Towards Automated Benchmarking of Atomistic Forcefields: Neat Liquid Densities and Static Dielectric Constants from the ThermoML Data Archive

Kyle A. Beauchamp+,[1, a)] Julie M. Behr+,[2] Ariën S. Rustenburg,[3] Christopher I. Bayly,[4] Kenneth Kroenlein,[5] and John D. Chodera[1, b)]

[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)] *OpenEye Scientific Software Inc., Santa Fe, NM*

[5)] *Thermodynamics Research Center, NIST, Boulder, CO*

(Dated: 23 June 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. 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. Compiling benchmark datasets of physical properties from non-machine-readable sources requires substantial human effort and is prone to the 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 IUPAC-standard 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 experimental measurements (specifically bulk liquid densities and static dielectric constants at ambient pressure) automatically extracted from the archive, and discuss the extent of data available for use in larger scale (or continuously performed) benchmarks. The results of even this limited initial benchmark highlight a general problem with fixed-charge forcefields in the representation low dielectric environments such as those seen in binding cavities or biological membranes.

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[2–4] and enhancements[5–7], with key outcomes including the ability to fold fast-folding proteins[8–10], improved fidelity of water thermodynamic properties[11], and improved prediction of NMR observables. Although small molecule forcefields have also been the subject of benchmarks[12–14] and improvements[15], 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[15]. 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 community agreement upon 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[16], FreeSolv[17], and the BMRB[18]). 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 large relevant chemical spaces, such as forcefields intended to describe a large variety of druglike small organic molecules.

In addition to inconvenience, the number and kind of

[a)] Corresponding author; Electronic mail: kyle.beauchamp@choderalab.org

[b)] Corresponding author; Electronic mail: john.chodera@choderalab.org

human-induced errors that can corrupt hand-compiled benchmarks are legion. A United States Geological Survey (USGS) case study examining the reporting and use of literature values of the aqueous solubility (S_w) and octanol-water partition coefficients (K_{ow}) for DDT and its persistent metabolite DDE provides incredible insight into a variety of common errors[19]. Secondary sources are often cited as primary sources—a phenomenon that occurred up to five levels deep in the case of DDT/DDE; citations for data are often incorrect, misattributed to unrelated publications, or omitted altogether; numerical data can be mistranscribed, transposed, or incorrectly converted among unit systems[19]. This occurs to such a degree that the authors note "strings of erroneous data compose as much as 41–73 percent of the total data"[19]. Given the incredible importance of these properties for human health and the environment, the quality of physicochemical datasets of far lesser importance is highly suspect.

To ameliorate problems of data archival, the NIST Thermodynamics Research Center (TRC) has developed an IUPAC standard XML-based format—ThermoML[20–22]—for storing physicochemical measurements, uncertainties, and metadata. Manuscripts containing new experimental measurements submitted to 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 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)[23,24] with the AM1-BCC charge model[25,26] 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 8 Apr. 2015. 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[27] dataframes, and the successive data filters described in Section III A were applied.

B. Simulation

To enable automated accuracy benchmarking of physicochemical properties of neat liquids such as mass density and dielectric constant, we developed a semi-automated pipeline for preparing simulations, running them on a standard computer cluster using a portable simulation package, and analyzing the resulting data. All code for this procedure is available at https://github.com/choderalab/LiquidBenchmark. Below, we describe the operation of the various stages of this pipeline and their application to the benchmark reported here.

1. Preparation

Chemical names were parsed from the ThermoML extract and converted to both CAS and smiles strings using cirpy[28]. Smiles strings were converted into molecular structures using the OpenEye Python Toolkit version 2015-2-3[29], as wrapped in openmoltools.

Simulation boxes containing 1000 molecules were constructed using PackMol version 14-225[30,31] wrapped in the Python automation library openmoltools. In order to ensure stable automated equilibration, PackMol box volumes were chosen to accommodate twice volume of the enclosed atoms, with atomic radii estimated as 1.06 Å and 1.53 Å for hydrogens and nonhydrogens, respectively.

For this illustrative benchmark, we utilized the generalized Amber small molecule forcefield (GAFF)[23,24] with the AM1-BCC charge model[25,26], which we shall refer to as the GAFF/AM1-BCC forcefield.

Canonical AM1-BCC[25,26,32] charges were generated with the OpenEye Python Toolkit version 2015-2-3[29], using the `oequacpac.OEAssignPartialCharges` module with the `OECharges_AM1BCCSym` option, which utilizes a conformational expansion procedure (using `oeomega.OEOmega`[33]) prior to charge fitting to minimize artifacts from intramolecular contacts. The `OEOmega` selected conformer was then processed using `antechamber` (with `parmchk2`) and `tleap` in AmberTools 14[34] to produce Amber-format `prmtop` and `inpcrd` files, which were then read into OpenMM to perform molecular simulations using the `simtk.openmm.app` module.

The simulations reported here used libraries openmoltools 0.6.4[35], OpenMM 6.3[36], and MDTraj 1.3[37]. Exact commands to install various dependencies can be found in Appendix A 1.

2. *Equilibration and production*

Simulation boxes were first minimized using the L-BFGS algorithm[38] 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 6.3[36] using a Langevin Leapfrog integrator[39] (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 (Fig. 4).

Equilibration and production simulations utilized a Monte Carlo barostat with a control pressure of 1 atm (101.325 kPa), utilizing molecular scaling and automated step size adjustment during equilibration, with volume moves attempted every 25 steps. The particle mesh Ewald (PME) method with conducting boundary conditions[40] was used with a long-range cutoff of 0.95 nm and a long-range isotropic dispersion correction. PME grid and spline parameters were automatically selected using the default settings in OpenMM 6.3 for the CUDA platform[36].

Automatic termination criteria. Production simulations were continued until automatic analysis showed standard errors in densities were less than 2×10^{-4} g / cm^3. Automatic analysis of the production simulation data was run every 1 ns of simulation time, and utilized the `detectEquilibration` method in the timeseries module of pymbar 2.1[41] to automatically discard the initial portion of the production simulation containing 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 using the fast adaptive statistical inefficiency computation method as implemented in the `timeseries.computeStatisticalInefficiency` method of pymbar 2.1 (where the algorithm is described in[42]). This approach is essentially the same as the fixed-width procedure described by eq. 7.12 of ref.[43], with n^* equal to 4000 and the sequential testing correction (n^{-1} term) ignored due to the large value of n. Statistical errors were computed by $\delta^2\rho \approx var(\rho)/N_{\mathrm{eff}}$, where $var(\rho)$ is the sample variance of the density and N_{eff} is the number of effectively uncorrelated samples. With this protocol, we found starting trajectory lengths of 12000 $(8000, 16000)$ frames (250 fs each), discarded regions of 28 $(0, 460)$, and statistical inefficiencies of 20 $(15, 28)$; reported numbers indicate (median, (25%, 75%)).

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

C. Timings

The wall time required for a given simulation depends on the number of atoms (3,000 - 29,000), the GPU used (GTX 680 or GTX Titan), and the time required for automated termination. For butyl acrylate (21,000 atoms) on a GTX Titan, the wall-clock performance is approximately 80 ns / day. Using 80 ns / day with approximately 3 ns of production simulation corresponds to 1 hour for the production segment of the simulation and 3 hours for the fixed equilibration portion of 10^7 steps.

1. *Data analysis and statistical error estimation*

Trajectory analysis was performed using OpenMM 6.3[36] and MDTraj 1.3[37].

Mass density. Mass density ρ 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. Static dielectric constants were calculated using the dipole fluctuation approach appropriate for PME with conducting ("tin-foil") boundary conditions[11,44], with the total system box dipole μ computed from trajectory snapshots using MDTraj 1.3[37].

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

where $\beta \equiv 1/k_B T$ is the inverse temperature.

Computation of expectations. Expectations were estimated by computing sample means over the production simulation after discarding the initial far-from-equilibrium portion to equilibration (as described in **Automatic termination criteria** above).

Statistical uncertainties. For density uncertainties, the Markov chain standard error (MCSE) was estimated as $\frac{\sigma}{\sqrt{N_{eff}}}$, where σ is the density standard deviation of the simulation not discarded to equilibration, $N_{eff} = \frac{N}{g}$ is the effective sample size, and g is the statistical inefficiency as estimated from the density time series. For dielectric uncertainties, the portion of the production simulation not discarded to equilibration was used as input to a circular block bootstrapping procedure[45] with block sizes automatically selected to maximize the error[46].

2. *Code availability*

Data analysis, all intermediate data (except configurational trajectories, due to their large size), and figure creation code for this work is available at https://github.com/choderalab/LiquidBenchmark.

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 sample 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 non-hydrogen atoms

4. The measurement was performed in a biophysically relevant temperature range $(270 \leq T$ [K] $\leq 330)$

5. The measurement was performed at ambient pressure $(100 \leq P$ [kPa] $\leq 102)$

6. Only measurements in liquid phase were retained

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 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 246 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[48,49]. We therefore used the present ThermoML extract as a benchmark of the GAFF/AM1-BCC forcefield (Fig. 1).

	Number of measurements remaining	
Filter step	**Mass density**	**Static dielectric**
1. Single Component	136212	1651
2. Druglike Elements	125953	1651
3. Heavy Atoms	71595	1569
4. Temperature	38821	964
5. Pressure	14103	461
6. Liquid state	14033	461
7. Aggregate T, P	3592	432
8. Density+Dielectric	246	246

TABLE I: **Successive filtration of the ThermoML Archive.** A set of successive filters were applied to all measurements in the ThermoML Archive 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 246 final measurements correspond to 45 unique molecules measured at several temperature conditions.

