

Surface Descriptors for Protein–Ligand Affinity Prediction

Ismael Zamora,^{*,†} Tudor Oprea,[‡] Gabriele Cruciani,[§] Manuel Pastor,^{||} and Anna-Lena Ungell[†]

DMPK & Bioanalytical Chemistry, AstraZeneca R & D Mölndal, S-431 83 Mölndal, Sweden, EST Chemical Computing, AstraZeneca R & D Mölndal, S-431 83 Mölndal, Sweden, Laboratory for Chemometrics, University of Perugia, I-06123 Perugia, Italy, and Institut Municipal d'Investigació Mèdica – IMIM/UPF, Doctor Aiguader, 80, E-08003 Barcelona, Spain

Received October 8, 2001

Molecular descriptors calculated by the VolSurf program have been extensively used to model pharmacokinetic properties, e.g., passive permeability through the gastrointestinal tract or through the blood–brain barrier. These descriptors quantify steric, hydrophobic, and hydrogen bond interactions between model compounds and different environments. Since these interactions are the same as those involved in the ligand–receptor binding, VolSurf descriptors could potentially be relevant in modeling this process as well. We obtained a significant model ($r^2 = 0.85$, $q^2 = 0.75$) using VolSurf descriptors derived from the ligand, the protein, and the ligand–protein complex for a diverse set of 38 structures previously used in the VALIDATE (ref 23) training set. Furthermore, a statistically significant model ($r^2 = 0.94$, $q^2 = 0.89$) was obtained using the same type of descriptors for a homogeneous set of glycogen phosphorylase inhibitors (ref 25). Using the VolSurf computational framework, both ligand–receptor binding and the ligand's pharmacokinetic behavior can be modeled simultaneously during the preclinical aspects of drug discovery.

Introduction

One of the major reasons for failure in the late stages of the drug discovery process (phases II and III of clinical trials) is the inadequate understanding of the pharmacokinetic behavior of drugs, and what constitutes a suitable pharmacokinetic profile for candidate drugs. To overcome these problems, *in vitro* or *in vivo* measurements are performed as early as possible in the drug discovery process. During the past decade, the number of synthetically accessible compounds has increased by several orders of magnitude due to combinatorial chemistry. As a result, high throughput methods to evaluate pharmacokinetic properties are needed. Recently, computational efforts have been made to obtain models that describe and predict the pharmacokinetic behavior of a compound.^{1–8} However, to streamline the drug discovery cycle, the efforts to predict and eventually optimize the pharmacokinetic properties should be coupled with the prediction and optimization of the binding affinity of the same analogue^{9,10} (see Figure 1). Several methods that model the ligand–receptor interaction and predict the binding affinity of a compound for a given protein target have been described.^{11–17}

Traditionally in the drug discovery process, the potency is optimized first, and the ADMET (administration, distribution, metabolism, excretion, and toxicology) properties are studied in later stages (Figure 1A). This methodology has limited success, as potency optimization

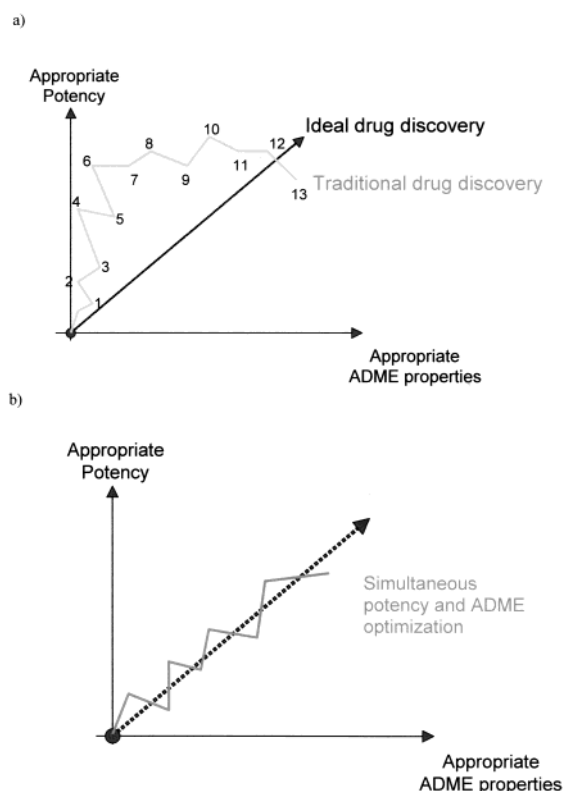


Figure 1. Drug discovery strategy: (a) first potency optimization and then ADME properties; (b) combine optimization of potency and ADME properties.

* To whom correspondence should be addressed: Lead Molecular Design, s.l., Francesc Cabanes i Alibau, 1–3 2^a-1^abb, 08190 Sant Cugat del Vallès, Barcelona, Spain, Tel: + 34 93 674 90 24, e-mail: ismael.zamora@telefonica.net.

[†] DMPK & Bioanalytical Chemistry, AstraZeneca R & D Mölndal.

[‡] EST Chemical Computing, AstraZeneca R & D Mölndal.

[§] University of Perugia.

^{||} Institut Municipal d'Investigació Mèdica – IMIM/UPF.

can yield compounds for which ADMET optimization becomes difficult, and often impossible. The efficiency of the drug discovery process is expected to improve if both aspects are considered at the same time, i.e., optimizing the chemical structure by considering both potency and ADMET properties (Figure 1B).

Table 1. Protein Data Bank Code and pK_i Values for the Validate Data Set. Some PKi Were Modified, Compared with the Original VALIDATE Publication³⁰

PDB Code	pK_i	PDB Code	pK_i
1AAQ	7.93	2SIN	11
1ABE	7.01	2TMN	5.89
1ABF	5.42	3SIC	10.2
1APB	5.82	3TMN	5.9
1DBB	9	4ER1	6.62
1DBJ	7.6	4ER4	6.8
1DBM	9.44	4PHV	9.15
1EED	4.79	4TMN	10.17
1HIV	9.15	5SIC	10.2
1HVI	10.5	5TMN	8.04
1SBN	10.3	6ABP	6.36
1TLP	8.55	6TMN	5.05
1TMN	7.3	7ABP	6.46
2DBL	8.7	7ABP	6.46
2ER0	6.38	7HVP	9.62
2ER6	7.22	7TLN	2.12
2ER7	9	8ABP	8
2ER9	7.4	9ABP	8
2PTC	13.3	9HVP	7.63

