Benchmarking Simulations against the ThermoML Database: Neat Liquid Densities and Static Dielectrics

Kyle A. Beauchamp[+,†] Julie M. Behr[+,†] Patrick B. Grinaway ,[†] Arien S. Rustenburg,[†] Kenneth Kroenlein,[‡] and John D. Chodera[*,†]

Memorial Sloan-Kettering Cancer Center, New York, NY, and NIST Thermodynamics Research Center, Boulder, CO

E-mail: john.chodera@choderalab.org

Abstract

Useful atomistic simulations require accurate depictions of solvent. Simple experimental observables, such as density and static dielectric constants, offer straightforward targets for evaluating forcefield quality. Here we examine the possibility of benchmarking atomistic models against the NIST ThermoML database of physicochemical measurements, which curates thousands of density, dielectric, and other measurements. We present a detailed benchmark of the GAFF forcefield against measurements extracted from ThermoML and discuss the extent of available data for neat liquids. We show that empirical polarizability models correct systematic biases inherent in predicting dielectric constants with fixed-charged forcefields.

[*]To whom correspondence should be addressed
[†]Memorial Sloan-Kettering Cancer Center, New York, NY
[‡]NIST Thermodynamics Research Center, Boulder, CO

Introduction

Recent advances in hardware and molecular dynamics software have provided routine access to 100 ns atomistic simulations. Leveraging these advances in combination with GPU clusters, distributed computing, or custom hardware has brought the microsecond and milliseconds within reach. These dramatic advances in sampling, however, have revealed forcefields as a critical barrier for truly predictive simulation.

Protein and water forcefields have been the subject of numerous benchmarks and enhancements, with key outcomes including the ability to fold fast-folding proteins, improved fidelity of water thermodynamic properties, and improved prediction of NMR observables. Although small molecule forcefields have also been the subject of benchmarks and improvements, such work has 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 dielectrics and hydration free energies. Other studies have found XYZ. There are also outstanding questions of generalizability of parameters. Will changes to a specific chemical moiety be compatible with orthogonal improvements? Addressing these questions requires agreement on shared benchmarks that can be easily replicated with proposed forcefield enhancements.

A key barrier in forcefield development is that many experimental datasets are heterogeneous, paywalled, and unavailable in machine-readable formats (although some counterexamples exist, such as FreeSolv and the BMRB). While this inconvenience is relatively minor for benchmarking a single target (e.g. water), it becomes prohibitive for studies spanning chemical space. To ameliorate problems of data archival, the NIST Thermodynamics Research Center has developed a IUPAC standard XML-based format—ThermomL—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 now guided through a data archival process that involves sanity checks and eventual archival

at the TRC (http://trc.nist.gov/ThermoML.html).

Here we examine the ThermoML archive as a potential source for neat liquid density and static dielectric measurements, with the goal of developing a standard benchmark for validating these properties in forcefields. These two observables provide sensitive tests of forcefield accuracy that are nonetheless straightforward to calculate. Using the ThermoML data, we evaluate the AM1-BCC GAFF forcefield and identify systematic biases that might be improved upon.

Results

Neat Liquid Measurements in ThermoML

We performed a number of queries to summarize the ThermoML content relevant for benchmarking organic molecule forcefields. Our aim is to explore neat liquid data with functional groups relevant to drug-like molecules. We therefore applied the following sequence of filters: has either density or static dielectric measurements, contains a single component, contains only druglike elements (H, N, C, O, S, P, F, Cl, Br), has low heavy atom count (≤ 10), has ambient temperature [K] ($270 \leq T \leq 330$), and has ambient pressure [kPA] ($100 \leq P \leq 102$). After applying these filters, we also assume that all pressures within this range are one atmosphere. We also assume that temperatures can be rounded to one decimal place. These approximations are motived by common data entry errors; 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. After the application of these filters (Table 1), we are left with 245 conditions for which both density and dielectric data are available.

Table 1: ThermoML Statistics

Filter	Mass Density	Static Dielectric
0. Single Component	130074	1649
1. Druglike Elements	120410	1649
2. Heavy Atoms	67897	1567
3. Temperature	36827	962
4. Pressure	13598	461
5. Aggregate T, P	**3591**	**432**
6. Density+Dielectric	**245**	**245**

Benchmarking GAFF against ThermoML: Mass Density

Mass density has been widely used as a critical ingredient for parameterizing and testing forcefields, particularly the Lennard Jones parameters. We therefore used the present ThermoML compilation as a benchmark of the Generalized Amber Force Field.

Benchmarking GAFF against ThermoML: Static Dielectric

As a measure of the electronic medium, the static dielectric constant of neat liquids provides a critical benchmark that is orthogonal to density and thermodynamic quantities. We therefore compare our GAFF simulations against the measurements in our ThermoML compilation. Overall, we find the dielectric constants to be qualitatively reasonable, but with clear deviations from experiment. In particular, the nonpolar organics show a clear discrepancy, with the GAFF predictions of 1.0 being substantially less polar than the measurements near 2.0. Because this deviation likely stems from the lack of electronic polarization, we added a simple empirical correction for polarization,[1] which leads to better agreement with experiment. A similar polarization correction was used in the development of the TIP4P-EW water model;[2] however, the need is much greater for the nonpolar organics, as the missing polarizability is the dominant contribution to the static dielectric constant. In the case of water, the Sales polarizability model predicts a dielectric correction of 0.52, while 0.79 was used for the TIP4P-EW model. For comparison, we also

