# Benchmarking Atomistic Simulations against the ThermoML Data Archive: Neat Liquid Densities and Static Dielectric Constants

Kyle A. Beauchamp[+],[1,*] Julie M. Behr[+],[2,†] Patrick B. Grinaway,[3,‡]
Ariën S. Rustenburg,[3,§] Kenneth Kroenlein,[4,¶] and John D. Chodera[1,**]

[1]*Computational Biology Program, Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, New York, NY*
[2]*Tri-Institutional Program in Computational Biology and Medicine, Weill Cornell Medical College, New York, NY*
[3]*Graduate Program in Physiology, Biophysics, and Systems Biology, Weill Cornell Medical College, New York, NY*
[4]*Themodynamics Research Center, NIST, Boulder, CO*

(Dated: March 18, 2015)

Atomistic molecular simulations are a powerful way to make quantitative predictions, but the accuracy of these predictions depends entirely on the quality of the forcefield employed for the task. While experimental measurements of fundamental physical properties offer a straightforward approach for evaluating forcefield quality, the bulk of this information has been tied up in formats that are not machine-readable. These formats require substantial human effort to compile benchmark datasets which are prone to accumulation of human errors, hindering the development of reproducible benchmarks of forcefield accuracy. Here, we examine the feasibility of benchmarking atomistic forcefields against the NIST ThermoML data archive of physicochemical measurements, which aggregates thousands of experimental measurements in a portable, machine-readable, self-annotating format. As a proof of concept, we present a detailed benchmark of the generalized Amber small molecule forcefield (GAFF) using the AM1-BCC charge model against measurements (specifically bulk liquid densities and static dielectric constants at ambient pressure) automatically extracted from the archive, and discuss the extent of available data. The results of this benchmark highlights a general problem with fixed-charge forcefields in the representation of liquids of low dielectric.

*Keywords: molecular mechanics forcefields; forcefield parameterization; forcefield accuracy; forcefield validation; mass density; static dielectric constant; biomolecular simulation*

## I. INTRODUCTION

Recent advances in hardware and software for molecular dynamics simulation now permit routine access to atomistic simulations at the 100 ns timescale and beyond [1]. Leveraging these advances in combination with consumer GPU clusters, distributed computing, or custom hardware has brought microsecond and millisecond simulation timescales within reach of many laboratories. These dramatic advances in sampling, however, have revealed deficiencies in forcefields as a critical barrier to enabling truly predictive simulations of physical properties of biomolecular systems.

Protein and water forcefields have been the subject of numerous benchmarks [2] and enhancements [3–5], with key outcomes including the ability to fold fast-folding proteins [6–8], improved fidelity of water thermodynamic properties [9], and improved prediction of NMR observables. Although small molecule forcefields have also been the subject of benchmarks [10] and improvements [11], such work has typically focused on small perturbations to specific functional groups. For example, a recent study found that modified hydroxyl nonbonded parameters led to improved prediction of static dielectric constants and hydration free energies [11]. There are also outstanding questions of generalizability of these targeted perturbations; it is uncertain whether changes to the parameters for a specific chemical moiety will be compatible with seemingly unrelated improvements to other groups. Addressing these questions requires establishing a community agreement on shared benchmarks that can be easily replicated among laboratories to test proposed forcefield enhancements and expanded as the body of experimental data grows.

A key barrier to establishing reproducible and extensible forcefield accuracy benchmarks is that many experimental datasets are heterogeneous, paywalled, and unavailable in machine-readable formats (although notable counterexamples exist, e.g. the RCSB [12], FreeSolv [13], and the BMRB [14]). While this inconvenience is relatively minor for benchmarking forcefield accuracy for a single target system (e.g. water), it becomes prohibitive for studies spanning the relevant chemical space. To ameliorate problems of data archival, the NIST Thermodynamics Research Center (TRC) has developed a IUPAC standard XML-based format—ThermoML [15]—for storing physicochemical measurements, uncertainties, and metadata. Experimental researchers publishing measurements in several journals (J. Chem. Eng. Data, J. Chem. Therm., Fluid Phase Equil., Therm. Acta, and Int. J. Therm.) are guided through a data archival process that involves sanity checks, conversion to a standard machine-readable format, and archival at the TRC

* kyle.beauchamp@choderalab.org
† julie.behr@choderalab.org
‡ patrick.grinaway@choderalab.org
§ bas.rustenburg@choderalab.org
¶ kenneth.kroenlein@nist.gov
** Corresponding author; john.chodera@choderalab.org

(http://trc.nist.gov/ThermoML.html).

Here, we examine the ThermoML archive as a potential source for providing the foundation for a reproducible, extensible accuracy benchmark of biomolecular forcefields. In particular, we concentrate on two important physical property measurements easily computable in many simulation codes—neat liquid density and static dielectric constant measurements—with the goal of developing a standard benchmark for validating these properties in fixed-charge forcefields of drug-like molecules and biopolymer residue analogues. These two properties provide sensitive tests of forcefield accuracy that are nonetheless straightforward to calculate. Using these data, we evaluate the generalized Amber small molecule forcefield (GAFF) [16, 17] with the AM1-BCC charge model [18, 19] and identify systematic biases to aid further forcefield refinement.

## II. METHODS

### A. ThermoML Archive retrieval and processing

A tarball archive snapshot of the ThermoML Archive was obtained from the the NIST TRC on 13 Sep 2014. [JDC: Because readers cannot easily extract a specific daily snapshot, I think we also want to make the archive subset we used in this paper available as Supplementary Information.] To explore the content of this archive, we created a Python (version 2.7.9) tool (ThermoPyL: https://github.com/choderalab/ThermoPyL) that formats the XML content into a spreadsheet-like format accessible via the Pandas (version 0.15.2) library. First, we obtained the XML schema (http://media.iupac.org/namespaces/ThermoML/ThermoML.xsd) defining the layout of the data. This schema was converted into a Python object via PyXB 1.2.4 (http://pyxb.sourceforge.net/). Finally, this schema was used to extract the data into Pandas [**?** ] dataframes, and the successive filters data filters described in Section III A were applied.

### B. Simulation

#### 1. Preparation

Simulation boxes containing 1000 molecules were constructed using PackMol version 14-225 [20] wrapped in an automated tool. [JDC: What density was used during the packmol step?] [JDC: Remind readers how they can obtain the tool?] AM1-BCC [18, 19] charges were generated with the OpenEye Python Toolkit version 2014-6-6 [21], using the `oequacpac.OEAssignPartialCharges` module with the `OECharges_AM1BCCSym` option, which utilizes a conformational expansion procedure prior to charge fitting to minimize artifacts from intramolecular contacts. The selected conformer was then processed using `antechamber` (with `parmchk2`) and `tleap` in AmberTools 14 [22] to produce `prmtop` and `inpcrd` files, which were then read into OpenMM using the `simtk.openmm.app` module. Simulation code used libraries gaff2xml 0.7 [23], OpenMM 6.3 [24], and MDTraj 1.3 [25].

