THE JOURNAL OF PHYSICAL CHEMISTRY B

Article

pubs.acs.org/JPCB

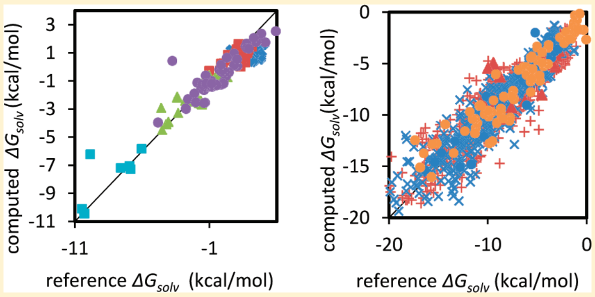
Fast Estimation of Solvation Free Energies for Diverse Chemical Species

Robert D. Boyer and Richard L. Bryan*

BioLeap, Inc., 238 West Delaware Avenue, Pennington, New Jersey 08534, United States

Ⓢ *Supporting Information*

ABSTRACT: The free energy of solvation can play an important or even dominant role in the accurate prediction of binding affinities and various other molecular-scale interaction phenomena critical to the study of biochemical processes. Many research applications for solvation modeling, such as fragment-based drug design, require algorithms that are both accurate and computationally inexpensive. We have developed a calculation of solvation free energy which runs fast enough for interactive applications, functions for a wide range of chemical species relevant to simulating molecules for biological and pharmaceutical applications, and is readily extended when data for new species becomes available. We have also demonstrated that the incorporation of ab initio data provides necessary access to sufficient reference data for a broad range of chemical features. Our empirical model, including an electrostatic term and a different set of atom types, demonstrates improvements over a previous, solvent-accessible surface area-only model by Wang et al.[1] when fit to identical training sets (mean absolute error of 0.513 kcal/mol versus the 0.538 kcal/mol reported by Wang). The incorporation of ab initio solvation free energies provides a significant increase in the breadth of chemical features for which the model can be applied by introducing classes of compounds for which little or no experimental data is available. The increased breadth and the speed of this solvation model allow for conformational minimization, conformational search, and ligand binding free energy calculations that economically account for the complex interplay of bonded, nonbonded, and solvation free energies as conformations with varying solvent-accessible surfaces are sampled.



INTRODUCTION

Solvation free energy (ΔG_{solv}) can play an important role in biological phenomena such as protein folding, and particularly in the presence of hydrophobic structures, it can significantly affect the ranking of binding affinity necessary for fragment-based drug design.[2−5] The computational demands of explicit incorporation of water molecules into atomic-scale simulation have fueled a variety of research into implicit solvation models.[6−13] However, functional descriptions of implicit solvation free energy remain an area of active research because the competing needs for computationally efficient and predicatively accurate models have not been sufficiently met.[14−18] Computing solvation effects in atomic-scale simulation at the protein−ligand interface, where bulk continuum models are inapplicable, has traditionally been done either by complicated, slow, external programs of acceptable accuracy, requiring specialized training and/or experience to operate;[19,20] or by a quick method[1,21] suffering from limited accuracy and not fully parametrized for chemistries of interest. Tabulating precomputed values is inadequate: when a ligand meets a protein, each is only partially desolvated; and one cannot precompute a dynamic system with constantly changing solvent exclusion. The need for computational efficiency comes both from the desire for solvation free energy in the context of a protein, and from the need for the integration of solvation models into molecular dynamics or large-scale screening simulations.

The standard theory of solvation requires consideration of electrostatic interactions, additional solute/solvent interface interactions and a cavitation or solvent displacement term for the free energy. Many solvation models incorporate precise but computationally expensive calculations of point charges interacting with the surrounding water. Formalisms such as the Poisson−Boltzmann equation[11,22] or its approximations such as the generalized Born[23] model can be leveraged to capture the electrostatic component of solvation free energy. Additional contributions, supported with much less satisfactory theory, are sometimes included via a linear set of fit parameters. Even though electrostatic effects can play a dominant role in calculation of the solvation free energy, the accuracy of the electrostatic term becomes less relevant when added to these low-resolution solutions to the calculation of the other contributions. Furthermore, the computational cost of these algorithms coupled to the need for recalculation in the face of changing protein environment or molecular configuration prevents the use of such methods for dynamic applications, particularly in the context of extended proteins.

Received: January 13, 2012
Published: February 17, 2012

Another approach to solvation modeling, which does address the need for computational efficiency, uses simple expressions to model ΔG_{solv}, with coefficients fit to large sets of experimentally measured data.[1,21] The accuracy of such models is heavily reliant on the availability of biologically relevant solvation free energies for training. Although there are extensive databases of experimentally measured ΔG_{solv},[1,24] the range of chemistries included in these databases does not fully reflect the needs of many research applications.

In the present work, we have fit an empirical expression containing both surface area and charge terms to an extensive database of experimental ΔG_{solv}. We have also explored quantum mechanical calculations of ΔG_{solv}[4,20] as a potential source of additional reference values from which the empirical models can be fit. Although these calculations are prohibitively time-consuming for direct integration into large atomistic simulations, the small-molecule solvation free energies produced by quantum mechanical, QM, methods provide a straightforward solution to the need for relevant reference data from which to create an accurate on-the-fly solvation model. In addition we have developed a set of atom types specifically designed to characterize the chemical features of interest for fragment-based drug design. The combination of additional parameters from both charge terms and specific atom types, along with increasingly broad chemical databases, have the potential to create a well-rounded, computationally efficient model for the calculation of solvation free energy in atomic-scale simulation.