In this paper, we describe a possible integrated framework⁹ that uses a common set of descriptors suitable for modeling both the pharmacokinetic and binding affinity aspects in drug design. The rationale behind this integrative approach is that those interactions responsible for passive permeability are the same as those involved in the ligand–protein binding process, i.e., steric, hydrophobic, electrostatic, and hydrogen bonding.^{11–17} VolSurf descriptors,^{18–20} based on the GRID^{21,22} interaction fields, were previously shown as suitable for modeling pharmacokinetic properties.^{7,9,18} VolSurf quantifies the interaction of a compound with a predefined environment, typically hydrophobic, and hydrogen bond donor and acceptor. The value of these volumes and surfaces of interaction provide a description of the potential interaction for the ligand and/or receptor and the interactions that were not used in making the complex for the complex ligand–receptor. To illustrate the applicability of VolSurf as an integrated drug discovery framework, we evaluate the use of VolSurf descriptors for modeling ligand–receptor binding affinity for two distinct data sets: a diverse set of 38 structures, previously used in VALIDATE²³ training set, and a homogeneous set of glycogen phosphorylase inhibitors.^{24,25} A number of articles illustrate the application of VolSurf in pharmacokinetics.^{26–28}

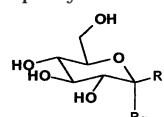
Materials and Methods

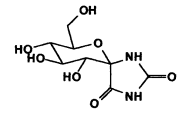
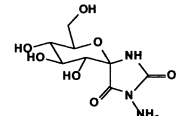
The QSAR analysis performed in this study is based on VolSurf²⁰ descriptors that quantify the interactions for a given compound, by evaluating them with three of the available chemical probes in the GRID program.²² The molecular interaction fields (MIF) of a water molecule, a sp^2 oxygen atom (which can accept, but not donate hydrogen bonds), and the DRY probe (representing a hydrophobic interaction) form the basis for VolSurf descriptors, as published elsewhere.^{18,19}

Two data sets were chosen for this analysis: • The VALIDATE data set^{14,23} contains a diverse set of 38 compounds from set of proteins (see Table 1). This would illustrate the usefulness of VolSurf descriptors for a general-case scoring scheme.

• The P-glycogen phosphorylase B data set^{24,25} contains 23 ligands cocrystallized with the same protein (see Table 2). This illustrates the usefulness of VolSurf descriptors in a typical drug discovery project, where most compounds interact with the same target, and in the same binding site.

Table 2. Structures of the Different Ligands and pK_i Values for the Glycogen Phosphorylase Data Set



Ligand	R α	R β	-logK i
GLUCOSE	OH	-	2.770
AAM	CH ₂ NH ₃ ⁺	-	1.462
ACH ₂ OH	CH ₂ OH	-	2.824
AZIDE	CONHNH ₂	-	2.523
BZIDE	-	CONHNH ₂	3.398
BMA	-	CH ₂ NH ₃ ⁺	1.775
BSUL1	-	CH ₂ OSO ₂ CH ₃	2.319
BCONH ₂	-	CONH ₂	3.357
BCONHME	-	CONHCH ₃	3.796
BNCONH	-	NHCONH ₂	3.854
BNHCOMET	-	NHCOCH ₃	4.495
BNCOPR	-	NHCOCH ₂ CH ₃	4.409
BNCOPR	-	NHCOCH ₂ CH ₂ CH ₃	4.027
BNHCO1	-	NHCOCH ₂ Cl	4.347
BNHCO2	-	NHCOCH ₂ Br	4.357
BNCHNH	-	NHCOCH ₂ NH ₂	3.432
BNHCOB	-	NHCO-Ph	4.092
BNHCOB	-	NHCOCH ₂ NHCOCH ₃	3.004
BNCOOP	-	NHCOOCH ₂ -Ph	3.456
CAR	CONH ₂	NHCOOCH ₃	4.796
HYDA		Id	5.523
NHYDAN		Id	3.836
5_THIO	OH	-	2.699

A. The VALIDATE Data Set. The molecules and corresponding receptors were obtained from the Protein Data Bank.²⁹ The pK_i values³⁰ and the four-letter PDB code are presented in Table 1. The ligand structure was extracted from each complex. Three separate GRID-VolSurf calculations were performed using the receptor, the ligand, and the ligand–receptor complex, respectively. The protein structures were converted to GRID format by using the GREAT and GRIN modules in the GRID package. Ligands were converted using the VolSurf interface by importing the SYBYL mol2 files. No further optimizations were performed in the receptor, ligands, or complexes.

Water molecules may play an important role in the ligand–receptor interaction.^{25,26,31,32} To study this possibility, two analyses were performed on this data set: One including the water molecules found in the X-ray structure, and the other excluding these waters. Another factor that needs to be considered in the protein modeling using GRID is the charge of the protein/complex after the atom-type assignment. This was analyzed considering the neutralized and non-neutralized forms of the proteins, respectively. Thus, four types of MIFs were calculated (a) with the crystal waters and not neutralizing the proteins, (b) with the crystal waters and neutralized proteins, (c) without the crystal waters but neutralizing the protein, and (d) excluding the crystal waters and not neutralizing the protein.

To find the positions of the counterions for the situations described under b and d above, the MINIM and FILMAP programs included in the GRID package²² were used. The atom types for the protein with the ligand in the crystal position were assigned using the GRIN program. GRIN also gives an indication about the overall charge of the protein and, depending on the sign, either a Na⁺ probe was used to neutralize a negatively charged complex, or a Cl⁻ probe was used to neutralize a positively charged one. The MIF was calculated for the whole protein, using the ionic probe with a 0.5 Å grid spacing. The MINIM program was used to locate the position

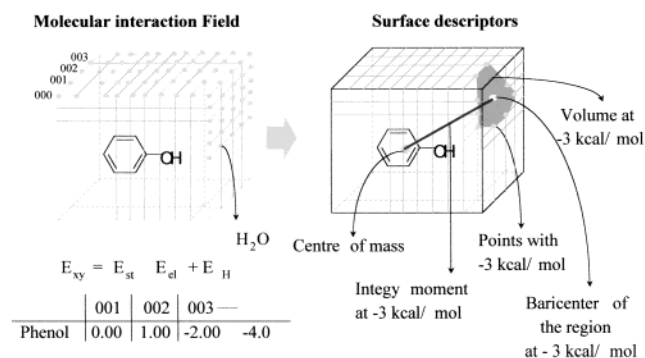


Figure 2. GRID-VolSurf calculation process: (1) The GRID program obtains the molecular interaction field. (2) The Volume and Surface descriptors are obtained at different interaction energy levels.

of the minima in the interaction field with a charge probe using 0.0 kcal/mol as the energy threshold and without interpolation. The FILMAP program was used to populate the different minima with simulated annealing (SA), taking the net complex charge plus 4 as the number of minima to populate. For example, if the net charge was 7, we used 11 minima, whereas a net charge of 3 would give 7 minima during the SA process. The counterion positions that were outside the ligand-derived GRID box plus 5 Å, were selected. Those counterion positions that were inside the GRID box were not considered, since it was assumed that such charges are important for the ligand–receptor interaction.

