UC Office of the President

Recent Work

Title

Infinite Dilution Activity Coefficients as Constraints for Force Field Parametrization and Method Development.

Permalink

Journal

ISSN

Authors

Duarte Ramos Matos, Guilherme
Calabrò, Gaetano
Mobley, David L

Publication Date

DOI

Copyright Information

Peer reviewed

Infinite dilution activity coefficients as constraints for force field parameterization and method development

Guilherme Duarte Ramos Matos,[†] Gaetano Calabrò,[‡] and David L. Mobley[*,¶,†]

†*Department of Chemistry, University of California, Irvine, CA*

‡*OpenEye Scientific Software, Santa Fe, NM*

¶*Department of Pharmaceutical Sciences, University of California, Irvine, CA*

E-mail: dmobley@uci.edu

Abstract

Molecular simulations begin with an underlying energy model or force field, and from this, can predict diverse physical properties. However, force fields were often developed with relatively limited datasets, yet accuracy for diverse properties across a broad chemical space is desirable, so tests of such accuracy are particularly important. Here, to this end, we calculate 237 infinite dilution activity coefficients (IDACs), comparing with experimental values from NIST's ThermoML database. We found that calculated IDAC values correlate strongly with experiment (Pearson R of 0.92 ± 0.01), and allow us to identify specific functional groups which appear to present challenges to the force field employed. One potentially valuable aspect of IDACs, as compared to solvation free energies which have been frequently employed as force field tests, is that the same molecules serve both as solutes and solvents in different cases, allowing us to ensure force fields are not overly tuned to one particular environment or solvent.

1 Introduction

Molecular simulations see widespread use in calculating various physical and biological properties of interest, with a key goal being predictive molecular design — the ability to design molecules which have desired physical properties and activity (e.g. adequate solubility while binding to a biological target, for example). Such simulations begin with an underlying energy model, or *force field*, giving the energy and forces as a function of the atomic positions. This energy model, then, determines the accuracy of calculated properties, thus considerable attention is paid to force field parameterization.

Force fields have long been parameterized based on both quantum mechanical calculations (e.g. for torsional potentials and in some cases electrostatics)[1] and condensed phase properties, including density, heat of vaporization, and others[1] including mixture data.[2] Solvation free energies of molecules in water (hydration free energies, for gas-to-water transfer) and solvation free energies in other solvents have become a standard test of force fields, and thus have comprised a component of various blind prediction challenges like the Statistical Assessment of Modeling of Proteins and Ligands (SAMPL) challenges.[3–7] This is in part because they can be calculated quite precisely, allowing quantitative tests of force field accuracy,[8,9] and even identification and repair of specific systematic errors.[10] Other force fields have actually used such free energies as part of their parameterization — especially the GROMOS family of force fields,[1,11] but also more recent versions of OPLS.[12,13]

While solvation free energies have proved valuable to the field, these are rather difficult to measure experimentally and few if any such experiments are still conducted.[3] In contrast, infinite dilution activity coefficients (IDACs) are more readily measured experimentally,[14–17] are available in relatively large quantities in databases such as NIST's ThermoML and DECHEMA's DETherm, and can yield insight into solvation by different solvents. Specifically, IDACs (also represented by γ^∞) are related to solvation free energies by:[18–20]

$$\gamma_i^\infty = \exp\left(\frac{\Delta G_i^{\text{solv}} - \Delta G_i^{\text{self solv}}}{k_B T}\right) \cdot \frac{\rho_{\text{molar}}^{\text{solvent}}}{\rho_{\text{molar}}^{\text{pure solute}}} \tag{1}$$

where ΔG_i^{solv} is the solvation free energy of a solute i, $\Delta G_i^{\text{self solv}}$ is the solvation free energy of a solute i in its bulk phase, k_B is the Boltzmann constant, T is the absolute temperature, and ρ_{molar} is the molar density of component of the mixture.

In addition to their experimental tractability, IDAC values have at least two additional potential advantages over solvation free energies. First, IDACs for a given solute of interest can be obtained in different solvents, allowing the potential to explore how well a force field represents molecules both as solutes and as solvents to ensure both contexts are treated accurately. Second, molecules become polarized when transferred from gas phase to water and hydration free energies with conventional force fields (excepting polarizable force fields) might not be able to describe this phenomenon well; parameterization to hydration free energies could even build in systematic error resulting from lack of treatment of polarization. This may be particularly important; while an IDAC calculation also involves a transfer between environments, it is a transfer between two condensed phases, which usually is associated to a much smaller change in polarization of the solute molecule in comparison to transfer from the gas phase, as in the case of hydration free energies. IDAC values thus could potentially be an even better way to test how a force field represents a condensed-phase environment than hydration free energies, and potentially a better source of parameterization data.

Here, we focus on the calculation of IDAC values using an automated protocol for the estimation of solvation free energies, and comparison of these values with experimental results given in NIST's ThermoML database. While IDAC values have been calculated before using molecular simulations, especially in the chemical engineering literature,[19,21] to our knowledge this is the first work calculating these on a large scale as a test of general-purpose small molecule force fields, and marks our first effort towards beginning to use these in force field parameterization. Additionally, this may also be the first study on IDACs outside the chemical engineering literature, and the first to use them to test general-purpose small

3

molecule force fields.

The overall usefulness of solvation free energies for molecular modeling has led to the availability of several databases of these values, such as FreeSolv,[22,23] ATB,[24,25] MNSol[26] and CompSol,[27] ΔG^{solv} and ΔG^{hyd}. We hope this work may spur a similar interest in infinite dilution activity coefficients.

2 Methods

In this work, we focus on automatically calculating a large number of infinite dilution activity coefficients for comparison with reference values automatically obtained from NIST's ThermoML database. Our focus is not on optimizing our protocol for calculation of these coefficients; rather, we use a protocol previously optimized for calculation of solvation/hydration free energies, and simply apply it to calculate the solvation free energies needed for IDAC calculations without further optimization. Thus, as we discuss further below, additional optimization may be possible.

Here, we briefly recap several key concepts behind activity coefficients and then detail the computational methods employed.

2.1 Conceptual background

Most readers will likely remember activity coefficients from introductory physical chemistry, where these are used to assess how, essentially, the effective concentration of a molecule in solution deviates from ideality. The activity of a component in a system, or its effective concentration, is the product between the component's activity coefficient and its concentration and is related to the chemical potential of a component in a mixture (μ_i) (Eq. 2). Thus, infinite dilution activity coefficients (IDACs) tell us how far an infinitely dilute mixture is from ideal solution conditions[14,15,28–30] and they are of considerable experimental and theoretical interest.[15,18,19] Deviations from ideality indicate whether a solvent is particularly good or

particularly poor for a given solute.

The chemical potential of a component is related to its activity coefficient and concentration via

$$\mu_i = \mu_i^0 + RT \ln \gamma \cdot \frac{[i]}{[i_0]}, \tag{2}$$

where γ is the activity coefficient, μ_i^0 is the standard chemical potential of the component, $[i]$ and $[i_0]$ are its concentration and standard concentration, R is the ideal gas constant, and T is the absolute temperature. If $\gamma = 1$, the mixture is ideal; if $\gamma > 1$ or $0 < \gamma < 1$, the mixture behaves non-ideally.

There are different ways to express activity coefficients, including in terms of concentration, mole fraction, or partial pressure. Here, we focus on activity coefficients expressed in terms of mole fraction (χ) henceforth, but it is worth remembering the connections to concentration and other forms.

Activity coefficients can also be defined relative to different reference states. Here, we define activity coefficients with reference to an ideal solution in the sense of Raoult's law, where, for each component in a mixture, $\gamma_i \rightarrow 1$ as $\chi_i \rightarrow 1$.[31] This defines the activity coefficient as 1 for the pure solution, and is called the Lewis-Randall standard state. Other reference states are also commonly employed. For example, a common textbook definition uses an ideal dilute solution as a reference state, in which case the activity coefficient is 1 for a solute at infinite dilution. Here, however, we use the Lewis-Randall reference state as it is the state employed by the database of experimental values we utilize.