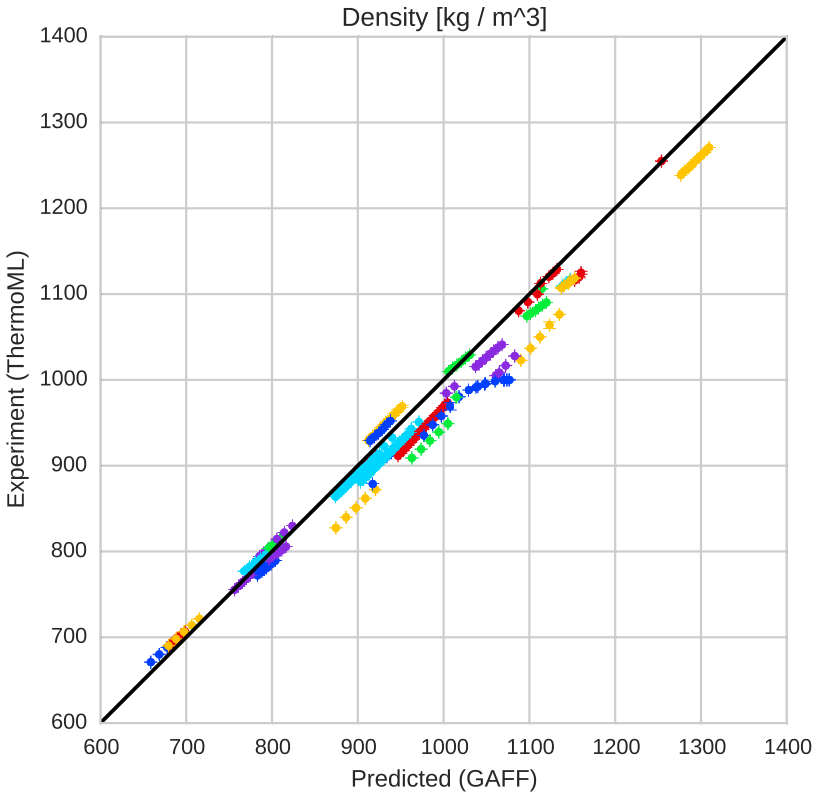


Figure 1: Measured (ThermoML) versus predicted (GAFF) densities. Color groupings represent identical chemical formulas. Simulation error bars represent one standard error of the mean, with the number of effective (uncorrelated) samples estimated using pymbar. Experimental error bars indicate the standard deviation between independently reported measurements, when available, or the authors reported standard deviations; for some measurements, neither uncertainty estimate is available. See section S2 for further discussion of error.

applied the same empirical correction to the VirtualChemistry dataset[3,4] and saw similarly improved agreement with experiment for both the GAFF and OPLS forcefields (Fig. S7).

