LUND UNIVERSITY

Accurate Predictions of Nonpolar Solvation Free Energies Require Explicit Consideration of Binding-Site Hydration

Genheden, Samuel; Mikulskis, Paulius; Hu, LiHong; Kongsted, Jacob; Söderhjelm, Pär; Ryde, Ulf

Link to publication

Total number of authors:
6

Accurate predictions of non-polar solvation free energies require explicit consideration of binding site hydration

Samuel Genheden[1], Paulius Mikulskis[1], LiHong Hu[1,2], Jacob Kongsted[3], Pär Söderhjelm[4], Ulf Ryde[1*]

[1]Department of Theoretical Chemistry, Lund University, Chemical Centre,
P. O. Box 124, SE-221 00 Lund, Sweden

[2]Faculty of Chemistry, North-east Normal University, Changchun, 130024, P. R. China

[3]Department of Physics and Chemistry, University of Southern Denmark, Campusvej 55,
5230 Odense M, Denmark

[4]Department of Chemistry and Applied Biosciences—Computational Science, ETH Zürich,
Via Giuseppe Buffi 13, CH-6900 Lugano, Switzerland

*Correspondence to Ulf Ryde, E-mail: Ulf.Ryde@teokem.lu.se,
Tel: +46 – 46 2224502, Fax: +46 – 46 2224543

2011-06-23

Abstract

Continuum-solvation methods are frequently used to increase the efficiency of computational methods to estimate free energies. In this paper, we have evaluated how well such methods estimate the non-polar solvation free-energy change when a ligand binds to a protein. Three different continuum methods at various levels of approximation were considered, viz. the polarised continuum model (PCM), a method based on cavity and dispersion terms (CD), and a method based on a linear relation to the solvent-accessible surface area (SASA). Formally rigorous double-decoupling thermodynamic integration was used as a benchmark for the continuum methods. We have studied four protein–ligand complexes with binding sites of varying solvent exposure, namely the binding of phenol to ferritin, a biotin analogue to avidin, 2-aminobenzimidazole to trypsin, and a substituted galactoside to galectin-3. For ferritin and avidin, which have relatively hidden binding sites, rather accurate non-polar solvation free energies could be obtained with the continuum methods if the binding site is prohibited to be filled by continuum water in the unbound state, even though experiments show that the ligand replaces several water molecules upon binding. For the more solvent exposed binding sites of trypsin and galectin-3, no accurate continuum estimates could be obtained, even if the binding site was allowed or prohibited to be filled by continuum water. This shows that continuum methods fail to give accurate free energies on a wide range of systems with varying solvent exposure, because they lack a microscopic picture of binding-site hydration as well as information about the entropy of water molecules that are in the binding site before the ligand binds. Consequently, binding-affinity estimates based upon continuum solvation methods will give absolute binding energies that may differ by up to 200 kJ/mol depending on the method used. Moreover, even relative energies between ligands with the same scaffold may differ by up to 75 kJ/mol. We have tried to improve the continuum-solvation methods by adding information about the solvent exposure of the binding site or of the hydration of the binding site and the results are promising at least for this small set of complexes.

Introduction

Solvation effects play a major role in many biochemical processes, such as molecular recognition, ligand binding, enzymatic catalysis, and protein folding. Therefore, much effort has been spent to reach a fundamental understanding of such effects and to obtain accurate estimates of the thermodynamics of solvation.[1,2,3] This is of particular interest in theoretical calculations, in which processes are typically studied in vacuum or with a restricted number of explicit solvent molecules: It would be desirable to have methods to extend the results from vacuum or a restricted solvation to a complete solvation.

The free energy of solvation can be calculated by rigorous microscopic free-energy perturbations or thermodynamic integration (TI), employing an appropriate thermodynamic cycle.[2,4,5,6,7] Unfortunately, these methods rely on energy estimates that are hard to converge[8] and the calculations are demanding in terms of computer time.[9,10,11] Therefore, macroscopic approximations have been developed that treat the solvent molecules as a structureless continuum.[7,6,5] The solvation process is typically decomposed into two steps:[12] First, a cavity is created within the solvent that can accommodate the solute and then the apolar solute is introduced in the cavity and interactions between the solute and the solvent are turned on. Three types of interactions are normally considered: electrostatics, dispersion, and exchange-repulsion. For most molecules, the electrostatic effects dominate and therefore the continuum-solvation energy is typically divided into two terms, viz. the polar solvation energy ΔG_{solv} (the electrostatic interactions with the solvent) and the non-polar solvation energy ΔG_{np} (the cavitation energy, as well as the dispersion and exchange-repulsion interactions). Polar continuum-solvation methods such as the Poisson–Boltzmann,[13] the polarised continuum model (PCM),[14] and the generalised Born[15] (GB) methods have been extensively studied and compared.[7,6,5,16,17,18,19] However, the non-polar part of the solvation energy has attracted less interest.[20,21,22,23,24,25,26]

The non-polar part can be treated at several levels of approximation. In the PCM method[14], all three terms are considered, so that ΔG_{np} is estimated from

$$\Delta G_{\text{np}}^{\text{PCM}} = \Delta G_{\text{cavity}} + \Delta E_{\text{rep}} + \Delta E_{\text{disp}} \tag{1}$$

The cavity term is calculated from expressions of the radius of each atom to the power of 0 to 3,[27] i.e., including both area and volume terms, fitted to the hydration energies of non-polar hydrocarbons. The other two terms are continuum approximations to the van der Waals interaction between the solute and solvent. They are calculated by volume integrals, using the average density of the solvent.[28]

It is often assumed that the cavity term depends linearly on some kind of measure of the molecular surface (MS) of the solute.[12,29] The repulsion energy may also show such a dependence and therefore a recent approach merged the repulsion and cavity terms:[23]

$$G_{\text{np}}^{\text{CD}} = \Delta G_{\text{CR}} + \Delta E_{\text{disp}} = \gamma_{\text{CD}} MS + b_{\text{CD}} + \Delta E_{\text{disp}} \tag{2}$$

where γ_{CD} and b_{CD} are fitted parameters. This model will be called the cavity–dispersion (CD) approach in the following. For small molecules, the choice of MS is not important,[23] because in this case areas and volumes are strongly correlated. However, for macromolecules, the solvent-accessible volume has been recommended.[30] The dispersion term is calculated from volume integrals, as in the PCM method.

Despite the fact that PCM and CD make less assumptions about the process, the most popular approach is to relate the entire non-polar solvation energy to the solvent-accessible surface-area (SASA):[31]

$$G_{np}^{SASA} = \gamma_{SASA} SASA + b_{SASA} \qquad (3)$$

where γ_{SASA} and b_{SASA} are fitted parameters, different from those in the CD method.

Recently, it has been observed that different continuum-solvation methods may give qualitatively different estimates of the non-polar solvation. For example, we have shown that SASA and PCM give anti-correlated estimates of ΔG_{np} with opposite signs for the binding of seven biotin analogues to avidin.[32] Calculations with the 3D-RISM method (three-dimensional reference interaction-site model) gave results that supported PCM.[19] In an attempt to explain this discrepancy, we have compared ΔG_{np} calculated by TI, PCM, CD, and SASA for a system that was small enough to be studied accurately with TI, viz. the binding of benzene to T4-lysozyme.[24] These calculations showed that SASA gave more accurate estimates of ΔG_{np} than PCM and CD. However, the reason for this is that all three continuum-solvation methods assume that the binding site in the ligand-free state is filled with continuum water, contrary to experimental observations. This could be avoided by placing a dummy ligand in the binding site that did not interact with the surroundings. Then, the PCM method gave the most accurate non-polar solvation energies.

However, it is not clear how general these results are, i.e. if they are applicable also to proteins with more solvent-accessible binding sites. Therefore, we study in this paper four protein–ligand complexes that have binding sites of varying solvent exposure and a varying number of water molecules in the ligand-free binding site, viz. phenol bound to ferritin, a biotin analogue bound to avidin, 2-aminobenzimidazole bound to trypsin, and a substituted galactoside bound to galectin-3. As can be seen in Figure 1, ferritin has a buried binding cavity that contains four water molecules in the ligand-free state. In avidin, the binding site is still buried, but the opening is larger. In trypsin, the ligand binds in a cleft that is partly solvent exposed, whereas in galectin, the ligand binds on the surface of the protein. The ligands were selected for computational convenience (small and neutral), so that accurate TI estimates can be calculated. For all four systems, well-determined experimental binding affinities are available.[33,34,35,36] We show that none of the continuum-solvent methods can accurately predict the non-polar solvation energies for such a wide range of binding sites, because of the lack of microscopic knowledge of the hydration state and the problem of estimating the entropy of water molecules in the empty binding site. The implication of these findings on binding free-energy predictions will be discussed.

Methods

System preparation. We have studied four protein–ligand systems: phenol bound to ferritin, imidazolidone (a biotin analogue, Btn7) bound to avidin, 2-aminobenzimidazole (ABI) bound to trypsin, and methyl-β-D-thiogalactopyranoside (L19) bound to the carbohydrate-binding domain of galectin-3. The four ligands are shown in Table 1. The phenol-bound ferritin simulations were based on the crystal structure 3f39 and the ligand-free simulations on 3f32.[33] The Btn7–avidin simulations were based on the 1avd crystal,[37] after the modification of the biotin molecule to Btn7 (by removing atoms), and the ligand-free simulations were based on the same crystal. We considered the binding of only a single Btn7 molecule to the A subunit of this tetrameric protein. The ABI–trypsin simulations were based on the 2fx6 crystal and the ligand-free trypsin simulations were based on the same crystal.[35] The L19–galectin-3 simulations were based on the 1kjr crystal,[38] where the co-crystallised ligand was modified to L19. The apo simulation of galectin-3 was taken from our previous study.[39]

The ferritin simulations required some special initial preparation. First, the PDB files

contain only a single monomer whereas in solution, ferritin is a multimer of almost 100 protein chains. However, to save computer time, only a dimer was simulated (as has successfully been done before[33]). The second protein chain was created by applying the appropriate symmetry operations. Second, Gly156 and Ser157 are missing from a flexible loop and they were built into the structure using Swiss-PdbViewer DeepView,[40] followed by a minimization with the sander module of Amber[41] (keeping all other atoms fixed). A few residues were also missing at the N and C terminals, but they were ignored.

In all systems, all Asp and Glu residues were supposed to have a negative charge and all Arg and Lys residues were considered to have a positive charge, although this was checked with PROPKA calculations.[42] The protonation of histidine residues was decided by a study of the local hydrogen-bonding network, the surrounding residues, and the solvent accessibility. In each subunit of ferritin, residues 49, 132 and 147 were doubly protonated, residue 114 was protonated on the ND1 atom, and residue 124 on the NE2 atom. In avidin, the single histidine residue in each subunit was protonated on the NE2 atom.[43] In trypsin, residues 40 and 57 were doubly protonated, whereas residue 91 was protonated on the NE2 atom. In galectin-3, residue 158 was protonated on the ND1 atoms, whereas the other three His residues where protonated on the NE2 atom. The proteins were described by the Amber99SB force field[44] and parameters for the ligands was taken from the generalised Amber force field,[45] except for Btn7 for which parameters were taken from the Amber99 force field.[43,46] Charges on the ligands were determined by the RESP procedure,[47] using ESP points calculated at the Hartree–Fock/6-31G* level and sampled with the Merz–Kollman scheme.[48] All the protein–ligand complexes, free proteins, and free ligands were solvated in a truncated octahedral box of TIP4P-Ewald water molecules,[49] extending at least 10 Å from the protein or ligand.

