JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

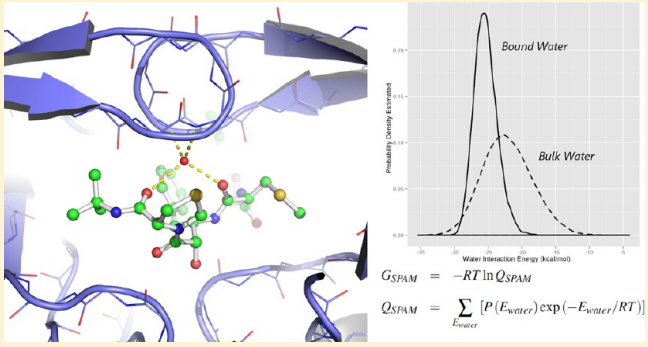
SPAM: A Simple Approach for Profiling Bound Water Molecules

Guanglei Cui,*[,†] Jason M. Swails,[‡] and Eric S. Manas[†]

[†]Computational Chemistry US, Platform Technology and Sciences, GlaxoSmithKline Pharmaceuticals, 1250 South Collegeville Road, Collegeville, Pennsylvania 19426, United States

[‡]Quantum Theory Project, University of Florida, Gainesville, Florida 32611, United States

ABSTRACT: A method that identifies the hydration shell structure of proteins and estimates the relative free energies of water molecules within that hydration shell is described. The method, which we call "SPAM" (*maps* spelled in reverse), utilizes explicit solvent molecular dynamics (MD) simulations to capture discrete hydration sites at the water−protein interface and computes a local free energy measure from the distribution of interaction energies between water and the environment at a specific site. SPAM is able to provide a qualitative estimate of the thermodynamic profile of bound water molecules that correlates nicely with well-studied structure−activity relationships and observed binding "hot spots". This is demonstrated in retrospective analyses of HIV1 protease and hen egg white lysozyme, where the effects of water displacement and solvent binding have been studied extensively. The simplicity and effectiveness of SPAM allow for prospective application during the drug discovery process.



INTRODUCTION

The solvation shell is an integral part of protein structure in aqueous solution. Water molecules in a protein solvation shell are thought to display unique structural and thermodynamic properties relative to the bulk solvent and play an important role in defining protein structure as well as in modulating the function and the dynamics of proteins.[1] When a protein forms a complex with another molecule (e.g., enzyme−substrate complexation, antibody−antigen recognition, receptor−ligand binding, etc.), the existing solvation shell structure may undergo significant changes, which result in unique thermodynamic or kinetic signatures that can be captured experimentally.[2−4]

Phenomenologically, the notion of hot spots in protein−ligand binding has been around for more than a decade.[5,6] The concept typically refers to discrete recognition elements that bear a large contribution to the overall process of binding and are known for unusually high binding efficiencies. First noted in the seminal paper by Kuntz et al.,[7] the efficiency of smaller ligands (or fragments) has also been evident in several more recent studies,[8,9] where the ligand efficiency (defined as binding free energy per non-hydrogen atom) can reach as high as 1.5 kcal mol^{-1} per heavy atom in extreme cases.[7] Understanding the nature of hot spots and being able to predict their locations would provide a valuable asset to structure-based drug discovery research, particularly given the current focus on improving ligand efficiency and avoiding the addition of unnecessary lipophilic moieties to compound designs, the ultimate goal being to improve the odds of success in the clinic.[10−12] One hypothesis is that certain hot spots may be determined not just by the favorable change in enthalpy associated with receptor−ligand interactions but also with favorable changes in both enthalpy and entropy associated with water displaced from the receptor−ligand interface to the bulk. Indeed, the hydrophobic effect itself has long been explained in terms of changes in the entropy and heat capacity of water,[13−16] i.e., the origin of the effect that is not due to enthalpy alone.

Solvent binding at the water−protein interface is a topic that has been extensively investigated by both experimental and computational techniques. Among the experimental techniques, crystallography is arguably the most definitive method of choice. The competitive binding between water and small organic solvent molecules may be directly observed with multisolvent protein crystallography,[17,18] which may shed light on the specific recognition elements at the water−protein interface. However, resolving water binding at the interface alone is not particularly informative from a thermodynamic or kinetic perspective, and the popular design strategy of displacing bound water molecules can potentially generate more misses than hits when applied with little discrimination.[19] Computational analysis of water or solvent binding can often be complementary to experimental approaches and provide insightful thermodynamic details of bound water molecules.

Computational solvent mapping techniques first appeared in 1985 with Goodford's GRID, which is capable of mapping energy minima of solvent molecules as well as numerous other probes.[20] Methods like GRID have been quite successful in identifying water and other solvent binding sites, but the significance of these sites is difficult to gauge since the degree to

Received: August 9, 2013
Published: October 28, 2013

which a favorable interaction energy is compensated by an entropic penalty is not directly assessed, although one can of course complement a water probe-based analysis in GRID with "dry" probes to account for the hydrophobic effect, albeit in a very approximate manner. While the energy functions in such methods have become increasingly sophisticated over time, the primary focus on the enthalpic component of binding has remained largely unchanged until fairly recently. One prominent example of these more recent methods is Watermap,[21] which is based on inhomogeneous fluid solvation theory (IFST)[22] and attempts to extract the excess entropy of bound water due to the translational and rotational restriction exerted by the local environment. Watermap has shown potential in pinpointing key structure−activity relationships (SAR) in lead optimization processes[3,23−27] and has helped to rekindle interest in the role of water in drug discovery research.[28] Additional statistical mechanics-based approaches have also been developed, such as the direct descendents of IFST—Grid Inhomogeneous Solvation Theory (GIST)[29] and Solvation Thermodynamics of Ordered Water (STOW),[30] 3D-RISM,[31] for which a graphical interface has recently been implemented by Chemical Computing Group (CCG),[32] SZMAP from OpenEye Scientific Software, which is based on a semi-continuum solvation model,[33,34] and a recent method combining IFST and 3D-RISM.[35] Besides the above-mentioned techniques, the binding affinity of specific water molecules near protein surfaces has also been estimated by double decoupling free energy calculations,[36] which have been used to rationalize potency changes in water-targeting ligand designs.[19,37−39] The computational cost and complexity of double decoupling-based techniques are well recognized and can make routine application in an industrial setting less practical.

Partly inspired by recent enthusiasm for characterizing protein hydration, a statistical-mechanics-based approach ("SPAM," maps spelled in reverse) that estimates the free energy difference between protein-bound and bulk water has been developed and evaluated both retrospectively and prospectively at GlaxoSmithKline (GSK). Through the evaluation, we found that SPAM is able to provide a qualitative estimate of the thermodynamic profile of water in hydration sites that is reasonably consistent with structure−activity relationships and known hot spots in protein binding sites. In the following sections, the theoretical basis of SPAM is presented in detail, along with two retrospective analyses on well studied protein targets—HIV1 protease and hen egg white lysozyme.

METHODS

SPAM Methodology. In order to explain the rationale behind the SPAM methodology, it is instructive to first consider IFST-based approaches, which estimate the local thermodynamic properties of solvation, i.e., the average interaction energy contribution to water binding as well as the entropic penalty due to increased ordering relative to the bulk. In IFST, this entropic contribution is calculated in an expansion by summing over the solute−solvent and solvent−solvent distribution functions of microscopic states described by solvent translational and rotational degrees of freedom. The integration of solute−solvent distribution functions is typically simplified by limiting the sampling of solute configurations, for example, by applying a restraining potential. However, it is still difficult to achieve sufficient sampling to evaluate the solvent−solvent contribution in the entropic expansion (i.e., the pairwise water−water correlation) with a relatively short molecular dynamics (MD) simulation, which has led to implementations that either include only a partial treatment of the two-body term or omit it entirely.[21,40] These previous studies have shown that such an omission is still able to produce interpretable thermodynamic quantities that are consistent with experimental findings.