The strategy for the VolSurf calculation is presented in Figure 2. In the first step, the interaction fields for the H₂O, O, and DRY probes were calculated on the four types of MIFs (a–d). A GRID box that extended 5 Å beyond the maximum and minimum atomic coordinates of the ligand was used with a 0.5 Å spacing for the receptor, the ligand, and the complex. In the second step, VolSurf descriptors were derived from the MIFs (see Table 3). These descriptors were analyzed by Partial Least Squares (PLS) using the GOLPE program with the autoscaling option before the statistical model was obtained.³³ The counterion positions that were outside the ligand-derived GRID box, plus 5 Å, were selected. Those counterion positions that were inside the GRID box were not considered, since it

was assumed that such charges are important for the ligand–receptor interaction.

The strategy for the VolSurf calculation is presented in Figure 2. In the first step, the interaction fields for the H₂O, O, and DRY probes were calculated on the four types of MIFs (a–d). A GRID box that extended 5 Å beyond the maximum and minimum atomic coordinates of the ligand was used with a 0.5 Å spacing for the receptor, the ligand, and the complex. In the second step, VolSurf descriptors were derived from the MIFs (see Table 3). These descriptors were analyzed by Partial Least Squares (PLS) using the GOLPE program with the autoscaling option before the statistical model was obtained.^{34,35}

B. The Glycogen Phosphorylase-b Data Set. The pK_a and the structures of the ligands used in this data set are presented in Table 2. The protein, ligand, and complex were prepared by using the nonneutralized protein without water, applying the VolSurf procedure, as detailed above.

Results and Discussion

The VALIDATE Data Set. Molecule Preparation: Counterion/Protein Neutralization. FILMAP was used to populate the counterion energy minima for the different proteins treated. These were filled by counterions, and the counterions positions were localized outside or inside the GRID box defined by the atomic coordinates of the ligand. Counterions were observed inside the ligand-defined box for the L-arabinose binding protein, for endothiapepsin and for two of the HIV-1 protease structures. To avoid the overlap with the ligands in the binding site, counterions were placed at FILMAP-identified locations only outside the ligand-defined boxes, even though these were found to be less populated during the SA procedure. In all instances where counterions were located inside the ligand-defined box for L-arabinose binding protein, the counterion was found at 2.0–2.5 Å from the side-chain of Asp²³⁵, which in turn was 4.5 Å away from the ligand. A counterion placed in that position would neutralize this aspartate, thus masking its potentially significant

Table 3. VolSurf Descriptors

	Descriptors Obtained from the Hydrophilic (H ₂ O) Interaction
V	volume of the water molecule interaction field at 0.2 kcal/mol energy level
S	surface of the water interaction field at the same 0.2 kcal/mol level
R	rugosity, e.g. the ratio between the volume and the surface
G	globularity, e.g., the ratio between the surface (S) and the surface of a sphere with the same volume (V)
W1–W15	the volume of the hydrophilic interactions at five different energy levels: –0.2, –0.5, –1.0, –2.0, –3.0, –4.0, –5.0, –6.0, –7.0, –8.0, –9.0, –10.0, –11.0, –12.0, –13.0, –14.0 kcal/mol
IW1–IW15	the integy moments at the same energy levels as W1–W15
CW1–CW15	the capacity factors, which are the ratio between the hydrophilic regions (W1–W15) and the molecular surface (S)
Emin1, Emin2, and Emin3	these descriptors express the energy values for the three lowest energy minima
D12, D13, and D23	the distances between the three minima
D1–D15	Descriptors Obtained from the Hydrophobic (DRY) Interaction the volume of the hydrophobic interactions at 15 energy levels: –0.2, –0.4, –0.6, –0.8, –1.0, –1.2, –1.4, –1.6, –1.8, –1.9, –2.0, –2.1, –2.2, –2.4 and –2.6 kcal/mol
ID1–ID15	the integy moment at the previous energy levels
HL1–HL2	hydrophilic–lipophilic balance: the ratio between the volume of the hydrophilic regions at –3 and –4 kcal/mol and the hydrophobic regions at –0.6 and –0.8 kcal/mol
A	the strength of the amphiphilic moment.
CP	critical packing parameter.
POL	polarizability
Wp1–Wp15	Descriptors Obtained from the Polar (O) Interaction the volume of the interaction with the O probe at 15 different energy levels
HB1–HB15	descriptors that represents the difference between the volume of the hydrophilic interaction (W1–W15) and the O probe interactions (Wp1–Wp15) and expresses the hydrogen bond donor capability of the target

Table 4. Validate Set Modeling Information: Water Molecules in the Complex and in the Receptor. Formal Charges

protein	charge _(complex)	charge _(receptor)	total no. water	inside no. water
1AAQ	+4	+4	1	1
1ABE	-4	-4	227	9
1ABF	-4	-4	191	11
1APB	-4	-4	168	11
1DBB	+2	+2	0	0
1DBJ	+2	+2	0	0
1DBM	+1	+2	0	0
1EED	-14	-14	278	41
1HIV	+4	+4	90	23
1HVI	+4	+4	1	1
1SBN	+4	+2	316	107
1TLP	-2	-2	162	0
1TMN	-2	-2	144	0
2DBL	+2	+2	0	0
2ER0	-14	-14	0	0
2ER6	-15	-14	321	50
2ER7	-20	-20	321	81
2ER9	-14	-14	321	41
2PTC	+14	+8	157	83
2SNI	+2	+3	168	80
2TMN	-2	-2	165	0
3SIC	-1	+2	273	148
3TMN	-2	-2	173	0
4ER1	-14	-14	260	33
4ER4	-13	-14	325	54
4PHV	+4	+4	104	23
4TMN	-2	-2	162	0
5SIC	-1	+2	288	169
5TMN	-2	-2	173	0
6ABP	-4	-4	205	10
6TMN	-2	-2	170	0
7ABP	-4	-4	193	11
4HVP	+4	+4	95	24
7TLN	-1	-2	166	9
8ABP	-4	-4	207	11
9ABP	-4	-4	207	9
9HVP	+4	+4	1	1

contribution to the ligand–protein interaction. Therefore, counterions inside the GRID box were ignored.

Molecule Preparation: Water Molecules. As water molecules play an important role in the binding process, their influence in the VolSurf models was examined. The number of water molecules present in the crystal structures available from the PDB are summarized in Table 4, along with the number of water molecules inside the ligand-defined GRID box.

Quantitative Structure-Binding Relationships. Four different models were obtained using the water/nonwater and the neutralized/non-neutralized conditions. The predicted versus experimental results for each of the models are shown in Figure 3. Three compounds could not be predicted by any model (9HVP, 2PTC, and 2ER7) and were found to be outliers.

