HHS Public Access

Author manuscript

J Chem Theory Comput. Author manuscript; available in PMC 2017 April 12.

Published in final edited form as:

J Chem Theory Comput. 2016 April 12; 12(4): 1930–1941. doi:10.1021/acs.jctc.5b00934.

Using MD simulations to calculate how solvents modulate solubility

Shuai Liu[†], **Shannon Cao**[†], **Kevin Hoang**[†], **Kayla L. Young**[‡], **Andrew S. Paluch**[‡], and **David L. Mobley**[*,¶]

[†]Department of Pharmaceutical Sciences, University of California, Irvine, Irvine, CA 92697

[‡]Department of Chemical, Paper and Biomedical Engineering, Miami University, Oxford, Ohio 45056, USA

[¶]Department of Pharmaceutical Sciences and Department of Chemistry, University of California, Irvine, Irvine, CA 92697

Abstract

Here, our interest is in predicting solubility in general, and we focus particularly on predicting how the solubility of particular solutes is modulated by the solvent environment. Solubility in general is extremely important, both for theoretical reasons – it provides an important probe of the balance between solute-solute and solute-solvent interactions – and for more practical reasons, such as how to control the solubility of a given solute via modulation of its environment, as in process chemistry and separations. Here, we study how the change of solvent affects the solubility of a given compound. That is, we calculate relative solubilities. We use MD simulations to calculate relative solubility and compare our calculated values with experiment as well as with results from several other methods, SMD and UNIFAC, the latter of which is commonly used in chemical engineering design. We find that straightforward solubility calculations based on molecular simulations using a general small-molecule force field outperform SMD and UNIFAC both in terms of accuracy and coverage of the relevant chemical space.

1 Introduction

Solubility is a fundamental property in industry, and is of particular interest in purification and separations. Thus, a good deal of research effort has been invested towards predicting solubility. However, in a recent blind test of current methods[1] on aqueous solubilities, predictions did not perform nearly as well as retrospective tests, suggesting substantial challenges remain. In part, there may be large issues with the transferability of these models, which are often fairly highly parameterized based on existing data. Challenges may be even

[*]To whom correspondence should be addressed, dmobley@mobleylab.org.

Supporting Information Available

A PDF file with additional figures and tables analyzing error for other methods; a .zip file containing GROMACS topology/coordinate files used as input for all calculations; sample GROMACS run input (.mdp) files for all calculations; sample run shell scripts for the calculations; and exact calculated values for all solvation free energies from GROMACS, UNIFAC, and SMD, in an Excel spreadsheet.

This material is available free of charge via the Internet at http://pubs.acs.org/.

worse when moving away from aqueous solubilities – for which substantial data is available for parameterization – to other solvents.

Several classes of methods have been employed in this area. One main category of methods is empirical methods based on molecular descriptions, like the Group Contribution (GC) method. In this category, one commonly employed method is UNIFAC[2–6] which uses a compound library to analyze the contribution of each functional group to the solute activity coefficient. When used with limited experimental data for the pure solid solute, the equilibrium solubility may be computed in a wide range of solvents. This approach is fast, and can produce acceptable results in many cases. However, a major potential drawback of this class of methods is that GC methods require a good deal of experimental data to calculate the contributions of each functional group. If a functional group does not exist in the experimental library, then solubility predictions for compounds with this functional group cannot be expected to be accurate.

A second category includes statistical methods like multiple linear regression (MLR) or Neural Network (NN) methods.[7] These methods use statistical or machine-learning tools to analyze existing data, build a model, polish the parameters of the model, test the model and then use the created model to predict solubility. Some of these methods have good results,[8–10] with RMS errors (RMSE) around 1.0 log unit and correlation coefficients (R^2) around 0.8. However, these models require a large amount of high quality input data for training, which can pose challenges. For example, high quality experimental data can be very difficult to obtain. Additionally, the physical interpretation of each model can be problematic. Specifically, the parameters in these models may not have simple physical interpretations, meaning that it can be difficult to understand why a particular prediction is made, or what ought to be done to change solubility in the desired direction. Overall, both major classes of method frequently suffer from problems of transferability, as illustrated by recent blind tests.[11] This is likely because these methods are highly dependent on the size and quality of the training set, and because of the degree of human input required in building the models.

There have been relatively few simulation-based efforts to calculate solubilities or relative solubilities from physical principles rather than the empirical training used in the studies above.[12–14] Here, we will call calculations based on physical principles “direct” solubility calculations, and in our view direct calculations are those which do not require training on solubility data [1], and do not require human interpretation of or adjustment of the model. Rather, direct calculations typically involve calculation of the underlying thermodynamic contributions to solubility (the chemical potentials of the solute in solid versus in solution) or approximations thereof. So here, we focus on using simulations to calculate solubilities, and in particular, relative solubilities.

[1]Direct calculations do not require training on solubility data, and are often based on a physical force field. However, force fields can be fitted to a wide variety of data. While we are not aware of a current force field which has been fitted to reproduce solubility data, some current-generation force fields *have* been fitted to reproduce solvation free energies,[15–18] though that is not the case for the General AMBER force field (GAFF) used here (though a reparameterization of GAFF that would include fitting to these has been proposed[19]).

We focus on relative solubility calculations because absolute calculations are still quite challenging. It is still difficult to compute the residual chemical potential of the solid[20–22] or related properties as needed for equilibrium solubility calculations.[23] Focusing on calculating the relative solubility of a solute in different solvents allows us to focus on solution-phase thermodynamics of the solute and how these are affected by the solvent. In other words, we can still directly calculate *relative* solubilities of the same solute in different solvents even without information about the chemical potential or free energy of the solid. Details of our approach can be found below in Methods. Here, we compute solubilities for eight solutes in 34 different solvents, for a total of 53 different solute-solvent pairs. Data for our test comes from the Open Notebook Challenge.[24] For each of these solute-solvent pairs, we compute the solvation free energy and other properties, allowing us to calculate the relative solubility for comparison with experiment.

We also compare our methods with two other commonly used methods, UNIFAC[2–6] and SMD,[25–27] and find that our calculations are more accurate than those from the stated methods on the present set, and also cover more of the compounds in our set.

While this study is the first we are aware of which applies a physical approach based on alchemical free energy calculations to calculate relative solubilities, there have been related studies on the solvation of small molecules in non-aqueous solvents;[15,17,28–31] it is calculations of solvation free energies that provide the foundation for our approach here. Following in the footsteps of earlier work,[29] one notable recent study[31] reported calculations of the solvation free energy for different solutes in a variety of organic solvents. Experimental solvation free energy data was obtained from the databases of Katritzky *et al.*,[32,33] which appears to draw both on direct measurements of solvation and on vapor pressure measurements[2]

2 Methods

2.1 Theory

To calculate the solubility of a single solute in a particular solvent directly, we need to know two pieces of information: the solvation free energy, and the fugacity of pure solid solute. Given these, the solubility can be calculated as was done in ref.:[38]

$$\ln \quad x_1^{\alpha} \quad = \quad -\beta\mu_1^{\alpha,res} \quad (T,p,x_1) - \ln \quad \left(\frac{RT}{v\,(T,p,x_1)}\right) + \ln \quad f_1^{S}\,(T,p) \qquad (1)$$

[2]The work of Katritzky *et al.* refers to “solubility”, as in Ostwald solubility (the relative concentration of a compound in gas versus solution) when discussing the solvation of molecules, which can create some confusion. But solvation free energies are particularly difficult to measure (some of the complexities are addressed by the work of Guthrie and collaborators on preparing the Statistical Assessment of Modeling of Proteins and Ligands (SAMPL) series of challenges[34–37]), and require a great deal of care in curating the experimental data, as Guthrie’s work indicates. Thus, solvation free energies are only available for a relatively small number (a few thousand[36]) of compounds, and new measurements require great care. Solvation free energies are perhaps one of the few physical properties where too much dynamic range poses a problem – if the solvation free energy is too favorable or too unfavorable, the concentration in the vapor phase or in solution will be extremely difficult to measure. As a result of these factors and others, few solvation free energies are available for drug-like or polyfunctional compounds[35] which are often of interest to simulators, making it difficult to test our force fields on these compounds. In contrast, solubility is a property of considerable interest in drug discovery and other areas, and is routinely measured for large numbers of compounds. Thus experimental solubility data is comparatively abundant, making the calculation of solubilities – even relative solubilities – particularly important.

where x_1^α is the equilibrium solubility of the solute in units of mole fraction, $\beta\mu_1^{\alpha,res}$ is the dimensionless residual chemical potential of the solute (denoted by the subscript 1) in solvent α, v is the molar volume of the mixture (solute 1 in solvent α), and f_1^S is the fugacity of pure solid solute.

