JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

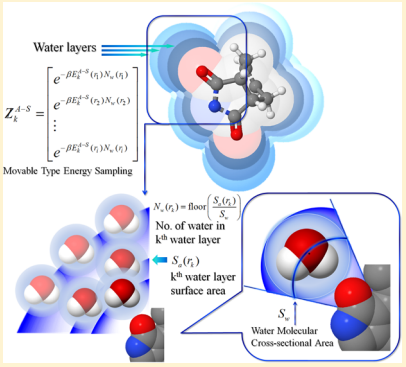
KECSA-Movable Type Implicit Solvation Model (KMTISM)

Zheng Zheng, Ting Wang, Pengfei Li, and Kenneth M. Merz, Jr.*

Institute for Cyber Enabled Research, Department of Chemistry and Department of Biochemistry and Molecular Biology, Michigan State University, 578 South Shaw Lane, East Lansing, Michigan 48824-1322, United States

Supporting Information

ABSTRACT: Computation of the solvation free energy for chemical and biological processes has long been of significant interest. The key challenges to effective solvation modeling center on the choice of potential function and configurational sampling. Herein, an energy sampling approach termed the "Movable Type" (MT) method, and a statistical energy function for solvation modeling, "Knowledge-based and Empirical Combined Scoring Algorithm" (KECSA) are developed and utilized to create an implicit solvation model: KECSA-Movable Type Implicit Solvation Model (KMTISM) suitable for the study of chemical and biological systems. KMTISM is an implicit solvation model, but the MT method performs energy sampling at the atom pairwise level. For a specific molecular system, the MT method collects energies from prebuilt databases for the requisite atom pairs at all relevant distance ranges, which by its very construction encodes all possible molecular configurations simultaneously. Unlike traditional statistical energy functions, KECSA converts structural statistical information into categorized atom pairwise interaction energies as a function of the radial distance instead of a mean force energy function. Within the implicit solvent model approximation, aqueous solvation free energies are then obtained from the NVT ensemble partition function generated by the MT method. Validation is performed against several subsets selected from the Minnesota Solvation Database v2012. Results are compared with several solvation free energy calculation methods, including a one-to-one comparison against two commonly used classical implicit solvation models: MM-GBSA and MM-PBSA. Comparison against a quantum mechanics based polarizable continuum model is also discussed (Cramer and Truhlar's Solvation Model 12).



INTRODUCTION

Modeling the solvation of molecular species has been of intense interest for many years, in part, because biological processes occur in an aqueous environment. Hence, through the experimental and theoretical understanding of the solvation process in terms of the aqueous structure around a solute and solvation free energies, insights into numerous chemical and biological processes have been obtained. In recent years, many methods have been developed to compute solvation free energies using explicit[1−57] and implicit[58−148] solvent models. Reviewing all the ongoing efforts in this field is beyond the present article, but numerous excellent reviews are available that cover this broad and continually evolving field.[23,66,129] Solvation energy calculations with explicit solvent models are usually carried out with some sort of sampling method, for example, Monte Carlo (MC), molecular dynamics (MD), etc., while implicit solvent models normally treat solvent as a continuum dielectric. Electrostatic energies are then calculated with either classical or QM based methods. In the context of force field based methods, the Poisson−Boltzmann (PB) and Generalized Born (GB) models coupled with the hydrophobic solvent accessible surface area (SA) are among the most commonly used classical implicit solvent models.[127−148] They provide reasonable solvation free energies for both small molecules and macromolecules. On the other hand, ab initio computation becomes feasible for solvation energy calculations when solvent is considered as a continuum model; namely, the polarizable continuum models (PCM) with different QM methods,[31,32,35,37,39,46] COSMO,[77] and Cramer and Truhlar's Solvation Model (SM) series[78−126] are frequently applied in small molecule/ion solvation energy calculations. However, given the central role solvation plays in protein−ligand binding among other biological processes, even modest errors in computed solvation free energies can adversely affect, for example, ligand binding affinity predictions. Hence, there is an ongoing need to improve our ability to model the solvation process using a broad range of methodologies. Herein, we describe an approach that employs statistical potential models coupled with fundamental statistical mechanical methodologies to compute solvation free energies.

Besides classical force field models of molecular structure and solvation, statistical potentials have been widely and successfully used to study biological systems.[149−186] Different from force field energy functions, which usually separate the pairwise potential into nonpolar (Lennard-Jones) and electrostatic (Coulomb) potentials, statistical potential methods directly convert observed pairwise probabilities into a pairwise potential by introducing a hypothetical reference state.[149] Chemical environment effects on the pairwise potential are embedded in

Received: August 27, 2014
Published: December 17, 2014

DOI: 10.1021/ct5007828
J. Chem. Theory Comput. 2015, 11, 667−682

the number density distributions instead of via electrostatic and Lennard-Jones pairwise interactions. This definition simplifies statistical potentials but increases the error by classifying atoms with significantly different electron densities into the same category.[168,169] Efforts have been made to address this issue by introducing detailed atom type classification schemes or through the use of fragment based statistical potentials.[168,169]

Another important issue for statistical potentials is that their performance rely crucially on available structural data. Substantial protein and small molecule structural databases support the success of statistical potentials in many application areas, for example, protein folding[170−186] and protein−ligand binding.[149−169] Structural databases with accurately positioned crystal waters including the Protein Databank (PDB) and the Cambridge Structural Database (CSD) are used herein to build pairwise models of solute−solvent interactions.

Statistical potentials, from the theoretical point of view build on the concept of the potential of mean force (PMF), are developed using structural information regarding structurally characterized molecular systems. A mean potential between specific atom pairs ($\omega^{(2)}(r_{12})$) is directly generated from the frequency of occurrence of the atom pairs contained in a large database of molecules:

$$\omega_{ij}^{(2)}(r_{12}) = -\frac{1}{\beta}\ln(g^{(2)}(r_{12})) = -\frac{1}{\beta}\ln\left(\frac{\rho_{ij}(r_{12})}{\rho_{ij}^{*}(r_{12})}\right) \quad (1)$$

where $g^{(2)}$ is called a correlation function. $\beta = 1/k_B T$ and k_B is the Boltzmann constant and T is the temperature. $\rho_{ij}(r)$ is the number density for the atom pairs of types i and j observed in the known protein or ligand structures, and $\rho_{ij}^{*}(r)$ is the number density of the corresponding pair in the background or reference state. A central problem for statistical potentials is to model specific atom pairwise interactions removed from the background energy. In protein−ligand complexes, geometric information, that is, atom pairwise radial distributions, represents an averaged effect of all interactions in chemical space, including bond, angle, torsion, and long-range noncovalent forces. Converting these radial distributions into energy functions is a challenge.[169]

Computing free energies requires the availability of accurate energy functions but also requires extensive phase space sampling. In the NVT ensemble, the Helmholtz solvation free energy is calculated using the ratio of partition functions given as

$$\Delta G_{\mathrm{solv}}^{\mathrm{L}} \approx \Delta A_{\mathrm{solv}}^{\mathrm{L}} = -RT\ln\left[\frac{Z_{\mathrm{LS}}}{Z_{\mathrm{L}}}\right] = -RT\ln\left[\frac{\int e^{-\beta E_{\mathrm{LS}}(r)}dr}{\int e^{-\beta E_{\mathrm{L}}(r)}dr}\right] \quad (2)$$

Equation 2 represents the free energy change of transferring a molecule (L) from vacuum to aqueous solution. Many sampling methods have proven effective, for example, MD and MC methods; however, thoroughly sampling phase space is challenging for brute-force methods. A new sampling method, which we call the Movable Type (MT) method, was developed by our group in an attempt to avoid some of the pitfalls encountered by the more computationally intensive sampling methods. Via sampling of all atom pairwise energies, at all possible distances, using prebuilt databases and then combining these energies for all atom pairs found in the molecular system of interest, the MT sampling method was able to accurately estimate binding free energies as well as protein−ligand poses.[187]

In the following paragraphs, we discuss in detail the data selection process, atom type recognition and a novel reference state model that aided in the development of a new solute−solvent statistical potential method, which when combined with the MT sampling method predicts solvation free energies. To validate our model, a curated set of 393 small molecule solvation free energies, collected in the MNSol database of Cramer and Truhlar, was used.[193] Our computed results were then compared to those obtained using the MM-GBSA and MM-PBSA models available in AMBER[146−148] against the same data set.

METHOD

Movable Type Continuum Solvation Model. The Movable Type (MT) method[187] is a free energy method that generates the ensemble of the molecular system of interest using pairwise energies and probabilities. The term "Movable Type" originates from the printing technique where a database of symbols (letters, numerals, etc.) is created and then assembled using a movable type system. Similarly, MT free energy calculations start from the construction of a large database containing interaction energies between all classes of atom pairs found in the chemical space under investigation. An atom pairwise energy function is required to create the database, and the modified KECSA model is employed herein.

With an atom pairwise energy database, molecular "printing" is then performed by assembling the pairwise energies using a "printing forme". A fixed-size matrix (Z-matrix) is introduced to represent the Boltzmann-weighted energy ensemble, in which atom pairwise energies at different distances are assembled to simultaneously represent the ensemble and free energies of the chemical space under investigation.

Creation of the Z-matrix starts from the first atom pair in the observed molecular system, with all elements in the matrix as Boltzmann-weighted energies of the observed atom pair at different distances, selected from the energy database.

$$Z_k^L = \begin{bmatrix} e^{-\beta E_k^L(r_1)} & e^{-\beta E_k^L(r_{i+1})} & \cdots & e^{-\beta E_k^L(r_{n-i+1})} \\ e^{-\beta E_k^L(r_2)} & e^{-\beta E_k^L(r_{i+2})} & \cdots & e^{-\beta E_k^L(r_{n-i+2})} \\ \vdots & \vdots & \ddots & \vdots \\ e^{-\beta E_k^L(r_i)} & e^{-\beta E_k^L(r_j)} & \cdots & e^{-\beta E_k^L(r_n)} \end{bmatrix} \quad (3)$$

Z_k^L (Z-matrix) in eq 3 represents a Boltzmann-weighted energy (Boltzmann factor) matrix for the kth atom pair in the observed molecule L containing energies ranging from distance r_1 to r_n. A pointwise product of Z-matrices results in the Boltzmann-weighted energy combinations between different atom pairs at different distances with a sampling size of n (matrix size), as is shown in eq 4.

DOI: 10.1021/ct5007828
J. Chem. Theory Comput. 2015, 11, 667−682

$$Z_{12}^{L} = Z_1^{L} \times Z_2^{L} = \begin{bmatrix} e^{-\beta(E_1^L(r_1)+E_2^L(r'_1))} & e^{-\beta(E_1^L(r_{i+1})+E_2^L(r'_{i+1}))} & \cdots & e^{-\beta(E_1^L(r_{n-i+1})+E_2^L(r'_{n-i+1}))} \\ e^{-\beta(E_1^L(r_2)+E_2^L(r'_2))} & e^{-\beta(E_1^L(r_{i+2})+E_2^L(r'_{i+2}))} & \cdots & e^{-\beta(E_1^L(r_{n-i+2})+E_2^L(r'_{n-i+2}))} \\ \vdots & \vdots & \ddots & \vdots \\ e^{-\beta(E_1^L(r_i)+E_2^L(r'_i))} & e^{-\beta(E_1^L(r_j)+E_2^L(r'_j))} & \cdots & e^{-\beta(E_1^L(r_n)+E_2^L(r'_n))} \end{bmatrix} \quad (4)$$

where the atom pairwise energy of 1 and 2 are sampled within their respective distance ranges (r_i and r'_i).