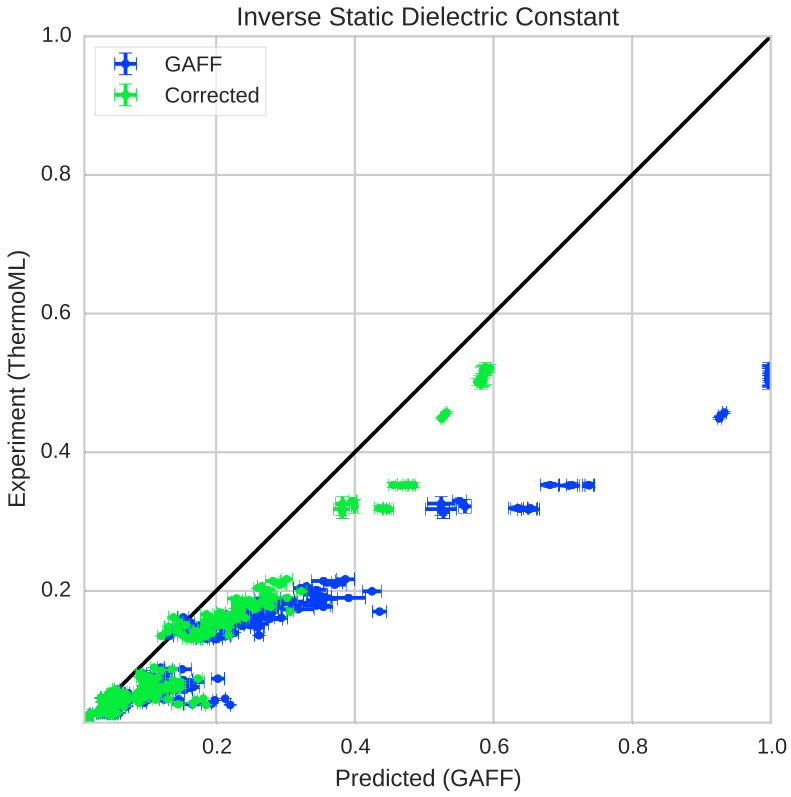
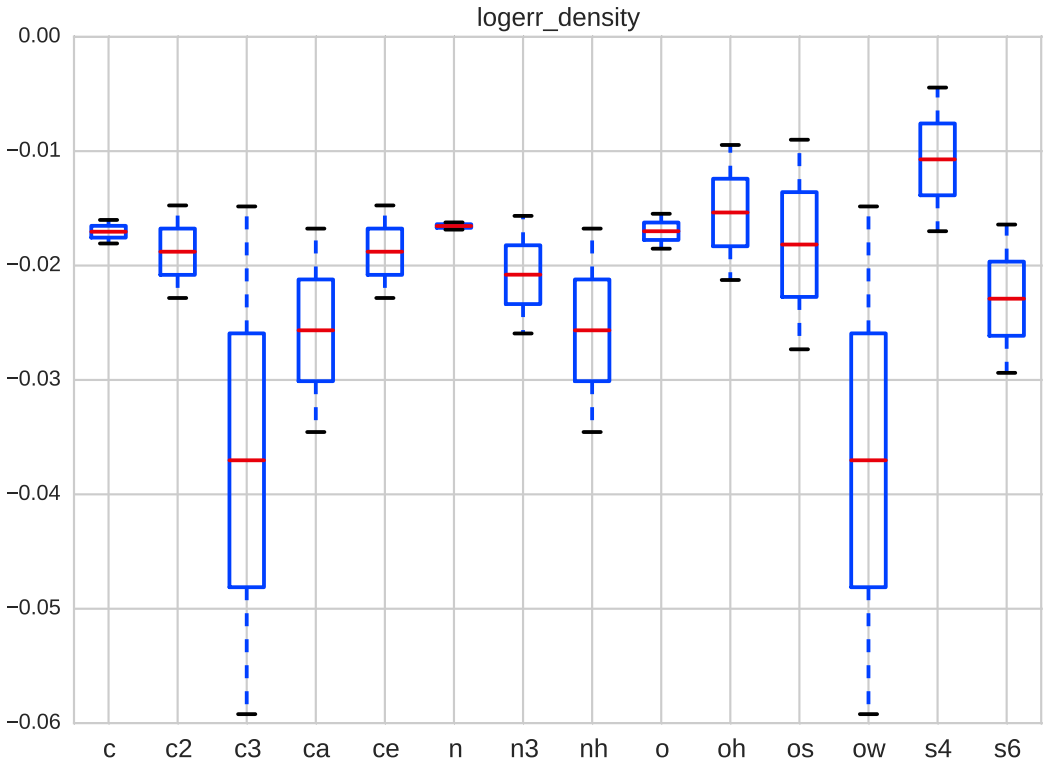
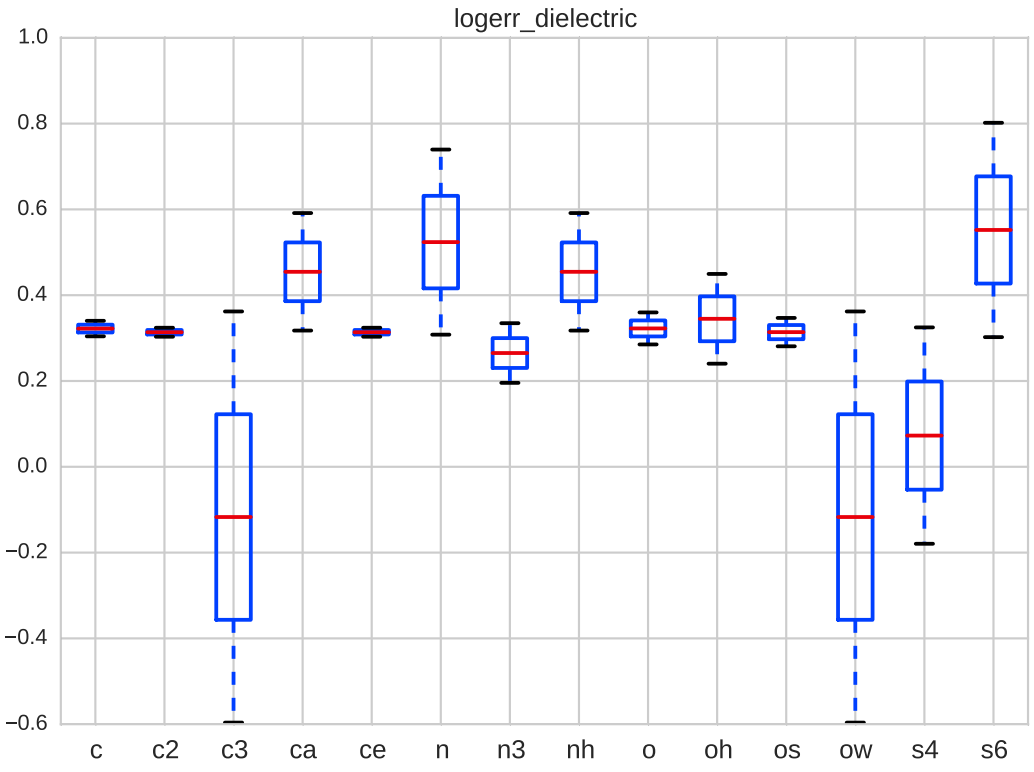


