# 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,‡]
Arien S. Rustenburg,[3,§] Kenneth Kroenlein,[4,¶] and John D. Chodera[1,**]

[1]*Computational Biology Program, 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]*Thermodynamics Research Center, NIST, Boulder, CO*

(Dated: February 23, 2015)

Useful atomistic simulations in the condensed phase require accurate depictions of solvent. 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 liquid densities and static dielectric constants at ambient pressure) automatically extracted from the archive, and discuss the extent of available data for neat liquids. 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*

## I. INTRODUCTION

Recent advances in hardware and software for molecular dynamics simulation now permits routine access to atomistic simulations at the 100 ns timescale and beyond. [JDC: Cite something here, like the Amber "routine microsecond" paper? http://pubs.acs.org/doi/abs/10.1021/ct400314y]. 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 [1] and enhancements [2–4], with key outcomes including the ability to fold fast-folding proteins [JDC: Cite Pande and Shaw papers?], improved fidelity of water thermodynamic properties [18], and improved prediction of NMR observables. Although small molecule forcefields have also been the subject of benchmarks [5] and improvements [6], 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 [6]. 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 [7], FreeSolv [8], and the BMRB [9]). While this inconvenience is relatively minor for benchmarking forcefield accuracy for a single target (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 [10]—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 (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.

* 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

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) [11] with the AM1-BCC charge model [12, 13] and identify systematic biases to aid further forcefield refinement.

## II. RESULTS

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

We retrieved a copy of the ThermoML Archive from the NIST TRC (`http://trc.nist.gov/ThermoML.html` accessed 13 Sep 2014) and performed a number of sequential filtering steps to produce an extract of the ThermoML Archive relevant for benchmarking forcefields describing small organic molecules. [JDC: This is the date I had on the `ThermoML.tar.gz` archive in GitHub. We should check to make sure this is accurate.] 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 the 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 [K] $(270 \leq T \leq 330)$
5. The measurement was performed at ambient pressure [kPA] $(100 \leq P \leq 102)$
6. Measured densities below 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

The temperature and pressure rounding step was motivated by common data reporting variations; for example, an experiment performed at water's freezing point at 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 one atmosphere. Temperatures were rounded to one decimal place. [JDC: Does this reflect the accuracy of reporting ambient temperatures?] 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.

| | Number of measurements remaining | |
|---|---|---|
| **Filter step** | **Mass density** | **Static dielectric** |
| 1. Single Component | 130074 | 1649 |
| 2. Druglike Elements | 120410 | 1649 |
| 3. Heavy Atoms | 67897 | 1567 |
| 4. Temperature | 36827 | 962 |
| 5. Pressure | 13598 | 461 |
| 6. Liquid state | 13573 | 461 |
| 7. Aggregate T, P | 3573 | 432 |
| 8. Density+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.

[JDC: It might be useful to point the users to the scripts that were used to do this extraction. Also, can we automate the downloading of the complete up-to-date archive, perhaps with Kenneth's help in identifying the least intrusive way to do so?]

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

[JDC: If we lead with a Results section before Methods, we have to start with a small summary of the calculation. We should tell readers the salient details—we ran simulations with a small timestep to minimize integrator error, we used stochastic thermal and pressure control, and we used an adaptive simulation scheme that ensured simulations ran long enough to achieve target accuracy. We can also mention that we used OpenMM, but these calculations can easily be adapted to other codes.]

#### 1. *Mass density*

Mass density has been widely used for parameterizing and testing forcefields, particularly the Lennard-Jones parameters representing dispersive and repulsive interactions [15, 16]. We therefore used the present ThermoML extract as a benchmark of the GAFF/AM1-BCC forcefield (Fig. 1).

[JDC: Remind readers how mass density is computed.]

Overall, the densities show reasonable accuracy, with a root-mean square (RMS) relative error over all measure-

| **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 [14]) is summarized here.

ments of 3±0.1% (with one standard error of the mean determined by bootstrapping over all measurements), especially encouraging given that this forcefield was not designed with the intention of modeling bulk liquid properties of organic molecules [11] [JDC: Sig figs issue—this should be 3.x±0.1%.] This is reasonably consistent with previous studies reporting agreement of 4% on a different benchmark set [5]. [JDC: Did that previous study report an uncertainty?]