In concentration units (molar), this can be rewritten as:

$$\ln \quad c_1^\alpha \quad = \quad \ln \quad \left(\frac{x_1^\alpha}{v\left(T,p,x_1\right)}\right) \quad = \quad -\beta\mu_1^{\alpha,res}\left(T,p,x_1\right) - \ln \quad \left(RT\right) + \ln \quad f_1^S\left(T,p\right) \tag{2}$$

where c_1^α is the molar concentration (at the equilibrium solubility) of solute 1 in solvent α.

From equation 2, since f_1^S is a solute dependent constant and RT is constant, we can compute the relative solubility of the solute 1 in solvent α relative to solvent ζ as

$$\ln \quad \left(\frac{c_1^\alpha}{c_1^\zeta}\right) \quad = \quad \ln \quad \left(\frac{x_1^\alpha}{x_1^\zeta}\frac{v^\zeta\left(T,p,x_1^\zeta\right)}{v^\alpha\left(T,p,x_1^\alpha\right)}\right) \quad = \quad \beta\mu_1^{\zeta,res}\left(T,p,x_1\right) - \beta\mu_1^{\alpha,res}\left(T,p,x_1\right) \tag{3}$$

where here v^α and v^ζ correspond to the molar volume of the binary mixture of the solute in solvent α and ζ, respectively.

If we assume that the solute is at infinite dilution, then solute-solute interactions can be ignored, so that the molar volume is independent of the solute concentration or mole fraction. In this case:

$$\ln \quad \left(\frac{c_1^\alpha}{c_1^\zeta}\right) \quad = \quad \beta\mu_1^{\zeta,res,\infty}\left(T,p\right) - \beta\mu_1^{\alpha,res,\infty}\left(T,p\right) \tag{4}$$

where the residual chemical potential is at infinite dilution (superscript ∞).

In this case, the residual chemical potential is equal to the Gibbs free energy of solvation of a single solute molecule:

$$\mu_1^{\alpha,res,\infty} \quad = \quad \Delta G_{1,\mathrm{solv}}^{\alpha,\infty} \tag{5}$$

So equation 4 allows us to estimate relative solubilities (on the left hand side) from solvation free energies readily obtained from molecular simulations (right hand side) at infinite dilution. Equation 4 is a relative formula, comparing the solubility of the same solute in different solvents. Thus, we can compute solvation free energies for a single solute in different solvents and calculate relative solubilities in different solvents for direct comparison with experiment. This approach can be used even in the absence of knowledge of the crystal structure of the solid, which can be difficult to calculate,[20–22] and its fugacity

(ln f_1^S (T, p) in equation 1), which can be even more difficult to calculate. The main assumption inherent in this approach is that the solubility is low enough that the solute can be treated as infinitely dilute. If this were not the case, then solute-solute interactions would need to be considered by calculating the residual chemical potentials in equation 3 would need to be calculated as a function of concentration, which is potentially feasible, but more computationally demanding.

2.2 Dataset selection

To compare calculated solubilities, we drew on the Open Notebook Science Solubility Challenge[24] which provides 9700 experimental solubility datasets, where in their terminology, a "dataset" consists of a set of experimental data resulting in a solubility measurement. We wanted a test set consisting of around 50 solubility measurements, so we filtered these 9700 measurements to select a sub-set based on four rules. First, we focused on relatively small solutes by picking cases where the number of solute heavy atoms was less than 15. Second, we focused on molecules only containing carbon, hydrogen, nitrogen, and oxygen. Third, we focused on molecules with a formal charge of zero. And fourth, we limited the number of rotatable bonds to three or less. While none of these rules represent fundamental limits of the methods we employ here, they do allow us to focus on a subset of available data, and specifically on cases where we expect conformational sampling to be relatively straightforward[39] and force field issues to be fairly well understood. Additionally, challenges relating to the calculation of solvation free energies of charged species[40,41] are avoided. We also required an experimental solubility under 0.1 mole fraction to meet our infinite dilution assumption as given in equation 4. This still left us with more solute-solvent pairs than needed, so we manually selected the final set, ensuring that each solute appears at least twice (to be able to calculate the relative solubility); that a wide range of topologies are considered (including chains, simple rings (both aromatic and non-aromatic), and polycyclic rings). We also deliberately avoided most carboxylic acids, as these could undergo a change of protonation state on transfer between different solvents, though we included two such molecules as a test. Our final set consists of 53 solute-solvent pairs, as detailed in Table 1. 2D structures are shown in Figure 1.

2.3 Simulation

Our approach here is to use alchemical free energy calculations based on molecular dynamics simulations[42,43] to compute solvation free energies for solutes in solution.

After construction of our test set, we generate input files for free energy calculations for all solute-solvent pairs in the set. For each solute or solvent, we take the SMILES string and generate 3D structures using OpenEye OEChem Python toolkit and Omega,[44] then assign AM1-BCC[45,46] partial charges. Antechamber[47] from AmberTools 13 was used to assign GAFF[48] atom types and then AmberTools' tleap was used to generate assign GAFF parameters[48] and write AMBER .prmtop and .crd files. The resulting files were converted to GROMACS format using acpype.[49] The individual solute and solvent GROMACS input files were stored, and packmol[50] was used to create solvated boxes consisting of one solute surrounded by many different solvent molecules. The simulation boxes were cubic, with at least 1.2 nm from the solute to the nearest box edge.

AMBER combination rules (arithmetic average for σ and geometric for ε) were used. Simulations were run using Langevin dynamics, as previously,[51–54] and the timestep was 1 fs. Lennard-Jones interactions were switched off between 0.9 and 1.0 nm, and an analytical correction was applied to the energy and pressure. PME was used for electrostatics, as previously. The real-space cutoff was 1.2 nm. LINCS constrained bonds to hydrogen.

We use λ as a parameter to control the transformation between end states, as is typical in alchemical calculations. λ ranges between 0 and 1, where 0 represents the unmodified system and 1 represents the end-state of the transformation. In this version of GROMACS, we use two separate λ values, one (λ_{chg}) which controls the solute-environment electrostatic interactions, and another (λ_{LJ}) which controls the Lennard-Jones interactions between the solute and its environment. We used λ_{chg} = [0.0, 0.25, 0.5, 0.75, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0] and λ_{LJ} = [0.0, 0.00, 0.0, 0.00, 0.0, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95, 1.0]. In this case, Coulombic interactions were turned off first, followed by the LJ interactions.

While in general we used the standard GAFF force field, we also ran a separate set of calculations to test the new GAFF-DC hydroxyl parameters,[55] a modification of the original GAFF parameter set, specifically, modification of the Lennard-Jones parameters and a rescaling of some of the AM1-BCC partial charges. This involved repeating our calculations for all hydroxyl-containing solute/solvent combinations.

For each λ value we first ran constant-pressure equilibration and the box sizes were adjusted at the end of equilibration (via an affine transformation) to set the box size to the correct average volume from equilibration. Then we ran additional 5 ns constant-pressure production simulation and discarded the first 100 ps as additional “equilibration”, as previously.[51] The Parrinello-Rahman barostat was used to modulate the pressure.

It is worth briefly remarking on the choice of AM1-BCC partial charges. In previous work, we found that for hydration free energies, these charges performed nearly as well as or better than RESP charges fit to a variety of much higher-level quantum mechanical calculations, with our without an SCRF treatment of solvent. MP2/cc-pVTZ SCRF calculations did yield small gains in accuracy, however.[56] But this was for hydration free energies, which involve transfer from gas to water. The dependence in calculated solubilities on charge set is expected to be somewhat smaller here, because the difference in the dielectric constant between environments is much less pronounced than in gas-to-water transfer. Therefore, in view of the computational expense and the lack of substantial accuracy gains expected, we retained AM1-BCC charges as we have in virtually all of our solvation free energy work since we studied this issue.[56,57]

2.4 Other methods

In addition to the free energy calculations discussed above, we also used the SMD and UNIFAC methods to serve as reference sets of predictions.