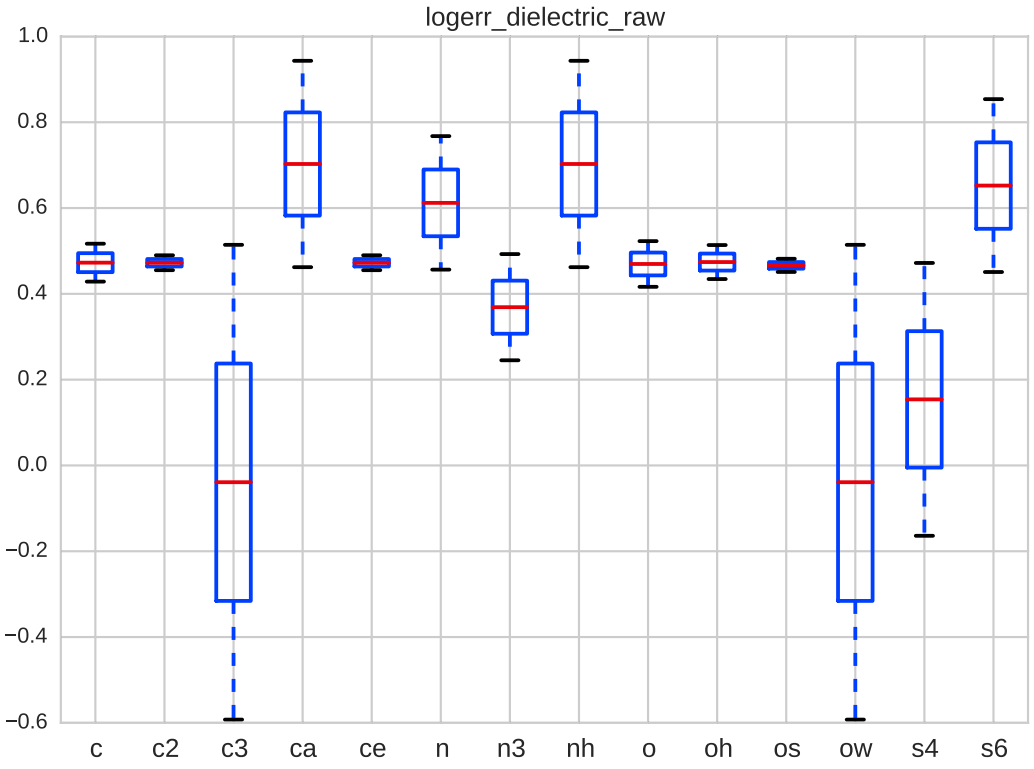
Figure 2: Measured (ThermoML) versus predicted (GAFF) static dielectrics (a). Color groupings represent identical chemical formulas. Simulation error bars represent one standard error of the mean estimated via block averaging with block sizes of 200 ps.[5] Experimental error bars indicate the larger of standard deviation between independently reported measurements and the authors reported standard deviations; for some measurements, neither uncertainty estimate is available. See section S2 for further discussion of error.

Discussion

Forcefield Accuracy Depends on Functional Group




logerr_dielectric
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
c
c2
c3
ca
ce
n
n3
nh
o
oh
os
ow
s4
s6


logerr_dielectric_raw
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
c
c2
c3
ca
ce
n
n3
nh
o
oh
os
ow
s4
s6

Fitting Forcefields to Dielectric Constants

Recent forcefield development has seen a resurgence of papers fitting dielectric constants as primary data.[6,7] However, a number of authors have pointed out potential challenges in constructing self-consistent fixed-charge force fields.[8,9] Interestingly, a recent work by Dill[8] pointed out that, for CCl_4, reasonable choices of point charges are incapable of recapitulating the observed dielectric of 2.2, instead producing dielectric constants in the range of $1.0 \leq \epsilon \leq 1.05$. Suppose, for example, that one attempts to directly fit the static dielectric constants of CCl_4, $CHCl_3$, CH_2Cl_2, CH_3Cl_1, CH_4. In moving from the tetrahedrally-symmetric CCl_4 to $CHCl_3$, it suddenly becomes possible to achieve the observed dielectric constant of 4.8. However, the model for $CHCl_3$ uses fixed point charges to account for *both* the net 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, CCl_4) have qualitatively different properties than closely related asymmetric molecules (e.g. toluene, $CHCl_3$). As a first-order fix, we suggest using empirical polarization corrections before directly comparing measured static dielectric constants to fixed-charge models—particularly when examining low-dielectric solvents. Separating the contributions of fixed charges and polarization may also lead to the development of improved models of electrostatics that account for the missing polarization physics.

ThermoML as a Data Source

The present work has focused on the neat liquid density and dielectric measurements present in ThermoML[10–12] as a target for molecular dynamics forcefield validation. While densities and dielectric constants have been widely used in forcefield work, several aspects of ThermoML make it a unique resource for the forcefield community. First, the curation, support, and dissemination of ThermoML is supported by NIST, whose mis-

sion makes these tasks a long-term priority. Second, ThermoML is actively growing, through partnerships with journals such as J. Chem. Thermo–new experimental measurements published in these journals are critically examined by the TRC and included in the archive. Finally, the files in ThermoML are machine readable via a formal XML schema, allowing facile access to thousands of measurements. In the future, we hope to examine additional measurement classes, including both mixture and two-phase data.

Methods

ThermoML Processing

ThermoML XML files were obtained from the the NIST TRC. To explore their content, we created a python (version 2.7.9) tool (ThermoPyl: https://github.com/choderalab/ThermoPyL) that munges 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 scheme was converted into a Python object via PyXB 1.2.4 (http://pyxb.sourceforge.net/). Finally, the scheme was used to extract the data.

Simulation

Boxes of 1000 molecules were constructed using PackMol.[13] AM1-BCC charges were generated using OpenEye toolkit 2014-6-6,[14] using the oequacpac.OEAssignPartialCharges module with parameter set OECharges_AM1BCCSym. The selected conformer was then processed using antechamber in AmberTools 14. The resulting AMBER files were converted to OpenMM[15] XML files. Simulation code used libraries gaff2xml 0.6, TrustButVerify 0.1, openmm 6.2, and MDTraj[16] 1.2.

