Konrad-Zuse-Zentrum
für Informationstechnik Berlin

Takustraße 7
D-14195 Berlin-Dahlem
Germany

T. BAUMEISTER, F. CORDES

A new Model for the free energy of solvation and its application in protein ligand scoring

ZIB-Report 04-51 (October 2004)

A new Model for the Free Energy of Solvation and its Application in Protein Ligand Scoring

Timm Baumeister, Frank Cordes

Abstract

A new and time efficient model to evaluate the free energy of solvation has been developed. The solvation free energy is separated into an electrostatic term, a hydrogen bond term, and a rest-term, combining both entropic and van der Waals effects. The electrostatic contribution is evaluated with a simplified boundary element method using the partial charges of the MMFF94 force field. The number of hydrogen bonds and the solvent excluded surface area over the surface atoms are used in a linear model to estimate the non-electrostatic contribution. This model is applied to a set of 213 small and mostly organic molecules, yielding an rmsd of 0.87kcal/mol and a correlation with experimental data of r=0.951. The model can be applied as a supplementary component of the free energy of binding to estimate binding constants of protein ligand complexes. The intermolecular interaction energy is evaluated by using the MMFF94 force field.

Key words. scoring, solvation, free energy.

Mathematics subject classification. 92C40 ,92-08 .

Introduction

Solvation and desolvation of biomolecules play a key role for a variety of important biological processes. For lipid bilayer assembly and protein folding, as well as protein-protein and protein-ligand interactions, the structures are stabilized by a reduction of hydrophobic interface area between the molecules and water [1, 2]. Modelling these interactions has become increasingly important with the growing application of virtual screening methods for computer aided pharmaceutical drug design. Estimating the interaction energy between a target and a specific ligand (the so called scoring) is the critical part of a screening algorithm in terms of quality and speed. Different levels of complexity can been applied to the scoring problem, such as QSAR [3] and rough models of surface and chemical complementarity (Pharmacophore) [4], or knowledge based scoring functions [5]. Most widely used, however, are methods that partition the free energy of binding into physically independent contributions. These components can either be parameterized empirically or modelled with semiempririca methods like force field energy calculations. As part of a docking application developed at the Konrad Zuse Institute Berlin (ZIB) [6], the intermolecular energy between protein and ligand is calculated using the Merck Molecular Force Field MMFF94 [7]. Solvation effects are not considered explicitly in this model. As the change of solvation free energy due to desolvation of the interface area between protein and ligand can be a significant contribution to the binding energy, an additional model for this component needs to be developed.

Modelling the free energy of Solvation

Theory

Partitioning the free energy of solvation

The complex nature of the solvation process has led to the application of a wide range of different models. The free energy of solvation is commonly decomposed into an electrostatic, a van der Waals, a hydrogen-bond, and an entropic component.

Water, with its high dipole moment, serves as a constant dielectric. This electrostatic component is completely enthalpic and negative, as the reorientation of water molecules and the resulting reaction field decreases the strength of the overall field of the molecule. This component is most often described by applying continuum electrostatics and solving the poisson equation, whereby the inside of the molecule is either modelled quantum mechanically or itself by a constant dielectric [8]. Methods that employ the finite element [9] or the boundary element approach [10] have been developed. Wide application has found the generalized born method [11, 12], a

generalized combination of the born and the onsager equations which are valid for spherical ions and dipoles, respectively.

Apart from this continuum component, the remaining contributions are locally confined to the interface area between solute and solvent and are therefore also called first solvation shell components.

The prime example for the entropic component is the hydrophobic effect between hydrocarbons. At room temperature, the positive free energy is nearly exclusively due to a decrease in entropy [13]. The molecular origin of this effect is still slightly controversial [14]. The most widely accepted explanation is that water molecules have to reorientate themselves around the molecular surface to maximise the number of possible hydrogen bridges. The accompanied decrease in entropy makes a smaller contribution to the free energy of solvation than the enthalpy increase which would be required for a breakup of the hydrogen bonds between the water molecules. This model explains observations, that for nearly planar molecular surfaces 3/4 of the hydrogen bridges are kept, while geometrically only a value of 1/2 would be expected [15]. Due to the computational costs of entropies derived from thermostatistical sampling, nearly all scoring functions methods use hydrophobicity parameters like atomic solvation parameters first described by Eisenberg [16].

Hydrogen Bridges can form between water molecules and atoms of the solute. Their strength depends on the type of the atoms which take part, the angle of binding and the local dielectric constant. Usually, their strength is in the range of -10 to $-40 kJ/mol$ [17]. This enthalpy, however, does not equal the contribution to the free energy of solvation. The reason is that such hydrogens usually replace existing hydrogen bonds between water molecules that find themselves broken due to the inserted interface area with the solute. Nevertheless, they are not thermodynamically neutral, and empirical models estimate the contribution of each solute-solvent hydrogen bridge to be in the range of -2.5 to $-7.5 kJ/mol$ [18].

Van der Waals component comprises interactions between permanent and induced dipoles of the molecule. The scale of this interaction therefore depends on the polarizability of the electron density around the involved atoms and is proportional to r^{-6}. Because of the sharp decrease with distance it is usually sufficient to only consider the interactions with molecules in the first solvation shell and the strength of these interactions is roughly proportional to the size of the molecular surface.

Boundary Element Method

The boundary element method can be applied to the solvation problem if the simplification is made that only two distinct dielectric environments exist. The problem can then be reformulated on the boundary surface between these environments. By starting from Gauss's Law of electrodynamics and

the continuity principle, it can be shown that the following equation holds [10]:

$$\sigma - f \oint_s \frac{\sigma_s \cdot (\mathbf{r} - \mathbf{r_s})\mathbf{n}}{(r - r_s)^3} ds = f \sum_i^{N_{atoms}} \frac{q_i \cdot (\mathbf{r} - \mathbf{r_i})\mathbf{n}}{\varepsilon_{in}(r - r_i)^3} \tag{1}$$

where sigma is the surface charge density, q are the partial charges of the atoms of the solute, the integral is over the complete boundary surface of the solute, and f is a constant whis is

$$f = \frac{\varepsilon_{in} - \varepsilon_{out}}{2\pi(\varepsilon_{in} + \varepsilon_{out})}$$

with the dielectric constant for the inside and the outside of the boundary surface being ε_{in} and ε_{out}, respectively.
The problem of solving the Poisson equation is therefore transformed into the problem of solving the integral equation 1 for σ. With known σ, the field and the potential can be derived and the reaction field energy is

$$\Delta G_{elec} = \frac{1}{2} \sum_{i=1}^{N_{atoms}} q_i \oint_S \frac{\sigma_H - \sigma_V}{\|\mathbf{r_i} - \mathbf{r_s}\|} ds \tag{3}$$

where σ_H and σ_V are the surface charges in solvation and in the gas phase. To solve equation (1) numerically for σ, the boundary surface has to be discretised into boundary elements with the surface charge densities σ_k, the normals n_k, and the areas A_k:

$$\sigma_k - f \sum_j^{N_{BE}} \frac{\sigma_j \cdot (\mathbf{r_k} - \mathbf{r_j})\mathbf{n_k}}{(r_k - r_j)^3} A_j = f \sum_i^{N_{atoms}} \frac{q_i \cdot (\mathbf{r_k} - \mathbf{r_i})\mathbf{n_k}}{\varepsilon_{in}(r_k - r_i)^3} \tag{4}$$

If N_{BE} is the number of boundary elements, equation (4) yield a system of N_{BE} linear equations for σ_k.

$$(\mathbf{I} - f\mathbf{K})\,\boldsymbol{\sigma} = \mathbf{e} \tag{5}$$

where $\mathbf{I}$ is the unity matrix, $\boldsymbol{\sigma}$ is the vector of the surface charge densities, $\mathbf{e}$ is the vector of the field caused by the atomic partial charges, and $\mathbf{K}$ is the coefficient matrix. The elements of $\mathbf{e}$ are :

$$e_k = f \sum_i^{N_{atoms}} \frac{q_i \cdot (\mathbf{r_k} - \mathbf{r_i})\mathbf{n_k}}{\varepsilon_{in}(r_k - r_i)^3} \tag{6}$$

The coefficients of $\mathbf{K}$ are:

$$K_{kj} = f \frac{(\mathbf{r_k} - \mathbf{r_j})\mathbf{n_k}}{(r_k - r_j)^3} A_j \tag{7}$$

This equation if evidently not valid for the diagonal elements ($i = j$). Zauhar [10] set these elements to zero. A more robust semianalytical approach has been introduced by Purisima [19] where the diagonal elements are a linear combination of nondiagonal elements:

$$K_{kk} = 2\pi - \sum_{j \neq i}^{N_{BE}} K_{jk} \frac{A_j}{A_k} \tag{8}$$

For equations (6)-(8) all variables are known and thus the system of equations (5) can be solved for $\boldsymbol{\sigma}$.

Atomic Solvation Parameter

The similarity between the macroscopic effect of surface tension and the microscopic hydrophobic effect, has let to a application of surface tensions for evaluating solvation energies [20, 21]. Eisenberg and McLachlan [16] first introduced hydrophobicity parameters on the basis of atomic surfaces. The solvation free energy is obtained by multiplying an atom type specific parameter with the surface area which are associated with the respective atom types.

$$\Delta G_{solv} = \sum_{i=1}^{numAtomTypes} ASP(atomtype_i) \cdot SASA(atomtype_i) \tag{9}$$

The parameters are determined by linear regression with experimental data.

The original idea behind atom based parameters was that they can account for atom specific effects, including electrostatic effects. Although this has proven to be only partially true, atom based parameters still have another important advantage. As the vdW radii of the atoms which are needed to compute the surface areas are ambiguously, a parameter per atom type can rescale the area and thus act as an empirical correction to the experimentally determined van der Waals radii which are used for surface calculation.

Molecular Surface

The previously described methods require the molecular surface for computation. There are three common definitions of the boundary surface.