Keeping in mind that meaningful information is embedded in the solute−solvent distribution function, SPAM similarly chooses to neglect higher order correlation between water molecules but takes a different approach to extract the entropic contribution from an MD simulation. Instead of describing the microscopic states of water in terms of the translational and rotational degrees of freedom and extracting the entropy of solvation directly from this distribution, SPAM considers the distribution of interaction energies (E_{water}) between water and the surrounding environment (solute and other solvent molecules) at a given site, and the perturbation to this distribution upon binding (depicted graphically in the following section on HIV1 protease). By neglecting water−water correlation, each solvation site in SPAM is conceptually equivalent to a single water molecule with a distribution function specific to that site but independent of others.[41] The free energy of such a water can be evaluated via the site partition function Q_{SPAM} defined in the following equations (eqs 1 and 2):

$$G_{SPAM} = -RT \ln Q_{SPAM} \tag{1}$$

$$Q_{SPAM} = \sum_{E_{water}} [P(E_{water}) \exp(-E_{water}/RT)] \tag{2}$$

where $P(E_{water})$ is the probability of a water within a site having an interaction energy with its surroundings equal to E_{water}. The treatment of a water site independent of others is analogous to the one-electron approximation in Hartree−Fock theory, in that we are considering the energies of each water site in the "average" field due to the surrounding protein and other water molecules. [The functional form of G_{SPAM} is reminiscent of the Widom insertion method,[42] except with SPAM, insertion would be into a preformed cavity.] The use of univariate energy distributions simplifies the calculation of the relative free energy difference between bulk and the bound water (ΔG_{SPAM}), as it avoids the clustering procedure and the need to estimate the translational and orientational probability distribution of water in IFST-based methods. However, this approach still allows the simultaneous evaluation of multiple water molecules using a single trajectory, a significant advantage over the double decoupling free energy methods. The energy and entropy components of ΔG_{SPAM} are then calculated using eq 3:

$$\begin{aligned} \Delta G_{SPAM} &\equiv G_{SPAM,bound} - G_{SPAM,bulk} \\ \Delta E_{SPAM} &\simeq \langle E_{water,bound} \rangle - \langle E_{water,bulk} \rangle \\ -T\Delta S_{SPAM} &\equiv \Delta G_{SPAM} - \Delta E_{SPAM} \end{aligned} \tag{3}$$

where $\langle E_{water} \rangle$ stands for the ensemble average of interaction energies of water sampled during the MD simulation.

The implementation of SPAM is straightforward. A general workflow consists of computing the location of hydration sites around the target protein from a multinanosecond explicit solvent MD simulation, extracting the interaction energy of water molecules that travel through hydration sites of interest and estimating the SPAM free energy of water according to eq

2. To reduce the difficulty in convergence, the protein conformation is restrained by a harmonic potential on the non-hydrogen atoms of the protein, while solvent molecules are allowed to move freely. The hydration sites are located by identifying the peaks in the computed water density map that are above a certain threshold (e.g., 0.05 water molecule per Å^3 or roughly 1.5 times of the water bulk density at room temperature). The volume of the hydration sites over which the summation in eq 2 is carried out should ideally be decided by the spatial span of the density peaks. However, to simplify the implementation, a cubic volume with an edge length of 2.5 Å centered around the hydration sites is used. An alternative strategy could, for example, use a spherical integration volume but would require more expensive distance calculations during the site assignment step. Any water molecules that travel to the volume of interest during the MD simulation need to be tracked so that their interaction energies with the surroundings can be calculated from the snapshots saved during the simulation. In the current implementation, ΔG_{SPAM} is calculated as if the volume is fully occupied by a single water molecule during the entire simulation. Snapshots in which no water molecule or multiple water molecules are present in the site are omitted from the analysis for that site. The latter situation where more than one water molecule occupies the same integration volume can occur, but this typically only represents a small proportion of all the sampled configurations (less than 1%) for high density water sites (greater than 0.10 Å^{-3}), which can be reduced by using a smaller integration volume.

Although rare, the local solvent entropic term ($-T\Delta S_{SPAM}$) can be favorable compared to the bulk, for example, in cases where a water is trapped in a homogeneous hydrophobic environment where no orientational preference is imposed (data not shown). This is in contrast to the entropy calculated by IFST-based methods, which assume a uniform translational and orientational distribution for water in the bulk. The notion of water gaining entropy inside a hydrophobic enclosure relative to the bulk may appear to conflict with the familiar view of water becoming more structured in a hydrophobic environment, but we point out that this phenomenon has been demonstrated by computer simulations of nonbiological systems, such as the filling of the hydrophobic channel of a carbon nanotube.[43]

System Preparation. The same protocol was followed to prepare both targets selected for this study, HIV1 protease (PDB ID: 1hpx)[44] and hen egg white lysozyme (PDB ID: 2lyo).[45] For the unliganded calculations, all existing ligands and crystallographic solvent molecules, including water and cosolvent used in the crystallization buffer, were first removed. The structures were then submitted to the Protein Preparation Wizard work flow provided in Maestro 9.2,[46] where protonation states of the proteins were assigned, and the hydrogen bond network was optimized. For HIV1 protease, one of the Asp25 residues was neutralized to be consistent with experimental observations.[47] To investigate the impact of ligand binding on the thermodynamic profile of bound water molecules, specifically Wat301 (known to be displaceable),[48] the simulation was also carried out for the HIV1 protease-inhibitor complex (KNI-272). All six crystallographic water molecules trapped in the KNI-272 binding site were included in the complex simulation.

Simulation and Analysis. Once prepared, the solute was solvated by an orthorhombic box of SPCE[49] water molecules with a margin of at least 12 Å on each side. No sophisticated measure was taken to place water molecules around the solute, although SPAM could potentially benefit from procedures such as Grand Canonical Monte Carlo[50,51] or 3D-RISM calculations,[52] particularly for water sites that are excluded from solvent access when only limited solute flexibility is allowed. Counterions were introduced as needed to neutralize the net charge of the system. AMBER force fields *parm99SB*[53] and *gaff*[54] were used to represent the proteins and KNI-272, respectively. AM1-BCC partial charges[55] were calculated for the bound conformation of KNI-272 using antechamber.[56] MD simulations with periodic boundary conditions and harmonic positional restraints on non-hydrogen atoms (force constant 1.0 kcal mol^{-1} Å^{-2}) were carried out with NAMD 2.8.[57] The long-range electrostatic interactions were handled by particle-mesh-Ewald (PME)[58] with a maximum grid spacing of 1.0 Å. A 10.0 Å cutoff was applied to short-range nonbond interactions (van der Waals and direct space electrostatic interaction) along with a switching function, which was turned on at 8.0 Å to avoid discontinuity in force evaluations. A common equilibration procedure was employed to bring the system to the desired temperature (300 K) and pressure (1 atm), which involved a 5000-step minimization and 30 ps constant temperature and volume (NVT) dynamics, followed by 200 ps constant temperature and pressure (NPT) dynamics. The production run from which the hydration shell configurations were sampled was then carried out for 10 ns with a time step of 2 fs, from which 5000 snapshots with a 2 ps interval were stored for the subsequent SPAM analysis. We found that such a simulation procedure was able to identify water binding sites, and up to 70−80% of all crystallographically resolved water molecules in structures with 2.0 Å resolution or better may be reproduced with sufficient accuracy (within 1.0 Å). The impact of omitting crystal contacts and using a restrained protein conformation on the accessibility of water sites is tolerably small and can be mitigated by including experimental water sites, or by applying complementary water placement techniques as mentioned above.

To prepare for the SPAM analysis, the unweighted volumeric water density map was calculated first as the time average on a grid (with a grid spacing of 0.5 Å) over the simulation snapshots by using the VolMap plugin of VMD 1.8.7.[59] After applying a density filter of 0.05 Å^{-3} to eliminate background peaks due to bulk density fluctuations, specific hydration sites corresponding to peaks in the calculated density were identified using mathematical morphology techniques common in multidimensional image processing.[60−62] For each hydration site of interest, the interaction of the water molecules within that site and the surrounding environment were calculated using the *pairInteraction* keyword in NAMD with a large nonbonded cutoff value (12.0 Å). The SPAM free energy for the hydration site was then calculated according to eq 2. The G_{SPAM} value of a bulk SPCE water molecule is found to be approximately −30.3 kcal mol^{-1}, and the average energy of water ($\langle E_{water} \rangle$) is −22.2 kcal mol^{-1}, which comes to a reasonable agreement with the heat of vaporization of water at room temperature (−10.52 kcal mol^{-1} at 298 K) multiplied by a factor of 2 to correct for double counting. The relative SPAM free energy of a solvation site relative to the bulk water (ΔG_{SPAM} along with the energy and entropy components) can then be calculated and used to assess the environmental effects due to solvation.