Measurement of IDAC can be conducted via a variety of techniques. IDACs are related to the slope of isothermal pressure-composition phase diagrams when the mole fraction tends to zero, and are proportional to the Henry's Law constant,[32] thus their measurement of depends on factors such as the volatility of solvent and solute.[30] Techniques such as gas-liquid chromatography,[30,33] high-performance liquid chromatography[33] and differential ebulliometry[30,33,34] are traditionally used to measure activities in extremely dilute systems at

varying concentrations, leading to the infinite dilution activity coefficient by extrapolation.[33] There has been considerable interest in predicting these coefficients[18–21,30,33–42] due to their use in phase equilibria studies in chemical engineering applications.[14,15,29,36]

2.2   Computational methods

We obtained experimental activity coefficients at infinite dilution from ThermoML,[43–45] an XML-based system for storage and exchange of thermochemical data. ThermoML was accessed on July 27, 2017 using `thermopyl`,[46] a Python tool that allows interaction with the database and provides access via a Pandas `Dataframe`. We made a search for IDACs of organic compounds containing less than 40 heavy atoms at temperatures between 250 K and 400 K. All the activity coefficients were obtained approximately at 101 kPa. The search was restricted to molecules containing no elements other than C, O, N, F, P, S, and Cl. The heaviest solute molecule of the set was hexadecane (226.44 Da) and the lightest was methanol (32.04 Da). The heaviest solvent molecule was tetradecanoic acid (228.37 Da) and the lightest was water (18.02 Da). Most molecules were fairly rigid with less than three rotatable bonds, but a few, such as hexadecane and undecane, had up to 13 rotatable bonds.

We found 263 coefficients but limited our study to 237 coefficients. The reduced set size resulted from problems building the simulation boxes for some systems with solutes or solvents with long chains, as well as parameterization issues for some tertiary amines. The final set contains a variety of combinations of a moderate number of different solvents and solutes. This allows us to look for trends in accuracy both as a function of solute and as function of solvent.

All solvation free energy calculations were performed using now relatively standard alchemical free energy calculations described further below, but automated via the OpenEye Orion cloud computing platform. The calculations could have been done on local computing resources using an identical protocol, but Orion allowed for higher throughput.

Setup of calculations began with processing the solute and solvent names from the data

obtained from ThermoML. From names, SMILES strings were generated using OpenEye's `OEChem` toolkits, and stored as `OEMol` objects,[47] with one `OEMol` for each solvation free energy calculation to be done (i.e. one for calculation of the solvation free energy of each solute in pure solute, and one for calculation of solvation free energy of the solute in pure solvent). In each case we attached the SMILES string of the solvent (generated with `OEChem`) to the `OEMol` for the solvent, along with the target temperature and pressure for each simulation (as these were required by the Orion workflow we constructed) and then output the resulting set of molecules to an OpenEye binary file (`.oeb`) for use on Orion.

The Orion workflow then conducted solvation free energy calculations from these input files in a straightforward manner, ultimately using Yank[48] to run free energy calculations as further detailed below. Before input into Yank, however, simulation boxes were built and parameters were assigned. Specifically, starting geometries for simulation boxes were built (in PDB format) from the solute and the specified solvent (as indicated by SMILES strings attached to the input molecule) using the OpenEye toolkits to generate molecular structures and conformers, and `PACKMOL` (version 17.221) to build boxes consisting of the solvated systems.[49] Force field parameters were then assigned via Antechamber and Ambertools (version 16.16.0), using the GAFF 1.8 small molecule force field[50] and AM1-BCC charges[51,52] (the latter as assigned by the OpenEye toolkits, version 2018.2.1) to describe solvents and solutes, with the exception of water, which was modeled by using TIP3P.[53] Once parameterized, the resulting systems were stored as `ParmEd`[54] (version 2.7.3) objects and attached to the OpenEye data record to progress through the workflow.

Following parameterization, equilibration stages were run using OpenMM[55–57] (version 7.1.1), followed by production free energy calculations done with Yank (version 0.20.1), using protocols that are now relatively standard (e.g. as in[23]). Nonbonded interactions were calculated for all inter-atomic distances under a cutoff of 9 $\mathring{A}$, with a long range dispersion correction applied to the energy and pressure.[58] Such corrections are, in isotropic systems, enough to typically make hydration/solvation free energies independent of cutoff well before

$9\mathring{A}$ [59,60] when combined with lattice-sum electrostatics. Electrostatic interactions were computed using particle mesh Ewald (PME). [61,62] In our simulations, bonds involving hydrogen were constrained. All simulations were conducted at the target temperature and pressure associated with the ThermoML data for the experiment, as provided by our input files.

Equilibration was done with standard equilibration workflows ("floes") from OpenEye on Orion, utilizing OpenMM. These consisted of 0.2 ns of NVT equilibration for each system with the solute heavy atoms harmonically restrained with $2.0 kcal/(mol \cdot \mathring{A}^2)$ spring constants, then an additional 0.2 ns of NPT equilibration with $0.1 kcal/(mol \cdot \mathring{A}^2)$ spring constant restraints on the same atoms, to equilibrate solvent molecules and bring the system to the correct temperature and pressure. Yank simulations were then started from these equilibrated structures, but additional data from the free energy simulations was then discarded as equilibration as discussed below.

Solvation free energies were computed with consistent protocol for every solute-solvent combination in an effort to test the overall accuracy of this approach when applied at scale; no attempt was made to ensure the protocol was ideally adapted to each individual case examined. In the approach employed here, each Hamiltonian replica exchange simulation run using Yank had 1000 iterations of 500 MD steps of 2 femtoseconds each at each of the 20 λ values we employed (λ values given in the Supporting Information), totaling one nanosecond per replica. Every iteration, exchanges were proposed between across all replicas using Gibbs sampling (rather than just pairwise swaps) to accelerate mixing, [63] as is standard with this protocol in Yank. Solvation free energies were estimated with the Multistate Bennett Acceptance Ratio [64] (MBAR), an extension of the Bennett Acceptance Ratio [65] that considers the overlap between a given state and all the others in the path between the end states, as provided by Yank. MBAR is the default free energy estimator in Yank, and the one we use. It is statistically optimal in a particular sense, [64] and also performs consistently well, [66] supporting this choice. Free energies and uncertainties are calculated via MBAR, following automatic equilibration detection within Yank and subsampling data

at intervals of the statistical inefficiency of the timeseries,[67] as is standard in Yank. Our systems were already equilibrated before Yank simulations began; however, in some cases Yank concluded discarding additional data to equilibration was needed. Ultimately this resulted in analysis of a varying amount of data (and significantly varying uncertainties) for each solute/solvent combination since automatic equilibration detection discarded varying amounts of data, *and* the autocorrelation time was different in each case. Since different solute-solvent combinations also have different overlap between λ values even given the same protocol, that also drove variation in uncertainties from case to case.

The protocol applied here for calculation of solvation free energies typically works relatively well/reliably for calculation of hydration free energies (e.g.[23]), even without Hamiltonian exchange. However, further optimizations of the protocol might be possible to improve efficiency and/or convergence. Additionally, as Hamiltonian exchange is additionally employed here with only 1000 iterations, mixing across all replicas might not be ideal — some solute-solvent combinations might pose particular challenges. An alternate approach to explore in future studies might be to run Yank until a specified statistical precision is reached for each calculation (which is available as an option); however, this would make the total compute time required unclear until the project's completion. Here, however, our focus here was instead on a large-scale test of the present fully-automated approach, saving optimizations of the protocol for future work.

