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 see widespread use in calculating various physical properties of interest, with a key goal being predictive molecular design. These simulations, including molecular dynamics (MD) simulations, begin with a underlying energy model or force field and then, based on this model, use simulations to compute properties of interest. However, one of the most significant challenges in molecular dynamics and modeling studies is ensuring that the force field is a good enough approximation of the underlying physics that computed quantities can be used to reproduce experimental properties with the desired level of accuracy. Parameterization of force fields depend on various experimental properties including as much of the chemistry of interest as possible. Physicochemical properties measurable in a relatively straightforward manner are particularly interesting for developers. Such properties can be measured for a relatively diverse chemical set and used to expand the parameterization dataset as needed. Here, we examine infinite dilution activity coefficients (IDACs) which are experimental quantities that can play this role. We retrieved 237 empirical IDACs from NIST's ThermoML, a database of measured thermodynamic properties, and we estimated the corresponding values using solvation free energy calculations. We found that calculated IDAC values correlate strongly with experiment. Specifically, the natural logarithm of calculated and experimental IDAC values shows a Pearson correlation coefficient of 0.85 ± 0.02. The calculated IDAC values allow us to identify strengths and potential weaknesses of force field parameters for specific functional groups in solutes and solvents, suggesting these may be a valuable source of data for force field parameterization, capturing some of the same type of information as hydration and solvation free energies and thus potentially providing a useful new source of experimental data.

1 Introduction

Infinite dilution activity coefficients (IDACs) tell us how far an infinitely dilute mixture is from ideal solution conditions[1–5] and they are of considerable experimental and theoretical interest.[4,6,7] Deviations from ideality indicate whether a solute is particularly good or particularly poor for a given solvent. This means that activity coefficients have a variety of downstream applications, such as for input to chemical engineering models studying liquid-vapor coexistence.[3,5]

Ideal solutions are mixtures where the interactions between two solvent molecules are equal to the interactions between two solute molecules and to the interactions between a solvent and a solute molecule.[1] Real solutions, however, do not satisfy this condition; in most cases, solute and solvent molecules have self-interactions which are not identical to their mutual interactions in solution. In this sense, activity coefficients can be interpreted as a measure of the propensity of a solute molecule to interact with the solvent.

Activity coefficients are also important because they help us determine the effective concentration of a particular component, or its propensity to react. Specifically, the activity of a component in a system, or its effective concentration, is the product between the component's activity coefficient and its real concentration and is related to the chemical potential of a component in a mixture (μ_i) (Eq. 1).

$$\mu_i = \mu_i^0 + RT \ln \gamma \cdot \frac{[i]}{[i_0]}, \tag{1}$$

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. The reference ideal state also plays an important part in the definition of activity coefficients. 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, $\gamma_i \to 1$ as $\chi_i \to 1$.[8] In this definition, the pure liquid is considered an ideal solution because all the interactions between its components have the same magnitude. With this reference state (called the Lewis-Randall standard state), the activity coefficient is 1 for the pure solution. Other reference states are also commonly employed. For example, a common textbook definition uses an ideal dilute solution as a reference state, where 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 will compare to.

An infinite dilution activity coefficient (IDAC) is the activity coefficient of a component when its concentration is infinitely small in a mixture. It is related to the slope of isothermal pressure-composition phase diagrams when the mole fraction tends to zero, and is proportional to the Henry's Law constant.[9] The measurement of IDACs depends on factors such as the volatility of solvent and solute.[5] Techniques such as gas-liquid chromatography,[5,10] high-performance liquid chromatography[10] and differential ebulliometry[5,10,11] are traditionally used to measure activities in extremely dilute systems at varying concentrations, leading to the infinite dilution activity coefficient by extrapolation.[10] There has been considerable interest in predicting these coefficients[5–7,10–21] due to their use in phase equilibria studies in chemical engineering applications.[2–4,14]

IDACs (also represented by γ^{∞}) are related to solvation free energies by the following equation:[6,7,13]

$$\gamma_i^{\infty} = \exp\left(\frac{\Delta G_i^{\text{solv}} - \Delta G_i^{\text{self solv}}}{k_B T}\right) \tag{2}$$

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, and T is the absolute temperature.