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	3
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 246 distinct (molecule, temperature, pressure) conditions, spanning 45 unique compounds. The functional groups represented in these compounds (as identified by the program `checkmol` v0.5[47]) is summarized here.

Overall accuracy. Overall, the densities show reasonable accuracy, with a root-mean square (RMS) relative error over all measurements of (3.0±0.1)%, especially encouraging given that this forcefield was not designed with the intention of modeling bulk liquid properties of organic molecules[23,24]. This is reasonably consistent with previous studies reporting agreement of 4% on a different benchmark set[12].

Temperature dependence. For a given compound, the signs of the errors typically do not change at different temperatures (Fig. 1, Fig. 7). Furthermore, the magnitudes of the error also remain largely constant (vertical lines in Fig. 1 B), although several exceptions do occur. It is possible that these systematic density offsets indicate correctable biases in forcefield parameters.

Outliers. The largest density errors occur for a number of oxygen-containing compounds: 1,4-dioxane; 2,5,8-trioxanonane; 2-aminoethanol; dimethyl carbonate; formamide; and water (Fig. 7). The absolute error on these poor predictions is on the order of 0.05 g/cm^3, which is substantially higher than the measurement error (≤ 0.008 g/cm^3; see Fig. 5).

We note that our benchmark includes a GAFF/AM1-BCC model for water due to our desire to automate benchmarks against a forcefield capable of modeling a large variety of small molecular liquids. Water—an incredibly important solvent in biomolecular systems—is generally treated with a special-purpose model (such as TIP3P[48] or TIP4P-Ew[11]) parameterized to fit a large quantity of thermophysical data. As expected, the GAFF/AM1-BCC model performs poorly in reproducing liquid densities for this very special solvent. We conclude that it remains highly advisable that the field continue to use specialized water models when possible.

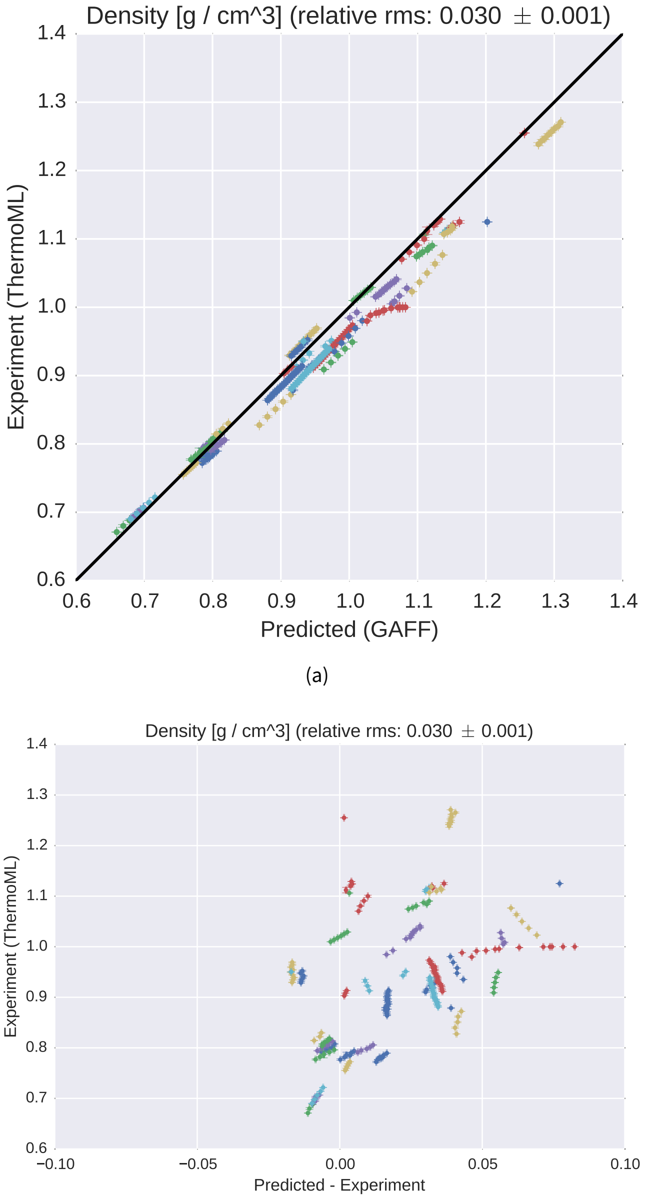
2. *Static dielectric constant*

Overall accuracy. 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. Discussing the accuracy in terms the ability of GAFF/AM1-BCC to reproduce the static dielectric constant ϵ is not necessarily meaningful because of the way that the solvent dielectric ϵ enters into the Coulomb potential between two point charges separated by a distance r,

$$U(r) = \frac{q_1 q_2}{\epsilon\, r} \propto \frac{1}{\epsilon}. \tag{3}$$

It is evident that $1/\epsilon$ is a much more meaningful quantity to compare than ϵ directly, as a 5% error in $1/\epsilon$ will cause a 5% error in the Coulomb potential between two point charges (assuming a uniform dielectric), while a 5% error in ϵ will have a much more complex ϵ-dependent effect on the Coulomb potential. We therefore compare simulations against measurements in our ThermoML extract on the $1/\epsilon$ scale in Fig. 2.

GAFF/AM1-BCC systematically underestimates the dielectric constants of nonpolar liquids. Overall, we find the dielectric constants to be qualitatively reasonable, but with clear deviations from experiment particularly for nonpolar liquids. This is not surprising given the complete neglect of electronic polarization which will be the dominant contribution for such liquids. In particular, GAFF/AM1-BCC systematically underestimates the dielectric constants for nonpolar liquids, with the predictions of $\epsilon \approx 1.0$ being substantially



(a)

(b)

FIG. 1: **Comparison of liquid densities between experiment and simulation.** (a). 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, although the color map repeats itself due to the large (45) number of unique compounds. Plots of density versus temperature grouped by chemical species are available in Fig. 7. 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 Fig. 5 for further discussion of error. (b). The same plot, but with the residual (predicted minus experiment) on the x axis. Note that the error bars are all smaller than the symbols.

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 α as a sum of elemental atomic polarizability contributions[50].

From the computed molecular electronic polarizability α, an additive correction to the simulation-derived static dielectric constant accounting for the missing electronic polarizability can be computed[11]

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

A similar polarization correction was used in the development of the TIP4P-Ew water model, where it had a minor effect[11] because almost all the high static dielectric constant for water comes from the configurational response of its strong dipole. However, the 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. Averaging all liquids in the present work leads to polarizability corrections to the static dielectric of 0.74 ± 0.08. Taking the dataset as a whole, we find that the relative error in uncorrected dielectric is on the order of -0.34 ± 0.02, as compared to -0.25 ± 0.02 for the corrected dielectric.

IV. DISCUSSION

A. Mass densities

Our simulations have indicated the presence of systematic density biases with magnitudes larger than the measurement error. Correcting these errors may be a low-hanging fruit for future forcefield refinements. As an example of the feasibility of improved accuracy in densities, a recent three-point water model was able to recapitulate water density with errors of less than 0.005 g / cm^3 over temperature range [280 K, 320 K][51]. This improved accuracy in density prediction was obtained alongside accurate predictions of other experimental observables, including static dielectric constant. We suspect that such accuracy might be obtainable for GAFF-like forcefields across some portion of chemical space. A key challenge for the field is to demarcate the fundamental limit of fixed-charge forcefields for predicting orthogonal classes of experimental observables. For example, is it possible to achieve a relative density error of 10^{-4} without sacrificing accuracy of other properties such as enthalpies? In our opinion, the best way to answer such questions is to systematically build forcefields with the goal of predicting various properties to within their known experimental uncertainties, similar to what has been done for water[11,51].

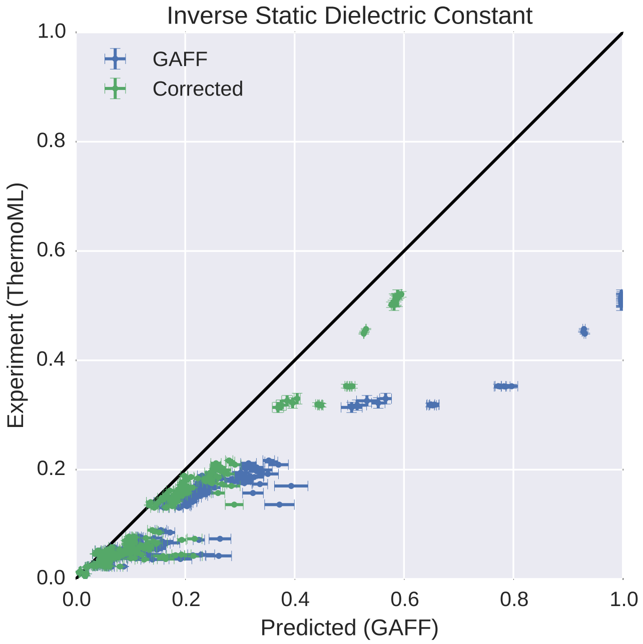


FIG. 2: **Measured (ThermoML) versus predicted (GAFF / AM1-BCC) inverse static dielectrics (a).** Simulation error bars represent one standard error of the mean. Experimental error bars indicate the larger of standard deviation between independently reported measurements and the authors reported standard deviations; for some measurements, neither uncertainty estimate is available. See Fig. 5 for further discussion of error. See Section III B 2 for explanation of why inverse dielectric constant (rather than dielectric constant) is plotted. For nonpolar liquids, it is clear that the forcefield predicts electrostatic interactions that are substantially biased by missing polarizability. Plots of dielectric constant versus temperature grouped by chemical species are available in Fig. 8.