The following shell commands can be used to install the necessary prerequisites via the `conda` package manager for Python:

```
# conda config -add channels http://conda.binstar.org/omnia

# conda install "gaff2xml>=0.7" "pymbar>=2.1" "mdtraj>=1.3"
"openmm>=6.3" packmol
```

[JDC: Maybe we can move these instructions on installing the exact versions to the Appendix or SI, where we can go into one-column mode? Also, we probably want to specify how to get the *exact* versions using == rather than the *latest* versions using >=.]

#### 2. Equilibration and production

Boxes were first minimized and equilibrated for $10^7$ steps with an equilibration timestep of 0.4 fs and a collision rate of 5 $ps^{-1}$. Production simulations were performed with OpenMM 6.2 [24] using a Langevin integrator (with collision rate 1 $ps^{-1}$) and a 1 fs timestep, as we found that timesteps of 2 fs timestep or greater led to a significant timestep dependence in computed equilibrium densities (Table 4). [JDC: Cite Langevin integrator used in OpenMM.] [KAB: please provide the reference.] Pressure control to 1 atm was achieved with a Monte Carlo barostat utilizing molecular scaling and automated step size adjustment during equilibration, with volume moves attempted every 25 steps. The particle mesh Ewald (PME) method [26] was used with a long-range cutoff of 0.95 nm and a long-range isotropic dispersion correction. [JDC: Can we report the automatically-selected PME parameters to aid reproducibility in other codes?]

Simulations were continued until automatic analysis showed standard errors in densities were less than $2 \times 10^{-4}$ g / mL. Automatic analysis was run every X ps of simulation time, and utilized the `detectEquilibration` method in the timeseries module of pymbar 2.1 [27] to automatically trim the initial portion of the simulation with strong far-from-equilibrium behavior by maximizing the number of effectively uncorrelated samples in the remainder of the production simulation as determined by autocorrelation analysis. Statistical errors were computed after subsampling the data to produce effectively uncorrelated equilibrium samples by $\delta^2 \rho \approx var(\rho)/N_{\mathrm{eff}}$, where $var(\rho)$ is the sample variance of the density and $N_{\mathrm{eff}}$ is the number of effectively uncorrelated samples. [JDC: I've tried to clarify this. Can you check if I did this correctly?]

Instantaneous densities were stored every 250 fs, while trajectory snapshots were stored every 5 ps.

#### 3. Data analysis and statistical error estimation

Trajectory analysis was performed using OpenMM 6.3 [24] and MDTraj 1.3 [25]. [JDC: Did we plan to make this data

available somewhere, or is it sufficient to put out the scripts?]

Mass density $\rho$ was computed via the relation,

$$\rho = \left\langle \frac{M}{V} \right\rangle, \tag{1}$$

where $M$ is the total mass of all particles in the system and $V$ is the instantaneous volume of the simulation box.

Static dielectric constants were calculated using the dipole fluctuation approach appropriate for PME with conducting ("tin-foil") boundary conditions [9], with the total system box dipole $\mu$ computed from trajectory snapshots using MDTraj 1.3 [25]. [JDC: I think the TIP4P-Ew paper cites the source of this equation, which we should cite as well.]

$$\epsilon = 1 + \frac{4\pi}{3} \frac{\langle \mu \cdot \mu \rangle - \langle \mu \rangle \cdot \langle \mu \rangle}{\langle V \rangle k_B T} \tag{2}$$

Statistical uncertainties were computed by bootstrapping uncorrelated samples following discarding the automatically-determined initial portion of the simulation to equilibration, as described in Section **??**. All reported uncertainties represent an estimate of one standard deviation of the mean unless otherwise reported.

#### 4. *Code availability*

All custom code is available from XXX.

## III. RESULTS

### A. Extracting neat liquid measurements from the NIST TRC ThermoML Archive

As described in Section II A, we retrieved a copy of the ThermoML Archive and performed a number of sequential filtering steps to produce an ThermoML extract relevant for benchmarking forcefields describing small organic molecules. As our aim is to explore neat liquid data with functional groups relevant to biopolymers and drug-like molecules, we applied the following ordered filters, starting with all data containing density or static dielectric constants:

1. The measured solution contains only a single component (e.g. no binary mixtures)
2. The molecule contains only druglike elements (defined here as H, N, C, O, S, P, F, Cl, Br)
3. The molecule has $\leq$ 10 heavy atoms
4. The measurement was performed in a biophysically relevant temperature range $(270 \leq T\,[\mathrm{K}] \leq 330)$
5. The measurement was performed at ambient pressure $(100 \leq P\,[\mathrm{kPA}] \leq 102)$
6. Measured mass densities $\leq$ 300 kg m$^{-3}$ were discarded to eliminate gas-phase measurements
7. The temperature and pressure were rounded to nearby values (as described below), averaging all measurements within each group of like conditions
8. Only conditions (molecule, temperature, pressure) for which *both* density and dielectric constants were available were retained

| | Number of measurements remaining | |
|---|---|---|
| **Filter step** | **Mass density** | **Static dielectric** |
| 1. single component | 130074 | 1649 |
| 2. only druglike elements | 120410 | 1649 |
| 3. $\leq$10 heavy atoms | 67897 | 1567 |
| 4. $(270 \leq T \leq 330)$ [K] | 36827 | 962 |
| 5. ambient pressure | 13598 | 461 |
| 6. liquid state | 13573 | 461 |
| 7. aggregate T, P | 3573 | 432 |
| 8. density and dielectric | 245 | 245 |

**TABLE I. Successive filtration of the ThermoML Archive.** A set of successive filters were applied to all measurements in the ThermoML Archive (accessed 13 Sep 2014) that contained either mass density or static dielectric constant measurements. Each column reports the number of measurements remaining after successive application of the corresponding filtration step.

The temperature and pressure rounding step was motivated by common data reporting variations; for example, an experiment performed at the freezing temperature of water and ambient pressure might be entered as either 101.325 kPA or 100 kPA, with a temperature of either 273 K or 273.15 K. Therefore all pressures within the range [kPA] $(100 \leq P \leq 102)$ were rounded to exactly 1 atm (101.325 kPa). Temperatures were rounded to one decimal place in K.