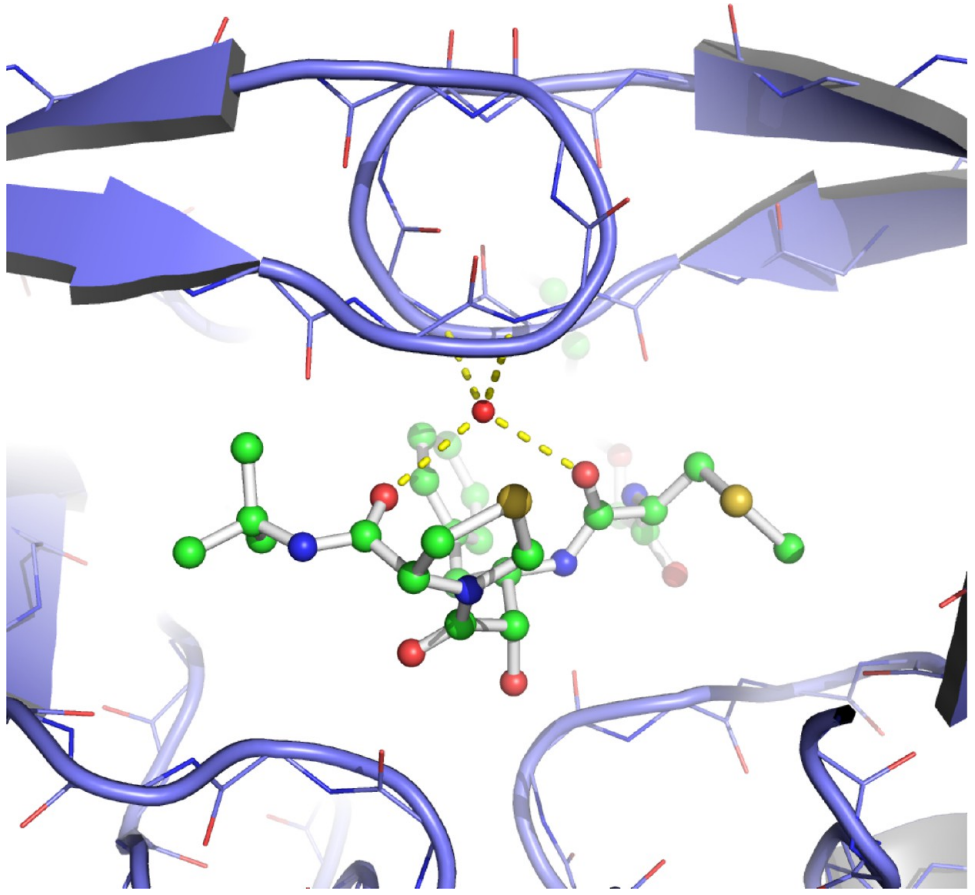


Figure 1. The hydrogen bond interactions that Wat301 (the center red sphere) of HIV1 protease (PDB ID: 1hpx) forms with the protein (shown in slate blue) and the HIV1 protease inhibitor KNI-272 (shown as the ball and stick model). Only the backbone atoms of the protein were shown for better clarity.

To facilitate the application of SPAM analysis, all the steps prior to the water interaction energy calculation were implemented in the *cpptraj* module of the freely available AmberTools 12 suite. A Python frontend driver, SPAM.py, was also written and can be used to automate the entire SPAM analysis with minimal user intervention.

RESULTS AND DISCUSSION

Our goal has been to investigate the ability of SPAM to identify hot spots in the context of protein−ligand binding by studying a number of GSK internal drug discovery targets as well as several well reported proteins in the public domain. However, it is difficult to validate energetic predictions made by SPAM or any other water profiling method as we do not have the means to measure the thermodynamic parameters of a single water site. Consequently, validation of the thermodynamic parameters generated by these methods tends to rely on indirect assessments, primarily consisting of agreement with structure−activity relationships, especially where significant potency changes upon water bridging and/or displacement are observed. We also remind the reader that while agreement with experimentally observed waters of crystallization is an important first step in validating computed water site locations, there is no reason to expect that such occupancies should correlate with local thermodynamic measures such as ΔG_{SPAM}, in contrast to system properties such as ΔG° and the chemical potential, which determine the spontaneity of processes such as binding under specific conditions.

The outcome of two retrospective studies on public systems is presented below to demonstrate the ability of SPAM to identify hot spots, which in principle could allude to the possibility of displacing or bridging to one or more waters with the goal being to introduce a net favorable contribution to the binding free energy of a ligand to a protein. The overall assessment of the method will also be discussed in the Conclusions.

HIV1 Protease. We initially focus our attention on HIV1 protease, which is one of the best known examples where a structural water molecule (Wat301) was successfully targeted for displacement with the aid of computer modeling.[48] Specifically, a cyclic urea scaffold was designed to displace Wat301 and replace its interactions with the amide NH groups of Ile50 from the protein dimer (Figure 1). In addition, literature SAR provides an interesting comparison with substrate-derived peptidic inhibitors such as KNI-272, which form bridging hydrogen bonds with Wat301 via two amide carbonyl groups as opposed to displacing the water. The energetics of Wat301 have also been considered in a number of different theoretical studies,[35,37−39,63] which will be discussed in more detail later in this section.

SPAM analysis was carried out on Wat301, with and without the bound ligand (Table 1). The entropic penalty relative to bulk due to interactions imposed by the protein environment is evident in the calculated probability density function of water interaction energies (Figure 2). The average interaction energy of Wat301 in the presence of KNI-272 is −25.4 kcal mol^{-1},

dx.doi.org/10.1021/ct400711g | J. Chem. Theory Comput. 2013, 9, 5539−5549

Table 1. The Thermodynamic Profiles of Wat301 (in kcal mol^{-1}) of HIV1 Protease Calculated in the Presence and Absence of KNI-272[a]

Wat301	ΔG_{SPAM}	σ	ΔE_{SPAM}	$-T\Delta S_{SPAM}$
KNI-272	3.0	0.043	−3.2	6.2
no KNI-272	1.8	0.21	−0.64	2.4

[a]The standard deviation (σ) of the computed ΔG_{SPAM} was estimated with a bootstrapping procedure (N = 20) with a subsample size half of the original.

which is −3.2 kcal mol^{-1} more favorable compared to the bulk. However, the overall distribution is substantially compressed due to the strong orientational restraint imposed by a total of four hydrogen bonds formed with the protein and the inhibitor, which incurs an entropic cost of 6.2 kcal mol^{-1} ($-T\Delta S_{SPAM}$). The orientational effect due to hydrogen bonding interactions is also noticeable in the absence of KNI-272. Without the inhibitor, the interaction energy of Wat301 is −0.6 kcal mol^{-1} relative to the bulk, and the distribution of water interaction energies still indicates a 2.4 kcal mol^{-1} entropic cost, consistent with the observation that the environmental restraint is less severe. The SPAM free energy change of Wat301 when KNI-272 is bound relative to the bulk is 3.0 kcal mol^{-1}, which is 1.2 kcal mol^{-1} less favorable than when KNI-272 is absent, suggesting that this bridging water has a net unfavorable contribution to the solvation of the complex relative to the unliganded protein.