B. Dielectric constants in forcefield parameterization

A key feature of the static dielectric constant for a liquid is that, for forcefield purposes, it consists of two very different components, distinguished by the dependence on the fixed charges of the forcefield and dynamic motion of the molecule. One component, the high-frequency dielectric constant, arises from the almost-instantaneous electronic polarization in response to the external electric field: this contributes a small component, generally around $\epsilon = 2$, which can be dominant for non-polar liquids but is completely neglected by the non-polarizable forcefields in common use for biomolecular simulations. The other component arises from the dynamical response of the molecule, through nuclear motion, to allow its various molecular multipoles to respond to the external electric field: for polar liquids such as water, this contributes the majority of the dielectric constant. Thus for polar liquids, we expect the parameterized atomic charges to play a major role in the static

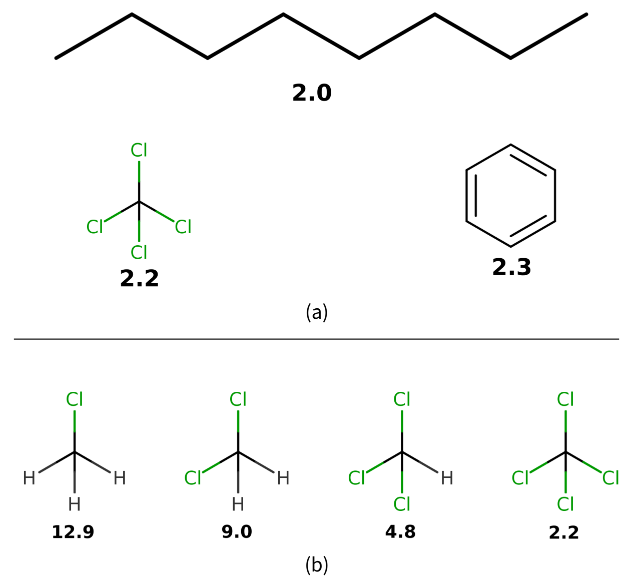


FIG. 3: **Typical experimental static dielectric constants of some nonpolar compounds.** (a). Measured static dielectric constants of various nonpolar or symmetric molecules[54,55]. Fixed-charge forcefields give $\epsilon \approx 1$ for each species; for example, we calculated $\epsilon = 1.0030 \pm 0.0002$ for octane. (b). A congeneric series of chloro-substituted methanes have static dielectric constants between 2 and 13. Reported dielectric constants are at near-ambient temperatures.

dielectric.

Recent forcefield development has seen a resurgence of papers fitting dielectric constants during forcefield parameterization[15,51]. However, a number of authors have pointed out potential challenges in constructing self-consistent fixed-charge forcefields[52,53].

Interestingly, recent work by Dill and coworkers[52] 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 profound consequences, which we discuss below.

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? We expect it to be important wherever we encounter the transfer of a polar molecule (such as a peptide, native ligand, or a pharmaceutical small molecule) from a polar environment (such as the cytosol, interstitial fluid, or blood) into a non-polar environment (such as a biological membrane or non-polar binding site of an enzyme or receptor). Thus we expect this to be implicated in biological processes ranging from ligand binding to absorption and distribution within the body. To understand this conceptually, consider the transfer of a polar small-molecule transfer from the non-polar interior of a lipid bilayer to the aqueous and hence very polar cytosol. 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 μ of radius a gives us a way to estimate the magnitude of error introduced by making an error of $\Delta\epsilon$ in the static dielectric constant of a solvent. The free energy of dipole solvation is given by this model 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 ϵ, 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 $\Delta\Delta G \sim -8$ kJ/mol (-2 kcal/mol).

Implications for transfer free energies. As another example, 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[17,56], 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 (-3.8 ± 0.3) kJ / mol $[(-0.91 \pm 0.07)$ 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. In other words, the use of a fixed-charge physics may lead to errors of 3.8 kJ / mol in cyclohexane transfer free energies. We conjecture that this missing physics will be important in the upcoming (2015) SAMPL challenge[57], which will examine transfer free energies in several low dielectric media.

Utility in parameterization. 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[58–60], point dipole methods[61,62], and Drude methods[63,64] 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[65] 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[20,21,66] 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. 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.

V. CONCLUSIONS

High quality, machine-readable datasets of physicochemical measurements will be required for the construction of next-generation small molecule forcefields. Here we have discussed the NIST/TRC ThermoML archive as a growing source of physicochemical measurements that may be useful for the forcefield community. From the NIST/TRC ThermoML archive, we selected a dataset of 246 ambient, neat liquid systems for which both densities and static dielectric constants are available. Using this dataset, we benchmarked GAFF/AM1-BCC, one of the most popular small molecule forcefields. We find systematic biases in densities and particularly static dielectric constants. Element-based empirical polarizabilty models are able to account for much of the systematic differences between GAFF/AM1-BCC and experiment. Non-polarizable forcefields may show unacceptable biases when predicting transfer and binding properties of non-polar environments such as binding cavities or membranes.

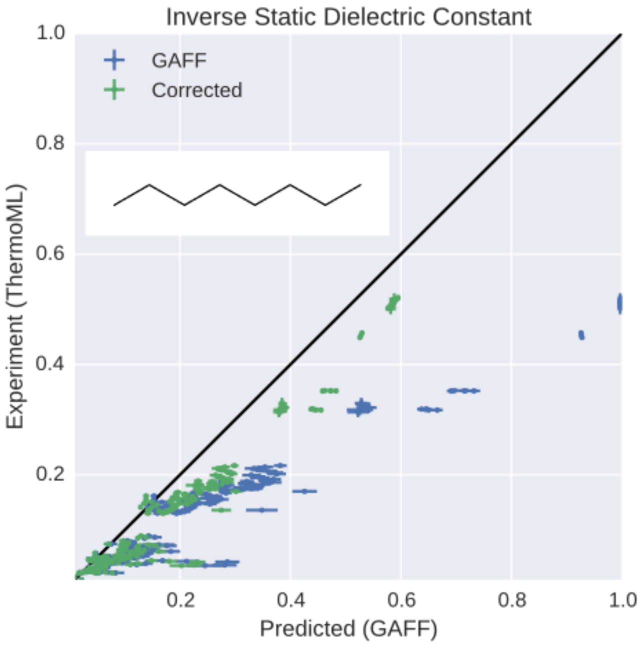
VI. ACKNOWLEDGEMENTS

We thank Patrick B. Grinaway (MSKCC), 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), Michael R. Shirts (University of Virginia), William C. Swope (IBM), Julia E. Rice (IBM), Hans Horn (IBM), Jed W. Pitera (IBM), 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). KAB was supported in part by Starr Foundation grant I8-A8-058. JDC and KAB acknowledge partial support from NIH grant P30 CA008748. KAB, JLB, ASR, and JDC acknowledge the generous support of this research by the Sloan Kettering Institute. The authors gratefully acknowledge OpenEye Scientific for generously providing their toolkit for use in noncommercial projects that generate results for the public domain.

VII. DISCLAIMERS

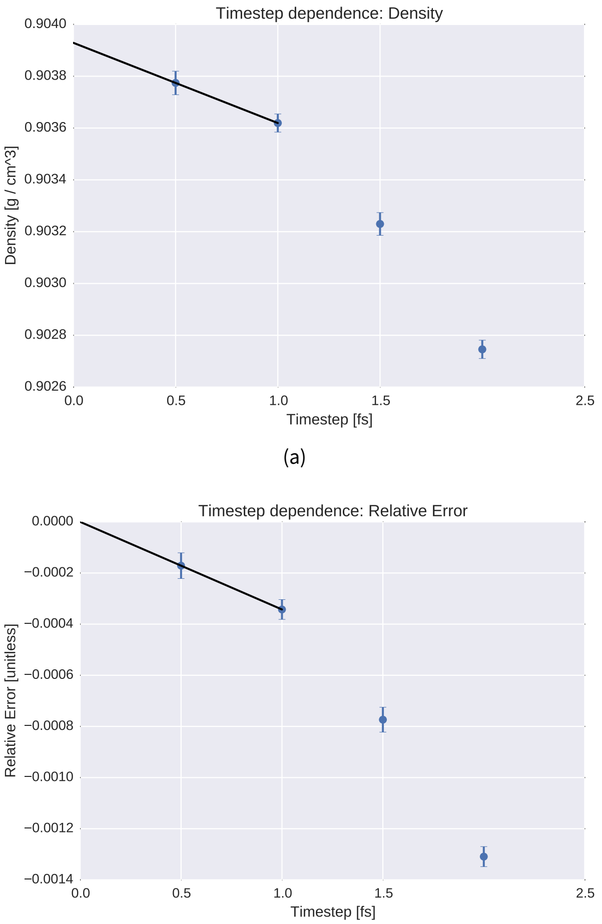
This contribution of the National Institute of Standards and Technology (NIST) is not subject to copyright in the United States. Products or companies named here are cited only in the interest of complete technical description, and neither constitute nor imply endorsement by NIST or by the U.S. government.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