[JDC: Discuss outliers here. There must be more things we can say about densities. Some of the densities are quite good, while others seem poor, with systematic bias toward higher densities than experiment. We can also point out that densities at different temperatures for a given molecule seem to be biased in a consistent way.]

### 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 [17]. [JDC: I've commented out the equation because I don't think it is central to our point. Essentially, it is just an atom-based linear model for computing the molecular polarizability, and the parameters can be looked up elsewhere.] 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 [18]

$$\Delta\epsilon = 4\pi N \frac{\alpha}{\langle V \rangle} \tag{1}$$

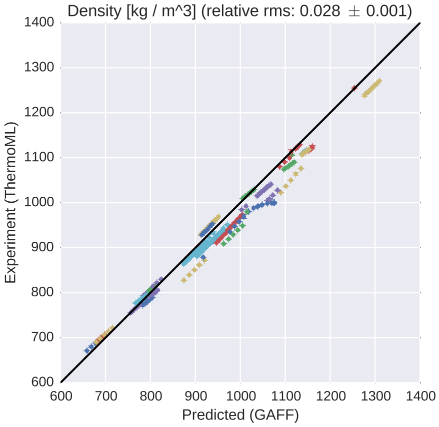

**FIG. 1**. **Comparison of liquid densities between experiment and simulation.** Liquid density measurements extracted from ThermoML are compared against densities predicted using the GAFF/AM1-BCC small molecule fixed-charge forcefield. Color groupings represent identical chemical species. Simulation 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 Appendix B for further discussion of error.

While a similar polarization correction was used in the development of the TIP4P-Ew water model, where it had a minor effect [18], 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. [JDC: What were the results for nonpolar organic molecules?] For comparison, we also applied the same empirical correction to the VirtualChemistry dataset [5, 19] and saw similarly improved agreement with experiment for both the GAFF and OPLS forcefields (Fig. 7). [JDC: Not sure if we should keep the VirtualChemistry stuff here other than to compare our computed pipeline results with Virtual Chemistry as validation to show we didn't screw up our pipeline.]

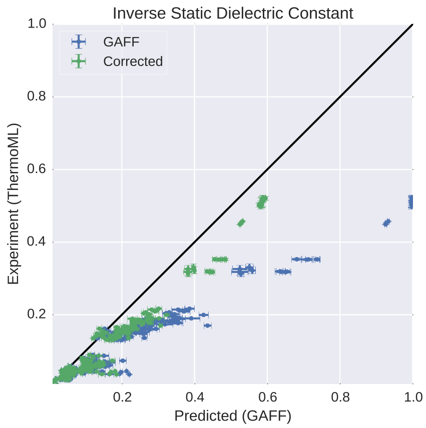

**FIG. 2. Measured (ThermoML) versus predicted (GAFF/AM1-BCC) inverse static dielectrics (a).** Simulation error bars represent one standard error of the mean estimated via block averaging with block sizes of 200 ps [20]. [JDC: Why are we using block averaging here? Why didn't we just use `timeseries.py`. We should not be using block averaging, especially without a justification that 200 ps is a reasonable block size for every specific system and condition Let's talk about this.] 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 Section B 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. $U(r) \propto q_1 q_2 / r \propto \epsilon^{-1}$]. [JDC: We need to trim the whitespace of all sides of the figures that you are outputting in order for the figure to actually fill the column width. There must be some option to set that. See the `figure.tight_layout()` option in `matplotlib`, along with `matplotlib.backends.backend_pdf.PdfPages`.]