In order to understand how one might use such results in the context of SAR exploration, it is instructive to consider the binding affinities (K_i) of C2 symmetric cyclic urea and peptidomimetic inhibitors with very similar size and peripheral groups, such that a very good spacial overlay and virtually identical interactions are obtained apart from the displacement of Wat301 by the cyclic urea compound (Figure 3). For example, the cyclic urea inhibitor CHEMBL284136 (K_i = 0.12 nM, MW = 538), which displaces Wat301, is a roughly 100-fold more potent binder relative to peptidomimetic inhibitor CHEMBL289188 (K_i = 22 nM, MW = 541), which forms a bridging interaction with Wat301. Admittedly, factors other than the displacement of Wat301 may contribute to the observed difference in binding affinities (~2.7 kcal mol^{-1}), such as solvation, internal strain, and conformational entropy of the ligands, as well as subtle differences in the interactions with the pocket. However, it is clear that the displacement of Wat301 is (1) achievable and (2) unlikely to represent a significant penalty, as such a penalty would have to be offset by highly favorable contributions to the protein−ligand binding free energy in order to produce the observed net 100-fold

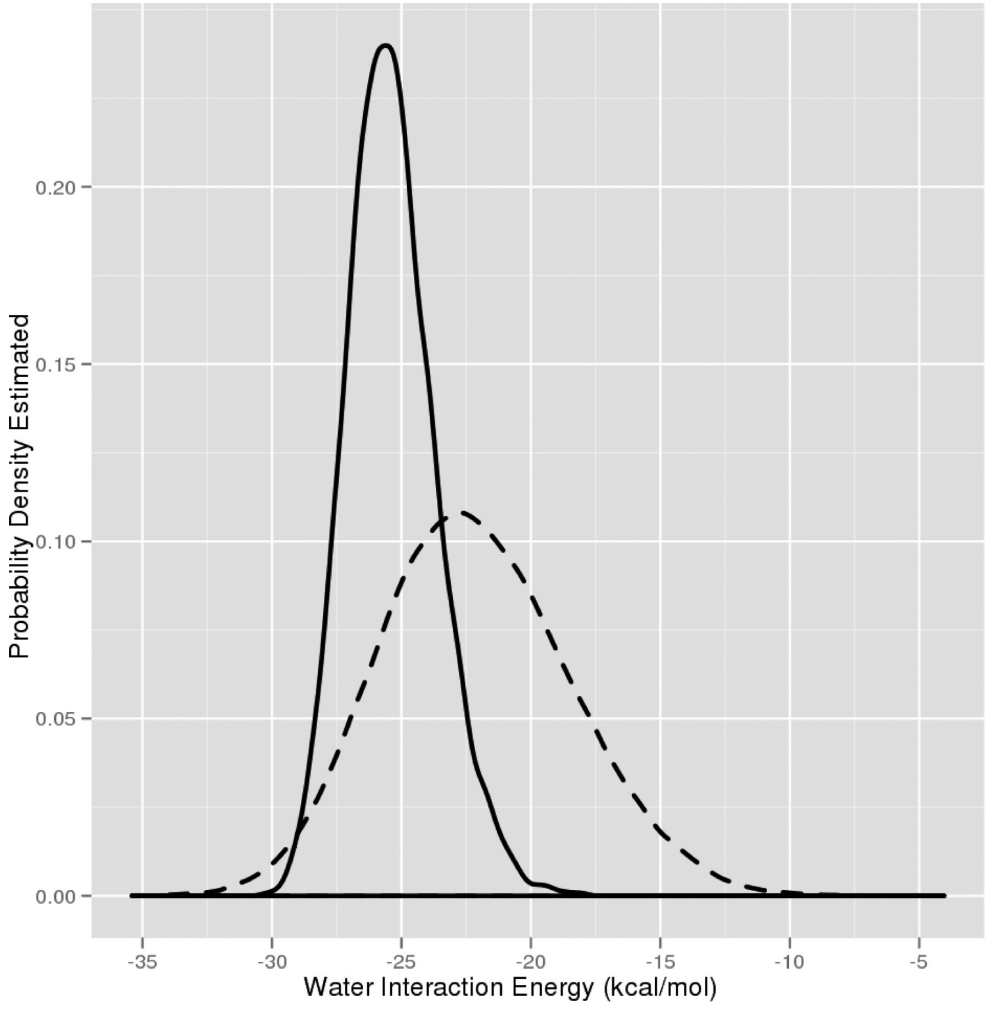


Figure 2. The probability distribution of water interaction energies estimated for Wat301 in the complex of HIV1 protease and KNI-272 (solid line), and a SPCE water in the bulk (dashed line).

dx.doi.org/10.1021/ct400711g | J. Chem. Theory Comput. 2013, 9, 5539−5549

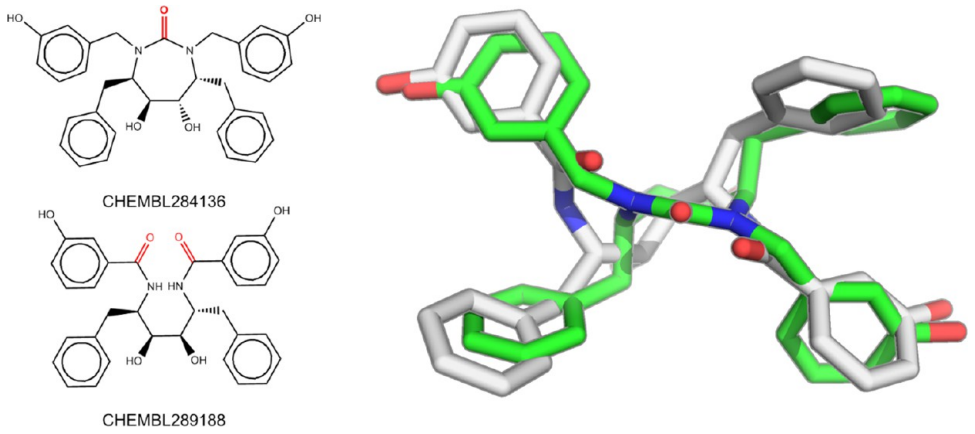


Figure 3. The chemical structures of CHEMBL284136 and CHEMBL289188 and the three-dimensional overlay of their predicted bound poses based on the existing cocrystal structures of HIV1 protease, PDB ID 1dmp and 1hvl.

improvement in potency. In the end, the calculated SPAM free energy profiles of Wat301 succeeded in capturing a favorable energetic contribution from hydrogen bonding to the protein and ligand, as well as an unfavorable entropic penalty from that same hydrogen bonding interaction. That these contributions result in a slightly net unfavorable free energy relative to the bulk for Wat301 in the presence of KNI-272 would have only supported the notion of displacing this water in a design setting, particularly in the absence of any clear opportunities to improve the bridging interaction.

Wat301 was also studied by IFST.[63] In the presence of KNI-272, the interaction energy between Wat301 and its surroundings is estimated to be −28.2 kcal mol^{-1}, while the entropy relative to bulk was 2.9 kcal mol^{-1} at room temperature. Qualitatively, both IFST and SPAM highlight how the hydrogen bonding interactions made by Wat301 are compensated by a significant entropic penalty, although the magnitudes differ due to the distinct formalisms, computational protocols, and assumptions made by each method. As initially pointed out with regard to the IFST results,[63] both the SPAM and IFST entropy estimates exceed the empirical upper limit provided by Dunitz[64] (about 2 kcal mol^{-1} at 300 K), which may be attributed to the stronger ordering effect of the specific environment at the binding interface.