Molecular dynamics simulations were performed using OpenMM 6.2 using a Langevin

integrator (friction $1ps^{-1}$) and a 1 fs timestep; interestingly, we found that a 2 fs timestep led to insufficient accuracy in equilibrium densities (Fig. S1). Pressure coupling was achieved with a Monte Carlo barostat applied every 25 steps. Particle mesh Ewald[17] was used with a long-range cutoff of 0.95 nm and an isotropic dispersion correction. Simulations were continued until density standard errors were less than 2×10^{-4} g / mL, as estimated using the equilibration detection module in pymbar 2.1.[18] Trajectory analysis was performed using OpenMM[15] and MDTraj.[16] Density data was output every 250 fs, while trajectory data was stored every 10 ps.

Conclusions

1. ThermoML is a potentially useful resource for the forcefield community 2. We have curated a subset of ThermoML for neat liquids with druglike atoms, with thousands of densities and hundreds of dielectrics 3. Empirical polarization models correct a systematic bias in comparing fixed-charge forcefields to static dielectric constants

Acknowledgements

We thank Vijay Pande, Lee-Ping Wang, Peter Eastman, Robert McGibbon, Jason Swails, David Mobley, Christopher Bayly, Michael Shirts, and members of Chodera lab for helpful discussions.

Supplementary Information

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

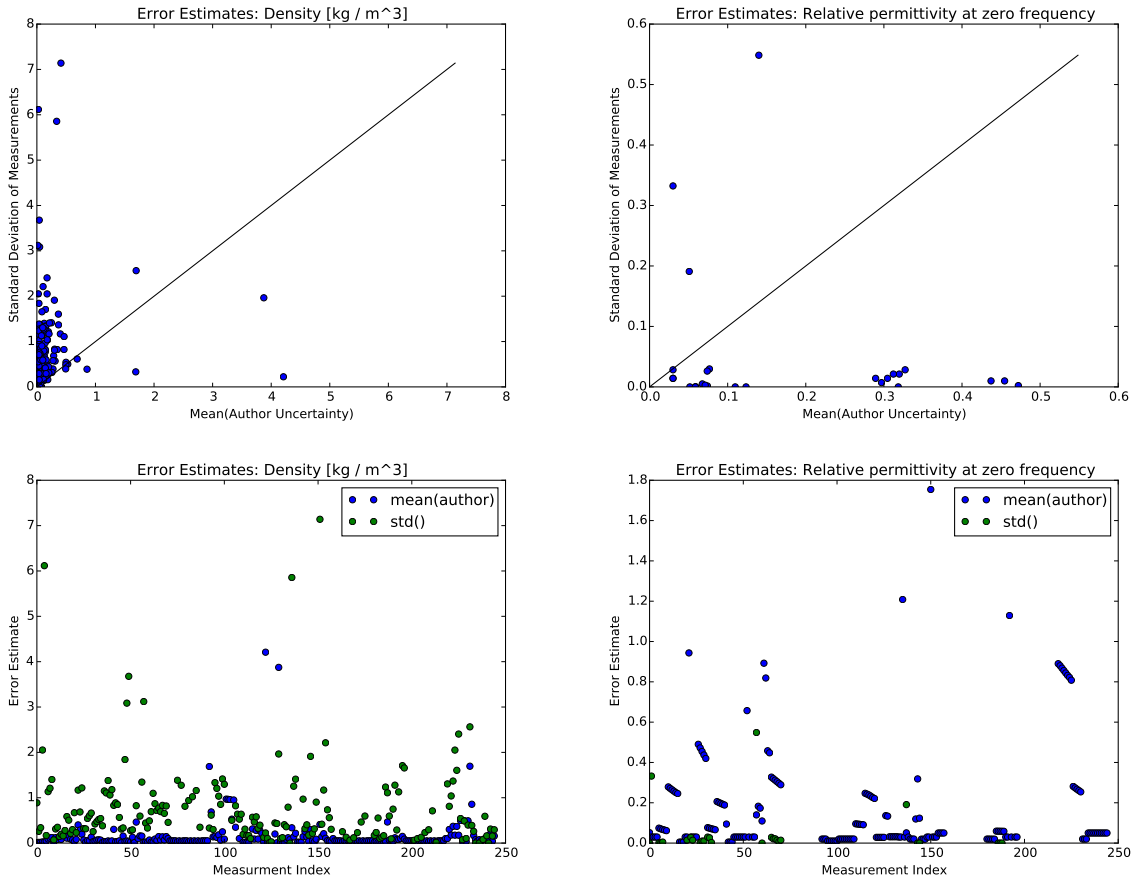


Figure 3: To assess the experimental error in our benchmark set, we consider two orthogonal error measurements. The first is the mean of the uncertainties reported by the measurement authors. The second is the standard deviation of measurements in ThermoML. We see that author-reported uncertainties appear to be overly optimistic for densities (a, c), but author-reported uncertainties of dielectrics (b, d) appear consistent with the standard deviations. 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[19]). However, this hardware limit is not achieved due to inconsistencies in sample preparation; see Appendix in ref.[20] Note that in panels (c, d) show the same information as (a, b) but as a function of the measurement index, rather than as a scatter plot—because not all measurements have author-supplied uncertainties, panels (c, d) contains slightly more data points than (a, b).