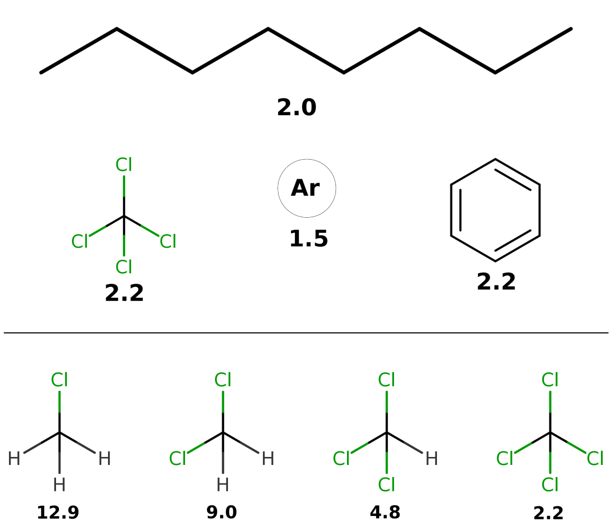

**FIG. 3. Typical experimental static dielectric constants of some nonpolar compounds.** (a). Measured static dielectric constants of various nonpolar or symmetric molecules [**?** ]; fixed-charge forcefields give $\epsilon \approx 1$ for each species. (b). A congeneric series of chlorosubstituted methanes have static dielectric constants between 2 and 13. [Can we use a better citable source for these numbers instead of Wikipedia? Also, what temperatures/pressures are these measurements cited at? Maybe we can just say "near ambient"?] [JDC: We should not use PNG files for figure graphics—only vector graphics (when possible). Can you use a vector graphics PDF instead?] [JDC: Can we show both experimental and GAFF/AM1-BCC computed dielectric constants for some of these compounds?]

## III. DISCUSSION

### A. Fitting Forcefields to Dielectric Constants

Recent forcefield development has seen a resurgence of papers fitting dielectric constants during forcefield parameterization [6, 21]. However, a number of authors have pointed out potential challenges in constructing self-consistent fixed-charge forcefields [22, 23].

Interestingly, recent work by Dill and coworkers [22] observed that, for $\mathrm{CCl_4}$, 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 unanticipated consequences. [JDC: Perhaps the free energy of binding to hydrophobic cavities in proteins could be relevant?]

Suppose, for example, that one attempts to fit forcefield parameters to match the static dielectric constants of $\mathrm{CCl_4}$, $\mathrm{CHCl_3}$, $\mathrm{CH_2Cl_2}$, and $\mathrm{CH_3Cl}$. In moving from the tetrahedrally-symmetric $\mathrm{CCl_4}$ to the asymmetric $\mathrm{CHCl_3}$, it suddenly becomes possible to achieve the observed dielectric constant of 4.8 by an appropriate choice of point charges. However, the model for $\mathrm{CHCl_3}$ uses fixed point charges to account for *both* the permanent dipole moment and the electronic polarizability, whereas the $\mathrm{CCl_4}$ 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 $\mathrm{CCl_4}$) have qualitatively different properties than closely related asymmetric molecules (e.g. toluene and $\mathrm{CHCl_3}$).

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. The Onsager model for the transfer free energy of a dipole (Eq. 2) gives an error of $\Delta\Delta G = \Delta G(\epsilon = 2.2) - \Delta G(\epsilon = 1)$ of $-2$ kcal / mol for the transfer of water ($a = 1.93$ Å, $\mu = 2.2$D) into a low dielectric medium such as tetrachloromethane or benzene.

$$\Delta G = -\frac{\mu^2}{a^3}\frac{\epsilon - 1}{2\epsilon + 1} \qquad (2)$$

Similarly, we calculated the mean polarization error for solvation free energies (gas to solvent transfer free energies) of druglike molecules in cyclohexane. For each molecule in the FreeSolv database [8] [JDC: Which version of FreeSolv?], we 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 $-0.9 \pm 0.07$ kcal / mol for the 643 molecules (where the standard error is computed from bootstrapping over measurements), [JDC: Need to fix reported sig figs: $-0.9x \pm 0.07$.] 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.

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 [24–26], point dipole methods [27, 28], and Drude methods [29, 30] 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 [31] may provide polarizability physics at a cost not much greater than fixed charge forcefields.

### B. ThermoML as a data source

The present work has focused on the neat liquid density and dielectric measurements present in the ThermoML Archive [10, 32, 33] 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.

## IV. METHODS

### A. ThermoML Processing

A tarball archive of the ThermoML Archive was obtained from the the NIST TRC on 13 Sep 2014. 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 and Pandas was used to extract the data and apply the successive data filters described in Section II A.