VIII. TOC FIGURE



Appendix A: Appendices

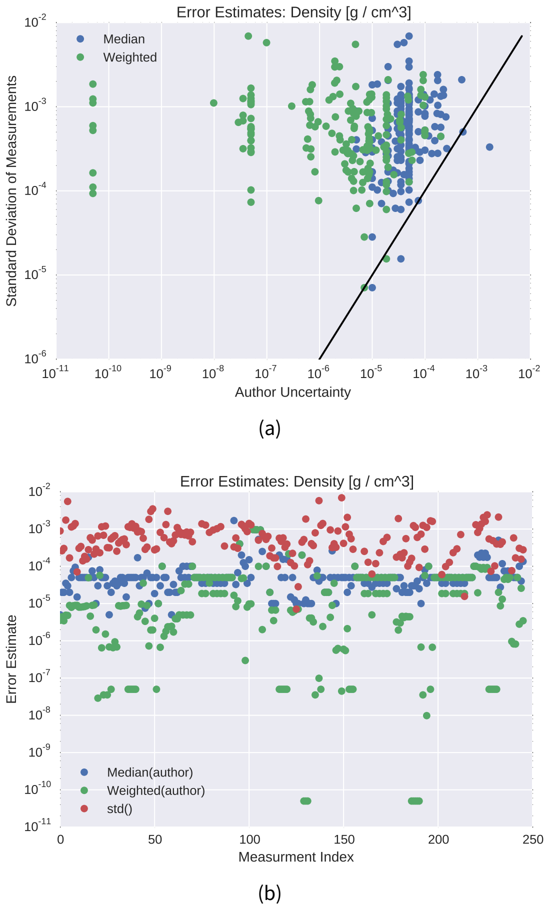
- Figure: Timestep-dependence of density
- Figure: Error analysis (density) for ThermoML dataset
- Figure: Error analysis (static dielectric constant) for ThermoML dataset
- Figure: Temperature Dependence: Density
- Figure: Temperature Dependence: Static Dielectric Constant
- Commands to install dependencies



(a)

(b)

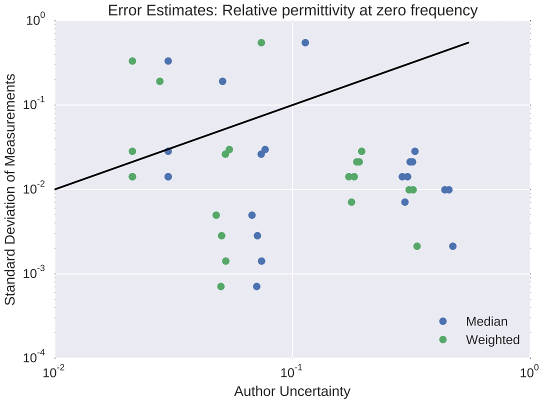
FIG. 4: **Dependence of computed density on simulation timestep.** To probe the systematic error from finite time-step integration, we examined the timestep dependence of butyl acrylate density. (a). The density is shown for several choices of timestep. (b). The relative error, as compared to the reference value, is shown for several choices of timestep. Error bars represent standard errors of the mean, with the number of effective samples estimated using pymbar's statistical inefficiency routine[41]. 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 0.13%, while 1 fs reduces the systematic bias to approximately 0.8%—we therefore selected a 1 fs timestep for the present work, where we aimed to achieve three digits of accuracy in density predictions.



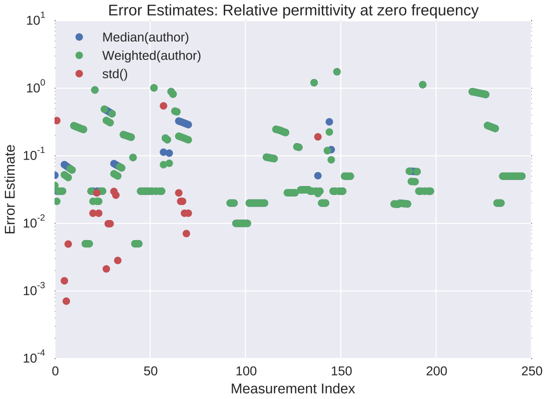
(a)

(b)

FIG. 5: **Assessment of experimental error: Density** To assess the experimental error in our ThermoML extract, we compared three different 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 uncertai8nty 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 repeatability stated by the manufacturer (e.g. 0.0001 g / cm^3 for a Mettler Toledo DM40[67]). However, this hardware limit is not achieved due to inconsistencies in sample preparation and experimental conditions; see Appendix in Ref.[22]. 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).



(a)



(b)

FIG. 6: **Assessment of experimental error: Static Dielectric Constant** To assess the experimental error in our ThermoML extract, we compared three different 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).

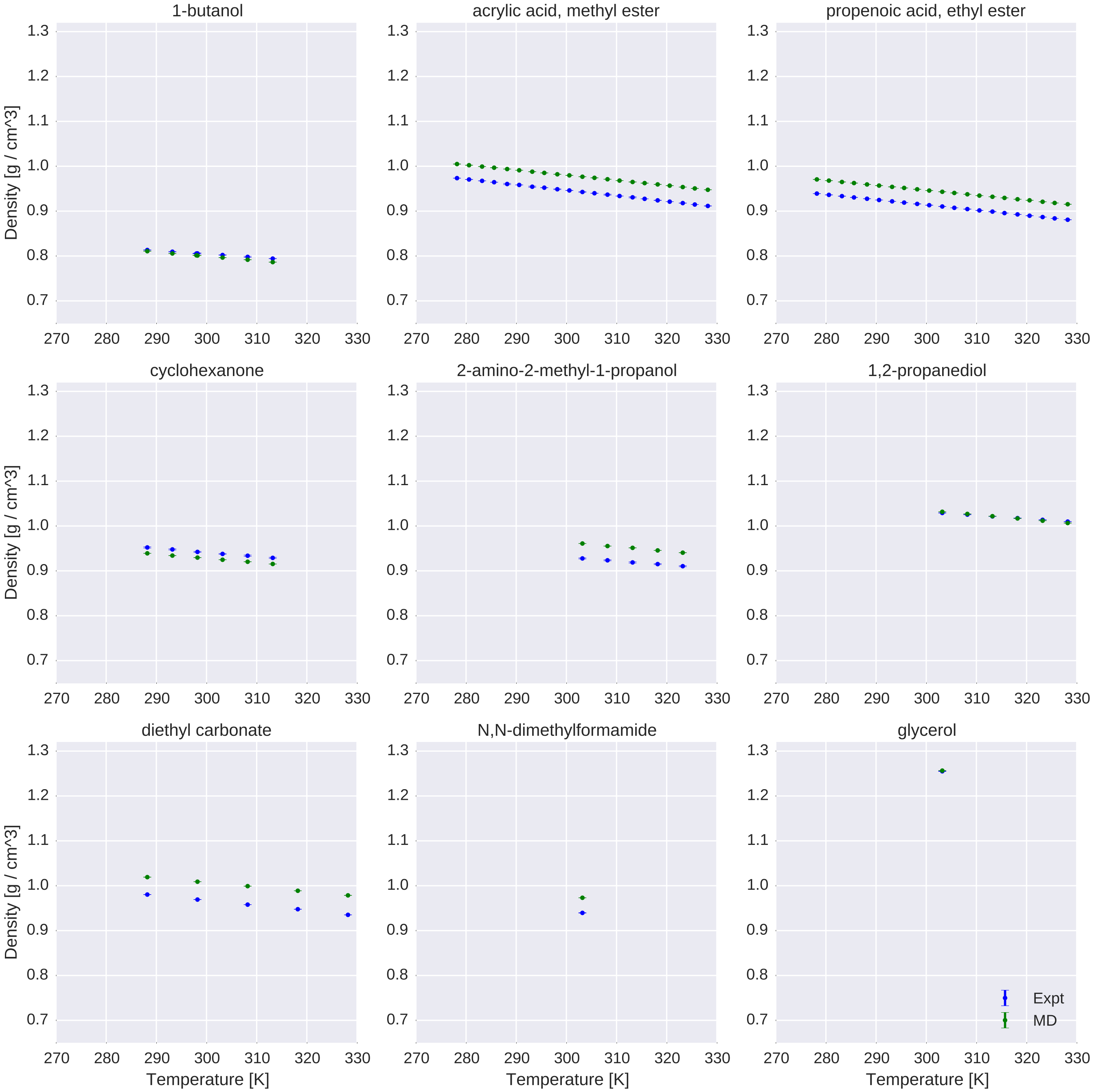


FIG. 7: **Comparison of simulated and experimental densities for all compounds.** Measured (blue) and simulated (green) densities are shown in units of g / cm^3.