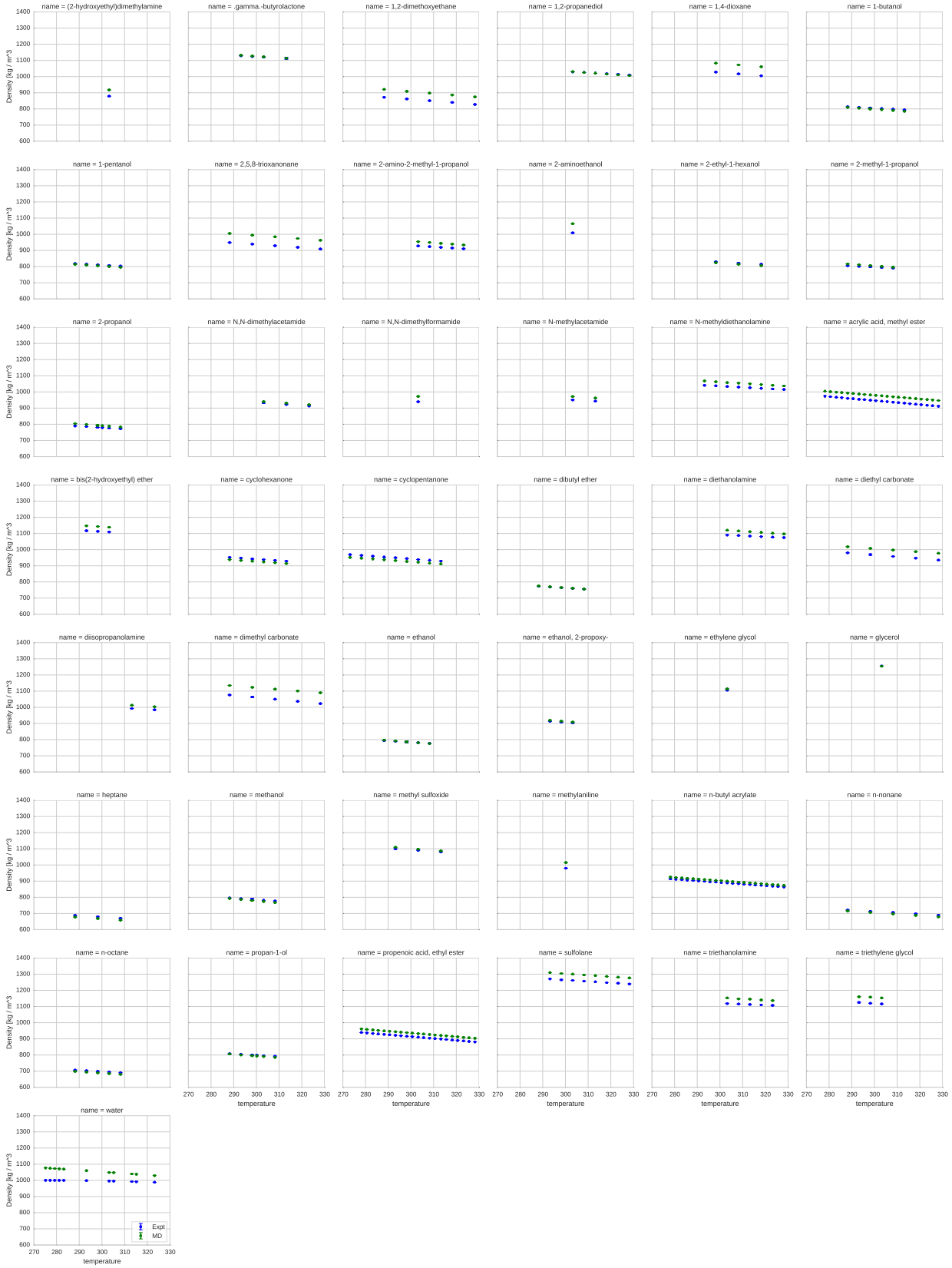


Figure 4: Measured (blue) and simulated (green) densities [kg / m3] for all compounds.

