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]*Themodynamics Research Center, NIST, Boulder, CO*

(Dated: February 22, 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 [CITE e.g. http://pubs.acs.org/doi/abs/10.1021/ct400314y]. 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 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 dielectrics and hydration free energies [?]. There are also outstanding questions of generalizability of parameters. Will changes to a specific chemical moiety be compatible with seemingly unrelated 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 notable counterexamples exist, e.g. RCSB [?], FreeSolv [?] and 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 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 constant measurements, with the goal of developing a standard benchmark for validating these properties in fixed-charge forcefields of drug-like molecules. These two observables provide sensitive tests of forcefield accuracy that are nonetheless straightforward to calculate. Using these data, we evaluate the generalized Amber

[*] 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

small molecule forcefield (GAFF) [?] with the AM1-BCC charge model [? ?] and identify systematic biases that might be improved upon.

II. RESULTS

A. Neat Liquid Measurements in the ThermoML Data Archive

We performed a number of sequential 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 druglike molecules. We therefore applied the following ordered sequence 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 (H, N, C, O, S, P, F, Cl, Br)
3. The molecule has fewer than or equal to 10 heavy atoms
4. The measurement was performed under ambient temperature [K] $(270 \leq T \leq 330)$
5. The measurement was performed under ambient pressure [kPA] $(100 \leq P \leq 102)$
6. Measured densities below 300 kg m^{-3} were discarded; this criterion eliminated all non-liquid data in the collection.
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. The application of these filters (Table I) leaves 245 conditions for which both density and dielectric data are available. The functional groups present are summarized in Table II.

Filter	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. **Number of ThermoML measurements matching sequentially applied filters.**

Functional Group	Counts
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 group counts present in the dataset. The number of unique compounds is 44. Functional group classification was performed using checkmol version 0.5 [?].

B. Benchmarking GAFF against the ThermoML Data Archive: Mass Density

Mass density has been widely used for parameterizing and testing forcefields, particularly the Lennard Jones parameters [? ?]. We therefore used the present ThermoML compilation as a benchmark of the GAFF / AM1-BCC forcefield (Fig. 1). Overall, the densities show reasonable accuracy (RMS percent error: 3 % ± 0.1%), consistent with previous studies [?] reporting agreement of 4 % on a different benchmark set.

C. Benchmarking GAFF / AM1-BCC against the ThermoML Data Archive: Static Dielectric

As a measure of the electronic medium, the static dielectric constant of neat liquids provides a critical benchmark of electrostatic models. We therefore compare simulations against the measurements in our ThermoML

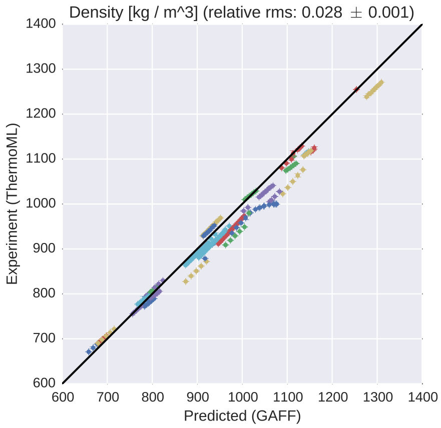


FIG. 1. **Comparison of liquid densities between experiment and simulation.** Liquid density measurements extracted from ThermoML are compared against densities predicted using the GAFF / AM1-BCC small molecule fixed-charge forcefield. Color groupings represent identical chemical species. Simulation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