The application of these filters (Table I) leaves 245 conditions—where a *condition* here indicates a (molecule, temperature, pressure) tuple—for which both density and dielectric data are available. The functional groups present in the resulting dataset are summarized in Table II; see Section II A for further description of the software pipeline used.

### B. Benchmarking GAFF/AM1-BCC against the ThermoML Archive

#### 1. *Mass density*

Mass densities of bulk liquids have been widely used for parameterizing and testing forcefields, particularly the Lennard-Jones parameters representing dispersive and repulsive interactions [29, 30]. We therefore used the present ThermoML extract as a benchmark of the GAFF/AM1-BCC forcefield (Fig. 1).

Overall, the densities show reasonable accuracy, with a root-mean square (RMS) relative error over all measurements of 2.8$\pm$0.1%, especially encouraging given that this

| Functional Group | Occurrences |
|---|---|
| 1,2-aminoalcohol | 4 |
| 1,2-diol | 3 |
| alkene | 3 |
| aromatic compound | 1 |
| carbonic acid diester | 2 |
| carboxylic acid ester | 4 |
| dialkyl ether | 7 |
| heterocyclic compound | 3 |
| ketone | 2 |
| lactone | 1 |
| primary alcohol | 19 |
| primary aliphatic amine (alkylamine) | 2 |
| primary amine | 2 |
| secondary alcohol | 4 |
| secondary aliphatic amine (dialkylamine) | 2 |
| secondary aliphatic/aromatic amine (alkylarylamine) | 1 |
| secondary amine | 3 |
| sulfone | 1 |
| sulfoxide | 1 |
| tertiary aliphatic amine (trialkylamine) | 3 |
| tertiary amine | 3 |

**TABLE II. Functional groups present in filtered dataset.** The filtered ThermoML dataset contained 245 distinct (molecule, temperature, pressure) conditions, spanning 44 unique compounds. The functional groups represented in these compounds (as identified by the program `checkmol` v0.5 [28]) is summarized here.

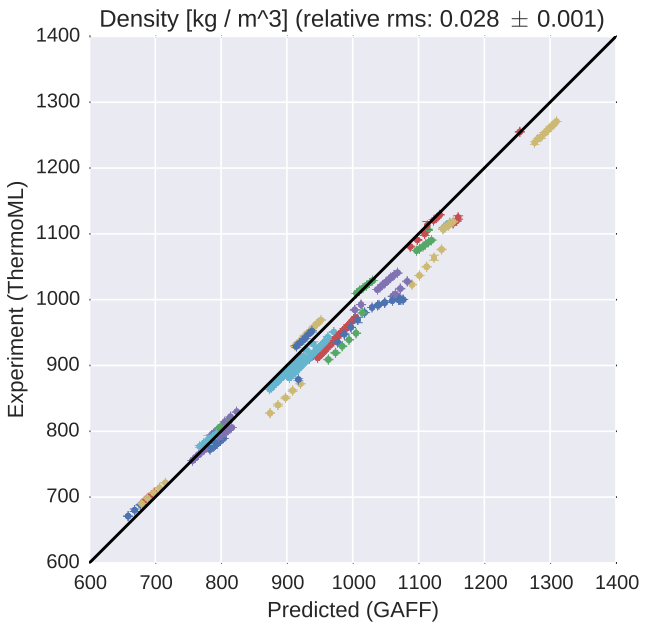



**FIG. 1. Comparison of liquid densities between experiment and simulation.** Liquid density measurements extracted from ThermoML are compared against densities predicted using the GAFF/AM1-BCC small molecule fixed-charge forcefield. Color groupings represent identical chemical species. Simulation error bars represent one standard error of the mean, with the number of effective (uncorrelated) samples estimated using pymbar. Experimental error bars indicate the standard deviation between independently reported measurements, when available, or author-reported standard deviations in ThermoML entries; for some measurements, neither uncertainty estimate is available. See SI Fig. 5 for further discussion of error.

forcefield was not designed with the intention of modeling bulk liquid properties of organic molecules [16, 17] This is reasonably consistent with previous studies reporting agreement of 4% on a different benchmark set [10].

[JDC: Discuss outliers from Fig. 7 here. There must be more things we can say about densities. Some of the densities are quite good, while others (e.g. water, sulfolane, dimethylcarbonate) seem poor, with systematic bias toward higher densities than experiment. We know, for example, for sulfolane, hypervalent sulfur atoms are a challenge for GAFF (from SAMPL challenges), and dimethyl carbonate has *three* oxygens, which had a vdW issue also observed in a previous SAMPL (and maybe addressed in the dielectric paper?). We can also point out that densities at different temperatures for a given molecule seem to be biased in a consistent way. Perhaps we could start by producing a figure showing OpenEye OEDepict images of the outliers, or embed those images in Fig. 7?]

### 2. *Static dielectric constant*

As a measure of the dielectric response, the static dielectric constant of neat liquids provides a critical benchmark of the accuracy electrostatic treatment in forcefield models. We therefore compare simulations against the measurements in our ThermoML extract. Overall, we find the dielectric constants to be qualitatively reasonable, but with clear deviations from experiment. In particular, GAFF/AM1-BCC systematically underestimates the dielectric constants for nonpolar organics, with the predictions of $\epsilon \approx 1.0 \pm 0.05$ being substantially smaller than the measured $\epsilon \approx 2$. Because this deviation likely stems from the lack of an explicit treatment of electronic polarization, we used a simple empirical polarization model that computes the molecular electronic polarizability $\alpha$ a sum of elemental atomic polarizability contributions [31]. From the computed molecular electronic polarizability $\alpha$, an additive correction to the simulation-derived static dielectric constant accounting for the missing electronic polarizability can be computed [9]

$$\Delta\epsilon = 4\pi N \frac{\alpha}{\langle V \rangle} \tag{3}$$

While a similar polarization correction was used in the development of the TIP4P-Ew water model, where it had a minor effect [9], missing polarizability is a dominant contribution to the static dielectric constant of nonpolar organic molecules; in the case of water, the empirical atomic polarizability model predicts a dielectric correction of 0.52, while 0.79 was used for the TIP4P-Ew model. Considering all predictions in the present work leads to polarizability corrections to the static dielectric of $0.74 \pm 0.08$.