For calculation of the density term, 5ns Langevin dynamics simulations were run using OpenMM,[56,57] with otherwise the same settings, to obtain densities of pure solvents and pure solutes to calculate the ratio in Eq. 1.

3   Results

IDACs (γ^∞), as defined in the Introduction, tell us how far from ideality a mixture is when the concentration of the solute is infinitely small. They are widely used as input for engineering

models, such as for prediction of liquid-vapor equilibria, and they can be calculated from solvation free energies (Eq. 1).[18–20] The natural logarithm of γ^∞ is proportional to the difference between the free energy of solvation of a solute i in a given solvent ($\Delta G_i^{\mathrm{solv}}$) and the free energy of solvation of the solute molecule in its pure bulk phase ($\Delta G_i^{\mathrm{self\ solv}}$, free energy of "self-solvation") plus the ratio between the densities of the solvent and the pure solute:

$$k_B T \ln \gamma_i^\infty = \Delta G_i^{\mathrm{solv}} - \Delta G_i^{\mathrm{self\ solv}} + k_B T \ln \left(\frac{\rho_{\mathrm{molar}}^{\mathrm{solvent}}}{\rho_{\mathrm{molar}}^{\mathrm{pure\ solute}}} \right) \tag{3}$$

where k_B is the Boltzmann constant and T is the absolute temperature.

Here, we calculated the solvation free energies using MBAR[64] and compared to experimental values as shown in Figure 1.
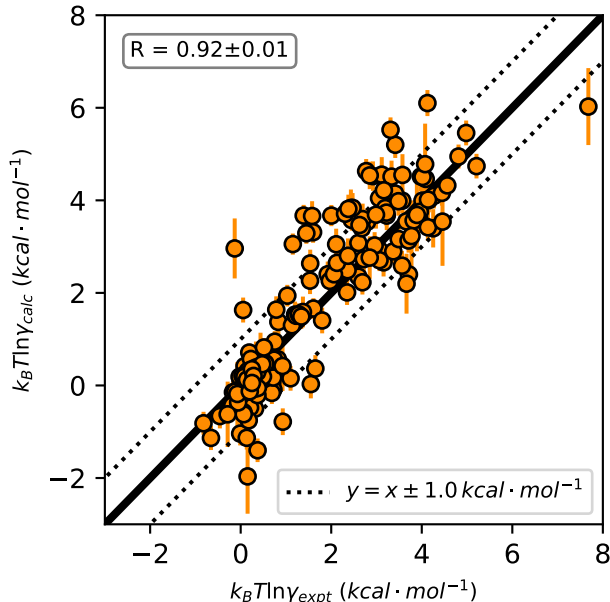

Figure 1: Calculated versus experimental $k_B T \ln \gamma^\infty$ for 237 solute – solvent pairs taken from ThermoML. Calculated values are on the vertical axis and experimental on the horizontal.

We found an average error (in free energy units, as in Equation 3) of 0.06 ± 0.05 kcal $\cdot$ mol^{-1}, a root-mean-square (RMS) error of 0.73 ± 0.05 kcal $\cdot$ mol^{-1}, an average unsigned error of 0.48 ± 0.04 kcal $\cdot$ mol^{-1}, a Kendall τ value of 0.67 ± 0.03, and a Pearson R value of

0.92 ± 0.01.

Here, we choose to report our error in free energy units rather than the dimensionless form used by γ itself in order to allow us to set our accuracy in the context of that typically seen for solvation free energies. In previous work on hydration free energies in a blind challenge context, the best half of methods participating typically yielded RMS errors around 1.5 kcal/mol,[60] consistent with retrospective studies such as that on FreeSolv[23] with somewhat larger errors seen for calculation of logD[60] and somewhat smaller for logP.[68] Thus it may be surprising that the RMS error here is actually *lower* than that on hydration free energies, despite the fact that these calculations involve a difference in solvation free energies as well as an excess density term. Perhaps this lower-than-expected error suggests a cancellation of error between solvents, or that hydration free energy calculations pose particular challenges.

Given that each IDAC tells us how well a solute molecule interacts with the solvent with respect to how well it interacts with itself, Fig. 1 can potentially also give us some idea whether a given force field underestimates or overestimates the intermolecular forces between solvent and solute. The diagonal line in Fig. 1 corresponds to the cases where the simulation agreed with the experiment. If a point is located below the diagonal line, the force field potentially underestimates solute – solvent interactions relative to solute – solute interactions. On the other hand, if a point is located above the diagonal line, the force field potentially overestimates solute – solvent interactions relative to solute – solute interactions.

Having an extensive set of IDAC values allows us to look for systematic errors in the force field and how it describes particular functional groups and solvents, as has been done previously in studies with hydration free energies.[9,69] Here, in order to detect possible issues with force field parameters, we partitioned our dataset by functional groups and by solvents. The absolute value of the average error of the free energy differences for functional groups with more than five occurrences in the set can be seen in Figure 2.

Here (Fig. 2), analysis of errors by functional group is slightly complicated by the fact that errors could depend on the identity of the solvent or the identity of the solute. In the
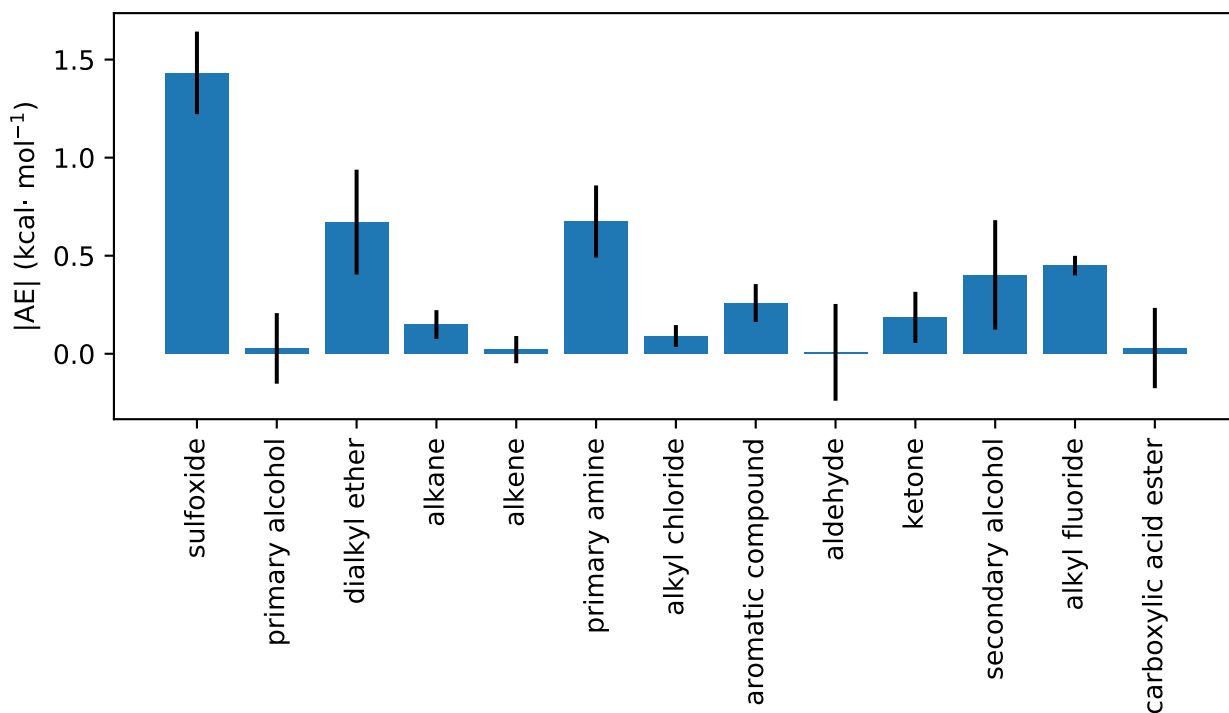
Figure 2: Absolute values of the average errors (AE) for solute functional groups with more than five occurrences in the set. Error bars denote the standard error in the mean of the quantity on the vertical axis.

limit of very large datasets this should be easily surmountable, because a large number of samples would ensure that analysis by solute would involve averaging over a large number of solutes, and analysis by solvent would involve averaging over a large number of solutes. Here, however, our set is relatively small, so it is important to not place too much confidence in any analysis of systematic errors. Still, such analysis can suggest likely targets needed for follow up studies to confirm potential problems, and some trends seem clear.