TI calculations. The TI protocol followed the method designed by Roux et al.,[9] as was thoroughly described in our previous article.[24] In short, we employed a double-decoupling method[9,50] in which the binding free energy is estimated by

$$\Delta G_{\text{bind}} = \Delta G_{\text{ele}}^{\text{free}} + \Delta G_{\text{vdW}}^{\text{free}} - \Delta G_{\text{ele}}^{\text{bound}} - \Delta G_{\text{vdW}}^{\text{bound}} + \Delta\Delta G_{\text{restr}} \tag{4}$$

where the superscripts free and bound indicate simulations of the ligand free in solution and when bound to the protein, respectively. ΔG_{ele} is the change in free energies when the charges of the ligand are zeroed, whereas ΔG_{vdW} is the change in free energy when the van der Waals parameters of the (non-polar) ligand are also zeroed. In the bound simulations, the ligand is restrained relative to the protein (to ensure that it remains in the binding site, even when all interactions with the surroundings are removed) and $\Delta\Delta G_{\text{restr}}$ is the free-energy difference of introducing and removing these restraints. The four first free energies on the right-hand side of Eqn. 4 were evaluated by TI simulations:[4]

$$\Delta G = \int_0^1 \left\langle \frac{\delta V}{\delta\lambda} \right\rangle_\lambda d\lambda \tag{5}$$

where the brackets indicate an ensemble average and $V(\lambda) = (1-\lambda)V_0 + \lambda V_1$. V_0 and V_1 are the potential energies of the initial and final states, respectively and λ is a coupling parameter describing the mixing of the two states. The integral in Eqn. 5 was estimated by simulating the systems at 11 distinct values of λ (0.05, 0.10, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80, 0.90, 0.95), followed by a trapezoid integration. The integrand at $\lambda = 0$ and 1 was estimated by linear extrapolation from the two nearest points. To avoid end-point problems in the van der Waals calculations, soft-core potentials were used, as implemented in Amber.[51]

$\Delta\Delta G_{\text{restr}}$ was estimated from an analytical approximation rather than from TI calculations.[52] The restraints involved three atoms in the protein (P1, P2, and P3) and three atoms in the ligand (L1, L2, and L3). The selected atoms for the four test systems are listed in Table 1. The distance P1–L1, the angles P2–P1–L1 and P1–L1–L2, and the dihedral angles P3–P2–P1–L1, P2–P1–L1–L2, and P1–L1–L2–L3 were harmonically restrained to their values in the crystal structure. Force constants of 41.84 kJ/mol/Å^2 and 836.8 kJ/mol/rad^2 were used for the distance and angle restraints, respectively. These restraints were found suitable for the binding of benzene to T4-lysozyme.[9] However, for the largest ligand (L19), we also tested two and three times larger force constants, but it did not result in any statistically significant differences. $\Delta\Delta G_{\text{restr}}$ was estimated as has been described previously.[24,52]

MM/GBSA calculations. Within the MM/GBSA framework, ΔG_{bind} is estimated from:[53]

$$\Delta G_{\text{bind}} = \langle G^{\text{PL}} - G^{\text{P}} - G^{\text{L}} \rangle_{\text{PL}} \tag{6}$$

where PL is the protein–ligand complex, P is the protein, and L is the ligand. The bracket indicates an ensemble average over an MD simulation of PL. Each free energy is calculated from

$$G = E_{\text{ele}} + E_{\text{vdW}} + G_{\text{solv}} + G_{\text{np}} - TS \tag{7}$$

where the first two terms are the electrostatics and van der Waals molecular mechanics (MM) energies, calculated with infinite cutoff. G_{solv} is the polar solvation energy, estimated either with PCM or GB. We used the GB model of Onufriev et al.,[54] model I (with α = 0.8, β = 0, and γ = 2.91). G_{np} was estimated with the PCM, CD, or SASA methods. For the SASA method, G_{np} was calculated according to Eqn. 3, using γ_{SASA} = 0.0227 kJ/mol/Å^2, b_{SASA} = 3.85 kJ/mol, and a probe radius of 1.4 Å. [55] For the CD method, the σ decomposition scheme was used for separating the repulsion and the dispersion, and radii optimised by Tan et al. was used.[56] The solvent-accessible volume was used for the cavity term, with γ_{CD} = 0.0378 kJ/mol/Å^2, b_{CD} = –0.5692 kJ/mol, and a probe radius of 1.3 Å. For the dispersion term, a probe radius of 0.557 Å was used and the water density was set to 1.129 kg/L. These are recommend values in the Amber software.[23,30] The final term in Eqn. 7, is the absolute temperature multiplied with an entropy estimate, obtained from harmonic frequencies at the MM level on a truncated and buffered system (8 + 4 Å), as described previously, to improve the statistical precision of the estimate.[57] All the terms in Eqn. 7 were averages over the last snapshot from 20 independent MD simulations. The MM/GBSA calculations were performed with the Amber 10 software.[41]

PCM calculations. For the PCM calculations, the integral-equation formulation of PCM, IEFPCM,[58] was used. Because of the large size of the systems, the PCM-induced charges were obtained using a direct inversion of the iterative subspace procedure,[59] as implemented in the GAMESS software.[60] UAKS radii as implemented in Gaussian[61] and a scaling factor for the polar part of 1.15 were used.

Correspondence between TI and MM/GBSA. According to Eqns. 6 and 7, the MM/GBSA non-polar solvation free energy for the binding of L to P is given by

$$\Delta G_{\text{np}}^{\text{MM/GBSA}} = G_{\text{np}}^{\text{PL}} - G_{\text{np}}^{\text{P}} - G_{\text{np}}^{\text{L}} \tag{8}$$

On the other hand, in TI, the non-polar solvation free-energy change is calculated as the free-energy difference when turning off the van der Waals interactions for the ligand free in solution and when it is bound to the protein:

$$\Delta G_{np}^{TI} = \Delta G_{vdW}^{free} - \Delta G_{vdW}^{bound} \qquad (9)$$

Here, the non-polar solvation energy is defined as all contributions to the van der Waals energy that involve solvent water molecules (not only the protein or the ligand).[24]
Consequently, we can identify

$$G_{vdW}^{free} = -\langle G_{np}^{L} \rangle_{PL} = G_{np}^{free} \qquad (10)$$

$$G_{vdW}^{bound} = \langle G_{np}^{P} - G_{np}^{PL} \rangle_{PL} = G_{np}^{bound} \qquad (11)$$

Here, we also define what will be called G_{np}^{free} and G_{np}^{bound} in the following.

Water molecules in the binding site. It is interesting to identify and count the number of water molecules within the binding site of the proteins. Then, we need to define the extent of the binding site in the flexible protein during the MD simulations. This was done by first superimposing each MD snapshot onto the ligand-bound crystal structures and then selecting water molecules within a specific radius from the centre of the ligand. For ferritin, a radius of 6.5 Å was used, so that there were four and two water molecules on the average in the binding site of the ligand-free and ligand-bound states, respectively (in accordance with the crystal structures[33]). For avidin and trypsin, the snapshots were visually inspected and the radius was varied until a proper number of water molecules was found. This resulted in radii of 6 and 4 Å, respectively. For the galectin-3, which does not have any distinct cavity, we selected a radius that gave no water molecules in the bound state, 4.5 Å.

MD simulations. All MD simulations were run with the Amber 10 software.[41] The SHAKE algorithm[62] was used to restrain bonds involving hydrogen atoms, allowing for a 2 fs time step. The temperature was kept constant at 300 K using a Langevin thermostat[63] with a collision frequency of 2.0 ps^{-1} and the pressure was kept constant at 1 atm using a weak-coupling isotropic algorithm[64] with a relaxation time of 1 ps. The non-bonded cut-off was set to 8 Å and the long-range electrostatics were treated by particle-mesh Ewald summation[65] with a fourth-order B spline interpolation and a tolerance of 10^{-5}. The non-bonded pair list was updated every 50 fs.

The MD simulations for the MM/GBSA estimates followed the solvent-induced independent-trajectory approach (SIIT):[66] Twenty independent simulations were initiated by solvating the systems in different water boxes and by giving the atoms different initial velocities. The systems were subjected to 500 cycles of steepest descent, with all atoms, except water molecules and hydrogen atoms, restrained to their initial position with a force constant of 418 kJ/mol/Å^2. This was followed by a 20 ps NPT simulation using the same constraints, a 100 (ferritin and avidin) or 1000 ps (trypsin and galectin-3) NPT equilibration without any restraints, and a 200 ps production run, still at constant pressure. The last snapshot from the production run was used for the MM/GBSA calculations. The equilibration times were selected on the basis of previous simulations of identical or similar systems.[17,66,67]

The TI simulations employed between five or ten independent simulations for each λ

value, and the starting structures were the last snapshot from the MD simulations for MM/GBSA. Ten independent simulations were used for avidin, trypsin, and galectin-3, whereas only five were used for ferritin because a sufficient precision was obtained already at that level. The systems were equilibrated for 20 ps in the NPT ensemble with constraints towards the starting structure on all atoms except water molecules and hydrogen atoms. This was followed by a equilibration of variable length without any restraints and still with constant pressure, and a 200 ps production run, where snapshots were extracted every tenth picosecond. For trypsin and galectin-3, an equilibration of 200 ps was enough to obtain stable energies and a proper number of water molecules in the active site for λ values larger than 0.5. However, for the more hidden cavities of ferritin and avidin, longer equilibration times were required to obtain a proper number of water molecules in the active site (i.e. similar to experiments and long MD simulations), 1 and 4 ns, respectively.

Results

Binding affinities. We have studied the binding of four ligands to four different proteins using rigorous TI calculations and three continuum-solvent methods based on the MM/GBSA framework.[53] From these simulations, the non-polar solvation energy was extracted according to Eqns. 8–9 and the TI results were used as a benchmark for the continuum-solvent methods. To validate such an approach, we must check that the TI calculations reproduce the experimentally determined net binding free energies. In Table 2 we compare the calculated and experimental free energies of binding for the four protein–ligand systems. It can be seen that the experimental and calculated binding energies differ by 6–9 kJ/mol. This is probably the accuracy that can expected for such demanding TI calculations of absolute binding affinities of ligands involving 12–27 atoms.[68] Moreover, the statistical uncertainty is rather large, 5–6 kJ/mol so none of the calculated energies are significantly different from the experimental ones. Fortunately, this precision is enough to study the non-polar contributions, because the standard error of the TI ΔG_{np} estimates is appreciably smaller, 1–2 kJ/mol, as can be seen in Table 3.

In the following, we will discuss the non-polar free energy for each of the four test systems in order of increasing solvent exposure of the binding site.