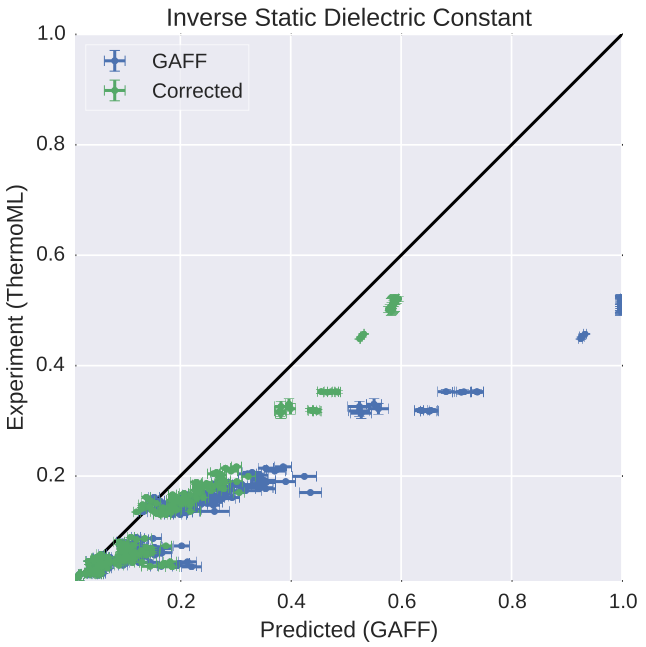



**FIG. 2. Measured (ThermoML) versus predicted (GAFF/AM1-BCC) inverse static dielectrics (a).** Simulation error bars represent one standard error of the mean estimated via circular block averaging [32] with block sizes automatically selected to maximize the error [33]. Experimental error bars indicate the larger of standard deviation between independently reported measurements and the authors reported standard deviations; for some measurements, neither uncertainty estimate is available. See SI Fig. 5 for further discussion of error. The inverse dielectric constant $\epsilon^{-1}$ is plotted instead of $\epsilon$ because $\epsilon^{-1}$ is directly proportional to the Coulomb interaction energy between point charges embedded in a dielectric material [e.g. $U(r) \propto q_1 q_2 / r \propto \epsilon^{-1}$].

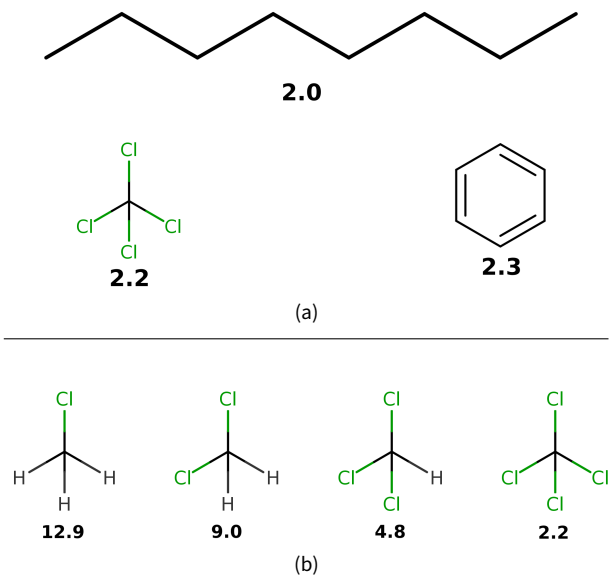



**FIG. 3. Typical experimental static dielectric constants of some nonpolar compounds.** (a). Measured static dielectric constants of various nonpolar or symmetric molecules [37, 38]. Fixed-charge forcefields give $\epsilon \approx 1$ for each species; for example, we calculated $\epsilon = 1.003 \pm 0.0002$ for octane. [JDC: Sig figs issue.] (b). A congeneric series of chloro-substituted methanes have static dielectric constants between 2 and 13. Reported dielectric constants are at near-ambient temperatures.

## IV. DISCUSSION

### A. Mass densities

JDC: Can we say anything about mass densities? For example, how accurate do we expect mass densities need to be? Are there other properties we expect to be highly correlated to mass densities?

For example, I suspect that the total system enthalpy (and therefore enthalpy per molecule) will be highly sensitive to total mass density. For a given forcefield, the sensitivity to a density error at fixed volume (NVT) is

$$\frac{\partial \langle U \rangle}{\partial \rho} = \frac{\partial \langle U \rangle}{\partial V} \frac{\partial V}{\partial \rho} = -\frac{\partial \langle U \rangle}{\partial V} M \rho^{-2} \tag{4}$$

Of course, it's possible the forcefield is parameterized to give a reasonable enthalpy per molecule (related to the enthalpy of vaporization) even though the density has larger deviations. But are there other properties we can think of that may be disrupted? Cavity formation free energies?

### B. Dielectric constants in forcefield parameterization

Recent forcefield development has seen a resurgence of papers fitting dielectric constants during forcefield parameterization [11, 34]. However, a number of authors have pointed out potential challenges in constructing self-consistent fixed-charge forcefields [35, 36].

Interestingly, recent work by Dill and coworkers [35] observed that, for $CCl_4$, reasonable choices of point charges are incapable of recapitulating the observed dielectric of $\epsilon = 2.2$, instead producing dielectric constants in the range of $1.0 \leq \epsilon \leq 1.05$. This behavior is quite general: fixed point charge forcefields will predict $\epsilon \approx 1$ for many nonpolar or symmetric molecules, but the measured dielectric constants are instead $\epsilon \approx 2$ (Fig. 3). While this behavior is well-known and results from missing physics of polarizability, we suspect it may have several profound consequences, which we discuss below.

Suppose, for example, that one attempts to fit forcefield parameters to match the static dielectric constants of $CCl_4$, $CHCl_3$, $CH_2Cl_2$, and $CH_3Cl$. In moving from the tetrahedrally-symmetric $CCl_4$ to the asymmetric $CHCl_3$, it suddenly becomes possible to achieve the observed dielectric constant of 4.8 by an appropriate choice of point charges. However, the model for $CHCl_3$ uses fixed point charges to account for *both* the permanent dipole moment and the electronic polarizability, whereas the $CCl_4$ model

contains no treatment of polarizability. We hypothesize that this inconsistency in parameterization may lead to strange mismatches, where symmetric molecules (e.g. benzene and $CCl_4$) have qualitatively different properties than closely related asymmetric molecules (e.g. toluene and $CHCl_3$).