Quantitative Structure-Binding Relationships: Outlier Analysis. The ligand for the 2PTC complex is bigger than any other compound in the training set and has the highest pK_i value. The 9HVP complex has a key water molecule, responsible for a hydrogen bond between the protein and the ligand. To investigate the influence of this water, the model in the neutralized form and without crystal waters was used to predict the activity of this compound with and without the water molecule. When the water molecule is considered, the binding affinity prediction improves. In this case, the structure of the compound is better described when the water molecule is present (see Figure 3). Therefore, we

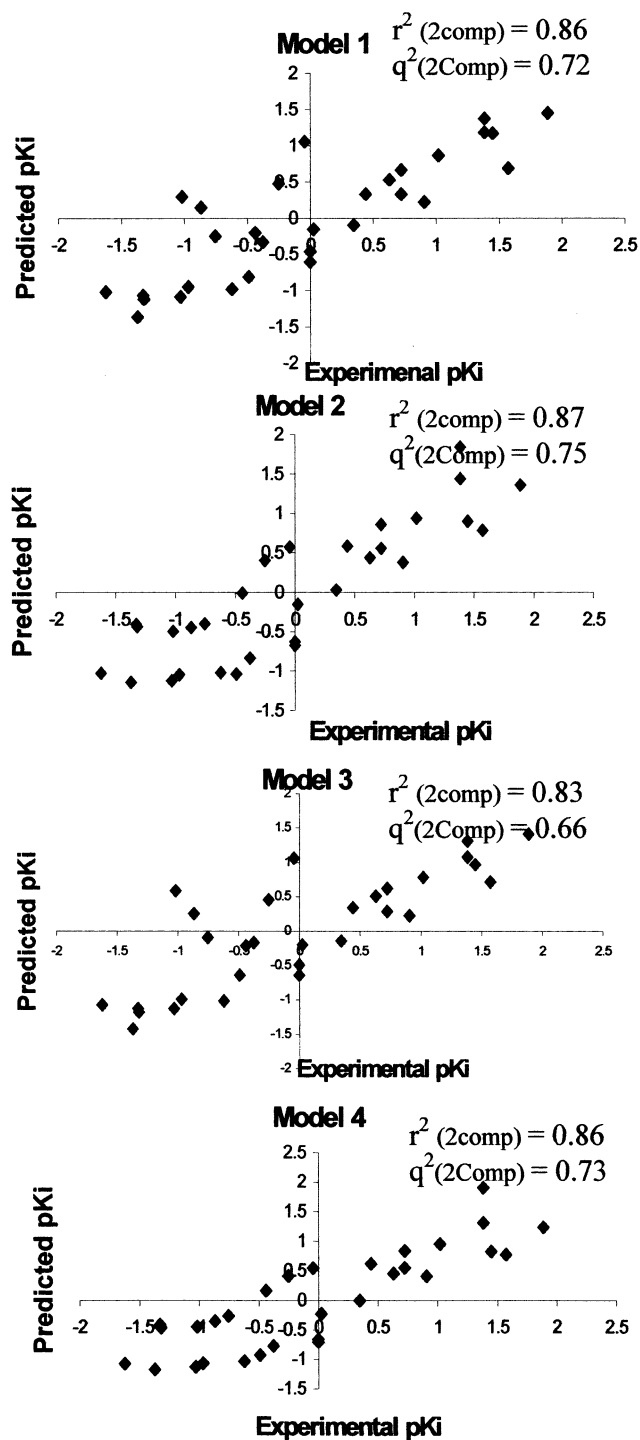


Figure 3. Experimental vs predicted pK_i values for the validate data set. Model 1: With H_2O and not neutralized. Model 2: Without H_2O and not neutralized. Model 3: With H_2O and neutralized. Model 4: Without H_2O and neutralized.

conclude that the structural water molecule plays an important role in this particular case.

However, we did not observe any influence on the quality of predictions when comparing the inclusion/exclusion of water molecules for the other HIV protease inhibitor complexes included in this study. This implies that displacing this water molecule in HIV protease may not necessarily be a beneficial effect.

The 2ER7 complex is the one with the highest charge and the position of the counterions in this particular case could have a great impact in the interaction energy,

Table 5. Influence of the Different Blocks of Variables on the Model

ligand	complex	receptor	r^2	q^{2a}	q^{2b}	SDEP ^b	SDEP (sdep) ^b
+	–	–	0.69	0.34	0.32	0.81	0.09
–	+	–	0.73	0.36	0.35	0.79	0.06
–	–	+	0.71	0.46	0.45	0.73	0.04
–	+	+	0.72	0.40	0.38	0.77	0.04
+	–	+	0.85	0.62	0.59	0.62	0.05
+	+	–	0.84	0.53	0.51	0.68	0.06
+	+	+	0.82	0.53	0.51	0.68	0.04

^a Leave one out. ^b Leave five random groups out 100 times.

this effect could perhaps account for the poor prediction of its binding affinity.

Water and Counterion Analysis. Since the counterion positions selected to neutralize the proteins were outside the box defined by the maximum and minimum coordinates of each ligand, no significant difference was observed between these models (Figure 3: models 1 and 3; and models 2 and 4). The introduction of the water molecules could not improve the statistical performance of the models, although the water-containing model could predict some compounds better than the model without water.

Statistical Analysis. The models obtained had two latent variables, in all four cases. The first PLS component explains, in part, the binding capabilities and the structural differences between the 1SBN, 2SNI, 3SIC, and 5SIC complexes, compared to those other complexes having lower inhibition constants. The second component explains the remaining dependent variable for the rest of the compounds.

To simplify the analysis and without considering the singularities for the already-mentioned compounds, a model was computed without these complexes, and by excluding various descriptor blocks originating from the ligand, the receptor or the ligand–receptor complex (Table 5). The models with only one block have the worse behavior, showing that information from both interaction partners, i.e., ligands and receptors, is required in order to derive acceptable models. This can be reasoned by considering that interaction fields calculated around the ligand do not consider which functional groups interact with the receptor, and to what extent. Nevertheless, when the descriptors are derived from the ligand, the receptor and/or the complex, PLS models show significantly improved statistical parameters (Table 5). This behavior was observed when the PLS model was applied to the entire set (data not shown).

On Model Interpretation. Before analyzing the models, it is important to define and to state the interpretation of the various interaction fields that have been considered:

(a) Interaction fields computed around the ligand represent all the possible interactions with the environment according to the GRID potential. However, when a ligand interacts with a receptor, only some of these interactions are actual in the binding process. Only some chemical groups of the ligand will be correctly interfaced to their receptor counterparts.

(b) In the case of the interaction fields calculated around the receptor, the information is conceptually similar to that extracted from the ligand, but it pertains

to the other interaction partner (i.e., water, counterions, or ligands). Thus, receptor-derived interaction fields represent all the possibilities for a chemical group of the receptor to interact with other partners, but as previously mentioned not all of these interactions become actual.

(c) The ligand–receptor interaction fields show regions where there is, potentially, a mismatch between the ligand and receptor. They can also show where a water molecule might bridge them.

Another important factor that should be considered when interpreting the models is the meaning of the different VolSurf descriptors (Table 3). To simplify the interpretation, the interaction types are classified according with their GRID potential energy:

(a) Steric interaction: between 0.0 and –3.0 kcal/mol.