The solvation free energy is the free energy change of transferring a molecule from an ideal gas state to a solvent.[22–26] If the solvent is water, we call the solvation free energy a hydration free energy (ΔG^{hyd}). Solvation free energies tell us which phase a given molecule prefers in a multiphasic system and also provide information on how a solute behaves in different environments.[24,27]

In this work, our interest is in physical models which can be used to calculate activity coefficients and related properties from molecular simulations. Particularly, molecular simulations begin with a description of the energy and forces in a physical system as a function of the coordinates — what is known as a “force field” — and allow calculation of numerous physical properties from simulations of such systems[28] . In addition to potentially providing predictions of various quantities like host-guest and protein-ligand binding affinities,[29–33] distribution and partition coefficients,[34,35] solvation free energies[36–38] or other physical properties for design applications, comparison of such results to experiment provides a quantitative test of the underlying physical model or force field.

Hydration and solvation free energies have proved particularly valuable in quantitatively testing all-atom molecular simulations and force fields and in highlighting systematic errors, in part because ΔG^{solv} and ΔG^{hyd} for small molecules can be calculated to a precision better than 0.1 kcal/mol.[39,40] The SAMPL series of blind challenges for computation have helped illustrate this, with several challenges focusing on solvation and hydration free energy prediction.[27,41–44] SAMPL1 through SAMPL4 featured blind prediction of hydration free energies.[27,41–43] SAMPL5 included the prediction of partition coefficients, which also required the calculation of solvation free energies in water, cyclohexane, and octanol.[44,45] Likewise, the FreeSolv database of calculated and experimental hydration free energies has been broadly useful for similar reasons.[26,46] Work on FreeSolv has helped highlight and resolve various force field problems, such as with hydroxyl group parameters.[47]

Even though there are experimental and calculated hydration free energy databases such as FreeSolv[26,46] and ATB,[48,49] ΔG^{solv} and ΔG^{hyd} measurements are difficult and require considerable expertise and specialized equipment, so few to no experimental measurements are presently made.[27] Infinite dilution activity coefficients, however, are a critical property of consideration when studying binary mixtures, and even more so as the number of components and/or the system size grows for industrial applications,[3] meaning that they are subject to considerable experimental attention. Additionally, they are easier to measure than ΔG^{solv} and ΔG^{hyd},[3,4,50,51] and can be calculated with similar precision as the aforementioned free energies. This means they are an ideal candidate as an alternative to hydration free energies for benchmarking computational chemistry methods and force fields.

2 Computational Methods

We obtained experimental activity coefficients at infinite dilution from ThermoML,[52–54] an XML-based system for storage and exchange of thermochemical data. ThermoML was accessed on July 27, 2017 using `thermopyl`,[55] 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,[56] 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[57] 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.[58] Force field parameters were then assigned via Antechamber and Ambertools (version 16.16.0), using the GAFF 1.8 small molecule force field[59] and AM1-BCC charges[60,61] (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.[62] Once parameterized, the resulting systems were stored as `ParmEd`[63] (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[64–66] (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[26]). Nonbonded interactions were calculated for all inter-atomic distances under a cutoff of 9 $\mathring{A}$. Electrostatic interactions were computed using particle mesh Ewald (PME).[67,68] Each Hamiltonian replica exchange simulation run using Yank had 1000 iterations of 500 MD steps of 2 femtoseconds each at each λ value, totaling one nanosecond per replica. 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. Solvation free energies were estimated with the Multistate Bennett Acceptance Ratio[69] (MBAR), an extension of the Bennett Acceptance Ratio[70] 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 most consistently well-performing free energy estimator[71] and is the default free energy estimator in Yank.

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. 2).[6,7,13] The natural logarithm of γ^{∞} is proportional to the difference between the free energy of solvation of a solute i in a given solvent (ΔG_i^{solv}) and the free energy of solvation of the solute molecule in its pure bulk phase ($\Delta G_i^{\text{self solv}}$, free energy of "self-solvation"):

$$k_B T \ln \gamma_i^{\infty} = \Delta G_i^{\text{solv}} - \Delta G_i^{\text{self solv}} \tag{3}$$

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

Here, we calculated the solvation free energies using MBAR[69] and compared to experimental values as shown in Figure 1.