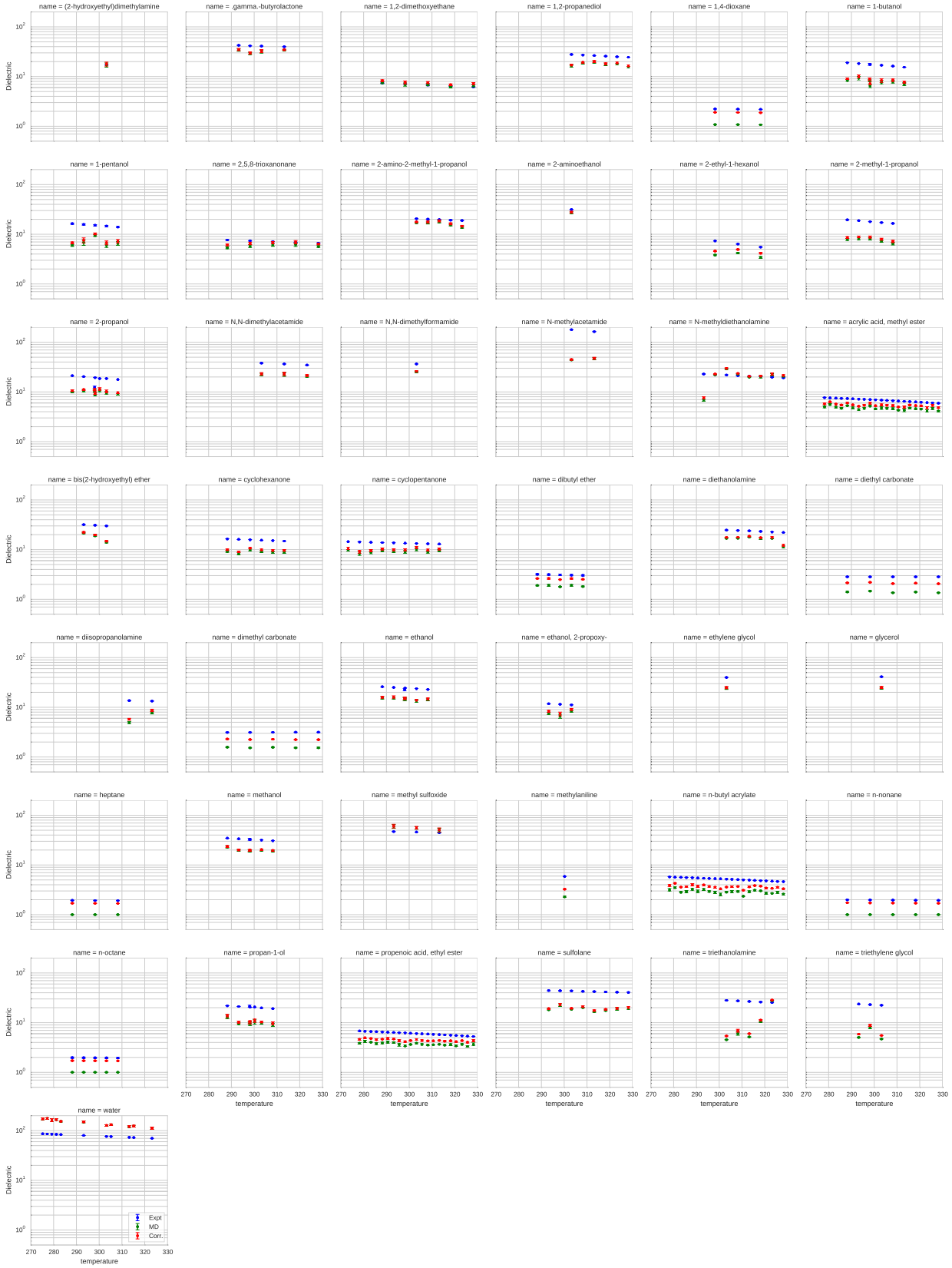
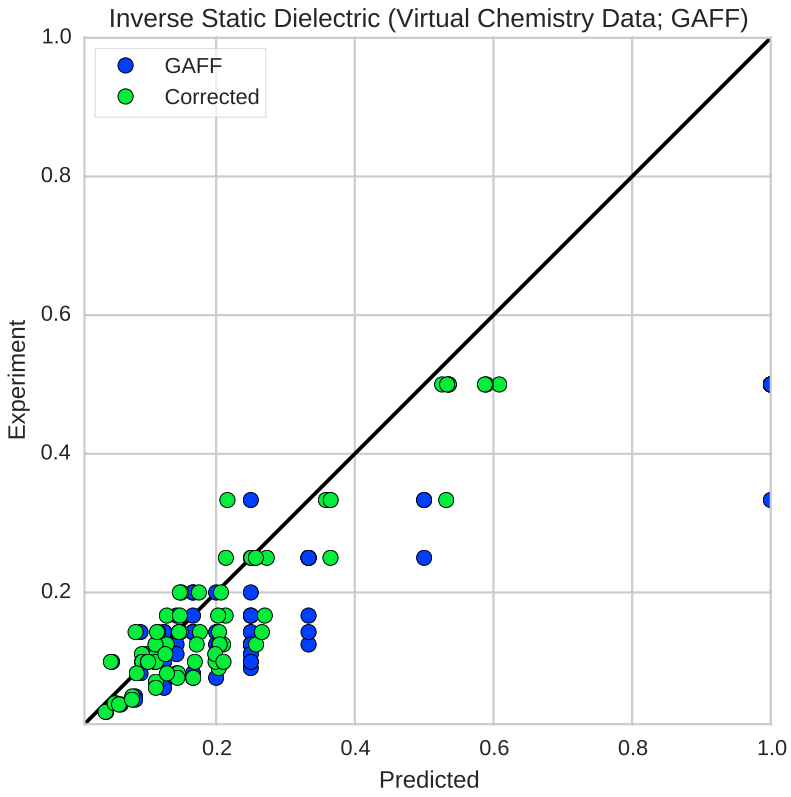


Figure 5: Measured (blue), MD (green), and MD + polarizability-corrected (red) dielectrics for all compounds.