(b) Hydrogen bond interaction (neutral species): between –3.0 and –10.0 kcal/mol.

(c) Charge–charge interaction (including hydrogen bonds): between –10.0 and –15.0 kcal/mol.

All these terms are calculated for the probes used, but they are weighted in different ways depending on the probe and/or atom interaction partner. So, the H₂O probe considers the steric, the hydrogen atom, and the charge interaction of a water molecule to the ligand and not only the hydrogen bond capabilities of the compound.

These above energy cutoff values are not exact. Rather, they serve as guidelines to simplify the model interpretation. For illustrative purposes, we present the water-probe GRID fields for the ligand, receptor, and complex for a high- and a low-affinity compound, 2TMN and 1HIV, respectively (see Figure 4).

Furthermore, when interpreting these models, one has to recall the different physicochemical processes that occur during binding:¹¹ (a) desolvation of the ligand; (b) desolvation of the receptor cavity; (c) freezing the conformational, rotational and translational degrees of freedom; (d) forming the new interactions between the ligand and receptor.

We have shown that VolSurf parameters correlate well with the free energy of solvation.³⁶ Since only one conformer was used in the VolSurf calculation, it is unlikely that VolSurf captures the entropic aspects that are involved, e.g., the loss of the degrees of freedom when transferring from the aqueous phase to the gas phase. However, the large majority of thermodynamic aspects (other than enthalpy), and indeed kinetic aspects, are not well captured by current molecular modeling methods (VolSurf included). One can only hope that soft models (such as QSAR) can compensate for the poorly understood aspects of ligand–receptor interaction. These models can evaluate factors that are not directly calculated by a correlation to the experimental data.

VolSurf-Based Binding Affinity Prediction for a Diverse Set of Targets. The PLS pseudo-coefficients (representing the influence of each variable into the model) for the neutralized model without water molecules are shown in Figure 5. For clarity, the influence of the different descriptors discussed separately for the ligand, the receptor, and the complex. We note that, with enough data, different trends are likely to be observed for receptor binding sites where polar interac-

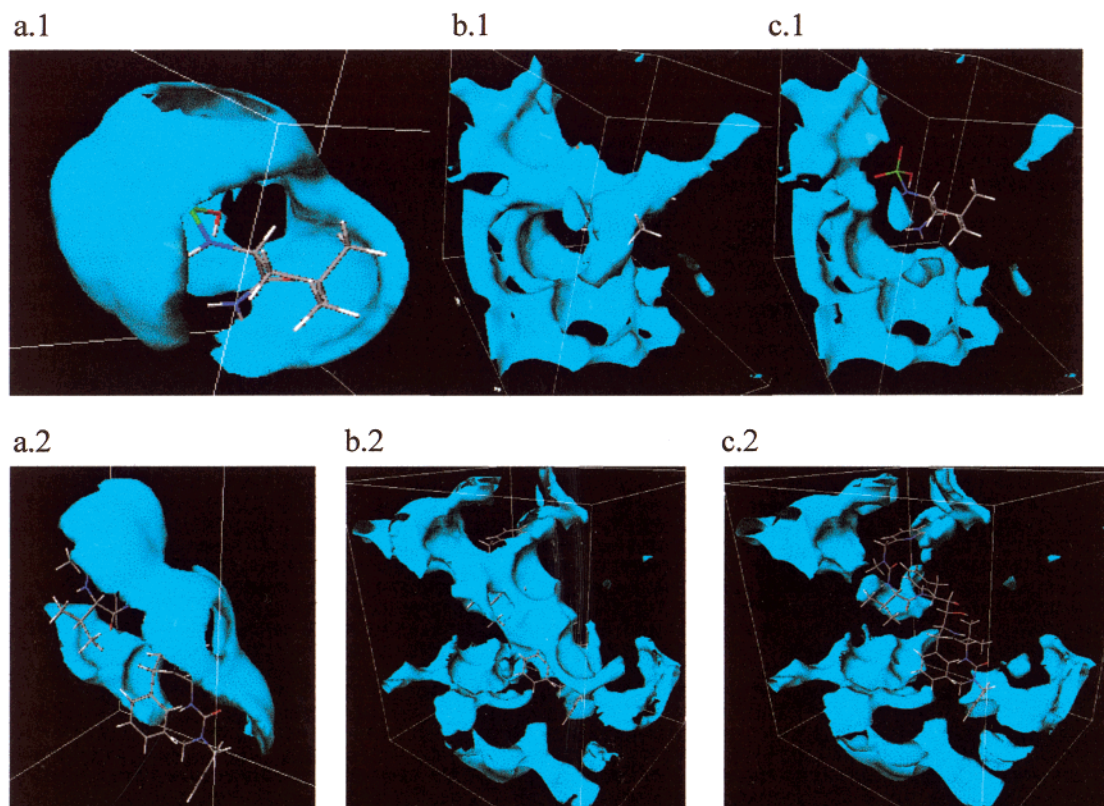


Figure 4. Interaction between the water probe and the ligand (a), receptor (b), and complex (c) for 2TMN (1), 1HIV (2) at -3 kcal/mol.

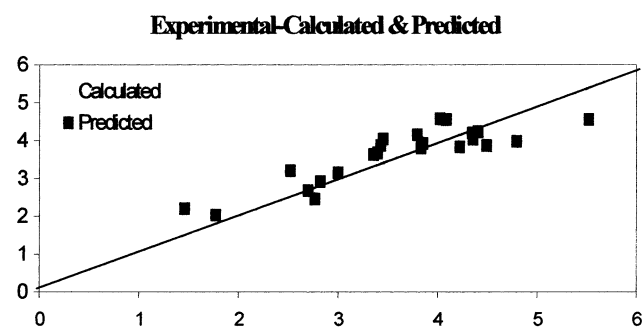


Figure 5. Experimental vs predicted values for the two sets of glycogen phosphorylase inhibitors.

tions are dominant (e.g., arabinose binding protein), compared to receptor binding sites where hydrophobic interactions play a major role (e.g., HIV protease). Below we provide the interpretation, as amenable to the entire VALIDATE data set.

Ligand. PLS coefficients computed using the water and polar probes (W1–W15, Wp1–Wp9, and HB1–HB15) have a positive contribution to binding: The greater the water and polar interaction volumes are, the better the binding affinities are, for this dataset. As all the levels for the water probe have the same sign, all the different interaction types (steric, H-bond, and electrostatic) contribute in the same manner. The negative PLS coefficients for CW1–CW6 indicate that a large ratio between polar regions and the total surface of the molecule is detrimental: A high volume of interaction per surface unit for the steric and H-bond (from 0 to -10 kcal/mol) interactions has thus a negative influence on the binding affinity. A small compound with high CW1–CW6 values would have large hydrophilic regions

in a small surface, resulting in a high desolvation free energy that would consequently decrease the binding. The hydrophobic interactions have a positive coefficient for the D1–D13 level; this implies that larger regions of hydrophobic interaction in the ligand improve the affinity.