How important is this effect? As a possible real-world example, we imagine that the missing atomic polarizability could be important in accurate transfer free energies involving low-dielectric solvents, such as the small-molecule transfer free energy from octanol or cyclohexane to water. The Onsager model for solvation of a dipole $\mu$ of radius $a$ gives us a way to estimate the magnitude of error introduced by making an error $\Delta\epsilon$ the static dielectric constant of a solvent. The free energy of dipole solvation is given by this odel as

$$\Delta G = -\frac{\mu^2}{a^3}\frac{\epsilon - 1}{2\epsilon + 1} \tag{5}$$

such that, for an error of $\Delta\epsilon$ departing from the true static dielectric constant $\epsilon$, we find the error in solvation is

$$\Delta\Delta G = -\frac{\mu^2}{a^3}\left[\frac{(\epsilon + \Delta\epsilon) - 1}{2(\epsilon + \Delta\epsilon) + 1} - \frac{\epsilon - 1}{2\epsilon + 1}\right] \tag{6}$$

For example, the solvation of water ($a = 1.93$ Å, $\mu = 2.2$ D) in a low dielectric medium such as tetrachloromethane or benzene ($\epsilon \sim 2.2$, but $\Delta\epsilon = -1.2$) gives an error of $\Delta\Delta G \sim -2$ kcal/mol.

The ramifications can be relevant for quantities of interest to drug discovery projects. Consider the transfer of small druglike molecules from a nonpolar solvent (such as cyclohexane) to water, a property often measured to indicate the expected degree of lipophilicity of a compound. To estimate the magnitude of error expected, for each molecule in the latest (Feb. 20) FreeSolv database [13, 39], we estimated the expected error in computed transfer free energies should GAFF/AM1-BCC be used to model the nonpolar solvent cyclohexane using the Onsager model (Eq. 6). We used took the cavity radius $a$ to be the half the maximum interatomic distance and calculated $\mu = \sum_i q_i r_i$ using the provided mol2 coordinates and AM1-BCC charges. This calculation predicts a mean error of $-0.91 \pm 0.07$ kcal / mol for the 643 molecules (where the standard error is computed from bootstrapping over FreeSolv compound measurements), suggesting that the missing atomic polarizabilty unrepresentable by fixed point charge forcefields could contribute substantially to errors in predicted transfer and solvation properties of druglike molecules. We also conjecture that [JDC: Unfinished thought?]

Given their ease of measurement and direct connection to long-range electrostatic interactions, static dielectric constants have high potential utility as primary data for forcefield parameterization efforts. Although this will require the use of forcefields with explicit treatment of atomic polarizability, the inconsistency of fixed-charge models in low-dielectric media is sufficiently alarming to motivate further study of polarizable forcefields. In particular, continuum methods [40–42], point dipole methods [43, 44], and Drude methods [45, 46] have been maturing rapidly. Finding the optimal balance of accuracy and performance remains an open question; however, the use of experimentally-parameterized direct polarization methods [47] may provide polarizability physics at a cost not much greater than fixed charge forcefields.

### C. ThermoML as a data source

The present work has focused on the neat liquid density and dielectric measurements present in the ThermoML Archive [15, 48, 49] as a target for molecular dynamics forcefield validation. While liquid mass densities and static dielectric constants have already been widely used in forcefield work, several aspects of ThermoML make it a unique resource for the forcefield community. First, the aggregation, support, and dissemination of ThermoML datasets through the ThermoML Archive is supported by NIST, whose mission makes these tasks a long-term priority. Second, the ThermoML Archive is actively growing, through partnerships with several journals, and new experimental measurements published in these journals are critically examined by the TRC and included in the archive. [JDC: Is the number of journal here also expanding?] Finally, the files in the ThermoML Archive are portable and machine readable via a formal XML schema, allowing facile access to hundreds of thousands of measurements. Numerous additional physical properties contained in ThermoML—including activity coefficients, diffusion constants, boiling point temperatures, critical pressures and densities, coefficients of expansion, speed of sound measurements, viscosities, excess molar enthalpies, heat capacities, and volumes—for neat phases and mixtures represent a rich dataset of high utility for forcefield validation and parameterization.

[JDC: Can we give some other statistics (maybe in a table) of the numbers of other kinds of measurements contained in ThermoML?]

## V. CONCLUSIONS

- ThermoML is a potentially useful resource for the forcefield community
- We have curated a subset of the ThermoML Data Archive for neat liquids with druglike atoms, with thousands of densities and hundreds of dielectrics
- Empirical polarization models correct a systematic bias in comparing fixed-charge forcefields to static dielectric constants

[JDC: Do we need an explicit Conclusions section? I am happy with a merged “Discussion and Conclusions” section.]

## VI. ACKNOWLEDGEMENTS

We thank Vijay S. Pande (Stanford University), Lee-Ping Wang (Stanford University), Peter Eastman (Stanford University), Robert McGibbon (Stanford University), Jason Swails (Rutgers University), David L. Mobley (University of California, Irvine), Christopher I. Bayly (OpenEye Software), Michael R. Shirts (University of Virginia), and members of Chodera lab for helpful discussions. Support for JMB was provided by the Tri-Institutional Training Program in Computational Biology and Medicine (via NIH training grant 1T32GM083937). [JDC: Need support acknowledgments for Patrick and Bas.]

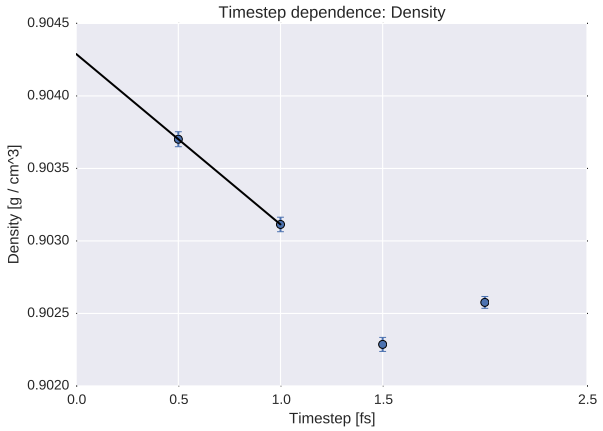
## VII. DISCLAIMERS

This contribution of the National Institute of Standards and Technology (NIST) is not subject to copyright in the United States. Products or companies named here are cited only in the interest of complete technical description, and neither constitute nor imply endorsement by NIST or by the U.S. government. Other products may be found to serve as well.

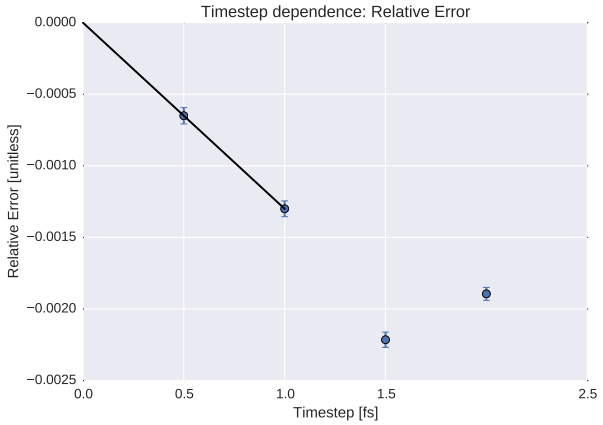
## Appendix A: Supplementary Information

All information below this point will eventually be pulled into a separate SI. This will happen closer to submission, as the formatting may be journal-specific. The references may be split in two as well, depending on journal. [JDC: It may be fine to leave this as an Appendix.]

