson, M. K.; Given, J. A.; Bush, B. L.; McCammon, J. A. The Statistical-Thermodynamic Basis for Computation of Binding Affinities: A Critical Review. *Biophys. J.* **1997**, *72*, 1047–1069.
- (4) (a) Searle, M. S.; Williams, D. H.; Gerhard, U. Partitioning of Free Energy Contributions in the Estimation of Binding Constants: Residual Motions and Consequences for Amide-Amide Hydrogen Bond Strengths. *J. Am. Chem. Soc.* **1992**, *114*, 10697–10704. (b) King, B. L.; Vajda, S.; DeLisi, C. Empirical Free Energy as a Target Function in Docking and Design: Application to HIV-1 Protease Inhibitors. *FEBS Lett.* **1996**, *384*, 87–91. (c) Weng, Z. S.; Vajda, S.; DeLisi, C. Prediction of Protein Complexes Using Empirical Free Energy Functions. *Protein Sci.* **1996**, *5*, 614–626. (d) Böhm, H.-J.; Keble, G. What Can We Learn from Molecular Recognition in Protein–Ligand Complexes for the Design of New Drugs? *Angew. Chem., Int. Ed. Engl.* **1996**, *35*, 2588–2614. (e) Janin, J. Elusive Affinities. *Proteins* **1995**, *21*, 30–39. (f) Jain, A. N. Scoring Noncovalent Protein–Ligand Interactions: A Continuous Differentiable Function Tuned to Compute Binding Affinities. *J. Comput.-Aided Mol. Design* **1996**, *10*, 427–440. (g) Novotny, J.; Bruccoleri, R. E.; Davis, M.; Sharp, K. A. Empirical free Energy Calculations: A Blind Test and Further Improvements to the Method. *J. Mol. Biol.* **1997**, *268*, 401–411.
- (5) Menziani, M. C.; De Benedetti, P. G.; Gago, F.; Richards, W. G. The Binding of Benzenesulfonamides to Carbonic Anhydrase Enzyme. A Molecular Mechanics Study and Quantitative Structure–Activity Relationships. *J. Med. Chem.* **1989**, *32*, 951–956.
- (6) (a) Holloway, M. K.; Wai, J. M.; Halgren, T. A.; Fitzgerald, P. M. D.; Vacca, J. P.; Dorsey, B. D.; Levin, R. B.; Thompson, W. J.; Chen, L. J.; deSolms, S. J.; Gaffin, N.; Ghosh, A. K.; Giuliani, E. A.; Graham, S. L.; Guare, J. P.; Hungate, R. W.; Lyle, T. A.; Sanders, W. M.; Tucker, T. J.; Wiggins, M.; Wiscourt, C. M.; Woltersdorf, O. W.; Young, S. D.; Darke, P. L. & Zugay, J. A. A Priori Prediction of Activity for HIV-1 Protease Inhibitors Employing Energy Minimization in the Active Site. *J. Med. Chem.* **1995**, *38*, 305–317. (b) Holloway, M. K.; Wai, J. M. Structure-Based Design of Human Immunodeficiency Virus-1 Protease Inhibitors. In *Computer-Aided Molecular Design. Applications in Agrochemicals, Materials, and Pharmaceuticals*; Reynolds, C. H., Holloway, M. K., Cox, H. K., Eds.; American Chemical Society: Washington, 1995; pp 36–50.
- (7) Grootenhuis, P. D. J.; van Galen, P. J. M. Correlation of Binding Affinities with Non-Bonded Interaction Energies of Thrombin-Inhibitor Complexes. *Acta Crystallogr.* **1995**, *D51*, 560–566.
- (8) Carson, M.; Yang, Z.; Babu, Y. S.; Montgomery, J. A. Calculation of Relative Binding Affinities of Purine Nucleoside Phosphorylase Inhibitors. *Acta Crystallogr.* **1995**, *D51*, 536–540.
- (9) Viswanadhan, V. N.; Reddy, M. R.; Wlodawer, A.; Varney, M. D.; Weinstein, J. N. An Approach to Rapid Estimation of Relative Binding Affinities of Enzyme Inhibitors: Application to Peptidomimetic Inhibitors of the Human Immunodeficiency Virus Type 1 Protease. *J. Med. Chem.* **1996**, *39*, 705–712.
- (10) Checa, A.; Ortiz, A. R.; de Pascual-Teresa, B.; Gago, F. Assessment of Solvation Effects on Calculated Binding Affinity Differences: Trypsin Inhibition by Flavonoids as a Model System for Congeneric Series. *J. Med. Chem.* **1997**, *40*, 4136–4145.
- (11) (a) Straatsma, T. P.; McCammon, J. A. Computational Alchemy. *Annu. Rev. Phys. Chem.* **1992**, *43*, 407–435. (b) Kollman, P. Free Energy Calculations: Applications to Chemical and Biochemical Phenomena. *Chem. Rev.* **1993**, *93*, 2395–2417. (c) Resat, H.; Marrone, T. J.; McCammon, J. A. Enzyme–Inhibitor Association Thermodynamics: Explicit and Continuum Solvent Studies. *Biophys. J.* **1997**, *72*, 522–532.