Performing random disordered permutations to the Z-matrices maximizes the variety of energy combinations at different distances. Based on the assumption that the molecular energy is separable into atom pairwise energies in the same molecular system, a pointwise product of disordered matrices over all atom pairs in the observed molecule derives the final Z-matrix representing the collection of Boltzmann-weighted energies with n configurations for the entire molecule L (eq 5).

$$Z_{\text{total}}^{L} = \text{disorder}(Z_1^{L}) \times \text{disorder}(Z_2^{L}) \times \cdots \times \text{disorder}(Z_k^{L}) = \begin{bmatrix} e^{-\beta(E_1^L(r_5)+E_2^L(r_3)+\cdots+E_k^L(r_l))} & e^{-\beta(E_1^L(r_i)+E_2^L(r_{n-1})+\cdots+E_k^L(r_{i+2}))} & \cdots & e^{-\beta(E_1^L(r_3)+E_2^L(r_{i+1})+\cdots+E_k^L(r_n))} \\ e^{-\beta(E_1^L(r_{i+1})+E_2^L(r_i)+\cdots+E_k^L(r_1))} & e^{-\beta(E_1^L(r_{i+1})+E_2^L(r_2)+\cdots+E_k^L(r_{l+1}))} & \cdots & e^{-\beta(E_1^L(r_{i-1})+E_2^L(r_{l-2})+\cdots+E_k^L(r_3))} \\ \vdots & \vdots & \ddots & \vdots \\ e^{-\beta(E_1^L(r_n)+E_2^L(r_l)+\cdots+E_k^L(r_i))} & e^{-\beta(E_1^L(r_l)+E_2^L(r_{i-1})+\cdots+E_k^L(r_2))} & \cdots & e^{-\beta(E_1^L(r_{i-2})+E_2^L(r_{n-2})+\cdots+E_k^L(r_l))} \end{bmatrix} \quad (5)$$

As the final Z-matrix may contain unreasonable distance combinations between different atom pairs, a Q-matrix of atom pairwise radial distribution probabilities is introduced in order to avoid physically unreasonable combinations between different atom pairs at certain bond lengths, angles, and torsions. The elements in the Q-matrix were collected from a large structural database containing 8256 protein crystal structures from PDBBind v2013[190−192] database and 44766 small molecules from both PDBBind v2013 and the CSD small molecule database. Corresponding to each element in an atom pairwise Z-matrix, there is a distance-dependent probability value chosen from the radial distribution probability database of the same atom pair type. Q-matrices matching the composition of the corresponding Z-matrix are also assembled using pointwise products. The final Q-matrix for the molecule of interest is normalized before being multiplied by the final Z-matrix, assuring that the overall probability is 1. The sum of the final matrix ($\mathcal{C}_{\text{total}}^{L}$) gives the ensemble average of the Boltzmann factors with a sampling size of n (matrix size).

$$\langle e^{-\beta E_L} \rangle = \text{sum}(\mathcal{C}_{\text{total}}^{L}) = \text{sum}(\overline{\mathcal{Q}}_{\text{total}}^{L} \mathcal{Z}_{\text{total}}^{L}) \quad (6)$$

Hence, with a predefined sampling size of n for the Z and Q matrices, the energies of different molecular conformations can be generated simultaneously via matrix products over all atom pairs. The solvation free energy is then calculated by incorporating the ensemble average of the Boltzmann factors into

$$\Delta G_{\text{solv}}^{L} \approx -RT \ln\left[\frac{Z_{LS}}{Z_L}\right] = -RT \ln\left[\frac{\int e^{-\beta E_{LS}(r)} dr}{\int e^{-\beta E_L(r)} dr}\right] = -RT \ln\left[\frac{\text{DOF}_{LS}\langle e^{-\beta E_{LS}(r)}\rangle}{\text{DOF}_L\langle e^{-\beta E_L(r)}\rangle}\right] \quad (7)$$

where the energy of the molecule in solution (E_{LS}) is modeled as

$$E_{LS}(r) = E_L(r) + E_{L-S\ \text{interaction}}(r) \quad (8)$$

DOF_{LS} and DOF_L indicate the degrees of freedom of the molecule in solution and in the gas phase, which were assigned the same value in the current implicit water model for simplicity.

The free energy is computed directly from the NVT ensemble avoiding issues related to the additivity of the free energy. Theoretically[194,195] and experimentally,[196] it can be shown that the energy can be decomposed, while the entropy and free energies cannot. Herein, we assemble the interaction energies using eq 8 and then place this into eq 7 to directly compute the free energy, thereby, avoiding issues related to the decomposition of the free energy. This is a real advantage of the MT method, and in future work, we will describe using this approach to compute both entropies and energies using statistical mechanics.

The MT energy sampling method can incorporate both an explicit and implicit water model into a solvation free energy calculation. Our previous attempt utilized a simple continuum ligand−solvent interaction model.[186] A new semicontinuum water model is developed herein, in which the solute−solvent interaction is calculated by placing water molecules around the solute. Water molecules were modeled as isotropic rigid balls with van der Waals radii of 1.6 Å.[197,198] Water molecules were placed into isometric solute-surrounding solvent layers, starting from the solute's water accessible surface until 8 Å away from the solute's van der Waals surface with an increment of 0.005 Å per layer. The number of water molecules was limited by comparing their maximum cross-sectional areas with the solvent accessible surface area at each solvent layer for each atom in the solute molecules. The number of water molecules (N_w) accessible to each atom at distance R away from the atomic center of mass is rounded down via filtering using the maximum cross-sectional area (S_w) of water with the atomic solvent accessible surface area (S_a) in the solvent layer at distance R.

DOI: 10.1021/ct5007828
J. Chem. Theory Comput. 2015, 11, 667−682

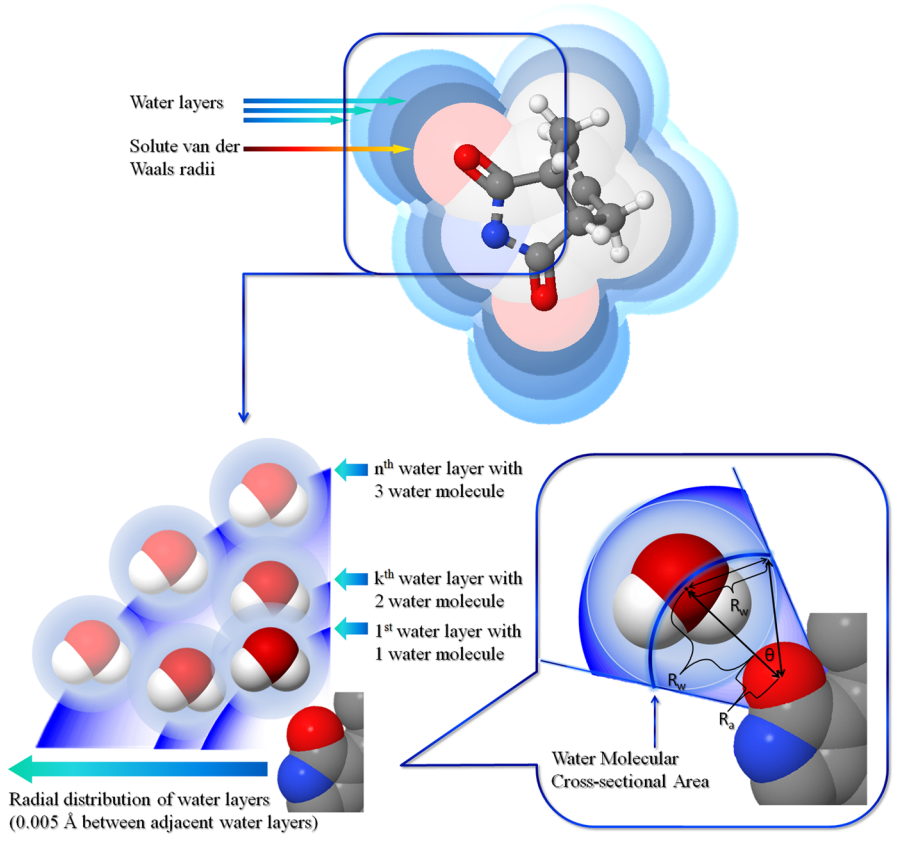


Figure 1. Modeling of the implicit solute−solvent model using the movable type method.

$$N_w(R) = \text{floor}\left(\frac{S_a(R)}{S_w}\right) \tag{9}$$

According to Figure 1, the maximum cross-sectional areas (S_w) of a water molecule is calculated as

$$\begin{aligned} S_w &= \int_{\pi/2-\theta}^{\pi/2} 2\pi(R_a + R_w)R_w \sin\left(\frac{\pi-\theta}{2}\right) d\left(\frac{\pi}{2} - \theta\right) \\ &= 2\pi(R_a + R_w)R_w \cos\left(\frac{\pi-\theta}{2}\right) \end{aligned} \tag{10}$$

where R_w and R_a are the van der Waals radii for water and the atom in the solute molecule, respectively.

The Boltzmann factor matrix for the kth solute atom−water ($\mathcal{Z}_k^{A-S}$) interaction is defined as a Boltzmann weighted solute atom-water energy multiplied by the number of accessible water molecules at the different distances. Multiplication of the Z-matrices for all solute atom−water interactions composes the final solute molecule-water Z-matrix ($\mathcal{Z}_{total}^{L-S}$), which when multiplied by the Z-matrix for the intrasolute molecular interactions ($\mathcal{Z}_{total}^{L}$) derives the final Z-matrix for the solute−solvent complex system ($\mathcal{Z}_{total}^{L-S}$). Multiplication of the final Z-matrix with its corresponding normalized Q-matrix generates the Boltzmann-weighted energy ensemble ($\mathcal{C}_{total}^{LS}$). With the energy ensembles for the solute molecule ($\mathcal{C}_{total}^{L}$) and solute−solvent complex ($\mathcal{C}_{total}^{LS}$), the solvation free energy is calculated using eq 14.

$$\mathcal{Z}_k^{A-S} = \begin{bmatrix} e^{-\beta E_k^{A-S}(r_1)N_w(r_1)} & e^{-\beta E_k^{A-S}(r_{i+1})N_w(r_{i+1})} & \cdots & e^{-\beta E_k^{A-S}(r_{n-i+1})N_w(r_{n-i+1})} \\ e^{-\beta E_k^{A-S}(r_2)N_w(r_2)} & e^{-\beta E_k^{A-S}(r_{i+2})N_w(r_{i+2})} & \cdots & e^{-\beta E_k^{A-S}(r_{n-i+2})N_w(r_{n-i+2})} \\ \vdots & \vdots & \ddots & \vdots \\ e^{-\beta E_k^{A-S}(r_i)N_w(r_i)} & e^{-\beta E_k^{A-S}(r_j)N_w(r_j)} & \cdots & e^{-\beta E_k^{A-S}(r_n)N_w(r_n)} \end{bmatrix} \tag{11}$$

$$\begin{aligned} \mathcal{Z}_{total}^{L-S} = \text{disorder}(\mathcal{Z}_1^{A-S}) \times \text{disorder}(\mathcal{Z}_2^{A-S}) \times \\ \cdots \times \text{disorder}(\mathcal{Z}_n^{A-S}) \end{aligned} \tag{12}$$

$$\mathcal{C}_{total}^{LS} = \overline{\mathcal{Q}}_{total}^{LS} \times \mathcal{Z}_{total}^{LS} = \overline{\mathcal{Q}}_{total}^{L-S} \times \overline{\mathcal{Q}}_{total}^{L} \times \mathcal{Z}_{total}^{L-S} \times \mathcal{Z}_{total}^{L} \tag{13}$$

$$\begin{aligned} \Delta G_{solv}^{L} &\approx -RT \ln\left[\frac{Z_{LS}}{Z_L}\right] = -RT \ln\left[\frac{\langle e^{-\beta E_{LS}(r)}\rangle}{\langle e^{-\beta E_L(r)}\rangle}\right] \\ &= -RT \ln\left[\frac{\text{sum}(\mathcal{C}_{total}^{LS})}{\text{sum}(\mathcal{C}_{total}^{L})}\right] = -RT \ln\left[\frac{\text{sum}(\overline{\mathcal{Q}}_{total}^{LS} \times \mathcal{Z}_{total}^{LS})}{\text{sum}(\overline{\mathcal{Q}}_{total}^{L} \times \mathcal{Z}_{total}^{L})}\right] \end{aligned} \tag{14}$$

DOI: 10.1021/ct5007828
J. Chem. Theory Comput. 2015, 11, 667−682

KECSA Energy Function. *Data Collection.* A collection of structural information is the first requirement to assemble a statistical potential. Many crystal structures in the Cambridge Structural Database (CSD) represent small molecules cocrystallized with water molecules.[188] Moreover, the Protein Data Bank (PDB) also contains a large number of protein−ligand complexes with water molecules at the interface between binding pockets and ligand molecules albeit the resolution of these structures are poorer than typically encountered in the CSD. Since our goal was to construct a solvation energy model focusing on small molecules, the CSD small molecule database became our primary resource for structural data. In order to data mine the CSD we only examined structures with (1) an R factor less than 0.1 and (2) all polymer structures and molecules with ions were excluded. The resulting data set selected contained 7085 small molecules surrounded by crystal water molecules.

Atom Type Recognition. Statistical potentials are derived by converting the number density distributions between two atoms or residues to energies; hence, they are "fixed-charge" models for the selected atom or residue pairs. In order to differentiate atoms of the same type but with different electron densities, a detailed atom type categorization has been employed where 21 atom types (shown in Table 1) were chosen from the database as having statistically relevant water molecule contact information contained within the CSD.

Table 1. List of 21 Atom Types in the Current Solvation Model

atom type	description
C1	sp^1 carbon
C2	sp^2 carbon
C3	sp^3 carbon
Car	aromatic carbon
N2	sp^2 nitrogen
N3	sp^3 nitrogen
N4	positively charged nitrogen
Nam	amide nitrogen
Nar	aromatic nitrogen
Npl3	trigonal planar nitrogen
Ow	water oxygen
O2	sp2 oxygen
O3	hydroxyl oxygen
OE	ether and ester sp3 oxygen
Oco2	carboxylate, sulfate, and phosphate oxygen
S2	sp2 sulfur
S3	sp3 sulfur
P3	sp3 phosphorus
F	fluorine
Cl	chlorine
Br	bromine

Energy Function Modeling. The pairwise radial distribution is a mixed consequence of direct pairwise contacts and indirect environmental effects, which is usually described as a "mean force" driven correlation function. For the atom pairwise radial distribution, the mean force potential is generally described by the following formula.

$$P_k = \frac{\int E_k(r) e^{-E_k(r)/RT} dr_{ak}\, dr_{bk} \cdots dr_{nk}}{\int e^{-E_k(r)/RT} dr_{ak}\, dr_{bk} \cdots dr_{nk}} \quad (15)$$

where P_k is the mean force potential on the kth particle in a system, $E_k(r)$ is the energy between the kth particle and any particle $i \in \{a, b, \cdots n\}$ in this system at a distance r separating these two particles.