Wat301 has been the subject of several double-decoupling free energy studies[37−39] where the standard free energy of binding was estimated to be between −1.9 and −10.0 kcal mol^{-1}, depending on the interacting ligands, the protonation states of Asp25, the force field parameters, and the protocols used. Although the enthalpy and entropy components were not calculated, it was speculated that the enthalpy of water binding would be favorable while the entropy of binding would be unfavorable due to the specific hydrogen bonding pattern that hinders the overall motion of water. It was also reasoned that the cyclic urea scaffold achieved improved potency by displacing a bound, entropically restrained water molecule while retaining the favorable enthalpic interactions. While this interpretation is qualitatively consistent with both SPAM and IFST results, the absolute free energy values are quite different. However, this is not surprising. As suggested by others,[19] the thermodynamic quantities captured by IFST and SPAM (as well as other "local" measures of water energetics such as SZMAP[65] and 3D-RISM[31,35]) are not technically comparable to absolute binding free energies of water. Besides concerns about convergence to a correct water distribution with MD (which could potentially be overcome via complementary techniques such as grand canonical Monte Carlo[51]), it has been noted that IFST does not take protein or ligand relaxation into account upon water displacement, although one could argue that accounting for relaxation of the surroundings is unnecessary if the water is going to be replaced by a ligand functional group in the course of SAR exploration. There are also differences specific to the double decoupling methodology. For example, Barillari et al.[39] suggested that double decoupling approaches that leave behind a cavity after the water is decoupled from its environment at the protein−ligand interface omit the contribution due to cavitation, which may explain why their result of −10.0 kcal mol^{-1} is much more stable than predicted by other double decoupling results. In addition, we point out that the standard free energy of binding ΔG°, i.e., the system free energy change relative to a standard concentration of 1 M (where water is treated as both the solvent and the ligand), is a fundamentally different quantity than local measures of water stability relative to the bulk, and thus there are theoretical reasons why the two types of results are not comparable.

Hen Egg White Lysozyme (HEWL). The interaction of HEWL with small organic solvents has been extensively studied both experimentally and computationally. Most recently, the mapping of hot spots in HEWL has been reported by three academic groups using a diverse set of computational techniques, including standard MD simulations in binary solvent,[66] grand canonical Monte Carlo simulations with different organic solvent probes,[67] and FTMap, a multiprobe, GRID-like solvent mapping technique, which utilizes continuum electrostatics.[68] HEWL is a small, rigid protein with 129 amino acids, four disulfide bonds, and a solvent-accessible surface area (SASA) of approximately 6700 Å^2. To demonstrate the ability of SPAM in identifying ligand binding site hot spots, a single calculation was carried out using the HEWL crystal structure 2LYO. A total of 388 hydration sites were identified by using a density cutoff of 0.05 Å^{-3}, roughly 0.06 per unit SASA (Å^2), from which a subset of 180 hydration sites with density greater than 0.07 Å^{-3} were selected for SPAM analysis. The calculated SPAM free energies relative to the bulk (ΔG_{SPAM}) range between −4.0 kcal mol^{-1} and 8.7 kcal mol^{-1} with a median of 0.2 kcal mol^{-1}. The overall interaction energy component is slightly more favorable than the bulk with a median of −0.7 kcal mol^{-1}. The SPAM mapping of the entire solvent accessible surface of HEWL in conjunction with the rich set of HEWL structural data (352 HEWL crystal structures

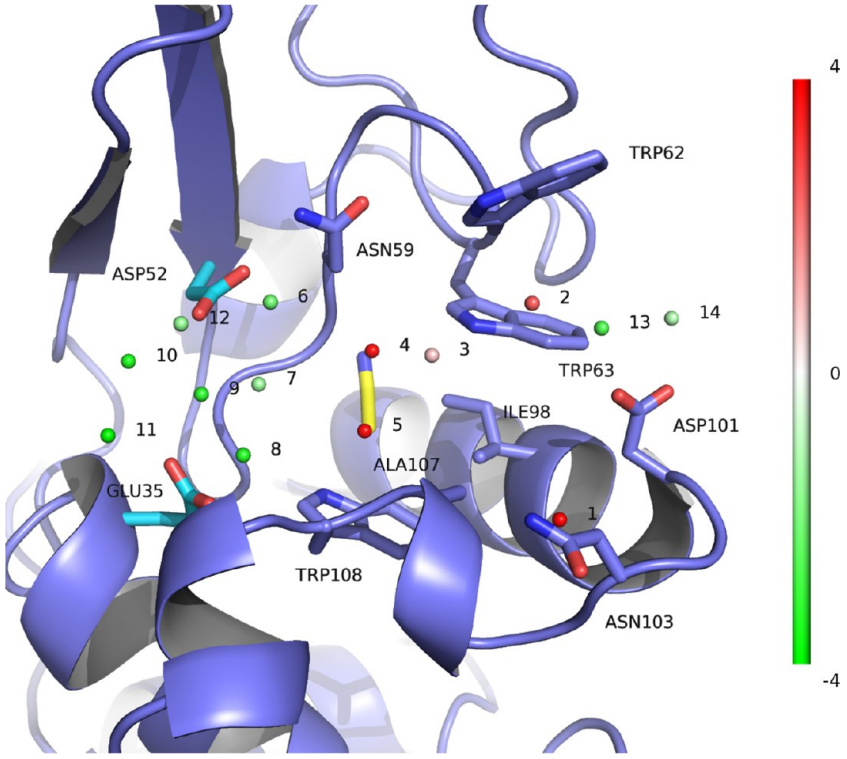


Figure 4. The NAG binding site of HEWL and the computed hydration sites discussed in this study. The residues identified by the NMR-based mapping technique[69] are shown as slate blue sticks, and the acidic residues involved in free arginine binding are shown in cyan. The centers of the computed hydration sites are shown as spheres and colored according to their relative SPAM free energy values, which are mapped to a green−white−red scale covering a range from −4 to +4 kcal mol^{-1}. The binding pose of acetonitrile determined in structure 2lyo was located in subsite C and was included as a reference.

and 93 ligand types; counting of ligand types is based on RCSB definitions) presents an ideal opportunity to scrutinize the connection between interfacial water properties and small molecule binding.

The most unfavorable hydration site is composed of a cluster of three conserved water molecules located near the intersection of three well-ordered loops at the N-terminus of the protein, distant from the most studied 2-N-acetyl-glycosamine (NAG) binding site. The site, enclosed by protein residues, is deeply recessed and just large enough to accommodate water molecules or monatomic ions (e.g., sodium and chloride). The ΔG_{SPAM} values of these three water sites (i.e., relative to the bulk) are 8.7, 6.5, and 3.5 kcal mol^{-1}. Although a glycerol molecule was modeled in this pocket of one HEWL crystal structure (PDB ID: 2b5z), examining the electron density map ($2F_o - F_c$) around the glycerol site as well as the geometry of the bound glycerol in 2b5z suggests that the electron density could have been interpreted as three water molecules or simple ions rather than as a distorted glycerol. It is not surprising that the size of the cavity precludes most small solvent molecules from populating this particular site.

At the second most unfavorably hydrated surface region, five water sites were found to cluster around the B and C subsites of NAG binding (water 1 through 5 in Figure 4). [The nomenclature of HEWL subsites was based on the location of each NAG repeating unit in crystal structures.] The SPAM free energies relative to the bulk are 6.4, 2.4, 0.9, 4.5, and 4.3 kcal mol^{-1}, with the first three in subsite B and the last two in subsite C. The interaction energy and entropy decomposition are summarized in Table 2. Most noticeably, the two water molecules labeled 4 and 5 in subsite C coincide precisely with

Table 2. The Thermodynamic Profiles of Selected Water Sites (in kcal mol^{-1}) in the NAG Binding Pocket

water sites	NAG subsite	ΔG_{SPAM}	σ	ΔE_{SPAM}	$-T\Delta S_{SPAM}$
1	B	6.4	0.052	1.1	5.3
2	B	2.4	0.21	1.3	1.1
3	B	0.92	0.22	−0.53	1.4
4	C	4.5	0.18	2.9	1.6
5	C	4.3	0.66	4.1	0.18
6	D	−2.1	0.21	−2.8	0.67
7	D	−1.5	0.17	−2.9	1.4
8	D	−3.3	0.28	−4.0	0.64
9	D	−3.2	0.41	−2.9	−0.28
10	D	−3.6	0.58	−2.8	−0.77
11	D	−4.0	0.19	−3.7	−0.39
12	D	−1.6	0.33	−2.6	1.0
13	A	−2.6	1.3	−1.1	−1.5
14	A	−2.3	0.79	−2.1	−0.19

the binding of acetonitrile. Although both water sites are nearly identical in terms of their relative SPAM free energy values, the decomposition of interaction energy and entropy components reveals differences that reflect on their individual environments. Water 4 is able to form a hydrogen bond with the backbone of Ile58, therefore having a less positive interaction energy term relative to the bulk compared to Water 5 (2.9 vs 4.1 kcal mol^{-1}). However, doing so results in a noticeable orientational restraint, which is captured by the entropy component (1.6 kcal mol^{-1}). The significance of organic solvent binding in subsite C had been probed experimentally using proton NMR spectroscopy by monitoring protein−ligand NOEs (nuclear Overhauser