- Figure: Timestep-dependence of density
- Figure: Error analysis for ThermoML dataset
- Table (CSV File): ThermoML Dataset used in present analysis.

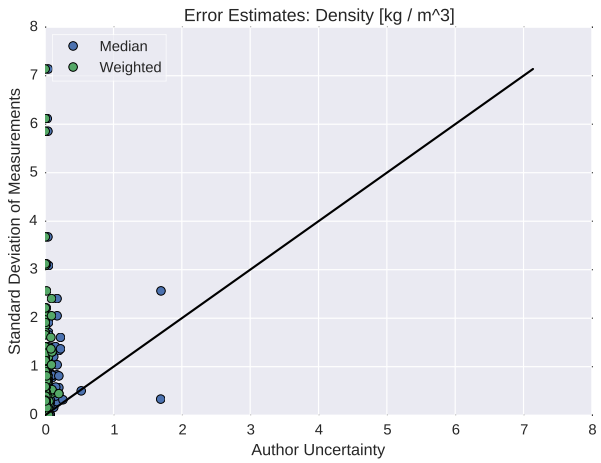



(a)

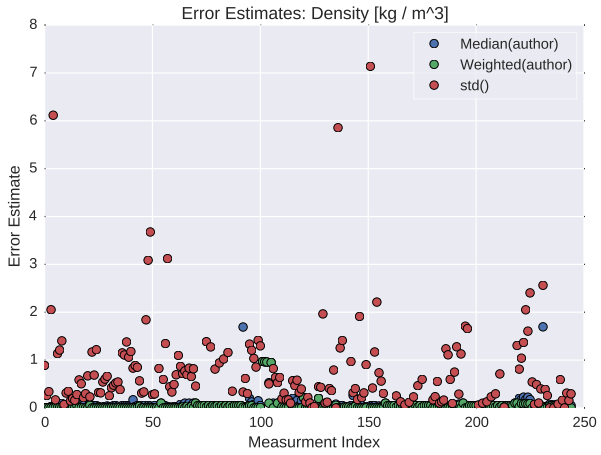



(b)

**FIG. 4. Dependence of computed density on simulation timestep.** To probe the systematic error from finite time-step integration, we examined the timestep dependence of butyl acrylate density. (a). The density is shown for several choices of timestep. (b). The relative error, as compared to the reference value, is shown for several choices of timestep. Error bars represent stand errors of the mean, with the number of effective samples estimated using pymbar's statistical inefficiency routine [27]. The reference value is estimated by linear extrapolation to 0 fs using the 0.5 fs and 1.0 fs data points; the linear extrapolation is shown as black lines. We find a 2 fs timestep leads to systematic biases in the density on the order of 0.19%, while 1 fs reduces the systematic bias to approximately 0.13%—we therefore selected a 1 fs timestep for the present work, where we aimed to achieve three digits of accuracy in density predictions. [JDC: This is really weird. Absent big numerical errors, this timestep dependence should be monotonic. Was this using the GPU in mixed precision?]

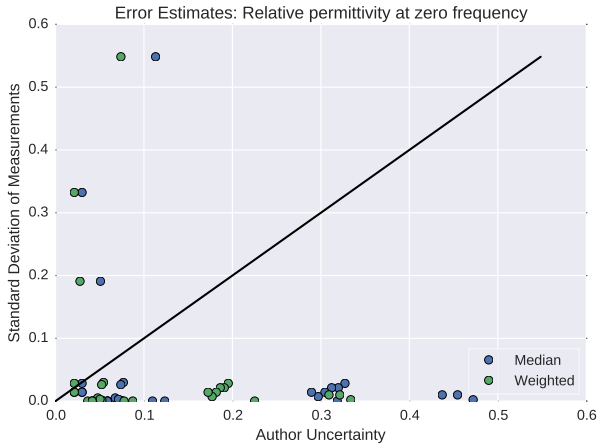



(a)




(b)

**FIG. 5**. **Assessment of experimental error: Density** To assess the experimental error in our ThermoML extract, we compared three difference estimates of uncertainty. In the first approach (Weighted), we computed the standard deviation of the optimally weighted average of the measurements, using the uncertainties reported by authors ($\sigma_{Weighted} = [\sum_k \sigma_k^{-2}]^{-0.5}$). This uncertainty estimator places the highest weights on measurements with small uncertainties and is therefore easily dominated by small outliers and uncertainty under-reporting. In the second approach (Median), we estimated the median of the uncertainties reported by authors; this statistic should be robust to small and large outliers of author-reported uncertainties. In the third approach (Std), we calculated at the standard deviation of independent measurements reported in the ThermoML extract, completely avoiding the author-reported uncertainties. Plot (a) compares the three uncertainty estimates. We see that author-reported uncertainties appear to be substantially smaller than the scatter between the observed measurements. A simple psychological explanation might be that because density measurements are more routine, the authors simply report the accuracy limit of their hardware (e.g. 0.0001 g / mL for a Mettler Toledo DM40 [50]). However, this hardware limit is not achieved due to inconsistencies in sample preparation; see Appendix in Ref. [51]. Panel (b) shows the same information as (a) but as a function of the measurement index, rather than as a scatter plot—because not all measurements have author-supplied uncertainties, panel (c) contains slightly more data points than (a, b). [JDC: Should discuss with Kenneth what kind of story to make out of this, and what to say in the main manuscript body.]




(a)


(b)

**FIG. 6**. **Assessment of experimental error: Static Dielectric Constant** To assess the experimental error in our ThermoML extract, we compared three difference estimates of uncertainty. In the first approach (Weighted), we computed the standard deviation of the optimally weighted average of the measurements, using the uncertainties reported by authors ($\sigma_{Weighted} = [\sum_k \sigma_k^{-2}]^{-0.5}$). This uncertainty estimator places the highest weights on measurements with small uncertainties and is therefore easily dominated by small outliers and uncertainty under-reporting. In the second approach (Median), we estimated the median of the uncertainties reported by authors; this statistic should be robust to small and large outliers of author-reported uncertainties. In the third approach (Std), we calculated at the standard deviation of independent measurements reported in the ThermoML extract, completely avoiding the author-reported uncertainties. Plot (a) compares the three uncertainty estimates. Unlike the case of densities, author-reported uncertainties appear to be somewhat larger than the scatter between the observed measurements. Panel (b) shows the same information as (a) but as a function of the measurement index, rather than as a scatter plot—because not all measurements have author-supplied uncertainties, panel (c) contains slightly more data points than (a, b).