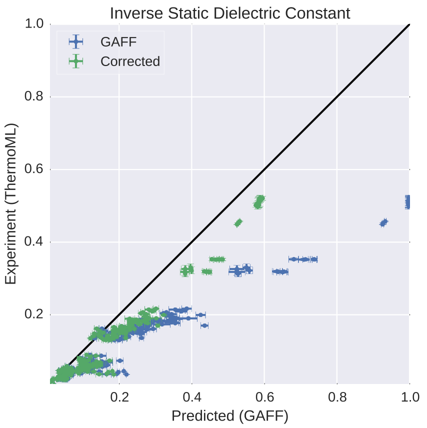


FIG. 2. **Measured (ThermoML) versus predicted (GAFF / AM1-BCC) inverse static dielectrics (a).** Simulation error bars represent one standard error of the mean estimated via block averaging with block sizes of 200 ps [**?**]. Experimental error bars indicate the larger of standard deviation between 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 $\frac{1}{\epsilon}$ is plotted instead of ϵ because $\frac{1}{\epsilon}$ is directly proportional to energy in continuum dielectric models: e.g. $U(r) = \frac{1}{4\pi\epsilon}\frac{q_1 q_2}{r} \propto \frac{1}{\epsilon}$.

compilation. 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 electronic polarization, we added a simple empirical correction for polarization [**?**] that is based on counting the elements in a molecule:

$$\frac{\alpha}{\mathrm{\AA}} = 1.53n_C + 0.17n_H + 0.57n_O + 1.05n_N + 2.99n_S + \\ 2.48n_P + 0.22n_F + 2.16n_{Cl} + 3.29n_{Br} + 5.45n_I + 0.32 \tag{1}$$

From the polarizability, one can correct the static dielectric using the following equation (from ref. [**?**]):

$$\epsilon_{corrected} = \epsilon_{MD} + 4\pi N \frac{\alpha}{\langle V \rangle}$$

A similar polarization correction was used in the development of the TIP4P-EW water model [**?**]; 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 applied the same empirical correction to the VirtualChemistry dataset [**? ?**] and saw similarly improved agreement with experiment for both the GAFF and OPLS forcefields (Fig. 7).

III. DISCUSSION

A. Fitting Forcefields to Dielectric Constants

Recent forcefield development has seen a resurgence of papers fitting dielectric constants as primary data [**? ?**]. However, a number of authors have pointed out potential challenges in constructing self-consistent fixed-charge forcefields [**? ?**].

Interestingly, a recent work by Dill [**?**] pointed out 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: a fixed charge monopole force field predicts $\epsilon \approx 1$ for several 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.

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 $CHCl_3$, it suddenly becomes possible to achieve the observed dielectric constant of 4.8 by an appropriate choice of point charges.

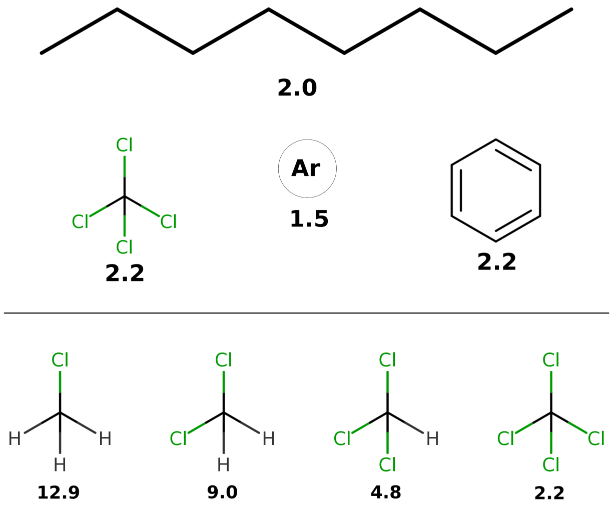


FIG. 3. (a). Measured static dielectric constants of various nonpolar or symmetric molecules [?]; fixed-charge forcefields give $\epsilon \approx 1$ for each species. (b). A congeneric series of chloromethanes show dielectric constants between 2 and 13.

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, CCl_4) have qualitatively different properties than closely related asymmetric molecules (e.g. toluene, $CHCl_3$).

As a possible real-world example, we imagine that the missing polarizability could be important in accurate transfer free energies involving low-dielectric solvents. Using the Onsager model for the transfer free energy of a dipole (Eqn. 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 of druglike molecules in cyclohexane. For each molecule in the FreeSolv database [?], we estimated a as 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 ± 0.07 kcal / mol for the 643 molecules, suggesting that the missing polarizabilty physics contributes substantially to errors in predicted solvation properties of druglike molecules.