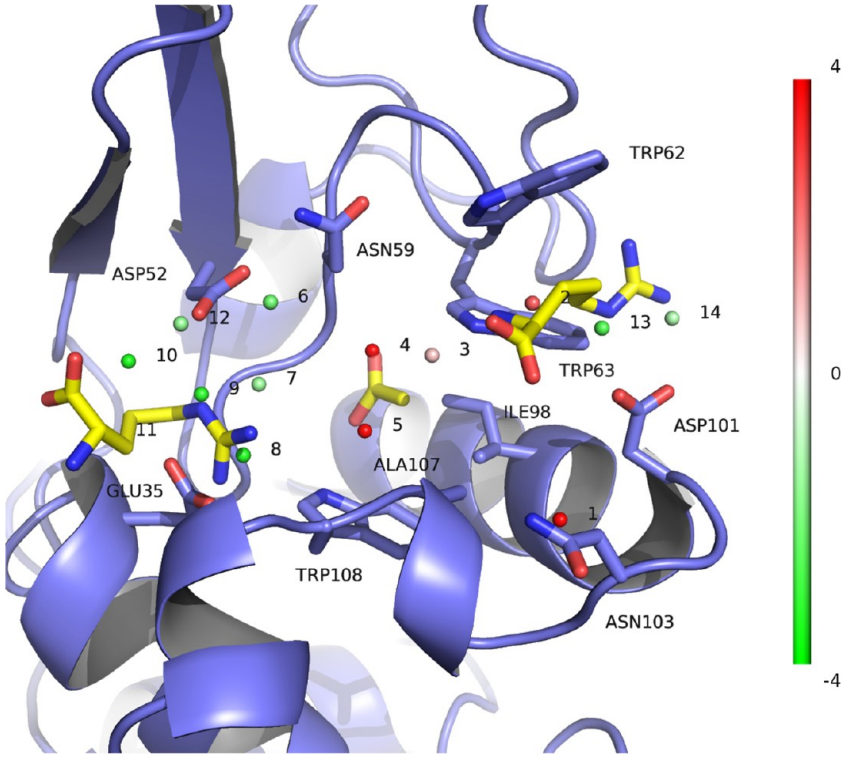


Figure 5. The free arginine binding in the NAG subsites A and D determined by crystallography (PDB ID: 3agi). The selected water sites from Figure 4 were superimposed for comparisons.

effects) and proton chemical shift changes.[69] The perturbed residues identified by Liepinsh and Otting, including Asp59, Trp62, Trp63, Ile98, Asp101, Asn103, Ala107, and Trp108 (Figure 4), are all located in subsites B and C and form an envelope around this cluster of water sites identified by the SPAM calculation.

As the only ligand binding sites confirmed by the NMR-based mapping technique, subsites B and C are significant in terms of judging the quality of computational hot spot mapping methods. Relying only on water molecules as probes, SPAM is able to identify both sites. The fact that the most prominent ligand binding sites of HEWL correlate well with hydration sites that are thermodynamically disadvantageous relative to the bulk seems to resonate well with the principle of hydrophobicity-driven ligand binding and the popular design strategy of displacing "unhappily" bound water molecules during lead optimization.

In contrast to subsites B and C, the significance of ligand binding in subsites A and D has been largely uncharacterized via experimental means and was not critically assessed in the early computational solvent mapping evaluations on HEWL. Our SPAM analysis of HEWL suggests that the solvation of subsites A and D is predominantly favorable compared to the bulk. A large cluster of seven water molecules is located in subsite D of the NAG binding pocket (Figure 4). All seven water sites share similar thermodynamic profiles, i.e., a favorable interaction energy component and relatively insignificant entropy component (water 6 through 12 in Figure 4 and Table 2), dictated by two key catalytic residues, Glu35 and Asp52. Similarly, a smaller cluster of two water molecules (waters 13 and 14 in Figure 4 and Table 2) is identified in subsite A centering around Asp101. Interestingly, in a recent study aiming to understand the structural basis of how free arginine promotes protein solubility and reduces protein aggregation, three arginine binding sites were found around HEWL where the binding appeared to be concentration dependent.[70] Two of these sites coincide with the two favorably hydrated regions of subsites A and D indicated by SPAM (Figure 5), where the arginine residues desolvate the pockets by complementing the acidic residues.

It is worth mentioning that the third arginine site is near Arg5 and Trp123 (about 20 Å from subsite C) and overlaps with an unfavorable hydration site instead (Figure 6, ΔG_{SPAM} = 2.5 kcal mol^{-1}, ΔE_{SPAM} = 1.4 kcal mol^{-1}, and $-T\Delta S_{SPAM}$ = 1.1 kcal mol^{-1}), which suggests small-molecule binding may occur at this location by simple dewetting as opposed to by being energetically complementary to the protein surface. This assessment is supported by the multitude of solvent binding events identified by HEWL crystallography. Besides arginine, many types of organic solvent, including TRIS (PDB ID: 1dpw), ethyl glycinate (PDB ID: 3ato), isopropanol (PDB ID: 3rt5), acetate (PDB ID: 3ojp), glycerol (PDB ID: 3a92), ethylene glycol (PDB ID: 1gwd), and ethanol (PDB ID: 3rw8), were also found in this third site, making no hydrogen bonds with the protein. On the other hand, the arginine binding in subsites A and D appears more selective, especially the one near Glu35 and Asp52 having a strong preference for basic moieties, such as the piperazine of HEPES (PDB ID: 2hub), the 2-amino group of TRIS (PDB ID: 1dpw), and the N-terminus of ethyl glycinate (PDB ID: 3ato). The significance of solvent binding sites observed in protein crystallography is often difficult to gauge, yet the preferential binding of charged species in these subsites is consistent with the SPAM analysis, which suggests that ligands need to make strong interactions (e.g., salt-bridge) to compete with these favorably bound water molecules, especially when there is little predicted entropic gain from desolvation. Reproducible sites from multisolvent protein crystallography similar to those in HEWL may carry additional