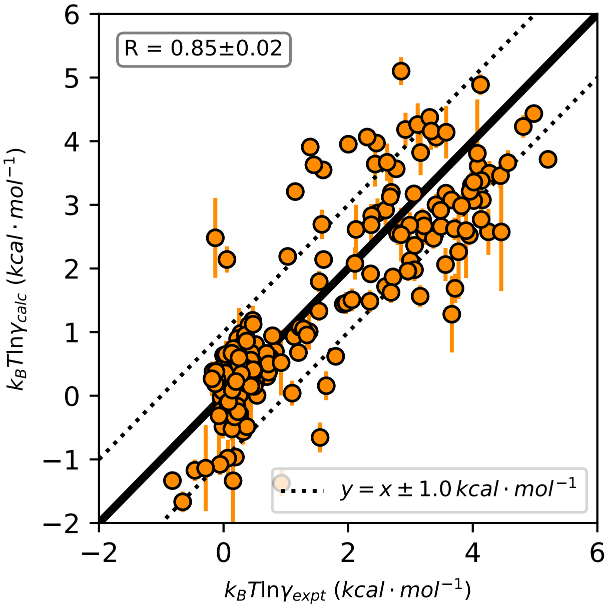

Figure 1: Calculated versus experimental $k_B T \ln \gamma^{\infty}$ for 237 solute – solvent pairs taken from ThermoML. Calculated values are on the vertical axis and experimental on the horizontal.

We found an average error (in free energy units, as in Equation 3) of -0.08 ± 0.05 $\mathrm{kcal \cdot mol^{-1}}$, a root-mean-square (RMS) error of 0.81 ± 0.05 $\mathrm{kcal \cdot mol^{-1}}$, an average unsigned error of 0.58 ± 0.04 $\mathrm{kcal \cdot mol^{-1}}$, a Kendall τ value of 0.61 ± 0.02, and a Pearson R value of 0.85 ± 0.02.

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.[40,72] 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 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