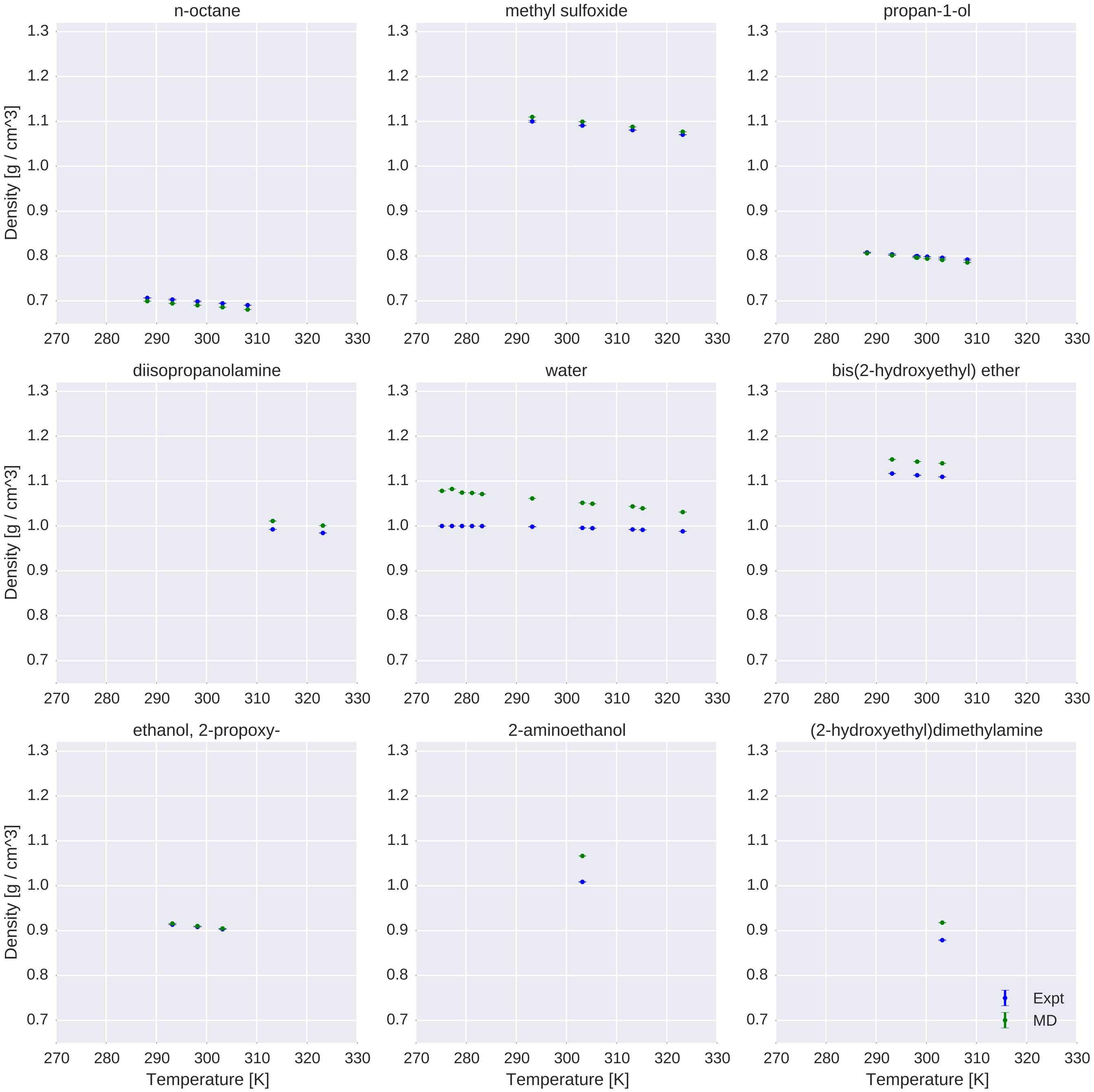


FIG. 7: **Comparison of simulated and experimental densities for all compounds.** Measured (blue) and simulated (green) densities are shown in units of g / cm^3.

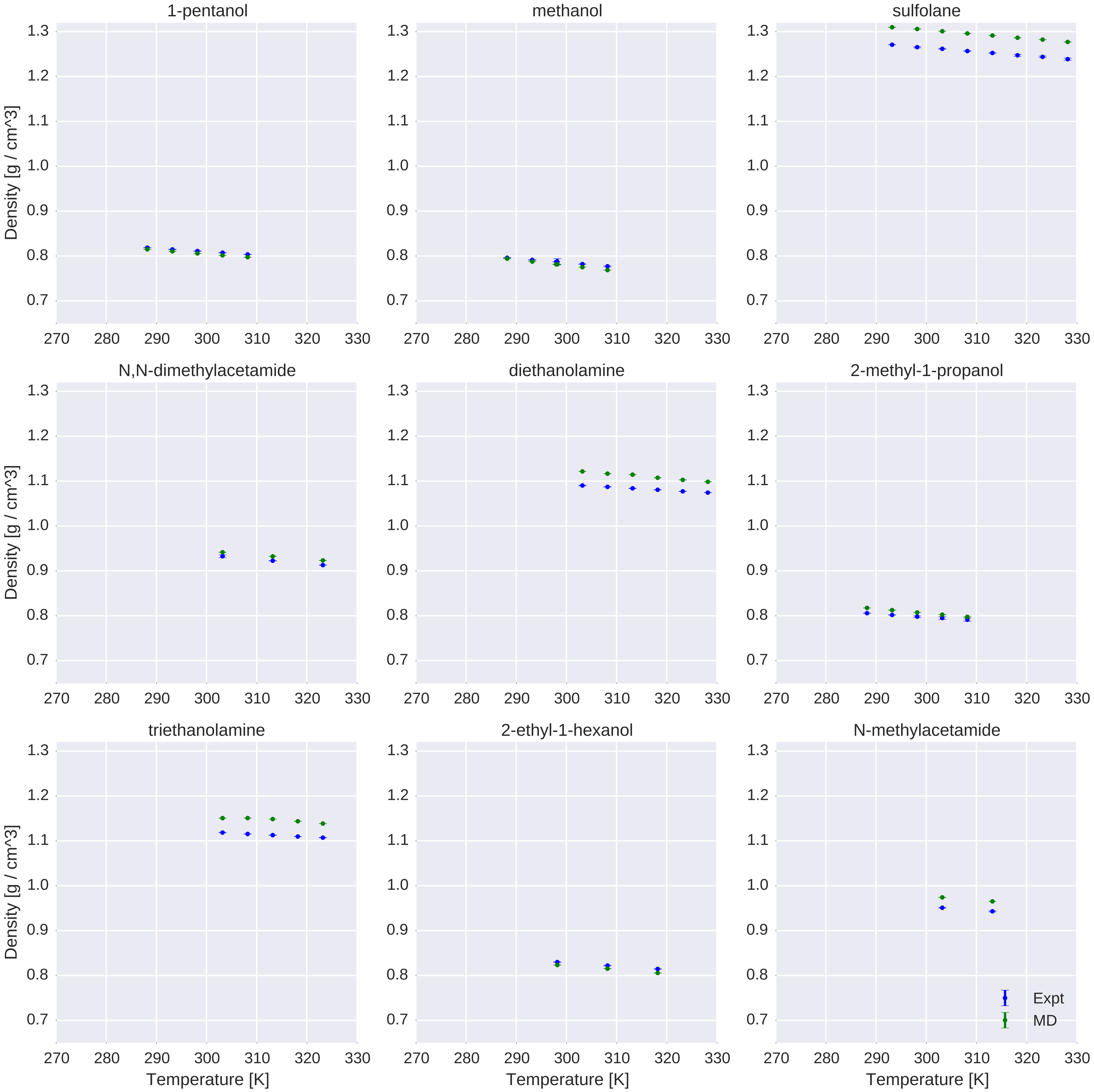


FIG. 7: **Comparison of simulated and experimental densities for all compounds.** Measured (blue) and simulated (green) densities are shown in units of g / cm^3.

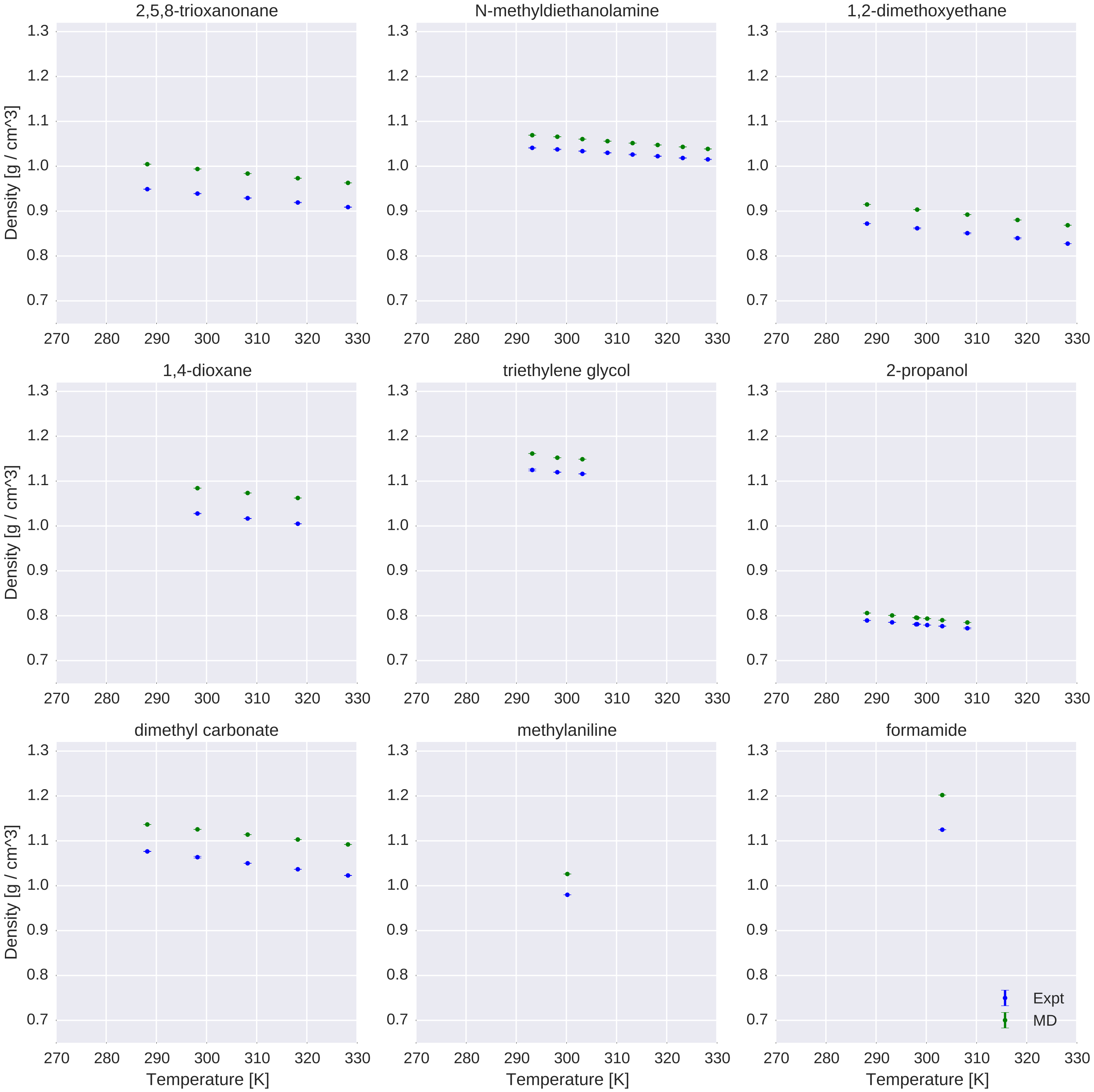


FIG. 7: **Comparison of simulated and experimental densities for all compounds.** Measured (blue) and simulated (green) densities are shown in units of g / cm^3.