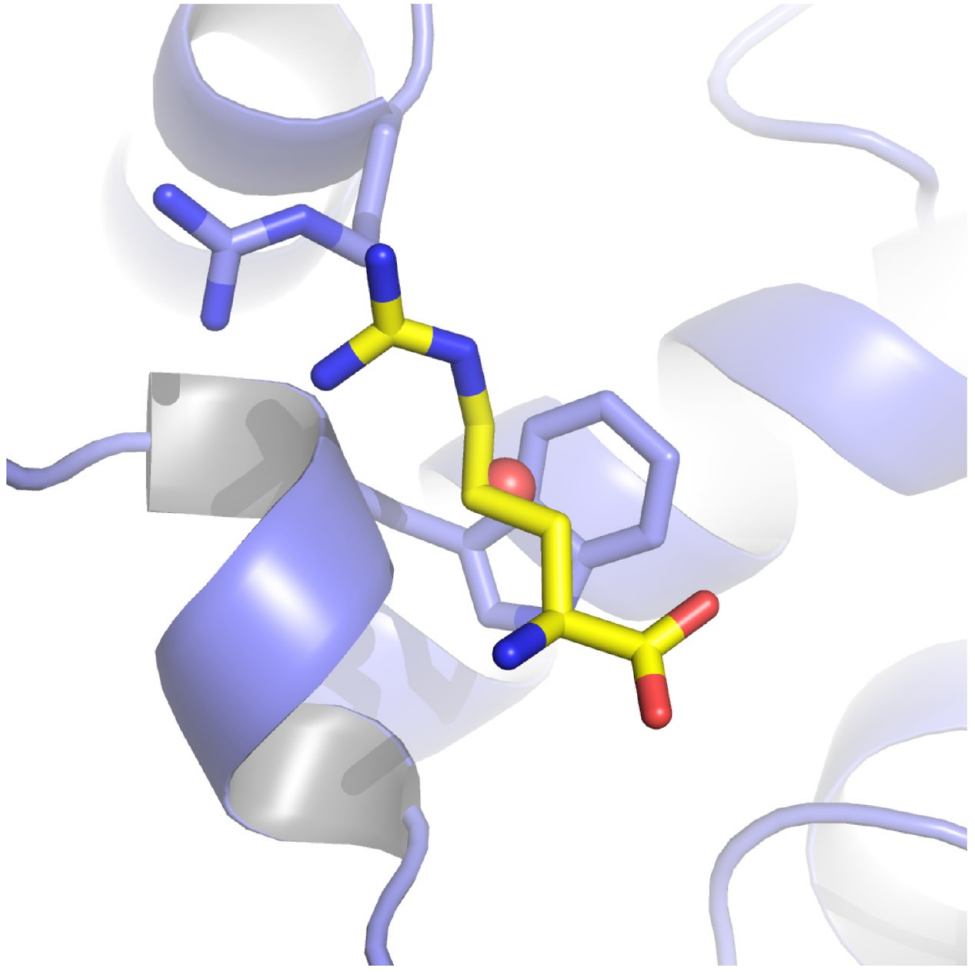


Figure 6. The third arginine binding site near Arg5 and Trp123 determined by crystallography (PDB ID: 3agi). The corresponding water site identified by SPAM is shown as a red sphere, colored according to its ΔG_{SPAM} value using the scale described in Figure 4.

interpretable information regarding the characteristic binding of protein surfaces. The SPAM analysis on HEWL demonstrates that extracting this information for the entire protein surface is possible. This initial SPAM analysis of the full HEWL surface presented here is preliminary but encouraging to warrant further studies on other systems.

CONCLUSIONS

The principles of solvation of large solutes with diverse shape, curvature, and surface composition are fundamental to understanding their behaviors in solution and are critical components of molecular association in aqueous environments. Motivated by the need for a design tool that identifies hot spots for a given target as well as the recent attention to computational solvent mapping and water characterization techniques, we devised SPAM as a simple statistical mechanics-based approach to extract both average interaction energy and entropy information for discrete hydration sites from the probability distribution of water interaction energies. The ability of SPAM to identify hot spots and provide additional information on the nature of individual hydration sites was demonstrated using two well studied systems, HIV1 protease and HEWL, as a small part of a more extensive case study that included both retrospective analysis and prospective applications. Despite the simplification of using a single type of probe molecule (water) and neglecting protein flexibility, SPAM was able to capture the energy−entropy compensation present in the flap water of HIV1 protease as well as all major hot spots in the NAG binding site of HEWL, illustrating the ability to provide reasonable free energy estimates. The latter study also provided the clear distinction between two types of arginine binding sites observed by X-ray crystallography, namely between sites where salt-bridge formation is required to compensate for highly favorable interactions between water and the protein, versus cases where neutral polar species were sufficient to displace water in the site. The performance is consistent for all retrospective targets that we have examined (data not shown). In the two prospective evaluations carried out at GSK that we are unable to publish at this point, SPAM was able to identify key regions of the binding pocket that led to significant potency improvements of up to 4 orders of magnitude. This outcome is encouraging and suggests that SPAM provides a reliable assessment of the locations and nature of hot spots in protein binding sites.

We also found that it is valuable to examine the breakdown of interaction energy and entropy components instead of focusing on a single thermodynamic measure. During the course of this study we often encountered varying degrees of cancellation between large favorable interaction energy and large unfavorable entropy components due to precise hydrogen bonding interactions between water and protein, and such water molecules were typically replaced by essential polar

functional groups in ligands. However, the balance between these compensating terms can be used to advise the nature of the displacing functional group, e.g., water with a small free energy difference relative to the bulk, and a large interaction energy with the site would likely require that its electrostatic interactions be mostly replaced by a displacing ligand (possibly requiring multiple hydrogen bonds and/or a charged functional group), whereas water with the same free energy difference and a small interaction energy might be readily displaced by a wider variety of groups, or perhaps by a less polar functional group. Similarly, when attempting to improve ligand binding affinity by making a less "happy" water "happier," it is useful to understand the balance between energetic interactions and entropy in order to produce the desired effect. Broadly speaking, we feel that the nature of such hydration sites may not be interpreted correctly if only a single thermodynamic measure is utilized, unless augmented with other means or user experience.

Designing small molecules that are sufficiently potent against their targets while possessing the right physical and chemical property profile in order to demonstrate *in vivo* efficacy is a daunting task, primarily owing to insufficient knowledge around the mechanisms of small molecule absorption, distribution, metabolism, excretion, and toxicity (ADMET). In the absence of mechanistic guidance, reducing molecular weight and lipophilicity by improving ligand binding efficiency may be a practical solution. Developing the ability to understand and predict protein binding site hot spots is a first step in this direction.

AUTHOR INFORMATION

Corresponding Author

*E-mail: guanglei.x.cui@gsk.com.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank the members of the global computational chemistry group at GSK for useful discussions, especially Colin M. Edge, Ian D. Wall, Kaushik Raha, Alan G. Graves, Neysa Nevins, and Marti Head. G.C. thanks the GSK Summer Talent Identification Program (STIP) for supporting J.M.S. and the development of SPAM.py. J.M.S. acknowledges financial support from the NSF GRFP award. E.S.M. thanks Woody Sherman and Robert Abel of Schrodinger, LLC as well as Anthony Nicholls of OpenEye Scientific Software, Inc. for useful discussions about water thermodynamics.