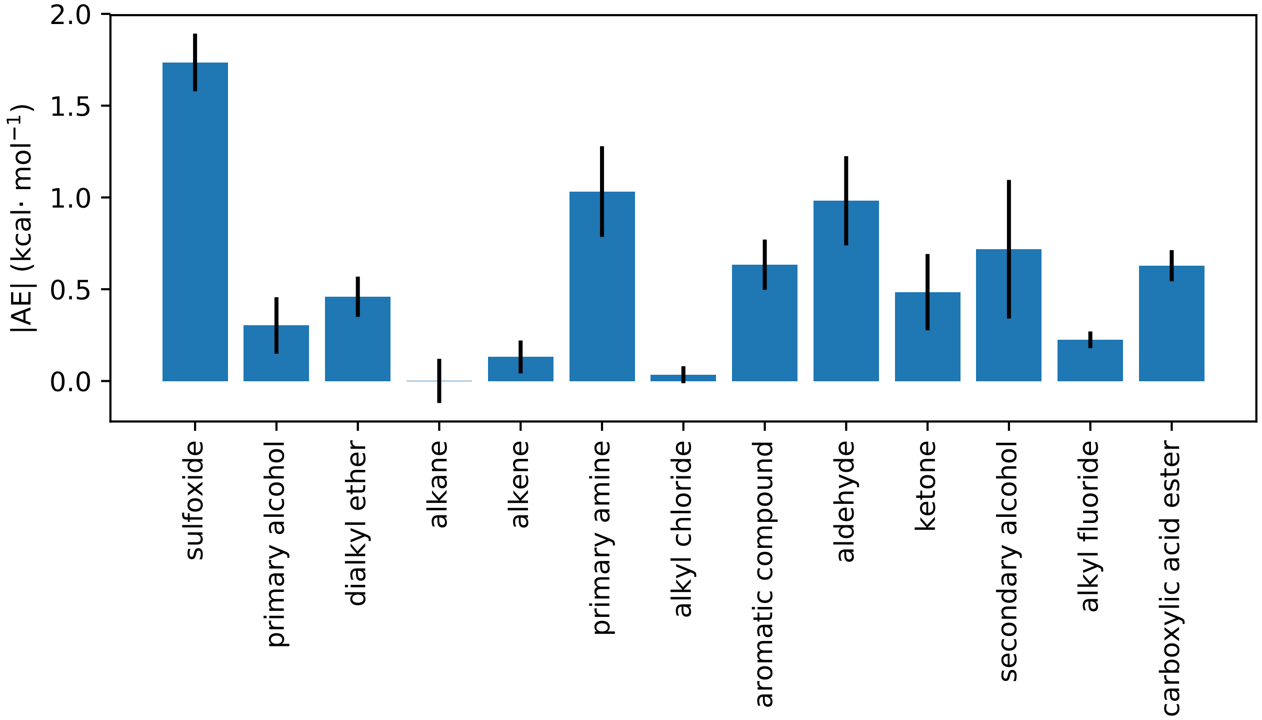

Figure 2: Absolute values of the average errors (AE) for functional groups with more than five occurrences in the set. Error bars denote the standard error in the mean of the quantity on the vertical axis.

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 +1.7 ± 0.2 kcal/mol suggests a systematic error in the GAFF description of DMSO (Fig. 3).

We did a similar analysis of IDAC values broken down by functional groups appearing in the 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

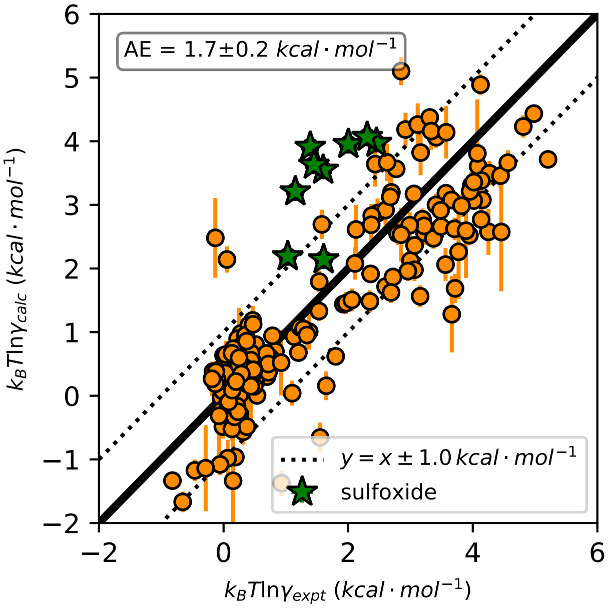

Figure 3: As a solute, dimethylsulfoxide (DMSO) shows a positive shift (average error of 1.7 ± 0.2 kcal·mol^{-1} for DMSO) with respect to the $y = x$ line, suggesting a potential systematic error in the force field. The set contained no measurements where DMSO was a solvent.

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 suggests that the differences with respect to the experimental IDACs are random in nature or are caused by solute parameters. This is perhaps expected, given that water models are typically given special attention and parameterized quite carefully. In contrast, Fig. 5 (a), (b), and (c) show clear systematic shifts away from the diagonal line, suggesting potential systematic errors for these solvents. Average errors were -1.2 ± 0.3 kcal·mol^{-1}, -1.1 ± 0.2 kcal·mol^{-1} and -0.9 ± 0.1 kcal·mol^{-1} for $k_B T \log \gamma$ for solutes in methanol, formamide and ethylene glycol, respectively. The plots in Fig. 5(a), (b) and (c) suggest the potential presence of systematic errors, but the size of our sample (7, 7 and 12 IDACs per solvent in (a), (b), and (c)) limits our ability to investigate in much detail. We believe, however, that the expansion of the data set can help confirm our analysis.