### B. Simulation

Using an automated tool, boxes of 1000 molecules were constructed using PackMol [34] [JDC: Which version?]. AM1-BCC [12, 13] charges were generated using OpenEye Toolkit 2014-6-6 [35], 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 in AmberTools 14 [36]. The resulting AMBER files were converted to OpenMM [37] ffxml forcefield XML files. Simulation code used libraries gaff2xml 0.6, TrustButVerify 0.1, OpenMM 6.2 [37], and MDTraj 1.2 [38]. [TODO: Provide a script to install all of these versions via `conda`.]

Molecular dynamics simulations were performed with OpenMM 6.2 [37] 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 III). [JDC: Cite Langevin integrator used in OpenMM.] 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 [39] 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 density standard errors were less than $2 \times 10^{-4}$ g / mL, as estimated using the equilibration detection module in pymbar 2.1 [40]. Trajectory analysis was performed using OpenMM [37] and MDTraj [38]. [JDC: Which versions?] Instantaneous densities were stored every 250 fs, while trajectory snapshots were stored every 10 ps. [JDC: Did we plan to make this data available somewhere, or is it sufficient to put out the scripts?]

## 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

## 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.]

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

## Appendix B: Assessment of experimental error in ThermoML measurements

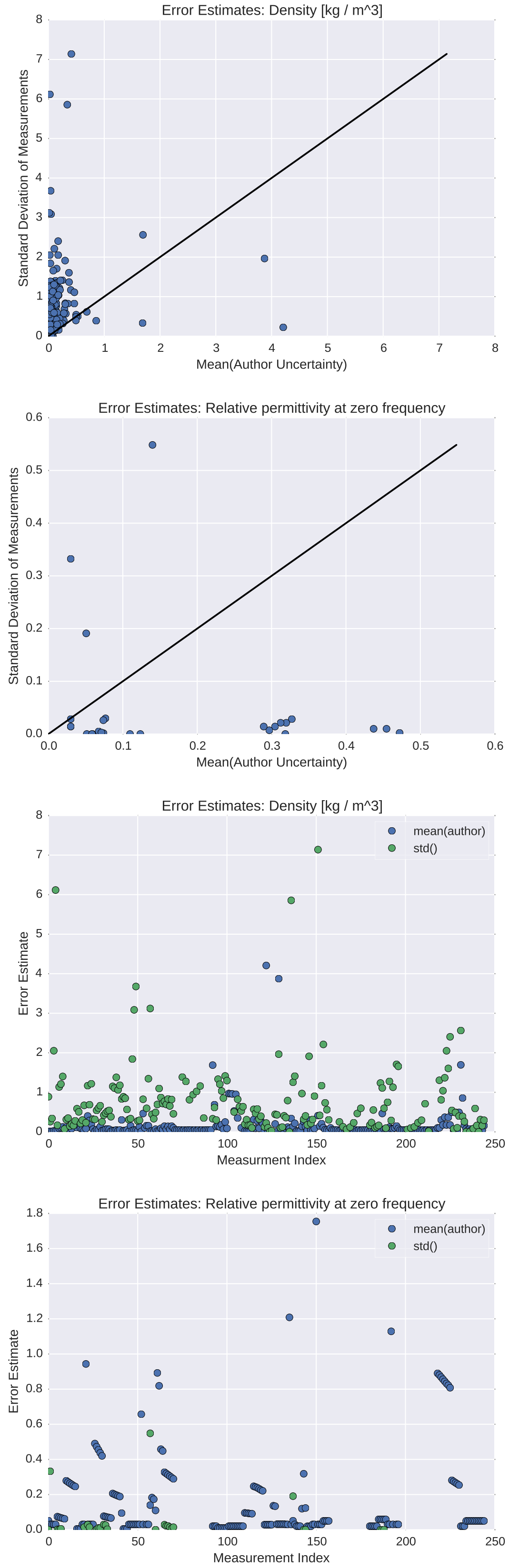

**FIG. 4. Assessment of experimental error in ThermoML data.** To assess the experimental error in our ThermoML extract, we considered two different approaches. In the first approach, we computed the mean of the uncertainties reported by the measurement authors. [JDC: This is an incorrect way to combine uncertainties. If you take the unweighted mean $\hat{x} = N^{-1}\sum_i x_i$ of $N$ experimental measurements $x_i$ with associated standard errors or uncertainties $\sigma_i$, the resulting uncertainty is $\hat{\sigma} = N^{-1}(\sum_i \sigma_i^2)^{1/2}$—the procedure you suggest where uncertainties are simply averaged is incorrect and should not be used. But I think we should

