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

Kyle A. Beauchamp[+],[1] Julie M. Behr[+],[1] Patrick B. Grinaway,[1]
Arien S. Rustenburg,[1] Kenneth Kroenlein,[2] and John D. Chodera[1,*]

[1]*Memorial Sloan-Kettering Cancer Center, New York, NY*
[2]*NIST Thermodynamics Research Center, Boulder, CO*

(Dated: February 3, 2015)

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 feasibility of benchmarking atomistic forcefields against the NIST ThermoML database of physicochemical measurements, which aggregates thousands of density, dielectric, and other measurements. We present a detailed benchmark of the GAFF AM1-BCC 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.

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

I. INTRODUCTION

Recent advances in hardware and molecular dynamics software have provided routine access to atomistic simulations at the 100 ns timescale and beyond. 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 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. RSCB [], 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 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 fixed-charge forcefields of drug-like molecules. 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.

II. RESULTS

A. 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 drug-like elements (H, N, C, O, S, P, F, Cl, Br), has low heavy atom count (≤ 10), has ambient temperature [K] ($270 \leq T \leq 330$), has ambient pressure [kPA] ($100 \leq P \leq 102$),

* Corresponding author; john.chodera@choderalab.org

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. Liquid state	13573	461
6. Aggregate T, P	3573	432
7. Density+Dielectric	245	245

TABLE I. ThermoML Statistics

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 [**?**].

and has density greater than 300 kg m^{-3} (a proxy for liquid state). After applying these filters, we also round all pressures within this range to exactly one atmosphere. We also round temperatures 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 I), we are left with 245 conditions for which both density and dielectric data are available. The functional groups present are summarized in Table I.

B. 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 GAFF AM1-BCC forcefield (Fig. 1). Overall,

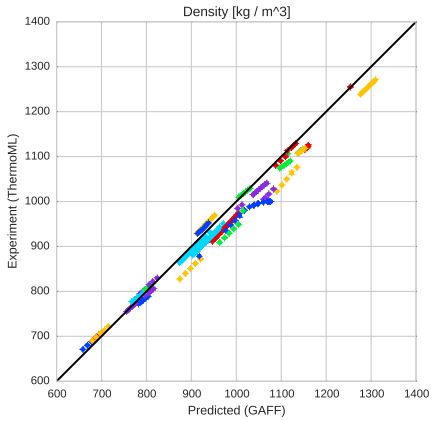


FIG. 1. Measured (ThermoML) versus predicted (GAFF) densities. Color groupings represent identical chemical formulas. Simulation error bars represent one standard error of the mean, with the number of effective (uncorrelated) samples estimated using pymbar. Experimental error bars indicate the standard 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.

the densities show reasonable accuracy (R2 + errobars), consistent with previous studies [**?**] reporting agreement of XYZ on a different benchmark set.

C. 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 somewhat orthogonal to density and thermodynamic quantities. We therefore compare 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, GAFF AM1-BCC systematically underestimates the dielectric constants for nonpolar organics, with GAFF 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 [**?**], which leads to better agreement with experiment. 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 agree-

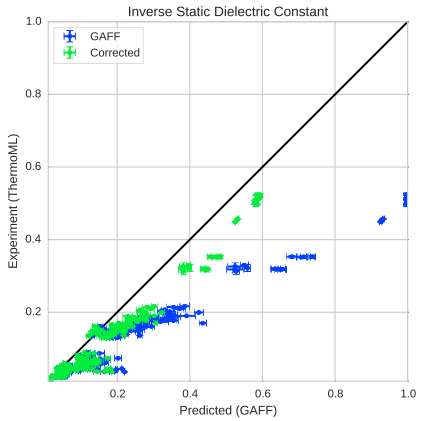
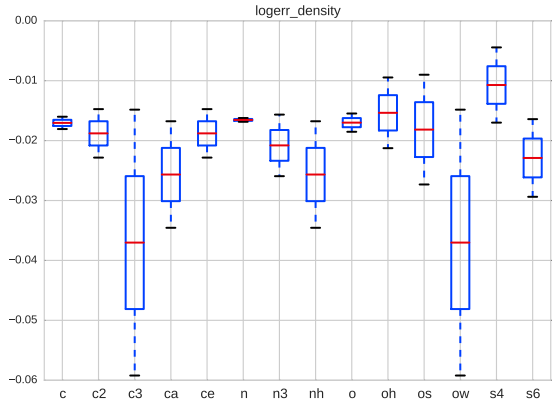