Phenol binding to ferritin. The ferritin binding site is a hidden cavity that is formed at the interface between two protein subunits as shown in Figure 1. ΔG_{np} calculated by TI and the continuum methods are shown in Table 3. It can be seen that for the bound state, the PCM and CD estimates are much too negative (–77 and –42 kJ/mol, respectively) compared to TI (+1 kJ/mol), whereas the SASA estimate is accurate (+3 kJ/mol).

However, by inspecting the components of PCM in Table 4, it can be seen that the energies of this system are similar to those in our previous investigation of the binding of benzene T4-lysozyme:[24] The PCM estimate is dominated by the cavity term (–54 kJ/mol, column P–PL; cf. Eqn. 11). This energy is clearly an overestimate, because the binding site is preformed even when the ligand is not bound; instead it should be essentially zero. The large cavity term comes from the fact that the continuum methods fill all empty space with continuum water.

As discussed before,[24] this deficiency can be corrected for by introducing a dummy phenol ligand in the PCM calculation of the ligand-free protein. The dummy molecule does not interact with the surroundings (the charges and the Lennard-Jones parameters are zeroed). Calculations with such a ligand are called P0 in the following. From Table 4, it can be seen that it gives a vanishing cavity contribution to the bound non-polar energy (column P0–PL) and the dispersion and repulsion contributions are also reduced, but not fully to zero. The

8

reason for this is that the cavity term depends only on the size and the shape of the cavity. The cavities in the PL and P0 states are identical and therefore, the P0–PL cavity term vanishes. On the other hand, the dispersion and repulsion terms represent the interaction energy between the continuum solvent and the solute. If we divide the solute into the protein (P) and the ligand (L; if present), and the solvent into that replaced by the ligand (S_L) and the rest (S_b), we have for the PL state two terms, PS_b and LS_b. In P, we still have the PS_b term, but the other term is replaced by PS_L because the ligand has been replaced by continuum solvent. Therefore, P–PL, contains two terms, $PS_L - LS_b$. On the other hand, P0 contains only one term, PS_b, because the PS_L term vanishes, owing to that the continuum water in the binding site is replaced with a dummy ligand with zeroed van der Waals parameters. Consequently, P0–PL contains only the $-LS_b$ term, i.e. the negative of the interaction energy between the ligand in the binding site and the surrounding bulk solvent, which is quite far away for this hidden binding cavity, explaining why this term is small. We can estimate the size of the PS_L term from the difference between the P and P0 estimates.

If this P0 approach is used, ΔG_{np}^{bound} estimated by PCM and CD is much improved, as can be seen in Table 3 (columns P0 for CD and PCM). In fact, the PCM result now reproduces the TI results within the uncertainty of the estimates, whereas the error of CD(P0) is reduced to 4 kJ/mol and that of SASA(P0) is 1 kJ/mol (SASA(P0) is always zero for the bound state).

The continuum estimates of ΔG_{np}^{free} for the ligand-free state are rather good, although they are all more negative than the TI results (Table 3). The PCM estimate is best (1 kJ/mol error) and the SASA estimate worst (6 kJ/mol error). The total TI ΔG_{np} is reproduced within 1 kJ/mol by PCM(P0), whereas CD(P0) and SASA(P0) give errors of 7 and 5 kJ/mol.

These good results of the P0 approach are quite unexpected: The P0 approach made sense for the binding of benzene to T4-lysozyme, because no water molecules enter the binding site in the ligand-free state.[24] However, this is not the case for ferritin: On the contrary, the crystal structures show that there are four water molecules in the ligand-free state, two of which are extruded by phenol.[33] This is reasonably reproduced in the TI simulations as can be seen in Table 5: The ligand-binding cavity contains 2.1 waters on average in the bound state and 5.7 water in the ligand-free state.

Therefore, we might expect that a mixture of the P and P0 approaches would be more appropriate, viz. that the cavity term should be taken from the P0 approach (i.e. it should vanish), because the cavity is preformed in the protein, whereas the dispersion and repulsion terms should be taken from the P approach, because there are water molecules that are expelled by the ligand. With PCM, we can test such an approach, because all terms are available in Table 4. However, the result, –23 kJ/mol (i.e. the sum of the dispersion and repulsion terms in the P–PL column in Table 4, shown in the PCM P/P0 column in Table 3) is much worse than that of the pure P0 approach.

Since the dispersion and repulsion terms in PCM are simple volume integrals (i.e. pure energy terms; this approximation has been confirmed in simulations[20,21]), we can compare the continuum estimate with the corresponding energies in the end-point TI simulations by simply calculating the average van der Waals interaction energy between all water molecules in the binding site (as defined in the Method section) and the protein. These energies are also included in Table 5 (column $\Delta E_{vdW}^{PS_L}$). It can be seen that the two water molecules in the bound state (on average) make an interaction energy of –5 kJ/mol with the protein, whereas the water molecules in the apo state make an average interaction energy of –28 kJ/mol. Thus, the difference between the free and bound states, –23 kJ/mol, is similar to the PCM(P) estimate of the PS_L energy term (sum of the dispersion and repulsion terms for P–P0 in Table 4, –24 kJ/mol, included in the PCM column in Table 5).

This shows that the continuum estimate of the PS_L interaction energy is quite accurate. The

reason why we get a poor estimate of ΔG_{np}^{bound} for the combined PCM(P/P0) approach is that we do not consider the entropy of the water molecules in the binding site. They should be included in the cavitation energy (implicitly by the parametrisation), but in the P0 approach, the cavity term vanishes by definition. Therefore, the entropy term is missing in the mixed P/P0 approach. On the other hand, we can conclude that the good results obtained with CD(P0) and PCM(P0) seem to be a coincidence, resulting from the cancellation of two errors, the neglect of the water molecules in the binding site and the entropy gain when these water molecules are released into bulk solvent.

Btn7 binding to avidin. The avidin binding site is also relatively hidden, as can be seen in Figure 1. Crystal structures of the apo state shows a varying number of water molecules in the ligand-free binding site:[69] two in one subunit, one in another, and no water molecules in the other two subunits. The reason for this could be the rather poor quality of the structure, that the water molecules in the binding site are disordered, or that water molecules diffuse readily between the binding site and the bulk solvent.

Therefore, we first studied how easy water molecules diffuse in to and out of the binding site in regular MD simulations. We run five different 5-ns MD simulations, starting with 2, 3, 4, 5, and 6 water molecules in the binding site by replacing the heavy atoms of Btn7 with water molecules. Figure 2 shows the time-evolution of the number of water molecules in these simulations. It can be seen that some of the inserted water molecules are rapidly expelled from the site, giving a minimum in the number of water molecules after ~100 ps. Then, the number increases again, converging to about 7 water molecules, but with a variation between 4 and 10 molecules, and requiring fairly long simulation times. This may indicate that the empty binding cavity is expanding. As a consequence of these results, we employed long equilibration times (4 ns) for the TI calculations to allow water molecules to diffuse into the water site and replace the disappearing ligand.

ΔG_{np}^{bound} from TI and the continuum methods is included in Table 3. Again, the SASA result (1 kJ/mol) reproduces the TI result (–3 kJ/mol) reasonably, whereas the PCM and CD estimates are much too negative (–50 and –29 kJ/mol, respectively). Using the P0 approach, we obtain strongly improved estimates with both PCM and CD, although they now give too positive results (by 4–6 kJ/mol).

Again, this good agreement is unexpected, considering that the binding cavity contains 4–10 more water molecules in the ligand-free state than in the bound state. These water molecules give average van der Waals interaction energies of –3 kJ/mol and –24 kJ/mol in the bound and free states, respectively (Table 5). This amounts to a PS_L interaction energy difference of –22 kJ/mol. Interestingly, this is 10 kJ/mol larger (more negative) than the continuum estimate (–12 kJ/mol; column PCM in Table 5). This indicates that there are more water molecules in the ligand-free binding site than what fits the bound binding site, i.e. that the binding site contracts when the ligand is bound. Some experimental studies show that one of the loops of the avidin binding site is disordered in the unbound state,[70,71,72] which could explain our results. Of course, the MM/GBSA approach, based on the simulations of only the bound state, can never include such an effect. In the passing, it should be mentioned that the predicted number of water molecules in the bound and unbound state matches what have been found in the related streptavidin protein, both by experiment and calculations.[73,74,75]

For the ligand-free state, the SASA estimate of ΔG_{np}^{free} is 11 kJ/mol too negative, whereas the PCM and CD estimates are 1–3 kJ/mol too positive. The total ΔG_{np} is well reproduced by both CD(P0) and PCM(P0) with errors of 2–3 kJ/mol, whereas SASA(P0) has an error of 13 kJ/mol. These results confirm our previous conclusion that PCM(P) gives worse non-polar solvation energies for avidin than SASA(P),[32] but we also see that the PCM(P0) results are better. 3D-RISM also gives much too positive non-polar solvation energies for the binding of

Btn7 to avidin.[19]

2-aminobenzimidazole binding to trypsin. Trypsin has three binding pockets and the ABI ligand binds to the most buried P1 pocket.[76] This pocket is a shallow cleft, as can be seen in Figure 1.

ΔG_{np} from TI and the continuum methods are shown in Table 3. For this protein, the SASA estimate of ΔG_{np}^{bound} is no longer accurate; instead it has the wrong sign and is 21 kJ/mol too positive. The PCM and CD methods predict the correct sign of the energy, but the estimates are too negative by 57 and 21 kJ/mol, respectively. With the P0 approach , the PCM and CD methods predict the wrong sign, although the PCM result is closer to the TI estimate. Thus, the TI non-polar solvation energy is in between the continuum P and P0 estimates. This indicates that the binding site has a specific hydration pattern that cannot be accurately represented by continuum-solvation methods.

Several crystal structures of trypsin show conserved water molecules in the binding site that can be displaced by ligands.[77,78] We performed a 5 ns simulation of ligand-free trypsin to investigate the binding-site hydration and found on average seven water molecules in the site. This is similar to what was observed in the TI calculations (Table 5). As the trypsin binding site is solvent exposed, the number of water molecules was stable during the course of the simulation. Using a similar definition for the bound state, 0.5 water molecules were found in the binding site when the ligand was bound.

The van der Waals interaction energies between these water molecules and the protein are also shown in Table 5. The water molecules in the apo and bound simulations had interaction energies of –29 kJ/mol and –2 kJ/mol, respectively. Thus, the PS_L energy difference amounts to –28 kJ/mol. This is rather close to the PCM continuum estimate in Table 5, –24 kJ/mol. In fact, for this protein, we obtain a reasonable estimate of ΔG_{np}^{bound} from the mixed P/P0 approach, viz. –17 kJ/mol (the sum of the dispersion and repulsion PCM energies in column P–PL of Table 4), compared to the TI estimate of –19 kJ/mol. However, for this solvent-exposed cleft, it is no longer evident that the cavitation energy of the bound state should vanish.

For the ligand-free state, the ΔG_{np}^{free} estimates are scattered. The SASA estimate is 11 kJ/mol too negative, the PCM estimate is 3 kJ/mol too negative, whereas the CD estimate is 2 kJ/mol too positive. Consequently, all the six P and P0 continuum-solvation methods give errors in the total non-polar solvation energy of 23–54 kJ/mol, whereas the error of the mixed P/P0 approach is 5 kJ/mol.