Given their ease of measurement and direct connection to long-range electrostatic interactions, static dielectric constants are potentially usable as primary data for forcefield parameterization efforts. Although this will require the use of forcefields with explicit 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 [? ? ?], induced dipole methods [? ?], and drude methods [? ?] 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 [?] 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 Data Archive [? ? ?] 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 aggregation, support, and dissemination of ThermoML is supported by NIST, whose mission makes these tasks a long-term priority. Second, ThermoML is actively growing, through partnerships with several journals—new experimental measurements published in these journals are critically examined by the TRC and included in the archive. Finally, the files in the ThermoML Data Archive 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.

IV. METHODS

A. 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 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 data filters described above.

B. Simulation

Boxes of 1000 molecules were constructed using PackMol [?]. AM1-BCC [? ?] charges were generated using OpenEye Toolkit 2014-6-6 [?], using the oequacpac.OEAssignPartialCharges module with the OECharges_AM1BCCSym. The selected conformer was then processed using antechamber in AmberTools 14 [?]. The resulting AMBER files were converted to OpenMM [?] ffxml forcefield XML files. Simulation code used libraries gaff2xml 0.6, TrustButVerify 0.1, OpenMM 6.2 [?], and MDTraj 1.2 [?]. [TODO: Provide a script to install all of these versions via conda.]

Molecular dynamics simulations were performed using OpenMM 6.2 [?] using a Langevin integrator (with collision rate 1 ps^{-1}) and a 1 fs timestep; interestingly, we found that a 2 fs timestep led to insufficient accuracy in equilibrium densities (Table III). [JDC: Cite Langevin integrator used in OpenMM.] Pressure coupling at 1 atmosphere was achieved with a Monte Carlo barostat utilizing molecular scaling and automated step size adjustment during equilibration, applied every 25 steps. Particle mesh Ewald [?] was used with a long-range cutoff of 0.95 nm and an long-range isotropic dispersion correction. [JDC: Can we report the automatically-selected PME parameters?] 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 [?]. Trajectory analysis was performed using OpenMM [?] and MDTraj [?]. Density data was output every 250 fs, while trajectory data was stored every 10 ps.

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).

VII. DISCLAIMERS

This contribution of the National Institute of Standards and Technology 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.

	mu	n	neff	sigma	stderr	error	
0.5	0.903701	145510	20357.973571	0.007362	0.000052	0.000000	0.0
1.0	0.903114	159515	21988.457281	0.007415	0.000050	-0.000588	-0.0
2.0	0.901811	108346	15964.072327	0.007494	0.000059	-0.001891	-0.0

TABLE III. 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 [**?**]. 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.

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.

