JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

Cite This: J. Chem. Inf. Model. 2020, 60, 249−258

pubs.acs.org/jcim

Benchmarking Electronic Structure Methods for Accurate Fixed-Charge Electrostatic Models

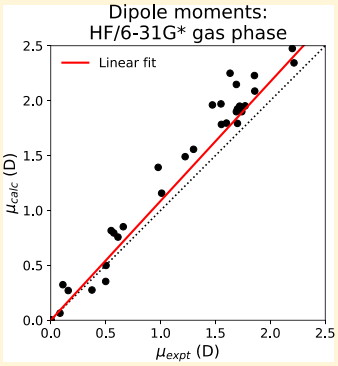
Alex Zhou,[†] Michael Schauperl,[‡] and Paul S. Nerenberg*[,†,§]

[†]Department of Physics & Astronomy and [§]Department of Biological Sciences, California State University, Los Angeles, Los Angeles, California 90032, United States

[‡]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego, La Jolla, California 92093, United States

Supporting Information

ABSTRACT: The accuracy of classical molecular mechanics (MM) force fields used for condensed phase molecular simulations depends strongly on the accuracy of modeling nonbonded interactions between atoms, such as electrostatic interactions. Some popular fixed-charge MM force fields use partial atomic charges derived from gas phase electronic structure calculations using the Hartree−Fock (HF) method with the relatively small 6-31G* basis set (HF/6-31G*). It is generally believed that HF/6-31G* generates fortuitously overpolarized electron distributions, as would be expected in the higher dielectric environment of the condensed phase. Using a benchmark set of 47 molecules, we show that HF/6-31G* overpolarizes molecules by just under 10% on average with respect to experimental gas phase dipole moments. The overpolarization of this method/basis set combination varies significantly though and, in some cases, even leads to molecular dipole moments that are lower than experimental gas phase measurements. We further demonstrate that using computationally inexpensive density functional theory (DFT) methods, together with appropriate augmented basis sets and a continuum solvent model, can yield molecular dipole moments that are both more strongly and more uniformly overpolarized. These data suggest that these methods—or ones similar to them—should be adopted for the derivation of accurate partial atomic charges for next-generation MM force fields.



INTRODUCTION

Fixed-charge molecular mechanics (MM) force fields have been a workhorse of biomolecular simulation since the first simulations of proteins were performed more than four decades ago.[1−4] Such force fields generally consist of relatively simple harmonic and periodic potentials to represent bonded interactions and Coulomb and Lennard-Jones potentials to represent nonbonded electrostatic and van der Waals interactions, respectively.[1] The key to the success of these models is their relatively low computational expense while providing nonbonded interactions that are reasonably accurate despite the neglect of electronic polarization, charge transfer, and a host of other important electrostatic phenomena. This is usually accomplished in a "mean field" way by utilizing partial atomic charges that correspond to molecular electron distributions that are overpolarized relative to the gas phase (i.e., more polar than would be expected for the gas phase), which is a sensible approach given that such simulations are generally performed in the condensed phase.[1,2]

There are many ways to achieve this overpolarization. Some force fields derive these partial atomic charges in an empirical way using condensed phase experimental data. Other force fields derive these charges in an ab initio way using some type of electronic structure calculation.[1,2] One specific set of force fields utilizing the latter approach are the AMBER family of biomolecular force fields based on AMBER ff94[5] and the GAFF small molecule force field.[6] This family of force fields relies on charges fit—in a two-step procedure with restraints[7]—to reproduce molecular electrostatic potentials (ESPs) generated using the Hartree−Fock (HF)/6-31G*[8−10] method and basis set combination. Alternatively, charges for GAFF may be generated using the AM1-BCC method that is parametrized to reproduce HF/6-31G* ESPs, but has some corrections that were hand-tuned to reproduce hydration free energies.[11,12]

It is commonly believed that HF/6-31G* combination results in charge distributions that are fortuitously 10−15% more polar than gas phase.[11] This is generally regarded as a desirable feature for condensed phase simulation, as any liquid (or solid) environment will have a dielectric constant greater than 1 and therefore overpolarize the molecule(s) in question relative to gas phase.