2.4.1 SMD—SMD is an electronic structure calculation method to compute $\Delta G^{\alpha,\infty}_{1,\mathrm{solv}}$ (see equation 5).[58] SMD employs an implicit solvent model that interacts with the charge density

of the solute molecule, which has been optimized to reproduce 2821 experimental solvation free energies.[25–27] So long as five parameters are available for a particular solvent (the dielectric constant, refractive index, bulk surface tension, and the acidity and basicity parameters), the solvation free energy of a solute (charged or neutral) may be estimated. The interested reader is directed to refs.[25–27] for further information.

For all of these calculations, the geometry of the solute was first optimized in vacuum at the M06-2X/cc-pVTZ level of theory/basis set followed by single point energy calculations at the M06-2X/6-31G(d) level of theory/basis.[59,60] Next, two approaches were used to compute the solvation free energy. First, single point energy calculations were performed on the vacuum optimized structure at the M06-2X/6-31G(d) level of theory/basis set in a self-consistent reaction field (SCRF) using the SMD universal solvation model for each solvent of interest. These calculations were performed following the work of refs.[26,27], and are labeled as “SMD vac” in this work.

Second, the solute geometry was re-optimized at the M06-2X/cc-pVTZ level of theory/basis in a SCRF using the SMD universal solvation model for each solvent of interest, followed by single point energy calculations at the M06-2X/6-31G(d) level of theory/basis. This work was motivated by the recent study of Klimovich and Mobley[51] that showed that the solute conformation in solution may be different than in vacuum, which in turn has an appreciable effect on the computed solvation free energy. These calculations will be labeled as “SMD” in this work.

In both cases, the single point energy calculation in each solvent combined with the single point energy calculation in vacuum may be used to estimate the solvation free energy in each solvent.

Additionally, to assess the sensitivity of the calculations to the chosen basis set, we repeated all of the single point energy calculations at the M06-2X/cc-pVTZ level of theory/basis set. These calculations will be labeled as “SMD vac cc-pVTZ” and “SMD cc-pvtz” for the use of vacuum and solvent optimized geometries, respectively.

The calculations were all performed with Gaussian 09, Revision B.01[61].

2.4.2 UNIFAC—UNIFAC[2–4] and mod-UNIFAC (Dortmund)[5,6] are predictive group contribution methods used extensively in chemical engineering design to model phase-equilibria. Within both models, one may estimate the composition dependent activity coefficient of the solute in solution, or in this study, we restrict ourselves to the composition independent infinite dilution activity coefficient. UNIFAC is parameterized around vapor-liquid equilibrium data. The mod-UNIFAC model makes minute empirical modifications to the functional form of UNIFAC to improve agreement with experiment. Additionally, mod-UNIFAC is fit to vapor-liquid equilibrium data, in addition to infinite dilution activity coefficient, excess enthalpy, excess heat capacity, liquid-liquid equilibrium, solid-liquid equilibrium, and azeotropic data. The interested reader is directed to refs.[2–6] for further information.

The infinite dilution activity coefficient is directly related to the infinite dilution residual chemical potential, allowing equation 4 to be re-written as[38,62]

$$\ln\left(\frac{c_1^{\alpha}}{c_1^{\zeta}}\right) = \ln\left(\frac{\gamma_1^{\zeta,\infty}}{\gamma_1^{\alpha,\infty}}\right) + \ln\left(\frac{v^{\zeta}(T,p)}{v^{\alpha}(T,p)}\right) \quad (6)$$

where $\gamma_1^{\alpha,\infty}$ and $\gamma_1^{\zeta,\infty}$ are the infinite dilution activity coefficient of the solute in solvent α and ζ, respectively, which are computed using UNIFAC or mod-UNIFAC, and v^{α} and v^{ζ} are the molar volume of pure solvent α and ζ, respectively. In this study the molar volume term makes only a minor contribution comparing to the infinite dilution activity coefficient. For example, for benzoic acid in solvents toluene and pentane, the value of $\ln\left(\frac{\gamma_1^{\zeta,\infty}}{\gamma_1^{\alpha,\infty}}\right)$ is 1.24, but the relative volume term ($\ln\left(\frac{v^{\zeta}(T,p)}{v^{\alpha}(T,p)}\right)$) contributes only −0.01. Similar contributions are found in most cases, as presented in the Supporting Information. For our UNIFAC/mod-UNIFAC relative solubility calculations, we use calculated infinite dilution activity coefficients in combination with experimental molar volumes for the pure solvents in order to obtain predicted relative solubilities.

2.4.3 Summary—In total, we used eight methods to calculate relative solubilities, which we label as follows:

1. GAFF: Alchemical free energy calculations with standard GAFF
2. GAFF-DC: Alchemical free energy calculations with GAFF-DC[55]
3. SMD: SMD using the solvent optimized geometry with M06-2X/6-31G(d) single point energy calculations
4. SMD vac: SMD using the original vacuum optimized geometry with M06-2X/6-31G(d) single point energy calculations[25–27]
5. SMD cc-pVTZ: SMD using the solvent optimized geometry with M06-2X/cc-pVTZ single point energy calculations
6. SMD vac cc-pVTZ: SMD using the original vacuum optimized geometry with M06-2X/cc-pVTZ single point energy calculations
7. UNIFAC: The UNIFAC approach[3,4]
8. mod-UNIFAC (Dortmund): A slightly modified the functional formal of UNIFAC[5,6]

Results from these approaches will be discussed below.

3 Results

Much of our previous work on solvation has focused on hydration free energies – the solvation of small molecules in water – but here, we instead study how small molecules

dissolve in a variety of different solvents by calculating how the choice of solvent modulates a solute's solubility. Our free energy calculations allow us to calculate the term on the right side of equation 3 – that is, the difference in dimensionless residual chemical potentials $\mu_1^{\zeta,res,\infty} - \mu_1^{\alpha,res,\infty}$, or the difference in solvation free energies between solvents α and ζ (equation 5). We call this value the calculated value. We can then directly compare with the experimental relative solubility – the term involving ln c_1 on the left side of equation 3. This is labeled the experimental value. The error for a particular solute-solvent pair is then taken as the difference between the calculated value and the experimental value.

Analysis is made slightly more complicated by the fact that we can actually calculate many different errors which are interrelated. So for one solute, if there are n different solvents (and, correspondingly, n solvation free energies), there are $n(n-1)/2$ solvent pairs leading to $n(n-1)/2$ potential errors which can be calculated (though only n of these are independent). Because all of these potential errors involve the same solute, they all provide data about how well that solute's solubility is predicted in different environments and thus are useful to consider as a unit. We therefore call this set of all possible pairwise errors for a given solute a 'dataset' and we number each dataset by the solute's PubChem Compound Identifier (CID). For each solute's dataset, we calculate and report the mean error and the mean absolute error for all pairs.

Figure 2 shows the average of these errors across all pairs for each solute, for each of the methods examined here. Tables 2—3 and the tables in the Supporting Information show error statistics for these methods. Results shown in the column "All Pairwise Errors" suggest that the simulation with new GAFF hydroxyl parameters in general performed best among all methods.

We are also interested in understanding not just the error in our calculated values, but how well they capture experimental trends. Thus, we plot experimental relative solubilities versus calculated ones - specifically, experimental vs calculated $\ln\left(c_1^{\alpha}/c_1^{\zeta}\right)$ – in Figure 3. We find that our approach based on full free energy calculations with GAFF or GAFF-DC performs best in terms of correlation (R^2) with experiment. In contrast, SMD yields very low correlation with experiment. While UNIFAC has fairly small errors, its R^2 IS smaller than the alchemical GAFF-based approaches (though higher than SMD). Additionally, for both SMD and UNIFAC (especially UNIFAC) compound coverage is not as good, so the size of the analyzed dataset is smaller. For UNIFAC, functional groups necessary to model a limited number of solutes and solvents were not available, a noted problem encountered when modeling solid-liquid equilibrium using UNIFAC.[63,64] With SMD, we were unable to model the solvents tert-butylcyclohexane and ethylamine, as they were not part of the solvent list in Gaussian 09.[65] These techniques simply do not cover all solute-solvent combinations examined here (Figure 2 and Tables 2—3 and SI Tables 1-6) because of their need for training data.