| $\Delta t$ | $\langle\rho\rangle$ (g/cm$^3$) | n | neff | stddev($\rho$) | stderr | abs error (g/cm$^3$) | rel error (%) |
|---|---|---|---|---|---|---|---|
| 0.5 | 0.903701 | 145510 | 20358.0 | 0.007362 | 0.000052 | 0.000000 | 0.0000 |
| 1.0 | 0.903114 | 159515 | 21988.5 | 0.007415 | 0.000050 | -0.000588 | -0.0650 |
| 2.0 | 0.901811 | 108346 | 15964.1 | 0.007494 | 0.000059 | -0.001891 | -0.2092 |

**TABLE III**. **Timestep dependence in computed equilibrium density of butyl acrylate.** To probe the systematic error from finite timestep integration, we examined the timestep dependence of butyl acrylate density. The number of effective samples was estimated using pymbar's statistical inefficiency routine [40]. To approximate the timestep bias, we compare the density expectation ($\langle\rho\rangle$) to values calculated with a 0.5 fs timestep. We find a 2 fs timestep leads to systematic biases in the density on the order of 0.2%, while 1fs reduces the systematic bias to less than 0.1%—we therefore selected a 1 fs timestep for the present work, where we aimed to achieve three digits of accuracy in density predictions. [JDC: I've reformatted this table a bit, paying more attention to sig figs. I think this might actually be better presented as a figure showing the timestep dependence, perhaps for 4 or 5 timesteps from 0.5 to 2.5 fs, rather than just 3.]