Evaluating the accuracy or suitability of a given charge derivation approach is difficult in that it is unclear what the "optimal" set of partial atomic charges for a given molecule should be. In reality, partial atomic charges are not a physical observable in any experiment. In addition, fixed-charge force fields necessarily neglect a great deal of physics from their description of nonbonded interactions and therefore depend

Received: October 14, 2019
Published: December 5, 2019

DOI: 10.1021/acs.jcim.9b00962
J. Chem. Inf. Model. 2020, 60, 249−258

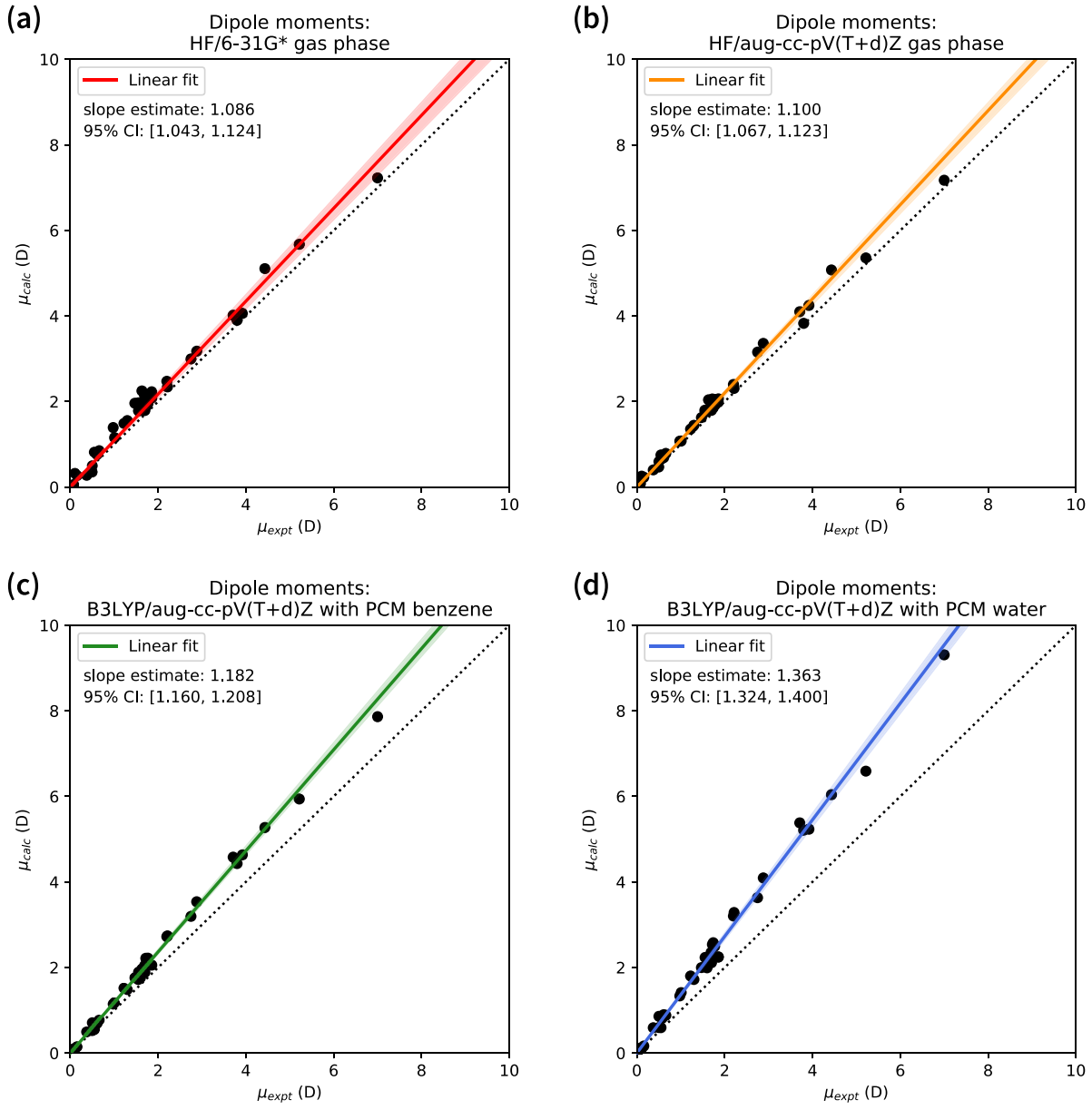


Figure 1. Calculated dipole moments versus experimental gas phase dipole moments for (a) HF/6-31G*, (b) HF/aug-cc-pV(T+d)Z, (c) B3LYP/aug-cc-pV(T+d)Z with PCM benzene, and (d) B3LYP/aug-cc-pV(T+d)Z with PCM water. The black dotted line indicates perfect agreement with experiment (i.e., no under- or overpolarization).