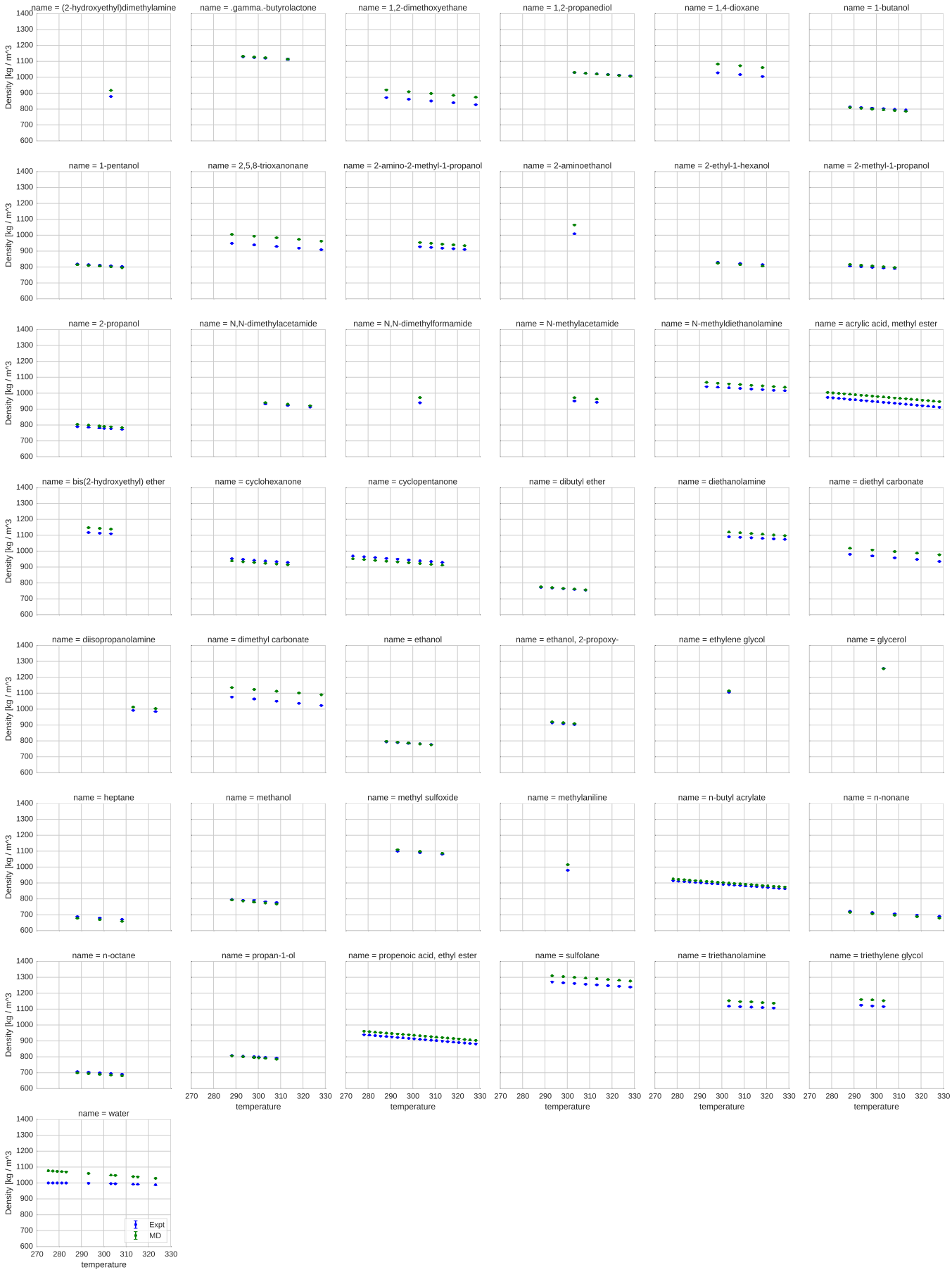



**FIG. 7**. **Comparison of simulated and experimental densities for all compounds.** Measured (blue) and simulated (green) densities are shown in units of kg/m$^3$.


**FIG. 8**. **Comparison of simulated and experimental static dielectric constants for all compounds.** Measured (blue), simulated (green), and polarizability-corrected simulated (red) static dielectric constants are shown for all compounds. Note that dielectric constants, rather than inverse dielectric constants, are plotted here. [JDC: Let's plot these as in Fig. 1 and Fig. 2, maybe only four plots across so they are larger and more legible. We can also shorten "name = compound" to just "compound".] KAB: We should discuss this more before I rebuild this figure several times.

[1] R. Salomon-Ferrer, A. W. GoÌĹtz, D. Poole, S. Le Grand, and R. C. Walker, Journal of Chemical Theory and Computation **9**, 3878 (2013).

[2] K. Lindorff-Larsen, P. Maragakis, S. Piana, M. Eastwood, R. Dror, and D. Shaw, PloS one **7**, e32131 (2012).

[3] D.-W. Li and R. Bruschweiler, J. Chem. Theory Comput. **7**, 1773 (2011).

[4] R. B. Best, X. Zhu, J. Shim, P. E. Lopes, J. Mittal, M. Feig, and A. D. MacKerell, J. Chem. Theory Comput. (2012).

[5] K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. Klepeis, R. Dror, and D. Shaw, Proteins: Struct., Funct., Bioinf. **78**, 1950 (2010).

[6] K. Lindorff-Larsen, S. Piana, R. Dror, and D. Shaw, Science **334**, 517 (2011).

[7] D. Ensign, P. Kasson, and V. Pande, J. Mol. Biol. **374**, 806 (2007).

[8] V. Voelz, G. Bowman, K. Beauchamp, and V. Pande, J. Am. Chem. Soc. **132**, 1526 (2010).

[9] H. Horn, W. Swope, J. Pitera, J. Madura, T. Dick, G. Hura, and T. Head-Gordon, J. Chem. Phys. **120**, 9665 (2004).

[10] C. Caleman, P. J. van Maaren, M. Hong, J. S. Hub, L. T. Costa, and D. van der Spoel, Journal of chemical theory and computation **8**, 61 (2011).

[11] C. J. Fennell, K. L. Wymer, and D. L. Mobley, The Journal of Physical Chemistry B (2014).