The van der Waals (vdW) surface of a molecule is the outer hull of the intersecting atomic spheres whose dimension is defined by the van der Waals radius of the respective atoms. The other two surface definitions take the dimension of the solvent molecules into account. The solvent is modelled as a spherical molecule. For water this is an acceptable approximation with the van der Waals radius of a water molecule being $r \approx 1.4\mathring{A}$.

The definition of the solvent accessible surface (SAS) [22] is equal to that of the van der Waals surface except that all spherical radii are expanded by

the van der Waals radius of the solvent molecule. The computed surface describes the center of the first solvation shell.

The solvent excluded surface (SES, also called 'Connolly Surface' or ' Molecular Surface') [23] is the contact surface if a sphere with the van der Waals radius of the solvent molecule is 'rolled' over the van der Waals surface of the solute molecule. The inner volume of this surface is therefore the volume into which no part of a solvent molecule can intrude.

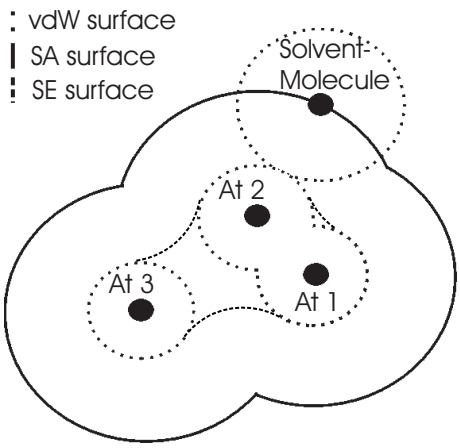


Figure 1: schematic representation of vdW, SA und SE surfaces

Methodology

For parameterizing, testing, and refining our models we used a set of 213 small and mostly organic molecules, whose free energy of solvation is experimentally known. The set is a subset of the molecules that were used by Chambers to parameterize his SMX models [24].

For the computation of the electrostatic component the boundary element method, as described in the previous section, has been used.

The linear system of equations is solved by carrying out Gauss-Seidel iteration. The convergence criterion was a deviation of less than $5cal/mol$ between two successive iterations, except if stated otherwise.

For the set of small molecules this method was well applicable. For surfaces of binding site dimensions, however, the demand of computing time becomes too large. We therefore simplified the method by using only a single iteration, which means that the mutual influence among the water dipoles is neglected. The resulting field energy is always more negative than the real value. To obtain the reaction field energy, the value is rescaled with a constant factor. This simplification assumes that the effect of influence of the dielectric onto itself scales with the size of the molecular field. To test this hypothesis we used the set of small molecules to compare the results of our method with the exact result when using unlimited iterations with a convergence criterion. The method uses a full atom model of the molecule. Both vdW and SE surfaces have been tested. For the final model the SE surface was chosen to represent the boundary surface.

First solvation shell effects were separated into a term which accounts for hydrogen bonds and a rest-term. Both terms use a unified atom model. The hydrogen bond term relates the number of hydrogen donors and acceptors

which have a nonzero solvent excludes surface to the respective free energy by assuming a linear relationship. Hydrogen donor and acceptor atoms are defined by the Merck Molecular Force Field. The remainder of the solvation free enthalpy includes both the entropic component and the van der Waals component. Both are proportional to the surface area and can thus be approximated with a set of atomic solvation parameters. As surface representation the solvent accessible surface has been used.

Hence, the complete solvation free energy is evaluated as a combination of three different processes:

1. Evaluate the reaction field energy with the simplified BE method and rescale it with a constant factor
2. Compute the number of hydrogen donors and acceptors and rescale it with a constant factor.
3. Evaluate the energy as given by the ASP set.

Two surfaces have to be generated in the process. The boundary element method requires the triangulated solvent excluded surface using a complete atom model. The other two methods require the analytical solvent accessible surface using a unified atom model. The surfaces were generated with the help of the visualization and molecular modelling systems Amira and AmiraMol [25].

The individual components of the model were first tested separately to gain more information about the applicability and problems of each single model. Quality was measured by Pearsons correlation r with experimental data. All linear regressions were carried out through the origin.

Results and discussion

BE method

The implementation of our BE method was tested by comparing the results with theoretical values. For a charge at the center of a spherical cavity, the energy can be computed with the Born equation [26]. The charge was set to one, the triangle edge length to $0.2\mathring{A}$, and the dielectric constants to $\varepsilon_{out} = 78.5$ and $\varepsilon_{in} = 1$. The radius was increased in steps of $0.1\mathring{A}$ in the range between $1.0\mathring{A}$ and $2.0\mathring{A}$. The relative error of the reaction field energy was 0.66%. To put this result into context we used the finite elements Poisson-Boltzmann solver APBS [27]. With the parameterization that was delivered as part of the package we obtained a relative error of 2.40%. The main disadvantage of the finite element method is that the discretization of the problem includes a set parameters for describing the grid box size, the spacing, and boundary conditions. For the BE method only the triangle

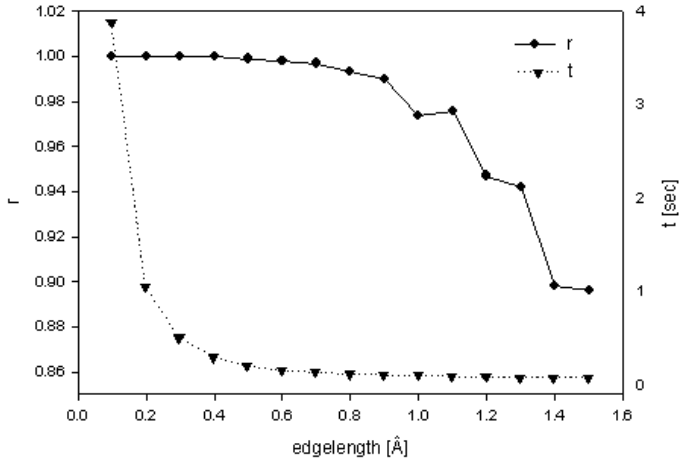


Figure 2: Mean CPU time per molecule t (in sec) and correlation coefficient r in relation to the triangle edge length (in $\mathring{A}$) for the ZIBBE Method.

edge length (or the related 'points per area') has to be used which is directly related to the accuracy of the method.

The time complexity of the BE method depends on the number of boundary elements N_{BE}: The evaluation of the matrix has the complexity $O(N_{BE}^2)$ which is the same as that of the Gauss-Seidel iteration. The memory demand for storing the matrix is $4 * N_{BE}^2$ if single precision floats are used. For a globular molecule with radius r the surface, and thus N_{BE}, increases with r^2. Both computing time and memory usage thus grow with r^4 which underlines the necessity to optimize the discretization of the surface. To get a reasonable estimate for the necessary accuracy of the surface triangulation, we used our test set of small molecules and evaluated the energy for edge lengths of the triangles ranging from 0.1 to 1.5 $\mathring{A}$. The resulting correlation between the respective data set and the data set with the highest accuracy can be considered to be a measure of quality and is shown in Figure 2. Also shown is the required time in seconds per molecule. For our test set, the correlation is nearly perfect for edge-lengths up to 0.5 $\mathring{A}$ and starts to drop off significantly for edge lengths over 1.0 $\mathring{A}$. A reasonable value should be chosen in this range. For out further use of the BE and the simplified BE method we have chosen an edge length of 0.7 $\mathring{A}$, unless stated otherwise.

Our simplified BE method is based on the assumption that the effect of influence of the dielectric onto itself scales with the size of the molecular field. To test this hypothesis we used the set of small molecules to compare the results of our method with the exact result when using unlimited iterations with a convergence criterion. We found a correlation between the two sets of results of r=0.996 which is sufficiently high to justify the simplification.

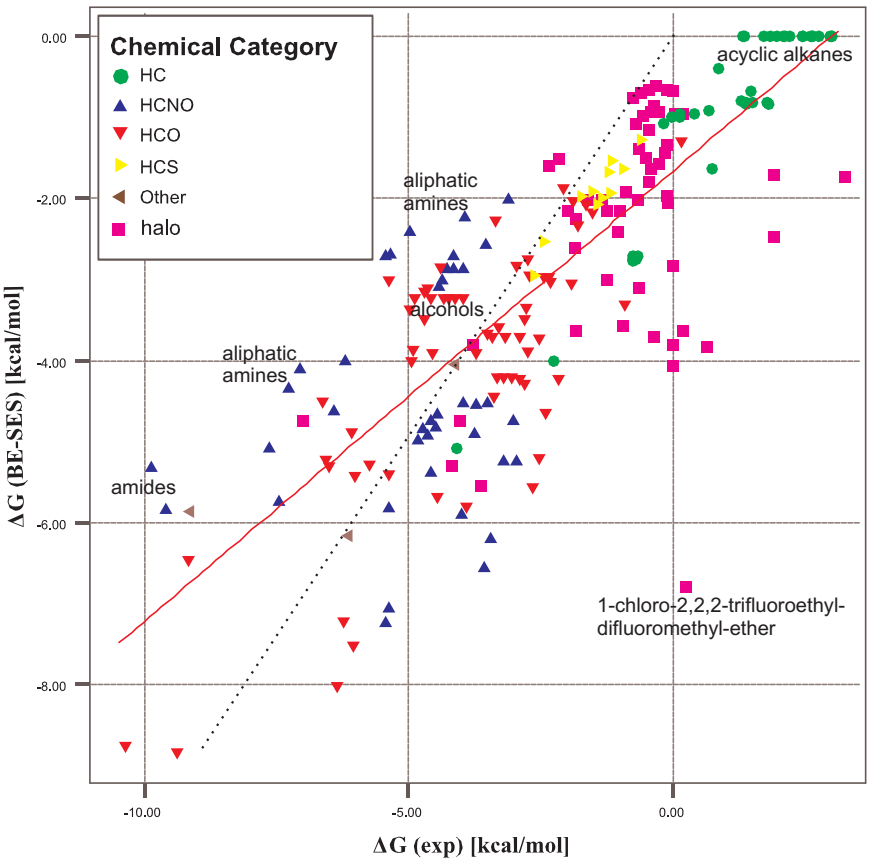


Figure 3: Correlation between experimental free energy of solvation and the value obtained with the ZIBBE method.