on some amount of error cancellation to return accurate results. Nonetheless, we posit that there are at least three desirable properties for any charge derivation method to be used for fixed-charge molecular simulations in the condensed phase:

1. The method should return a molecular charge distribution that is at least as overpolarized as it would be if the molecule were surrounded by a completely nonpolar dielectric medium.
2. The method should return a molecular charge distribution that is less overpolarized than it would be if the molecule were surrounded by aqueous solvent (or other highly polar dielectric media) because of the need to account for the cost of electronic polarization.[13−16]
3. The method should be consistent in predicting the relative overpolarization of the molecular charge distribution relative to the gas phase.

The first property ensures that the molecular charge distributions are overpolarized enough to be representative of dielectric environments likely to be encountered in condensed phase simulations. One possibility to satisfy properties 1 and 2 is for the method to utilize a polarizing medium/solvent that has a dielectric constant between the lowest (e.g., $\varepsilon \approx 1.8-2.0$ for alkanes) and highest (e.g., $\varepsilon \approx 80$ for water) dielectric constants likely to be encountered in typical biomolecular simulation scenarios. Indeed, previous studies have used solvents with intermediate dielectric constants to generate charges for biomolecular simulations in the aqueous phase to balance the effects of electronic distortion and polarization energies, thereby facilitating comparisons between free energies obtained from fixed-charge molecular dynamics (MD) simulations and experiments.[16] The third property ensures that the relative strength of the