Receptor. The coefficients for the descriptors calculated for the receptor follow similar trends to the ligand-based coefficients, except for the water-probe based volumes (W8–W15) and capacity factors (CW1–CW15). The large negative coefficients of the W8–W15 and CW8–CW15 descriptors are consistent with the following interpretation: The displacement of tightly bound waters that are present in protein cavities has a negative effect on the binding affinity. The positive PLS coefficients for CW1–CW7, that are consistent with high affinity, are more likely to be related to loosely bound water molecules that are also present in protein cavities.

Complex. The coefficients for the descriptors calculated for the ligand–receptor complexes follow the same trends as the receptor-based coefficients. Hence, the interpretation of the variables is similar to the one for the receptor case. From Table 5, it is apparent that the complex-derived descriptor block is redundant, i.e., that significant PLS models can be achieved by using just the ligand and the receptor variable blocks. Therefore, one might be able to derive predictive models without having to establish the bound conformations, or without doing the docking of the ligand into the binding site. This is not the case, since the VolSurf models use the receptor-bound conformation for the ligand and the ligand-bound conformation for the receptor, hence the apparent redundancy in Table 5.

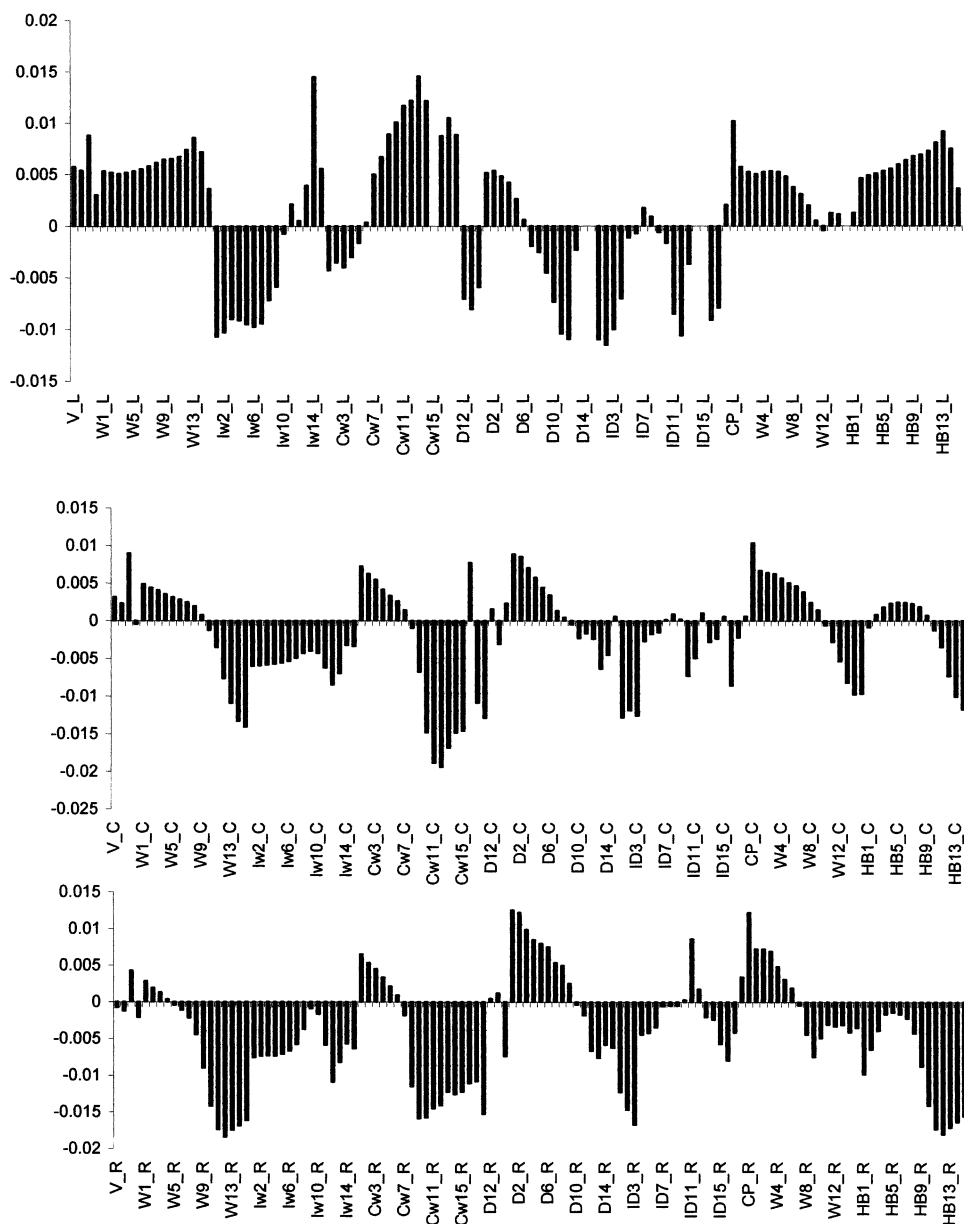


Figure 6. Partial least squares pseudo-coefficients (two components) using the Validate data set.

Glycogen Phosphorylase-b Set. Molecule Preparation. The proteins were prepared as described above for the nonneutralized, no water case. VolSurf descriptors for the ligand, receptor, and complex were computed using the H₂O, DRY and O probes in a box that exceeded the maximum and minimum coordinates of each ligand by 5 Å. No water or charge study was conducted on this set of compounds. Twenty-three different proteins of similar size and binding mode were selected from the original dataset.²⁵

Quantitative Structure-Binding Relationships. Plots comparing predicted and calculated vs experimental values are shown in Figure 6. A statistically significant model ($r^2 = 0.94$, $q^2 = 0.89$) of the inhibition constants was obtained after variable selection (Figure 7) (the statistical parameters before variable selection were appropriate to use this technique $q^2 > 0.3$) was performed using the FFD technique, as implemented in GOLPE (ratio combination = 5; number of components = 2).^{27,28}

VolSurf-Based Binding Affinity Prediction for a Single Target. Glycogen phosphorylase-b is an enzyme that has a very hydrophilic binding pocket, since it has evolved to bind mainly polar ligands, i.e., glucose derivatives. This explains why many of the variables in the hydrophobic interaction have been eliminated during the variable selection process. It can be seen that the contribution of each descriptor differs significantly from the results obtained from the previous dataset, which contains targets with mostly hydrophobic binding pockets. As it often happens in homologous analogue series, those descriptors that determine the size of the binding site and the steric interaction are not relevant to this PLS model. In contrast, descriptors related to the polar interaction (ability to accept H-bonds) are very important for these dataset, having a negative contribution to the global polar interaction and a positive contribution to the hydrogen bond acceptor capabilities, as shown in Figure 6.