Intrinsically, statistical potentials have difficulty in separating out various chemical environment effects on the observed atoms, thereby generating a major source of error in this class of models.[169] As noted by Thomas and Dill,[184] over-represented contacts in a structural database could mask the presence of other contacts. With the reference state presupposing a uniform availability of all contacts, quantitatively minor contacts are always underestimated in statistical potentials. A new statistical potential energy function called KECSA developed in our group defines a new reference state attempting to eliminate the contact masking due to quantitative preferences.[189]

Unlike traditional statistical potentials using a reference state mimicking the ideal gas state, KECSA defines a reference state energy or background energy as the energy contributed by all atoms surrounding the observed atom pairs. It introduces a reference state number distribution modeled by a linear combination of the number distribution under mean force ($\langle n_{ij}(r)\rangle$) and the number distribution of an ideal gas state ($(N_{ij}/V)4\pi r^2\Delta r$):

$$n_{ij}^{**}(r) = \left(\frac{N_{ij}}{V} 4\pi r^2 \Delta r\right) x + \langle n_{ij}(r)\rangle (1 - x) \quad (16)$$

where x indicates the intensity of the observed atom pairwise interaction in the chemical space V. This definition puts the number distribution of one certain observed atom pair in the reference state somewhere between the ideal gas state and the "mean force" state, depending on its relative strength. Stronger interactions have background energies closer to an ideal gas state while weaker interactions have background energies approaching the mean force state energy contributed by all atoms in the chemical space.

To build a KECSA energy function modeling solvent, solute and solvent−solute interactions requires us to define an "x term" for each atom pairwise interaction. Several approaches have been used to model x in our knowledge-based energy function. In the original KECSA, we simply used the number ratio of the chosen atom pair i and j over the total atom pairs in the chemical space to represent the intensity x. Meanwhile, based on the assumption that all contacts are uniformly available in the chemical space given by the selected database,[154] we assigned an identical x for every atom pair found in the given chemical space.

$$\begin{aligned} n_{ij}^{**}(r) &= \left(\frac{N_{ij}}{V} 4\pi r^2 \Delta r\right) x + \langle n_{ij}(r)\rangle (1 - x) \\ &= \left(\frac{N_{ij}}{V} 4\pi r^2 \Delta r\right) \frac{1}{N_t} + \langle n_{ij}(r)\rangle \left(1 - \frac{1}{N_t}\right) \end{aligned} \quad (17)$$

where N_t is the total atom type number in the chemical space.

The original model of $n_{ij}^{**}(r)$ is based on the notion that every atom pair has an equal contact opportunity in a background energy contributed by the other atom pairs, while neglecting the fact that the background energies have different effects on atom pairwise distributions with different interaction strengths (say atom i and j under a covalent bond constraint compared to atom k and l under a nonbond interaction constraint).

DOI: 10.1021/ct5007828
J. Chem. Theory Comput. 2015, 11, 667−682

Table 2. Performance of KMTISM, MM-GBSA, and MM-PBSA for the Prediction of the Solvation Free Energies of Neutral Molecules

	total neutral molecule set			amide set		
	KMTISM	MM-GBSA	MM-PBSA	KMTISM	MM-GBSA	MM-PBSA
R^2	0.792	0.734	0.804	0.660	0.493	0.509
Kendall's τ	0.755	0.708	0.793	0.568	0.484	0.465
raw RMSE (kcal/mol)	2.597	4.629	4.647	4.368	8.666	9.717
scaled RMSE (kcal/mol)	2.248	2.634	2.160	3.852	4.885	4.663
	hydrocarbon set			halocarbon set		
	KMTISM	MM-GBSA	MM-PBSA	KMTISM	MM-GBSA	MM-PBSA
R^2	0.699	0.906	0.954	0.648	0.004	0.594
Kendall's τ	0.663	0.748	0.887	0.656	0.091	0.625
raw RMSE (kcal/mol)	0.858	1.179	0.925	1.052	2.768	1.148
scaled RMSE (kcal/mol)	0.845	0.498	0.332	1.030	2.063	1.109
	oxygenated molecule set			organosulfur and organophosphorus set		
	KMTISM	MM-GBSA	MM-PBSA	KMTISM	MM-GBSA	MM-PBSA
R^2	0.829	0.881	0.916	0.762	0.751	0.777
Kendall's τ	0.657	0.723	0.754	0.680	0.626	0.618
raw RMSE (kcal/mol)	2.104	4.232	3.868	4.337	8.297	9.179
scaled RMSE (kcal/mol)	1.578	1.613	1.186	3.500	4.316	3.992
	nitrogenous molecule set			heterocycle set		
	KMTISM	MM-GBSA	MM-PBSA	KMTISM	MM-GBSA	MM-PBSA
R^2	0.615	0.485	0.795	0.604	0.528	0.552
Kendall's τ	0.420	0.412	0.592	0.652	0.622	0.646
raw RMSE (kcal/mol)	2.384	2.416	1.690	4.314	7.584	8.722
scaled RMSE (kcal/mol)	2.276	2.555	1.797	3.721	4.413	4.217
	oxygenated and nitrogenous molecule set			polyfunctional molecule set		
	KMTISM	MM-GBSA	MM-PBSA	KMTISM	MM-GBSA	MM-PBSA
R^2	0.545	0.747	0.694	0.736	0.615	0.650
Kendall's τ	0.565	0.663	0.621	0.726	0.577	0.609
raw RMSE (kcal/mol)	3.259	4.282	5.043	4.688	10.138	11.132
scaled RMSE (kcal/mol)	2.991	2.794	2.484	3.597	5.335	4.804

A more accurate $n_{ij}^{**}(r)$ model is introduced herein that takes every atom pairwise contact as a energy state distributed between an ideal gas state energy and mean force state energy following a Boltzmann distribution in the reference state. In this way, the x factor is defined as

$$x = \frac{N_{ij}(r)\mathrm{e}^{-\beta E_{ij}(r)}}{\sum_i^n \sum_j^n N_{ij}(r)\mathrm{e}^{-\beta E_{ij}(r)}} \tag{18}$$

where $\mathrm{e}^{-\beta Eij(r)}$ is the Boltzmann factor and $N_{ij}(r)$ is the degeneracy factor (contact number) for atom type pair i and j.

With the x term built up as a probability of all contacts, the number distribution of the observed atom pair in the background state $n_{ij}^{**}(r)$ is modeled as

$$n_{ij}^{**}(r) = \left(\frac{N_{ij}}{V}4\pi r^2\Delta r\right)\frac{N_{ij}(r)\mathrm{e}^{-\beta E_{ij}(r)}}{\sum_i^n \sum_j^n N_{ij}(r)\mathrm{e}^{-\beta E_{ij}(r)}} + (n_{ij}(r))\left(1 - \frac{N_{ij}(r)\mathrm{e}^{-\beta E_{ij}(r)}}{\sum_i^n \sum_j^n N_{ij}(r)\mathrm{e}^{-\beta E_{ij}(r)}}\right) \tag{19}$$

Hence, we can build the energy function for each atom type pair as

$$\begin{aligned} E_{ij}(r) &= -\frac{1}{\beta}\ln\left(\frac{n_{ij}(r)}{n_{ij}^{*}(r)}\right) \\ &= -\frac{1}{\beta}\ln\left[\frac{n_{ij}(r)}{\left(\frac{N_{ij}}{V}4\pi r^2\Delta r\right)x + (n_{ij}(r))(1-x)}\right] \\ &= \frac{1}{\beta}\ln\left[x\left(\frac{N_{ij}3r^2\Delta r}{R^3 n_{ij}(r)}\right) + (1-x)\right] \\ &= \frac{1}{\beta}\ln\left[\frac{N_{ij}(r)\mathrm{e}^{-\beta E_{ij}(r)}}{\sum_i^n \sum_j^n N_{ij}(r)\mathrm{e}^{-\beta E_{ij}(r)}}\left(\frac{N_{ij}3r^2\Delta r}{R^3 n_{ij}(r)}\right) + \left(1 - \frac{N_{ij}(r)\mathrm{e}^{-\beta E_{ij}(r)}}{\sum_i^n \sum_j^n N_{ij}(r)\mathrm{e}^{-\beta E_{ij}(r)}}\right)\right] \end{aligned} \tag{20}$$

In eq 20, with the energy functions built up in the chosen chemical space, each $E_{ij}(r)$ can be derived iteratively at discrete distance points. Using this model, every $E_{ij}(r)$ derived using the KECSA energy function is never a mean force potential between atom pair i and j as found in traditional statistical potentials.[189] Instead, $E_{ij}(r)$ represents a pure atom pairwise interaction energy between i and j because the reference state

DOI: 10.1021/ct5007828
J. Chem. Theory Comput. 2015, 11, 667−682

energy defined in KECSA is a background energy contributed by all other atom pairs and not just the ideal gas state energy.

Test Set Selection. Two major differences between KMTISM and other continuum solvation models are (1) the MT method calculates the free energy change using a ratio of partition functions in the NVT ensemble, while traditional continuum solvation models separate the Gibbs free energies into linear components with enthalpy and entropy components.[139] (2) Electrostatic interactions are implicit via the categorization of pairwise atom-types in the KECSA model while they are calculated explicitly using classical or QM based energy calculation approaches. In this manner, KMTISM can be viewed as the null hypothesis for the addition of explicit electrostatic interactions. If electrostatic interactions are added to a model, it should outperform the knowledge-based approach; if not, the explicit electrostatic model is not an improvement over the implicit inclusion of this key interaction. A key concern for the KMTISM method is the validity of using preconstructed atom-type pairwise energy data in free energy calculations for molecules with similar atoms, which differ in their chemical environments. Hence, KMTISM was examined with test compounds containing C, O, N, S, P, and halogen atoms within different functional groups. Given that the KECSA energy function was parametrized using organic structural data, validation focused on reproducing the aqueous solvation free energy for drug-like molecules. The Minnesota Solvation Database is a well-constructed data set, including aqueous solvation free energies for 391 neutral molecules and 144 ions. This data set was filtered down to 372 neutral molecules and 21 ions in our test set via the exclusion of (1) inorganic molecules, and (2) molecules with atom types not represented in the KECSA potential. This test set, including various hydrocarbons, mono- and polyfunctional molecules with solvation free energies ranging from −85 to 4 kcal/mol, was further classified into different subsets based on the functional groups within the molecules. Some molecules were included in several subsets due to their polyfunctional nature.

Carbon, nitrogen, and oxygen are essential elements in organic molecules. More than one-half of the compounds in the neutral test set (219 out of 372 compounds) were composed exclusively of C, N, and O atoms. From the Minnesota Solvation Database we created four subsets from these 219 molecules: 41 hydrocarbons, 91 molecules with oxygen based functional groups, 44 molecules with nitrogen based functional groups, and 43 molecules with mixed N and O functional groups. Validation also focused on molecules with sulfur, phosphorus, and halogen atoms, which play important roles in organic molecules. A test set with only halocarbons was created for the purpose of avoiding interference from other polar atoms. Sulfur and phosphorus, on the other hand, are often contained in oxyacid groups in organic molecules. Collected from the neutral data set, a test set with sulfur or phosphorus-containing molecules was composed. Heterocycles, amides, and their analogs are pervasive in drug-like molecules and are well represented in the Minnesota Solvation Database. Thirty-seven heterocyclic compounds and 33 amides and their analogs were categorized into two subsets. In addition, 28 molecules containing three or more different functional groups were selected to provide a challenging test with complex and highly polar molecules. The ion test set was limited to biologically relevant ions herein resulting in positively charged nitrogen and negatively charged carboxylate oxygen subsets. In this way, 21 ions were chosen from Minnesota Solvation Database (11 cations and 10 anions). Alkoxide ions among others present in the Minnesotae Solvation Data set will be examined in the future with the aid of molecular dynamics simulation of ion−water interactions for these ions but were excluded herein. Calculation results using KMTISM, MM-GBSA, and MM-PBSA for all test sets are contained in the Supporting Information. Only statistical data are given in Table 2 and Table 3 in this manuscript.

Table 3. Performance of KMTISM, MM-GBSA, and MM-PBSA for the Prediction of the Solvation Free Energies of Ions

	ion set		
	KMTISM	MM-GBSA	MM-PBSA
R^2	0.351	0.000	0.003
Kendall's τ	0.258	−0.057	−0.067
RMSE (kcal/mol)	5.777	11.736	10.481
	carboxylate set		
	KMTISM	MM-GBSA	MM-PBSA
R^2	0.239	0.161	0.166
Kendall's τ	−0.090	−0.180	−0.180
RMSE (kcal/mol)	5.337	11.918	11.252
	charged amine set		
	KMTISM	MM-GBSA	MM-PBSA
R^2	0.557	0.008	0.009
Kendall's τ	0.491	−0.127	−0.127
RMSE (kcal/mol)	6.149	11.569	9.727

RESULTS AND DISCUSSION

Comparison with MM-GBSA and MM-PBSA Results. Data analysis covered solvation free energy calculations for all subsets using KMTISM along with the corresponding MM-GBSA and MM-PBSA results. Both MM-GBSA and MM-PBSA calculation were performed using AMBER with the General AMBER force field (GAFF). GB parameters were set as igb = 2 and saltcon = 0.100. In the PB calculation, istrng = 0.100.