■ METHODS

Model. Implicit solvation models require consideration of polar electrostatic interactions, additional nonpolar, solute/solvent van der Waals interactions and the free energy of cavitation or solvent displacement. Often these phenomena are modeled independently since the solvation free energy can be written as the sum of contributions for the various interactions:

$$\Delta G_{solv} = \Delta G_{electrostatic} + \Delta G_{vdw} + \Delta G_{cavitation} \qquad (1)$$

Our basic analytical form assumes that each term in eq 1 can be formulated as the sum of contributions from each of the atoms in the molecule or protein of interest. For a single molecule, we first accumulate the solvent-accessible surface areas, a_i, and the products of the partial charge, q_i, and solvent-accessible surface area for each atom of the same type, t, as follows:

$$A_t = \sum_{i=1}^{N_t} a_i \qquad (2)$$

$$B_t = \sum_{i=1}^{N_t} a_i \cdot q_i \qquad (3)$$

where N_t is the number of atoms of each type contained in the molecular structure of interest. An expression for ΔG_{solv} can then be evaluated as the sum of the contributions from the solvent-accessible surface areas, A_t, and the accumulated area-mediated charges, B_t, for each of the atom types contained in a molecule, ligand or protein.

$$\Delta G_{solv} = \sum_{t=1}^{N} \alpha_t A_t + \beta_t B_t \qquad (4)$$

where N is the number of atom types; the α_t are the fitting parameters for the contribution to ΔG_{solv} of each atom type's nonpolar surface interaction; and the β_t are the fitting parameters for each atom type's contribution to ΔG_{solv} from electrostatic interactions.

We explored extending eq 4 to include an additional term for the molecular volume, to capture the energy difference due to creating a cavity in the solvent. However, the overall fit to our reference data was not significantly improved by the addition of a volume term. The effects based on the size of the molecule are largely incorporated into the area term, which is shown to be linearly related to volume for small molecules[6] (Figure 1).

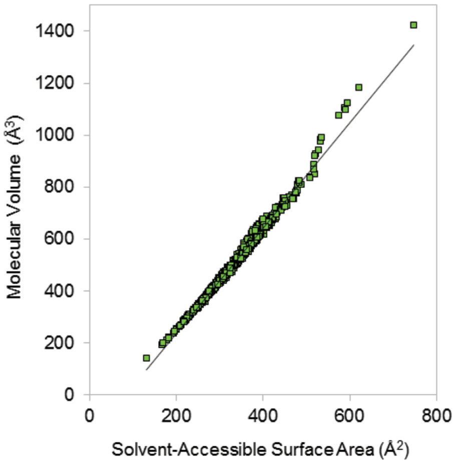


Figure 1. Molecular volume scales linearly with the solvent-accessible surface area for small molecules. As a result, effects that scale with either quantity will be captured with a single fitting parameter.

Evaluating eq 4 requires the calculation of both partial charge and solvent-accessible surface area for each atom in the system of interest. Typically, our atom partial charges would already have been computed for use with the Amber force field.[25] Accordingly, our approach to partial charge calculation for solvation parallels that practice. The major components used were GAMESS and RESP.[26] GAMESS, version Apr112008R1, was used to do the ab initio quantum mechanics at the 6-31G(d) level of theory. GAMESS was instructed to produce an array of electrostatic potential values which were processed by various components of Antechamber,[27] version 1.27, notably including RESP, to compute partial charges at the atom centers. The common GAMESS input file headers appear in the Supporting Information.

Computation of the solvent-accessible surface is purely geometric, using only sphere centers and radii. The solvent-accessible surface is defined[28] as the locus of the center of a probe sphere as it is conceptually rolled around the van der Waals surface of the molecule. This is equivalent to the molecular surface that would result if each atom's van der Waals sphere were inflated by the probe's radius. We have used the Bondi van der Waals radii[29] inflated by a probe radius of 1.4 Å, which is a standard approximation of the size of a water molecule.[1]

Various methods including those of Connolly,[30] of Lee and Richards,[31] and of Sanner[32] were explored for SASA calculation, but eventually a simple method was independently developed: simple point counting.[33] Examining a nearly uniform distribution of points on each atom's inflated sphere,[34] any point which is buried within another atom is rejected. The area of the entire sphere scaled by the fraction of points remaining is the solvent-accessible surface area. While this introduces a quantization error (controllable by increasing the number of points on the sphere) it is generally modest when compared with errors and approximations elsewhere in the model. Using, e.g., 1000 points, the quantization error is on the order of 1 part in 1000; but when an atom is nearly buried, the relative error due to quantization increases significantly.

Multistage Fitting of Neutral and Ionic Parameters. Neutral and ionic fragments contain many of the same chemical features and reasonably similar solvent-accessible surface areas (SASAs) and partial charges at the per-atom scale. Therefore, the nearly order-of-magnitude difference in their solvation free energies (ΔG_{solv}) indicates a need for additional parameter space in order to simultaneously model both neutral and charged species. The standard least-squares methodology simultaneously fits all $2N$ coefficients in eq 4, reproducing reference ΔG_{solv} values using the computed charges and SASAs for all molecules in our training database. However, even with many atom types left exclusively for ionic features, the inclusion of a large set of ions into the training database—while fitting the entire database in one step—causes significant error in both neutral and ionic species. The desire to separate the two data sets is further increased by the known difficulty of experimentally measuring ionic solvation free energies.