REFERENCES

(1) Ball, P. *Chem. Rev.* **2007**, *108*, 74−108.
(2) Grossman, M.; Born, B.; Heyden, M.; Tworowski, D.; Fields, G. B.; Sagi, I.; Havenith, M. *Nat. Struct. Mol. Biol.* **2011**, *18*, 1102−1108.
(3) Snyder, P. W.; Mecinović, J.; Moustakas, D. T.; Thomas, S. W.; Harder, M.; Mack, E. T.; Lockett, M. R.; Héroux, A.; Sherman, W.; Whitesides, G. M. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 17889−17894.
(4) Li, I. T. S.; Walker, G. C. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 16527−16532.
(5) Clackson, T.; Wells, J. A. *Sciences (New York)* **1995**, *267*, 383−386.
(6) Mattos, C.; Ringe, D. *Nat. Biotechnol.* **1996**, *14*, 595−599.
(7) Kuntz, I. D.; Chen, K.; Sharp, K. A.; Kollman, P. A. *Proc. Natl. Acad. Sci. U. S. A.* **1999**, *96*, 9997−10002.
(8) Bembenek, S. D.; Tounge, B. A.; Reynolds, C. H. *Drug Discovery Today* **2009**, *14*, 278−283.
(9) Hajduk, P. J. *J. Med. Chem.* **2006**, *49*, 6972−6976.
(10) Leeson, P. *Nature* **2012**, *481*, 455−456.
(11) Leeson, P. D.; Empfield, J. R. *Reducing the Risk of Drug Attrition Associated with Physico-Chemical Properties*; Elsevier: New York, 2010; Vol. *45*, pp 393−407.
(12) Leeson, P. D.; St-Gallay, S. A. *Nat. Rev. Drug Discovery* **2011**, *10*, 749−765.
(13) Tanford, C. *The Hydrophobic Effect: Formation of Micelles and Biological Membranes*, 2nd ed.; Wiley: New York, 1980.
(14) Madan, B.; Sharp, K. *J. Phys. Chem.* **1996**, *100*, 7713−7721.
(15) Chandler, D. *Nature* **2005**, *437*, 640−647.
(16) Berne, B. J.; Weeks, J. D.; Zhou, R. *Annu. Rev. Phys. Chem.* **2009**, *60*, 85−103.
(17) Mattos, C.; Bellamacina, C. R.; Peisach, E.; Pereira, A.; Vitkup, D.; Petsko, G. A.; Ringe, D. *J. Mol. Biol.* **2006**, *357*, 1471−1482.
(18) Mattos, C.; Ringe, D. *Curr. Opin. Struct. Biol.* **2001**, *11*, 761−764.
(19) Michel, J.; Tirado-Rives, J.; Jorgensen, W. L. *J. Am. Chem. Soc.* **2009**, *131*, 15403−15411.
(20) Goodford, P. J. *J. Med. Chem.* **1985**, *28*, 849−857.
(21) Young, T.; Abel, R.; Kim, B.; Berne, B. J.; Friesner, R. A. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104*, 808−813.
(22) Lazaridis, T. *J. Phys. Chem. B* **1998**, *102*, 3531−3541.
(23) Abel, R.; Young, T.; Farid, R.; Berne, B. J.; Friesner, R. A. *J. Am. Chem. Soc.* **2008**, *130*, 2817−2831.
(24) Beuming, T.; Farid, R.; Sherman, W. *Protein Sci.* **2009**, *18*, 1609−1619.
(25) Higgs, C.; Beuming, T.; Sherman, W. *ACS Med. Chem. Lett.* **2010**, *1*, 160−164.
(26) Abel, R.; Salam, N. K.; Shelley, J.; Farid, R.; Friesner, R. A.; Sherman, W. *ChemMedChem.* **2011**, *6*, 1049−1066.
(27) Beuming, T.; Che, Y.; Abel, R.; Kim, B.; Shanmugasundaram, V.; Sherman, W. *Proteins* **2012**, *80*, 871−883.
(28) Wilson, E. K. *Chem. Eng. News* **2012**, *90*, 64−65.
(29) Nguyen, C. N.; Young, T. K.; Gilson, M. K. *J. Chem. Phys.* **2012**, *137*, 044101+.
(30) Li, Z.; Lazaridis, T. In *Computational Drug Discovery and Design*; Baron, R., Ed.; Springer: New York, 2012; Methods in Molecular Biology, Vol. *819*, pp 393−404.
(31) Imai, T.; Hiraoka, R.; Kovalenko, A.; Hirata, F. *J. Am. Chem. Soc.* **2005**, *127*, 15334−15335.
(32) MOE - Structure-Based Design. http://www.chemcomp.com/MOE-Structure_Based_Design.htm.
(33) Tanger, J. C.; Pitzer, K. S. *J. Phys. Chem.* **1989**, *93*, 4941−4951.
(34) Rashin, A. A.; Bukatin, M. A. *J. Phys. Chem.* **1991**, *95*, 2942−2944.
(35) Sindhikara, D. J.; Hirata, F. *J. Phys. Chem. B* **2013**, *117*, 6718−6723.
(36) Gilson, M. K.; Given, J. A.; Bush, B. L.; McCammon, J. A. *Biophys. J.* **1997**, *72*, 1047−1069.
(37) Hamelberg, D.; McCammon, J. A. *J. Am. Chem. Soc.* **2004**, *126*, 7683−7689.
(38) Lu, Y.; Yang, C.-Y.; Wang, S. *J. Am. Chem. Soc.* **2006**, *128*, 11830−11839.
(39) Barillari, C.; Taylor, J.; Viner, R.; Essex, J. W. *J. Am. Chem. Soc.* **2007**, *129*, 2577−2587.
(40) Huggins, D. J. *J. Comput. Chem.* **2012**, *33*, 1383−1392.
(41) Henchman, R. H. *J. Chem. Phys.* **2007**, *126*, 064504+.
(42) Widom, B. *J. Chem. Phys.* **1963**, *39*, 2808−2812.
(43) Hummer, G.; Rasaiah, J. C.; Noworyta, J. P. *Nature* **2001**, *414*, 188−190.
(44) Baldwin, E. T.; Bhat, T. N.; Gulnik, S.; Liu, B.; Topol, I. A.; Kiso, Y.; Mimoto, T.; Mit-suya, H.; Erickson, J. W. *Structure (Oxford, U. K.)* **1995**, *3*, 581−590.
(45) Wang, Z.; Zhu, G.; Huang, Q.; Qian, M.; Shao, M.; Jia, Y.; Tang, Y. *Biochim. Biophys. Acta* **1998**, *1384*, 335−344.
(46) *Maestro*, version 9.2; Schrodinger, LLC: Portland, OR, 2011.

(47) Adachi, M.; Ohhara, T.; Kurihara, K.; Tamada, T.; Honjo, E.; Okazaki, N.; Arai, S.; Shoyama, Y.; Kimura, K.; Matsumura, H.; Sugiyama, S.; Adachi, H.; Takano, K.; Mori, Y.; Hidaka, K.; Kimura, T.; Hayashi, Y.; Kiso, Y.; Kuroki, R. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 4641−4646.
(48) Lam, P. Y. S.; Jadhav, P. K.; Eyermann, C. J.; Hodge, C. N.; Ru, Y.; Bacheler, L. T.; Meek, J. L.; Otto, M. J.; Rayner, M. M.; Wong, Y. N.; Chang, C.-H.; Weber, P. C.; Jackson, D. A.; Sharpe, T. R.; Erickson-Viitanen, S. *Science* **1994**, *263*, 380−384.
(49) Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. *J. Phys. Chem.* **1987**, *91*, 6269−6271.
(50) Michel, J.; Tirado-Rives, J.; Jorgensen, W. L. *J. Phys. Chem. B* **2009**, *113*, 13337−13346.
(51) Wang, L.; Berne, B. J.; Friesner, R. A. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 1326−1330.
(52) Sindhikara, D. J.; Yoshida, N.; Hirata, F. *J. Comput. Chem.* **2012**, *33*, 1536−1543.
(53) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. *Proteins* **2006**, *65*, 712−725.
(54) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. *J. Comput. Chem.* **2004**, *25*, 1157−1174.
(55) Jakalian, A.; Bush, B. L.; Jack, D. B.; Bayly, C. I. *J. Comput. Chem.* **2000**, *21*, 132−146.
(56) Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A. *J. Mol. Graphics Modell.* **2006**, *25*, 247−260.
(57) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K. *J. Comput. Chem.* **2005**, *26*, 1781−1802.
(58) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. *J. Chem. Phys.* **1995**, *103*, 8577−8593.
(59) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graphics* **1996**, *14*, 33−38.
(60) Chen, Y.; Duan, H. A QRS Complex Detection Algorithm Based on Mathematical Morphology and Envelope. *Engineering in Medicine and Biology Society, 2005. IEEE-EMBS 2005. 27th Annual International Conference of the IEEE*; Institute of Electrical and Electronics Engineers: New York, 2006; pp 4654−4657.
(61) Zhang, F.; Lian, Y. *IEEE Trans. Biomed. Circuits Syst.* **2009**, *3*, 220−228.
(62) Zhang, F.; Lian, Y. *J. Signal Processing Syst.* **2011**, *64*, 187−194.
(63) Li, Z.; Lazaridis, T. *J. Am. Chem. Soc.* **2003**, *125*, 6636−6637.
(64) Dunitz, J. D. *Science* **1994**, *264*, 670.
(65) *SZMAP*; OpenEye Scientific Software, Inc.: Santa Fe, NM, 2011.
(66) Lexa, K. W.; Carlson, H. A. *J. Am. Chem. Soc.* **2010**, *133*, 200−202.
(67) Kulp, J. L.; Kulp, J. L.; Pompliano, D. L.; Guarnieri, F. *J. Am. Chem. Soc.* **2011**, *133*, 10740−10743.
(68) Hall, D. H.; Grove, L. E.; Yueh, C.; Ngan, C. H.; Kozakov, D.; Vajda, S. *J. Am. Chem. Soc.* **2011**, *133*, 20668−20671.
(69) Liepinsh, E.; Otting, G.; Liepinsh, E.; Otting, G. *Nat. Biotechnol.* **1997**, *15*, 264−268.
(70) Ito, L.; Shiraki, K.; Matsuura, T.; Okumura, M.; Hasegawa, K.; Baba, S.; Yamaguchi, H.; Kumasaka, T. *Protein Eng., Des. Sel.* **2011**, *24*, 269−274.