Here, based on our analysis of errors in activity coefficients broken down by the functional group observed in the solute, we found that sulfoxide had the largest absolute value of the average error of the set (Figure 2). In fact, the only sulfoxide present in the set was dimethylsulfoxide (DMSO), and all IDACs involving DMSO were for this molecule in different solvents (DMSO was never present as a solvent in the simulations). The average error of $+1.4 \pm 0.2$ kcal/mol suggests a systematic error in the GAFF description of DMSO (Fig.

3). The data for primary amines and dialkyl ethers also suggests significant issues, but is less pronounced than the case of DMSO.
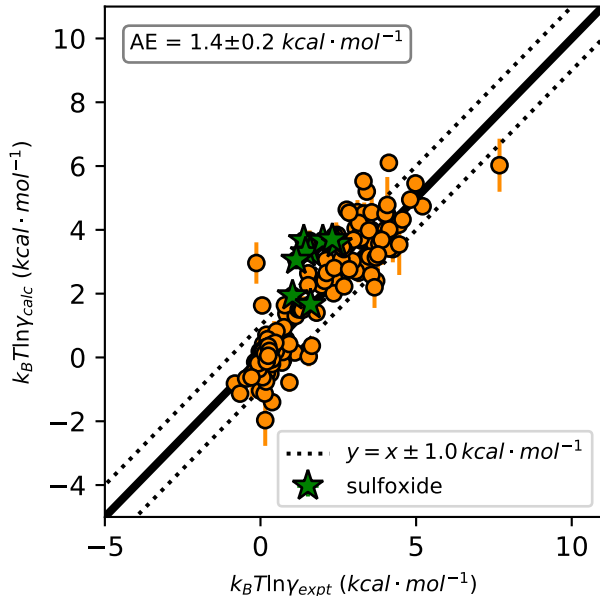

Figure 3: As a solute, dimethylsulfoxide (DMSO) shows a positive shift (average error of 1.7 $\pm$ 0.2 kcal·mol^{-1} for DMSO) with respect to the $y = x$ line, suggesting a potential systematic error in the force field. The set contained no measurements where DMSO was a solvent.

We did a similar analysis of IDAC values broken down by particular solvent. The absolute average error of the free energy differences by solvents with more than five occurrences can be seen in Figure 4.

Given that solvents tended to occur many times in IDAC measurements, our analysis by solvent provided more data concerning potential systematic errors than did our analysis by solute. Methanol, formamide, and ethylene glycol were the solvents whose IDACs showed the largest average absolute errors of the set (Figure 5).

Figure 5 singles out solvents containing four different functional groups for particular analysis, and highlights several potentially important trends. For instance, when examining water as a solvent, IDAC values are nearly evenly spread around the $x = y$ line for IDACs in water (Fig. 5(d)), which indicates that water undersolvates and oversolvates solutes nearly equally often. This is perhaps expected, given that water models are typically given special
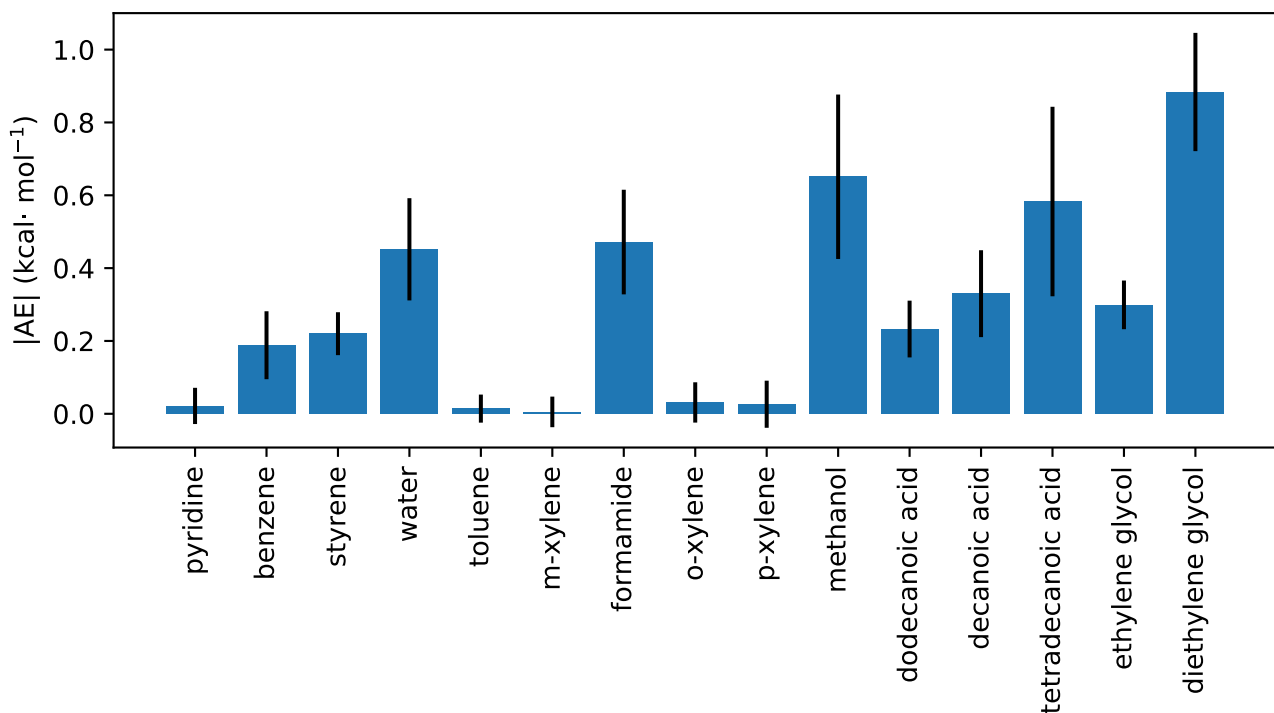
Figure 4: Absolute values of the average errors (AE) for solvents with more than five occurrences in the set. Error bars denote the standard error in the mean of the quantity on the vertical axis.