Against the neutral molecule test set, KMTISM and MM-PBSA gave comparable correlation coefficients (R^2) and both were better than MM-GBSA. According to Kendall's τ values, MM-PBSA outperformed the other two methods in ranking ability, with KMTISM as the second best. In terms of accuracy of the models, KMTISM has the lowest root-mean-square error (RMSE), while the RMSE values for MM-GBSA and MM-PBSA were almost twice as large. A plot of the experimental data vs calculated data is given in Figure 2 while the statistical results are summarized in Tables 2 and 3.

For the purpose of a more thorough analysis, a linear scaling model was applied to all three methods using eq 21 in order to bring their respective regression lines closer to $y = x$. Linear scaling, due to its data set dependence, did not improve the performance of the methods, but instead, it provided a way to examine the deviation of the calculated results from their regression lines.

$$y_{\text{corrected}} = \frac{(y_{\text{raw}} - b)}{a} \tag{21}$$

Here, a and b are the slope and the intercept of the regression line between experimental data and computed data, respectively.

DOI: 10.1021/ct5007828
J. Chem. Theory Comput. 2015, 11, 667−682

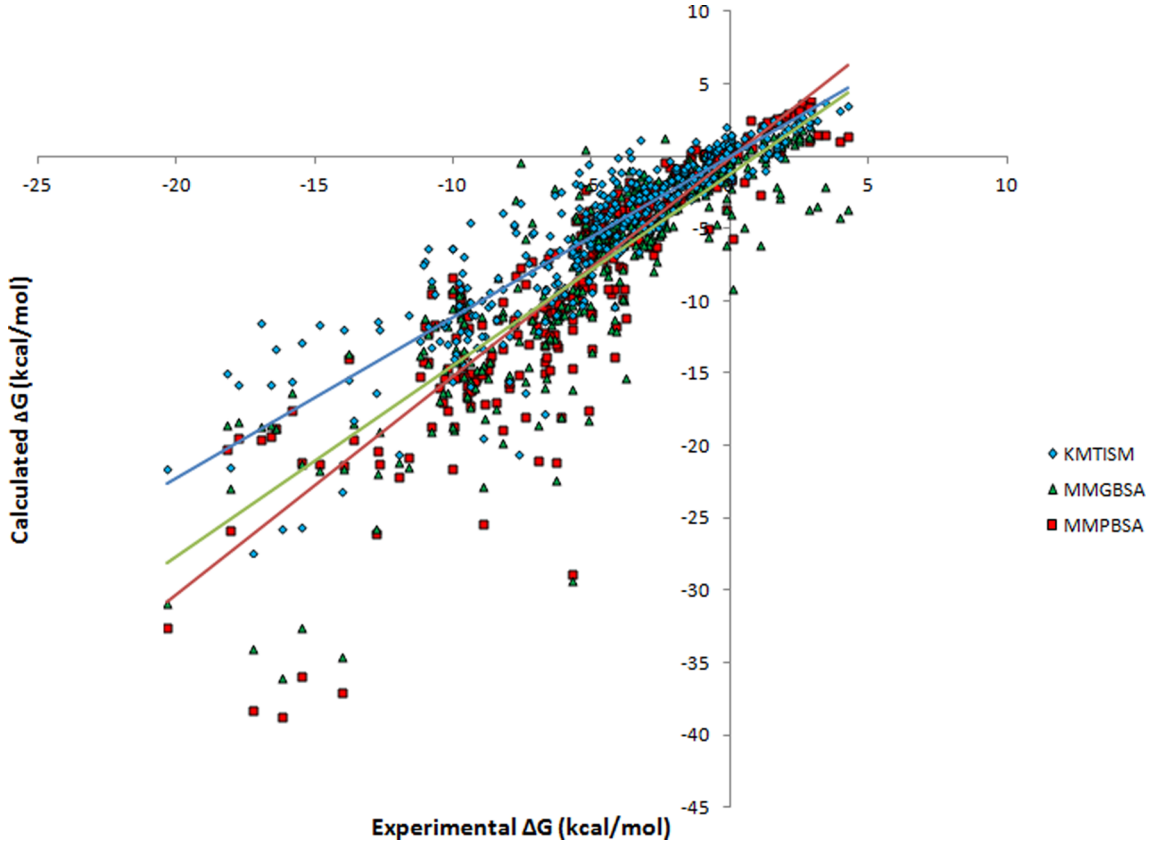


Figure 2. KMTISM, MM-GBSA, and MM-PBSA calculated vs experimental solvation free energies (kcal/mol) for 372 neutral molecules (kcal/mol).

The MM-GBSA and MM-PBSA results suggested a biased regression against the experimental solvation energies (y = 1.3186x − 1.2902 for MM-GBSA and y = 1.5095x − 0.1556 for MM-PBSA, where x and y represent the experimental and calculated solvation free energies). The slopes of their regression lines indicated an overestimation of the solvation free energies using these two methods. The significant improvement in RMSE values for MM-GBSA and MM-PBSA after the linear scaling as well as their correlation coefficient (R^2 and Kendall's τ) values indicate that they have a better ranking ability than free energy prediction. On the other hand, KMTISM's regression function (y = 1.1078x − 0.0811) affected the RMSE to a lesser extent.

Results for different test sets classified by functional groups provide deeper insights into the prediction abilities of the three computational models. For all three approaches, errors increased with the complexity of the functional groups. Solvation free energies of hydrocarbons, halocarbons, and oxygen and nitrogen containing molecules were better reproduced than molecules with other functional groups (see Figure 3), while amides and mixed polyfunctional groups resulted in the largest RMSEs (see Figure 4).

Among all data sets, the hydrocarbon set was reproduced with the lowest RMSE value for all of the models, while unsaturated hydrocarbons proved more difficult for KMTISM than the other two methods. The drop in R^2 for KMTISM was due to overestimation of the solvation free energies of unsaturated hydrocarbons, for example, the two compounds with the largest error (~2 kcal/mol) were ethene and s-trans-1,3-butadiene, where all heavy atoms were sp^2 hybridized. In the KECSA training set, which includes mostly drug-like molecules, different adjacent polar functional groups significantly altered the electron densities of adjacent unsaturated carbon atoms (via delocalization, for example) and this varies the electrostatic characteristics of these carbon atoms more than that seen in the case of sp^3 hybridized carbon.

On the other hand, polar atom types in the KECSA energy function were classified according to their corresponding hydrophilic functional groups and were less affected by adjacent functional groups. Polar atom type-water radial probabilities were driven by a more fine grained atom pairwise set of interactions, thereby, improving the performance of the KECSA energy function for these groups. The oxygenated molecule set and halocarbon set were among the top three test sets based on KMTISM's performance according to RMSE. Against the oxygen containing molecule set, KMTISM gave a correlation coefficient comparable to MM-PBSA, while its RMSE was better than MM-GBSA. For the halocarbon set, KMTISM outperformed the MM-PB/GBSA methods according to the RMSE and correlation coefficients. Especially for fluorocarbons whose solvation free energies were much better reproduced by KMTISM compared to the MM-PB/GBSA methods. For the data set of 8 fluorocarbons, the RMSE for KMTISM was 1.1 kcal/mol compared to RMSE values as 5.8 kcal/mol for MM-GBSA and 2.2 kcal/mol for MM-PBSA.

The variety of atom types in different molecules with different chemical environments make the atom type classification process an inherent source of error for any statistical energy function. The use of atom types in classical potentials is also an issue, but it is typically mitigated by an explicit electrostatic model, which takes into account environmental effects. Collecting similar atom types into the same category can reduce the predictive ability of a statistical potential. For example, the sp^3 oxygen atom in ethers and

DOI: 10.1021/ct5007828
J. Chem. Theory Comput. 2015, 11, 667−682

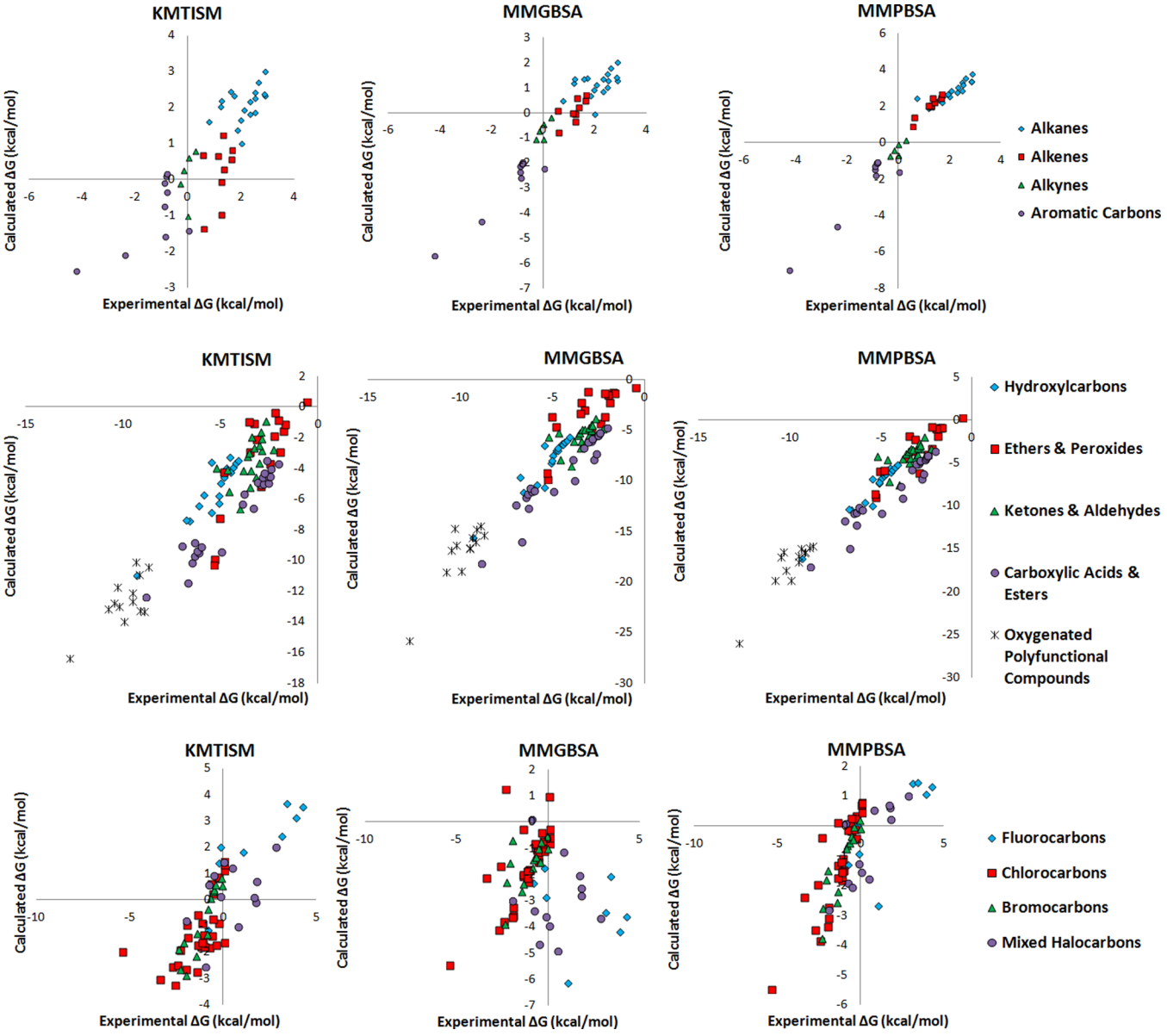


Figure 3. KMTISM's top three performing test sets according to RMSE. KMTISM, MM-GBSA, and MM-PBSA calculated solvation free energies (kcal/mol) vs experimental data are listed from left to right. From the top to bottom, the test sets are the hydrocarbon, oxygen containing, and halocarbon test sets.

peroxides were modeled using the same atom type within KECSA. This resulted in the solvation free energies for the two peroxides to be overestimated by KMTISM, that is, the ΔG_{sol} for methylperoxide was −9.90 kcal/mol or −8.86 kcal/mol (scaled) vs the experimental value of −5.28 kcal/mol and the ΔG_{sol} for ethylperoxide was −10.27 kcal/mol or −9.20 kcal/mol (scaled) vs the experimental value of −5.32 kcal/mol. In comparison with the MM-GB/PBSA methods the ΔG_{sol} for methylperoxide was −9.89 kcal/mol or −6.51 kcal/mol (scaled) using MM-GBSA and −9.07 kcal/mol or −5.90 kcal/mol (scaled) using MM-PBSA; the ΔG_{sol} for ethylperoxide was −9.21 kcal/mol or −6.00 kcal/mol (scaled) using MM-GBSA and −8.59 kcal/mol or −5.59 kcal/mol (scaled) using MM-PBSA. Hence, none of the methods examined particularly did well modeling the solvation free energy of peroxides.