The standard method of simultaneously fitting the parameters for all types is sufficient when training only the neutral molecules; however, separate fitting steps are employed for the inclusion of ionic molecules in the training sets. We first fit coefficients for all atom types contained in neutral molecules from our training database. With these coefficients, ΔG_{solv} for neutral molecules containing a wide variety of biologically relevant chemical features can be calculated. Furthermore, because the ΔG_{solv} for each molecule is calculated as a sum of contributions from each atom, we can also use the coefficients to determine the contribution from atoms of the *neutral* types to the solvation free energy of each *ion* in our training database. Subtracting the contribution of the nonionic types from an ion's reference solvation free energy results in the desired contribution to ΔG_{solv} from the remaining atoms that have been assigned "ionic" types. The fitting procedure is then applied to determine a best-fit set of coefficients for these ionic atom types, considering only the atoms associated with those types in each molecule and only their expected contributions to the solvation free energy.

By fitting the ions in a separate stage, we effectively hold constant the coefficients associated with the atom types contained in neutral species. Changing the charge state of a molecule does have extended effects on the ΔG_{solv} contributions of every atom in the molecule in terms of the atom types assigned as well as the redistribution of partial charge throughout the molecule. However, we allow the few types contained only in ionic species to account for the bulk of the order of magnitude increase in ΔG_{solv} for ionic species compared to their neutral counterparts. Phenomenologically, we can therefore confine the larger energy penalty for desolvating these ionic molecules to the specific chemical features responsible for the nonzero charge. The alternative would allow the least-squares fitting to smear the effect throughout the molecule and, worse, throughout every molecule which shares an atom type with the ion.

Atom Types. The effectiveness of an empirically fit solvation model depends in part upon the choice of a well-posed set of fitting parameters. Each atom type adds two parameters to our model and effectively sorts the molecule into bins of atoms whose contribution to ΔG_{solv} can be well-described by a common set of model coefficients. We started with the GAFF atom types used by AMBER[35] for atom typing in the solvation model, because the model is implemented within a suite of software which uses the AMBER empirical potential for configurational energy calculations. Our set of solvation atom types was then refined when specific chemical features were poorly represented by the model, and when a chemically intuitive rationale could be applied to the merger of GAFF types or the creation of new types. The GAFF types being used, as well as our added types are listed in Table 1.

Table 1. Atom Types for Solvation Model

atom type	description
br cl f	GAFF Types
c c1 c2 c3 ca cc ce cg cp cu cv cx cy cz	
h1 h2 h3 h4 h5 ha hc hn ho hp hs hx	
n n1 n2 n3 n4 na nb nc ne nh no	
o oh os	
p2 p4 p5 pb pc px py	
s s2 s4 s6 sh ss sx sy	
c-	carbon in a carboxylate ion
o-	oxygen in a carboxylate ion
nz	protonated nitrogen in charged azole ring
nz+	nitrogen in charged azole ring -no hydrogen bond
on	oxygen in a nitro group
n9	any nitrogen double bonded to C3, P2 or N2
oa	oxygen anion with a single bond
sa	sulfur anion with a single bond
sf	sulfur in sulfonium ion
na+	protonated nitrogen in a six-membered ring
nc-	nitrogen in ionized tetrazole
n+	default protonated nitrogen

Several pairs of redundant types used by the GAFF type definitions to encode bond typing information for series of atoms in resonant structures were unnecessary for solvation. These pairs (*cc*—*cd*, *ce*—*cf*, *cg*—*ch*, and *nc*—*nd*) have been combined into a single type per pair (the first listed) for the solvation model.

Additional types have been included when a particular chemical feature is poorly represented by the broader parent type offered by GAFF. One such case was the oxygen atoms in nitro groups, which are cast by GAFF simply as oxygen with only one bond. The relative complexity of the resonant bond structure for the two oxygen atoms bonded to a nitrogen in a nitro group and a poor showing of nitro molecules in initial model training led to the inclusion of another oxygen type, *on*.

Exploration for potential new types was performed using plots of charge versus the area to look for significant clustering in the measured data for each atom type. When multiple distinct clusters were observed in the atomic properties and a chemical explanation for the behavior could be determined, the effect of splitting the types was examined. As an example, analysis showed that the protonated nitrogen atoms in positively charged, five-membered, aromatic rings were poorly fit because their area depended too starkly on whether the additional electron was involved in bonding with a hydrogen atom or with some larger group such as an alkane chain. As a result the additional types *nz* and *nz+* were formed to differentiate between the two cases.

Some ionic chemical features, such as a sulfur anion, are not described by the GAFF types. In these cases additional atom types were required to accommodate a fragment whose inclusion was desired in the training, or because a specific application required the use of the particular ionic feature it represents.