- (12) Åqvist, J.; Medina, C.; Samuelsson, J.-E. A New Method for Predicting Binding Affinity in Computer-Aided Drug Design. *Protein Eng.* **1994**, *7*, 385–391.
- (13) Jones-Hertzog, D. K.; Jorgensen, W. L. Binding Affinities for Sulfonamide Inhibitors with Human Thrombin Using Monte Carlo Simulations with a Linear Response Method. *J. Med. Chem.* **1997**, *40*, 1539–1549.
- (14) Reddy, M. R.; Varney, M. D.; Kalish, V.; Viswanadhan, V. N.; Appelt, K. Calculation of Relative Differences in the Binding Free Energies of HIV-1 Protease Inhibitors: A Thermodynamic Cycle Perturbation Approach. *J. Med. Chem.* **1994**, *37*, 1145–1152.
- (15) (a) Ortiz, A. R.; Pisabarro, M. T.; Gago, F.; Wade, R. C. Prediction of Drug Binding Affinities by Comparative Binding Energy Analysis. *J. Med. Chem.* **1995**, *38*, 2681–2691. (b) Wade, R. C.; Ortiz, A. R.; Gago, F. Comparative Binding Energy Analysis. In *3D QSAR in Drug Design, Vol. II, Ligand-Protein Interactions and Molecular Similarity*; Kubinyi, H., Folkers, G., Martin, Y. C., Eds.; Kluwer Academic Publishers: Dordrecht, in press.
- (16) (a) Wold, S.; Johansson, E.; Cocchi, M. PLS-Partial Least-Squares Projections to Latent Structures. In *3D QSAR in Drug Design. Theory, Methods and Applications*; Kubinyi, H., Ed.; ESCOM Science Publishers B.V.: Leiden, 1993; pp 523–550. (b) Cramer, R. D., III. Partial Least Squares (PLS): Its Strengths and Limitations. *Perspect. Drug Discovery Design* **1993**, *1*, 269–278.
- (17) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. Semianalytical Treatment of Solvation for Molecular Mechanics and Dynamics. *J. Am. Chem. Soc.* **1990**, *112*, 6127–6129.
- (18) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.
- (19) Besler, B. H.; Merz, K. M.; Kollman, P. A. Atomic Charges Derived from Semiempirical Methods. *J. Comput. Chem.* **1990**, *11*, 431–439.
- (20) Thompson, W. J.; Fitzgerald, P. M.; Holloway, M. K.; Emini, E. A.; Darke, P. L.; McKeever, B. M.; Schleif, W. A.; Quintero, J. C.; Zugay, J. A.; Tucker, T. J.; Schwering, J. E.; Hommick, C. F.; Nunberg, J.; Springer, J. P.; Huff, J. R. Synthesis and Antiviral Activity of a Series of HIV-1 Protease Inhibitors with Functionality Tethered to the P1 or P1' Phenyl Substituents: X-ray Crystal Structure Assisted Design. *J. Med. Chem.* **1992**, *35*, 1685–1701.
- (21) Wlodawer, A.; Erickson, J. W. Structure-Based Inhibitors of HIV-1 Protease. *Annu. Rev. Biochem.* **1993**, *62*, 543–585.
- (22) Wang, Y.-X.; Freedberg, D. I.; Yamazaki, T.; Wingfield, P. T.; Stahl, S. J.; Kaufman, J. D.; Kiso, Y.; Torchia, D. A. Solution NMR Evidence That the HIV-1 Protease Catalytic Aspartyl Groups Have Different Ionization States in the Complex Formed with the Asymmetric Drug KNI-272. *Biochemistry* **1996**, *35*, 9945–9950.
- (23) Insight II, release 95.0 (1995), Biosym/Molecular Simulations, 9685 Scranton Road, San Diego, CA 92121-3752.
- (24) Stewart, J. J. P. MOPAC 93 (1993), Fujitsu Ltd., Tokyo, Japan.
- (25) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. AM1: A New General Purpose Quantum Mechanical Molecular Model. *J. Am. Chem. Soc.* **1985**, *107*, 3902–3909.
- (26) AMBER (UCSF): *Assisted Model Building with Energy Refinement*, version 4.1, 1995. Department of Pharmaceutical Chemistry, University of California, San Francisco.
- (27) Nicholls, A.; Honig, B. A Rapid Finite Difference Algorithm, Utilizing Successive Over-Relaxation to Solve the Poisson-Boltzmann Equation. *J. Comput. Chem.* **1991**, *12*, 435–445.
- (28) Weiner, S. J.; Kollman, P. A.; Case, D. A.; Singh, U. C.; Ghio, C.; Alagona, C.; Profeta, S.; Weiner, P. A New Force Field for Molecular Mechanical Simulation of Nucleic Acids and Proteins. *J. Am. Chem. Soc.* **1984**, *106*, 765–784.