As an alternative simplification of the BE method, we tested the idea of Totrov to join all triangles of an atom into a single boundary element [28]. Our test for 213 small molecules only resulted in a correlation of 0.904 with the values of the complete method.

The method was applied with a triangle edge length of $0.7\mathring{A}$, a dielectric constant for the interior of the molecule of $\varepsilon_{in} = 2$, and for the outside of the molecule of $\varepsilon_{out} = 78.3$. Convergence criterion was a deviation of ΔG from the last iteration of less than $5cal/mol$.

The boundary element method only accounts for the electrostatic component of the solvation free energy. In spite of this limitation there should still be a clearly visible correlation between the electrostatic component and the experimental values for the free energy of solvation. Figure 3 shows the scatter plot of the relation and the obtained correlation with experimental date is 0.812 for both the full and the simplified BE method.

One of the problems of the BE method is the ambiguity of the molecular surface. Most methods in literature use the vdW surface as boundary

surface, mainly because of the simplicity of its computation. Our results show that the method yields better results for the SE surface. When using the vdW surface, we obtain a correlation coefficient of only $r = 0.775$ for both methods. The cause for this significant difference can most likely be found in small crevices that occur in the vdW surfaces. If these crevices are smaller than the diameter of a water molecule, they should be considered to be part of the solute rather than the solvent, which is only the case when using the SE surface.

ASP

Many molecular modelling packages use atomic solvation parameters as the only method for evaluating the free energy of solvation. For very restricted sets of molecules this can be quite successful. As an example, Cramer showed in his review that for a set of 26 very simple organic molecules the ASP set of Ooi et al. had a correlation of r=0.967 with experimental data. We evaluated the values anew with our own ASP implementation and Oii's parameters and found a similar result of r=0.969. To judge the transferability of this ASP set, we extended our test to the complete test set except those molecules which contain elements not included in Ooi's ASP set. For these 159 molecules we obtained a correlation of 0.482. The model yields unsatisfying results especially for highly charged species like aldehydes and nitro-hydrocarbons. For aldehydes we obtain strong positive deviations. This is due to the high ASP values for acylic carbons of 427 $cal/mol\mathring{A}^2$. Carbons in this position carry a high partial charge and thus one would, expect from a physical point of view, that their asp should be negative to account for the electrostatic component. Nitrohydrocarbons on the other hand show a strong negative deviation. This is due to the strongly negative ASP for oxygens and nitrogens. It can be concluded that electrostatic effects of oxygens and nitrogens are weighted too highly in this parameterization, an error which is partly compensated by the absurdly high parameter for acylic carbons.

Atomtyp	r	ASP
CH	1.55	-2.1
CYL	1.75	21.4
CA	2.00	-9.9
N	1.55	-86.2
OYL	1.4	-76.5
OH	1.4	-73.7
S	2.00	-12.3
F	1.46	4.2
CL	1.76	-3.2
BR	1.87	-7.0
I	2.03	-5.7

Table 1: Parameterization of the ZIBASP1 model. Radii in $\mathring{A}$ and ASP in $cal/mol\mathring{A}^2$

We parameterized a new ASP set with our complete set of 213 molecules by supplementing the 7 Oons atoms types with four halogen atom types for Fluorine, Chlorine, Bromine, and Iodine. For each of these types the SAS was computed and the results were fitted with experimental data by using

equation 9 for the regression through the origin

$$\Delta G_{solv} = \sum_{i=1}^{11} ASP(atomtype_i) \cdot SASA(atomtype_i) \tag{10}$$

The parameters $ASP(atomtype_i)$ are shown in table 1 and the results of the fit are shown in figure 4.

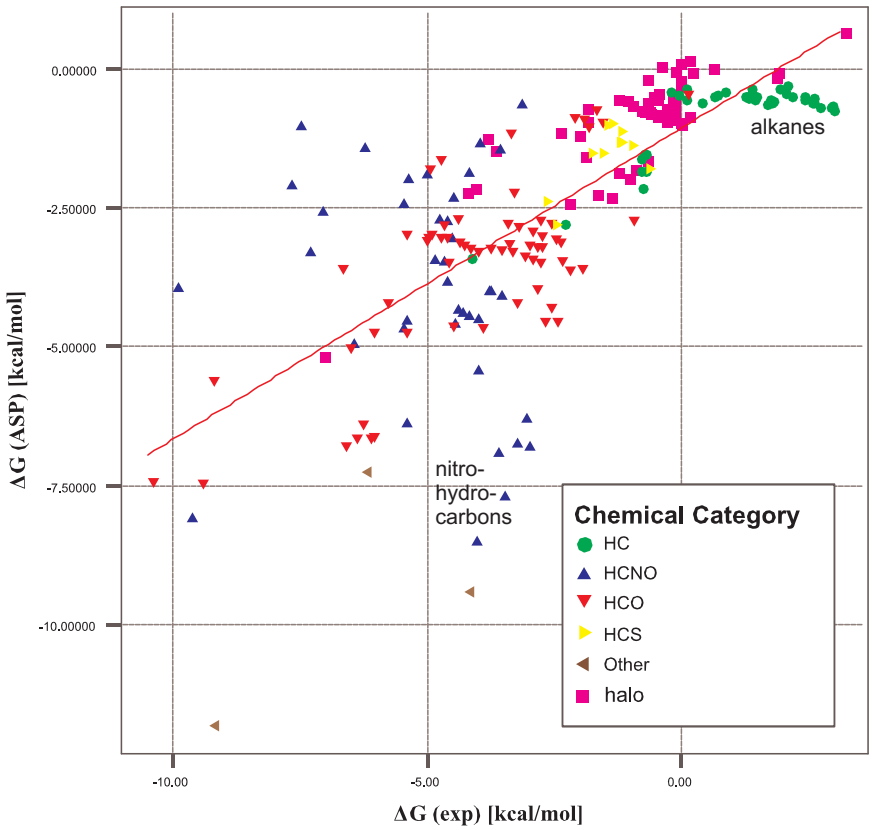


Figure 4: Correlation between ZIBASP1 result and experimental free energy of solvation

The correlation of this parameterization with experimental data is r=0.746. The high ASP for acylic carbons of the oons parameterization has been decreased by about factor 20, while on the other hand the quality for uncharged species has decreased.

Combined Model

Our final model combines the two methods described in the previous sections. As atom type definitions for the ASP set, all 29 atom types of the Merck force-field were used that were contained in the test set. The basis for the computation of the molecular surface can be the complete molecule or a unified atom model where hydrogen atoms are neglected. A complete atom model has the advantage that the effect of hydrogen bonds can be modelled explicitly with the parameters for the involved hydrogen atoms while for a complete atom model this has to happen either over the associated donors and acceptors or by adding another explicit hydrogen bond term. We found that the disadvantage of the complete atom model is a high sensitivity to conformational changes. As most surface area of the heavy atoms is covered by hydrogens, any conformational change concerning the hydrogens can increase these areas by a high factor, thus leading to a lack of stability for the associated parameters.

	Koef.
PCM	0.721
h-num	-2.511
C	0.010
C=O	0.054
CB	0.001
N	0.286
O	0.177
F	0.153
S	0.194
Cl	0.132
Br	0.082
I	0.071

Table 2: ZIBSM1 parameterization

We found that the best correlation with the least number of parameters could be reached when using a unified atom model together with an explicit consideration of hydrogen bridges. Therefore, we used the number of atoms defined as hydrogen donors and acceptors by the force-field which had a nonzero surface area and used it as a descriptor in our linear model. We reduced the number of atom types by analyzing the mutual correlation of surface areas and similarities between the parameters. The final model includes 10 atom types: Carbonyl carbons (C=O), aromatic carbons (CB), other carbons (C), as well as nitrogen, oxygen, fluorine, sulfur, chlorine, bromine, and iodine.

Element	r
H	1.30
C	1.90
N	1.50
O	1.50
F	1.75
S	1.60
Cl	1.60
Br	1.75
I	1.95

Table 3: Optimal van der Waals radii

$$\begin{aligned}\Delta G_{solv} = &\sum_{i=1}^{10} ASP(atomtype_i) \cdot SASA(atomtype_i) \\ &+ p(hNum) \cdot hNum \\ &+ p(\text{ZIBPCM}) \cdot E_{\text{ZIBPCM}}.\end{aligned}$$

The van der Waals radius of the elements was optimized by analyzing the correlation with experimental data in dependance of the vdW radius of the individual atoms. The radii were increased in steps of $0.05\mathring{A}$. Table 3

shows the results.

Table 2 shows the resulting parameterization of the complete model. Figure 5 shows the correlation between the predicted values and the experimental values in a scatterplot.