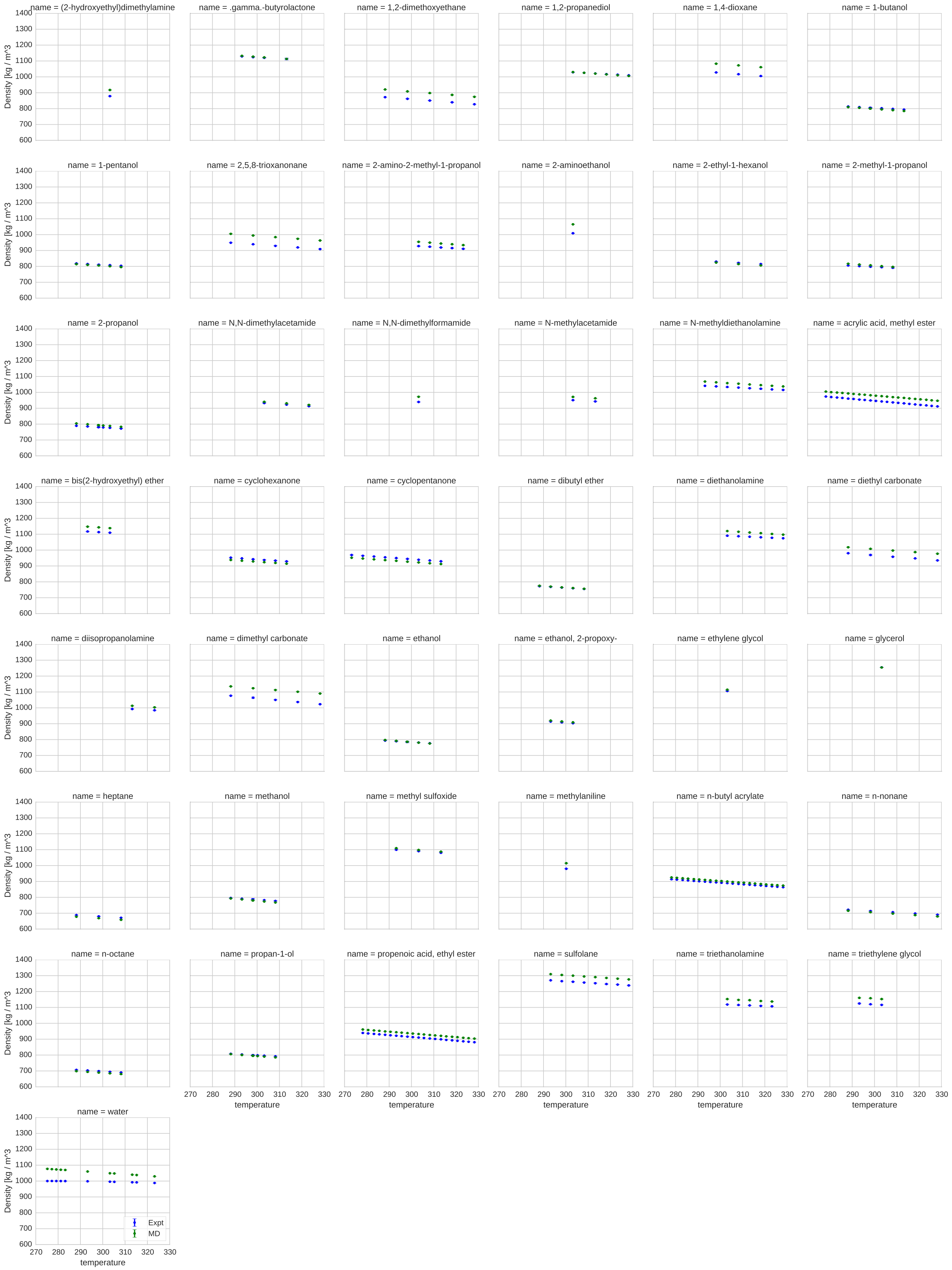

**FIG. 5. Comparison of simulated and experimental densities for all compounds.** Measured (blue) and simulated (green) densities are shown in units of kg/m$^3$.