- (29) (a) Gilson, M. K.; Sharp, K. A.; Honig, B. H. Calculating the Electrostatic Potential of Molecules in Solution: Method and Error Assessment. *J. Comput. Chem.* **1987**, *9*, 327–335. (b) Gilson, M. K.; Honig, B. Calculation of the Total Electrostatic Energy of a Macromolecular System: Solvation Energies, Binding Energies, and Conformational Analysis. *Proteins* **1988**, *4*, 7–18.
- (30) Jackson, R. M.; Sternberg, M. J. E. A Continuum Model for Protein-Protein Interactions: Application to the Docking Problem. *J. Mol. Biol.* **1995**, *250*, 258–275.
- (31) Ortiz, A. R.; Pastor, M.; Palomer, A.; Cruciani, G.; Gago, F.; Wade, R. C. Reliability of CoMFA Models: Effects of Data Scaling and Variable Selection Using a Set of Human Synovial Fluid Phospholipase A₂ Inhibitors. *J. Med. Chem.* **1997**, *40*, 1136–1148.
- (32) Richards, F. M. Areas, Volumes, Packing and Protein Structure. *Annu. Rev. Biophys. Bioeng.* **1977**, *6*, 151–176.
- (33) Connolly, M. L. Analytical Molecular Surface Calculation. *J. Appl. Crystallogr.* **1983**, *16*, 548–558.
- (34) Klapper, I.; Hagstrom, R.; Fine, R.; Sharp, K.; Honig, B. Focusing of Electric Fields in the Active Site of Cu-Zn Superoxide Dismutase: Effects of Ionic Strength and Amino-acid Modification. *Proteins* **1986**, *1*, 47–59.
- (35) Baroni, M.; Constantino, G.; Cruciani, G.; Riganelli, D.; Valigi, R.; Clementi, S. Generating Optimal Linear PLS Estimations (GOLPE): An Advanced Chemometric Tool for Handling 3D QSAR Problems. *Quant. Struct.-Act. Relat.* **1993**, *12*, 9–20.
- (36) GOLPE, version 3.0 (1996), Multivariate Infometric Analysis, Viale dei Castagni 16, Perugia, Italy.
- (37) Cruciani, G.; Clementi, S.; Baroni, M. Variable Selection in PLS Analysis. In *3D QSAR in Drug Design. Theory, Methods and Applications*; Kubinyi, H., Ed.; ESCOM Science Publishers B.V.: Leiden, 1993; pp 551–564.
- (38) Cuevas, C.; Pérez, C.; Pastor, M.; Gago, F. In preparation.
- (39) Pisabarro, M. T.; Ortiz, A. R.; Palomer, A.; Cabré, F.; García, L.; Wade, R. C.; Gago, F.; Mauleón, D.; Carganico, G. Rational modification of human synovial fluid phospholipase A₂ inhibitors. *J. Med. Chem.* **1994**, *37*, 337–341.
- (40) West, M. L.; Fairlie, D. P. Targeting HIV-1 Protease: A Test of Drug Design Methodologies. *Trends Pharmacol. Sci.* **1995**, *16*, 67–75.
- (41) Bardi, J. S.; Luque, I.; Freire, E. Structure-Based Thermodynamic Analysis of HIV-1 Protease Inhibitors. *Biochemistry* **1997**, *36*, 6588–6596.
- (42) Rosenfeld, R.; Vajda, S.; DeLisi, C. Flexible Docking and Design. *Annu. Rev. Biophys. Biomol. Struct.* **1995**, *24*, 677–700.
- (43) Strange, P. G. The Energetics of Ligand Binding at Catecholamine Receptors. *Trends Pharmacol. Sci.* **1996**, *17*, 238–244.
- (44) (a) Baldwin, E. T.; Bhat, T. N.; Liu, B.; Pattabiraman, N.; Erickson, J. W. Structural Basis of Drug Resistance for the V82A Mutant of HIV-1 Proteinase. *Nature Struct. Biol.* **1995**, *2*, 244–249. (b) Hong, L.; Treharne, A.; Hartsuck, J. A.; Foundling, S.; Tang, J. Crystal Structures of Complexes of a Peptidic Inhibitor with Wild-Type and Two Mutant HIV-1 Proteases. *Biochemistry* **1996**, *35*, 10627–10633. (c) Ala, P. J.; Huston, E. E.; Klabe, R. M.; McCabe, D. D.; Duke, J. L.; Rizzo, C. J.; Korant, B. D.; DeLoskey, R. J.; Lam, P. Y.; Hodge, C. N.; Chang, C. H. Molecular Basis of HIV-1 Protease Drug Resistance: Structural Analysis of Mutant Proteases Complexed with Cyclic Urea Inhibitors. *Biochemistry* **1997**, *36*, 1573–1580.

JM970535B