An algorithm for automatic atom type assignment was developed for use in the solvation model. It uses only atom element identities, bond connectivity, and the formal charge of the molecule to assign the solvation atom types. The typer also produces atom- and bond-types relevant to computations with the AMBER force field. Unlike an earlier method[36] it starts with ring perception, accounts for aromaticity and works outward from correctly assigned ring bonds. This method allows the typer to be more accurate for fragments typically relevant to drug design and to determine results independent of the order in which atoms are input to the algorithm.

RESULTS

Training Data Sets. A comprehensive database of reference solvation free energies is required to calculate the coefficients for each atom type in eq 4. The ideal reference database would contain a wide variety of chemistries, and for each chemical feature there would be sufficient examples to provide a general, context-independent representation of its effect on solvation. Two previous works by Wang et al.[1] and Mobley et al.[24] have compiled extensive databases of experimental solvation free energies. These databases contain 401 and 505 small-molecule free energies, respectively, although their significant overlap yields only 582 distinct neutral molecules and 14 ions. An additional set of ten experimental solvation free energies for carboxylate ions was measured by Lee et al.[37] Experimental measurement of ΔG_{solv} seems biased toward small neutral molecules with relatively simple chemistries. Even with the nearly six hundred fragments contained in these databases, there are significant gaps in the chemistries relevant to pharmaceutical molecules. The most notable deficiencies are found in ionic molecules, key nitrogen and sulfur-containing molecules such as azoles and sulfones, and molecules with fused aromatic rings such as indoles and benzimidazoles.

Many chemical features not included or barely represented in these experimental databases are necessary for the accurate modeling of the in-house database of small molecules that we have developed for fragment-based drug design. In the absence of experimental data, we have turned to ab initio calculations in order to increase the breadth of our training databases. Our in-house database is an accumulation of small molecules (with few, if any, rotatable bonds) which are suitable for use as building blocks in fragment-based drug design. GAMESSPLUS,[20] version 2010−2, was used to calculate QM solvation values for a subset of the in-house database. As an extension to GAMESS, this package uses a charge model to compute a self-consistent reaction field for QM energy calculation, and a corresponding solvation model to compute the solvation free energy from the resulting electron populations. Calculated values are listed in Table S1, Supporting Information. The common GAMESSPLUS input file headers also appear in the Supporting Information.

As validation we have also calculated QM solvation values as described above for the Wang and Mobley data sets and compared them to the experimentally measured ΔG_{solv}. QM solvation is reasonably correlated with the experimental data (Figure 2). The mean absolute error (MAE) for neutral molecules is 0.82 kcal/mol with an rms error of 0.91 kcal/mol. The ions, with significantly larger magnitude ΔG_{solv}, have an MAE of 3.55 kcal/mol with an rms error of 1.88 kcal/mol. We deemed that to be accurate enough for our purposes.

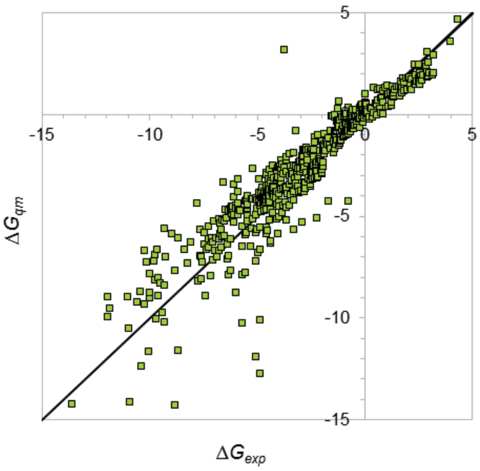


Figure 2. QM solvation free energies calculated with GAMESSPLUS exhibit unsigned mean absolute error of 0.82 kcal/mol compared to experimental values for neutral molecules (shown in figure) and a mean absolute error of 3.55 kcal/mol for ionic molecules.

Solvation Free Energy of Small Molecules. The solvation free energies calculated with our model for a database of 1896 molecules are reported in the Supporting Information. We have fit the model in eq 4 with three different training sets, and evaluated all molecules with each set of coefficients. The resulting free energy values for each molecule, as well as the reference values used in training, are detailed in Table S1 of the Supporting Information, and summarized in Table 2. For comparison with previous work, we first trained the model with the neutral experimental values found in the Wang data set.[1] The resulting mean absolute error for the Wang data set compared to the experimental values is 0.513 kcal/mol; this is better than the 0.538 kcal/mol reported by Wang et al., for their model, trained with the same database. That our model adheres more closely to the training data than does the model by Wang et al. is unsurprising, because our model adds a charge term for each atom type, and has more atom types; as a result it has significantly more parameters than the Wang model. Unfortunately, fitting with only the data in the Wang database is insufficient to capture the chemistries contained in our larger in-house database of small molecules. Because experimental data trends toward simpler molecules, evaluation of the 1,724 neutral molecules contained in the larger database shows an MAE of 16.6 kcal/mol, due to molecules such as azoles and sulfonamides that are not well-represented by the Wang data. This motivates training with a larger data set including examples of additional chemistries.