attention and parameterized quite carefully. Still, the average error for water *is* nonzero by a significant margin (0.5 ± 0.1 kcal·mol^{-1}) suggesting a modest systematic error may remain. In contrast, Fig. 5 (a) and (b) show more systematic shifts away from the diagonal line, suggesting larger potential systematic errors for these solvents. Average errors were 0.9 ± 0.1 kcal·mol^{-1}, -0.7 ± 0.2 kcal·mol^{-1} for $k_B T \log \gamma$ for solutes in diethylene glycol and methanol, respectively, backing up the trend visible by inspection. In both cases, the average error is more than three times the statistical uncertainty. The case of tetradecanoic acid (panel (c) is less clear. Still the size of our sample (11, 7 and 11 IDACs per solvent in (a), (b), and (c)) limits our ability to investigate in much detail, and leaves open other possibilities – for example, for the case of diethylene glycol, all solutes are hydrophobic, so a systematic error in treatment of hydrophobic compounds could result in such a trend, and for methanol, all solutes are amines, so a systematic error in treatment of amines could result in this trend.
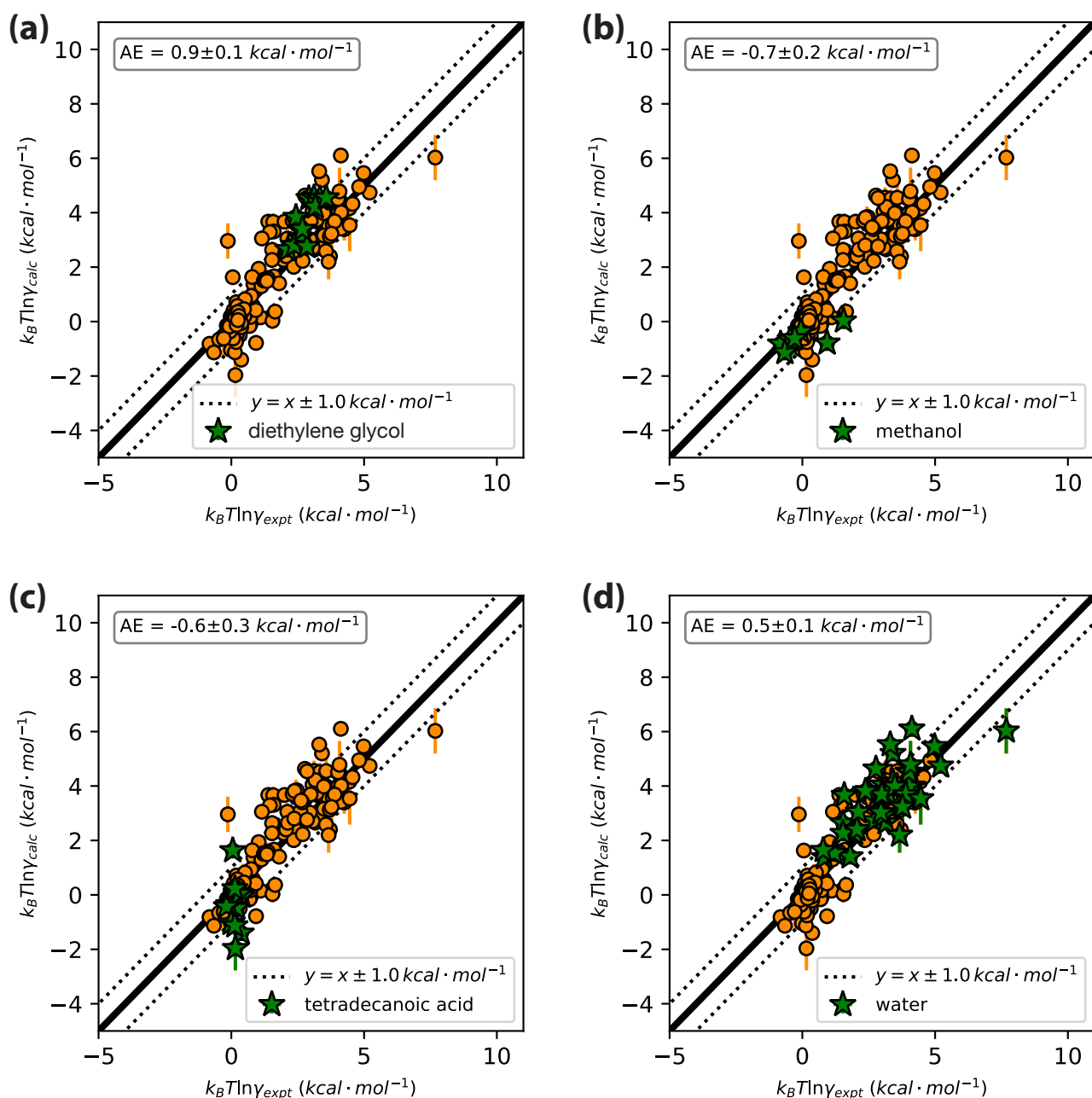
Figure 5: Plots highlighting the IDACs for solutes in diethylene glycol (a), methanol (b), tetradecanoic acid (c), and water (d), four of the cases solutes with the largest average error. AE stands for the average error of computed values for the green star-shaped points. Several of these plots suggest possible systematic errors in the treatment of particular solvents; for example, by eye, panels (a) and (b) suggest at least a modest systematic offset, whereas (c) and (d) are less clear (though the average error is nonzero by a statistically significant margin. Additional data can be found in the Supporting Information.

We believe, however, that the expansion of the data set can yield further insights.

4   Discussion

Here, we calculated a large number of infinite dilution activity coefficients and compared with experimental values extracted automatically from NIST's ThermoML database. We used relatively standard (if computationally demanding), easily-automated approaches for calculation of solvation free energies, and performed the calculations in a high-throughput manner on OpenEye's Orion cloud computing platform. This allowed the calculations to run in a highly parallel and relatively efficient manner and complete on an overnight time-frame. Interestingly, agreement between calculated and experimental values is actually quite good, and these calculations are also able to highlight clear systematic issues for particular functional groups or types of solutes/solvents, suggesting promising areas for investigation of possible force field deficiencies. The fact that infinite dilution activity coefficients can also be measured in a relatively straightforward manner means these will likely serve as a valuable source of data for future tests of computational methods.

Our calculations were kept rather short for computational efficiency, resulting in some-what high statistical errors. For computational efficiency, we ran only 1 ns per lambda value, allowing each $k_B T \log \gamma^\infty$ to be computed quite quickly. With additional sampling at each lambda value we could further reduce the statistical error and better ensure that sampling is adequate, so extending the simulations may be something to explore in future work. Advances in hardware have already provided considerable gains in this area, already making it possible to perform the large number of calculations reported here in a relatively short amount of time, in part due to the availability of GPUs.[70]

Each calculated $k_B T \log \gamma^\infty$ value requires two solvation free energy calculations, which modestly increases the computational cost in comparison to our traditional approach of us-ing hydration free energies, which requires a single free energy calculation[23] . This small

increase is worthwhile given that hydration free energies involve gas-to-water transfers while many events computational chemistry seeks to predict (e.g. binding, solubility, partitioning, permeation, etc.) involve transfer between condensed phases. Thus IDACs may be particularly appealing for force field parameterization since IDAC calculations involve the transfer of molecules between condensed phases, similar to biological and pharmacological events which typically involve transfer from one condensed phase environment to another.

An additional interesting avenue for future inquiry is to assess the accuracy of the solvation free energy and self-solvation free energy separately to determine the exact origins of any systematic errors. Specifically, the self-solvation free energy is directly related to the vapor pressure,[71–73] and vapor pressures for many compounds have been measured. Thus, follow up work could separately examine accuracies of IDACs and vapor pressures, allowing any discrepancies to be traced back to solute-solute or solute-solvent interactions. We plan to revisit this topic in future work as it may prove invaluable for force field development.

The abundance of $k_B T \ln \gamma^\infty$ values around zero (Figure 1) is noteworthy and is potentially an artifact of the type of data which is available in ThermoML. Recall that a value of 0 here corresponds to an activity coefficient of 1 (see Section 1). Specifically, a large portion of the available data is for transfer of solute molecules to solvents of similar polarity — for example, transfer of a polar solute to a polar solvent, or (more commonly) transfer of a nonpolar solute to a nonpolar solvent. If the dataset contained more cases of transfer of a nonpolar solute to a polar solvent, or a polar solute to a nonpolar solvent, we would expect to see more values substantially different from 0. Thus, we believe that the IDAC data should be expanded to include more activity coefficients for compounds of very different polarity than the solvent, to capture more features of transfers between nonpolar (or weakly polar) environments to very polar environments.

5 Conclusion

Here we reported our results calculating some 237 different infinite dilution activity coefficients (IDACs) for small molecules in various solvents, and comparing to experiment. In general, results were quite promising and showed considerable predictive power over a range of 6 $kcal/mol$ in free energy units.