- 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

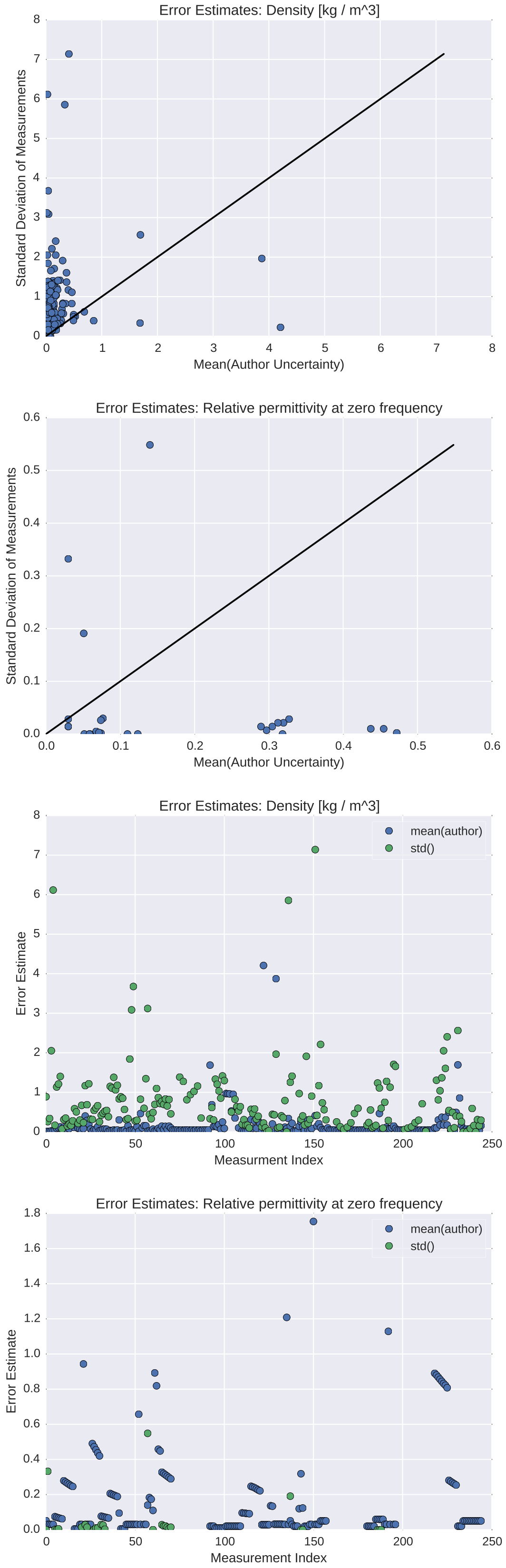


FIG. 4. **Assessment of experimental error in ThermoML data.** To assess the experimental error in our benchmark set, we consider two orthogonal error measurements. The first is the mean of the uncertainties reported by the measurement authors. The second is the standard deviation of measurements in 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