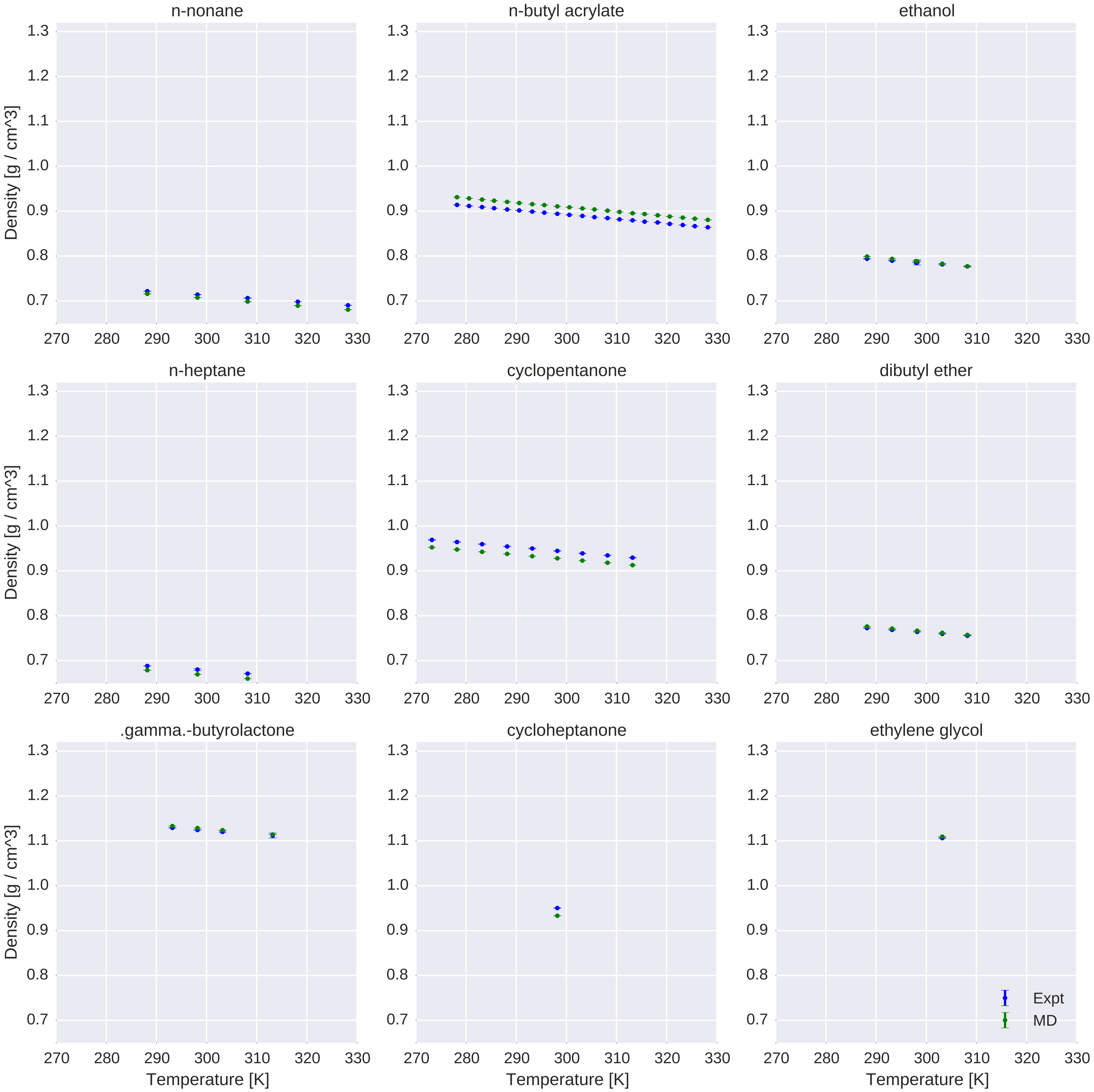


FIG. 7: **Comparison of simulated and experimental densities for all compounds.** Measured (blue) and simulated (green) densities are shown in units of g / cm^3.

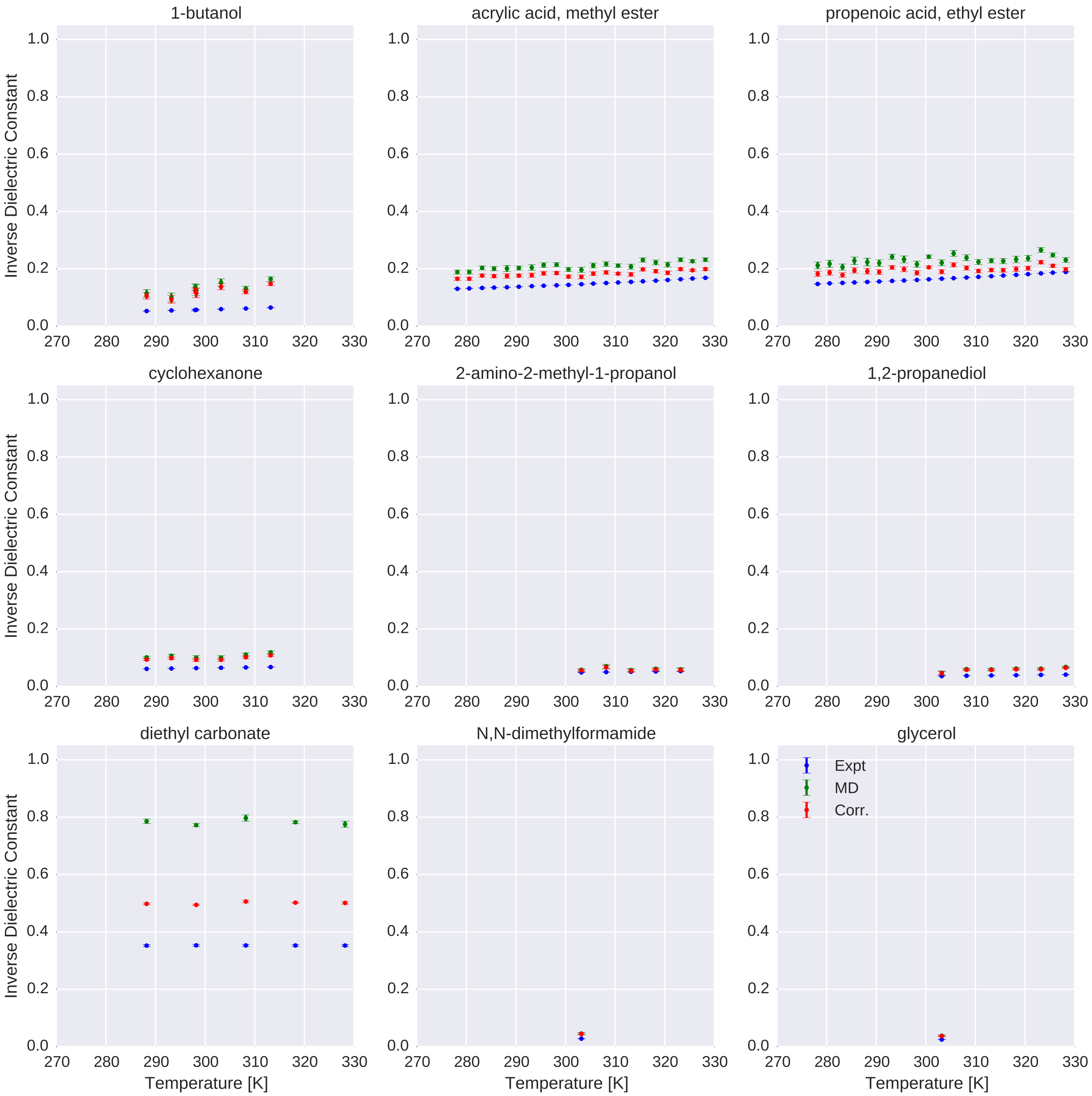


FIG. 8: **Comparison of simulated and experimental static dielectric constants for all compounds.** Measured (blue), simulated (green), and polarizability-corrected simulated (red) static dielectric constants are shown for all compounds.