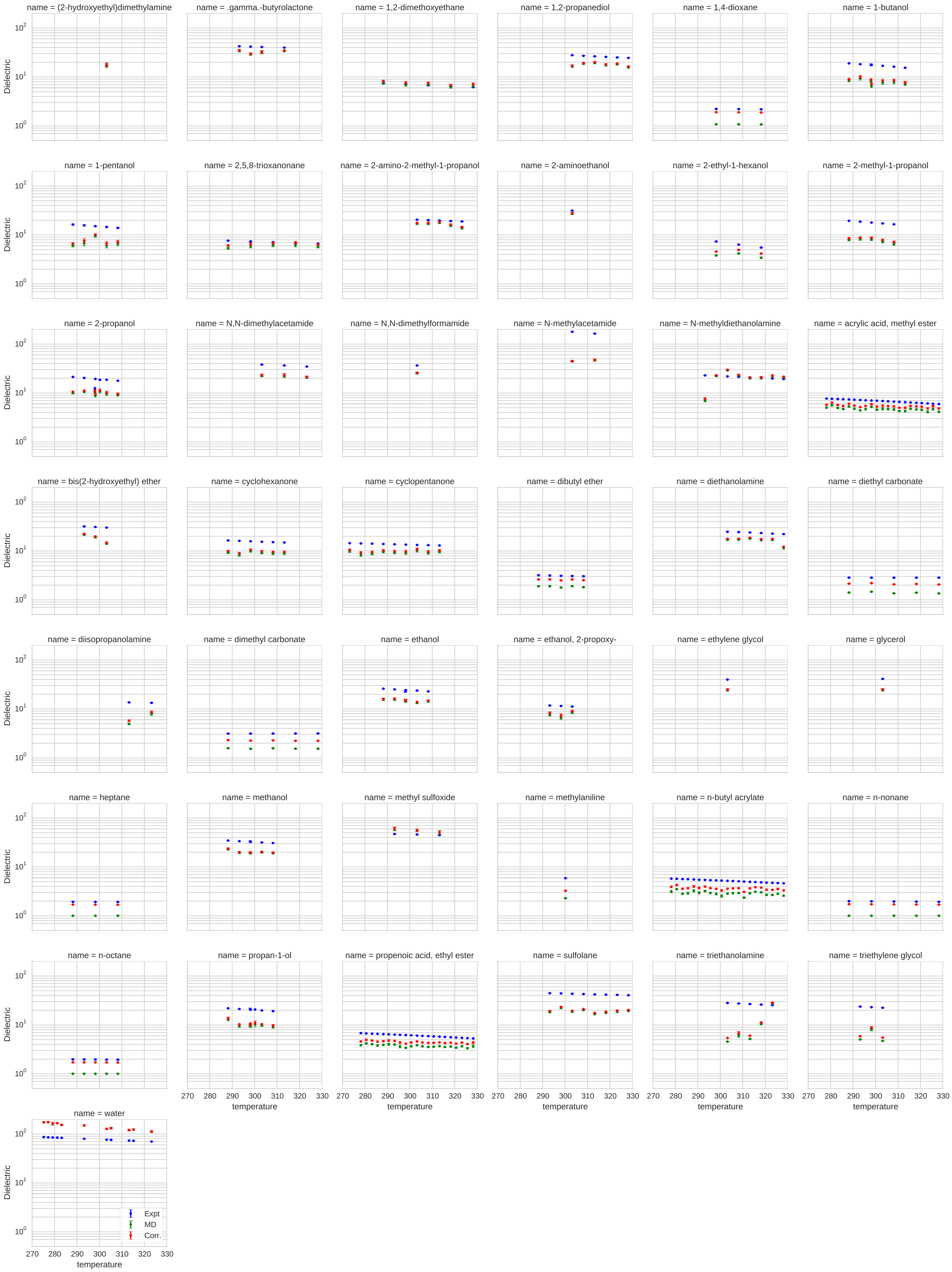

**FIG. 6**. **Comparison of simulated and experimental static dielectric constants for all compounds.** Measured (blue), simulated (green), and polarizability-corrected simulated (red) static dielectric constants are shown for all compounds. Note that dielectric constants, rather than inverse dielectric 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".]

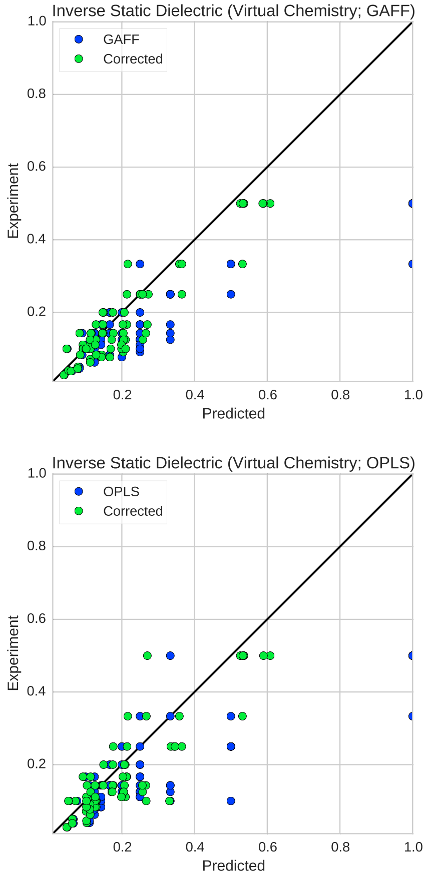

**FIG. 7. Atomic polarizability corrections for the Virtual Chemistry dataset.** A comparison of measured (*blue*) and simulated dielectric constants from the Virtual Chemistry dataset [5, 19]. with (*red*) and without (*green*) atomic polarizability corrections are shown. [JDC: What from the Virtual Chemistry dataset did we end up using? The computed values? The parameterized files? Since the story is about the ThermoML Archive, do we really want this here, or do we just want to do a comparison between our computed results for some molecules and the Virtual Chemistry results as a validation or sanity check on our computational pipeline?] [JDC: Plot needs to be adjusted to be more legible if we decide to keep it.]

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

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

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

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

[5] 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).

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

[7] 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).

[8] 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.

[9] 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).

[10] 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).

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

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

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

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

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

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

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

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

[19] D. van der Spoel, P. J. van Maaren, and C. Caleman, Bioinformatics **28**, 752 (2012).

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

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

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

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

[24] 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).

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

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

[27] 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).

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

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

[30] 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).

[31] 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).

[32] 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).

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

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

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

[36] 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).

[37] 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).

[38] 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).

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

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

[41] *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.

[42] 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).