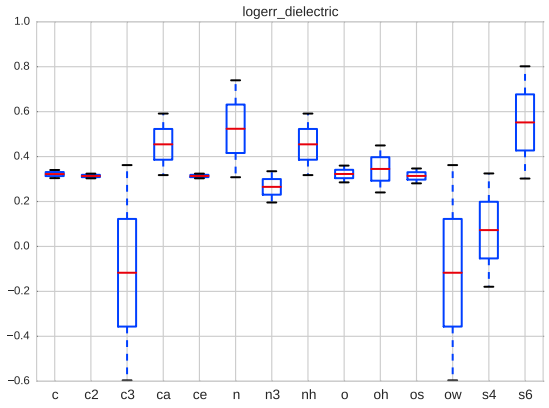
FIG. 2. Measured (ThermoML) versus predicted (GAFF) inverse static dielectrics (a). Color groupings represent identical chemical formulas. Simulation error bars represent one standard error of the mean estimated via block averaging with block sizes of 200 ps [**?**]. Experimental error bars indicate 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. 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. $U(r) = \frac{1}{4\pi\epsilon}\frac{q_1 q_2}{r} \propto \frac{1}{\epsilon}$.

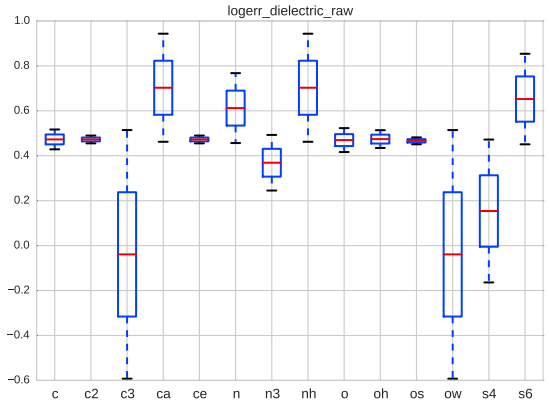
ment with experiment for both the GAFF and OPLS forcefields (Fig. 6).

III. DISCUSSION

A. Forcefield Accuracy Depends on Functional Group???







B. 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 force fields [**? ?**]. 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 $1.0 \leq \epsilon \leq 1.05$. 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 con-

stants 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; some such models have been proposed recently [**?**].

C. ThermoML as a Data Source

The present work has focused on the neat liquid density and dielectric measurements present in ThermoML [**? ? ?**] 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 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.

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 parameter set OECharges_AM1BCCSym. The selected conformer was then processed using antechamber in AmberTools 14 [**?**]. The resulting AMBER files were converted to OpenMM [**?**] XML files. Simulation code used libraries gaff2xml 0.6, TrustButVerify 0.1, openmm 6.2, and MDTraj [**?**] 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 (Table III). Pressure coupling was achieved with a Monte Carlo barostat applied every 25 steps. Particle mesh Ewald [**?**] 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 [**?**]. 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

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

VI. 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.

VII. DISCLAIMERS

	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.