DOI: 10.1021/acs.jcim.9b00962
J. Chem. Inf. Model. 2020, 60, 249−258

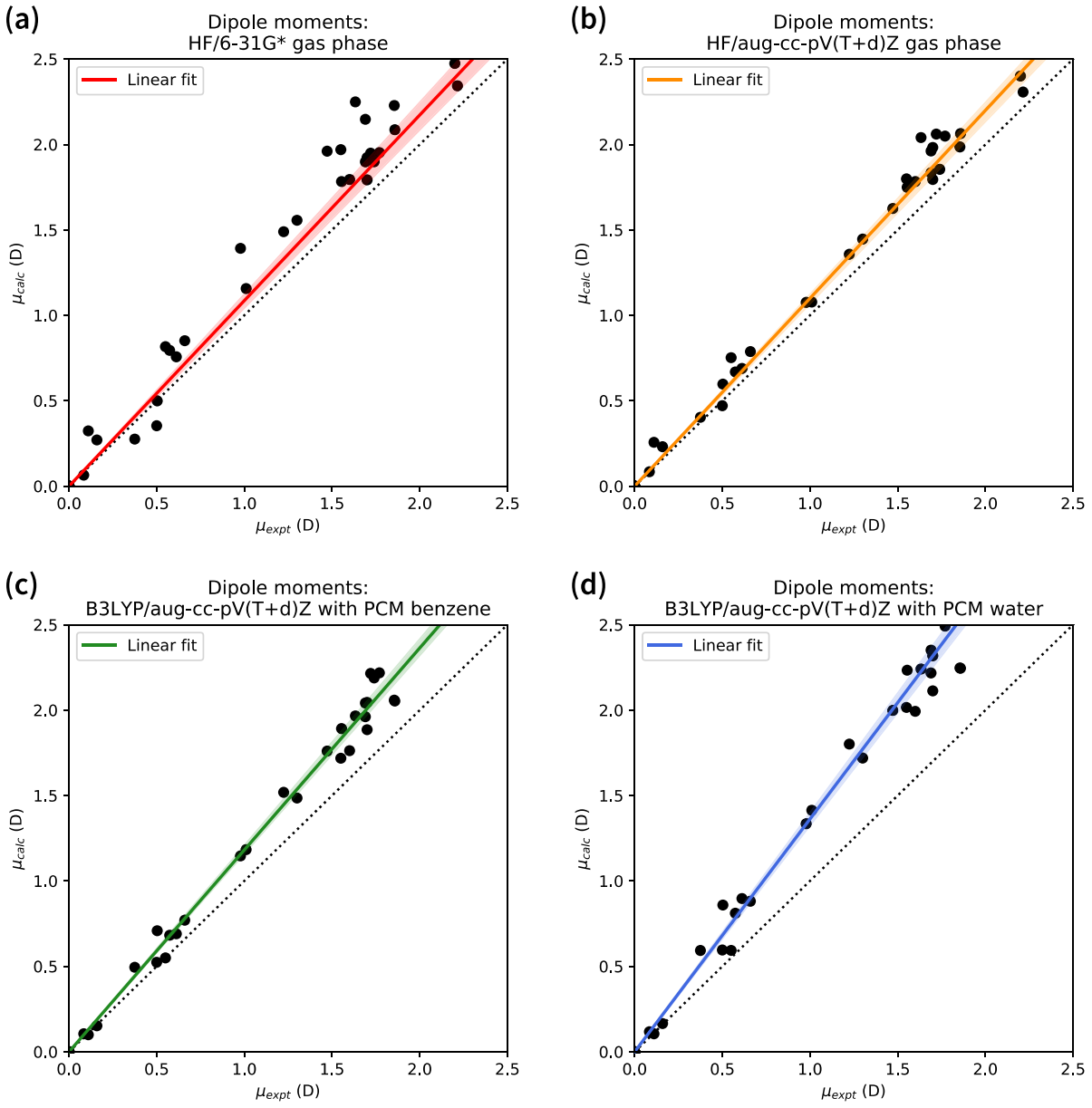


Figure 2. Calculated dipole moments versus experimental gas phase dipole moments—limited to 2.5 D to show detail—for (a) HF/6-31G*, (b) HF/aug-cc-pV(T+d)Z, (c) B3LYP/aug-cc-pV(T+d)Z with PCM benzene, and (d) B3LYP/aug-cc-pV(T+d)Z with PCM water. The black dotted line indicates perfect agreement with experiment (i.e., no under- or overpolarization).

predicted overpolarization will be the same across different chemical moieties and not result in systematic errors for certain moieties (i.e., it will be transferable).

In this work we demonstrate that HF/6-31G* yields a characteristic overpolarization that is in agreement with previous studies,[11,17] but that this overpolarization is not consistent across all chemical moieties. In particular, for some moieties the overpolarization is >35%, while other moieties are underpolarized relative to the gas phase (i.e., unphysically low dipole moments). In contrast, the combination of a relatively simple hybrid generalized gradient approximation (GGA) (or meta-GGA) functional, augmented basis set, and implicit solvent yields dipole moments with greater characteristic overpolarization and much greater consistency across all moieties. We then discuss these results in the context of both historical and recent work in the force field development community, as well as their implications for future force field development.

METHODS

Benchmark Set. For this work we utilized the 46 molecule set of Hickey and Rowley[17] that has experimentally measured dipole moments[18] ranging between 0 and 7.0 D. Despite its relatively small size, this benchmark set includes a variety of inorganic and organic molecules with moieties that span much of biochemistry. We added one additional molecule to this set, ammonia borane, which provides an experimentally measured dipole moment[19,20] (5.22 D) that is intermediate to the two highest dipole moments in this set, dimethylsulfone (4.43 D) and cytosine (7.00 D), which otherwise have a relatively large gap between them. All molecular structures were built from

DOI: 10.1021/acs.jcim.9b00962
J. Chem. Inf. Model. 2020, 60, 249−258

SMILES strings using a Python interface to Open Babel 2.4.1.[21] In the case of acetic acid, the structure was modified to the *syn* conformer (the preferred conformer in gas phase) using Avogadro 1.2.0.[22]

Geometry Optimization. All geometry optimizations were performed using Psi4 1.3[23] and the B3LYP[24−27]/cc-pV(T+d)Z[28−30] combination of method and basis set. The geometry optimization convergence criteria were the Psi4 default (i.e., "QChem"). This methodology is in contrast to some previous studies, which have allowed the method/basis set combination to vary for geometry optimization along with property calculation. We believe that using a single set of optimized structures is important to avoid conflating the behavior of the method/basis set combination with respect to intramolecular interactions (which in turn determine the minimum energy structure) versus molecular properties.