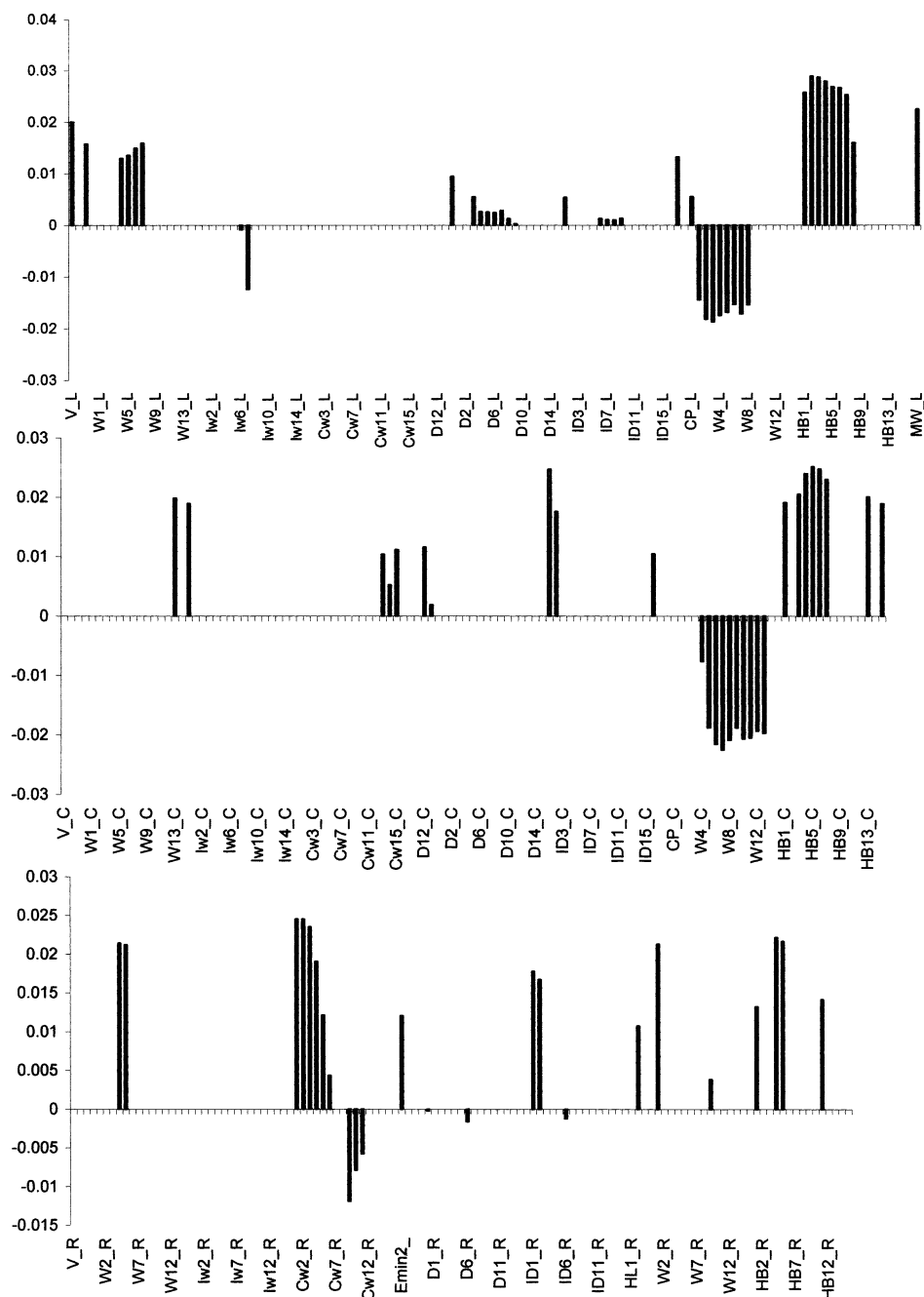


Figure 7. Partial least squares pseudo-coefficients (two components) after variable selection.

Conclusions

With the goal of optimizing both the receptor binding affinity and the pharmacokinetic properties of ligands⁹ during drug discovery-oriented lead optimization (Figure 1B),¹⁰ we have chosen to estimate the suitability of a descriptor set initially established for modeling pharmacokinetic properties, with respect to binding affinity. This integrated framework, based on GRID and VolSurf, was utilized to estimate the binding affinity for a diverse set of 38 ligand–protein complexes, as well as for a set of 23 glycogen phosphorylase-b inhibitors. A database of 195 ligand–protein complexes, derived from 51 targets, has recently been compiled.³⁷ These are worth investigating in the manner described here.

For binding affinity models using VolSurf, we established that information from both the ligand, as well

as from the receptor and/or the corresponding ligand–receptor complex, is required. There was little or no influence from the water molecules in the VolSurf derived models. However, since the data set is diverse and the positions of the water molecules are not always defined in the X-ray structures, it is not possible to generalize these conclusions outside the scope of these dataset. Furthermore, the interpretation of GRID-based interaction fields, as quantified in the PLS models, can be used to understand the factors that govern the ligand–receptor interaction not only in the context of binding, but in the context of pharmacokinetic properties as well. Being alignment-independent, this strategy has the advantage that various chemical possibilities can be explored for each series of ligands. The disadvantage of this methodology is that model interpretation

in terms of chemical groups is more complex compared to typical 3D-QSAR analyses, since the structure of the complex and the receptor are also needed. However, this approach is, to our knowledge, the first to demonstrate that it is possible to integrate both binding affinity and passive pharmacokinetic properties such as passive permeability and solubility, at the molecular descriptor level.