VIII. 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.

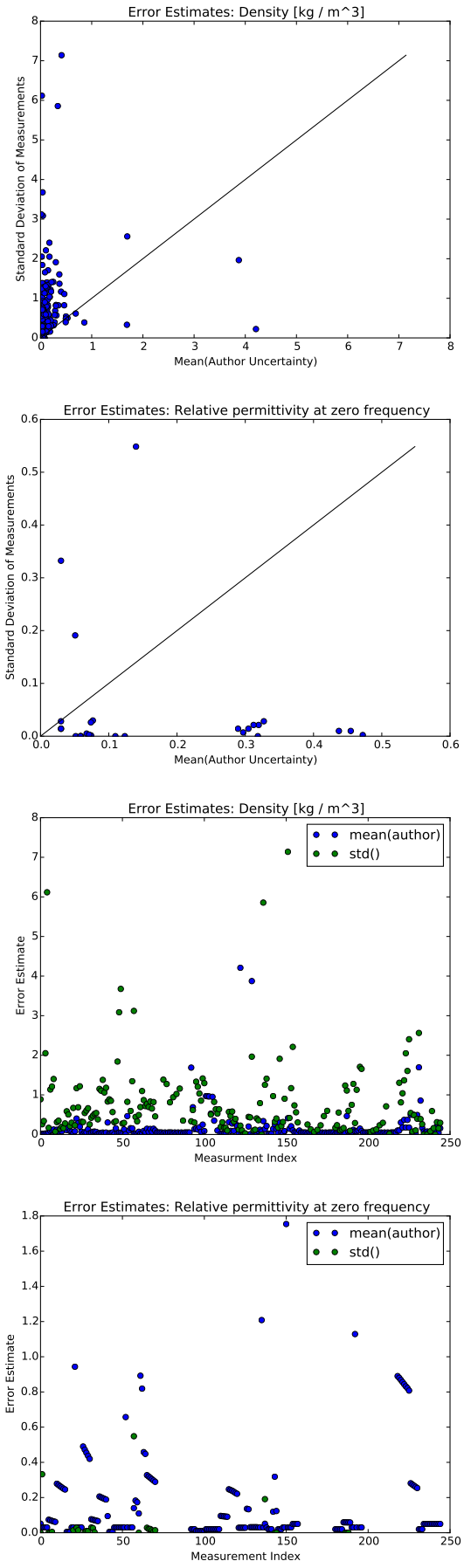


FIG. 3. To assess the experimental error in our benchmark set, we consider two orthogonal error measurements. The first is the mean of the uncertainties reported by the measurement authors. The second is the standard deviation of measurements

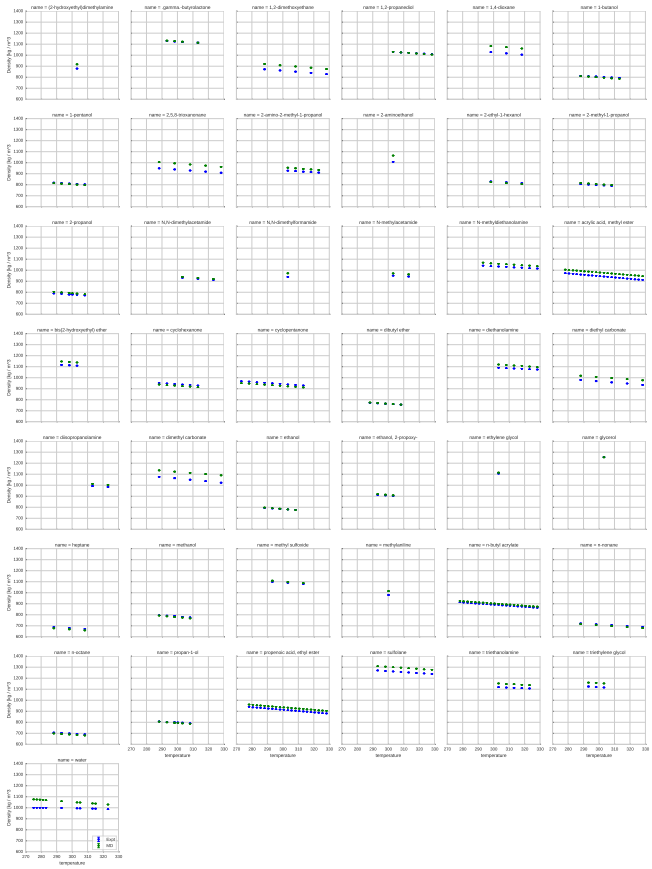


FIG. 4. Measured (blue) and simulated (green) densities [kg / m3] for all compounds.