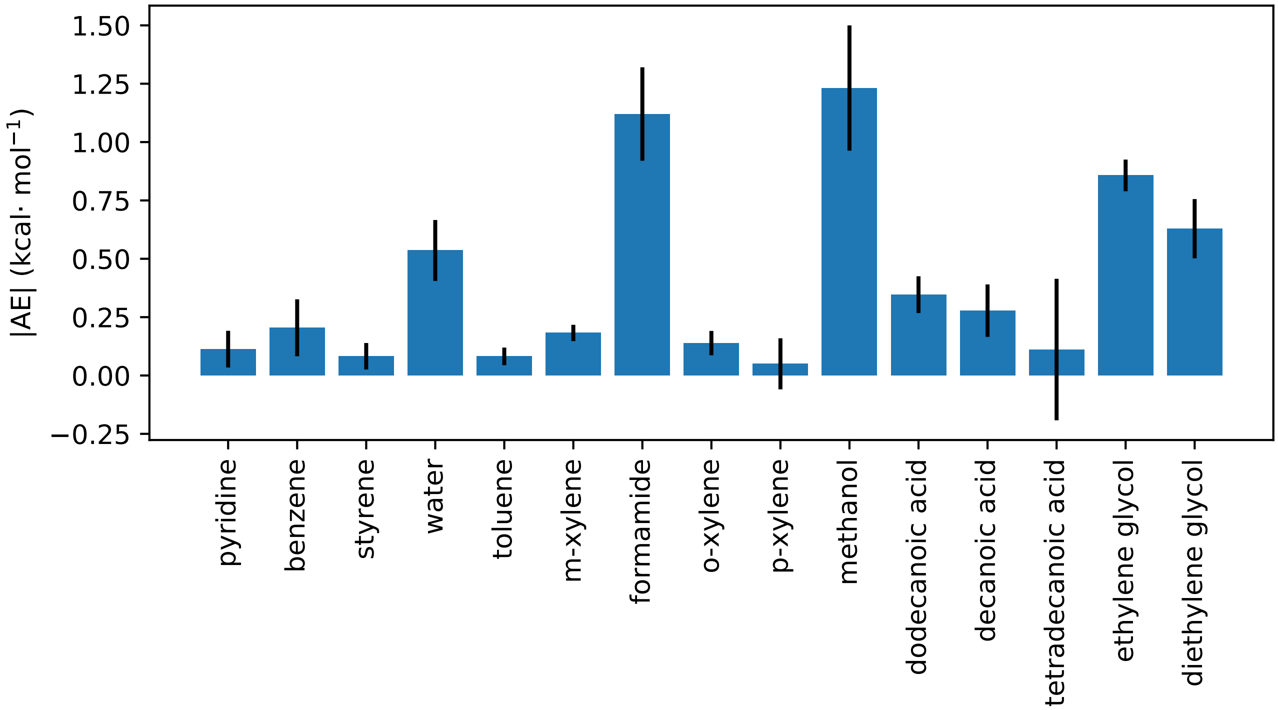

Figure 4: Absolute values of the average errors (AE) for solvents with more than five occurrences in the set. Error bars denote the standard error in the mean of the quantity on the vertical axis.

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. 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