L19 binding to galectin-3. Finally, we considered the binding of L19 to galectin-3. The binding site of galectin-3 is located on the surface of the protein as can be seen in Figure 1. From experiments and MD simulations it is known that the positions of the water oxygen atoms of the ligand-free binding site of galectin-3 mirror the oxygen positions of the ligand in the bound state.[79] For L19, this would correspond to at least six water molecules.

ΔG_{np} from TI and the continuum-solvation methods is shown in Table 3. In the bound state, the SASA estimate is 16 kJ/mol too positive. The negative sign of the SASA estimate shows that this is the first case for which the SASA of the PL complex is larger than the SASA of P. The PCM and CD estimates are, as in the trypsin case, too negative but with the correct sign. Using the P0 approach, PCM and CD give energies that are too positive, but the PCM result is closer to the TI results.

The components of the PCM estimate in Table 4 show that ΔG_{np}^{bound} is almost entirely dominated by a cavity term of –98 kJ/mol, whereas the dispersion is only 4 kJ/mol and the

exchange-repulsion is zero. As discussed above, the latter two terms contain two contributions, viz. $PS_L − LS_b$. According to the P0–PL column, the $−LS_b$ terms amount to 45 and −9 kJ/mol, respectively, so both terms are sizeable. Thus, the near cancellation is coincidental.

We investigated the hydration of the ligand-free state of galectin-3, using MD trajectories from a previous study.[79] It was found that on average ~10 water molecules occupied the same space as L19 in the apo state. This is reproduced by the TI simulations (Table 5). The difference in the van der Waals interaction between these water molecules and the protein before and after ligand binding is −16 kJ/mol, which is 16 kJ/mol less negative than the PCM continuum estimate. This large difference is quite unexpected for this solvent-exposed surface.

The estimates of ΔG_{np}^{free} are again scattered: Only the CD method gives reasonable results with an error of 3 kJ/mol, whereas the SASA and PCM estimates are too negative by 12 and 20 kJ/mol, respectively. It is not clear why SASA and PCM give so poor results for this ligand. Consequently, again all methods give large errors for the total non-polar solvation energy, 22–73 kJ/mol. The best results are obtained for the CD(P) approach.

A mixture of the P0 and P approaches based on the solvent exposure of the ligand. The results in Table 3 showed that for the two systems with relatively hidden binding sites, ferritin and avidin, as well as for the binding of benzene to lysozyme,[24] accurate results were obtain using the P0 cavity for the unbound protein, especially with the non-polar terms from PCM. However, for trypsin and galectin-3, with their more solvent-exposed binding sites, the TI results are somewhere between the P or P0 estimates. Apparently, no single method gives reliable results for all systems. However, the results in Table 3 indicate that the TI results come closer to the P results the more solvent-exposed the binding site is.

This indicates that we may obtain more reliable estimates by a linear combination of the P and P0 estimates:

$$G_{np}^{bound} = \xi\, G_{np}^{bound}(P) + (1-\xi)\, G_{np}^{bound}(P0) \tag{13}$$

This equation can be solved for the ξ parameter, using the TI results for each protein on the left-hand side. The results for the PCM and CD are shown in Table 6, where we have included also our previous results for the benzene–lysozyme system.[24] SASA is excluded because G_{np}^{bound} for SASA(P0) is always zero and therefore ξ becomes large and negative for many systems. We see that for PCM, T4-lysozyme, ferritin, and avidin are optimal at $\xi \approx 0$, trypsin at $\xi = 0.32$, and galectin-3 at $\xi = 0.40$. For CD, the optimal ξ values are consistently higher, mainly because the magnitude of the CD results is always smaller (cf. Table 3).

Such an approach is meaningful only if we can relate ξ to some property of the protein–ligand system so that ξ could be predicted beforehand. We expect that the property should describe the solvent exposure of the ligand in the complex. One simple approach is to calculate the ratio between the SASA of the ligand when it is bound to the protein and when it is free in solution, which we will call the solvent exposure (SE) in the following. The SE for each protein–ligand complex is also shown in Table 6. The correlation between ξ (fitted to TI using the data from CD or PCM) and SE is shown in Figure 3. The correlations are fair with correlation coefficients (r^2) of 0.70 and 0.74 for PCM and CD, respectively, but trypsin seems to fall outside the lines. The van der Waals area was also tested, but the correlations were worse.

If we estimate ξ from SE, using the linear relations in Figure 3 and insert these $\xi(SE)$ values into Eqn. 13, we can calculate G_{np}^{bound} for both PCM and CD. The results are shown in

the last two columns of Table 6 (presented as the differences from the TI results). It can be seen that this approach gives good results for all proteins (errors 0–7 kJ/mol), except for trypsin, for which the error is 15–22 kJ/mol. Thus, this approach is promising, although not yet fully satisfying.

Explicitly-sampled interactions PCM. We have seen that it is hard to find a continuum-solvation method that gives accurate predictions for the non-polar solvation energy during ligand binding for all four protein–ligand complexes. Therefore, we tested whether the results can be improved by using some data from the MD simulations. We formulate this explicitly-sampled interactions PCM (ESI–PCM) method as

$$\Delta G_{np}^{bound} = \Delta G_{np}^{PCM}(P0) - \Delta G_{np}^{PCM}(PL) + \langle E_{vdW}^{PS_L} \rangle_P - \langle E_{vdw}^{PS_L} \rangle_{PL} + \Delta G_{np}^{S_L} \tag{14}$$

Here, the first two terms on the right-hand side are the PCM(P0) estimate of ΔG_{np}^{bound} (shown in Table 3). Such an estimate does not consider the water molecules in the binding site, and therefore is incomplete. The next three terms try to estimate the contribution from those water molecules. The two $\langle E_{vdW}^{PS_L} \rangle$ terms are the van der Waals interaction between the binding-site water molecules and the protein, taken from the TI simulations of P or PL, shown in Table 5. Thus, we use in this approach the explicitly sampled interaction energies, rather than the continuum estimate, used in the mixed P/P0 approach (the $\Delta E_{vdW}^{PS_L}$ column in Table 5). To these terms, we add the last term, which attempts to estimate the entropy of the binding-site water molecules.

We tried different approaches to estimate this term. The difference in the number of water molecules in the binding site with and without the ligand gave a poor correlation ($r^2 = 0.23$). With SE instead, the agreement with TI was improved ($r^2 = 0.77$). Finally, we estimated the entropy from a linear relation to the SASA of the ligand, SASA(L):

$$\Delta G_{np}^{S_L} = \gamma_{ESI} SASA(L) + b_{ESI} \tag{15}$$

The coefficients were determined by a linear regression to the TI estimate of G_{np}^{bound}, using the other terms in Eqn. 14. This gave $\gamma_{ESI} = -0.32$ kJ/mol/Å^2 and $b_{ESI} = 84.1$ kJ/mol. As can be seen in Figure 4, this gave a reasonable correlation ($r^2 = 0.83$) and a maximum deviation (for lysozyme) of 15 kJ/mol.

The difference between the ESI-PCM and TI results are shown in Table 7. It is clear that the parametrisation using the SASA of the ligand gives the best overall results. The method works very well for the avidin, trypsin, and galectin-3 with deviations less than 3.5 kJ/mol, whereas for T4-lysozyme and ferritin, the deviations are larger than 10 kJ/mol.

Overall performance. In Table 8, we list the mean absolute deviation (MAD) of ΔG_{np} calculated with the three continuum methods relative to the TI results. It can be seen that SASA (with both P and P0) on average gives a much better result for the bound state than any of the other methods. This could explain why such a simple model has been so successful. Usually, the SASA estimates are small (~1 kJ/mol). This is favourable for the ferritin and avidin cases, for which TI also gives small results. However, for trypsin and galectin-3, the deviations are much larger. The SE approach gives MADs of 7 and 6 kJ/mol for PCM and CD, respectively, i.e. slightly better than the SASA results. The ESI-PCM approach (using the SASA of the ligand) also gives a MAD of 7 kJ/mol. However, both the SE and ESI–PCM approaches need to be tested with more data to ensure that the parametrisations are reliable.

For the free state, the SASA estimates are worse (MAD = 9 kJ/mol) than both PCM and

CD (MAD = 5 and 3 kJ/mol, respectively), highlighting that the two latter methods were parametrised on small molecules. Yet, this does not compensate the errors for the bound state, so that SASA still gives the best estimates of the total ΔG_{np}, although all methods give MADs of 16–58 kJ/mol, except the PCM(P/P0) approach that gives a MAD of 11 kJ/mol. This is probably the most important conclusion of this investigation, viz. that no continuum method gives reliable results for ΔG_{np} for binding sites of varying solvent-accessibility. The SE-CD and SE-PCM methods give MADs of 5 and 9 kJ/mol, and ESI-PCM gives a MAD 10 kJ/mol, again indicating that they are promising approaches.

Seven biotin analogues and relative affinities. The present results allow us to understand the anti-correlation observed previously for the non-polar solvation-energy contribution to the binding energy between SASA on the one hand and PCM or 3D-RISM on the other hand.[19,32] In Figure 5, the non-bonded contributions to the binding energies are shown for the SASA, CD, and PCM, using both the P and P0 approaches. The test case is the same as studied before, i.e. the binding of seven biotin analogues (Btn1–Btn7, also shown in Figure 5) to avidin. We see that PCM(P) and CD(P) give large and positive energies, whereas the other four approaches give small and negative energies. Moreover, the data of these two groups of methods give roughly anti-correlated results, which is most pronounced for the largest (Btn4) and smallest (Btn7) ligands.

The reason for this behaviour is that all the individual non-polar solvation terms are correlated to the size of the ligand, but they have different signs. The PCM(P) and CD(P) results are strongly dominated by the PL–P term, which is large and positive (cf. Table 4 for Btn7). The PCM dispersion and cavity terms dominate and have the same sign, whereas for CD(P), the dispersion term is largest. On the other hand, for the other methods, the PL–P or PL–P0 term nearly cancels (only the small LS_b term remains). Therefore, the total non-polar energy is essentially the negative of the non-polar solvation energy of the ligand, which is small and positive for all ligands (except for Btn7 with PCM(P0) and CD(P0), for which it is positive, but close to zero).

However, the most interesting conclusion from Figure 5 is that even for ligands with a similar scaffold, there will be a qualitative difference between different non-polar methods in general, and between the P and P0 approaches in particular. For avidin, the P0 approach probably gives the most accurate results, in accordance with our previous findings.[19,32] However, for more solvent-accessible binding sites, we have seen that neither the P approach, nor the P0 approach give accurate results. In such cases, we will not even know whether the non-polar solvation energy should increase or decrease with the size of the ligand. Thus, the results in Figure 5 shows that continuum-solvation method will not give reliable results even for relative binding affinities of related ligands.