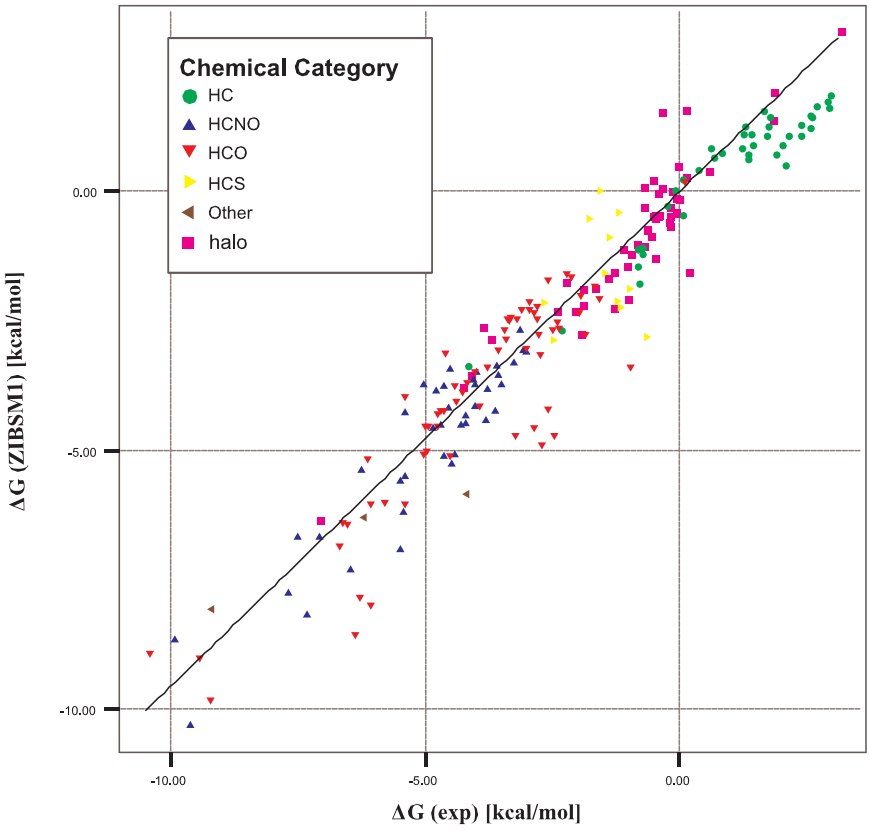


Figure 5: Solvation free energy, as predicted with the ZIBSM method, plotted against experimental values

Solvation model as a scoring function

Theory

Free energy of binding

The strength of association between a protein and a ligand can be described thermodynamically by an association constant k_a or the related inhibition and dissociation constants k_i and k_d.

$$k_a = k_i^{-1} = k_d^{-1} = \frac{[\text{P'L'}]}{[\text{P}]\,[\text{L}]} \tag{11}$$

The free energy of binding can be derived as

$$\Delta G_{bind} = -RT \cdot \ln k_a \tag{12}$$

Another related measure is the binding affinity A which is defined as $A = -\Delta G$.

The change of the binding free energy can be driven by a decrease of enthalpy or an increase in entropy.

$$\Delta G_{bind} = \Delta H - T\Delta S \tag{13}$$

The free energy is a state function and can be divided into arbitrary partial processes. This is used to separate ΔG into components which can be modelled separately.

1. Intermolecular interaction: Decrease of enthalpy due to electrostatic and vdW interaction between the molecules
2. Intramolecular energies: Increase of enthalpy due to a change of internal energies of the molecules during binding.
3. Solvation component: Increase of entropy of water due to desolvation of the interaction surface between the two molecules as well as a decrease in enthalpy due to a decrease of electrostatic and dispersive interactions between the molecules and the solvent.
4. Conformational Entropy: Decrease of entropy due to complexation and the related decrease of of freedoms of translation, rotation and vibration as well as a freezing of bond torsions.

Using this model $\Delta G_{binding}$ can be written as

$$\Delta G_{binding} = \Delta G_{interaction} + \Delta G_{conformation} + \Delta G_{solvent} + \Delta G_{entropy} \tag{14}$$

The components $\Delta G_{interaction}$ and $\Delta G_{conformation}$ can be estimated by the minimum energy conformation of a force field approach under the assumption that the energetic minimum conformation is a good statistical representative for the conformational ensemble. $\Delta G_{solvent} + \Delta G_{entropy}$ need to be modelled separately. In the following section we will neglect $\Delta G_{entropy}$.

Evaluation of the free energy of binding in solvation

The free energy of binding in solvation ΔG^S_{bind} can be decomposed into contributions which correspond to the in-vacuo binding ΔG^V_{bind} and the solvation of the seperate molecules and the complex.

Figure 6 show the dependance of ΔG^S_{bind} on the components which can be evaluated directly. L^V, L^S, P^V and P^S stand for the ligand and the protein in vacuum and in their solvated form. LP^V and LP^S denote the complex in the two environments.

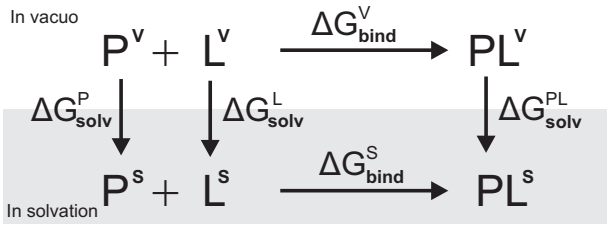


Figure 6: Berechnung von ΔG^S_{bind}

If it is possible to evaluate the solvation free energy for the Protein ΔG^P_{solv}, the ligand ΔG^L_{solv} and the complex ΔG^{LP}_{solv}, all terms are known to evaluate ΔG^S_{bind}:

$$\Delta G^S_{bind} = \Delta G^V_{bind} + \Delta G^{LP}_{solv} - \Delta G^P_{solv} - \Delta G^L_{solv} \tag{15}$$

Methodology

For parameterization we used a set of 66 protein ligand complexes of the protein ligand database (PLD) [29, 30]. At the time of this study the PDL contained some inconsistencies between enthalpies and binding constants. We derived all experimental binding enthalpies from the given binding constants using equation 12.

Evaluation of the MMFF interaction energy

First the interaction energy as given by the Merck Molecular Force Field has to be computed. Before binding the two molecules have intramolecular energies of ΔG^P_{intra} and ΔG^L_{intra}. During binding both protein and ligand change their conformation in a way that maximizes the interaction energy. Hence, a part of the negative intermolecular energy is compensated by an increase of intramolecular energy. After binding the complex has the intramolecular energy ΔG^{PL}_{intra} and the intermolecular energy ΔG^{PL}_{inter}. The

enthalpy of binding can thus be obtained by subtracting the increase of intramolecular energies from the decrease of intermolecular energy.

$$\Delta G_{bind} = \Delta G^{PL}_{inter} + \Delta G^{PL}_{intra} - \Delta G^{P}_{intra} - \Delta G^{L}_{intra} \tag{16}$$

The 4 components are approximated by local minima of the internal energy E, one for each single molecule and one for the complex.

$$\Delta G_{bind} \approx E^{PL}_{inter} + E^{PL}_{intra} - E^{P}_{intra} - E^{L}_{intra} \tag{17}$$

Ideally a global optimization should be applied. For efficiency reasons we used the conjugate gradient method as a local minimization method. This procedure can be justified for the complex whose PDB structure should be very near to the MMFF minimum structure. Minimizing the ligand and the protein starting from their structure in the complex can on the other hand easily lead into the trapping problem where the algorithm gets stuck in a local minimum.

Our results for the 3 minimizations in fact are unreliable, and the change of intermolecular energy of the single protein and the protein in the complex can be numerically so high that the other components are negligible. The consideration of conformation flexibility therefore leads to errors which seem to outweigh the advantages of such a procedure. Hence, we did not consider conformational flexibility of the protein in our study and thus equation 17 can be simplified to

$$\Delta G_{bind} = E^{PL}_{inter} + E^{LC}_{intra} - E^{L}_{intra} \tag{18}$$

where E^{LC}_{intra} is the intramolecular energy of the ligand in the complex. Due to this simplification only two energy minimizations are required: one for the ligand and one for the complex.

For minimization, we used a conjugate gradient method, implemented in accordance to the algorithm given in [31]. The following convergence criteria were used:

$$\begin{aligned} & |E^{intra}_i - E^{intra}_{i-1}| < .01 kJ/mol \\ \wedge \quad & |E^{inter}_i - E^{inter}_{i-1}| < .01 kJ/mol \\ \wedge \quad & \parallel g_i \parallel < 1 kN/mol \end{aligned}$$

where E^{intra}_i is the intramolecular energy of the i-th step of the iteration, E^{inter}_i is the energy of the noncovalent interaction between protein and ligand and g_i is the gradient. For minimizing the ligand without the protein the first criterion was omitted.

The computation of the noncovalent interaction is the component that determines the requirement in CPU time, as it has to be computed for all

atom pairs. To limit the computational demand we have choosen a cutoff for noncovalent interaction between atom i and j of

$$9990 \cdot |q_i \cdot q_j| + 10 < \| r_i - r_j \|$$

where q is the partial charge and r the position of the atom. As the protein is kept rigid during computation all noncovalent interactions between protein atoms can be omitted.

Results and Discussion

MMFF94

The implementation of the force field and the minimization algorithm has been validated using 10 reference structures of known MMFF minimum energy by applying small perturbations of $0.1\mathring{A}$ to the minimum structure. After minimization, the original structures with the corresponding minimum energies were again obtained.

The minimization of the complexes results in the intermolecular energies between protein and ligand. Table 8 shows the obtained values. Some of the complexes have positive binding energies which is caused by a larger increase of internal energy of the ligand than decrease of intermolecular interaction. These are cases of the trapping problem as in the worst case scenario the minimization should remove the ligand from the target resulting in a zero interaction energy and zero change in conformational energy.

Component	r
E_{intra}	-0.288
E_{elec}	0.055
E_{vdw}	0.649
E_{total}	-0.064

Table 4:

Table 4 shows the correlation between the computed components of the binding enthalpy and the experimental value of the free energy of binding. A significant correlation can only be found for the van der Waals component of the force field. The electrostatic component is completely uncorrelated. It might be argued that the protein ligand interaction is largely governed by van der Waals interactions but the results show that numerically the electrostatic component outweighs the van der Waals component. In order for this theory to be true one of the following thesis should be valid:

1. The electrostatic component in the MMFF model is overestimated by a large factor

2. the electrostatic component is compensated by another component which is not yet considered in the model

Such a component could be the electrostatic component of the free energy of solvation which we will discuss in the following section. A reason for the

overestimation of the electrostatic component by the MMFF model could be the electrostatic shielding due to polarization. The force field computes the Coulomb interaction in vacuo with a dielectric constant of $\varepsilon = 1$. Polarizable groups inside of the molecule ($\varepsilon \approx 2 - 4$) as well as the surrounding water $\varepsilon \approx 80$ can result in a strong reduction of the interaction. This effect can partially find consideration in our model by dividing ΔG_{elec} by a mean dielectric constant. The simplest way to do this is to include the seperated form of noncovalent interaction in the linear model such that the regression parameter will account for this effect. We carried out a linear regression with three independent variables E_{elec}, E_{vdw} und E_{intra}. With r=0.669 the increase in correlation over the pure van der Waals model is negligible. The components E_{intra} and E_{elec} are taken out of the model by very small regression parameters, for example 0.0018 for E_{elec} while E_{vdw} is scaled with 0.1451. Thus, the electrostatic component is scaled down by a factor of about 80, in relation to the vdW component, which is significantly higher than the value of the mean dielectric constant should be.