References

- (1) Palm, K.; Luthman, K.; Ungell, A.-L.; Strandlund, G.; Artursson, P. Correlation of drug absorption with molecular surface properties. *J. Pharm. Sci.* **1996**, *85*, 32–39.
- (2) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3–25.
- (3) Van de Waterbeemd, H.; Camenisch, G.; Folkers, G.; Raevsky, O. A. Estimation of Caco-2 cell permeability using calculated molecular descriptors. *Quant. Struct.-Act. Relat.* **1996**, *15*, 480–490.
- (4) Stenberg, P.; Luthman, K.; Artursson, P.; Prediction of membrane permeability to peptides from calculated dynamic molecular surface properties. *Pharm. Res.* **1999**, *16* (2), 205–212.
- (5) Palm, K.; Luthman, K.; Ungell, A.-L.; Strandlund, G.; Beigi, F.; Lundahl, P.; Artursson, P. Evaluation of dynamic Polar Surface Area as predictor of drug absorption: comparison with other computational and experimental predictors. *J. Med. Chem.* **1998**, *41*, 5382–5392.
- (6) Oprea, T. I.; Gottfries, J. Toward Minimalistic modeling of oral drug absorption. *J. Mol. Graphics Mod.* **1999**, *17*, 261–274.
- (7) Zamora, I.; Ungell, A.-L. Correlation between drug absorption and molecular surface descriptors: Comparison between different experimental models. *Eur. J. Pharm. Sci.*, submitted.
- (8) Oprea, T. I.; Zamora, I.; Ungell, A.-U. Pharmacokinetic based mapping device for chemical space navigation. *J. Comb. Chem.* **2002**, *4*, in press.
- (9) Oprea, T. I.; Zamora, I.; Svensson, P. Quo Vadis, Scoring Functions? Toward an Integrated Pharmacokinetic and Binding Affinity Prediction Framework. In *Combinatorial Library Design and Evaluation for Drug Design*; Ghose, A. K., Viswanadhan, V. N., Eds., Marcel Dekker Inc.: New York, 2001; pp 233–266.
- (10) Oprea, T. I. Virtual screening in lead discovery: A viewpoint. *Molecules* **2002**, *7*, 51–62.
- (11) Williams, D. H.; Cox J. P. L.; Doig, A. J.; Gardner, M.; Gerhard, U.; Kaye, P. T.; Lal, A. R.; Nicholls, I. A.; Salter, C. J.; Mitchell, R. C. Toward the semiquantitative estimation of binding constants. Guides for peptide-peptide binding in aqueous solution. *J. Am. Chem. Soc.* **1991**, *113*, 7020–7030.
- (12) Böhm, H. J. The development of a simple empirical scoring function to estimate the binding constant for a protein–ligand complex of known three-dimensional structure. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 243–256.
- (13) Oprea, T. I.; Waller, C. L. Theoretical and Practical aspects of three-dimensional Quantitative Structure–Activity Relationships. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; Wiley-VCH: New York, 1997; Vol 11, pp 127–182.
- (14) Oprea, T. I.; Marshall, G. R.; Receptor-Based Prediction of Binding Affinities. *Perspect. Drug Discovery Des.* **1998**, *9*, 35–61.
- (15) Böhm, H. J.; Stahl, M. Rapid empirical scoring functions in virtual screening applications. *Med. Chem. Res.* **1999**, *9*, 445–462.
- (16) Muegge, I.; Martin, Y. C. A general and fast scoring function for protein–ligand interactions: A simplified potential approach. *J. Med. Chem.* **1999**, *42*, 791–804.
- (17) Gohlke, H.; Hendlich, M.; Klebe, G. Knowledge-based scoring function to predict protein–ligand interactions. *J. Mol. Biol.* **2000**, *295*, 337–356.
- (18) Cruciani, G.; Crivori, P.; Carrupt, P.-A.; Testa B. Molecular Fields in Quantitative Structure-Permeation Relationships: The VolSurf Approach. *J. Mol. Struct. (THEOCHEM)* **2000**, *503*, 17–30.
- (19) Cruciani, G.; Pastor, M.; Clementi, S. Handling information from 3D grid maps for QSAR studies of bioactivity. Gundertofte, G., Jogerssen, F. E., Eds.; Kluwer Academic Plenum Publishers: New York, 2000; pp 73–82.
- (20) VolSurf version 2.0 is available from Molecular Discovery Ltd., West Way House, Elms Parade, Oxford OX2 9LL, UK, www.moldiscovery.com.
- (21) Goodford, P. J. Computational procedure for determining energetically favourable binding sites on biologically important macromolecules. *J. Med. Chem.* **1985**, *28*, 849–857.
- (22) GRID available from Molecular Discovery Ltd, www.moldiscovery.com.
- (23) Head, R. D.; Smythe, M. L.; Oprea, T. I.; Walter, C. L.; Green, S. M.; Marshall, G. R. VALIDATE: A new method for the receptor-based prediction of binding affinities of novel ligands. *J. Am. Chem. Soc.* **1996**, *118*, 3959–3969.
- (24) Watson, K. A.; Mitchell, E. P.; Johnson, L. N.; Cruciani, G.; Son, J. C.; Bichard, C. J. F.; Fleet, G. W. J.; Oikonomakos, N. G.; Kontou, M.; Zographos, S. E. Glucose Analogue Inhibitors of Glycogen Phosphorylase: from Crystallographic Analysis to Drug Prediction using GRID Force-Field and GOLPE Variable Selection. *Acta Crystallogr.* **1995**, *D51*, 458–472.
- (25) Pastor, M.; Cruciani, G.; Watson, K. A. A strategy for the incorporation of water molecules present in a ligand binding site into a three-dimensional quantitative structure–activity relationship analysis. *J. Med. Chem.* **1997**, *40* (25), 4089–4102.
- (26) Cruciani, G.; Crivori, P.; Carrupt, P.-A.; Testa B. Molecular Fields in Quantitative Structure-Permeation Relationships: The VolSurf Approach. *J. Mol. Struct. (THEOCHEM)* **2000**, *503*, 17–30.
- (27) Cruciani, G.; Pastor, M.; Clementi, S. Handling information from 3D grid maps for QSAR studies of bioactivity. Gundertofte, G., Jogerssen, F. E., Eds., Kluwer Academic Plenum Publishers: New York, 2000; pp 73–82.
- (28) VolSurf 2.0 available from Molecular Discovery, LD, www.moldiscovery.com.
- (29) Bernstein, F. C.; Koetzle, T. F.; Williams, G. J. B.; Meyer, E. F.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. The protein data bank: A computer-based archival file for macromolecular structures. *J. Mol. Biol.* **1977**, *112*, 535–542. The Protein Data Bank web site is <http://www.rcsb.org/pdb/>.
- (30) Ho, C. M. W.; Head, R. D.; Marshall, G. R. personal communication, 2000.
- (31) Renzoni, D. A.; Zvebil, M. J. J. M.; Lundbäck, T.; Ladbury, J. E. Exploring uncharted waters: Water molecules in drug design strategies. In *Structure-Based Drug Design. Thermodynamics, Modeling and Strategy*; Ladbury, J. E., Connelly, P. R., Eds., Berlin: Springer 1997; pp 161–180.
- (32) Connelly, P. R. The cost of releasing site-specific, bound water molecules from proteins: Toward a quantitative guide for structure-based drug design. In *Structure-Based Drug Design. Thermodynamics, Modeling and Strategy*; Ladbury, J. E., Connelly P. R., Eds.; Berlin: Springer, 1997; pp 143–159.
- (33) For example, if the net charge was 7, we used 11 minima, whereas a net charge of 3 would give 7 minima during the SA process.
- (34) Baroni, M.; Costantino, 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.
- (35) GOLPE version 4.5, MIA sraL, Perugia, Italy, 2000.
- (36) Zamora, I.; Oprea, T. I.; Norinder, U., manuscript in preparation.
- (37) Roche, O.; Kiyama, R.; Brooks, C. L., III. Ligand-Protein DataBase: Linking Protein–Ligand Complex Structures to Binding Data. *J. Med. Chem.* **2001**, *44*, 3592–3598.