Conclusions

We have evaluated how well three continuum-solvation methods, PCM, CD, and SASA, estimate the non-polar solvation free-energy change upon ligand binding. As reference data, we used microscopic TI calculations. We have studied four different protein–ligand systems with varying solvent exposure of the bound ligand: galectin-3, which binds its ligand on the protein surface, trypsin, which bind the ligand in a partly solvent-exposed cleft, and avidin and ferritin, which bind their ligands in an increasingly hidden cavity, displacing a varying number of water molecules, cf. Figure 1. Together with our previous study of T4-lysozyme,[24] which binds its ligand in a hidden cavity that is empty before the ligand binds, this should include most typical topologies of ligand binding.

For the systems with hidden binding sites, lysozyme, ferritin and avidin, we obtain rather accurate non-polar solvation free energies by the SASA method (error less than 14 kJ/mol for

the total ΔG_{np}), especially for the bound state (error less than 4 kJ/mol). The reason for this is mainly that ΔG_{np}^{bound} is small. CD and PCM give very poor results for ΔG_{np}^{bound}, because they assume that the cavity is filled with continuum solvent when the ligand is not bound. This problem can be avoided by filling the binding site with a non-interacting ligand (the P0 approach). Then, PCM gives both total and bound ΔG_{np} that are within 4 kJ/mol of the TI result. However, for ferritin and avidin, these good results are fortuitous, because they neglect the water molecules that are displaced by the ligand.

For the other two proteins with solvent-exposed binding sites, trypsin and galectin-3, none of the continuum solvation methods give accurate results, neither if the cavity was filled with continuum water, nor with a non-interacting ligand (errors of 22–73 kJ/mol for the total ΔG_{np}). Instead, the TI results were between the P and P0 results. We have tried to quantify this mixing by a parameter, correlated to the SE of the bound ligand. Such an approach gave errors of 1–7 kJ/mol for all proteins, except trypsin (errors of 15–23 kJ/mol). Still, this is the best result obtained for any of the continuum-solvation methods tested (MADs of 5–9 kJ/mol). It is notable that neither of the continuum-solvation methods works well for the solvent-exposed binding site of galectin-3, which could be expected to be easiest. This shows that even in a solvent-exposed binding site, the properties of water molecules are not bulk-like.

We have also shown that the PCM estimates can be improved by including interaction energies from explicit MD simulations (ESI–PCM; MAD = 10 kJ/mol). However, the parametrisation of an entropic term is non-trivial, and the test set is too small to be reliable. Despite these drawbacks, we have shown that ESI–PCM could be used to reduce the error of the PCM calculations. It is also clear that the PCM approach with its three non-polar energy terms is easier to connect to explicit simulations, because there is a direct correspondence between the dispersion and repulsion terms and the explicit van der Waals interaction energy. If the number of terms should be reduced, it seems more natural to join the dispersion and repulsion terms than the repulsion and cavity terms (as in the CD approach). This would avoid the need of separating these two terms in the simulations, which is ambiguous.[24] Moreover, the dispersion and repulsion typically have opposite signs, which would make the term smaller and therefore more stable and easier to parametrise. However, it must not be forgotten that the cavity term includes the entropy of displaced water molecules.

It should be noted that the problems observed in this paper for the non-polar solvation energy also apply for the polar part of the solvation energy. In particular, the P0 approach will also change the polar solvation energy, because the site is filled with a non-polar dummy ligand, instead of polar water molecules. Thus, if the binding site is (partly) filled with water molecules that are displaced by the ligand, the electrostatic interaction energy of these ligands needs also to be considered (i.e. an electrostatic PS_L term). This term can be appreciably larger than its non-polar counterpart, exaggerating these problems.

The present results have strong bearings on all methods that use continuum-solvation methods to estimate ligand-binding affinities, e.g. MM/PB(GB)SA.[53] Our results show that such methods will have severe problems to give accurate affinity predictions. In particular, absolute affinities will be poor for many proteins, because we have seen that different continuum-solvation methods give non-polar energies that differ by up to almost 200 kJ/mol. Similar problems with absolute MM/PBSA energies have been reported also for the entropy[80,81] and polar solvation terms.[19] Even worse, the results for the seven biotin analogues in Figure 5 show that relative energies will also vary by up to 75 kJ/mol between different methods. For proteins with a hidden cavity, we have seen that the SASA and P0 approaches give the best results, but for more solvent-exposed cavities, the best result is somewhere between the P and P0 approaches. Then, it is not even clear whether the non-polar contribution to the binding energy should increase or decrease with the size of the ligand (and

the ideal balance between P and P0 probably depends on the ligand). This is a most unfortunate result for continuum-solvation-based ligand-affinity methods.

Supporting Information

Listed authors of refs. 35 and 46. The starting structure of the ferritin dimer after minimisation of the added residues. This material is available free of charge via the Internet at http://pubs.acs.org.

References

1   Cramer, C. J.; Truhlar, D. G. *Chem. Rev.* **1999**, *99*, 2161-2200.
2   Orozco, M.; Luque, F. J. *Chem. Rev.* **2000**, *100*, 4187-4225.
3   Tomasi, J.; Mennucci, B.; Cammi, R. *Chem. Rev.* **2005**, *105*, 2999-3093.
4   Kirkwood, J. H. *J. Chem. Phys.* **1935**, *3,* 300-314.
5   Jorgensen, W. L. *J. Phys. Chem.* **1983**, *87*, 5304-5314.
6   Bhalachandra, L. T.; McCammon, J. A. *Comput. Chem.* **1984**, *4*, 281-283.
7   Jorgensen, W. L. *Acc. Chem. Res.* **1989**, *22*, 184-189.
8   Warshel, A.; Sharma, P. K.; Kato, M.; Parson, W. W. *Biochim. Biophys. Acta* **2006**, *1764*, 1647-1676.
9   Deng, Y.; Roux, B. *J. Chem. Theory Comput.* **2006**, *2*, 1255-1273.
10 Fujitani H.; Tanida, Y.; Ito, M.; Jayachandran G.; Snow, C. D.; Shirts, M. R.; Sorin, E. J.; Pande, V. S. *J. Chem. Phys.,* **2005**, *123*, 804108.
11 Zhao, L.; Caplan, D. A.; Noskov, S. Y. *J. Chem. Theory. Comput.* **2010**, *6,* 1900-1914.
12 Westergren, J.; Lindfors, L.; Höglund, T.; Lüder, K.; Nordholm, S.; Kjellander, R. *J. Phys. Chem. B* **2007**, *111*, 1872-1882.
13 Sharp, K. A.; Honig, B. *Annu. Rev. Biophys. Biophys. Chem.* **1990**, *19*, 301-332.
14 Barone, V.; Cossi, M.; Tomasi, J. *J. Chem. Phys.* **1997,** *107*, 3210-3221.
15 Bashford, D.; Case, D. A. *Annu. Rev. Phys. Chem.* **2000**, *51*, 129-152.
16 Klamt, A.; Mennucci, B.; Tomasi, J.; Barone, V.; Curutchet C.; Orozco, M.; Luque, F. J. *Acc. Chem. Res.* **2009**, *42,* 489-492.
17 Gohlke, H.; Case, D. A. *J. Comput. Chem.* **2003**, *25*, 238-250.
18 Kongsted, J.; Söderhjelm, P.; Ryde U. *J. Comp.-Aided Mol. Design* **2009,** *23*, 395-409.
19 Genheden, S.; Luchko, T.; Gusarov, S.; Kovalenko, A.; Ryde, U. *J. Phys. Chem. B* **2010**, *114*, 8505-8516.
20 Gallicchio, E.; Kubo, M. M.; Levy, R. M. *J. Phys. Chem. B* **2000**, *104*, 6271-6285.
21 Levy, R. M.; Zhang, L. Y.; Gallicchio, E.; Felts, A. K. *J. Am. Chem. Soc.* **2003**, *125*, 9523-9530.
22 Mobley, D. L.; Bayly, C. I.; Cooper, M. D.; Shirts, M. R.; Dill, K. A. *J. Chem. Theory Comput.* **2009**, *5*, 350-358.
23 Tan, C.; Tan, Y.-H.; Luo, R. *J. Chem. Phys. B* **2007**, *111*, 12263-12274.
24 Genheden, S.; Kongsted, J.; Söderhjelm, P.; Ryde, U. *J. Chem. Theory Comput.* **2010**, *114* 3558–3568.
25 Wagoner, J. A.; Bake, N. A. *Proc. Nat. Am. Soc.* **2006**, *103*, 8331-8336.
26 Sitkoff, D.; Sharp, K. A.; Honig, B. *J. Phys. Chem.* **1994**, *98*, 1978-1988.

27 Cossi, M.; Tomasi, J.; Cammi, R. *Int. J. Quant. Chem. Quant. Chem. Symp.* **1994**, *29*, 695-702.

28 Floris, F.; Tomasi, J. *J. Comput. Chem.* **1989**, *10*, 616-627

29 Westergren, J.; Lindfors, L.; Höglund, T.; Lüder, K.; Nordholm, S.; Kjellander, R. *J. Phys. Chem. B* **2007**, *111*, 1872-1882.

30 Wang, J.; Cai, Q.; Ye X.; Hsieh, M.-J., Tan, C.; Luo, R. *Amber Tools User's Manual, Version 1.4* **2010**, 143-150.

31 Hermann, R. B. *J. Phys. Chem.* **172**, *76*, 2754-2759.

32 Söderhjelm, P.; Kongsted, J.; Ryde U. *J. Chem. Theory Comput.* **2010,** *6*, 1726-1737.

33 Vedula, L. S.; Brannigan, G.; Economou, N. J.; Xi, J.; Hall, M. A.; Liu, R.; Rossi, M. J.; Dailey, W. P.; Grasty, K. C.; Klein, M. L; Eckenhoff, R. G.; Loll, P. J. *J. Bio. Chem.* **2009**, *284*, 24176-24184.

34 Green, N. M. *Adv. Prot. Chem.* **1975**, *29*, 85-133.

35 McGrath, M. E.; et al. *Biochem.* **2006**, *45*, 5964-5973.

36 Öberg, C. T.; Leffler, H.; Nilsson, U. J. *J. Med. Chem.* **2008**, *51*, 2297-2301.

37 Pugliese, L.; Coda, A.; Malcovati, M.; Bolognesi, M. *J. Mol. Biol.* **1993**, *231*, 698-710.

38 Sorme, P.; Arnoux, P.; Kahl-Knutsson, B.; Leffler, H.; Rini, J. M.; Nilsson, U. J. *J. Am. Chem. Soc.* **2005**, *127*, 1737-1743.