Including the solvation model

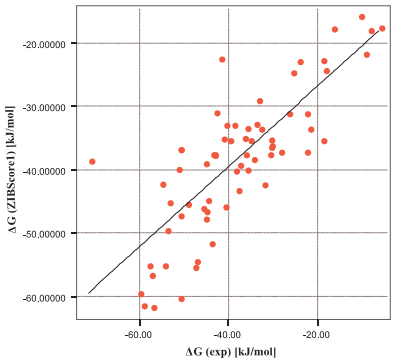
Deskriptor	r
PCM	-0.078
H-Num	0.441
C 0.591	
C=0	0.169
CB	0.412
N	0.269
O	0.545
F	0.112
P	-0.119
S	0.044
ΔG_{zibsm1}	-0.078
ΔSES	0.626

Table 5:

We will first apply the solvation model as a separate scoring function. Table 5 shows the correlation between each of the descriptors and the experimental value for ΔG_{bind}.

There is no significant correlation between the total ZIBSM1 parameter and experimental results. As already shown for the Merck force field, it is necessary to separate the model into its individual components. Of all components the changes of solvent accessible surface area of carbon and oxygen atoms have the best correlation. The change of these surface areas are largely proportional to the van der Waals interaction. To investigate these connections we used the sum of all changes of surface areas as a new descriptor which results in a correlation of 0.626 with ΔG_{exp} and of 0.797 with the van der Waals component of the Merck force field. The electrostatic component of the simplified BE method shows no correlation which is comparable to the results obtained for the intermolecular Coulomb interaction of the Merck force field.

We now want to answer the question whether the electrostatic interaction between the molecule and the electrostatic solvation component cancel each other out. The correlation between the two values is $r = 0.837$. Considering the statistical deviations of the intermolecular Coulomb energies due to the trapping problem of the minimization this is a very high value.



Deskriptor	r	σ	sign.
(Constant)	-16.820	4.282	.000
E_{elect}	.006	.003	.028
E_{vdw}	.147	.048	.003
E_{intra}	.001	.003	.644
PCM	.018	.011	.125
H-NUM	-.405	.357	.262
C	-.007	.016	.653
C=O	-.254	.106	.020
CB	.009	.028	.761
N	.082	.129	.529
O	.166	.059	.007
Cl	.333	.237	.166
P	-.644	2.106	.761
S	.974	.293	.002

Figure 7: Ergebnisse und Parametrisierung der ZIBSCORE1 Scoringfunktion

Thus, both components in fact seem to scale linearly and cancel each other out at least partially. We parameterized the complete model with its 13 descriptors. Table 7 shows the results. The new scaling factor between electrostatic MMFF interaction energy and the van der Waals component is now $0.147/0.006 \approx 20$ which is a appropriate value for a mean dielectric constant.

The RMSD between the predicted value and the experimental value is $9.41 kJ/mol$, the correlation is $r = 0.797$. The crossvalidated correlation coefficient is $r^2_{cv} = 0.543$ and thus considerably smaller than r^2 mit 0.634. Thus, the parameterization of the model should be repeated with a larger amount of experimental data in the future.

Other non-solvation components

Apart from the interaction enthalpy and the change in solvation free energy the remaining entropic components are not explicitly considered in our model. These are the loss of entropy due to the loss of freedoms of translation rotation and vibration of the two molecules as well as the freezing of bond torsions in the complex

By combining thermostatistical argumentation with experimental data, Yu [32] concludes that the contribution of the change of translational and rotational entropies for association per subunit is $G^{\circ}_{tr} = 0 \pm 5RT$. In the case of protein ligand docking two subunits aggregate and the contribution is therefore $\Delta G_{tr} = 0 \pm 5RT$. Hence, the entropic contribution due to a loss of freedoms of translation and rotation is both small and independent of the

properties of the molecules. This component is thus implicitely modelled by the constant of the regression.

To estimate the effect of torsion freezing the conformation flexibility of the single molecules and the complex needs to be analyzed. Some scoring function which have been parameterized with experimental data estimate this contribution per fixation of a torsion as about $1.4kJ/mol$ [33, 34]. This component depends on the size of the ligand, as larger ligands will cause a higher number of side chain fixation of the protein and can become significant for larger ligands. We will consider this component in one of our future studies. For the moment we wish to point out that this component is partly accounted for by the atomic solvation parameters, as the number of frozen bonds can be considered to be roughly proportional to the change of the solvent accessible surface area of the two molecules.

Conclusion and outlook

We have applied one ASP model from the literature and 3 newly developed models to a set of 213 small molecules. Table 6 shows the cross-correlation of the predicted values. While the new ASP parameterization ZIBASP1 shows a significant improvement over the Oons model, it still fails for strongly charged molecules. It has been shown by Juffer [35] that the prediction of different ASP sets does not only differ in magnitude but also in sign. Our results underline that ASP sets are only appropriate to estimate the complete free energy of solvation for uncharged species. A pure electrostatic term shows a better correlation while failing for uncharged molecules like hydrocarbons. The combination of the two models and an additional hydrogen bond term leads to a good agreement with experimental data (ZIBSM1) proving that an independent modelling of the underlying physical effects of solvation can yield excellent results.

	Exp	Oons	ZIBasp1	ZIBpcm1	ZIBsm1
Exp	1.000	0.509	0.746	0.812	0.951
Oons	0.509	1.000	0.679	0.488	0.553
ZIBasp1	0.746	0.679	1.000	0.763	0.796
ZIBpcm1	0.812	0.488	0.763	1.000	0.839
ZIBsm1	0.951	0.553	0.796	0.839	1.000

Table 6: cross correlation table for experimental free energies of solvation and the values predicted by the 4 models

The 3 newly developed models ZIBASP, ZIBBE, and ZIBSM have been implemented as a part of the visualization software Amira [25]. In our implementation of the model the required processor time is in the range of one second per molecule on a common desktop PC. The rmsd of the model is $0.8 kcal/mol$ and therefore only slightly worse than that of the more costly SMX models of Chambers [24].

The intermolecular energy between protein and ligand has been estimated with the Merck Molecular Force Field. For energy minimization we used a conjugate gradient method. For a set of molecules we obtained positive binding energies which are caused by a trapping of the conformation of the complex in local minima. There is no significant correlation between the energies. Charifson found in a comparative study of 13 scoring functions that the scoring with the MMFF94 energy belonged to the group with the worst hitrates [36]. Charifson also used a conjugate gradient method with a rigid protein.

As an alternative scoring function we used our solvation model ZIBSM1. In its original form which was parameterized for small molecules the correlation is similarly bad.

By separating the MMFF or solvation energies into their individual physical components the correlation becomes significant, while the dominating contribution is the van der Waals component. The correlation between electrostatic interaction energy and the electrostatic component of the solvation energy is large, thus indicating a partial elimination of the two components. This is a common problem of models which rely on the partition of the free energy. If strongly correlated contributions are sufficiently high, a high precision for their computation is needed, as small inaccuracies can outweigh the scale of those components which determine the process.

The combination of the solvation model with the MMFF model results in a correlation of $r = 0.797$ and a rmsd of $9.41 kJ/mol$. As a comparison, the knowledge based scoring function BLEEP reaches a correlation of $r = 0.624$ for our parameterization set. With only one parameter (the van der Waals component of the MMFF force field or the total change of solvent accessible surface area) we obtain a better correlation. The scoring functions with a test set of more than 50 molecules which were listed by Gohlke have an rmsd between 6 and $10 kJ/mol$ [37].

The computational effort of each scoring is high due to the energy minimization (in the range of one hour). Without the minimization only about half a minute is needed. As the solvation model reacts less sensitively to conformational changes it should be seperately used for an initial rough scoring while the complete model should only be applied to the final conformation.

The development of an additional solvation term for protein ligand scoring has been done as part of the development of a docking suite at the Konrad Zuse Institute Berlin [6]. For the future development several problems need to be addressed:

- High computational effort: This method will be part of a hierachical scoring scheme. For the early phases of the conformational search simpler methods like knowledge based scoring functions will be used.
- Trapping: Global optimization like simulated annealing or genetic algorithm [38, 34] need to be implemented.
- Entropic contributions: A model for the remaining entropic contributions needs to be developed.
- Anticorrelation of components: Further mutual compensations between components need to be investigated. Eventually an empirical scoring function which is not directly based on the Merck force field has to be developed to combine such contributions into single terms.

Acknowledgement

The work of F. Cordes has been supported by the German Federal Ministry of education and research (grant no. 031U109A/031U209A, Berlin Center for Genome Based Bioinformatics). The authors would like to thank Steffen Eisenhardt, Andreas May, and Sebastian Moll for help during this study.