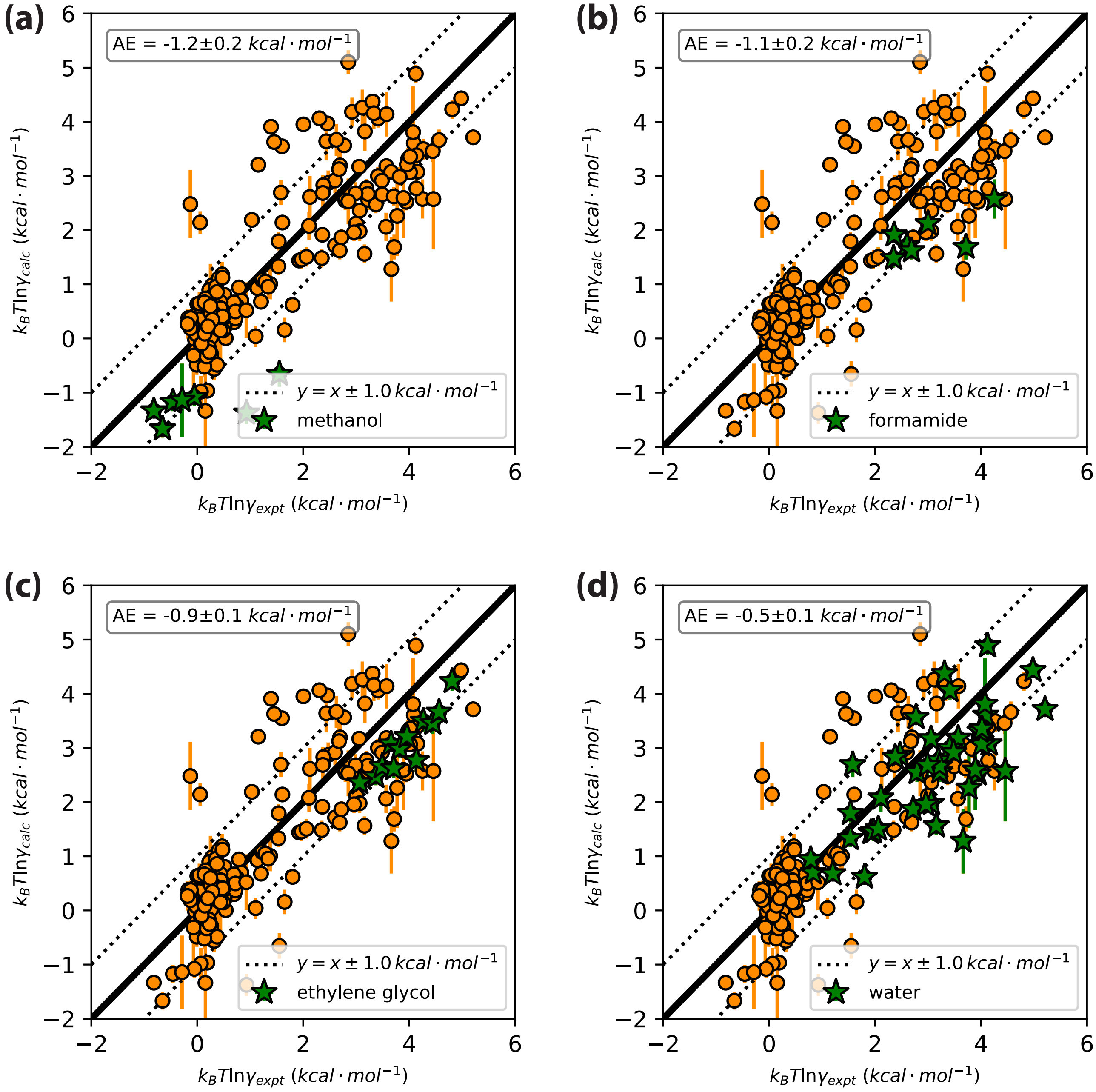

Figure 5: Plots highlighting the IDACs for solutes in methanol (a), formamide (b), ethylene glycol (c), and water (d) simulations. While points in (d) are consistently spread around the $y = x$ line, the remaining plots suggest systematic errors in the description of methanol, formamide and ethylene glycol. Additional data can be found in the Supporting Information. AE stands for the average error of the green star-shaped points.

tests of computational methods.

Our calculations were kept rather short for computational efficiency, resulting in somewhat 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.[73]

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 using hydration free energies, which requires a single free energy calculation[26] . 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.

We also believe that there are two other advantages of using IDACs: 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. 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.

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.

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

GDRM would like to thank CAPES (BEX 3932-13-3) for the Science without Borders scholarship. DLM and GDRM appreciate the financial support from the National Science Foundation (CHE 1352608), and computing support from the UCI GreenPlanet cluster, supported in part by NSF Grant CHE-0840513. DLM also appreciates support from the NIH (1R01GM108889-01). We would also like to thank OpenEye, Christopher Bayly, Anthony Nicholls and the Orion team, especially Quinn Bailey and Craig Bruce.

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

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

(3) 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.

(4) 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.

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

(6) 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.

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

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

(9) 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.

(10) 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.

(11) Raal, J. D. Characterization of Differential Ebulliometers for Measuring Activity Coefficients. *AIChE J.* **2000**, *46*, 210–220.

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

(13) 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.

(14) 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.

(15) 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.

(16) 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.

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

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

(19) 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 Infinite Dilution by Gas–Liquid Chromatography. *Fluid Phase Equilib.* **2013**, *353*, 93–100.

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

(21) 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.

(22) Yu, H.-A.; Karplus, M. A Thermodynamic Analysis of Solvation. *J. Chem. Phys.* **1988**, *89*, 2366–2379.

(23) Cramer, C. J.; Truhlar, D. G. A Universal Approach to Solvation Modeling. *Acc. Chem. Res.* **2008**, *41*, 760–768.