39 Genheden, S.,; Diehl, C.; Akke, M.; Ryde, U *J. Chem. Theory. Comput.* **2010**, *6*, 2176–2190.

40 Guex, N.; Peitsch, M. C. *Electrophoresis* **1997**, *18*, 2714-2723.

41 Case, D. A.; et al. Amber 10, University of California, San Francisco, 2008.

42 Li, H.; Robertson, A. D.; Jensen, J. H. *Proteins* **2005**, *61*, 704-721.

43 Weis, A.; Katebzadeh, K.; Söderhjelm, P.; Nilsson, I.; Ryde U. *J. Med. Chem.* **2006,** *49*, 6596–6606.

44 Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. *Proteins: Struct. Funct., Bioinform.* **2006**, *65*, 712-725

45 Wang, J. M.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. *J. Comput. Chem.* **2004**, 25, 1157-1174.

46 Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem Soc.* **1995**, *117*, 5179-5197

47 Bayly, C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. A. *J. Phys. Chem.* **1993**, *97*, 10269-10280.

48 Besler, B. H.; Merz, K. M.; Kollman, P. A. *J. Comput. Chem.* **1990**, *11*, 431-439.

49 Horn, H. W.; Swope, W. C.; Pitera, J. W.; Madura, J. D.; Dick, T. J.; Hura, G.; Head-Gordon, T. *J. Chem. Phys.* **2004**, *120*, 9665-9678.

50 Gilson, M. K.; Given, J. A.; Bush, B. L.; McCammon, J. A. *Biophys. J.* **1997**, *72*, 1047-

1069.

51 Steinbrecher, T.; Mobley, D. L.; Case, D. A. *J. Chem. Phys.* **2007**, *127*, 214108-13.

52 Wang, J.; Deng, Y.; Roux, B. *Biophys. J.* **2006**, *91*, 2798-2814.

53 Kollman, P. A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W.; Donini, O.; Cieplak, P.; Srinivasan, J.; Case, D.A.; Cheatham III, T. E. *Acc. Chem. Res.* **2000**, *33*, 889-897.

54 Onufriev, A.; Bashford, D. ; Case, D. A. *Proteins* **2004**, *55*, 383-394.

55 Kuhn, B.; Kollman, P. A. *J. Med. Chem.* **2000**, *43*, 3786-3791.

56 Tan, C. H.; Tang, L. J.; Luo, R. *J. Phys. Chem. B* **2006**, *110*, 18680-18687

57 Kongsted, J.; Ryde, U. *J. Comput.-Aided Mol. Des.* **2009**, *23*, 63-71.

58 Cances, E.; Mennucci, B.; Tomasi, J. *J. Chem. Phys.* **1997**, *107*, 3032-3041.

59 Li, H.; Pomelli, C. S.; Jensen, J. H. *Theor. Chem. Acc.* **2003**, *109*, 71-84.

60 Schmidt, M. W.; Baldridge, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. H.; Koseki, S.; Matsunaga, N.; Nguyen, J. A.; Su, S.; Windus, T. L.; Dupius, M.; Montgomery, J. A. *J. Comput. Chem.* **1993**, *14*, 1347-1363.

61 Frisch, A. E.; Frisch, M. J.; Trucks, G. W. *Gaussian 03 User's Reference*; Gaussian Inc.: Wallingford, CT, USA, 2003, p. 205.

62 Ryckaert, J. P; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327-341.

63 Wu, X.; Brooks, B. R. *Chem. Phys. Lett.* **2003**, *381*, 512-518.

64 Berendsen, H.J.C.; Postma, J.P.M.; van Gunsteren, W. F.; DiNola, A.; Haak, J.R. *J Chem Phys* **1984**, *81*, 3684–3690.

65 Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089-10092.

66 Genheden S., Ryde, U. *J. Comput. Chem.,* **2011**, *32*, 187-195.

67 Genheden, S., Ryde, U. *J. Comput. Chem.* **2010**, *31*, 837–846.

68 Merz, K. M. *J. Chem. Theory Comput.* **2010**, *6*, 1769-1776.

69 Pugliese, L.; Malcovati, M.; Coda, A.; Bolognesi, M. *J. Mol. Biol.* **1994**, *235*, 42-46

70 Livnah, O.; Bayer, E. A.; Wilcheck, M.; Sussman, J. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 5076-5080

71 Repo, S.; Paldanius, T. A.; Hytönen, V. P.; Nyholm, T. K. M.; Halling, K. K.; Huuskonen, J.; Pentikäinen, O. T.; Rissanen, K.; Peter Slotte, J.; Airenne, T. T.; Salminen, T. A.; Kulomaa, M. S.; Johnson, M. S. *Chem. Biol.* **2006**, *13*, 1029-1039.

72 Nardone, E.; Rosano, C.; Santambrogio, P.; Curnis, F.; Corti, A.; Magni, F.; Siccardi, A. G.; Paganelli, G.; Losso, R.; Apreda, B.; Bolognesi, M.; Sidoli, A. Arosio, P. *Eur. J. Biochem.* **1998,** *256*, 453-460.

73 Weber, P. C.: Ohlendorf. D. H.; Wendoloski, J. J.; Salemme, F. R. *Science* **1989**, *243*, 85.

74 DeChancie, J.; Houk, K. N. *J. Am. Chem. Soc.* **2007,** *129*, 5419.

75 Li, Q.; Gusarov, S., Evoy, S., Kovalenko, A. *J. Phys. Chem. B* **2009**, *113*, 9958-9967.

76 Katz, B. A.; Mackman, R.; Luong, C.; Radika, K.; Martelli, A., Sprengeler, P. A.; Wang, J.; Chan, H.; Wong, L. *Chem. Bio.* **2000**, *7*, 299-312.

77 Mackman, R. L.; Katz, B. A.; Breitenbucher, J. G.; Hui, H. C.; Verner, E.; Luong, C.; Liu, L.; Sprengeler, P. A. *J. Med. Chem.* **2001**, *44*, 3856-3871.

78 Matter, H.; Defossa, E.; Heinelt, U.; Blohm, P.-M.; Schneider, D.; Müller, A.; Hreok, Si.; Schreuder, H.; Liesum, A.; Brachvogel, V.; Lönze, P.; Walser, A.; Al-Obeidi, F.; Wildgoose, P. *J. Med. Chem.* **2002**, *45*, 2749-2769

79 Genheden, S.; Diehl, C.; Akke, M.; Ryde, U. *J Chem. Theory Comput.* **2010**, *6*, 2176-2190.

80 Gohlke, H.; Case, D. A *J. Comput. Chem.* **2004**, *25*, 238-250.

81 Singh, N.; Warshel, A. *Prot.* **2010**, *78*, 1705-1723.

Table 1. Considered systems and atoms used in the restraints.

Protein	Ligand	P1	P2	P3
Ferritin	(phenol structure; L1, L2, L3, OH)	Arg57 CA	Ser2B N	Gly45B N
Avidin	(imidazolidinone structure; L1 with O, HN, NH L3, L2)	Val115B CG2	Glu53A C	Gly21D N
Trypsin	(benzimidazole structure; L1–NH₂, L2, L3, N, NH)	Lys109 C	Gly18 C	Ser96 C
Galectin-3	(sugar structure; OH, OH, O, L1 SMe, L2, L3, HO, OH)	Phe159 CB	Leu242 CA	Lys226 O
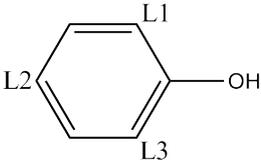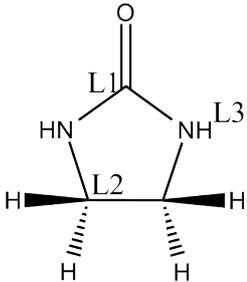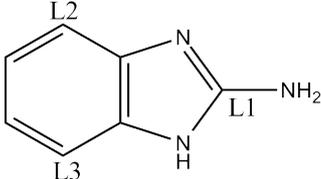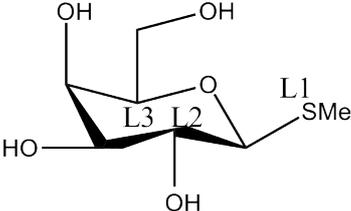
The ligand atoms that were used in the restraints (LI, L2, and L3) are shown in the ligand picture. The corresponding protein atoms are designated P1, P2, and P3. The letters after the residue number for ferritin and avidin indicate the subunit of the protein.

Table 2. Binding free energies calculated with TI, compared to experimental estimates (in kJ/mol).

	TI	Experimental	Difference
Ferritin	-21.7±4.6	-27.3[33]	5.6
Avidin	-12.8±5.6	-18.8[34]	6.0
Trypsin	-13.3±5.7	-4[35]	-9.3
Galectin-3	-3.4±4.7	-12.8[36]	9.4

Table 3. ΔG_{np} calculated with different methods (in kJ/mol).

Protein	Term	SASA P	SASA P0[c]	CD P	CD P0[c]	PCM P	PCM P0[c]	PCM P/P0[d]	TI
Ferritin	Bound[a]	2.5±0.0	0.0±0.0	-42.1±0.6	5.3±0.2	-77.0±1.3	1.2±0.1	-22.6±0.8	1.1±1.0
	Free[b]	-9.7±0.0	-9.7±0.0	-5.8±0.1	-5.8±0.1	-4.4±0.1	-4.4±0.1	-4.4±0.1	-3.4±0.9
	Total	-12.2±0.0	-9.7±0.0	36.3±0.6	-11.1±0.2	72.6±1.3	-5.6±0.1	18.2±0.8	-4.4±1.4
Avidin	Bound[a]	1.3±0.0	0.0±0.0	-29.0±0.7	3.1±0.1	-50.1±0.9	1.2±0.2	-11.1±0.6	-2.5±0.9
	Free[b]	-9.2±0.0	-9.2±0.0	3.8±0.1	3.8±0.1	2.5±0.1	2.5±0.1	2.5±0.1	1.3±0.7
	Total	-10.6±0.0	-9.2±0.0	32.8±0.7	0.7±0.2	52.6±0.9	1.3±0.2	13.6±0.6	3.7±1.1
Trypsin	Bound[a]	1.8±0.0	0.0±0.0	-39.7±0.5	11.2±0.3	-75.6±1.1	7.8±0.7	-16.6±0.4	-18.8±2.0
	Free[b]	-10.7±0.0	-10.7±0.0	2.8±0.1	2.8±0.1	-2.5±0.1	-2.5±0.1	-2.5±0.1	0.6±0.7
	Total	-12.5±0.0	-10.7±0.0	42.5±0.6	-8.4±0.4	73.1±1.1	-10.4±0.7	14.1±0.4	19.4±2.1
Galectin-3	Bound[a]	-0.7±0.0	0.0±0.0	-36.1±0.7	47.7±0.6	-94.8±1.3	36.0±0.5	3.7±0.9	-16.9±1.4
	Free[b]	-12.7±0.0	-12.7±0.0	1.7±0.2	1.7±0.2	-20.9±0.3	-20.9±0.3	-20.9±0.3	-0.8±0.9
	Total	-12.0±0.0	-12.7±0.0	37.8±0.8	-46.1±0.6	73.9±1.3	-56.9±0.6	-24.6±0.9	16.1±1.6

[a] For TI this is G_{vdW}^{bound}, for the other methods it is the difference $\langle G_{np}(P)-G_{np}(PL)\rangle_{PL}$ (Eqn. 11).

[b] For TI this is G_{vdW}^{free}, for the other methods it is $-\langle G_{np}(L)\rangle_{PL}$ (Eqn. 10).

[c] Here, we assume that the cavity in the free-protein calculations is filled with a dummy ligand that does not interact with the surroundings.