Property Calculations. Molecular dipole moments were calculated using several different combinations of method and basis set: HF/6-31G*, HF/aug-cc-pV(T+d)Z, B3LYP/aug-cc-pV(T+d)Z, and PW6B95[31]/def2-TZVPD.[32,33] In addition, for the B3LYP and PW6B95 combinations we calculated dipole moments using polarizable continuum models (PCMs)[34] of benzene (ε = 2.25) and water (ε = 78.39), as implemented by the CPCM method[35,36] using UFF atomic radii[37] in the PCMSolver module.[38]

Statistical Analysis. All linear regression fits were performed using the nonparametric Theil−Sen estimator[39,40] for slopes, as implemented in the Python library SciPy.[41] The use of a Theil−Sen estimator is important because rather than attempting to minimize the sum of squared residuals—a framework that also requires that the residuals have constant variance and be normally distributed—it simply finds the median slope among all pairs of points in the data. This methodology enables the Theil−Sen estimator to be robust to outliers. The 95% confidence intervals for the slope estimates were computed using Sen's method.[40]

RESULTS

Characteristic Overpolarization. To determine the characteristic overpolarization of each method relative to gas phase experimental data, we fit a robust, nonparametric linear regression model[39,40] (see Methods) to the dipole moments obtained using each method/basis set combination and the experimental gas phase dipole moments. The use of a robust method is warranted here to minimize the impacts of outliers, including cytosine, which in previous studies[17] has been noted to have a potentially problematic experimental measurement. The point estimate provided by the Theil−Sen estimator that we used can be regarded as the median overpolarization of the method/basis set combination with respect to the experimental gas phase data.

As shown in Figure 1a, we found that the point estimate for the slope of HF/6-31G* dipole moments relative to experimental gas phase dipole moments is 1.086 with a 95% confidence interval (CI) of [1.043, 1.124]. In other words, these HF/6-31G* dipole moments are—on average—expected to be somewhere between 4.3 and 12.4% larger than experimental gas phase dipole moments. Figure 2a, which focuses on molecules with experimental dipole moments below 2.5 D, reveals that there is considerable variation about the regression line, suggesting that there is a large variability in the predicted overpolarization of the HF/6-31G* combination.

The HF/aug-cc-pV(T+d)Z combination yields dipole moments that are 1.100 times larger (95% CI [1.067, 1.123]) (Figure 1b). This is a somewhat smaller CI than the one obtained for HF/6-31G*, indicating that HF/aug-cc-pV(T+d)Z likely yields more precise predictions of this overpolarization (Figure 2b). Nonetheless, both the HF/6-31G* and HF/aug-cc-pV(T+d)Z combinations yield point estimates for the characteristic overpolarization of ∼10% relative to experimental gas phase measurements, suggesting that this is the amount of overpolarization inherent to the HF method itself.

We then examined the overpolarization using B3LYP/aug-cc-pV(T+d)Z, a combination of a widely used density functional method and widely used augmented basis set. When using no implicit solvent (i.e., in gas phase), this method/basis set combination yields a slope estimate of 1.023 with a 95% CI of [1.004, 1.037] (Figure S1). This indicates that there is a small amount of overpolarization relative to experiment inherent to this method/basis set, but that it is ∼4 times smaller than the overpolarization of HF/6-31G* or HF/aug-cc-pV(T+d)Z.

In concert with a PCM model of benzene (ε = 2.25), however, this combination yields dipole moments that are 1.182 times larger (95% CI [1.160−1.208]) than experimental gas phase measurements (Figure 1c). We note that this 95% CI does not overlap with the interval for HF/6-31G* (or HF/aug-cc-pV(T+d)Z). Moreover, the point estimate of 18.2% is more than twice as large as that of HF/6-31G*. This is a surprising observation given that HF/6-31G* gas phase calculations are thought to mimic the condensed phase and that benzene has a relatively low dielectric constant consistent with only nonpolar condensed phase environments. Additionally, we find that the variance of data from the regression line is comparable to the HF/aug-cc-pV(T+d)Z data (Figure 2c).

Finally, we examined the overpolarization of B3LYP/aug-cc-pV(T+d)Z with PCM water (ε = 78.4). In this higher dielectric environment, the calculated dipole moments are 1.363 times larger (95% CI [1.324, 1.400]) than experimental gas phase dipole moments (Figure 1d), suggesting a characteristic overpolarization that is more than 4 times the HF/6-31G* result. It is additionally apparent that some of the calculated dipole moments are further from the regression line for PCM water than for PCM benzene (Figure 2d).

Consistency of Predicted Overpolarization. One might be tempted to use the common goodness-of-fit metric R^2 to describe the consistency of the overpolarization predictions, but it is important to note that the correlation—or strength of linear relationship—between the response variable (the calculated dipole moments, μ_{calc}) and explanatory variable (the experimental dipole moment, μ_{expt}) is very strong regardless of the electronic structure method employed. Therefore, such a metric is unlikely to reveal meaningful differences between various electronic structure methods. To examine the consistency of the predicted overpolarization of each method, we instead elected to perform an analysis of the residuals arising from each linear regression model. Methods that yield more consistent overpolarization should yield residuals that are on average smaller than methods for which the strength of the overpolarization varies considerably.