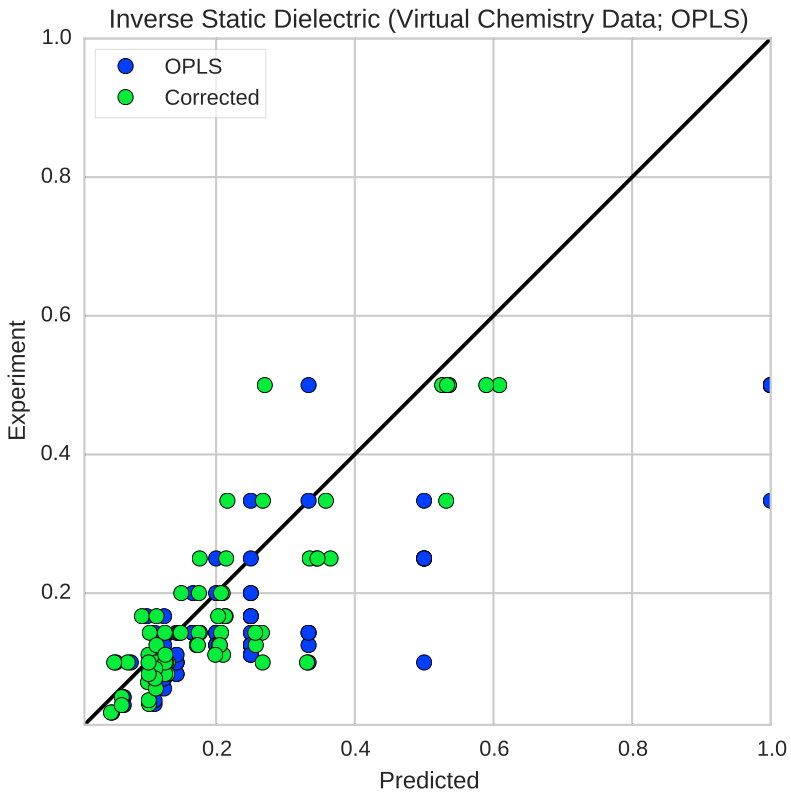


Figure 6: Measured (blue), MD (green), and MD + polarizability-corrected (red) dielectrics for the virtualchemistry dataset.[3,4] 15

Table 2: To probe the systematic error from finite time-step integration, we examined the timestep dependence of butyl acrylate density. The number of effective samples was estimated using pymbar's statistical inefficiency routine.[18] To approximate the timestep bias, we compare the density expectation ($\langle\rho\rangle$) to values calculated with a 0.5fs timestep. We find a 2fs 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 1fs timestep for the present work, where we aimed to achieve three digits of accuracy in density predictions.

	mu	n	neff	sigma	stderr	error	relerr
0.5	0.903701	145510	20357.973571	0.007362	0.000052	0.000000	0.000000
1.0	0.903114	159515	21988.457281	0.007415	0.000050	-0.000588	-0.000650
2.0	0.901811	108346	15964.072327	0.007494	0.000059	-0.001891	-0.002092

References

(1) Bosque, R.; Sales, J. *Journal of chemical information and computer sciences* **2002**, *42*, 1154–1163.

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

(3) Caleman, C.; van Maaren, P. J.; Hong, M.; Hub, J. S.; Costa, L. T.; van der Spoel, D. *Journal of chemical theory and computation* **2011**, *8*, 61–74.

(4) van der Spoel, D.; van Maaren, P. J.; Caleman, C. *Bioinformatics* **2012**, *28*, 752–753.

(5) Flyvbjerg, H.; Petersen, H. G. *J. Chem. Phys.* **1989**, *91*, 461.

(6) Wang, L.-P.; Martínez, T. J.; Pande, V. S. *The Journal of Physical Chemistry Letters* **2014**,

(7) Fennell, C. J.; Wymer, K. L.; Mobley, D. L. *The Journal of Physical Chemistry B* **2014**,

(8) Fennell, C. J.; Li, L.; Dill, K. A. *The Journal of Physical Chemistry B* **2012**, *116*, 6936–6944.

(9) Leontyev, I. V.; Stuchebrukhov, A. A. *The Journal of chemical physics* **2014**, *141*, 014103.

(10) Frenkel, M.; Chiroco, R. D.; Diky, V.; Dong, Q.; Marsh, K. N.; Dymond, J. H.; Wakeham, W. A.; Stein, S. E.; Königsberger, E.; Goodwin, A. R. *Pure and applied chemistry* **2006**, *78*, 541–612.

(11) 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. *Journal of Chemical & Engineering Data* **2003**, *48*, 2–13.

(12) Chirico, R. D.; Frenkel, M.; Diky, V. V.; Marsh, K. N.; Wilhoit, R. C. *Journal of Chemical & Engineering Data* **2003**, *48*, 1344–1359.

(13) Martínez, L.; Andrade, R.; Birgin, E. G.; Martínez, J. M. *Journal of computational chemistry* **2009**, *30*, 2157–2164.

(14) OpenEye Toolkits 2014. http://www.eyesopen.com.

(15) 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.; Pande, V. S. *J. Chem. Theory Comput.* **2012**, *9*, 461–469.

(16) McGibbon, R. T.; Beauchamp, K. A.; Schwantes, C. R.; Wang, L.-P.; Hernández, C. X.; Harrigan, M. P.; Lane, T. J.; Swails, J. M.; Pande, V. S. *bioRxiv* **2014**, 008896.

(17) Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089.

(18) Shirts, M. R.; Chodera, J. D. *J. Chem. Phys.* **2008**, *129*, 124105.

(19) Mettler Toledo Density Meters. http://us.mt.com/us/en/home/products/Laboratory_Analytics_Browse/Density_Family_Browse_main/DE_Benchtop.tabs.models-and-specs.html, [Online; accessed 15-Jan-2015].

(20) others,, et al. *Journal of Chemical & Engineering Data* **2013**, *58*, 2699–2716.