Our results suggest that IDAC values can potentially play an important role in testing force fields and assisting with force field parameterization. They are frequently measured for applications in engineering while other commonly used quantities, such as hydration free energies (ΔG^{hyd}), are not routinely measured. Furthermore, since IDACs can be calculated in a straightforward manner using solvation free energy calculations, they can be calculated with essentially the same degree of precision as solvation free energies, and with the same procedures. IDACs actually could be even better than solvation free energies, which involve transfer between gas and liquid phases, since they are related to transfer between two condensed phases – pure solute and pure solvent – which makes them an ideal candidate to test how a force field represents condensed-phase environments.

While we have made the case that IDAC values are potentially superior to solvation free energies for force field testing and parameterization, there are other properties which have some of the same advantages. For example, partition and distribution coefficients can be estimated from differences in solvation free energies in different solvents, avoiding the issues with phase transfer which plague solvation free energies. Thus these have seen some use in the SAMPL series of challenges — particularly, distribution between water, cyclohexane and octanol.[7,74] And water-octanol partition coefficients are particularly abundant as they are commonly measured in pharmaceutical discovery. Another potentially useful property is relative solubility, examining how well different solvents solvate the same solute[75] However, relative solubilities and partition/distribution coefficients are typically available for relatively few solvents, meaning they lack the advantage of IDACs wherein the same compounds are present both as solute and solvent.

Overall, not only are IDAC calculations appealing in principle, but our results suggest that these calculations can indeed be helpful in identifying force field issues needing attention. Specifically, graphic analysis of experimental and calculated $k_B T \ln \gamma^\infty$ values enabled the identification of possible systematic errors in the force field used in this study. We hope that the evidence shown in this work drives future research in expanding the number of experimental activity coefficients at infinite dilution in the literature, and in using γ^∞ as a new source of constraints for force field parameterization and method development.

Acknowledgement

Supporting Information Available

Jupyter notebooks that generated the input dataset for Orion, generated the plots, and searched ThermoML for the activity values; curated ThermoML entries in `pickle` and `json` formats; YANK's `.yaml` file containing the details of the simulations; a results set in `pickle` and `json` formats; a Python script containing the statistical analysis tools used in this work.

References

(1) Dauber-Osguthorpe, P.; Hagler, A. T. Biomolecular Force Fields: Where Have We Been, Where Are We Now, Where Do We Need to Go and How Do We Get There? *J*

Comput Aided Mol Des **2018**,

(2) Lin, B.; Lopes, P. E. M.; Roux, B.; MacKerell, A. D. Kirkwood-Buff Analysis of Aqueous N-Methylacetamide and Acetamide Solutions Modeled by the CHARMM Additive and Drude Polarizable Force Fields. *J. Chem. Phys.* **2013**, *139*, 084509.

(3) Guthrie, J. P. A Blind Challenge for Computational Solvation Free Energies: Introduction and Overview. *J. Phys. Chem. B* **2009**, *113*, 4501–4507.

(4) Geballe, M. T.; Skillman, A. G.; Nicholls, A.; Guthrie, J. P.; Taylor, P. J. The SAMPL2 Blind Prediction Challenge: Introduction and Overview. *J. Comput. Aided Mol. Des.* **2010**, *24*, 259–279.

(5) Geballe, M. T.; Guthrie, J. P. The SAMPL3 Blind Prediction Challenge: Transfer Energy Overview. *J. Comput. Aided Mol. Des.* **2012**, *26*, 489–496.

(6) Guthrie, J. P. SAMPL4, a Blind Challenge for Computational Solvation Free Energies: The Compounds Considered. *J Comput Aided Mol Des* **2014**, *28*, 151–168.

(7) Yin, J.; Henriksen, N. M.; Slochower, D. R.; Shirts, M. R.; Chiu, M. W.; Mobley, D. L.; Gilson, M. K. Overview of the SAMPL5 Host–Guest Challenge: Are We Doing Better? *J. Comput. Aided Mol. Des.* **2017**, *31*, 1–19.

(8) Shirts, M. R.; Pitera, J. W.; Swope, W. C.; Pande, V. S. Extremely Precise Free Energy Calculations of Amino Acid Side Chain Analogs: Comparison of Common Molecular Mechanics Force Fields for Proteins. *J. Chem. Phys.* **2003**, *119*, 5740–5761.

(9) Mobley, D. L.; Bayly, C. I.; Cooper, M. D.; Shirts, M. R.; Dill, K. A. Small Molecule Hydration Free Energies in Explicit Solvent: An Extensive Test of Fixed-Charge Atomistic Simulations. *J. Chem. Theory Comput.* **2009**, *5*, 350–358.

(10) Fennell, C. J.; Wymer, K. L.; Mobley, D. L. 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.

(11) Oostenbrink, C.; Villa, A.; Mark, A. E.; van Gunsteren, W. F. A Biomolecular Force Field Based on the Free Enthalpy of Hydration and Solvation: The GROMOS Force-Field Parameter Sets 53A5 and 53A6. *J. Comput. Chem.* **2004**, *25*, 1656–1676.

(12) Shivakumar, D.; Harder, E.; Damm, W.; Friesner, R. A.; Sherman, W. Improving the Prediction of Absolute Solvation Free Energies Using the Next Generation OPLS Force Field. *Journal of Chemical Theory and Computation* **2012**, *8*, 2553–2558.

(13) Harder, E.; Damm, W.; Maple, J.; Wu, C.; Reboul, M.; Xiang, J. Y.; Wang, L.; Lupyan, D.; Dahlgren, M. K.; Knight, J. L.; Kaus, J. W.; Cerutti, D. S.; Krilov, G.; Jorgensen, W. L.; Abel, R.; Friesner, R. A. OPLS3: A Force Field Providing Broad Coverage of Drug-like Small Molecules and Proteins. *J. Chem. Theory Comput.* **2016**, *12*, 281–296.

(14) Eckert, C. A.; Newman, B. A.; Nicolaides, G. L.; Long, T. C. Measurement and Application of Limiting Activity Coefficients. *AIChE J.* **1981**, *27*, 33–40.

(15) Bader, M. S. H.; Gasem, K. A. M. Determination of Infinite Dilution Activity Coefficients for Organic-Aqueous Systems Using a Dilute Vapor-Liquid Equilibrium Method. *Chem. Eng. Commun.* **1995**, *140*, 41–72.

(16) Krummen, M.; Gruber, D.; Gmehling, J. Measurement of Activity Coefficients at Infinite Dilution in Solvent Mixtures Using the Dilutor Technique. *Ind. Eng. Chem. Res.* **2000**, *39*, 2114–2123.

(17) Martínez, R.; Sanz, M. T.; Beltrán, S.; Corcuera, E. Activity Coefficients at Infinite Dilution of Volatile Compounds in Water: Effect of Temperature and Salt Concentration. *J. Chem. Eng. Data* **2012**, *57*, 1480–1485.

(18) Chialvo, A. A. Determination of Excess Gibbs Free Energy by the Single-Charging-Integral Approach. Infinite Dilution Activity Coefficients and Related Quantities. *J. Phys. Chem.* **1991**, *95*, 6683–6687.

(19) da Silva, E. F. Use of Free Energy Simulations to Predict Infinite Dilution Activity Coefficients. *Fluid Phase Equilib.* **2004**, *221*, 15–24.

(20) da Silva, E. F. Erratum to "Use of Free Energy Simulations to Predict Infinite Dilution Activity Coefficients" [Fluid Phase Equilb. 221 (2004) 15]. *Fluid Phase Equilib.* **2005**, *231*, 252–253.