Appendix

Results of the solvation models

	G_{exp}	oons	zibasp1	zibpcm1	zibsm1
brominated hydrocarbons					
1-bromobutane	-0.41		-0.94	-0.85	1.36
1-bromopentane	-0.08		-1.01	-0.87	-0.23
1-bromopropane	-0.56		-0.88	-0.86	-0.57
2-bromopropane	-0.48		-0.85	-0.79	-0.52
bromobenzene	-1.46		-2.31	-1.85	-1.59
bromoethane	-0.70		-0.82	-0.90	-0.79
bromomethane	-0.82		-0.79	-0.99	-1.09
dibromomethane	-2.11		-1.21	-1.97	-2.33
p-dibromobenzene	-2.30		-2.45	-1.39	-1.67
tribromomethane	-1.98		-1.59	-2.41	-2.80
rmsd			0.56	0.48	0.66
chlorinated hydrocarbons					
1,1,1-trichloroethane	-0.25		-0.72	-1.81	-0.60
1,1,2-trichloroethane	-1.95		-0.74	-3.33	-2.06
1-chloropropane	-0.27		-0.57	-1.31	-0.58
2-chloropropane	-0.25		-0.55	-1.22	-0.48
3-chloropropene	-0.57		-0.56	-1.64	-0.79
chlorobenzene	-1.12		-1.99	-1.97	-1.25
chloroethane	-0.63		-0.50	-1.38	-0.84
chloroethene	-0.59		-0.50	-1.06	-0.21
chloromethane	-0.56		-0.44	-1.51	-1.22
dichloromethane	-1.36		-0.57	-2.77	-2.12
e-1,2-dichloroethene	-0.76		-0.62	-1.27	-0.24
o-dichlorobenzene	-1.36		-1.87	-1.97	-1.37
p-dichlorobenzene	-1.01		-1.82	-1.75	-1.04
trichloroethene	-0.39		-0.72	-1.45	-0.18
trichloromethane	-1.07		-0.68	-3.28	-1.93
z-1,2-dichloroethene	-1.17		-0.60	-2.21	-1.38
rmsd			0.55	1.12	0.41
fluorinated hydrocarbons					
1,1-difluoroethane	-0.11		0.07	-2.60	0.45
fluorobenzene	-0.78		-1.67	-1.85	-0.19
fluoromethane	-0.22		-0.07	-1.88	-0.02
rmsd			0.53	1.84	0.48

iodinated hydrocarbons					
1-iodobutane	-0.25		-0.92	-0.60	-0.36
1-iodopentane	-0.12		-0.99	-0.62	-0.20
1-iodopropane	-0.59		-0.86	-0.61	-0.54
2-iodopropane	-0.46		-0.83	-0.56	-0.50
diiodomethane	-2.49		-1.15	-1.47	-2.32
iodobenzene	-1.73		-2.26	-1.84	-1.72
iodoethane	-0.72		-0.80	-0.64	-0.75
iodomethane	-0.89		-0.76	-0.70	-1.03
rmsd			0.66	0.44	0.10
other halo					
1,1,1,3,3,3-hexafluoropropan-2-ol	-3.77		-1.48	-5.09	-2.74
1,1,1-trifluoropropan-2-ol	-4.16		-2.15	-4.36	-3.51
1,1,2-trichloro-1,2,2-trifluoroethane	1.77		-0.18	-2.26	1.39
1-bromo-1,2,2,2-tetrafluoroethane	0.52		0.01	-3.52	0.37
1-bromo-1-chloro-2,2,2-trifluoroethane	-0.13		-0.30	-3.50	-0.36
1-bromo-2-chloroethane	-1.95		-0.95	-2.07	-1.85
1-chloro-2,2,2-trifluoroethane	0.06		0.13	-3.33	0.32
1-chloro-2,2,2-trifluoroethyl-difluoromethyl-ether	0.11		-0.09	-6.24	-1.48
2,2,2-trifluoroethanol	-4.31		-2.25	-4.88	-3.76
2,2,2-trifluoroethyl-vinyl-ether	-0.12		-0.22	-3.73	-0.82
bis(2-chloroethyl)sulfide	-3.92		-1.25	-3.49	-2.53
bromotrifluoromethane	1.79		-0.09	-1.57	1.77
chlorodifluoromethane	-0.50		0.03	-3.41	0.00
chlororfluoromethane	-0.77		-0.21	-2.86	-0.98
p-bromophenol	-7.13		-5.19	-4.35	-6.20
tetrachloroethene	0.05		-0.86	-0.88	1.33
tetrafluoromethane	3.11		0.63	-1.59	2.96
rmsd			1.55	3.12	0.76
alkene					
1-butene	1.38	1.86	-0.49	-0.76	0.47
1-hexene	1.68	2.32	-0.62	-0.76	0.80
1-pentene	1.66	2.09	-0.56	-0.76	0.64
2-methylpropene	1.16	1.86	-0.50	-0.74	0.50
cyclopentene	0.56	1.91	-0.51	-0.84	0.45
ethene	1.27	1.39	-0.37	-0.76	0.08
propene	1.27	1.64	-0.44	-0.75	0.29
trans-1,3-butadiene	0.61	1.82	-0.48	-1.51	-0.12
trans-2-pentene	1.34	2.09	-0.56	-0.62	0.76
rmsd		0.77	1.76	2.07	0.84

alkine					
1-butyne	-0.16	1.84	-0.49	-0.91	0.26
1-hexyne	0.29	2.3	-0.61	-0.88	0.62
1-pentyne	0.01	2.07	-0.55	-0.88	0.45
ethyne	-0.01	1.37	-0.36	-0.91	-0.11
propyne	-0.31	1.62	-0.43	-0.98	-0.01
rmsd		1.89	0.53	0.89	0.35
arene					
anthracene	-4.23	-2.75	-3.41	-4.66	-3.00
benzene	-0.87	-1.75	-2.17	-2.49	-1.57
ethylbenzene	-0.80	-0.33	-1.84	-2.50	-1.05
m-xylene	-0.84	0.41	-1.53	-2.52	-0.95
naphthalene	-2.39	-2.25	-2.79	-3.67	-2.37
o-xylene	-0.90	0.1	-1.64	-2.54	-1.00
p-xylene	-0.81	0.41	-1.53	-2.52	-0.95
toluene	-0.89	-0.67	-1.86	-2.52	-1.27
rmsd		0.95	0.87	1.52	0.53
branched alkane					
2,2,4-trimethylpentane	2.85	2.6	-0.69	0.00	1.58
2,2-dimethylpropane	2.50	2.04	-0.54	0.00	1.13
2,4-dimethylpentane	2.88	2.49	-0.66	0.00	1.47
2-methylpentane	2.52	2.31	-0.61	0.00	1.33
2-methylpropane	2.32	1.86	-0.49	0.00	0.99
rmsd		0.37	3.23	2.62	1.32
cycloalkane					
cis-1,2-dimethylcyclohexane	1.58	2.4	-0.64	0.00	1.44
cyclohexane	1.23	2.04	-0.54	0.00	1.15
cyclopentane	1.20	1.89	-0.50	0.00	1.02
cyclopropane	0.75	1.57	-0.42	-0.36	0.54
methylcyclohexane	1.71	2.25	-0.60	0.00	1.31
rmsd		0.74	1.88	1.39	0.22
unbranched alkane					
butane	2.08	1.88	-0.50	0.00	1.00
ethane	1.83	1.42	-0.38	0.00	0.66
heptane	2.62	2.59	-0.69	0.00	1.52
hexane	2.49	2.36	-0.63	0.00	1.35
methane	2.00	1.16	-0.31	0.00	0.47
octane	2.89	2.83	-0.75	0.00	1.70

pentane	2.33	2.13	-0.56	0.00	1.18
propane	1.96	1.66	-0.44	0.00	0.83
rmsd		0.36	2.85	2.3	1.19
aliphatic amines					
azetidine	-5.56	-1.66	-2.42	-2.47	-3.92
butylamine	-4.29	-4.49	-4.42	-2.62	-4.46
diethylamine	-4.07	0.8	-1.33	-2.02	-3.40
dimethylamine	-4.29	-0.77	-1.84	-2.47	-4.35
dipropylamine	-3.66	1.29	-1.44	-2.35	-3.29
ethylamine	-4.50	-4.96	-4.30	-2.75	-5.14
methylamine	-4.56	-5.77	-4.56	-2.81	-5.33
n,n-dimethylpiperazine	-7.58	1.77	-1.01	-5.25	-6.31
n-methylpiperazine	-7.77	-0.36	-2.09	-4.65	-7.45
pentylamine	-4.10	-4.26	-4.49	-2.62	-4.29
piperazine	-7.40	-2.69	-3.28	-3.97	-7.98
piperidine	-5.11	-0.27	-1.88	-2.20	-3.63
propylamine	-4.39	-4.73	-4.37	-2.61	-4.61
pyrrolidine	-5.48	-0.66	-1.97	-2.46	-4.15
trimethylamine	-3.23	1.62	-0.61	-1.83	-2.50
rmsd		4.56	3.17	2.3	0.85
amide					
ethanamide	-9.71	-6.99	-8.07	-5.34	-7.88
n-methylacetamide	-10.00	-0.5	-3.94	-4.86	-7.28
rmsd		6.99	4.44	4.77	2.32
aromatic amines					
2,4-dimethylpyridine	-4.86	-1.48	-2.70	-4.43	-4.18
2,5-dimethylpyridine	-4.72	-1.49	-2.71	-4.34	-4.09
2,6-dimethylpyridine	-4.60	-0.81	-2.30	-4.27	-3.82
2-ethylpyrazine	-5.51	-4.65	-4.50	-6.46	-7.00
2-methylpyrazine	-5.57	-5.23	-4.64	-6.63	-7.64
2-methylpyridine	-4.63	-2.57	-3.02	-4.40	-4.45
3-methylpyridine	-4.77	-3.3	-3.46	-4.51	-4.73
4-methylpyridine	-4.94	-3.24	-3.42	-4.57	-4.76
aniline	-5.49	-8.44	-6.37	-5.34	-7.20
pyridine	-4.70	-4.46	-3.82	-4.92	-5.31
rmsd		2.35	1.55	0.53	1.06
nitrile					
benzonitrile	-4.10	-6.5	-5.41	-4.13	-3.08
butanonitrile	-3.64	-3.95	-4.05	-4.15	-3.01