[12] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, Nucleic Acids Res. **28**, 235 (2000).

[13] D. L. Mobley, *Experimental and calculated small molecule hydration free energies*, Retrieved from: http://www.escholarship.org/uc/item/6sd403pz, uC Irvine: Department of Pharmaceutical Sciences, UCI.

[14] E. Ulrich, H. Akutsu, J. Doreleijers, Y. Harano, Y. Ioannidis, J. Lin, M. Livny, S. Mading, D. Maziuk, and Z. Miller, Nucleic Acids Res. **36**, D402 (2008).

[15] M. Frenkel, R. D. Chiroco, V. Diky, Q. Dong, K. N. Marsh, J. H. Dymond, W. A. Wakeham, S. E. Stein, E. Königsberger, and A. R. Goodwin, Pure and applied chemistry **78**, 541 (2006).

[16] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case, J. Comput. Chem. **25**, 1157 (2004).

[17] J. Wang, W. Wang, P. A. Kollman, and D. A. Case, J. Mol. Graph Model. **25**, 247260 (2006).

[18] A. Jakalian, B. L. Bush, D. B. Jack, and C. I. Bayly, J. Comput. Chem. **21**, 132 (2000).

[19] A. Jakalian, D. B. Jack, and C. I. Bayly, J. Comput. Chem. **23**, 1623 (2002).

[20] L. Martínez, R. Andrade, E. G. Birgin, and J. M. Martínez, Journal of computational chemistry **30**, 2157 (2009).

[21] *Openeye toolkits 2014*, URL http://www.eyesopen.com.

[22] D. Case, V. Babin, J. Berryman, R. Betz, Q. Cai, D. Cerutti, T. Cheatham III, T. Darden, R. Duke, H. Gohlke, et al., University of California, San Francisco (2014).

[23] URL http://github.com/choderalab/gaff2xml.

[24] P. Eastman, M. S. Friedrichs, J. D. Chodera, R. J. Radmer, C. M. Bruns, J. P. Ku, K. A. Beauchamp, T. J. Lane, L.-P. Wang, D. Shukla, et al., J. Chem. Theory Comput. **9**, 461 (2012).

[25] R. T. McGibbon, K. A. Beauchamp, C. R. Schwantes, L.-P. Wang, C. X. Hernández, M. P. Harrigan, T. J. Lane, J. M. Swails, and V. S. Pande, bioRxiv p. 008896 (2014).

[26] T. Darden, D. York, and L. Pedersen, J. Chem. Phys. **98**, 10089 (1993).

[27] M. R. Shirts and J. D. Chodera, J. Chem. Phys. **129**, 124105 (2008).

[28] N. Haider, Molecules **15**, 5079 (2010).

[29] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, The Journal of chemical physics **79**, 926 (1983).

[30] W. L. Jorgensen, J. D. Madura, and C. J. Swenson, Journal of the American Chemical Society **106**, 6638 (1984).

[31] R. Bosque and J. Sales, Journal of chemical information and computer sciences **42**, 1154 (2002).

[32] K. Sheppard, *Arch toolbox for python* (2015), GitHub repository: https://github.com/bashtage/arch, URL http://dx.doi.org/10.5281/zenodo.15681.

[33] H. Flyvbjerg and H. G. Petersen, J. Chem. Phys. **91**, 461 (1989).

[34] L.-P. Wang, T. J. Martínez, and V. S. Pande, The Journal of Physical Chemistry Letters (2014).

[35] C. J. Fennell, L. Li, and K. A. Dill, The Journal of Physical Chemistry B **116**, 6936 (2012).

[36] I. V. Leontyev and A. A. Stuchebrukhov, The Journal of chemical physics **141**, 014103 (2014).

[37] A. D'Aprano and I. D. Donato, Journal of Solution Chemistry **19**, 883 (1990).

[38] W. M. Haynes, *CRC handbook of chemistry and physics* (CRC Press, 2011).

[39] D. L. Mobley, *Experimental and calculated small molecule hydration free energies*, Retrieved from: https://github.com/choderalab/FreeSolv, uC Irvine: Department of Pharmaceutical Sciences, UCI.

[40] J.-F. Truchon, A. Nicholl's, J. A. Grant, R. I. Iftimie, B. Roux, and C. I. Bayly, Journal of computational chemistry **31**, 811 (2010).

[41] J.-F. Truchon, A. Nicholls, B. Roux, R. I. Iftimie, and C. I. Bayly, Journal of chemical theory and computation **5**, 1785 (2009).

[42] J.-F. Truchon, A. Nicholls, R. I. Iftimie, B. Roux, and C. I. Bayly, Journal of chemical theory and computation **4**, 1480 (2008).

[43] J. Ponder, C. Wu, P. Ren, V. Pande, J. Chodera, M. Schnieders, I. Haque, D. Mobley, D. Lambrecht, R. DiStasio Jr, et al., J. Phys. Chem. B **114**, 2549 (2010).

[44] P. Ren and J. W. Ponder, The Journal of Physical Chemistry B **108**, 13427 (2004).

[45] G. Lamoureux and B. Roux, The Journal of Chemical Physics **119**, 3025 (2003).

[46] V. M. Anisimov, G. Lamoureux, I. V. Vorobyov, N. Huang, B. Roux, and A. D. MacKerell, Journal of Chemical Theory and Computation **1**, 153 (2005).

[47] L.-P. Wang, T. L. Head-Gordon, J. W. Ponder, P. Ren, J. D. Chodera, P. K. Eastman, T. J. Martínez, and V. S. Pande, J. Phys. Chem. B **117**, 9956 (2013).

[48] M. Frenkel, R. D. Chirico, V. V. Diky, Q. Dong, S. Frenkel, P. R. Franchois, D. L. Embry, T. L. Teague, K. N. Marsh, and R. C. Wilhoit, Journal of Chemical & Engineering Data **48**, 2 (2003).

[49] R. D. Chirico, M. Frenkel, V. V. Diky, K. N. Marsh, and R. C. Wilhoit, Journal of Chemical & Engineering Data **48**, 1344 (2003).

[50] *Mettler toledo density meters*, [Online; accessed 15-Jan-2015], URL http://us.mt.com/us/en/home/products/Laboratory_Analytics_Browse/Density_Family_Browse_main/DE_Benchtop.tabs.models-and-specs.html.

[51] R. D. Chirico, M. Frenkel, J. W. Magee, V. Diky, C. D. Muzny, A. F. Kazakov, K. Kroenlein, I. Abdulagatov, G. R. Hardin, and W. E. Acree Jr, Journal of Chemical & Engineering Data **58**, 2699 (2013).