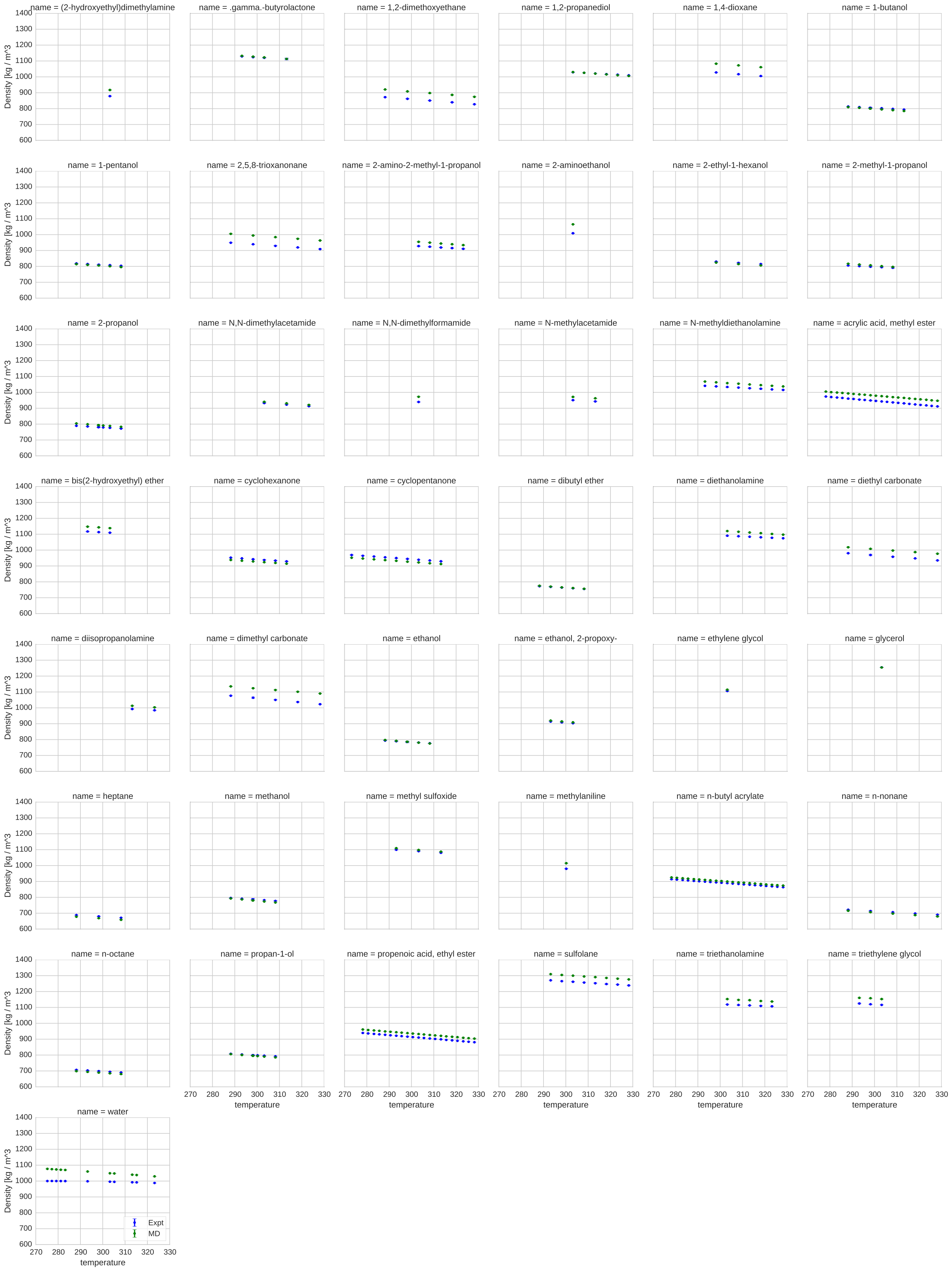


FIG. 5. **Comparison of simulated and experimental densities for all compounds.** Measured (blue) and simulated (green) densities are shown in units of kg/m^3.