As the structural complexity of a molecule increased so did the computed RMSE. Possible long-range polar interactions add additional difficulty to accurate solvation free energy calculations using the methods described herein. The largest errors were found in the amide set, organosulfur and organophosphorus set, and polyfunctional molecule set for all three methods. With lower errors for most individual polar functional groups based on the analysis of the monofunctional test set results, KMTISM had less cumulative error against these three test sets when compared with the MM-GB/PBSA methods for both the raw RMSE and scaled RMSE values (see Table 2). This result suggests that KMTISM has an advantage over the MM-GB/PBSA methods for the prediction of the solvation free energy of polyfunctional molecules. This advantage will have a significant effect on the ability of this model to predict, for example, protein−ligand binding affinities, where the solvation free energy of the ligand can have a significant impact on binding affinity prediction.

Although the magnitude of the errors in the solvation free energies for the ion test set were relatively poor for all three methods (see Figure 5), KMTISM showed better correlations

DOI: 10.1021/ct5007828
J. Chem. Theory Comput. 2015, 11, 667−682

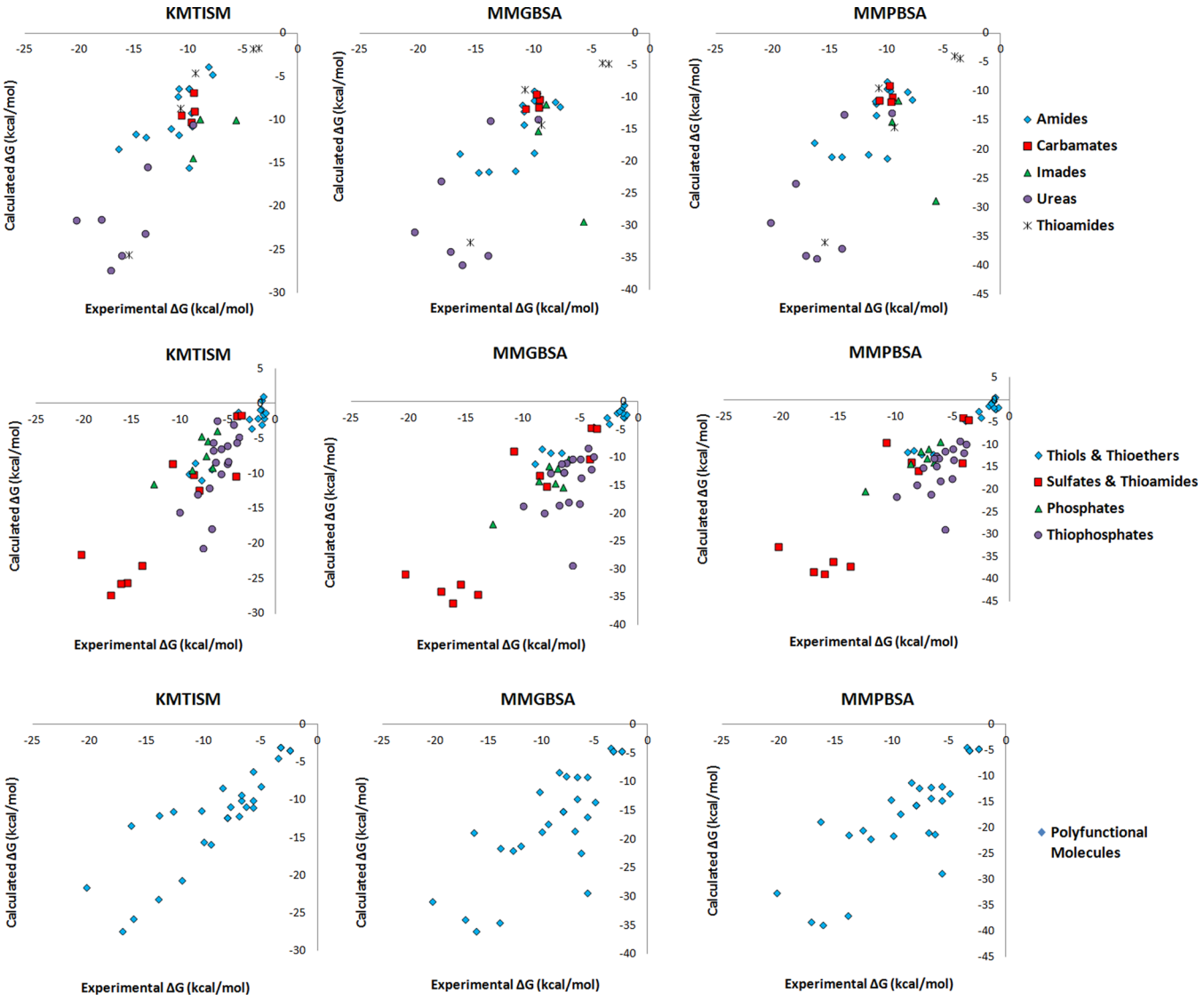


Figure 4. KMTISM's worst three performing test sets according to RMSE. KMTISM, MM-GBSA, and MM-PBSA calculated solvation free energies (kcal/mol) vs experimental data are listed from left to right. From the top to bottom the test sets are the amide, organosulfur and organophosphorus, and polyfunctional test sets.

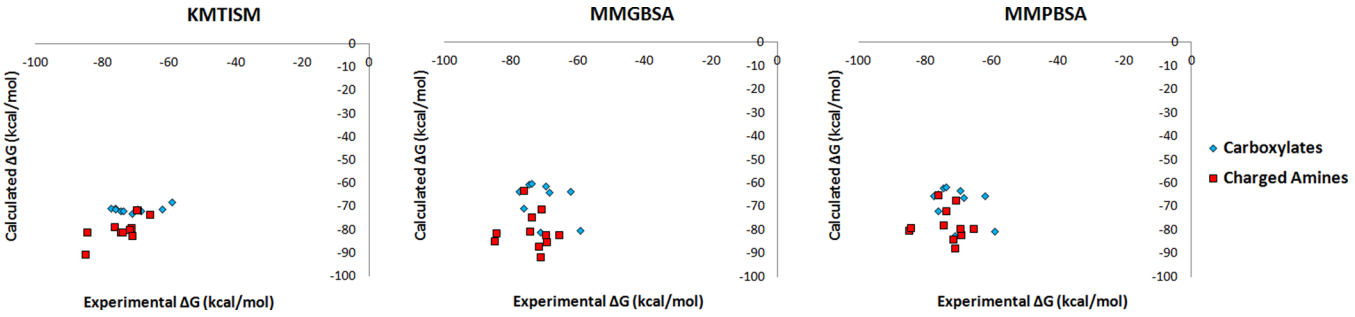


Figure 5. Solvation free energy results for KMTISM, MM-GBSA, and MM-PBSA against the ion test sets.

and RMSE than the other two implicit water models, especially for the charged amine test set (see Table 3). While the error magnitude was large for all methods the percentage error is comparable to the neutral set. The carboxylate functional group, which is conjugated, lowered the accuracy of KMTISM's calculation, while charged amines, on the other hand, whose electron densities are more localized, were better reproduced by the KMTISM method.

Comparison with SMX Results. Overall, QM based solvation models have had limited application to the study of macromolecular systems due to its higher computational expense, but are routinely used to understand the effect of solvation on small molecules. A thorough analysis of KMTISM

DOI: 10.1021/ct5007828
J. Chem. Theory Comput. 2015, 11, 667−682

against QM based solvation models was not the focus of the present research, but a general comparison helps put the present work in perspective relative to more advanced models.

Cramer and Truhlar's Solvation Model (SMX) series has been developed over several decades and is considered to be one of the best methods available to calculate solvation free energies of small molecules.[78−126] Their most up-to-date model reported mean absolute errors (MAE) for solvation free energy (ΔG_{sol}) prediction ranging from 0.57 to 0.84 kcal/mol using different QM methods against 274 neutral molecules and calculated ΔG_{sol} MAEs ranged from 2.7 to 3.8 kcal/mol against 112 ions.[126] Against a similar small molecule and ion test set, KMTISM reproduced a calculated ΔG_{sol} MAE of 1.8 kcal/mol against 372 neutral molecules and a MAE of 5.1 kcal/mol against 21 ions. Even though the data sets examined were different, the trend is clear that the latest SMX models are more accurate than KMTISM (and MM-GBSA and MM-PBSA) by ∼1 kcal/mol for both the neutral molecules and the ions as measured by MAE. Nonetheless, the performance of our first generation KMTISM model is impressive and future versions of this model should yield even better accuracy.

CONCLUSION

MM-GBSA and MM-PBSA are two broadly used implicit solvation models. KMTISM, using a new sampling method (MT method), combined with a statistical energy function (KECSA), is found to have a comparable or a better ability to predict the solvation free energy for several test sets selected from the Minnesota Solvation Database. Though all of these methods perform worse than the most recent SMX model reported by Cramer and Truhlar. It is important to appreciate that without using the approximation that the free energy of solvation is a collection of linearly combined free energies, as is employed in many traditional continuum solvent models, KMTISM uses computed energies to directly determine free energies. Hence, the Helmholtz free energy is calculated by the construction of the relevant partition functions. These partition functions are efficiently assembled using a new sampling method, the MT method, which is able to rapidly estimate free energy, enthalpy, as well as entropy changes. Drawbacks of the KMTISM model lie in its use of a statistical energy function, whose parametrization can have weak spots for atom types with high polarizabilites and uncommon atom types whose lack of available experimental data can pose issues. Future work includes (1) a detailed study of enthalpy changes and entropy changes using the MT method, (2) improving the statistical energy terms by data collection from MD simulations of atom types with high polarizability and uncommon atom types in structural databases, and (3) replacing the statistical energy function with different force field based energy functions and combine them with the MT sampling method in order to affect the rapid evaluation of thermodynamic quantities.

ASSOCIATED CONTENT

Supporting Information

Table S1: KMTISM, MM-GBSA, and MM-PBSA calculated solvation free energy (in kcal/mol) results against the test set with 372 neutral compounds. Table S2: calculated solvation free energy (in kcal/mol) results using these three methods against the test set with 21 charged compounds. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: merzjrke@msu.edu.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank Dave Case for useful discussions of the KECSA model as the null hypothesis for a model including electrostatics. We thank the National Institutes of Health (NIH) (GM066859) for supporting this research in part. We thank the High Performance Computing Center at the University of Florida for computing support.

REFERENCES

(1) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926.
(2) Schaefer, M.; van Vlijmen, H. W.; Karplus, M. Electrostatic contributions to molecular free energies in solution. *Adv. Protein Chem.* **1998**, *51*, 1.
(3) Mobley, D. L.; Liu, S.; Cerutti, D. S.; Swope, W. C.; Rice, J. E. Alchemical prediction of hydration free energies for SAMPL. *J. Comput.-Aided Mol. Des.* **2012**, *26*, 551.
(4) Mobley, D. L.; Dumont, E.; Chodera, J. D.; Dill, K. A. Comparison of charge models for fixed-charge force fields: Small-molecule hydration free energies in explicit solvent. *J. Phys. Chem. B* **2007**, *111*, 2242.
(5) Horn, H. W.; Swope, W. C.; Pitera, J. W.; Madura, J. D.; Dick, T. J.; Hura, G. L.; Head-Gordon, T. Development of an improved four-site water model for biomolecular simulations: TIP4P-Ew. *J. Chem. Phys.* **2004**, *120*, 9665.
(6) García, A. E.; Sanbonmatsu, K. Y. Exploring the energy landscape of a β hairpin in explicit solvent. *Proteins* **2001**, *42*, 345.
(7) Gibas, C. J.; Subramaniam, S. Explicit solvent models in protein pK_a calculations. *Biophys. J.* **1996**, *71*, 138.
(8) Gonçalves, P. F. B.; Stassen, H. Calculation of the free energy of solvation from molecular dynamics simulations. *Pure Appl. Chem.* **2004**, *76*, 231.
(9) Shi, Y.; Wu, C.; Ponder, J. W.; Ren, P. Multipole electrostatics in hydration free energy calculations. *J. Comput. Chem.* **2011**, *32*, 967.
(10) Klimovich, P. V.; Mobley, D. L. Predicting hydration free energies using all-atom molecular dynamics simulations and multiple starting conformations. *J. Comput.—Aided Mol. Des.* **2010**, *24*, 307.
(11) Shivakumar, D.; Williams, J.; Wu, Y.; Damm, W.; Shelley, J.; Sherman, W. Prediction of absolute solvation free energies using molecular dynamics free energy perturbation and the OPLS force field. *J. Chem. Theory Comput.* **2010**, *6*, 1509.
(12) Wang, L.; Voorhis, T. V. A polarizable QM/MM explicit solvent model for computational electrochemistry in water. *J. Chem. Theory Comput.* **2012**, *8*, 610.
(13) Mobley, D. L.; Bayly, C. I.; Cooper, M. D.; Shirts, M. R.; Dill, K. A. Small molecule hydration free energies in explicit solvent: An extensive test of fixed-charge atomistic simulations. *J. Chem. Theory Comput.* **2009**, *5*, 350.
(14) Shirts, M. R.; Pande, V. S. Solvation free energies of amino acid side chain analogs for common molecular mechanics water models. *J. Chem. Phys.* **2005**, *122*, 134508.
(15) Zhou, R.; Berne, B. J.; Germain, R. The free energy landscape for β hairpin folding in explicit water. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 14931.
(16) Scarsi, M.; Apostolakis, J.; Caflisch, A. Comparison of a GB solvation model with explicit solvent simulations: Potentials of mean force and conformational preferences of alanine dipeptide and 1,2-dichloroethane. *J. Phys. Chem. B* **1998**, *102*, 3637.