ethanonitrile	-3.89	-4.5	-3.99	-4.48	-3.83
propanonitrile	-3.85	-4.15	-3.97	-4.16	-3.27
rmsd		1.26	0.69	0.42	0.67
nitrohydrocarbons					
1-nitrobutane	-3.08	-13.19	-6.77	-4.80	-2.92
1-nitropropane	-3.34	-13.42	-6.71	-4.80	-3.12
2-methyl-1-nitrobenzene	-3.59	-14.48	-7.67	-5.68	-4.07
2-nitropropane	-3.14	-12.45	-6.28	-4.33	-2.71
nitrobenzene	-4.12	-16.82	-8.50	-5.41	-3.96
nitroethane	-3.71	-14.2	-6.89	-6.00	-4.40
rmsd		10.65	3.67	1.72	0.41
other hcno					
2-methoxyethanamine	-6.55	-5.89	-4.93	-4.23	-7.35
morpholine	-7.17	-2.06	-2.56	-3.76	-6.53
N-methylmorpholine	-6.34	0.2	-1.41	-3.66	-5.19
rmsd		4.81	4.01	2.84	0.89
alcohol					
1,2-ethanediol	-9.30	-11.55	-5.63	-5.95	-9.76
1-butanol	-4.72	-4.46	-3.07	-2.98	-4.26
1-heptanol	-4.24	-3.76	-3.25	-2.99	-3.75
1-hexanol	-4.36	-4	-3.19	-2.99	-3.92
1-octanol	-4.09	-3.53	-3.31	-2.99	-3.58
1-pentanol	-4.47	-4.23	-3.13	-2.99	-4.09
1-propanol	-4.83	-4.81	-3.06	-2.92	-4.32
2-methyl-2-propanol	-4.51	-3.65	-2.73	-2.63	-3.81
2-propanol	-4.76	-4.27	-2.83	-2.86	-4.25
cyclopentanol	-5.49	-4.25	-2.98	-2.78	-4.02
ethanol	-5.01	-5.05	-3.00	-2.97	-4.55
m-crescol	-5.49	-7.9	-4.75	-4.98	-5.87
methanol	-5.11	-5.78	-3.12	-3.11	-5.05
o-crescol	-5.87	-6.59	-4.23	-4.87	-5.82
p-crescol	-6.14	-7.9	-4.75	-4.99	-5.89
phenol	-6.62	-8.92	-5.05	-4.89	-6.15
rmsd		1.21	1.82	1.82	0.57
aldehydes					
benzaldehyde	-4.02	8.22	-4.69	-5.35	-4.35
butanal	-3.18	8.81	-3.38	-3.87	-2.71
ethanal	-3.50	9.93	-3.16	-4.10	-3.29
octanal	-2.29	9.73	-3.64	-3.89	-2.08

pentanal	-3.03	9.02	-3.44	-3.91	-2.57
propanal	-3.44	8.59	-3.32	-3.87	-2.91
rmsd		12.3	0.66	1.01	0.39
acids					
butanoic acid	-6.36	-6.41	-6.42	-6.65	-7.78
ethanoic acid	-6.70	-7.15	-6.82	-4.82	-6.41
hexanoic acid	-6.21	-5.84	-6.68	-4.51	-5.24
pentanoic acid	-6.16	-6.52	-6.65	-6.91	-7.88
propanoic acid	-6.47	-6.76	-6.68	-7.38	-8.46
rmsd		0.33	0.32	1.26	1.41
ester					
butyl ethanoate	-2.55	-0.11	-4.57	-4.29	-4.61
ethyl ethanoate	-3.10	0.2	-3.19	-2.62	-3.02
ethyl methanoate	-2.65	8.77	-4.33	-4.79	-4.69
methyl butanoate	-2.83	0.95	-3.02	-2.72	-3.08
methyl ethanoate	-3.32	-1.42	-4.23	-3.87	-4.47
methyl hexanoate	-2.49	1.08	-3.14	-2.75	-2.54
methyl methanoate	-2.78	7.41	-4.57	-5.12	-5.41
methyl octanoate	-2.04	1.12	-3.61	-2.81	-2.26
methyl pentanoate	-2.57	0.86	-3.08	-2.75	-2.68
methyl propanoate	-2.93	-1.64	-3.99	-3.95	-4.27
pentyl ethanoate	-2.45	1.2	-3.47	-2.80	-2.64
propyl ethanoate	-2.86	0.39	-3.22	-2.54	-2.76
rmsd		5.24	1.17	1.14	1.24
ether					
1,2-dimethoxyethane	-4.84	-0.44	-1.65	-3.22	-4.54
1,4-dioxane	-5.05	-1.45	-1.83	-3.57	-4.97
anisole	-1.04	-2.68	-2.74	-3.06	-3.20
diethylether	-1.76	1.55	-0.76	-1.93	-1.86
dimethyl ether	-1.92	-0.04	-1.05	-2.16	-2.73
methylisopropylether	-2.01	1.05	-0.93	-1.88	-2.02
methylpropylether	-1.66	0.99	-0.97	-2.01	-2.08
t-butylmethylether	-2.21	1.33	-0.90	-1.75	-1.67
tetrahydrofuran	-3.47	0.02	-1.19	-2.11	-2.47
rmsd		3.17	1.94	1.11	0.88
ketone					
1-phenylethanone	-4.58	-1.41	-4.64	-5.24	-4.65
2-heptanone	-3.04	1.39	-2.93	-3.43	-2.13
2-hexanone	-3.29	1.21	-2.87	-3.43	-2.30

2-octanone	-2.88	1.55	-3.50	-3.58	-1.97
2-pentanone	-3.53	0.98	-2.81	-3.43	-2.49
3,3-dimethybutanone	-2.89	0.94	-2.73	-3.08	-2.11
3-pentanone	-3.41	1.28	-2.25	-3.32	-2.29
4-heptanone	-2.93	1.06	-3.21	-3.22	-1.94
5-nonanone	-2.67	2.03	-2.80	-3.44	-1.61
butanone	-3.64	0.26	-3.29	-3.38	-2.80
cyclopentanone	-4.68	0.3	-3.51	-3.60	-2.93
propanone	-3.85	0.14	-3.24	-3.61	-3.14
rmsd		4.29	0.61	0.51	1.00
other hco					
2-methoxyethanol	-6.77	-5.93	-3.61	-4.15	-6.85
2-propen-1-ol	-5.08	-4.84	-3.05	-3.70	-4.92
butenyne	0.04	1.83	-0.48	-1.20	0.02
m-hydroxybenzaldehyde	-9.51	1.64	-7.47	-8.12	-9.23
p-hydroxybenzaldehyde	-10.48	1.71	-7.45	-8.05	-9.16
rmsd		7.44	2.35	1.9	0.61
sulfide					
diethyl disulfide	-1.63	0.12	-1.50	-1.75	-0.17
diethyl sulfide	-1.43	1.06	-0.99	-1.83	-0.92
dimethyl disulfide	-1.83	-0.8	-1.52	-1.81	-0.70
dimethyl sulfide	-1.54	0.21	-1.00	-1.90	-1.56
dipropyl sulfide	-1.27	1.5	-1.12	-1.78	-0.50
hydrogen sulfide	-0.70	-3.05	-1.79	-1.17	-2.75
thioanisole	-2.73	-1.69	-2.38	-2.71	-2.01
rmsd		1.99	0.53	0.33	1.14
thiols					
1-propanethiol	-1.05	-0.53	-1.39	-1.50	-1.86
ethanethiol	-1.30	-0.76	-1.32	-1.53	-2.08
methanethiol	-1.24	-1.19	-1.32	-1.41	-2.18
thiophenol	-2.55	-3.23	-2.79	-2.32	-2.68
rmsd		0.51	0.2	0.3	0.73
other					
amonia	-4.29	-14.44	-9.42	-3.71	-6.23
hydrazine	-9.30	-18.11	-11.82	-5.37	-8.69
water	-6.31	-16.95	-7.26	-5.65	-8.88
rmsd		9.9	3.35	2.32	1.89
total rmsd		4.1	1.87	1.75	0.87

Table 7: Number of hydrogen donors and acceptors (h-Num), experimental free energy of solvation (Gexp) and values computed with the 4 models (all values in kcal/mol). For each type of molecules an rmsd is given under the individual data.

Results for the protein ligand complexes

PDB	ΔG_{exp}	ΔG_{inta}	ΔG_{elec}	ΔG_{vdW}	$\Delta G_{ZIBscore1}$
1a07	-22.82	272.06	-1464.39	-36.21	-31.24
1aaq	-47.98	379.73	-400.24	-123.72	-55.56
1abe	-40.11	73.14	-477.50	-15.59	-35.53
1abf	-30.97	85.37	-419.67	-19.04	-36.57
1anf	-31.13	204.17	-523.29	-13.13	-37.72
1apb	-33.26	74.11	-382.66	-36.24	-33.66
1apt	-53.68	563.09	-2523.12	-81.33	-45.32
1apu	-43.98	496.76	-327.50	-122.21	-37.71
1apv	-51.34	282.58	-455.47	-111.07	-47.41
1apw	-45.64	245.56	-385.08	-118.95	-47.89
1ba8	-51.34	471.39	-846.10	-79.91	-36.95
1bap	-39.19	62.33	-465.09	-24.31	-33.10
1c83	-19.23	183.67	-2502.01	31.01	-35.47
1cbs	-41.07	144.41	-316.35	-96.97	-33.07
1cho	-60.28	4397.74	-961.63	-185.50	-59.65
1ejn	-32.51	263.02	-120.74	-63.51	-42.43
1epo	-45.40	662.36	-574.88	-97.90	-46.75
1fkf	-55.37	384.60	-167.33	-121.85	-42.31
1fkg	-36.86	316.55	-67.05	-109.09	-35.09
1hbv	-36.34	508.86	-280.58	-115.04	-40.14
1hew	-34.23	656.74	-515.16	-56.73	-32.96
1hsb	-51.67	386.17	-1105.76	-17.11	-40.00
1htf	-46.21	184.50	-255.71	-110.27	-46.17
1hvi	-57.50	209.08	-365.75	-196.19	-61.86
1hvj	-59.67	190.75	-260.70	-210.81	-61.57
1hvk	-57.73	271.30	-311.25	-206.82	-56.87
1hvl	-51.40	245.66	-284.93	-204.03	-60.46
1hvr	-54.25	180.56	-291.39	-180.25	-49.69
1jao	-33.78	236.43	890.57	-12.11	-29.20
1jap	-26.95	291.00	-1021.92	-35.12	-31.27
1mmp	-35.60	152.51	-365.23	-74.82	-35.48
1mmq	-51.34	266.11	-211.38	-88.70	-36.92
1nco	-44.34	1297.47	-113.08	-140.93	-51.82