Analyzing the residuals from these linear model fits requires some care. In particular, one of the core assumptions of ordinary least squares (OLS) linear regression is that the residuals ($\mu_{calc} - \hat{\mu}_{calc}$) should have a constant variance. In

DOI: 10.1021/acs.jcim.9b00962
J. Chem. Inf. Model. 2020, 60, 249−258

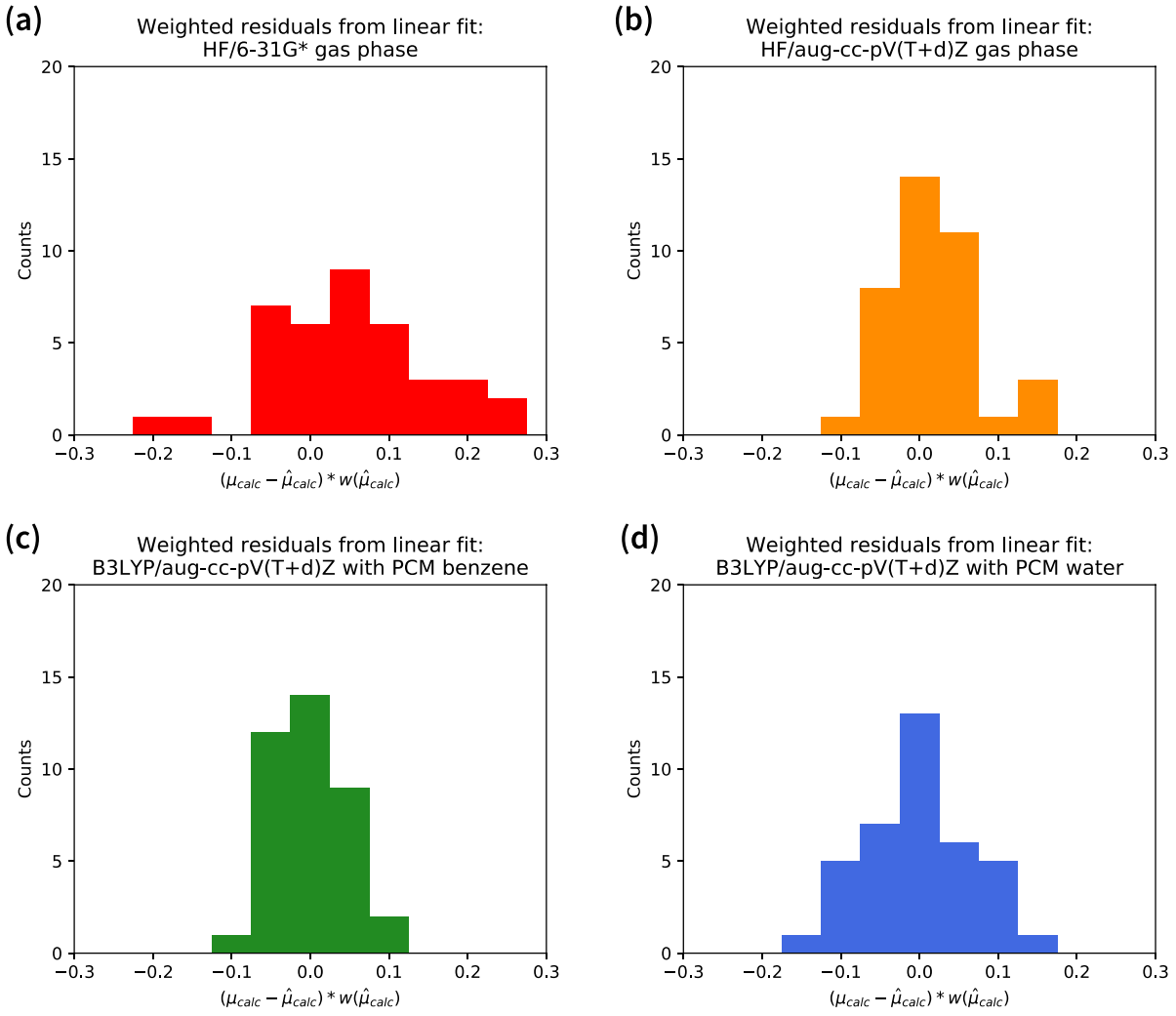


Figure 3. Weighted residuals for robust linear regression fits to experimental gas phase dipole moments for (a) HF/6-31G*, (b) HF/aug-cc-pV(T+d)Z, (c) B3LYP/aug-cc-pV(T+d)Z with PCM benzene, and (d) B3LYP/aug-cc-pV(T+d)Z with PCM water. A weighted residual of 0 indicates a prediction with no deviation from the regression line or, in other words, a prediction that is perfectly consistent with a linear model for overpolarization. All molecules with zero dipole moment are excluded from this analysis.