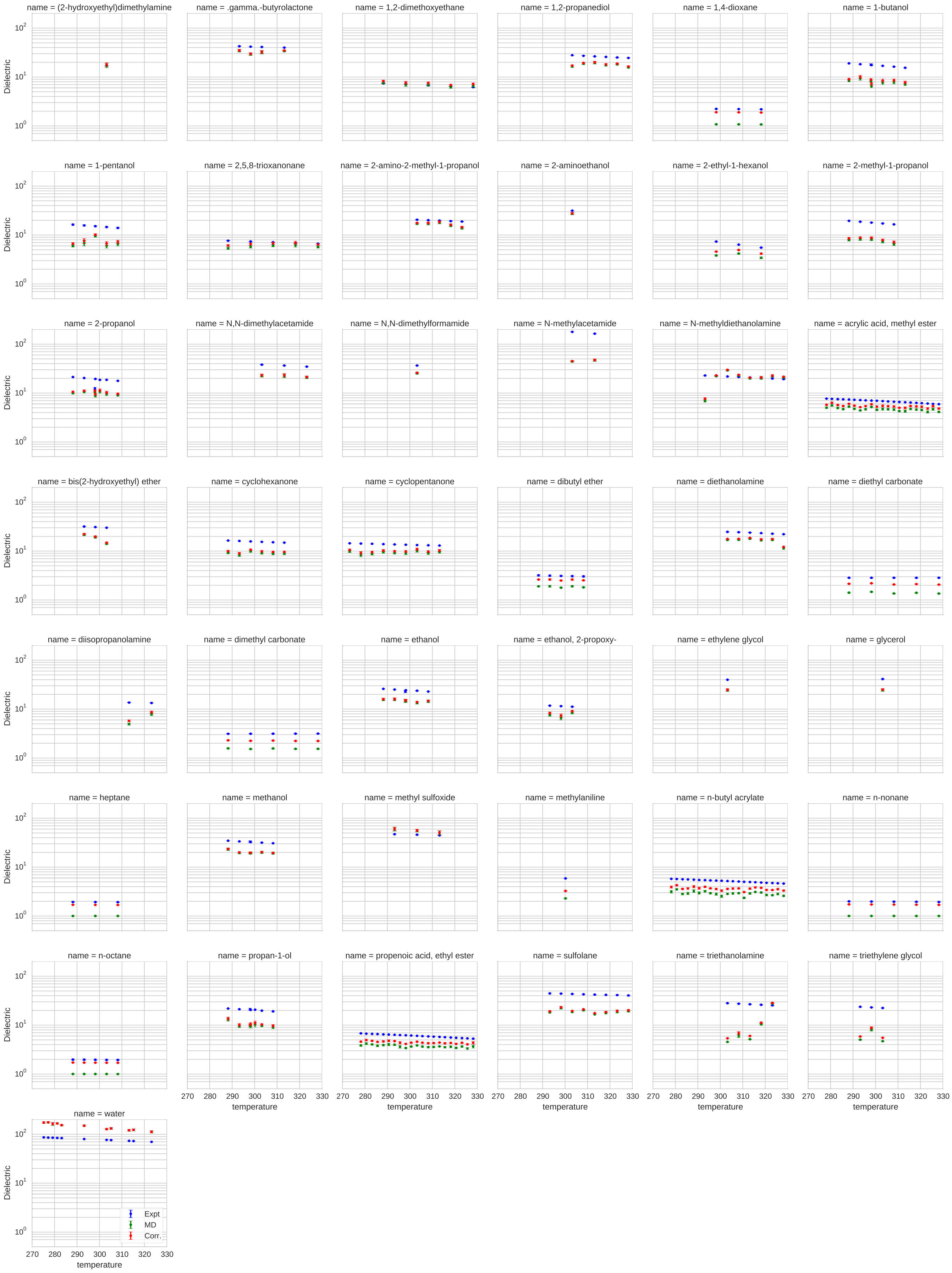
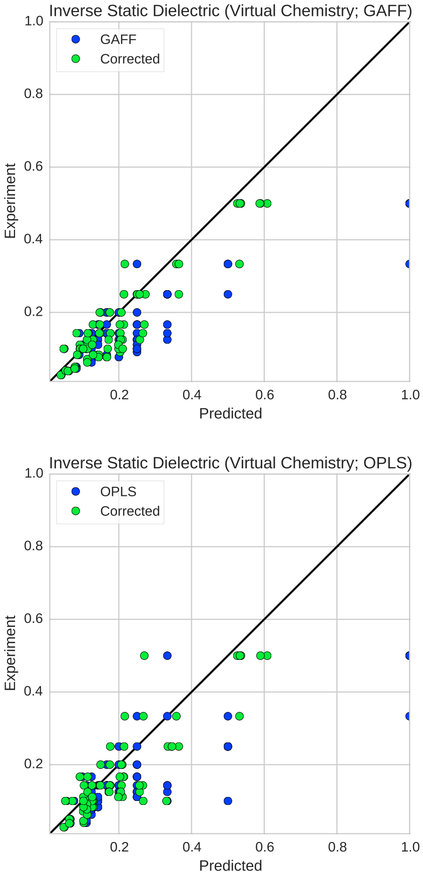


FIG. 6. **Comparison of simulated and experimental static dielectric constants for all compounds.** Measured (blue), simulated (green), and polarizability-corrected simulated (red) static dielectric constants are shown for all compounds. Note that dielectric constants, rather than inverse dielectric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



Inverse Static Dielectric (Virtual Chemistry; GAFF)

Inverse Static Dielectric (Virtual Chemistry; OPLS)

FIG. 7. **Comparison of measured and simulated dielectric constants from the virtualchemistry dataset with and without polarizability correction.** Measured (blue), MD (green), and MD + polarizability-corrected (red) dielectrics for the Virtual Chemistry dataset [**? ?**].