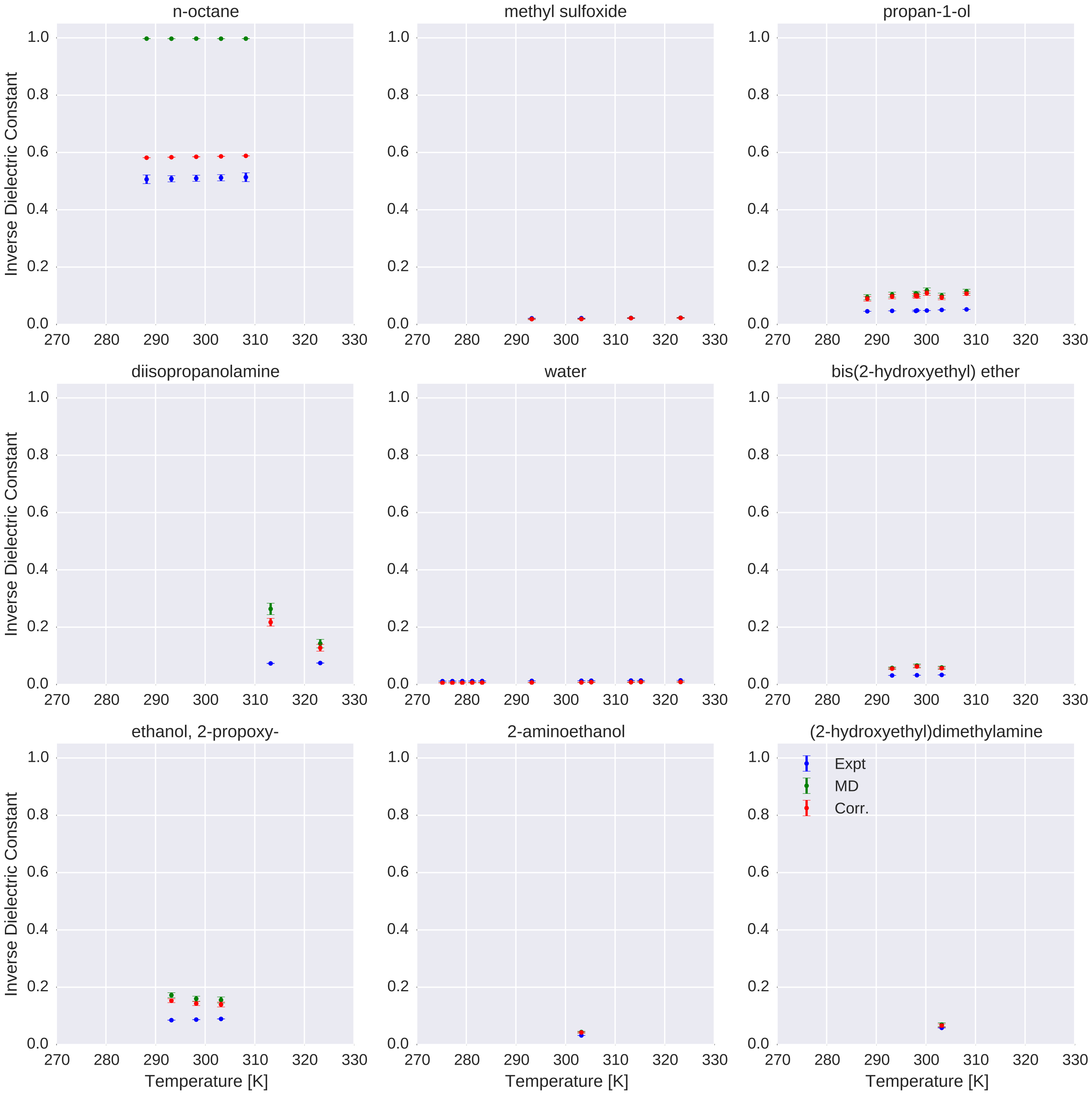


FIG. 8: **Comparison of simulated and experimental static dielectric constants for all compounds.** Measured (blue), simulated (green), and polarizability-corrected simulated (red) static dielectric constants are shown for all compounds.

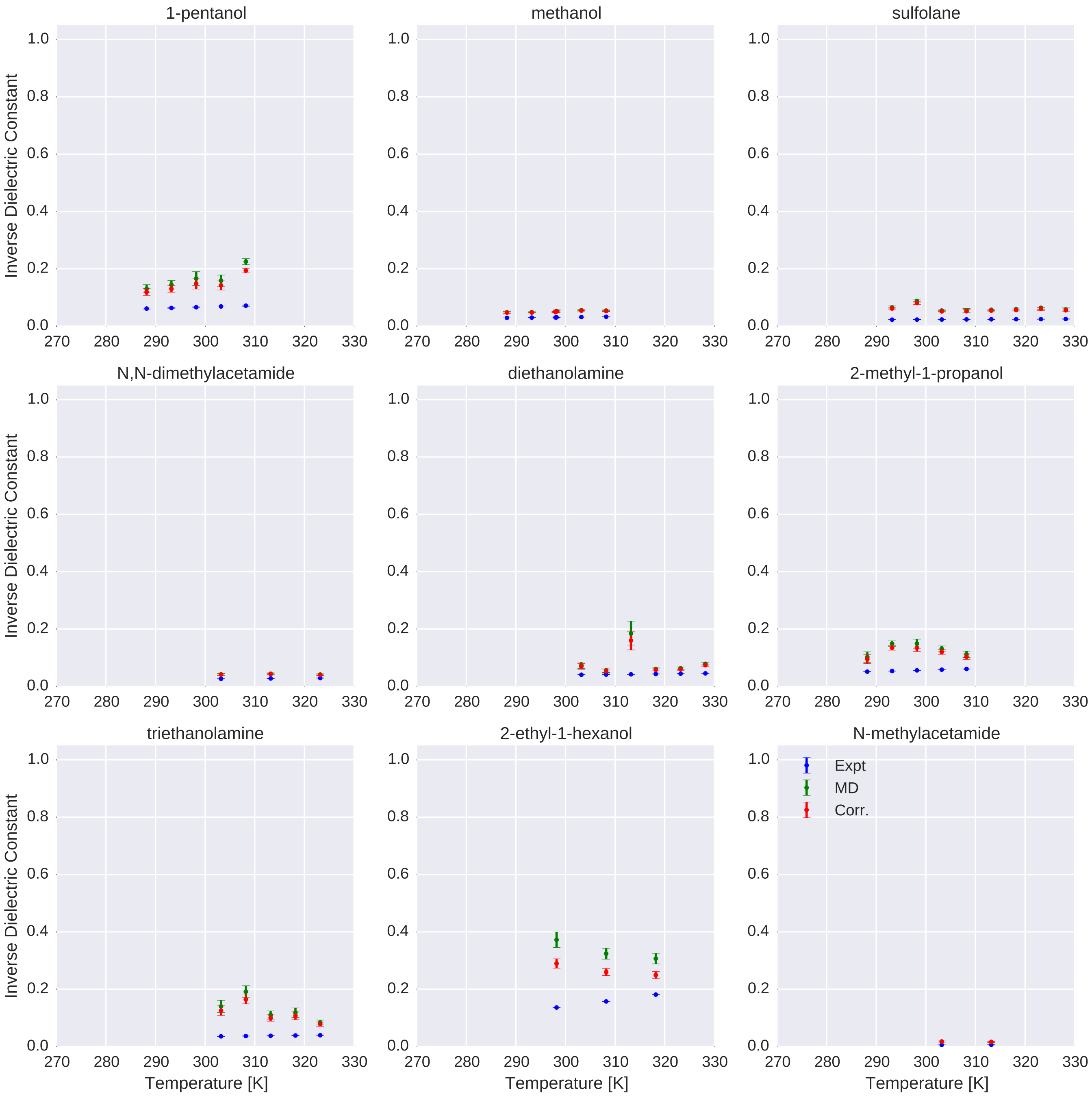


FIG. 8: **Comparison of simulated and experimental static dielectric constants for all compounds.** Measured (blue), simulated (green), and polarizability-corrected simulated (red) static dielectric constants are shown for all compounds.