(21) Dhakal, P.; Roese, S. N.; Stalcup, E. M.; Paluch, A. S. Application of MOSCED To Predict Limiting Activity Coefficients, Hydration Free Energies, Henry's Constants, Octanol/Water Partition Coefficients, and Isobaric Azeotropic Vapor–Liquid Equilibrium. *J. Chem. Eng. Data* **2018**, *63*, 352–364.

(22) Mobley, D. L.; Guthrie, J. P. FreeSolv: A Database of Experimental and Calculated Hydration Free Energies, with Input Files. *J. Comput. Aided Mol. Des.* **2014**, *28*, 711–720.

(23) Duarte Ramos Matos, G.; Kyu, D. Y.; Loeffler, H. H.; Chodera, J. D.; Shirts, M. R.; Mobley, D. L. Approaches for Calculating Solvation Free Energies and Enthalpies Demonstrated with an Update of the FreeSolv Database. *J. Chem. Eng. Data* **2017**, *62*, 1559–1569.

(24) Malde, A. K.; Zuo, L.; Breeze, M.; Stroet, M.; Poger, D.; Nair, P. C.; Oostenbrink, C.; Mark, A. E. An Automated Force Field Topology Builder (ATB) and Repository: Version 1.0. *J. Chem. Theory Comput.* **2011**, *7*, 4026–4037.

(25) Koziara, K. B.; Stroet, M.; Malde, A. K.; Mark, A. E. Testing and Validation of the Automated Topology Builder (ATB) Version 2.0: Prediction of Hydration Free Enthalpies. *J. Comput. Aided Mol. Des.* **2014**, *28*, 221–233.

(26) Marenich, A. V.; Kelly, C. P.; Thompson, J. D.; Hawkins, G.; Chambers, C. C.; Giesen, D. J.; Winget, P.; Cramer, C. J.; Truhlar, D. G. Minnesota Solvation Database. University of Minnesota, 2012.

(27) Moine, E.; Privat, R.; Sirjean, B.; Jaubert, J.-N. Estimation of Solvation Quantities from Experimental Thermodynamic Data: Development of the Comprehensive CompSol Databank for Pure and Mixed Solutes. *Journal of Physical and Chemical Reference Data* **2017**, *46*, 033102.

(28) Ben-Naim, A. *Molecular Theory of Solutions*; Oxford University Press: Oxford, 2006.

(29) Gautreaux, M. F.; Coates, J. Activity Coefficients at Infinite Dilution. *AIChE J.* **1955**, *1*, 496–500.

(30) Eckert, C. A.; Sherman, S. R. Measurement and Prediction of Limiting Activity Coefficients. *Fluid Phase Equilib.* **1996**, *116*, 333–342.

(31) Prausnitz, J. M.; Lichtenthaler, R. N.; Gomes de Azevedo, E. *Molecular Thermodynamics of Fluid-Phase Equilibria*, 2nd ed.; Prentice Hall, 1986.

(32) Miyano, Y.; Kimura, A.; Kuroda, M.; Matsushita, A.; Yamasaki, A.; Yamaguchi, Y.; Yoshizawa, A.; Tateishi, Y. Henry's Law Constants and Infinite Dilution Activity Coefficients of Propane, Propene, Butane, 2-Methylpropane, 1-Butene, 2-Methylpropene, Trans-2-Butene, Cis-2-Butene, 1,3-Butadiene, Dimethyl Ether, Chloroethane, and 1,1-Difluoroethane in Benzene, Toluene, o-Xylene, m-Xylene, p-Xylene, and Styrene. *J. Chem. Eng. Data* **2007**, *52*, 291–297.

(33) Kojima, K.; Zhang, S.; Hiaki, T. Measuring Methods of Infinite Dilution Activity Coefficients and a Database for Systems Including Water. *Fluid Phase Equilib.* **1997**, *131*, 145–179.

(34) Raal, J. D. Characterization of Differential Ebulliometers for Measuring Activity Co-
efficients. *AIChE J.* **2000**, *46*, 210–220.

(35) Diedenhofen, M.; Eckert, F.; Klamt, A. Prediction of Infinite Dilution Activity Coeffi-
cients of Organic Compounds in Ionic Liquids Using COSMO-RS. *J. Chem. Eng. Data*
2003, *48*, 475–479.

(36) Schacht, C. S.; Zubeir, L.; de Loos, T. W.; Gross, J. Application of Infinite Dilution
Activity Coefficients for Determining Binary Equation of State Parameters. *Ind. Eng.*
Chem. Res. **2010**, *49*, 7646–7653.

(37) Gerber, R. P.; Soares, R. d. P. Prediction of Infinite-Dilution Activity Coefficients Using
UNIFAC and COSMO-SAC Variants. *Ind. Eng. Chem. Res.* **2010**, *49*, 7488–7496.

(38) Lisboa, F. M.; Pliego, J. R. Infinite Dilution Activity Coefficient from SMD Calcu-
lations: Accuracy and Performance for Predicting Liquid-Liquid Equilibria. *J. Mol.*
Model. **2018**, *24*, 56.

(39) Ahmadian Behrooz, H.; Boozarjomehry, R. B. Prediction of Limiting Activity Co-
efficients for Binary Vapor-Liquid Equilibrium Using Neural Networks. *Fluid Phase*
Equilib. **2017**, *433*, 174–183.

(40) Tabar Heydar, K.; Gharavi, H. G.; Nazifi, M.; Mirzaei, M.; Sharifi, A. Using Binary
Mixtures of Dicationic Ionic Liquids for Determination of Activity Coefficients at Infi-
nite Dilution by Gas–Liquid Chromatography. *Fluid Phase Equilib.* **2013**, *353*, 93–100.

(41) Xu, Y.; Qian, W.; Gao, Q.; Li, H. Prediction of Vapor–Liquid Equilibria of Alco-
hol+hydrocarbon Systems by 1H NMR Spectroscopy. *Chem. Eng. Sci.* **2012**, *74*, 211–
218.

(42) Estrada, E.; Díaz, G. A.; Delgado, E. J. Predicting Infinite Dilution Activity Coefficients

of Organic Compounds in Water by Quantum-Connectivity Descriptors. *J. Comput. Aided Mol. Des.* **2006**, *20*, 539–548.

(43) Frenkel, M.; Chirico, R. D.; Diky, V. V.; Dong, Q.; Frenkel, S.; Franchois, P. R.; Embry, D. L.; Teague, T. L.; Marsh, K. N.; Wilhoit, R. C. ThermoML: An XML-Based Approach for Storage and Exchange of Experimental and Critically Evaluated Thermophysical and Thermochemical Property Data. 1. Experimental Data. *J. Chem. Eng. Data* **2003**, *48*, 2–13.

(44) Frenkel, M. D.; Chirico, R. D.; Diky, V.; Dong, Q.; Marsh, K. N.; Dymond, J. H.; Wakeham, W. A.; Stein, S. E.; Koenigsberger, E.; Goodwin, A. R. XML-Based IUPAC Standard for Experimental, Predicted, and Critically Evaluated Thermodymic Property Data Storage and Capture (ThermoML): IUPAC Recommendations 2005. *Pure Appl. Chem.* **2006**, *78*.

(45) Frenkel, M.; Chirico, R. D.; Diky, V.; Brown, P. L.; Dymond, J. H.; Goldberg, R. N.; Goodwin, A. R. H.; Heerklotz, H.; Königsberger, E.; Ladbury, J. E.; Marsh, K. N.; Remeta, D. P.; Stein, S. E.; Wakeham, W. A.; Williams, P. A. Extension of ThermoML: The IUPAC Standard for Thermodynamic Data Communications (IUPAC Recommendations 2011). *Pure Appl. Chem.* **2011**, *83*, 1937–1969.

(46) Beauchamp, K. A.; Behr, J. M.; Rustenburg, A. S.; Bayly, C. I.; Kroenlein, K.; Chodera, J. D. Towards Automated Benchmarking of Atomistic Forcefields: Neat Liquid Densities and Static Dielectric Constants from the ThermoML Data Archive. *J. Phys. Chem. B* **2015**, *119*, 12912–12920.