It's also important to understand how the performance of the different methods compares, so we plotted errors on each solute (across all solvents) for different methods. Specifically, Figure 4 shows the error for each solute from our standard alchemical GAFF approach on

the horizontal axis, versus the error on the same compounds with an alternate approach on the vertical axis. If both methods performed equally well or equally poorly, all data points would fall on the blue $x = y$ line. On the other hand, whenever the method showing on the vertical axis performs better than that on the horizontal axis, the data point will fall below $x = y$ (between $x = y$ and the x axis), and vise versa. In general there are far more points above the line than below, indicating that the GAFF approach typically outperforms the other approaches studied, except GAFF-DC.

Another way to examine our results is to use the experimental solubility for a solute in one or more specific solvents to determine an estimate of the fugacity term in equation 2, then compare that to the estimates of the fugacity term which we would have obtained if we had done the same with other solvents. The downside of this, however, is that we have to pick one or more particular experimental solubility to use to estimate the fugacity term. But this approach also allows us to examine whether the average error for a particular compound across all solvents might appear unusually large simply because of a large error for just one individual solvent. To investigate this, for each compound we selected one solvent to use as a test case, and used the remaining solvents as a "training set" to determine a best estimate of the fugacity term in equation 2. A schematic of this is shown in Figure 5. Here, we consider a specific solute A, solvated in solvents B, C and D, in turn. So first we pick solvent B as the test case, and use solvents C and D as the training solvents to determine the fugacity term. We then estimate the fugacity term as $\ln f_{\text{ave}} = 1/2\ (\ln f_{\text{C}} + \ln f_{\text{D}})$, where f_{C} and f_{D} are the fugacities as estimated from finding f_{C} such that

$$\ln\ c^{\text{C}}_{\text{A,expt}} = \ln\ c^{\text{C}}_{\text{A,calc}} = \ln\left(\frac{x^{\text{C}}_{\text{A}}}{v\,(T,p,x_{\text{C}})}\right) - \beta\mu^{\text{C,res}}_{\text{A}}\,(T,p,x_{\text{C}}) - \ln\ (RT) + \ln\ f^{\text{S}}_{\text{C}}\,(T,p) \tag{7}$$

where $c^{\text{C}}_{\text{A,expt}}$ is the experimental solubility for A in C, and $c^{\text{C}}_{\text{A,calc}}$ is the calculated solubility for A in C. We do the same to obtain f_{D}. We then calculate the error in the fugacity for our test solvent as $\delta \ln f_a = \ln f_{\text{ave}} - \ln f_a$ (where a denotes the selected solvent), so for example for solvent B, $\delta \ln f_{\text{B}} = \ln f_{\text{ave}} - \ln f_{\text{B}}$. This is a fair test, since B was not included when obtaining $\ln f_{\text{ave}}$. We can also calculate $\delta \ln f_{\text{C}}$ and $\delta \ln f_{\text{D}}$, though these will obviously underestimate of the true error in the calculated fugacity since solvents C and D were included in obtaining $\ln f_{\text{ave}}$. Still, we can determine the average or RMS error (RMSE) for compounds in the "training set". In this case, the RMSE on the training set is the RMS error across $\delta \ln f_{\text{C}}$ and $\delta \ln f_{\text{D}}$. We define the "training set error" as this RMSE. This whole process of examining a particular solute, picking a particular solvent as a test case, and evaluating training set and test set errors, can be iterated across all choices of solvent. In our example of three solvents, each of B, C, and D serve as the test case in turn. This allows us to obtain three different estimates of the test set error, and three estimates of the RMSE on the training set.

In what follows, the RMS error across all test cases is reported as the final error estimate for each particular solute dataset, and the average of the training set RMS errors is labeled the training set error. These are shown in Tables 2—3 (and SI Tables 1-6) and suggest that the GAFF-based alchemical results are the most accurate overall.

This procedure allows us a way to test how well our calculated solvation free energies do at yielding consistent estimates of the fugacity of each solid when coupled with the experimental solubility. If the solvation free energies were perfectly predicted (and the assumption of an infinitely dilute solute met) then all fugacity estimates for a particular solute ought to be identical, at least within experimental error. In this case, both the test and training set errors would be zero (within uncertainty).

Because the SMD calculations were done using one particular choice of basis set/level of theory, there is the possibility that SMD could appear to perform poorly solely because of that choice. Therefore, we repeated all of our calculations with an alternate approach. Particularly, SMD calculations were performed following the work of ref.[26,27], specifically using vacuum optimized geometries with single point energy calculations performed at the M06-2X/6-31G(d) level of theory/basis set. In addition, calculations were performed wherein the geometry was re-optimized in each solvent to assess the sensitivity to the solute geometry. Also, all of the single point energy calculations were repeated at the M06-2X/cc-pVTZ level of theory/basis set. Overall, we found that the later two changes had only a minute effect on the predictions. The results for these new SMD sets can be found in the Supporting Information.

4 Discussion

Our results indicate that alchemical free energy calculations based on molecular simulations can be a powerful approach for estimating relative solubilities of solutes in different solvent environments, with accuracies exceeding those of more highly-parameterized methods considered here such as UNIFAC and the SMD solvation model. This is especially interesting given that the force field employed, GAFF, has had no empirical tuning to reproduce solvation free energies in non-aqueous solvents, and no prior testing on relative solubilities that we are aware of. Thus, the techniques employed here may be of interest for prediction of relative solubilities.

Our work made one major assumption to simplify our calculations - that the solubility of a target solute in a particular solvent is low enough that the solution can be assumed to be ideal (i.e. that solute-solute interactions are negligible). When this is not the case, our general framework may still be useful, but additional simulations at different solute concentrations will be required in order to deal with non-ideality.

While our results agree fairly well with experimental relative solubility estimates, there is certainly room for improvement, and our data suggest that relative solubility measurements may provide a valuable (though indirect) source of experimental information on non-aqueous solvation free energies. Relative solubility data, then, may be an excellent tool to help improve force fields. One example of this is the performance of the GAFF-DC

parameters in this experiment – the accuracy of GAFF-DC appears superior to standard GAFF, despite the fact that it was developed for entirely orthogonal reasons.[55]

Previous work on solvation free energies has highlighted how errors can often be traced to particular functional groups.[31,66] Indeed, systematic errors for hydroxyl-containing compounds led to the development of GAFF-DC.[55] In principle, relative solubility studies should be able to highlight similar features – if particular solute functional groups are always poorly predicted, regardless of solvent environment, it likely means there is a systematic force field problem for that particular functional group. Thus, this will likely make a promising avenue for follow-up work. However, the present dataset of eight solutes in 53 solute-solvent combinations is not enough for us to be able to draw any meaningful conclusions about likely systematic errors. This especially true here, where systematic errors can result from *either* the solute or the solvent, whereas in hydration free energy calculations the solvent has already been carefully parameterized in its own right.

The recent reported work of Zhang et al.[31] also examined solvation free energies in non-aqueous solvents, comparing alchemical techniques with an empirical technique based on quantitative structure property relations, and quantum mechanical calculations with COSMO-RS. It found that the alchemical approach was not a clear winner, with the other two models in fact performing slightly better. Thus the authors concluded that further force field improvements are needed. While both studies rely on solvation free energies in non-aqueous solvents, the Zhang et al. work compared calculated solvation free energies with experimental values directly, whereas we calculate relative solubilities. While solubility measurements can be converted to estimates of the solvation free energy if vapor pressure data is available, this data is often not available[34–37] and is difficult to measure. Solvation free energies themselves can also be very difficult to measure, as discussed above. In contrast, solubilities are measured routinely and solubility measurements are abundant, even for drug-like compounds. Thus, the ability of this study to directly connect with solubility data is important.

While our results are far from indicating that further forcefield improvements are unwarranted, our method does outperform the other methods tested here. It seems likely that this may be precisely because of the relative abundance of solubility data compared to solvation free energy data. Specifically, the QSPR and COSMO-RS methods employed by Zhang et al.[31] have both been specifically fit at least in part to reproduce solvation free energies, and given the relatively small amount of solvation free energy data available, their training may have involved some of the same compounds on which they were tested. In contrast, the vast amount of solubility data available – and the lack of training of the methods tested on solubility data – means that the present test gives less of an advantage to empirical or semi-empirical methods.