[d] Here, we take the cavity term from the P0 approach and the dispersion and repulsion terms from the P approach.

Table 4. PCM components of the non-polar energy in kJ/mol, i.e. cavitation, dispersion, and repulsion energy. ΔG_{np}^{PCM} is the sum of these three terms. The components are given for the complex (PL), the free protein (L), and the free ligand (L), as well as for the free protein in the cavity of the complex (P0) and various differences of the terms.

		PL	P	L	P–PL	PL–P–L	P0[a]	P0–PL	PL–P0–L
Ferritin	ΔG_{cavity}	19616	19562	58	-54	-4	19616	0	-58
	ΔE_{disp}	-5890	-5919	-68	-29	97	-5888	2	66
	ΔE_{rep}	1429	1435	14	6	-20	1428	0	-14
	ΔG_{np}^{PCM}	15155	15078	4	-77	73	15156	1	-6
Avidin	ΔG_{cavity}	27659	27620	46	-39	-7	27659	0	-46
	ΔE_{disp}	-7913	-7926	-75	-14	89	-7911	1	74
	ΔE_{rep}	2083	2085	26	3	-29	2083	0	-26
	ΔG_{np}^{PCM}	21830	21779	-2	-50	53	21831	1	1
Trypsin	ΔG_{cavity}	11655	11596	71	-59	-12	11655	0	-71
	ΔE_{disp}	-3683	-3713	-88	-29	118	-3674	9	79
	ΔE_{rep}	982	995	20	13	-33	981	-1	-19
	ΔG_{np}^{PCM}	8954	8879	3	-76	73	8962	8	-10
Galectin-3	ΔG_{cavity}	8063	7964	108	-98	-9	8063	0	-108
	ΔE_{disp}	-2830	-2826	-124	4	120	-2785	45	79
	ΔE_{rep}	690	690	37	0	-36	681	-9	-28
	ΔG_{np}^{PCM}	5923	5828	21	-95	74	5959	36	-57

[a] P0 is the free protein in the cavity of the complex.

Table 5. Average number of water molecules in the binding site (#Wat) and the van der Waals interaction energy between these water molecules and the protein ($\langle E_{\mathrm{vdW}}^{PS_L} \rangle$; kJ/mol) in the TI simulations with ($\lambda = 0.05$) and without ($\lambda = 0.95$) the ligand bound for the various proteins, as well as the difference and the PCM estimate of $\langle E_{\mathrm{vdW}}^{PS_L} \rangle$.

	Bound ligand		No ligand		Difference		PCM[a]
	#Wat	$\Delta E_{\mathrm{vdW}}^{PS_L}$	#Wat	$\Delta E_{\mathrm{vdW}}^{PS_L}$	#Wat	$\Delta E_{\mathrm{vdW}}^{PS_L}$	$\Delta E_{\mathrm{vdW}}^{PS_L}$
T4-lysozyme[24]	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0	0.0	-8.0±0.4
Ferritin	2.1±0.2	-5.3±2.9	5.7±0.7	-28.2±7.7	3.6	-22.9	-23.8±0.7
Avidin	0.4±0.1	-2.6±1.0	5.6±0.8	-26.1±5.0	5.2	-23.4	-12.2±0.6
Trypsin	0.5±0.1	-1.7±0.7	6.8±0.2	-29.3±1.2	6.3	-27.6	-24.4±0.9
Galectin-3	1.8±0.1	8.6±1.2	9.7±0.1	-7.9±2.0	7.9	-16.5	-32.3±0.5

[a] The PCM estimate of E_{vdW} is the P–P0 difference for the sum of the ΔE_{disp} and ΔE_{rep} energies in Table 4, i.e. the continuum PS_L term.

Table 6. ξ value for the linear combination of P and P0 estimates (Eqn. 13) from the PCM and CD methods fitted to the TI results or to the solvent exposure (SE).[a]

	ξ fitted to TI		SE	ξ from SE		Deviation from TI[b]	
	PCM	CD		PCM	CD	PCM	CD
T4-lysozyme	-0.02	0.00	0.00	0.03	0.10	2.6	3.2
Ferritin	0.00	0.09	0.06	0.09	0.22	7.3	5.9
Avidin	0.07	0.17	0.04	0.08	0.18	0.3	0.3
Trypsin	0.32	0.59	0.10	0.14	0.30	-14.9	-22.8
Galectin-3	0.40	0.77	0.38	0.44	0.83	4.6	4.9

[a] A **ξ** value of 0 indicates that the continuum method is optimal when using the P0 approach, whereas a value of 1 indicates that the method is optimal when using the P approach (Eqn. 13). SE is the ratio between the SASA of the ligand when it is bound to the protein and when it is free in solution.

[b] This is the deviation of G_{np}^{bound} from the TI results (kJ/mol), when it is calculated from Eqn. 13, using ξ calculated from SE. A positive value indicates that the continuum estimate is too negative.

Table 7. Difference between ESI-PCM and TI results for G_{np}^{bound} in kJ/mol, obtained by fits to SASA, SE or ΔN(wat).

	SASA[a]	SE[b]	ΔN(wat)[c]
T4-lysozyme	-14.7	-16.3	-17.6
Ferritin	11.9	13.2	17.3
Avidin	1.9	7.9	20.2
Trypsin	3.2	-2.8	5.8
Galectin-3	-2.3	-2.1	-25.8

A positive value indicates that the continuum estimate is too negative.

[a] ESI-PCM was parametrised using Eqn. 15, i.e. through a linear relation to the SASA of the ligand, shown in Figure 4.

[b] Parametrised using the SE of the ligand

[c] Parametrised using the difference in the number of water molecules between the bound and unbound protein.

Table 8. Mean absolute deviation of the ΔG_{np} estimates of the continuum methods from the TI results in kJ/mol, including the data also for T4-lysozyme.[24]

Term	SASA		CD		PCM			SE		ESI-
	P	P0	P	P0	P	P0	P/P0	CD	PCM	PCM
Bound	8.6	8.3	28.0	20.9	61.5	16.9	13.3	5.9	7.4	6.7
Free	8.9	8.9	2.6	2.6	5.2	5.2	5.2	2.6	5.2	5.2
Total	16.6	15.9	28.9	20.0	56.9	21.5	10.7	5.3	8.9	10.1

Figure 1. The studied protein–ligand complexes. The solvent-accessible surface of the protein is shown in green and the ligands are shown as space-filled models. a) ferritin–phenol (the channels to the binding sites are indicated with arrows), b) avidin–Btn7, c) trypsin–ABI, and d) galectin-3–L19.
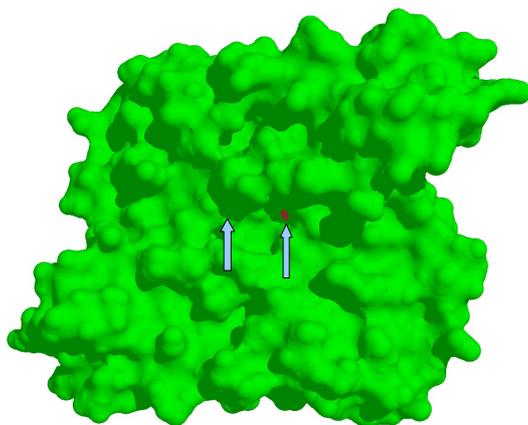

a)
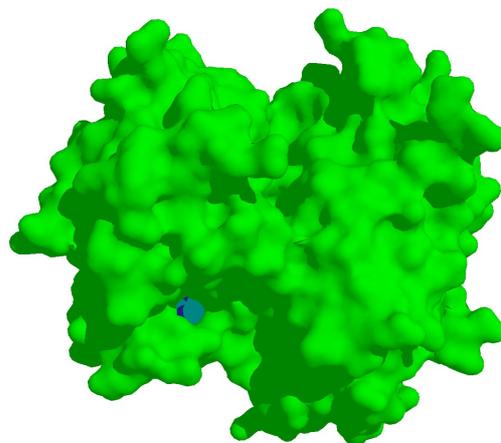
b)
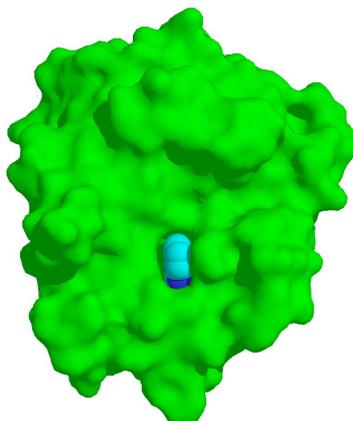
c)
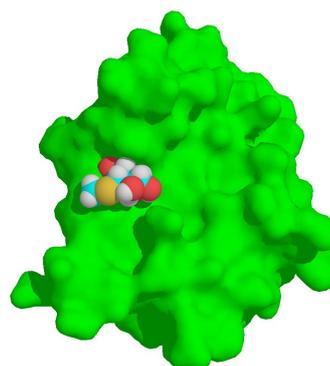
d)