(47) OpeneEye Scientific Software, I. OEChem Toolkit. 2010. (accessed June 16, 2015).

(48) Rizzi, A.; Grinaway, P. B.; Parton, D. L.; Shirts, M. R.; Wang, K.; Eastman, P. M.; Friedrichs, M. S.; Pande, V. S.; Branson, K.; Mobley, D. L.; Chodera, J. D. YANK: A GPU-Accelerated Platform for Alchemical Free Energy Calculations.

(49) Martínez, L.; Andrade, R.; Birgin, E. G.; Martínez, J. M. PACKMOL: A package for building initial configurations for molecular dynamics simulations. *J. Comp. Chem.* **2009**, *30*, 2157–2164.

(50) Wang, J. M.; Wolf, R. M.; Caldwell, J. W.; a Kollman, P.; a Case, D. Development and Testing of a General Amber Force Field. *J. Comput. Chem.* **2004**, *25*, 1157–1174.

(51) Jakalian, A.; Bush, B. L.; Jack, D. B.; Bayly, C. I. Fast, Efficient Generation of High-Quality Atomic Charges. AM1-BCC Model: I. Method. *J. Comput. Chem.* **2000**, *21*, 132–146.

(52) Jakalian, A.; Jack, D. B.; Bayly, C. I. Fast, Efficient Generation of High-Quality Atomic Charges. AM1-BCC Model: II. Parameterization and Validation. *J. Comput. Chem.* **2002**, *23*, 1623–1641.

(53) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, *79*, 926–935.

(54) Swails, J.; Hernandez, C.; Mobley, D. L.; Nguyen, H.; Wang, L.-P.; Janowski, P. Parmed. (accessed October 9, 2015).

(55) Eastman, P.; Pande, V. S. Efficient Nonbonded Interactions for Molecular Dynamics on a Graphics Processing Unit. *J. Comput. Chem.* **2009**, *31*, 1268–1272.

(56) Eastman, P.; Pande, V. OpenMM: A Hardware-Independent Framework for Molecular Simulations. *Comput. Sci. Eng.* **2010**, *12*, 34–39.

(57) Eastman, P.; Friedrichs, M. S.; Chodera, J. D.; Radmer, R. J.; Bruns, C. M.; Ku, J. P.; Beauchamp, K. A.; Lane, T. J.; Wang, L.-P.; Shukla, D.; Tye, T.; Houston, M.; Stich, T.; Klein, C.; Shirts, M. R.; Pande, V. S. OpenMM 4: A Reusable, Exten-

sible, Hardware Independent Library for High Performance Molecular Simulation. *J. Chem. Theory Comput.* **2013**, *9*, 461–469.

(58) Shirts, M. R.; Mobley, D. L.; Chodera, J. D.; Pande, V. S. Accurate and Efficient Corrections for Missing Dispersion Interactions in Molecular Simulations. *J. Phys. Chem. B* **2007**, *111*, 13052–13063.

(59) Parameswaran, S.; Mobley, D. L. Box Size Effects Are Negligible for Solvation Free Energies of Neutral Solutes. *J Comput Aided Mol Des* **2014**, *28*, 825–829.

(60) Bannan, C. C.; Burley, K. H.; Chiu, M.; Shirts, M. R.; Gilson, M. K.; Mobley, D. L. Blind Prediction of Cyclohexane-Water Distribution Coefficients from the SAMPL5 Challenge. *J Comput-Aided Mol Des* **2016**, *30*, 927–944.

(61) Darden, T.; York, D.; Pedersen, L. Particle Mesh Ewald: An N Log(N) Method for Ewald Sums in Large Systems. *J. Chem. Phys.* **1993**, *98*, 10089–10092.

(62) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. A Smooth Particle Mesh Ewald Method. *J. Chem. Phys.* **1995**, *103*, 8577–8593.

(63) Chodera, J. D.; Shirts, M. R. Replica Exchange and Expanded Ensemble Simulations as Gibbs Sampling: Simple Improvements for Enhanced Mixing. *J. Chem. Phys.* **2011**, *135*, 194110.

(64) Shirts, M. R.; Chodera, J. D. Statistically Optimal Analysis of Samples from Multiple Equilibrium States. *J. Chem. Phys.* **2008**, *129*, 124105.

(65) Bennett, C. H. Efficient Estimation of Free Energy Differences from Monte Carlo Data. *J. Comp. Phys.* **1976**, *22*, 245–268.

(66) Paliwal, H.; Shirts, M. R. A Benchmark Test Set for Alchemical Free Energy Transformations and Its Use to Quantify Error in Common Free Energy Methods. *J. Chem. Theory Comput.* **2011**, *7*, 4115–4134.

(67) Chodera, J. D. A Simple Method for Automated Equilibration Detection in Molecular Simulations. *J. Chem. Theory Comput.* **2016**, *12*, 1799–1805.

(68) Bannan, C. C.; Burley, K. H.; Chiu, M.; Shirts, M. R.; Gilson, M. K.; Mobley, D. L. Blind Prediction of Cyclohexane–Water Distribution Coefficients from the SAMPL5 Challenge. *J.Comput. Aided Mol. Des.* **2016**, *30*, 1–18.

(69) Knight, J. L.; Yesselman, J. D.; Brooks III, C. L. Assessing the Quality of Absolute Hydration Free Energies among CHARMM-Compatible Ligand Parameterization Schemes. *J. Comput. Chem.* **2013**, *34*, 893–903.

(70) Friedrichs, M. S.; Eastman, P.; Vaidyanathan, V.; Houston, M.; Legrand, S.; Beberg, A. L.; Ensign, D. L.; Bruns, C. M.; Pande, V. S. Accelerating Molecular Dynamic Simulation on Graphics Processing Units. *J. Comput. Chem.* **2008**, *30*, 864–872.

(71) Winget, P.; Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. Prediction of Vapor Pressures from Self-Solvation Free Energies Calculated by the SM5 Series of Universal Solvation Models. *J. Phys. Chem. B* **2000**, *104*, 4726–4734.

(72) Horn, H. W.; Swope, W. C.; Pitera, J. W.; Madura, J. D.; Dick, T. J.; Hura, G. L.; Head-Gordon, T. Development of an Improved Four-Site Water Model for Biomolecular Simulations: TIP4P-Ew. *The Journal of Chemical Physics* **2004**, *120*, 9665–9678.

(73) Fuerst, G. B.; Ley, R. T.; Paluch, A. S. Calculating the Fugacity of Pure, Low Volatile Liquids via Molecular Simulation with Application to Acetanilide, Acetaminophen, and Phenacetin. *Ind. Eng. Chem. Res.* **2015**, *54*, 9027–9037.

(74) Yin, J.; Henriksen, N. M.; Slochower, D. R.; Gilson, M. K. The SAMPL5 Host-Guest Challenge: Computing Binding Free Energies and Enthalpies from Explicit Solvent Simulations by the Attach-Pull-Release (APR) Method. *J. Comput. Aided Mol. Des.* **2017**, *31*, 133–145.

(75) Liu, S.; Cao, S.; Hoang, K.; Young, K. L.; Paluch, A. S.; Mobley, D. L. Using MD Simulations To Calculate How Solvents Modulate Solubility. *Journal of Chemical Theory and Computation* **2016**, *12*, 1930–1941.

Graphical TOC Entry

$$k_B T \ln \gamma_i^{\infty} = \Delta G_i^{\text{solv}} - \Delta G_i^{\text{self solv}} + k_B T \ln \left(\frac{\rho_{molar}^{\text{solvent}}}{\rho_{molar}^{\text{pure solute}}} \right)$$