The difficulty of measuring solvation free energies in general[34–37] has led the Statistical Assessment of Modeling of Proteins and Ligands (SAMPL) challenge to shift its solvation component to the calculation of partition/distribution coefficients[67] for SAMPL5, rather than solvation free energies which had previously formed the core of this part of the challenge in SAMPL1-4.[34–37] Like solubility data, partition/distribution data appears substantially more

straightforward to obtain experimentally than solvation free energies, and thus it may prove an even better opportunity for force field testing and development.

5 Conclusions

We used alchemical free energy calculations based on molecular simulations to calculate the relative solubilities of particular solutes solvated in a variety of different solvents, achieving average absolute errors of around 1 log unit in relative solubility.

We also compared our results with those obtained from SMD and UNIFAC solvation models applied to the essentially the same set, and found that our alchemical approach is more accurate in calculating relative solubilities on this set, especially when using the new GAFF-DC parameters for hydroxyl-containing compounds. Additionally, GAFF with alchemical techniques at present covers a wider range of chemical space than SMD and UNIFAC, in part because of the empirical tuning these techniques have required. We also found that overall, the GAFF-DC parameters out-perform standard GAFF parameters for relative solubilities in this set. It is interesting to note that relative solubility calculations - which essentially amount to calculating a difference in solvation free energies - may be a valuable source of experimental solvation data which can perhaps be used to further test and improve force fields for molecular simulations.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgement

S.L., S.C., K.H., and D.L.M. appreciate financial support from the National Institutes of Health (1R15GM096257-01A1, 1R01GM108889-01) and the National Science Foundation (CHE 1352608), and computing support from the UCI GreenPlanet cluster, supported in part by NSF Grant CHE-0840513. K.L.Y. and and A.S.P. acknowledge computing support by the Ohio Supercomputer Center and Miami University's Research Computing Support Group.

References

(2). Hopfinger AJ, Esposito EX, Llinaàs A, Glen RC, Goodman JM. Findings of the Challenge To Predict Aqueous Solubility. J. Chem. Inf. Model. 2009; 49:1–5. [PubMed: 19117422]

(2). Fredenslund A, Jones RL, Prausnitz JM. Group-contribution estimation of activity coefficients in nonideal liquid mixtures. AIChE J. 1975; 21:1086–1099.

(3). Hansen HK, Rasmussen P, Fredenslund A, Schiller M, Gmehling J. Vapor-liquid equilibria by UNIFAC group contribution. 5. Revision and extension. Ind. Eng. Chem. Res. 1991; 30:2352–2355.

(4). Wittig R, Lohmann J, Gmehling J. VaporLiquid Equilibria by UNIFAC Group Contribution. 6. Revision and Extension. Ind. Eng. Chem. Res. 2003; 42:183–188.

(5). Gmehling J, Li J, Schiller M. A modified UNIFAC model. 2. Present parameter matrix and results for different thermodynamic properties. Ind. Eng. Chem. Res. 1993; 32:178–193.

(6). Gmehling J, Lohmann J, Jakob A, Li J, Joh R. A Modified UNIFAC (Dortmund) Model. 3. Revision and Extension. Ind. Eng. Chem. Res. 1998; 37:4876–4882.

(7). Hewitt M, Cronin MTD, Enoch SJ, Madden JC, Roberts DW, Dearden JC. In SilicoPrediction of Aqueous Solubility: The Solubility Challenge. J. Chem. Inf. Model. 2009; 49:2572–2587. [PubMed: 19877720]

(8). Hughes LD, Palmer DS, Nigsch F, Mitchell JBO. Why are some properties more difficult to predict than others? A study of QSPR models of solubility, melting point, and Log P. J. Chem. Inf. Model. 2008; 48:220–232. [PubMed: 18186622]

(9). Votano JR, Parham M, Hall LH, Kier LB, Oloff S, Tropsha A, Xie Q, Tong W. Three new consensus QSAR models for the prediction of Ames genotoxicity. Mutagenesis. 2004; 19:365–377. [PubMed: 15388809]

(10). Ran Y, Jain N, Yalkowsky SH. Prediction of aqueous solubility of organic compounds by the general solubility equation (GSE). J. Chem. Inf. Comput. Sci. 2001; 41:1208–1217. [PubMed: 11604020]

(11). Hopfinger AJ, Esposito EX, Llinaàs A, Glen RC, Goodman JM. Findings of the Challenge To Predict Aqueous Solubility. J. Chem. Inf. Model.

(12). Palmer DS, McDonagh JL, Mitchell JBO, van Mourik T, Fedorov MV. First-Principles Calculation of the Intrinsic Aqueous Solubility of Crystalline Druglike Molecules. J. Chem. Theory Comput. 2012; 8:3322–3337. [PubMed: 26605739]

(13). Aguilar B, Onufriev AV. Efficient Computation of the Total Solvation Energy of Small Molecules via the R6 Generalized Born Model. J. Chem. Theory Comput. 2012; 8:2404–2411. [PubMed: 26588972]

(14). Chebil L, Chipot C, Archambault F, Humeau C, Engasser JM, Ghoul M, Dehez F. Solubilities Inferred from the Combination of Experiment and Simulation. Case Study of Quercetin in a Variety of Solvents. J. Phys. Chem. B. 2010; 114:12308–12313. [PubMed: 20715800]

(15). Oostenbrink C, Villa A, Mark AE, van Gunsteren WF. A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force-field parameter sets 53A5 and 53A6. J. Comp. Chem. 2004; 25:1656–1676. [PubMed: 15264259]

(16). Baker CM, Lopes PEM, Zhu X, Roux B, MacKerell AD Jr. Accurate Calculation of Hydration Free Energies using Pair-Specific Lennard-Jones Parameters in the CHARMM Drude Polarizable Force Field. J Chem Theory Comput. 2010; 6:1181–1198. [PubMed: 20401166]

(17). Horta BAC, Fuchs PFJ, van Gunsteren WF, Huünenberger PH. New Interaction Parameters for Oxygen Compounds in the GROMOS Force Field: Improved Pure-Liquid and Solvation Properties for Alcohols, Ethers, Aldehydes, Ketones, Carboxylic Acids, and Esters. J Chem Theory Comput. 2011; 7:1016–1031. [PubMed: 26606351]

(18). Harder E, Damm W, Maple J, Wu C, Reboul M, Xiang JY, Wang L, Lupyan D, Dahlgren MK, Knight JL, Kaus JW, Cerutti DS, Krilov G, Jorgensen WL, Abel R, Friesner RA. OPLS3: a force field providing broad coverage of drug-like small molecules and proteins. J Chem Theory Comput. 2015 acs.jctc.5b00864.

(19). Jaämbeck JPM, Lyubartsev AP. Update to the General Amber Force Field for Small Solutes with an Emphasis on Free Energies of Hydration. J Phys Chem B. 2014; 118:3793–3804. [PubMed: 24684585]

(20). Schnieders MJ, Baltrusaitis J, Shi Y, Chattree G, Zheng L, Yang W, Ren P. The Structure, Thermodynamics and Solubility of Organic Crystals from Simulation with a Polarizable Force Field. J. Chem. Theory Comput. 2012; 8:1721–1736. [PubMed: 22582032]

(21). Vega C, Noya EG. Revisiting the Frenkel-Ladd method to compute the free energy of solids: The Einstein molecule approach. J. Chem. Phys. 2007; 127:154113. [PubMed: 17949138]

(22). Noya EG, Conde MM, Vega C. Computing the free energy of molecular solids by the Einstein molecule approach: Ices XIII and XIV, hard-dumbbells and a patchy model of proteins. J. Chem. Phys. 2008; 129:104704. [PubMed: 19044935]

(23). Ferrario M, Ciccotti G, Spohr E, Cartailler T. Solubility of KF in water by molecular dynamics using the Kirkwood integration method. J. Chem. Phys. 2002; 117:4947.

(24). Bradley, J-C.; Friesen, B.; Mancinelli, J.; Bohinski, T.; Mirza, K.; Bulger, D.; Moritz, M.; Federici, M.; Rein, D.; Tchakounte, C.; Bradley, J-C.; Truong, H.; Neylon, C.; Guha, R.; Williams, A.; Hooker, B.; Hale, J.; Lang, A. Open Notebook Science Challenge: Solubilities of Organic Compounds in Organic Solvents. Nature Precedings preprint archive. 2010. Available from http://dx.doi.org/10.1038/npre.2010.4243.3, accessed Jan. 27, 2014

(25). Marenich AV, Cramer CJ, Truhlar DG. 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; 113:6378–6396. [PubMed: 19366259]

(26). Marenich AV, Cramer CJ, Truhlar DG. Performance of SM6, SM8, and SMD on the SAMPL1 Test Set for the Prediction of Small-Molecule Solvation Free Energies. J. Phys. Chem. B.