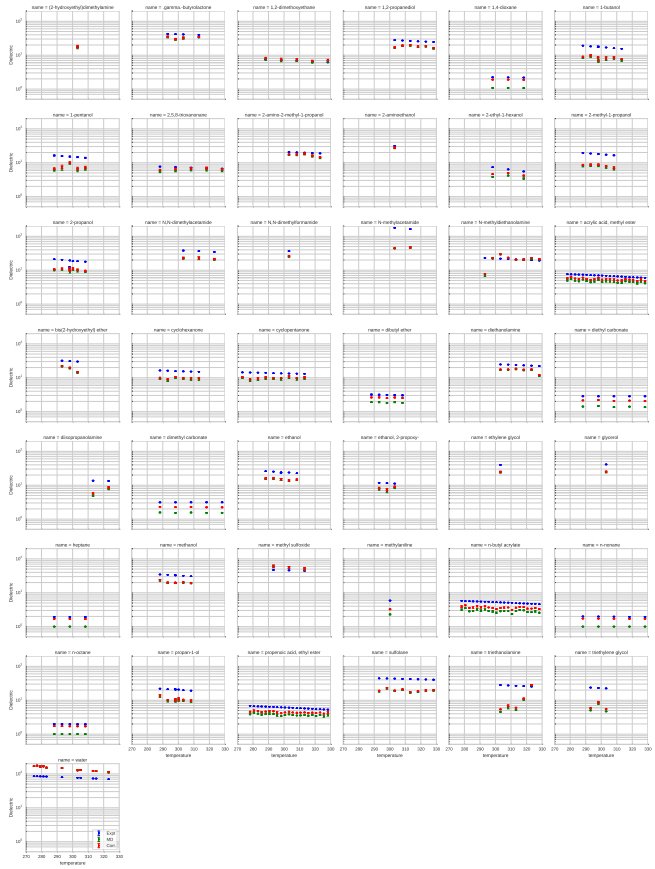


FIG. 5. Measured (blue), MD (green), and MD + polarizability-corrected (red) dielectrics for all compounds. Note that these are dielectrics, *not* inverse dielectrics.

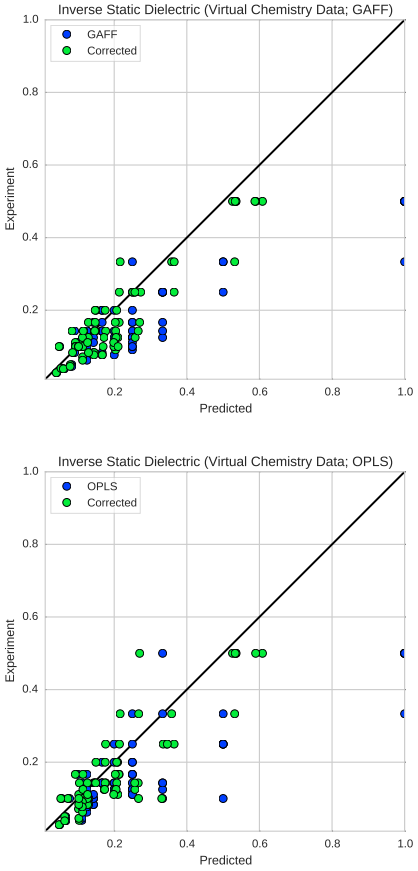


FIG. 6. Measured (blue), MD (green), and MD + polarizability-corrected (red) dielectrics for the virtualchemistry dataset [**? ?**].