Table 2. Agreement of Model with Reference ΔG_{solv} with Different Training Sets

	MAE of ΔG_{solv} (kcal/mol)		
training set	Wang	Wang + Mobley	all
Wang	0.51	1.41	16.64
Wang + Mobley	0.61	0.66	2.98
all neutrals	0.83	0.93	1.11
all neutrals + ions	1.41	1.31	1.85

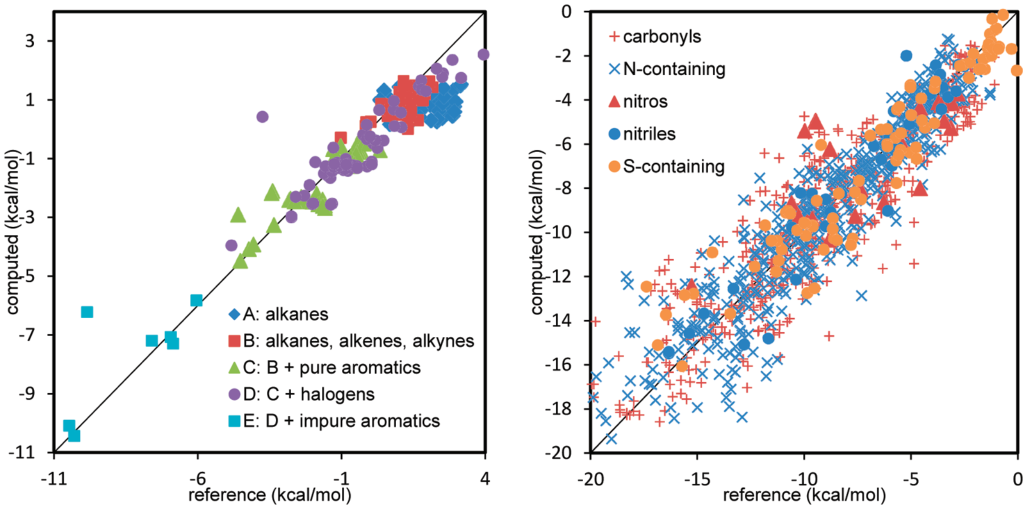


Figure 3. Computed solvation free energy compared with reference values for the neutral fragments in the in-house database. The distribution of chemical features within the database is indicated by varied data markers. Each molecule was plotted with the symbol of the first chemical class in the legend to which it belongs. (In the right-hand plot, *N-containing* was the last category considered.)

Table 3. Agreement of Model with Reference Δ*G*solv by Chemical Class

	trained with Wang			trained with all experimental			trained with all databases		
chemistry	sample size	MAE kcal/mol	RMSE kcal/mol	sample size	MAE kcal/mol	RMSE kcal/mol	sample size	MAE kcal/mol	RMSE kcal/mol
all molecules	1724	16.64	48.95	1724	2.98	5.63	1896	1.85	4.34
alkanes	51	1.14	1.29	51	1.11	1.28	51	1.21	1.38
alkanes, alkenes, alkynes	88	0.82	1.03	88	0.82	1.03	88	0.90	1.12
plus pure aromatics	141	0.68	0.88	141	0.74	0.93	141	0.72	0.95
plus impure aromatics	225	4.88	24.94	225	1.41	4.11	225	0.68	1.00
plus halogens	216	0.61	0.88	216	0.66	0.88	216	0.65	0.89
nitriles	42	19.86	58.19	42	2.65	4.52	43	1.10	1.66
carbonyl	563	11.52	34.97	563	2.71	4.18	603	1.84	3.26
nitros	26	7.06	23.31	26	1.73	2.67	26	1.79	2.22
sulfur-containing	187	11.98	33.17	187	3.5	5.88	203	1.84	4.37
phosphorus-containing	12	0.74	0.96	12	0.56	0.78	12	1.11	1.29
nitrogen-containing	1054	26.55	62.57	1054	4.33	7.12	1140	1.77	3.34
all ions	–	–	–	–	–	–	172	9.22	13.51
carboxylates	–	–	–	–	–	–	20	11.87	15.45
protonated nitrogens	–	–	–	–	–	–	80	9.58	14.03
protonated azoles	–	–	–	–	–	–	9	1.95	2.54

The model was then trained with the all of the available experimental reference data by combining the solvation free energies for neutral molecules found in both the Wang and the Mobley[24] data sets. The MAE for the entire database decreased to 2.98 kcal/mol for 1,724 molecules, because the additional experimental database contains a number of critical chemical features not included in the Wang data. There are also additional nitrogen-containing molecules in the Mobley data which dramatically increases the quality of our model's calculation of the larger database.

Finally, the model was trained with all of the available reference data including both experimental and QM-calculated reference values for ions. The neutral fragments exhibit a MAE of 1.11 kcal/mol; including the ions increases the MAE to 1.85 kcal/mol. The inclusion of QM reference data allows for calculations with significantly more accuracy than the experimental data alone for a broad range of chemistries relevant to drug discovery. The increased breadth of the training set does come at a cost: evaluating the Wang data set when the model is trained with all the possible reference data increases the MAE to 0.832 kcal/mol. The experimental data sets contain large numbers of relatively simple molecules such as short chain alkanes and simple aromatic rings. The broader set includes many more examples of these atom types, but in increasingly complex molecules. Although training the most frequently occurring types in a large range of contexts allows for important gains in the portability of the model, we do lose some accuracy for the simple molecules.

Figure 3 compares the computed solvation free energies of all neutral fragments in the in-house database to their experimental or QM reference values. Symbols indicate different chemical classes and exhibit the clustering of solvation free energy by chemistry. For neutral fragments the model performs about equally well for the various chemistries, although the smallest molecules with positive ΔG_{solv} do exhibit somewhat closer adherence to the reference data.