other words, the size of the residual should not depend on the magnitude of the fitted predicted dipole moment ($\hat{\mu}_{calc}$), which in this case is just μ_{expt} multiplied by the slope of the linear model fit. Such behavior is readily violated in electronic structure calculations because, for example, nonpolar molecular species that have experimental dipole moments of exactly 0 are reliably predicted to have dipole moments of exactly 0 as well (i.e., 0 variance in the prediction). [As described under Methods, this is another reason why we have not employed OLS linear regression in this work and have instead used nonparametric Theil−Sen linear regression for our statistical inference.] A more reasonable assumption might be that electronic structure calculations yield predictions that have some inherent percent or fractional error. If this were the case, then the residuals should be divided by the fitted dipole moments to yield a fractional residual (i.e., $(\mu_{calc} - \hat{\mu}_{calc})/\hat{\mu}_{calc}$) before comparing different electronic structure methods. This approach, however, can lead to molecules with small dipole moments erroneously dominating the residual analysis.

Recent work by Hait and Head-Gordon reached a similar conclusion to our discussion above.[42] Their solution to a similar problem (i.e., errors relative to a CCSD(T)/CBS gas phase reference) was to use exact differences for small dipole moment species (≤1.0 D) and fractional differences for large dipole moment species (>1.0 D). While this does "regularize" the errors to some extent, it introduces an arbitrary transition in how to treat the differences at 1.0 D, a dipole moment that has no special physical meaning. Our solution to this problem is to multiply the residuals by a weighting function:

$$w(\hat{\mu}_{calc}) = \mathrm{erf}\left(\frac{\sqrt{\pi}\,\hat{\mu}_{calc}}{2}\right)/\hat{\mu}_{calc}$$

where $\hat{\mu}_{calc}$ is the predicted dipole moment from the linear regression model and erf(x) is the error function. [We note that the $\hat{\mu}_{calc}$ inside the error function is assumed to be multiplied by a constant of 1 D^{-1} to make the argument of the function dimensionless.] As the dipole moment of a molecule tends toward 0, our weighting function approaches 1. In other words, the weighted residual tends toward being an exact residual. As the dipole moment of a molecule becomes large, our weighting function approaches $1/\hat{\mu}_{calc}$, and therefore the weighted residual tends toward being a fractional residual. For example, by 2.0 D, the weighting function has a value that is