(27). Ribeiro RF, Marenich AV, Cramer CJ, Truhlar DG. Prediction of SAMPL2 aqueous solvation free energies and tautomeric ratios using the SM8, SM8AD, and SMD solvation models. J. Comput.-Aided Mol. Des. 2010; 24:317–333. [PubMed: 20358259]

(28). Duffy EM, Jorgensen WL. Prediction of properties from simulations: Free energies of solvation in hexadecane, octanol, and water. J Am Chem Soc. 2000; 122:2878–2888.

(29). Geerke DP, van Gunsteren WF. Force Field Evaluation for Biomolecular Simulation: Free Enthalpies of Solvation of Polar and Apolar Compounds in Various Solvents. ChemPhysChem. 2006; 7:671–678. [PubMed: 16514695]

(30). Garrido NM, Jorge M, Queimada AJ, Macedo EA, Economou IG. Using molecular simulation to predict solute solvation and partition coefficients in solvents of different polarity. Phys. Chem. Chem. Phys. 2011; 13:9155–9164. [PubMed: 21487617]

(31). Zhang J, Tuguldur B, van der Spoel D. Force Field Benchmark of Organic Liquids. 2. Gibbs Energy of Solvation. J. Chem. Inf. Model. 2015; 55:1192–1201. [PubMed: 26010106]

(32). Katritzky AR, Oliferenko AO, Oliferenko PV, Petrukhin R, Tatham DB, Maran U, Lomaka A, Acree WE Jr. A General Treatment of Solubility. 1. The QSPR Correlation of Solvation Free Energies of Single Solutes in Series of Solvents. J. Chem. Inf. Model. 2003; 43:1794–1805.

(33). Katritzky AR, Tulp I, Fara DC, Lauria A, Maran U, Acree WE. A General Treatment of Solubility. 3. Principal Component Analysis (PCA) of the Solubilities of Diverse Solutes in Diverse Solvents. J. Chem. Inf. Model. 2005; 45:913–923. [PubMed: 16045285]

(34). Guthrie JP. A Blind Challenge for Computational Solvation Free Energies: Introduction and Overview. J. Phys. Chem. B. 2009; 113:4501–4507. [PubMed: 19338360]

(35). Geballe MT, Skillman AG, Nicholls A, Guthrie JP, Taylor PJ. The SAMPL2 blind prediction challenge: introduction and overview. J Comput Aided Mol Des. 2010; 24:259–279. [PubMed: 20455007]

(36). Geballe MT, Guthrie JP. The SAMPL3 blind prediction challenge: transfer energy overview. J Comput Aided Mol Des. 2012; 26:489–496. [PubMed: 22476552]

(37). Guthrie JP. SAMPL4, a blind challenge for computational solvation free energies: the compounds considered. J Comput Aided Mol Des. 2014; 28:151–168. [PubMed: 24706106]

(38). Paluch AS, Maginn EJ. Predicting the Solubility of Solid Phenanthrene: A Combined Molecular Simulation and Group Contribution Approach. AIChE J. 2013; 59:2647–2667.

(39). Klimovich PV, Mobley DL. Predicting hydration free energies using all-atom molecular dynamics simulations and multiple starting conformations. J. Comput.-Aided Mol. Des. 2013; 24:307–316. [PubMed: 20372973]

(40). Kastenholz MA, Huünenberger PH. Computation of methodology-independent ionic solvation free energies from molecular simulations. I. The electrostatic potential in molecular liquids. J. Chem. Phys. 2006; 124:124106. [PubMed: 16599661]

(41). Kastenholz MA, Huünenberger PH. Computation of methodology-independent ionic solvation free energies from molecular simulations. II. The hydration free energy of the sodium cation. J. Chem. Phys. 2006; 124:224501. [PubMed: 16784292]

(42). Shirts, MR.; Mobley, DL. Biomolecular Simulations. Monticelli, L.; Salonen, E., editors. Humana Press; New York, NY: 2013. p. 271-311.

(43). Shirts, MR.; Mobley, DL.; Brown, SP. Drug Design: Structure and Ligand-based Approaches. Merz, KM., Jr; Ringe, D.; Reynolds, CH., editors. Cambridge University Press; 2010.

(44). Hawkins PCD, Skillman AG, Warren GL, Ellingson BA, Stahl MT. Conformer Generation with OMEGA: Algorithm and Validation Using High Quality Structures from the Protein Databank and Cambridge Structural Database. J. Chem. Inf. Model. 2010; 50:572–584. [PubMed: 20235588]

(45). Jakalian A, Bush BL, Jack DB, Bayly CI. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: I. Method. J. Comput. Chem. 2000; 21:132–146.

(46). Jakalian A, Jack DB, Bayly CI. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. J. Comp. Chem. 2002; 23:1623–1641. [PubMed: 12395429]

(47). Wang J, Wang W, Kollman PA, Case DA. Automatic atom type and bond type perception in molecular mechanical calculations. J. Mol. Graphics Modell. 2006; 25:247–260.

(48). Wang J, Wolf RM, Caldwell JW, Kollman PA, Case DA. Development and testing of a general AMBER force field. J. Comput. Chem. 2004; 25:1157–1174. [PubMed: 15116359]

(49). Sousa da Silva AW, Vranken WF. ACPYPE - AnteChamber PYthon Parser interfacE. BMC Res Notes. 2012; 5:367–374. [PubMed: 22824207]

(50). MartÍnez L, Andrade R, Birgin EG, MartÍnez JM. PACKMOL: A package for building initial configurations for molecular dynamics simulations. J. Comput. Chem. 2009; 30:2157–2164. [PubMed: 19229944]

(51). Klimovich PV, Mobley DL. Predicting hydration free energies using all-atom molecular dynamics simulations and multiple starting conformations. J. Comput.-Aided Mol. Des. 2010; 24:307–316. [PubMed: 20372973]

(52). Mobley DL, Liu S, Cerutti DS, Swope WC, Rice JE. Alchemical prediction of hydration free energies for SAMPL. J. Comput.-Aided Mol. Des. 2012; 26:551–562. [PubMed: 22198475]

(53). Liu S, Wu Y, Lin T, Abel R, Redmann JP, Summa CM, Jaber VR, Lim NM, Mobley DL. Lead optimization mapper: automating free energy calculations for lead optimization. J. Comput.-Aided Mol. Des. 2013; 27:755–770. [PubMed: 24072356]

(54). Liu S, Wang L, Mobley DL. Is ring breaking feasible in relative binding free energy calculations? J. Chem. Inf. Model. 2015; 55:727–735. [PubMed: 25835054]

(55). Fennell CJ, Wymer KL, Mobley DL. A Fixed-Charge Model for Alcohol Polarization in the Condensed Phase, and Its Role in Small Molecule Hydration. J. Phys. Chem. B. 2014; 118:6438–6446. [PubMed: 24702668]

(56). Mobley DL, Dumont É, Chodera JD, Dill KA. Comparison of charge models for fixed-charge force fields: Small-molecule hydration free energies in explicit solvent. J. Phys. Chem. B. 2007; 111:2242–2254. [PubMed: 17291029]

(57). Mobley DL, Wymer KL, Lim NM, Guthrie JP. Blind prediction of solvation free energies from the SAMPL4 challenge. J Comput Aided Mol Des. 2014; 28:135–150. [PubMed: 24615156]

(58). It is useful to put the calculation of $\Delta G^{\alpha,\infty}_{1,\mathrm{solv}}$ using SMD in this study in the context/language of a conventional molecular simulation free energy calculation. When coupling/decoupling a single solute molecule when performing a molecular simulation free energy calculation, the SMD calculations here assume that the simulation box is approximately the same size when the solute is fully coupled and fully decoupled (i.e., $\Delta V^{\alpha,\infty}_{1,\mathrm{solv}} = 0$ such that the molar concentration of the solute is the same in both states.) This results in $\Delta G^{\circ}_{\mathrm{conc}} = 0$.

(59). Zhao Y, Truhlar DG. The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other functionals. Theor. Chem. Account. 2008; 120:215–241.

(60). Cramer, CJ. Essentials of Computational Chemistry. John Wiley & Sons Ltd; Chichester, West Sussex, England: 2002.