Table 4. Coefficients Trained with Wang, Wang and Mobley and All Reference Data

	Wang		Wang + Mobley		all			Wang		Wang + Mobley		all	
atom type	α_i	β_i	α_i	β_i	α_i	β_i	atom type	α_i	β_i	α_i	β_i	α_i	β_i
c	0.023	0.202	−0.246	0.473	−0.073	0.200	nb	0.051	0.251	−0.044	0.101	−0.039	0.066
c1	−0.023	−0.003	−0.005	0.042	−0.022	−0.027	nc	−1.511	−0.512	−0.230	0.604	−0.005	0.204
c2	−0.037	0.018	−0.132	−0.217	−0.120	−0.242	ne	–	–	–	–	0.643	1.117
c3	0.160	0.077	0.111	−0.030	0.092	−0.007	nh	0.531	0.521	−0.052	0.015	−0.109	−0.027
ca	−0.070	0.034	−0.076	0.049	−0.067	−0.039	no	−3.235	4.129	−0.189	0.504	2.709	−2.890
cc	0.730	3.403	0.106	0.853	−0.090	−0.036	n9	−0.037	0.041	−0.027	0.030	−0.182	−0.113
ce	−0.068	0.050	−0.148	−0.299	−0.103	−0.240	o	−0.267	−0.259	0.031	0.207	0.068	0.305
cg	−0.053	0.350	0.013	0.182	0.027	−0.079	oh	−0.167	−0.124	−0.158	−0.119	−0.061	0.019
cp	0.088	−0.399	0.054	−0.206	0.011	−0.277	os	0.082	0.236	−0.029	−0.070	0.115	0.290
cx	0.047	−0.013	−1.030	−3.905	−0.280	−1.153	on	0.377	0.852	−0.026	0.037	−0.494	−0.903
cy	−0.310	3.564	−0.237	2.731	−0.014	0.289	p5	3.208	−3.265	2.828	−2.200	4.052	−2.609
h1	−0.041	0.057	−0.040	0.058	−0.032	0.005	py	4.672	1.969	3.550	1.496	5.533	2.332
h2	−0.052	−0.052	−0.045	−0.077	−0.034	−0.118	s	0.299	0.880	0.143	0.429	0.187	0.629
h3	−0.144	0.202	−0.097	0.060	−0.136	0.306	s4	–	–	−0.102	−0.038	−0.364	0.674
h4	0.017	−0.082	0.017	0.089	0.022	0.041	s6	–	–	2.663	−1.589	−5.863	5.173
h5	−0.023	1.051	−0.045	2.218	0.097	−0.402	sh	−0.032	−0.083	−0.035	−0.091	−0.033	−0.091
ha	0.022	0.017	0.032	−0.025	0.026	−0.050	ss	−0.007	−0.060	−0.022	−0.117	0.001	−0.018
hc	0.001	−0.029	0.003	−0.051	0.003	−0.051	sx	–	–	–	–	−0.267	−0.065
hn	0.000	0.000	0.000	0.000	0.000	0.000	sy	–	–	–	–	−0.400	−0.559
ho	0.000	0.000	0.000	0.000	0.000	0.000	cz	–	–	–	–	−8.202	0.820
hp	0.000	0.000	0.000	0.000	0.000	0.000	c−	–	–	–	–	−13.22	16.53
hs	0.000	0.000	0.000	0.000	0.000	0.000	hx	–	–	–	–	−1.209	2.141
br	−0.003	−0.019	−0.003	−0.021	−0.006	−0.045	o−	–	–	–	–	5.757	8.052
cl	0.000	−0.017	0.001	−0.018	0.001	−0.024	oa	–	–	–	–	1.846	4.446
f	0.019	0.049	0.017	0.047	0.016	0.040	n4	–	–	–	–	−0.471	0.770
n	−0.394	−0.254	−0.367	−0.231	−0.050	0.031	nz	–	–	–	–	−0.837	0.142
n1	−0.090	−0.077	−0.029	0.055	0.141	0.370	nz+	–	–	–	–	−14.98	19.77
n2	−0.004	0.004	−0.003	0.003	−0.095	−0.011	na+	–	–	–	–	−1.821	0.445
n3	−0.067	−0.014	−0.063	−0.012	−0.061	−0.001	n+	–	–	–	–	−0.961	0.116
n4	-	-	-	-	−0.471	0.770	nc−	–	–	–	–	−0.407	0.175
na	3.222	−6.089	0.151	1.356	0.029	0.308	sa	–	–	–	–	−1.123	−0.331

Incorporating ab initio reference data into the training set of our solvation model produces marked improvement in the calculation of solvation free energy for certain vital chemical classes (Table 3). Simple chemistries which are extremely well represented in the experimental databases do show slight increases in error, but a number of previously unusable chemistries such as nitriles and carbonyls are reasonably addressed by the solvation model. In addition, the new data sets make possible the evaluation of ionic types needed not only for accurate calculation of solvation free energy in the context of most proteins, but also for charged fragments.