DOI: 10.1021/ct5007828
J. Chem. Theory Comput. 2015, 11, 667−682

(17) Marrone, T. J.; Gilson, M. K.; McCammon, J. A. Comparison of continuum and explicit models of solvation: Potentials of mean force for alanine dipeptide. *J. Phys. Chem.* **1996**, *100*, 1439.
(18) Shivakumar, D.; Deng, Y.; Roux, B. Computations of absolute solvation free energies of small molecules using explicit and implicit solvent model. *J. Chem. Theory Comput.* **2009**, *5*, 919.
(19) Zhang, L. Y.; Gallicchio, E.; Friesner, R. A.; Levy, R. M. Solvent models for protein−ligand binding: Comparison of implicit solvent poisson and surface generalized Born models with explicit solvent simulations. *J. Comput. Chem.* **2000**, *22*, 591.
(20) Vorobjev, Y. N.; Hermans, J. ES/IS: Estimation of conformational free energy by combining dynamics simulations with explicit solvent with an implicit solvent continuum model. *Biophys. Chem.* **1999**, *78*, 195.
(21) Huißmann, S.; Likos, C. N.; Blaak, R. Explicit vs implicit water simulations of charged dendrimers. *Macromolecules* **2012**, *45*, 2562.
(22) Zhou, R. Free energy landscape of protein folding in water: Explicit vs implicit solvent. *Proteins* **2003**, *53*, 148.
(23) Levy, R. M.; Gallicchio, E. Computer simulations with explicit solvent: Recent progress in the thermodynamic decomposition of free energies and in modeling electrostatic effects. *Annu. Rev. Phys. Chem.* **1998**, *49*, 531.
(24) Nymeyer, H.; García, A. E. Simulation of the folding equilibrium of α-helical peptides: A comparison of the generalized Born approximation with explicit solvent. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 13934.
(25) Nicholls, A.; Mobley, D. L.; Guthrie, J. P.; Chodera, J. D.; Bayly, C. I.; Cooper, M. D.; Pande, V. S. Predicting small-molecule solvation free energies: An informal blind test for computational chemistry. *J. Med. Chem.* **2008**, *51*, 769.
(26) Miertuš, S.; Scrocco, E.; Tomasi, J. Electrostatic interaction of a solute with a continuum. A direct utilization of ab initio molecular potentials for the prevision of solvent effects. *Chem. Phys.* **1981**, *55*, 117.
(27) Miertuš, S.; Tomasi, J. Approximate evaluations of the electrostatic free energy and internal energy changes in solution processes. *Chem. Phys.* **1982**, *65*, 239.
(28) Pascual−Ahuir, J. L.; Silla, E.; Tuñón, I. GEPOL: An improved description of molecular−surfaces. 3. A new algorithm for the computation of a solvent−excluding surface. *J. Comput. Chem.* **1994**, *15*, 1127.
(29) Cossi, M.; Barone, V.; Cammi, R.; Tomasi, J. Ab initio study of solvated molecules: A new implementation of the polarizable continuum model. *Chem. Phys. Lett.* **1996**, *255*, 327.
(30) Barone, V.; Cossi, M.; Tomasi, J. A new definition of cavities for the computation of solvation free energies by the polarizable continuum model. *J. Chem. Phys.* **1997**, *107*, 3210.
(31) Cancès, E.; Mennucci, B.; Tomasi, J. A new integral equation formalism for the polarizable continuum model: Theoretical background and applications to isotropic and anistropic dielectrics. *J. Chem. Phys.* **1997**, *107*, 3032.
(32) Mennucci, B.; Tomasi, J. Continuum solvation models: A new approach to the problem of solute's charge distribution and cavity boundaries. *J. Chem. Phys.* **1997**, *106*, 5151.
(33) Mennucci, B.; Cancès, E.; Tomasi, J. Evaluation of solvent effects in isotropic and anisotropic dielectrics, and in ionic solutions with a unified integral equation method: Theoretical bases, computational implementation, and numerical applications. *J. Phys. Chem. B* **1997**, *101*, 10506.
(34) Barone, V.; Cossi, M. Quantum calculation of molecular energies and energy gradients in solution by a conductor solvent model. *J. Phys. Chem. A* **1998**, *102*, 1995.
(35) Cossi, M.; Barone, V.; Mennucci, B.; Tomasi, J. Ab initio study of ionic solutions by a polarizable continuum dielectric model. *Chem. Phys. Lett.* **1998**, *286*, 253.
(36) Barone, V.; Cossi, M.; Tomasi, J. Geometry optimization of molecular structures in solution by the polarizable continuum model. *J. Comput. Chem.* **1998**, *19*, 404.
(37) Cammi, R.; Mennucci, B.; Tomasi, J. Second-order Møller−Plesset analytical derivatives for the polarizable continuum model using the relaxed density approach. *J. Phys. Chem. A* **1999**, *103*, 9100.
(38) Cossi, M.; Barone, V.; Robb, M. A. A direct procedure for the evaluation of solvent effects in MC-SCF calculations. *J. Chem. Phys.* **1999**, *111*, 5295.
(39) Tomasi, J.; Mennucci, B.; Cancès, E. The IEF version of the PCM solvation method: An overview of a new method addressed to study molecular solutes at the QM ab initio level. *J. Mol. Struct.* **1999**, *464*, 211.
(40) Cammi, R.; Mennucci, B.; Tomasi, J. Fast evaluation of geometries and properties of excited molecules in solution: A Tamm−Dancoff model with application to 4-dimethylaminobenzonitrile. *J. Phys. Chem. A* **2000**, *104*, 5631.
(41) Cossi, M.; Barone, V. Solvent effect on vertical electronic transitions by the polarizable continuum model. *J. Chem. Phys.* **2000**, *112*, 2427.
(42) Cossi, M.; Barone, V. Time−dependent density functional theory for molecules in liquid solutions. *J. Chem. Phys.* **2001**, *115*, 4708.
(43) Cossi, M.; Rega, N.; Scalmani, G.; Barone, V. Polarizable dielectric model of solvation with inclusion of charge penetration effects. *J. Chem. Phys.* **2001**, *114*, 5691.
(44) Cossi, M.; Scalmani, G.; Rega, N.; Barone, V. New developments in the polarizable continuum model for quantum mechanical and classical calculations on molecules in solution. *J. Chem. Phys.* **2002**, *117*, 43.
(45) Cossi, M.; Rega, N.; Scalmani, G.; Barone, V. Energies, structures, and electronic properties of molecules in solution with the C-PCM solvation model. *J. Comput. Chem.* **2003**, *24*, 669.
(46) Tomasi, J.; Mennucci, B.; Cammi, R. Quantum mechanical continuum solvation models. *Chem. Rev.* **2005**, *105*, 2999.
(47) Chipman, D. M. Reaction field treatment of charge penetration. *J. Chem. Phys.* **2000**, *112*, 5558.
(48) Cancès, E.; Mennucci, B. Comment on 'Reaction field treatment of charge penetration'. *J. Chem. Phys.* **2001**, *114*, 4744.
(49) Foresman, J. B.; Keith, T. A.; Wiberg, K. B.; Snoonian, J.; Frisch, M. J. Solvent effects. 5. The influence of cavity shape, truncation of electrostatics, and electron correlation on ab initio reaction field calculations. *J. Phys. Chem.* **1996**, *100*, 16098.
(50) Kirkwood, J. G. Theory of solutions of molecules containing widely separated charges with special application to zwitterions. *J. Chem. Phys.* **1934**, *2*, 351.
(51) Onsager, L. Electric moments of molecules in liquids. *J. Am. Chem. Soc.* **1936**, *58*, 1486.
(52) Wong, M. W.; Frisch, M. J.; Wiberg, K. B. Solvent effects 1. The mediation of electrostatic effects by solvents. *J. Am. Chem. Soc.* **1991**, *113*, 4776.
(53) Wong, M. W.; Wiberg, K. B.; Frisch, M. J. Hartree−Fock second derivatives and electric field properties in a solvent reaction field—Theory and application. *J. Chem. Phys.* **1991**, *95*, 8991.
(54) Wong, M. W.; Wiberg, K. B.; Frisch, M. J. Solvent effects 2. Medium effect on the structure, energy, charge density, and vibrational frequencies of sulfamic acid. *J. Am. Chem. Soc.* **1992**, *114*, 523.
(55) Wong, M. W.; Wiberg, K. B.; Frisch, M. J. Solvent effects 3. Tautomeric equilibria of formamide and 2-pyridone in the gas phase and solution: An ab initio SCRF study. *J. Am. Chem. Soc.* **1992**, *114*, 1645.
(56) Mobley, D. L.; Ii, A. E.; Fennell, C. J.; Dill, K. A. Charge asymmetries in hydration of polar solutes. *J. Phys. Chem. B* **2008**, *112*, 2405.
(57) Su, Y.; Gallicchio, E. The non-polar solvent potential of mean force for the dimerization of alanine dipeptide: The role of solute-solvent van der Waals interactions. *Biophys. Chem.* **2004**, *109*, 251.
(58) Sitkoff, D.; Sharp, K. A.; Honig, B. Accurate calculation of hydration free energies using macroscopic solvent models. *J. Phys. Chem.* **1994**, *98*, 1978.
(59) Honig, B.; Nicholls, A. Classical electrostatics in biology and chemistry. *Science* **1995**, *268*, 1144.