(61). Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Scalmani G, Barone V, Mennucci B, Petersson GA, Nakatsuji H, Caricato M, Li X, Hratchian HP, Izmaylov AF, Bloino J, Zheng G, Sonnenberg JL, Hada M, Ehara M, Toyota K, Fukuda R, Hasegawa J, Ishida M, Nakajima T, Honda Y, Kitao O, Nakai H, Vreven T, Montgomery JA, Peralta JE, Ogliaro F, Bearpark M, Heyd JJ, Brothers E, Kudin KN, Staroverov VN, Kobayashi R, Normand J, Raghavachari K, Rendell A, Burant JC, Iyengar SS, Tomasi J, Cossi M, Rega N, Millam JM, Klene M, Knox JE, Cross JB, Bakken V, Adamo C, Jaramillo J, Gomperts R, Stratmann RE, Yazyev O, Austin AJ, Cammi R, Pomelli C, Ochterski JW, Martin RL, Morokuma K, Zakrzewski VG, Voth GA, Salvador P, Dannenberg JJ, Dapprich S, Daniels AD, Farkas Ö, Foresman JB, Ortiz JV, Cioslowski J, Fox DJ. Gaussian 09, Revision B.01. 2009

(62). Paluch AS, Vitter CA, Shah JK, Maginn EJ. A comparison of the solvation thermodynamics of amino acid analogues in water, 1-octanol and 1-n-alkyl-3-methylimidazolium bis(trifluoromethylsulfonyl)imide ionic liquids by molecular simulation. J. Chem. Phys. 2012; 137:184504. [PubMed: 23163380]

(63). Gracin S, Brinck T, Rasmuson AC. Prediction of Solubility of Solid Organic Compounds in Solvents by UNIFAC. Ind. Eng. Chem. Res. 2002; 41:5114–5124.

(64). Tanveer S, Hao Y, Chen C-C. Introduction to Solid-Fluid Equilibrium Modeling. Chem. Eng. Progress. 2014; 110:37–47.

(65). Gaussian 09 User's Reference – SCRF. {http://www.gaussian.com/g_tech/g_ ur/k_scrf.htm}, (accessed August 5, 2015)

(66). Mobley DL, Bayly CI, Cooper MD, Shirts MR, Dill KA. Small Molecule Hydration Free Energies in Explicit Solvent: An Extensive Test of Fixed-Charge Atomistic Simulations. J. Chem. Theory Comput. 2009; 5:350–358. [PubMed: 20150953]

(67). Drug Design Data Resource, SAMPL5. 2016. https://drugdesigndata.org/ about/sampl5, Accessed Feb. 5, 2016.

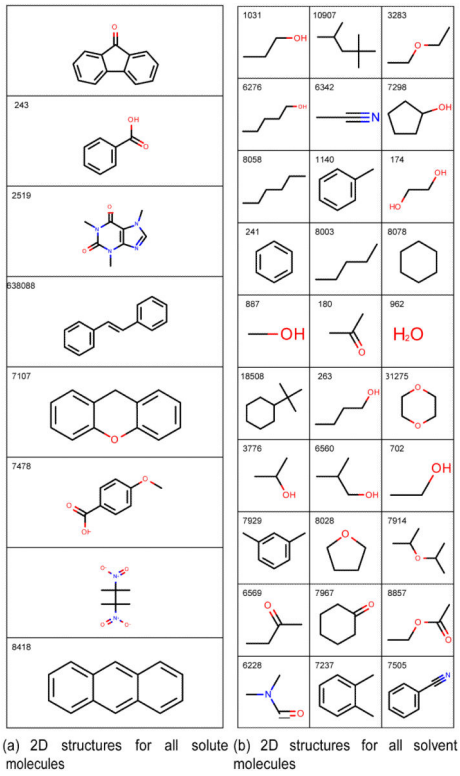
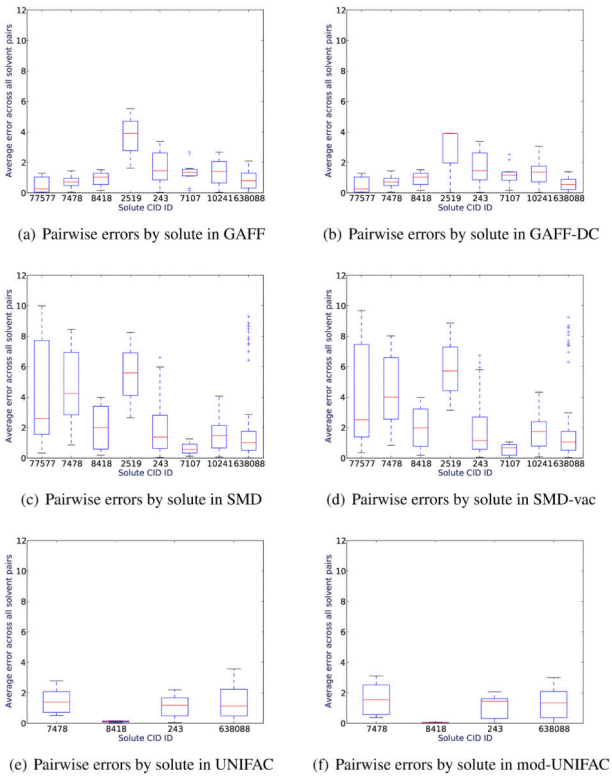


Figure 1.

2D structures for all solute and solvent molecules. The corresponding CIDs are showing on the left upper corner of each panel.



(a) Pairwise errors by solute in GAFF (b) Pairwise errors by solute in GAFF-DC

(c) Pairwise errors by solute in SMD (d) Pairwise errors by solute in SMD-vac

(e) Pairwise errors by solute in UNIFAC (f) Pairwise errors by solute in mod-UNIFAC

Figure 2.

The average error in $\ln\left(\frac{c_1^{\alpha}}{c_1^{\zeta}}\right)$ by solute, across all possible solvent pairs for each solute for the different methods considered (a-f). The vertical axis shows the error in the log ratio (unitless), and the horizontal axis shows the solvent considered. The plot is a box and whisker plot, with the box showing the lower and upper quartiles of the data, and the red line marking the median. The whiskers show the range of the data.

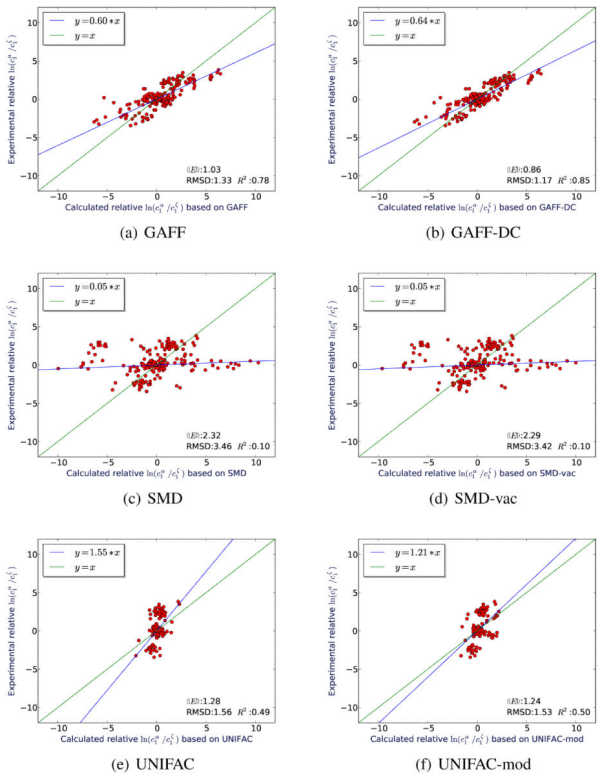


Figure 3.

Comparison of calculated relative solubilities with the experimental relative solubilities for all solute-solvent pairs and all methods.