The coefficients developed for each of the cases discussed above are shown in Table 4. Coefficients are listed as zero when molecules containing the atom type are actually included in the database but the atoms have no SASA and as a result do not participate in the calculation of ΔG_{solv}. Conversely, when the model is fit with a subset of the total data, there are certain types not represented in the training set and as a result no coefficient is listed. For the purposes of evaluating the above data sets, atoms with no representation in the training set were effectively ignored in the calculation of ΔG_{solv}. Although there is a significant effect on the accuracy of individual ΔG_{solv} calculations, retaining the unrepresented molecules gives a more accurate picture of the increased functionality of the model with improved training sets.

■ DISCUSSION

Calculation of the solvation free energy is a persistent challenge in the field of computational biochemistry. Although there is a strong grasp of some of the varied phenomena which contribute to solvation effects on the free energy of a system, more work is required to incorporate all of them into a physically intuitive model generically applicable at the atomic scale. A variety of techniques can accurately predict electrostatic interactions in implicit solvent models. QM solvation calculations are also adequate in many cases for predicting small-molecule solvation free energies. However, such techniques are not computationally efficient enough for integration into large simulations. The most notable disconnect between the theory of solvation and standard practice when calculating the solvation free energy is the incorporation of the energy differences due to solvent cavitation, which are often ignored. These effects are likely to be small in general, particularly relative to the error inherent in most solvation models; however, a full description of the solvation free energy across broad classes of biochemical molecules will eventually require incorporation of these volume effects. Many other problems in computational biochemistry provide further challenges, notably the protein context necessary for application of implicit solvent models to fragment-based drug design.

Recent work by Mobley et al.[24] highlights that experimental solvation free energies, typically measured at room temperature, are an average over the varied conformations a molecule exhibits at temperature. The disconnect between the dynamic ensemble of configurations sampled for the experimental measurements and the single, rigid configurations used in our fitting scheme likely constitutes a significant source of error in the present work. The distribution of conformations will also change between free solution and the constrained environment of "induced fit" to a protein, yielding a further, nonstandard change to the entropic contribution to free energy. Mobley et al. calculate a mean error of −0.500 ± 0.089 kcal/mol in the free energy of solvation for molecules relaxed in vacuum, such as our databases, versus experimental conditions. The error would be mitigated in an unlikely context where no conformational change was allowed, particularly when the single conformation matches that used for QM calculations of solvation free energies (noting, though, that the QM solvation methods were parametrized to match experimental results). Even ignoring dynamics, however, the use of our model in the protein context typically requires relaxation of the configuration. In some cases, the relaxation exposes an oversensitivity to changes in solvent-accessible surface area, particularly for atom types which are largely buried in the equilibrium, small-molecule context. We intend to pursue the assignment of a statistically relevant sampling of nonequilibrium configurations to each experimental solvation free energy as a method for incorporating increased robustness to this model.

It also remains for future investigation to explore whether prediction accuracy could be improved by training the atom radii and probe radius.[38] Such training may prove particularly fruitful for atoms which are currently substantially screened by their neighboring atoms. Training the radii may also provide a path to reintegrating the hydrogen atoms attached to atoms other than carbon, currently ignored by our model because they have no surface area which is solvent-accessible. Our current linear algebra methods are inadequate to these nonlinear tasks; additional fitting schemes need to be adopted.

Currently the state of the art requires a researcher to determine whether accuracy or computational efficiency of the solvation model is more critical to the work being performed. Our goal was a solvation free energy model sufficiently fast for incorporation into large-scale simulations over many iterations or for real-time evaluation of fragment-based drug designs. In this we are undoubtedly successful. On the other hand, the instantaneous calculation of dramatically wrong solvation free energies would obviously be of no value: we also show reasonable agreement with the available experimental data for small-molecule solvation free energies. In addition the application of solvation free energies from QM methods as training data for this model has allowed us to dramatically improve the accuracy for important chemical classes such as azoles, sulfonamides and molecules with multiple fused rings. We have also provided a flexible enough set of atom types that the inclusion of further QM reference data would allow future workers to tailor the model to perform well with chemistries not currently addressed.

The end goal of an all-inclusive model accurately describing chemical features relevant to a broad range of potential applications in biochemistry and computed with relevant performance is a grand challenge. We have described a method that allows experimental reference data to be supplemented with ab initio calculations and that can in principle be used to provide fast, reasonably accurate solvation free energies for any chemistry of interest.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

A full listing of small molecules used in training sets, their experimental or QM-derived ΔG_{solv} reference values and computed ΔG_{solv} for each model discussed in the text; input header files for GAMESS partial charges and for GAMESSPLUS QM solvation calculations. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: rbryan@bioleap.com.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank John L. Kulp III and John L. Kulp, Jr., for comments on the manuscript and Dennis G. Sprous for assistance with QM solvation.

■ REFERENCES

(1) Wang, J.; Wang, W.; Huo, S.; Lee, M.; Kollman, P. A. *J. Phys. Chem. B* **2001**, *105*, 5055−5067.

(2) Mecinovic, J.; Snyder, P. W.; Mirica, K. A.; Bai, S.; Mack, E. T.; Kwant, R. L.; Moustakas, D. T.; Heroux, A.; Whitesides, G. M. *J. Am. Chem. Soc.* **2011**, *133*, 14017−14026.

(3) Kuntz, I. D.; Chen, K.; Sharp, K. A.; Kollman, P. A. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 9997−10002.

(4) Cramer, C. J.; Truhlar, D. G. *Acc. Chem. Res.* **2008**, *41*, 760−768.