(24) Klimovich, P. V.; Mobley, D. L. Predicting Hydration Free Energies Using All-Atom Molecular Dynamics Simulations and Multiple Starting Conformations. *J. Comput. Aided Mol. Des.* **2010**, *24*, 307–316.

(25) Ratkova, E. L.; Palmer, D. S.; Fedorov, M. V. Solvation Thermodynamics of Organic Molecules by the Molecular Integral Equation Theory: Approaching Chemical Accuracy. *Chem. Rev.* **2015**, *115*, 6312–6356.

(26) 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.

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

(28) Karplus, M.; McCammon, J. A. Molecular Dynamics Simulations of Biomolecules. *Nat. Struct. Mol. Biol.* **2002**, *9*, 646–652.

(29) Jiao, D.; Golubkov, P. A.; Darden, T. A.; Ren, P. Calculation of Protein–Ligand Binding Free Energy by Using a Polarizable Potential. *PNAS* **2008**, *105*, 6290–6295.

(30) Hansen, N.; van Gunsteren, W. F. Practical Aspects of Free-Energy Calculations: A Review. *J. Chem. Theory Comput.* **2014**, *10*, 2632–2647.

(31) Mishra, S. K.; Calabró, G.; Loeffler, H. H.; Michel, J.; Koča, J. Evaluation of Selected Classical Force Fields for Alchemical Binding Free Energy Calculations of Protein-Carbohydrate Complexes. *J. Chem. Theory Comput.* **2015**, *11*, 3333–3345.

(32) Wang, L.; Wu, Y.; Deng, Y.; Kim, B.; Pierce, L.; Krilov, G.; Lupyan, D.; Robinson, S.; Dahlgren, M. K.; Greenwood, J.; Romero, D. L.; Masse, C.; Knight, J. L.; Steinbrecher, T.; Beuming, T.; Damm, W.; Harder, E.; Sherman, W.; Brewer, M.; Wester, R.; Murcko, M.; Frye, L.; Farid, R.; Lin, T.; Mobley, D. L.; Jorgensen, W. L.; Berne, B. J.; Friesner, R. A.; Abel, R. Accurate and Reliable Prediction of Relative Ligand Binding Potency in Prospective Drug Discovery by Way of a Modern Free-Energy Calculation Protocol and Force Field. *J. Am. Chem. Soc.* **2015**, *137*, 2695–2703.

(33) Mobley, D. L.; Gilson, M. K. Predicting Binding Free Energies: Frontiers and Benchmarks. *Annu Rev Biophys* **2017**, *46*, 531–558.

(34) Yang, L.; Ahmed, A.; Sandler, S. I. Comparison of Two Simulation Methods to Com-

pute Solvation Free Energies and Partition Coefficients. *J. Comput. Chem.* **2013**, *34*, 284–293.

(35) Bannan, C. C.; Calabró, G.; Kyu, D. Y.; Mobley, D. L. Calculating Partition Coefficients of Small Molecules in Octanol/Water and Cyclohexane/Water. *J. Chem. Theory Comp.* **2016**, *12*, 4015–4024.

(36) Christ, C. D.; Mark, A. E.; van Gunsteren, W. F. Feature Article Basic Ingredients of Free Energy Calculations: A Review. *J. Comput. Chem.* **2010**, *31*, 1569–1582.

(37) Pohorille, A.; Jarzynski, C.; Chipot, C. Good Practices in Free-Energy Calculations. *J. Phys. Chem. B* **2010**, *114*, 10235–10253.

(38) Mohamed, N. A.; Bradshaw, R. T.; Essex, J. W. Evaluation of Solvation Free Energies for Small Molecules with the AMOEBA Polarizable Force Field. *J. Comput. Chem.* **2016**, n/a–n/a.

(39) 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.

(40) 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.

(41) 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.

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

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

(44) 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.

(45) 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.

(46) 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.

(47) 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.

(48) 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.

(49) 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.

(50) 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.

(51) 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.

(52) 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.

(53) 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*.

(54) 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.

(55) 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.

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

(57) 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.

(58) 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.

(59) 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.

(60) 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.

(61) 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.

(62) 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.

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

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

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

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

(67) 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.

(68) 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.

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

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

(71) 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.

(72) 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.

(73) 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.

Graphical TOC Entry