DOI: 10.1021/ct5007828
J. Chem. Theory Comput. 2015, 11, 667−682

(60) Beroza, P.; Case, D. A. Calculation of proton binding thermodynamics in proteins. *Methods Enzymol.* **1998**, *295*, 170.
(61) Madura, J. D.; Davis, M. E.; Gilson, M. K.; Wade, R. C.; Luty, B. A.; McCammon, J. A. Biological applications of electrostatic calculations and Brownian dynamics. *Rev. Comput. Chem.* **1994**, *5*, 229.
(62) Gilson, M. K. Theory of electrostatic interactions in macromolecules. *Curr. Opin. Struct. Biol.* **1995**, *5*, 216.
(63) Scarsi, M.; Apostolakis, J.; Caflisch, A. Continuum electrostatic energies of macromolecules in aqueous solutions. *J. Phys. Chem. A* **1997**, *101*, 8098.
(64) Pei, J.; Wang, Q.; Zhou, J.; Lai, L. Estimating protein−ligand binding free energy: Atomic solvation parameters for partition coefficient and solvation free energy calculation. *Proteins* **2004**, *57*, 651.
(65) Zhou, R.; Berne, B. J. Can a continuum solvent model reproduce the free energy landscape of a β-hairpin folding in water? *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 12777.
(66) Tomasi, J.; Persico, M. Molecular interactions in solution: An overview of methods based on continuous distributions of the solvent. *Chem. Rev.* **1994**, *94*, 2027.
(67) Cossi, M.; Rega, N.; Scalmani, G.; Barone, V. Energies, structures, and electronic properties of molecules in solution with the C-PCM solvation model. *J. Comput. Chem.* **2003**, *24*, 669.
(68) Rizzo, R. C.; Aynechi, T.; Case, D. A.; Kuntz, I. D. Estimation of absolute free energies of hydration using continuum methods: Accuracy of partial charge models and optimization of nonpolar contributions. *J. Chem. Theory Comput.* **2006**, *2*, 128.
(69) Felts, A. K.; Harano, Y.; Gallicchio, E.; Levy, R. M. Free energy surfaces of β-hairpin and α-helical peptides generated by replica exchange molecular dynamics with the AGBNP implicit solvent model. *Proteins* **2004**, *56*, 310.
(70) Roux, B.; Simonson, T. Implicit solvent models. *Biophys. Chem.* **1999**, *78*, 1.
(71) Kolář, M.; Fanfrlík, J.; Hobza, P. Ligand conformational and solvation/desolvation free energy in protein−ligand complex formation. *J. Phys. Chem. B* **2011**, *115*, 4718.
(72) Shoichet, B. K.; Leach, A. R.; Kuntz, I. D. Ligand solvation in molecular docking. *Proteins* **1999**, *34*, 4.
(73) Rees, D. C.; Wolfe, G. M. Macromolecular solvation energies derived from small molecule crystal morphology. *Protein Sci.* **1993**, *2*, 1882.
(74) Eisenberg, D.; McLachlan, A. D. Solvation energy in protein folding and binding. *Nature* **1986**, *319*, 199.
(75) Lazaridis, T.; Karplus, M. Effective energy function for proteins in solution. *Proteins* **1999**, *35*, 133.
(76) Wesson, L.; Eisenberg, D. Atomic solvation parameters applied to molecular dynamics of proteins in solution. *Protein Sci.* **1992**, *1*, 227.
(77) Klamt, A.; Schüürmann, G. COSMO: A new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient. *J. Chem. Soc., Perkin Trans. 2.* **1993**, *5*, 799.
(78) Cramer, C. J.; Truhlar, D. G. General parameterized SCF model for free energies of solvation in aqueous solution. *J. Am. Chem. Soc.* **1991**, *113*, 8305.
(79) Cramer, C. J.; Truhlar, D. G. Molecular orbital theory calculations of aqueous solvation effects on chemical equilibria. *J. Am. Chem. Soc.* **1991**, *113*, 9901.
(80) Cramer, C. J.; Truhlar, D. G. An SCF solvation model for the hydrophobic effect and absolute free energies of aqueous solvation. *Science* **1992**, *256*, 213.
(81) Cramer, C. J.; Truhlar, D. G. PM3-SM3: A general parameterization for including aqueous solvation effects in the PM3Molecular orbital model. *J. Comput. Chem.* **1992**, *13*, 1089.
(82) Cramer, C. J.; Truhlar, D. G. AM1-SM2 and PM3-SM3 parameterized SCF solvation models for free energies in aqueous solution. *J. Comput.—Aided Mol. Des.* **1992**, *6*, 629.
(83) Cramer, C. J.; Truhlar, D. G. Polarization of the nucleic acid bases in aqueous solution. *Chem. Phys. Lett.* **1992**, *198*, 74.
(84) Cramer, C. J.; Truhlar, D. G. Quantum chemical conformational analysis of glucose in aqueous solution. *J. Am. Chem. Soc.* **1993**, *115*, 5745.
(85) Cramer, C. J.; Truhlar, D. G. Correlation and solvation effects on heterocyclic equilibria in aqueous solution. *J. Am. Chem. Soc.* **1993**, *115*, 8810.
(86) Giesen, D. J.; Cramer, C. J.; Truhlar, D. G. Entropic contributions to free energies of solvation. *J. Phys. Chem.* **1994**, *98*, 4141.
(87) Cramer, C. J.; Truhlar, D. G. Quantum chemical conformational analysis of 1,2-ethanediol: Correlation and solvation effects on the tendency to form internal hydrogen bonds in the gas phase and aqueous solution. *J. Am. Chem. Soc.* **1994**, *116*, 3892.
(88) Liotard, D. A.; Hawkins, G. D.; Lynch, G. C.; Cramer, C. J.; Truhlar, D. G. Improved methods for semiempirical solvation models. *J. Comput. Chem.* **1995**, *16*, 422.
(89) Giesen, D. J.; Storer, J. W.; Cramer, C. J.; Truhlar, D. G. A general semiempirical quantum mechanical solvation model for nonpolar solvation energies. *n*-Hexadecane. *J. Am. Chem. Soc.* **1995**, *117*, 1057.
(90) Giesen, D. J.; Cramer, C. J.; Truhlar, D. G. A semiempirical quantum mechanical solvation model for solvation free energies in all alkane solvents. *J. Phys. Chem.* **1995**, *99*, 7137.
(91) Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. Pairwise solute screening of solute charges from a dielectric medium. *Chem. Phys. Lett.* **1995**, *246*, 122.
(92) Barrows, S. E.; Cramer, C. J.; Truhlar, D. G.; Elovitz, M. S.; Weber, E. J. Factors controlling regioselectivity in the reduction of polynitroaromatics in aqueous solution. *Environ. Sci. Technol.* **1996**, *30*, 3028.
(93) Chambers, C. C.; Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. Model for aqueous solvation based on class IV atomic charges and first-solvation shell effects. *J. Phys. Chem.* **1996**, *100*, 16385.
(94) Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. Parameterized models of aqueous free energies of solvation based on pairwise descreening of solute atomic charges from a dielectric medium. *J. Phys. Chem.* **1996**, *100*, 19824.
(95) Cramer, C. J.; Truhlar, D. G.; French, A. D. Exo-anomeric effects on energies and geometries of different conformations of glucose and related systems in the gas phase and aqueous solution. *Catal. Rev.* **1997**, *298*, 1.
(96) Giesen, D. J.; Chambers, C. C.; Cramer, C. J.; Truhlar, D. G. A solvation model for chloroform based on class IV atomic charges. *J. Phys. Chem. B* **1997**, *101*, 2061.
(97) Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. New methods for potential functions for simulating biological molecules. *J. Chim. Phys.* **1997**, *94*, 1448.
(98) Giesen, D. J.; Chambers, C. C.; Cramer, C. J.; Truhlar, D. G. What controls the partitioning of nucleic acid bases between chloroform and water? *J. Phys. Chem. B* **1997**, *101*, 5084.
(99) Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. Parameterized model for aqueous free energies of solvation using geometry-dependent atomic surface tensions with implicit electrostatics. *J. Phys. Chem. B* **1997**, *101*, 7147.
(100) Cramer, C. J.; Truhlar, D. G.; Falvey, D. E. Singlet-triplet splittings and 1,2-hydrogen shift barriers for methylphenylborenide, methylphenylcarbene, and methylphenylnitrenium in the gas phase and solution. What a difference a charge makes. *J. Am. Chem. Soc.* **1997**, *119*, 12338.
(101) Giesen, D. J.; Hawkins, G. D.; Liotard, D. A.; Cramer, C. J.; Truhlar, D. G. A universal solvation model for the quantum mechanical calculation of free energies of solvation in non-aqueous solvents. *Theor. Chem. Acc.* **1997**, *98*, 85.
(102) Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. Universal quantum mechanical model for solvation free energies based on gas-phase geometries. *J. Phys. Chem. B* **1998**, *102*, 3257.
(103) Hawkins, G. D.; Liotard, D. A.; Cramer, C. J.; Truhlar, D. G. OMNISOL: Fast prediction of free energies of solvation and partition coefficients. *J. Org. Chem.* **1998**, *63*, 4305.

DOI: 10.1021/ct5007828
J. Chem. Theory Comput. 2015, 11, 667−682

(104) Zhu, T.; Li, J.; Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. Density functional solvation model based on CM2 atomic charges. *J. Chem. Phys.* **1998**, *109*, 9117.
(105) Li, J.; Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. Universal reaction field model based on ab initio Hartree−Fock theory. *Chem. Phys. Lett.* **1998**, *288*, 293.
(106) Sullivan, M. B.; Brown, K.; Cramer, C. J.; Truhlar, D. G. Quantum chemical analysis of para-substitution effects on the electronic structure of phenylnitrenium ions in the gas phase and aqueous solution. *J. Am. Chem. Soc.* **1998**, *120*, 11778.
(107) Zhu, T.; Li, J.; Liotard, D. A.; Cramer, C. J.; Truhlar, D. G. Analytical energy gradients of a self-consistent reaction-field solvation model based on CM2 charges. *J. Chem. Phys.* **1999**, *110*, 5503.
(108) Li, J.; Cramer, C. J.; Truhlar, D. G. Application of a universal solvation model to nucleic acid bases. Comparison of semiempirical molecular orbital theory, ab initio Hartree−Fock theory, and density functional theory. *Biophys. Chem.* **1999**, *103*, 3802.
(109) Li, J.; Zhu, T.; Hawkins, G. D.; Winget, P.; Liotard, D. A.; Cramer, C. J.; Truhlar, D. G. Extension of the platform of applicability of the SM5.42R universal solvation model. *Theor. Chem. Acc.* **1999**, *103*, 9.
(110) Cramer, C. J.; Truhlar, D. G. Implicit solvation models: Equilibria, structure, spectra, and dynamics. *Chem. Rev.* **1999**, *99*, 2161.
(111) Li, J.; Zhu, T.; Cramer, C. J.; Truhlar, D. G. A universal solvation model based on class IV charges and the intermediate neglect of differential overlap for spectroscopy molecular orbital method. *J. Phys. Chem. A* **2000**, *104*, 2178.
(112) Dolney, D. M.; Hawkins, G. D.; Winget, P.; Liotard, D. A.; Cramer, C. J.; Truhlar, D. G. A universal solvation model based on the conductor-like screening model. *J. Comput. Chem.* **2000**, *21*, 340.
(113) Winget, P.; Weber, E. J.; Cramer, C. J.; Truhlar, D. G. Computational electrochemistry: Aqueous one-electron oxidation potentials for substituted anilines. *Phys. Chem. Chem. Phys.* **2000**, *2*, 1231.
(114) Winget, P.; Cramer, C. J.; Truhlar, D. G. Parameterization of a universal solvation model for molecules containing silicon. *J. Phys. Chem. A* **2002**, *106*, 5160.
(115) Thompson, J. D.; Cramer, C. J.; Truhlar, D. G. New universal solvation model and comparison of the accuracy of the SM5.42R, SM5.43R, C-PCM, D-PCM, and IEF-PCM continuum solvation models for aqueous and organic solvation free energies and for vapor pressures. *J. Phys. Chem. A* **2004**, *108*, 6532.
(116) Thompson, J. D.; Cramer, C. J.; Truhlar, D. G. Density-functional theory and hybrid density-functional theory continuum solvation models for aqueous and organic solvents: Universal SM5.43 and SM5.43R solvation models for any fraction of Hartree−Fock exchange. *Theor. Chem. Acc.* **2005**, *113*, 107.
(117) Kelly, C. P.; Cramer, C. J.; Truhlar, D. G. SM6: A density functional theory continuum solvation model for calculating aqueous solvation free energies of neutrals, ions, and solute−water clusters. *J. Chem. Theory Comput.* **2005**, *1*, 1133.
(118) Chamberlin, A. C.; Cramer, C. J.; Truhlar, D. G. Predicting aqueous free energies of solvation as functions of temperature. *J. Phys. Chem. B* **2006**, *110*, 5665.
(119) Kelly, C. P.; Cramer, C. J.; Truhlar, D. G. Adding explicit solvent molecules to continuum solvent calculations for the calculation of aqueous acid dissociation constants. *J. Phys. Chem. A* **2006**, *110*, 2493.
(120) Marenich, A. V.; Olson, R. M.; Kelly, C. P.; Cramer, C. J.; Truhlar, D. G. Self-consistent reaction field model for aqueous and nonaqueous solutions based on accurate polarized partial charges. *J. Chem. Theory Comput.* **2007**, *3*, 2011.
(121) Chamberlin, A. C.; Cramer, C. J.; Truhlar, D. G. Extension of a temperature-dependent aqueous solvation model to compounds containing nitrogen, fluorine, chlorine, bromine, and sulfur. *J. Phys. Chem. B* **2008**, *112*, 3024.
(122) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. Universal solvation model based on solute electron density and on a continuum model of the solvent defined by the bulk dielectric constant and atomic surface tensions. *J. Phys. Chem. B* **2009**, *113*, 6378.
(123) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. Universal solvation model based on the generalized Born approximation with asymmetric descreening. *J. Chem. Theory Comput.* **2009**, *5*, 2447.
(124) Liu, J.; Kelly, C. P.; Goren, A. C.; Marenich, A. V.; Cramer, C. J.; Truhlar, D. G.; Zhan, C.-G. Free energies of solvation with surface, volume, and local electrostatic effects and atomic surface tensions to represent the first solvation shell. *J. Chem. Theory Comput.* **2010**, *6*, 1109.
(125) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. Sorting out the relative contributions of electrostatic polarization, dispersion, and hydrogen bonding to solvatochromic shifts on vertical electronic excitation energies. *J. Chem. Theory Comput.* **2010**, *6*, 2829.
(126) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. Generalized Born solvation model SM12. *J. Chem. Theory Comput.* **2013**, *9*, 609.
(127) Holst, M.; Baker, N.; Wang, F. Adaptive multilevel finite element solution of the Poisson−Boltzmann equation I. Algorithms and examples. *J. Comput. Chem.* **2000**, *21*, 1319.
(128) Baker, N.; Holst, M.; Wang, F. Adaptive multilevel finite element solution of the Poisson−Boltzmann equation II. Refinement at solvent-accessible surfaces in biomolecular systems. *J. Comput. Chem.* **2000**, *21*, 1343.
(129) Lu, B. Z.; Zhou, Y. C.; Holst, M. J.; McCammon, J. A. Recent progress in numerical methods for the Poisson−Boltzmann equation in biophysical applications. *Commun. Comput. Phys.* **2008**, *3*, 973.
(130) Nicholls, A.; Honig, B. A rapid finite difference algorithm, utilizing successive over-relaxation to solve the Poisson−Boltzmann equation. *J. Comput. Chem.* **1990**, *12*, 435.
(131) Baker, N. A. Improving implicit solvent simulations: A Poisson-centric view. *Curr. Opin. Struct. Biol.* **2005**, *15*, 137.
(132) Fogolari, F.; Brigo, A.; Molinari, H. The Poisson−Boltzmann equation for biomolecular electrostatics: A tool for structural biology. *J. Mol. Recognit.* **2002**, *15*, 377.
(133) Baker, N. A.; Sept, D.; Joseph, S.; Holst, M. J.; McCammon, J. A. Electrostatics of nanosystems: Application to microtubules and the ribosome. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 10037.
(134) Luo, R.; David, L.; Gilson, M. K. Accelerated Poisson−Boltzmann calculations for static and dynamic systems. *J. Comput. Chem.* **2002**, *23*, 1244.
(135) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.* **1990**, *112*, 6127.
(136) Schaefer, M.; Froemmel, C. A precise analytical method for calculating the electrostatic energy of macromolecules in aqueous solution. *J. Mol. Biol.* **1990**, *216*, 1045.
(137) Srinivasan, J.; Trevathan, M. W.; Beroza, P.; Case, D. A. Application of a pairwise generalized Born model to proteins and nucleic acids: Inclusion of salt effects. *Theor. Chem. Acc.* **1999**, *101*, 426.
(138) Sagui, C.; Darden, T. A. Molecular dynamics simulations of biomolecules: Long-range electrostatic effects. *Annu. Rev. Biophys. Biomol. Struct.* **1999**, *28*, 155.
(139) Massova, I.; Kollman, P. A. Combined molecular mechanical and continuum solvent approach (MM-PBSA/GBSA) to predict ligand binding. *Perspect. Drug Discovery Des.* **2000**, *18*, 113.
(140) Tsui, V.; Case, D. A. Theory and applications of the generalized born solvation model in macromolecular simulations. *Biopolymers* **2001**, *56*, 275.
(141) Bashford, D.; Case, D. A. Generalized Born models of macromolecular solvation effects. *Annu. Rev. Phys. Chem.* **2000**, *51*, 129.
(142) Zou, X.; Sun, Y.; Kuntz, I. D. Inclusion of solvation in ligand binding free energy calculations using the generalized Born model. *J. Am. Chem. Soc.* **1999**, *121*, 8033.
(143) Onufriev, A.; Bashford, D.; Case, D. A. Modification of the generalized Born model suitable for macromolecules. *J. Phys. Chem. B* **2000**, *104*, 3712.