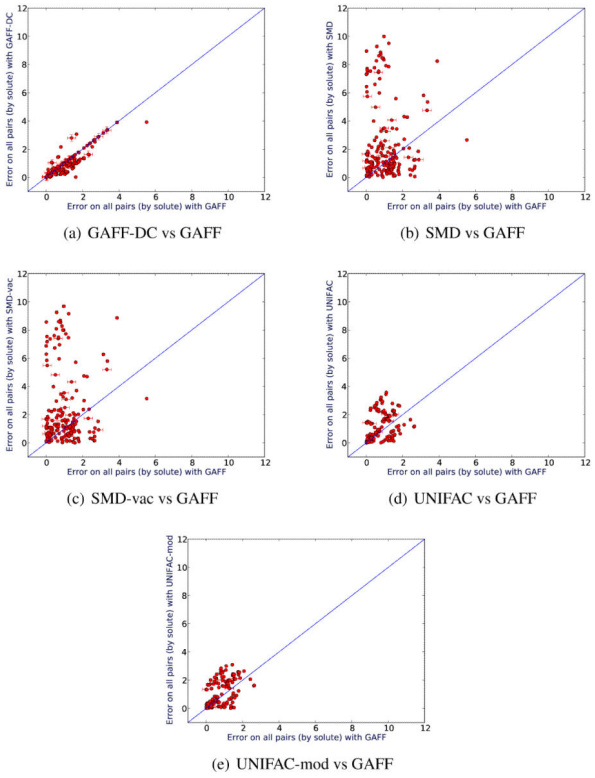


Figure 4.

Comparison of errors for different methods for each solute in all pairs of solvents. The $x = y$ line divides the figure into two regions, the left-top region and right-bottom region. If a particular datapoint is in the left-top region, then the method shown on the x-axis performs better for that particular case, and if the point is in the right-bottom region, the method shown on the y-axis performs better.

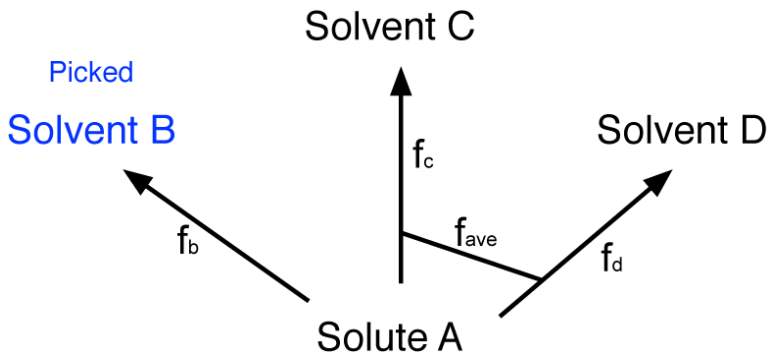


Figure 5.

An example of how we calculate the test and training set errors. Here, we examine a particular solute (A) in three solvents (B-D). As discussed in the text, we pick one particular solvent (B) in which to "predict" the solubility of the compound, and use the other solvents to calculate the best estimate of the fugacity (f_{ave}) of the solute by comparison to the experimental solubilities. From this estimate, we can then calculate solubility of the solute in solvent B, or (nearly equivalently) the fugacity term for B. This allows us to calculate the error in the fugacity for our test case, B (the test set error), and the error in the fugacity for the other cases (the training set error).

Table 1

Solute-solvent pairs studied here. Here, we use PubChem compound identifiers to track our compounds as these eliminate confusion due to different naming conventions, and also are more convenient for some of our tools to handle. Traditional compound names are listed as well.

Solute ID in Solvent ID	Solute Name	Solvent Name
10241 in 1031	9-fluorenone	1-propanol
10241 in 10907	9-fluorenone	2,2,4-trimethylpentane
10241 in 3283	9-fluorenone	diethyl ether
10241 in 6276	9-fluorenone	1-pentanol
10241 in 6342	9-fluorenone	acetonitrile
10241 in 7298	9-fluorenone	cyclopentanol
10241 in 8058	9-fluorenone	n-hexane
243 in 1140	benzoic acid	toluene
243 in 174	benzoic acid	ethylene glycol
243 in 241	benzoic acid	benzene
243 in 6342	benzoic acid	acetonitrile
243 in 8003	benzoic acid	pentane
243 in 8058	benzoic acid	n-hexane
243 in 8078	benzoic acid	cyclohexane
243 in 887	benzoic acid	methanol
2519 in 180	caffeine	acetone
2519 in 887	caffeine	methanol
2519 in 962	caffeine	water
638088 in 1031	trans-stilbene	1-propanol
638088 in 10907	trans-stilbene	2,2,4-trimethylpentane
638088 in 1140	trans-stilbene	toluene
638088 in 18508	trans-stilbene	tert-butylcyclohexane
638088 in 241	trans-stilbene	benzene
638088 in 263	trans-stilbene	1-butanol
638088 in 31275	trans-stilbene	1,4-dioxane
638088 in 3776	trans-stilbene	2-propanol
638088 in 6276	trans-stilbene	n-pentanol
638088 in 6560	trans-stilbene	isobutyl alcohol
638088 in 702	trans-stilbene	ethanol
638088 in 7929	trans-stilbene	3-xylene
638088 in 8028	trans-stilbene	tetrahydrofuran
638088 in 8058	trans-stilbene	n-hexane
638088 in 887	trans-stilbene	methanol
7107 in 3283	xanthene	diethyl ether
7107 in 702	xanthene	ethanol
7107 in 7914	xanthene	isopropyl ether
7107 in 8078	xanthene	cyclohexane

Solute ID in Solvent ID	Solute Name	Solvent Name
7107 in 887	xanthene	methanol
7478 in 31275	4-methoxybenzoic acid	1,4-dioxane
7478 in 6276	4-methoxybenzoic acid	n-pentanol
7478 in 6560	4-methoxybenzoic acid	isobutyl alcohol
7478 in 8028	4-methoxybenzoic acid	tetrahydrofuran
77577 in 180	2,3-dimethyl-2,3-dinitrobutane	acetone
77577 in 31275	2,3-dimethyl-2,3-dinitrobutane	1,4-dioxane
77577 in 6342	2,3-dimethyl-2,3-dinitrobutane	acetonitrile
77577 in 6569	2,3-dimethyl-2,3-dinitrobutane	methylethyl ketone
77577 in 7967	2,3-dimethyl-2,3-dinitrobutane	cyclohexane
77577 in 8028	2,3-dimethyl-2,3-dinitrobutane	tetrohydrofuran
77577 in 8857	2,3-dimethyl-2,3-dinitrobutane	ethyl acetate
8418 in 6228	anthracene	dimethylformamide
8418 in 7237	anthracene	2-xylene
8418 in 7505	anthracene	benzonitrile
8418 in 7929	anthracene8	3-xylene

Table 2

GAFF errors: in log S, by solute, across all pairs of solvents for each solute; and in ln f_C, across training and test sets for each solute

Solute ID	data size	average error, all pairs	average absolute error, all pairs	training set error	test set error
77577	7	−0.057(1)	0.561(1)	0.5(1)	0.6(4)
7478	4	−0.3859(9)	0.7289(9)	0.47(3)	0.7(1)
8418	4	−0.3284(6)	0.9130(6)	0.59(3)	0.8(1)
2519	3	−2.5980(5)	3.6839(5)	1.84(9)	3.5(3)
243	8	−0.1839(2)	1.6187(2)	1.26(6)	1.5(2)
7107	5	−0.8687(1)	1.3623(1)	0.95(5)	1.2(2)
10241	7	0.65786(8)	1.29910(8)	0.98(5)	1.2(2)
638088	15	0.20307(2)	0.83181(2)	0.68(4)	0.7(2)
Average		0.04012(4)	1.03129(4)	0.852(2)	1.089(8)

Table 3

GAFF-DC errors: in log S, by solute, across all pairs of solvents for each solute; and in ln f_C, across training and test sets for each solute

Solute ID	data size	average error, all pairs	average absolute error, all pairs	training set error	test set error
77577	7	−0.057(1)	0.561(1)	0.5(1)	0.6(4)
7478	4	−0.3859(9)	0.7289(9)	0.47(3)	0.7(1)
8418	4	−0.3284(6)	0.9130(6)	0.59(3)	0.8(1)
2519	3	−2.5980(5)	2.6114(5)	1.31(7)	2.8(3)
243	8	−0.1839(2)	1.6187(2)	1.26(6)	1.5(2)
7107	5	−0.8146(1)	1.2005(1)	0.83(4)	1.1(1)
10241	7	0.76278(8)	1.40036(8)	1.05(9)	1.2(3)
638088	15	0.18366(2)	0.53873(2)	0.44(3)	0.5(1)
Average		0.04366(4)	0.86389(4)	0.751(2)	0.972(9)