1nnb	-22.83	347.20	-883.15	-8.47	-37.26
1ppk	-43.70	480.25	-432.14	-99.71	-37.71
1ppl	-45.03	455.81	-350.45	-138.71	-44.95
1qbu	-58.43	327.88	-283.89	-156.20	-55.29
1rbp	-38.33	337.08	-50.75	-153.28	-43.34
1rgk	-24.59	211.80	-449.56	25.96	-22.92
1stp	-71.47	195.01	-477.07	-34.05	-38.76
1thl	-36.63	237.77	-246.19	-72.19	-37.71
1tmn	-41.67	944.27	-721.96	-53.97	-35.24
1tng	-16.75	49.99	295.30	-26.71	-17.84
1tnh	-19.22	34.52	221.71	-23.27	-22.88
1tni	-9.69	101.12	247.10	-15.11	-21.76
1tnj	-6.15	61.22	271.81	-10.51	-17.64
1tnk	-8.50	114.80	248.29	-13.70	-18.08
1tnl	-10.70	80.87	299.94	7.26	-15.93
1uvs	-30.81	500.84	-32.66	-130.03	-36.28
1uvt	-43.60	259.71	-54.83	-116.39	-37.82
2cmd	-26.10	305.06	-2129.85	35.01	-24.75
2er6	-41.22	1265.01	-519.48	-95.87	-45.99
2gbp	-43.36	124.31	-511.34	4.10	-31.05
2h4n	-49.65	89.63	-129.78	-49.39	-45.55
2ifb	-30.97	182.62	-277.02	-45.68	-35.35
3cpa	-22.13	201.33	-970.22	-29.63	-33.74
4er4	-38.78	904.71	-7821.44	-43.05	-40.29
4hvp	-34.85	1444.33	-711.19	-49.39	-38.42
4sga	-18.65	278.18	-683.25	-56.34	-24.33
5abp	-37.94	100.06	-494.97	-27.33	-39.34
6abp	-36.34	147.01	-477.90	-15.87	-33.52
6tmn	-28.82	425.45	-397.44	-86.13	-37.33
7dfr	-42.21	494.92	-856.81	-20.34	-22.62
7hvp	-54.94	1394.28	-544.64	-130.49	-55.27
9abp	-45.70	69.59	-531.95	-20.39	-39.16
9hvp	-47.64	451.59	-349.26	-162.02	-54.61

Table 8: Experimental free solvation enthalpy ΔG_{exp}, change of conformational energy of the ligand computed with MMFF94 ΔG_{intra}, electrostaic interaction ΔG_{elec}, and van der Waals interaction ΔG_{vdw}, as well as estimated free energy of binding of the ZIBSCORE1 model (all values in kJ/mol)

References

[1] K. A. Dill. Dominant forces in protein folding. *Biochemistry*, 29:7133–7155, 1990.

[2] F. Eisenhaber. Hydrophobic regions on protein surfaces. *Perspectives in Drug Discovery and Design*, 17:27–42, 1999.

[3] J.K. Seydel. *QSAR and Strategies in the Design of Bioactive Compounds*. VCH, Weinheim, 1985.

[4] Osman F. Guner. *Pharmacophore Perception, Development, and Use in Drug Design*. International University Line, 2000.

[5] Holger Gohlke, Manfred Hendlich, and Gerhard Klebe. Knowledge based scoring function to predict protein-ligand interactions. *J. Molec. Biol.*, 295:337–356, 2000.

[6] Andreas May, Steffen Eisenhardt, Johannes Schmidt-Ehrenberg, and Frank Cordes. Rigid body docking for virtual screening. Technical report, 2003. 03-47 ZIB.

[7] Thomas A. Halgren. Merck Molecular Force Field I-V. *J.Comp.Chem.*, 17:490–640, 1996.

[8] A. Nicholls B. Honig. Classical electrostatics in biology and chemistry. *Science*, 26:1144–1149, 1995.

[9] M. Holst, N. Baker, and F. Wang. Adaptive multilevel finite element solution of the poisson-boltzmann equation I: algorithms and examples. *J. Comput. Chem.*, 21:1319–1342, 2000.

[10] R.J. Zauhar and R.S. Morgan. A new method for computing the macromolecular electric potential. *J. Mol. Biol.*, 186:815–820, 1985.

[11] W.C. Still, A. Tempczyrk, R.C. Hawley, and T. Hendrickson. Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.*, 112:6127–6129, 1990.

[12] C.J. Cramer and D.G. Truhlar. Continuum solvation models: Classical and quantum mechanical implementations. *Rev. in Comp. Chemistry*, VI, 1995.

[13] N.T. Southall, K.A. Dill, and A.D.J. Haymet. A view of the hydrophobic effect. *J. Phys. Chem. B*, 106:521–533, 2002.

[14] T. Lazaridis. Solvent size vs cohesive energy as the origin of hydrophobicity. *Accounts of Chemical Research*, 34:931–937, 2001.

[15] L.R. Pratt and A. Pohorille. Hydrophobic effects and modeling of biophysical aqueous solution interfaces. *Chem Rev.*, 102:2671–2692, 2002.

[16] D. Eisenberg and A.D. McLachlan. Solvation energy in protein folding and binding. *Nature*, 319:199– 203, 1986.

[17] G.A. Jeffrey and W. Sanger. *Hydrogen Bonding in Biological Structures*. Springer Verlag, 1991.

[18] H.J. Boehm and G. Klebe. What can we learn from molecular recognition in protein-ligand complexes for the design of new drugs? *Angew. Chem. Int. Ed. Engl.*, 35:2588–2614, 1996.

[19] Enrico O. Purisima and Shahul H. Nilar. A simple yet accurate boundary element method for continuum dielectic calculations. *J. Comp. Chem.*, 16:681–689, 1995.

[20] C. Chothia. Hydrophobic bonding and accessible surface areas in proteins. *Nature*, 248:338– 339, 1974.

[21] K.A. Sharp, A. Nicholls, R. Fine, and B. Honig. Reconciling the magnidtude of microscopic and macroscopic hydrophobic effects. *Science*, 252:106–109, 1991.

[22] B. Lee and F. M. Richards. The interpretation of protein structures: Estimation of static accessibility. *J. Mol. Biol.*, 55:379–400, 1971.

[23] F.M. Richards. Areas, volumes, packing and protein structure. *Ann. Rev. Biophys. Bioeng.*, 6:151–176, 1977.

[24] Candee C. Chambers, Gregory D. Hawkins, Christopher J. Cramer, and Donald G. Truhlar. Models for aeqeous solvation based on class iv atomic charges and first solvation shells effects. *J. Phys. Chem.*, 100:16385–16398, 1996.

[25] Detlev Stalling, Malte Westerhoff, and Hans-Christian Hege. Amira - a highly interactive system for visual data analysis. 2003. To appear in: Christopher R. Johnson and Charles D. Hansen (eds.), Visualization Handbook (2004), Academic Press, also available under http://www.zib.de/visual/publications/sources/amira-overview.pdf.

[26] J. O'M. Bockris and A. K. N Reddy. *Modern Electrochemistry 1*. Plenum Press New York, 1977.

[27] Nathan A. Baker, David Sept, Simpson Joseph, Michael J. Holst, and J. Andrew McCammon. Electrostatics of nanosystems: application to microtubules and the ribosom. *Proc. Natl. Acad. Sci. USA*, 98:10037–10041, 2001.

[28] Maxim Totrov and Ruben Abagyan. Rapid boundary element solvation electrostatics calculations in folding simulations. *Biopolymers*, 60:124–133, 2001.

[29] *http://www.rcsb.org/pdb/.*

[30] Dushyanthan Puvanendrampillai and John B. O. Mitchell. Protein ligand database (pld): additional understanding of the nature and specificity of protein ligand complexes. *Bioinformatics*, 19:1856–1857, 2003.

[31] William H. Press. *Numerical Recipes in C.* Camebridge University Press, 1988.

[32] Y. Bruce Yu, Peter L. Privalov, and Robert S. Hodges. Contribution of translational and rotational motions to molecular association in aqueous solution. *Bioph. J.*, 81:1632–1642, 2001.

[33] H.J. Böhm. The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. *J. Comp.-Aided Mol. Des.*, 8:243–256, 1994.

[34] G.M. Morris, D.S. Goodsell, R.S. Halliday, R. Huey, W.E. Hart, R.K. Belew, and A.J. Olson. Automated docking using a lamarckian genetic algorithm and an empirical binding free energy function. *J.Comp.Chem.*, 14:1639–1662, 1998.

[35] A.H. Juffer, F. Eisenhaber, S.J. Hubbard, D. Walther, and P. Argos. Comparison of atomic solvation parametric sets. *Protein Science*, 4:2499–2509, 1995.

[36] Paul S. Charifson, Joseph J. Corkery, Mark A. Murcko, and Patrick Walters. Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *J. Med. Chem.*, 42:5100–5109, 1999.

[37] Holger Gohlke and Gerhard Klebe. Ansätze zur Beschreibung und Vorhersage der Bindungsaffinität niedermolekularer Liganden an makromolekulare Rezeptoren. *Angew. Chem.*, 114:2764–2798, 2002.

[38] D.R. Westhead, D.E. Clark, and C.W. Murray. A comparison of heuristic search algorithms for molecular docking. *J.Comp.Aided Molec. Design*, 11:209–228, 1997.