DOI: 10.1021/ct5007828
J. Chem. Theory Comput. 2015, 11, 667−682

(144) Onufriev, A.; Bashford, D.; Case, D. A. Exploring protein native states and large-scale conformational changes with a modified generalized Born model. *Proteins* **2004**, *55*, 383.
(145) Dzubiella, J.; Swanson, J. M.; McCammon, J. A. Coupling nonpolar and polar solvation free energies in implicit solvent models. *J. Chem. Phys.* **2006**, *124*, 084905.
(146) Hou, T.; Wang, J.; Li, Y.; Wang, W. Assessing the performance of the MM/PBSA and MM/GBSA methods. 1. The accuracy of binding free energy calculations based on molecular dynamics simulations. *J. Chem. Inf. Model.* **2011**, *51*, 69.
(147) Hou, T.; Wang, J.; Li, Y.; Wang, W. Assessing the performance of the MM/PBSA and MM/GBSA methods: II. The accuracy of ranking poses generated from docking. *J. Comput. Chem.* **2011**, *32*, 866.
(148) Bello, M. Binding free energy calculations between bovine β-lactoglobulin and four fatty acids using the MMGBSA method. *Biopolymers* **2014**, *101*, 1010.
(149) Sippl, M. J. Calculation of conformational ensembles from potentials of mean force. *J. Mol. Biol.* **1990**, *213*, 859.
(150) Miyazawa, S.; Jernigan, R. L. Estimation of effective interresidue contact energies from protein crystal structures: Quasi-chemical approximation. *Macromolecules* **1985**, *18*, 534.
(151) Hendlich, M.; Lackner, P.; Weitckus, S.; Floeckner, H.; Froschauer, R.; Gottsbacher, K.; Casari, G.; Sippl, M. J. Identification of native protein folds amongst a large number of incorrect models. The calculation of low energy conformations from potentials of mean force. *J. Mol. Biol.* **1990**, *216*, 167.
(152) Jones, D. T.; Taylor, W. R.; Thornton, J. M. A new approach to protein fold recognition. *Nature* **1992**, *358*, 86.
(153) Thomas, P. D.; Dill, K. A. An iterative method for extracting energy-like quantities from protein structures. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 11628.
(154) Thomas, P. D.; Dill, K. A. Statistical potentials extracted from protein structures: How accurate are they? *J. Mol. Biol.* **1996**, *257*, 457.
(155) Lu, H.; Skolnick, J. A distance-dependent atomic knowledge-based potential for improved protein structure selection. *Proteins: Struct. Funct. Genet.* **2001**, *44*, 223.
(156) Muegge, I.; Martin, Y. C. A general and fast scoring function for protein−ligand interactions: A simplified potential approach. *J. Med. Chem.* **1999**, *42*, 791.
(157) Muegge, I. A knowledge-based scoring function for protein−ligand interactions: Probing the reference state. *Perspect. Drug Discovery Des.* **2000**, *20*, 99.
(158) Muegge, I. Effect of ligand volume correction on PMF scoring. *J. Comput. Chem.* **2001**, *22*, 418.
(159) Gohlke, H.; Hendlich, M.; Klebe, G. Knowledge-based scoring function to predict protein−ligand interactions. *J. Mol. Biol.* **2000**, *295*, 337.
(160) Velec, H. F. G.; Gohlke, H.; Klebe, G. DrugScore(CSD)—Knowledge-based scoring function derived from small molecule crystal data with superior recognition rate of near-native ligand poses and better affinity prediction. *J. Med. Chem.* **2005**, *48* (20), 6296.
(161) DeWitte, R. S.; Shakhnovich, E. I. SMoG: de Novo design method based on simple, fast, and accurate free energy estimate. 1. Methodology and supporting evidence. *J. Am. Chem. Soc.* **1996**, *118*, 11733.
(162) Ishchenko, A. V.; Shakhnovich, E. I. Small molecule growth 2001 (SMoG2001): An improved knowledge-based scoring function for protein−ligand interactions. *J. Med. Chem.* **2002**, *45*, 2770.
(163) Mitchell, J. B. O.; Laskowski, R. A.; Alex, A.; Thornton, J. M. BLEEP—Potential of mean force describing protein−ligand interactions: I. Generating potential. *J. Comput. Chem.* **1999**, *20*, 1165.
(164) Mitchell, J. B. O.; Laskowski, R. A.; Alex, A.; Forster, M. J.; Thornton, J. M. BLEEP—Potential of mean force describing protein−ligand interactions: II. Calculation of binding energies and comparison with experimental data. *J. Comput. Chem.* **1999**, *20* (11), 1177.
(165) Huang, S.-Y.; Zou, X. An iterative knowledge-based scoring function to predict protein−ligand interactions: II. Validation of the scoring function. *J. Comput. Chem.* **2006**, *27*, 1876.
(166) Huang, S.-Y.; Zou, X. Inclusion of solvation and entropy in the knowledge-based scoring function for protein−ligand interactions. *J. Chem. Inf. Model.* **2010**, *50*, 262.
(167) Kirkwood, J. G. Statistical mechanics of fluid mixtures. *J. Chem. Phys.* **1935**, *3*, 300.
(168) Huang, S.; Grinter, S. Z.; Zou, X. Scoring functions and their evaluation methods for protein−ligand docking: Recent advances and future directions. *Phys. Chem. Chem. Phys.* **2010**, *12*, 12899.
(169) Huang, S. Y.; Zou, X. Mean-force scoring functions for protein−ligand binding. *Annu. Rep. Comput. Chem.* **2010**, *6*, 280.
(170) Sippl, M. J. Knowledge-based potentials for proteins. *Curr. Opin. Struct. Biol.* **1995**, *5*, 229.
(171) Samudrala, R.; Moult, J. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *J. Mol. Biol.* **1998**, *275*, 895.
(172) Zhang, J.; Zhang, Y. A novel side-chain orientation dependent potential derived from random-walk reference state for protein fold selection and structure prediction. *PloS One* **2010**, *5*, e15386.
(173) Xu, D.; Zhang, J.; Roy, A.; Zhang, Y. Automated protein structure modeling in CASP9 by I-TASSER pipeline combined with QUARK-based ab initio folding and FG-MD-based structure refinement. *Proteins* **2011**, *79*, 147.
(174) Xu, D.; Zhang, Y. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins* **2012**, *80*, 1715.
(175) Zhou, H.; Zhou, Y. Single-body residue-level knowledge-based energy score combined with sequence-profile and secondary structure information for fold recognition. *Proteins* **2004**, *55*, 1005.
(176) Melo, F.; Sánchez, R.; Sali, A. Statistical potentials for fold assessment. *Protein Sci.* **2002**, *11*, 430.
(177) Benkert, P.; Biasini, M.; Schwede, T. Toward the estimation of the absolute quality of individual protein structure models. *Bioinformatics* **2011**, *27*, 343.
(178) Skolnick, J.; Jaroszewski, L.; Kolinski, A.; Godzik, A. Derivation and testing of pair potentials for protein folding. When is the quasichemical approximation correct? *Protein Sci.* **1997**, *6*, 676.
(179) Krishnamoorthy, B.; Tropsha, A. Development of a four-body statistical pseudopotential to discriminate native from non-native protein conformations. *Bioinformatics* **2003**, *19*, 1540.
(180) Frishman, D.; Argos, P. Incorporation of nonlocal interactions in protein secondary structure prediction from the amino acid sequence. *Protein Eng.* **1996**, *9*, 133.
(181) Lu, H.; Skolnick, J. A distance-dependent atomic knowledge-based potential for improved protein structure selection. *Proteins* **2001**, *44*, 223.
(182) Jernigan, R. L.; Bahar, I. Structure-derived potentials and protein simulations. *Curr. Opin. Struct. Biol.* **1996**, *6*, 195.
(183) Shen, M.; Sali, A. Statistical potential for assessment and prediction of protein structures. *Protein Sci.* **2006**, *15*, 2507.
(184) Thomas, P. D.; Dill, K. A. Statistical potentials extracted from protein structures: How accurate are they? *J. Mol. Biol.* **1996**, *257*, 457.
(185) Liu, S.; Zhang, C.; Zhou, H.; Zhou, Y. A physical reference state unifies the structure-derived potential of mean force for protein folding and binding. *Proteins* **2004**, *56*, 93.
(186) Tobi, D.; Elber, R. Distance−dependent, pair potential for protein folding: Results from linear optimization. *Proteins* **2000**, *41*, 40.
(187) Zheng, Z.; Ucisik, M. N.; Merz, K. M., Jr. The movable type method applied to protein−ligand binding. *J. Chem. Theory Comput.* **2013**, *9*, 5526.
(188) Allen, F. H. The Cambridge Structural Database: A quarter of a million crystal structures and rising. *Acta Crystallogr.* **2002**, *B58*, 380.
(189) Zheng, Z.; Merz, K. M., Jr. Development of the knowledge-based and empirical combined scoring algorithm (KECSA) to score protein−ligand interactions. *J. Chem. Inf. Model.* **2013**, *53*, 1073.
(190) Cheng, T.; Li, X.; Li, Y.; Liu, Z.; Wang, R. Comparative assessment of scoring functions on a diverse test set. *J. Chem. Inf. Model.* **2009**, *49*, 1079.
(191) Wang, R.; Fang, X.; Lu, Y.; Yang, C. Y.; Wang, S. The PDBbind database: Methodologies and updates. *J. Med. Chem.* **2005**, *48*, 4111.

DOI: 10.1021/ct5007828
J. Chem. Theory Comput. 2015, 11, 667−682

(192) Wang, R.; Fang, X.; Lu, Y.; Wang, S. The PDBbind database: Collection of binding affinities for protein−ligand complexes with known three-dimensional structures. *J. Med. Chem.* **2004**, *47*, 2977.
(193) Marenich, A. V.; Kelly, C. P.; Thompson, J. D.; Hawkins, G. D.; Chambers, C. C.; Giesen, D. J.; Winget, P.; Cramer, C. J.; Truhlar, D. G. *Minnesota Solvation Database*, version 2012; University of Minnesota: Minneapolis, 2012.
(194) Mark, A. E.; van Gunsteren, W. F. Decomposition of the free energy of a system in terms of specific interactions. Implications for theoretical and experimental studies. *J. Mol. Biol.* **1994**, *240*, 167.
(195) Bren, M.; Florián, J.; Mavri, J.; Bren, U. Do all pieces make a whole? Thiele cumulants and the free energy decomposition. *Theor. Chem. Acc.* **2007**, *117*, 535.
(196) Baum, B.; Muley, L.; Smolinski, M.; Heine, A.; Hangauer, D.; Klebe, G. Non-additivity of functional group contributions in protein−ligand binding: A comprehensive study by crystallography and isothermal titration calorimetry. *J. Mol. Biol.* **2010**, *397*, 1042.
(197) Guillot, B. A reappraisal of what we have learnt during three decades of computer simulations on water. *J. Mol. Liq.* **2002**, *101*, 219.
(198) Narten, A. H.; Danford, M. D.; Levy, H. A. X-ray diffraction study of liquid water in the temperature range 4−200 °C. *Faraday Discuss.* **1967**, *43*, 97.

DOI: 10.1021/ct5007828
J. Chem. Theory Comput. 2015, 11, 667−682