Figure 2. Time evolution of the number of water molecules in the binding site of avidin, averaged over the four subunits. Each simulation was started with a different number of water molecules in the binding site. The 100-ps equilibration period is excluded.
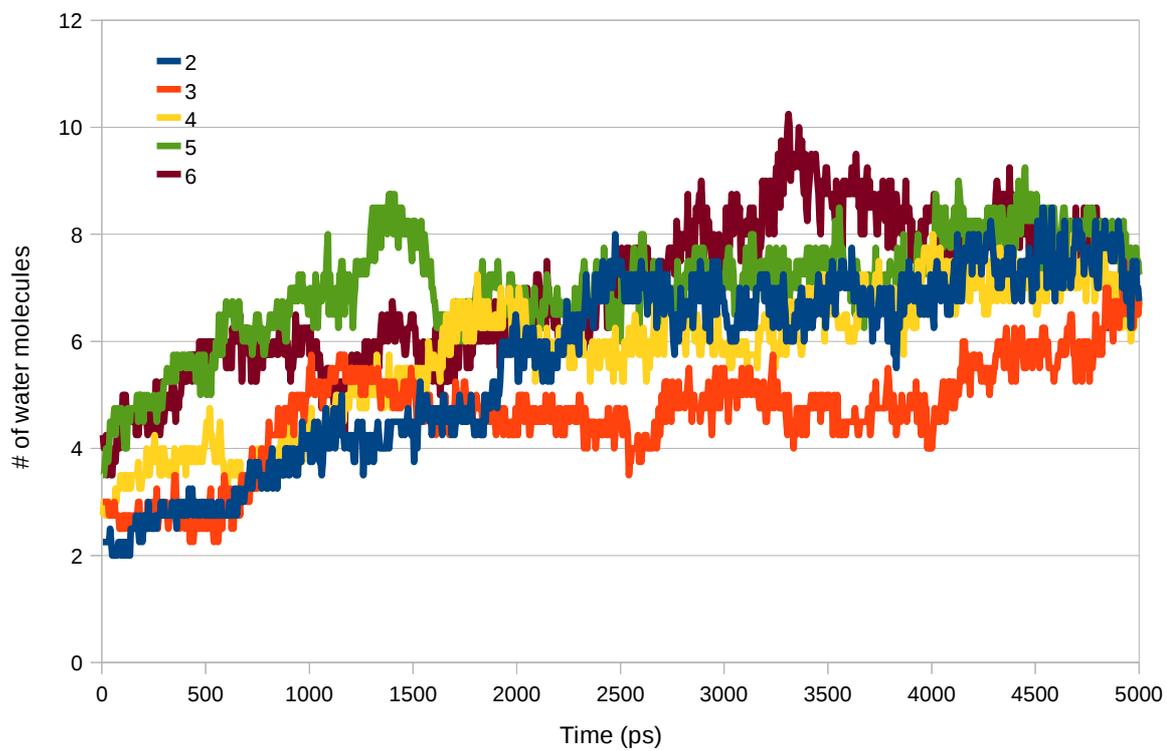
Figure 3. Correlation between optimal mixing parameter ξ fitted to TI (Eqn. 13) and the solvent exposure. Estimates for PCM are shown with squares and a dashed line, whereas the estimates for CD are shown as triangles and a solid line.
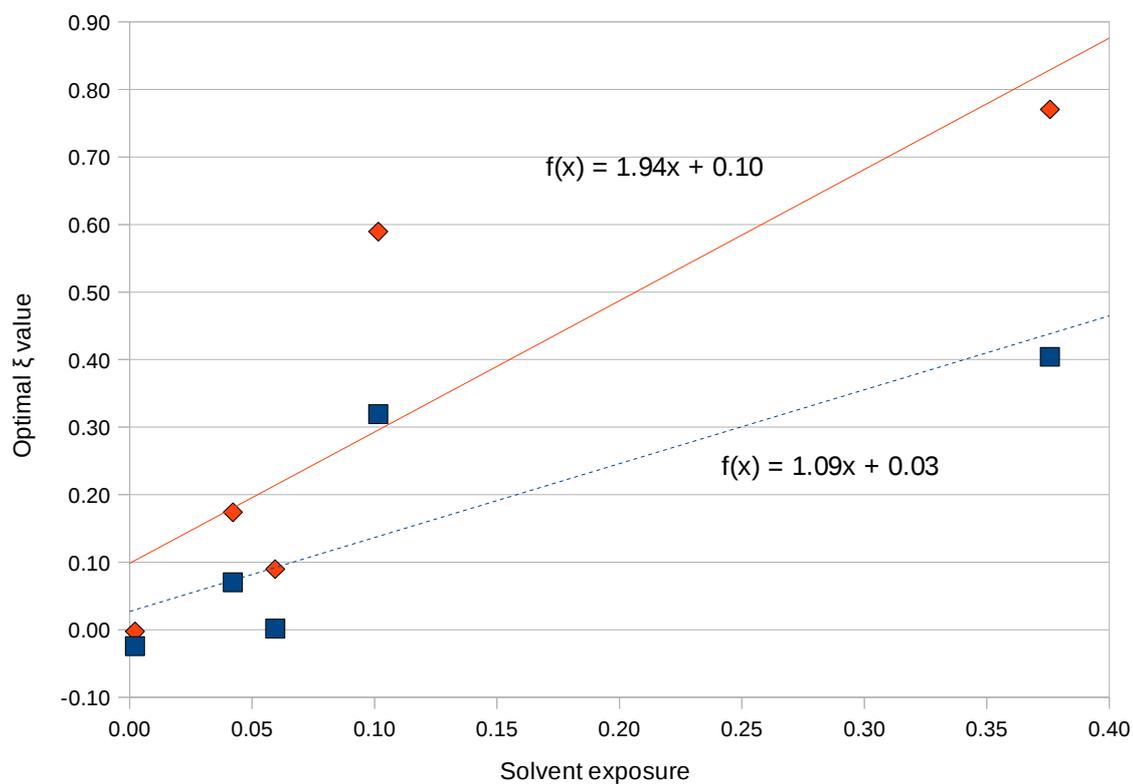
Figure 4. The fit of $\Delta G_{np}^{S_L}(\text{TI})$ against SASA (Eqn. 15) for the ESI-PCM method. The difference between the squares and the line also represents the error of the ESI-PCM estimate of ΔG_{np}^{bound} for each ligand.


f(x) = -0.32x + 84.11
R² = 0.83
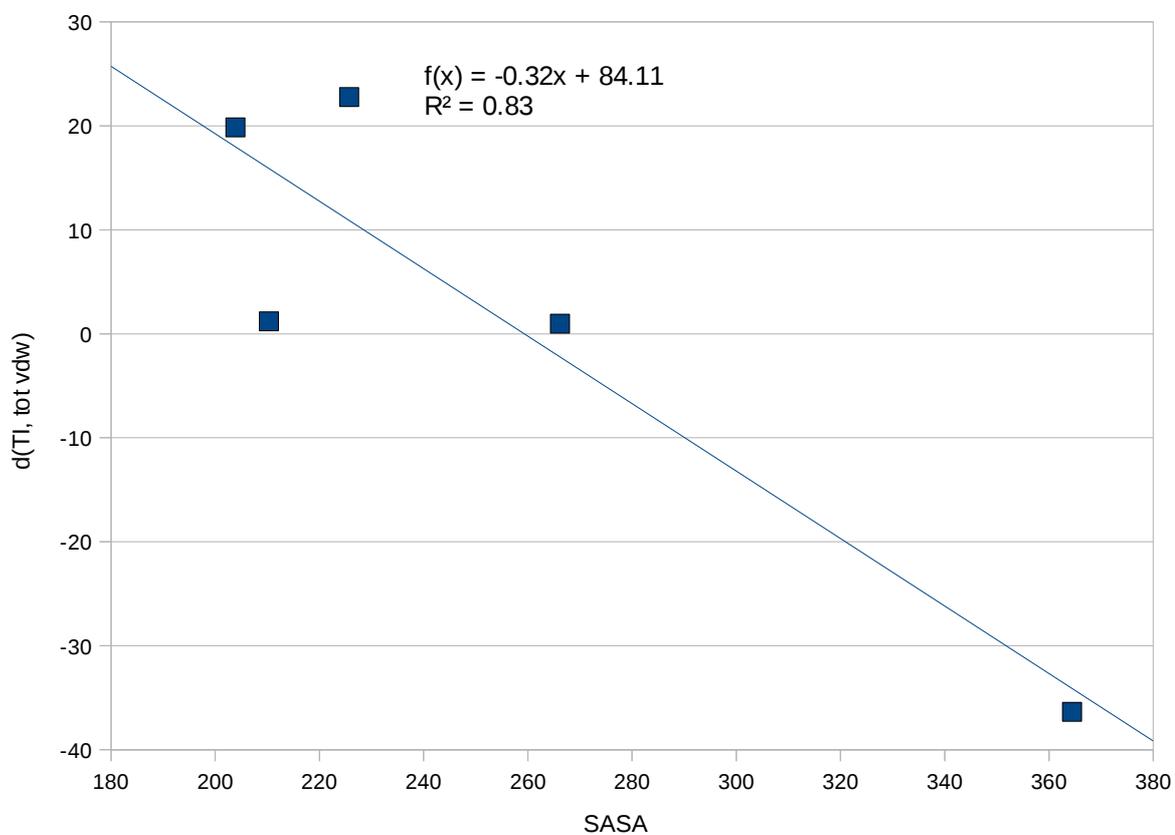
Figure 5. Then non-polar contribution to the binding affinity of seven biotin analogues (Btn1–Btn7 shown at the bottom of the figure) to avidin, calculated with three different continuum methods and the P and P0 approaches.
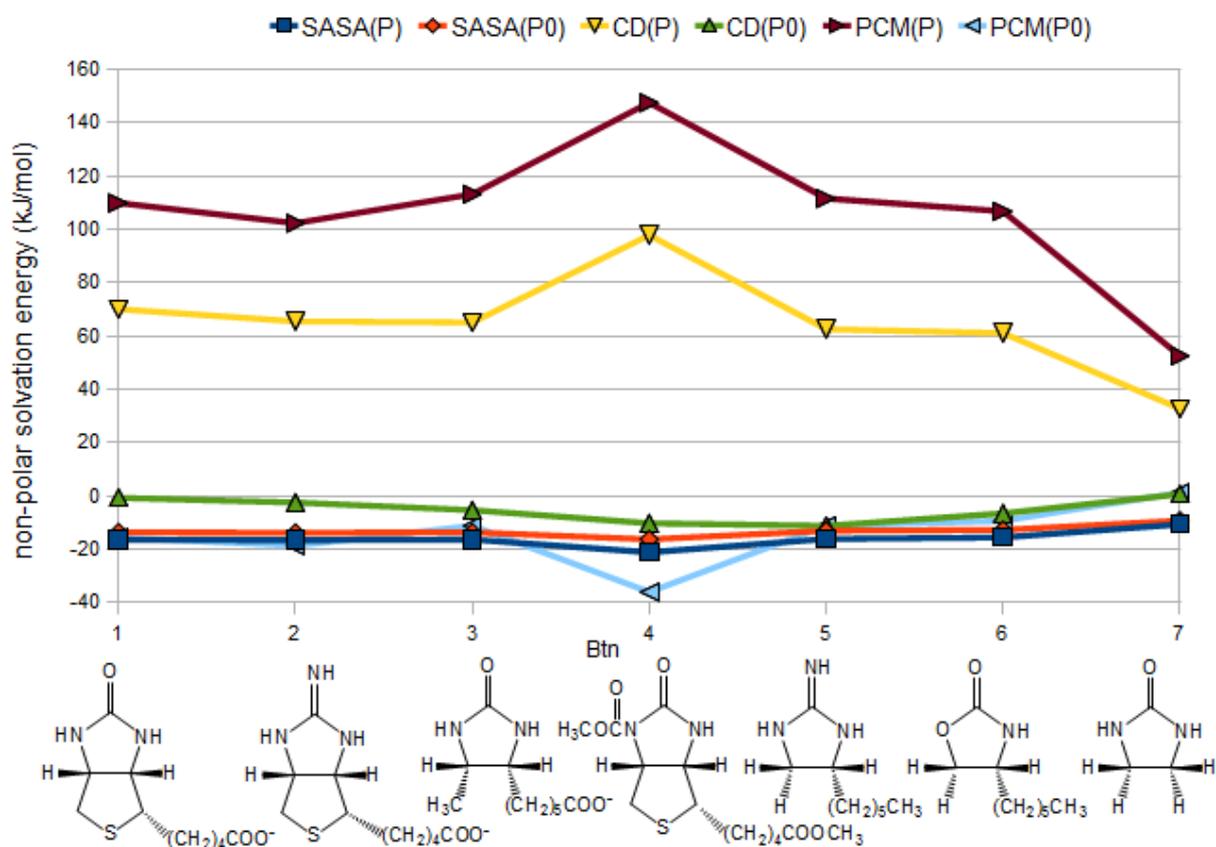
Table of content graphics

TI estimates of ΔG_{np}^{bound} **on a relative scale.** The scale goes from 0 to 1, and at 0 the TI results is equal to continuum estimates that employs the P0 cavity and at 1 the TI results is equal to default continuum estimates (P). Results are shown for both PCM and CD.