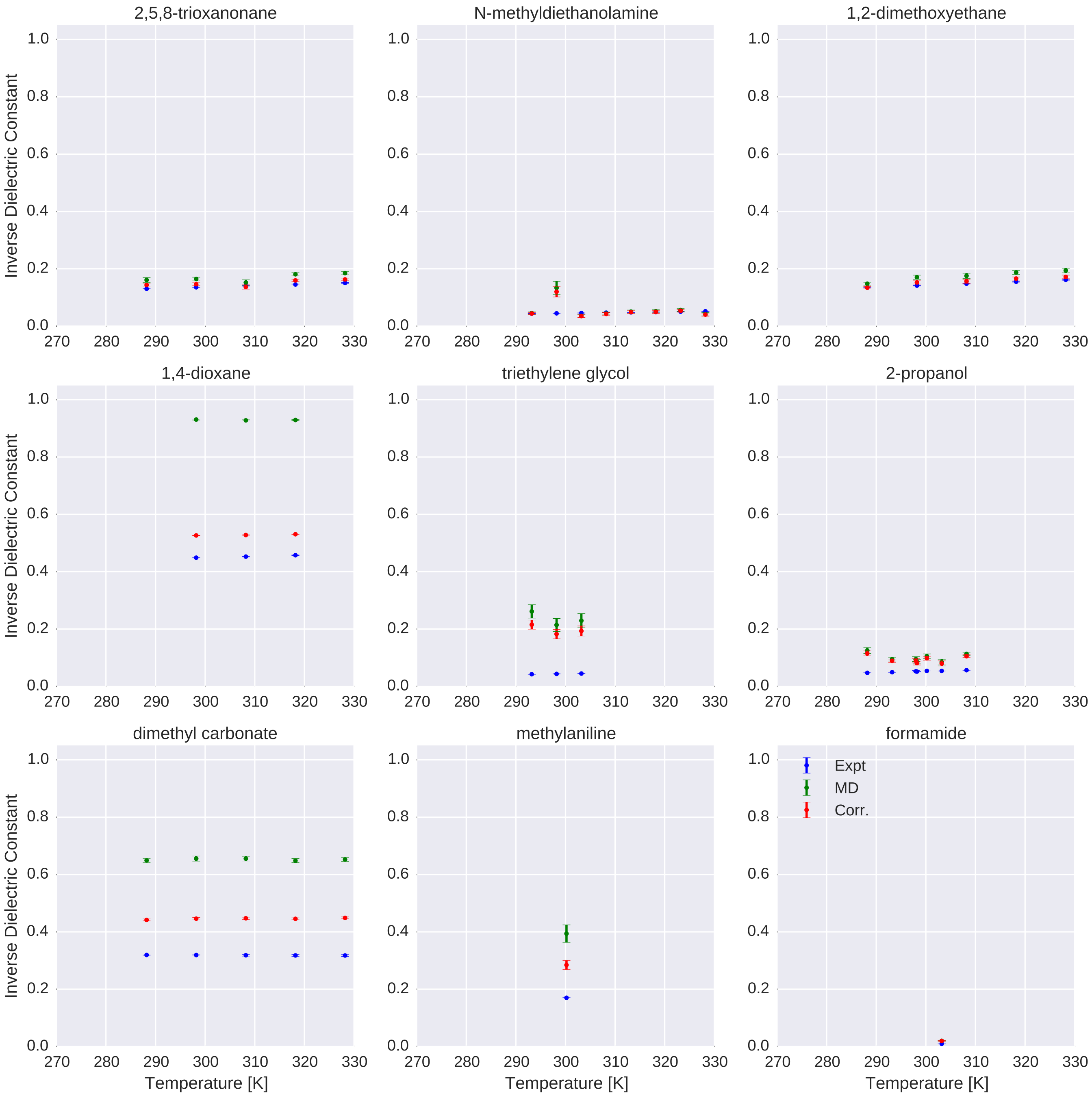


FIG. 8: **Comparison of simulated and experimental static dielectric constants for all compounds.** Measured (blue), simulated (green), and polarizability-corrected simulated (red) static dielectric constants are shown for all compounds.

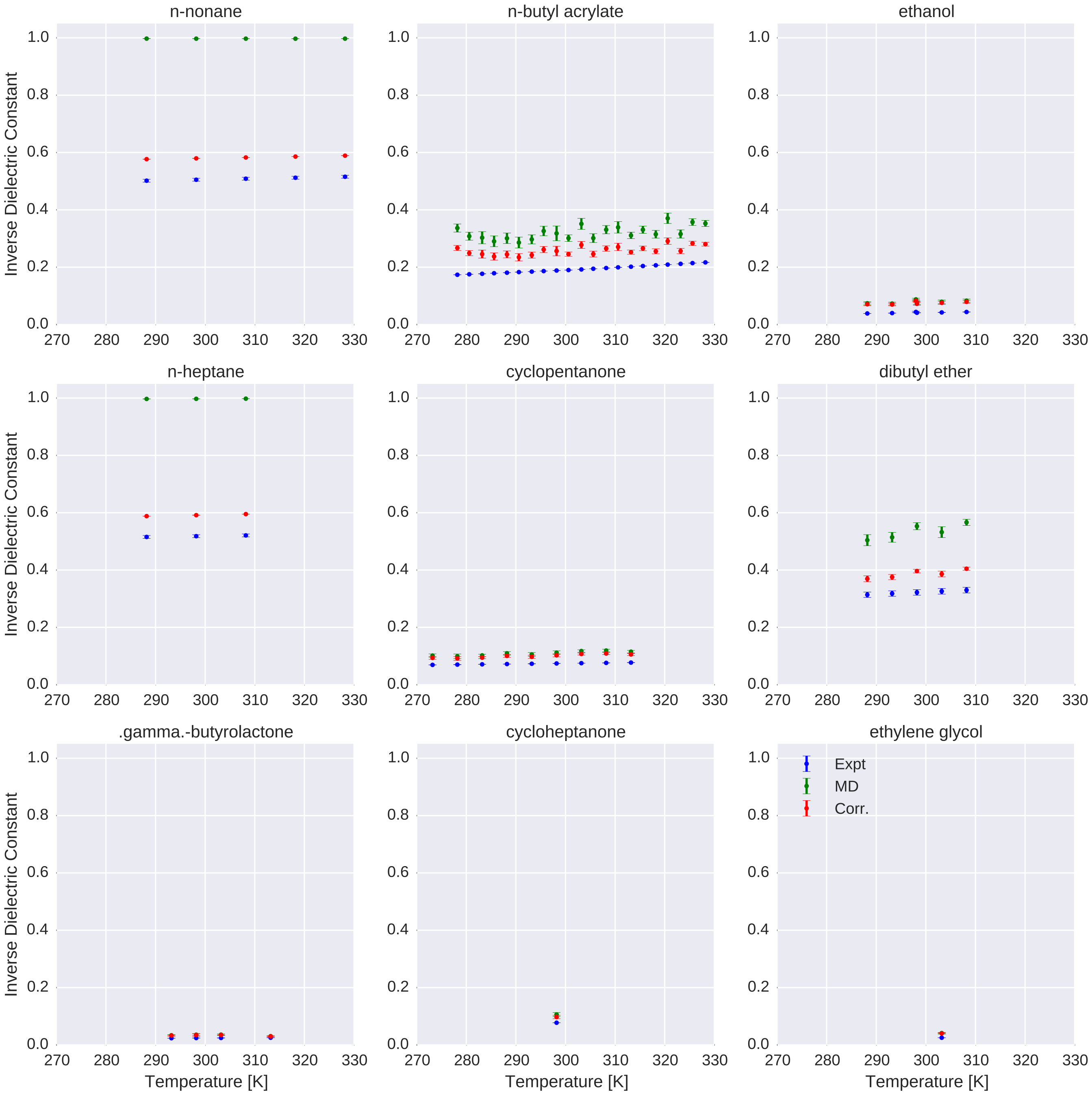


FIG. 8: **Comparison of simulated and experimental static dielectric constants for all compounds.** Measured (blue), simulated (green), and polarizability-corrected simulated (red) static dielectric constants are shown for all compounds.

1. Dependency Installation

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 "openmoltools" "pymbar==2.1" "mdtraj==1.3" "openmm==6.3" packmol
%
```

Note that this command installs the exact versions used in the present study, with the exception of openmoltools for which only a more recent package is available. However, for authors interested in extending the present work, we suggust using the most up-to-date versions available instead, which involves replace the equality symbols == with >=.

[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] K. Beauchamp, Y. Lin, R. Das, and V. Pande, J. Chem. Theory Comput. **8**, 1409 (2012).

[4] R. Best, N. Buchete, and G. Hummer, Biophys. J. **95**, L07 (2008).

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

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

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

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

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

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

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

[12] 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).

[13] N. M. Fischer, P. J. van Maaren, J. C. Ditz, A. Yildirim, and D. van der Spoel, Journal of Chemical Theory and Computation (2015).

[14] J. Zhang, B. Tuguldur, and D. van der Spoel, Journal of Chemical Information and Modeling (2015).

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

[16] 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).

[17] 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.

[18] 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).

[19] J. Pontolillo and R. P. Eganhouse, Tech. Rep. Water-Resources Investigations Report 01-4201, U.S. Geological Survey, Reston, Virginia (2001).

[20] 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).

[21] 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).

[22] 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).

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

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

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

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

[27] W. McKinney, in *Proceedings of the 9th Python in Science Conference*, edited by S. van der Walt and J. Millman (2010), pp. 51 – 56.

[28] M. Swain, *Cirpy-a python interface for the chemical identifier resolver (cir)*, URL https://github.com/mcs07/CIRp.

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

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

[31] URL http://www.ime.unicamp.br/~martinez/packmol/.

[32] C. Velez-Vega, D. J. McKay, V. Aravamuthan, R. Pearlstein, and J. S. Duca, Journal of chemical information and modeling **54**, 3344 (2014).

[33] P. C. Hawkins and A. Nicholls, Journal of chemical information and modeling **52**, 2919 (2012).

[34] 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).

[35] URL http://github.com/choderalab/openmoltools.

[36] 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).

[37] 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).

[38] D. C. Liu and J. Nocedal, Mathematical programming **45**, 503 (1989).

[39] J. A. Izaguirre, C. R. Sweet, and V. S. Pande, Pacific Symposium on Biocomputing **15**, 240 (2010).

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

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

[42] J. D. Chodera, N. Singhal, V. S. Pande, K. A. Dill, and W. C. Swope, J. Chem. Phys. **126**, 155101 (2007).

[43] S. Brooks, A. Gelman, G. Jones, and X.-L. Meng, *Handbook of Markov Chain Monte Carlo* (CRC press, 2011).

[44] M. Neumann, Molecular Physics **50**, 841 (1983).

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

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

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

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

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

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

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

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

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

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

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

[56] 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.

[57] J. Newman, V. J. Fazio, T. T. Caradoc-Davies, K. Branson, and T. S. Peat, Journal of biomolecular screening (2009).

[58] 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).

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

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

[61] 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).

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

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

[64] 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).

[65] 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).

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

[67] *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.