(5) Kulp, J. L.; Blumenthal, S. N.; Wang, Q.; Bryan, R. L.; Guarnieri, F. *J. Comput.-Aided Mol. Des.* **2012**, in press.

(6) Mobley, D. L.; Bayly, C. I.; Cooper, M. D.; Shirts, M. R.; Dill, K. A. *J. Chem. Theory Comput* **2009**, *5*, 350−358.

(7) Genheden, S.; Mikulskis, P.; Hu, L.; Kongsted, J.; Söderhjelm, P.; Ryde, U. *J. Am. Chem. Soc.* **2011**, *15*, 13081−13092.

(8) Thompson, J. D.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. A* **2004**, *108*, 6532−6542.

(9) Shivakumar, D.; Williams, J.; Wu, Y.; Damm, W.; Shelley, J.; Sherman, W. *J. Chem. Theory Comput.* **2010**, *6*, 1509−1519.

(10) Shivakumar, D.; Deng, Y.; Roux, B. *J. Chem. Theory Comput.* **2009**, *5*, 919−930.

(11) Hassan, S. A.; Guarnieri, F.; Mehler, E. L. *J. Phys. Chem. B* **2000**, *104*, 6478−6489.

(12) Sitkoff, D.; Sharp, K. A.; Honig, B. *J. Phys. Chem.* **1994**, *98*, 1978−1988.

(13) Cozzini, P.; Fornabaio, M.; Marabotti, A.; Abraham, D. J.; Kellogg, G. E.; Mozzarelli, A. *Curr. Med. Chem.* **2004**, *11*, 3093−3118.

(14) Nicholls, A.; Mobley, D. L.; Guthrie, J. P.; Chodera, J. D.; Bayly, C. I.; Cooper, M. D.; Pande, V. S. *J. Med. Chem.* **2008**, *51*, 769−779.

(15) Knight, J. L.; Brooks, C. L. *J. Comput. Chem.* **2011**, 32, 2909−2923.

(16) Guthrie, J. P. *J. Phys. Chem. B* **2009**, *113*, 4501−4507.

(17) Guvench, O.; MacKerell, A. D. *Curr. Opin. Struct. Biol.* **2009**, *19*, 56−61.

(18) Yang, P.-K.; Lim, C. *J. Phys. Chem. BB* **2008**, *112*, 14863−14868.

(19) Shirts, M. R.; Pande, V. S. *J. Chem. Phys.* **2005**, *122*, 134508−134521.

(20) Higashi, M.; Marenich, A. V.; Olson, R. M.; Chamberlin, A. C.; Pu, J.; Kelly, C. P.; Thompson, J. D.; Xidos, J. D.; Li, J.; Zhu, T.; et al. GAMESSPLUS, version 2010−2, University of Minnesota, Minneapolis, 2010, based on the General Atomic and Molecular Electronic Structure System (GAMESS) as described in Schmidt, M. W.; Baldridge, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. H.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S. J.; et al. *J. Comput. Chem.* **1993**, *14*, 1347−1363.

(21) Ooi, T.; Oobatake, M.; Némethy, G.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **1987**, *84*, 3086−3090.
(22) Sharp, K. A.; Honig, B. *Annu. Rev. Biophys. Biol.* **1990**, *19*, 301−332.
(23) Cramer, C. J.; Truhlar, D. G. *Science* **1992**, *256*, 213−217.
(24) Mobley, D. L.; Dill, K. A.; Chodera, J. D. *J. Phys. Chem. B* **2008**, *112*, 938−946.
(25) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. *J. Comput. Chem.* **2004**, *25*, 1157−1174.
(26) Schmidt, M. W.; Baldridge, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. H.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S. J.; Windus, T. L.; Dupuis, M.; Montgomery, J. A. *J. Comput. Chem.* **1993**, *14*, 1347−1363.
(27) Wang, J.; Wang, W.; Kollman, P.; Case, D. *Molecules* **2001**, *222*, U403−U403.
(28) Cramer, C. J. *Elements of Computational Chemistry*, 2nd ed.; Wiley: New York, 2004; pp 407−408.
(29) Bondi., A. Van der Waals Volumes and Radii. *J. Phys. Chem.* **1964**, *68*, 441.
(30) Connolly, M. L. *J. Am. Chem. Soc.* **1985**, *107*, 1118−1124.
(31) Lee, B.; Richards, F. M. *J. Mol. Biol.* **1971**, *55*, 379−400.
(32) Sanner, M. F.; Olson, A. J.; Spehner, J. C. *Biopolymers* **1996**, *38*, 305−20.
(33) Shrake, A.; Rupley, J. A. *J. Mol. Biol.* **1973**, *79*, 351−371.
(34) Bauer, R. *J. Guid. Control. Dynam.* **2000**, *23*, 130−137.
(35) Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. *J. Comput. Chem.* **2005**, *26*, 1668−1688.
(36) Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A. *J. Mol. Graphics Modell.* **2006**, *25*, 247−260.
(37) Lee, S.; Cho, K.-H.; Lee, C. J.; Kim, G. E.; Na, C. H.; In, Y.; No, K. T. *J. Chem. Inf. Model.* **2011**, *51*, 105−114.
(38) Nicholls, A.; Wlodek, S.; Grant, J. A. *J. Phys. Chem. B* **2009**, *113*, 4521−4532.