DOI: 10.1021/acs.jcim.9b00962
J. Chem. Inf. Model. 2020, 60, 249−258

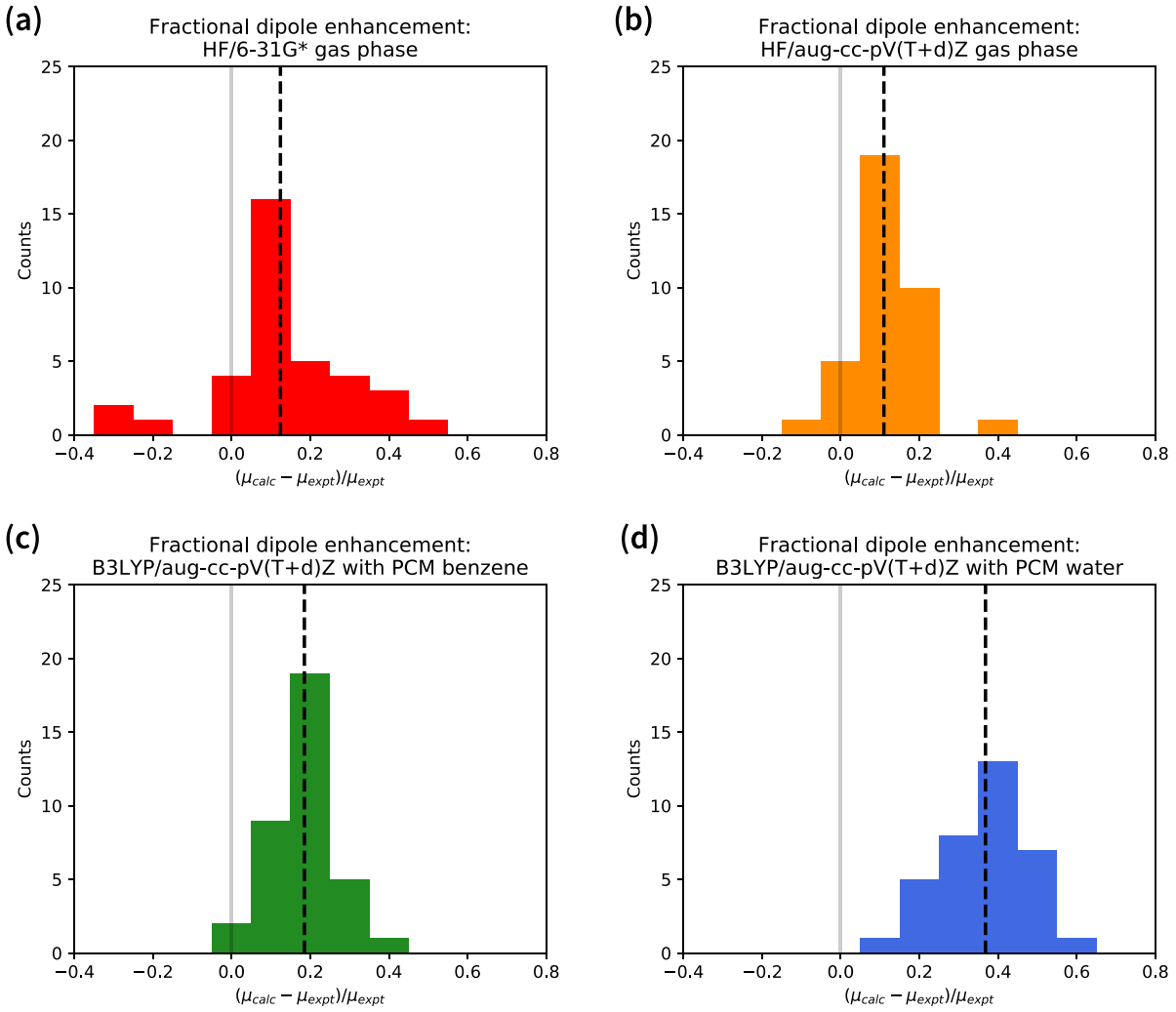


Figure 4. Fractional dipole enhancement relative to experimental gas phase data for (a) HF/6-31G*, (b) HF/aug-cc-pV(T+d)Z, (c) B3LYP/aug-cc-pV(T+d)Z with PCM benzene, and (d) B3LYP/aug-cc-pV(T+d)Z with PCM water. The solid gray line indicates perfect agreement with experiment; the black dashed line indicates the median for each method/basis set combination. All molecules with zero dipole moment are excluded from this analysis.

98.8% of $1/\hat{\mu}_{calc}$. In addition, this weighting function is continuous, smoothly varying, and does not require choosing any particular transition value for the "switch" between exact and fractional residuals.

The weighted residuals for each of the method/basis set combinations are shown in Figure 3. In concordance with the considerable variation of data about the regression line displayed in Figures 1a and 2a, the weighted residuals of the HF/6-31G* method show the broadest distribution of all four method/basis set combinations that we examined. In particular, the maximum and median absolute weighted residuals for HF/6-31G* are 0.261 and 0.052, respectively (Figure 3a). For comparison, the maximum and median absolute weighted residuals for HF/aug-cc-pV(T+d)Z are 0.135 and 0.040, respectively (Figure 3b). This complements our earlier observation regarding the 95% CI for the slope estimate for this method versus HF/6-31G* and suggests that the use of a more sophisticated basis set is indeed useful in improving the consistency of the overpolarization predictions, regardless of the underlying electronic structure method employed. The predictions of B3LYP/aug-cc-pV(T+d)Z with PCM benzene are even more consistent, however, with maximum and median absolute weighted residuals of 0.104 and 0.036, respectively (Figure 3c). In other words, the maximum absolute weighted residual for this combination is approximately 2.5 times smaller than for HF/6-31G*, and the median absolute weighted residual is 30% smaller. There is a reversal of this trend for B3LYP/aug-cc-pV(T+d)Z with PCM water, though. The summary statistics for the absolute weighted residuals increase to 0.154 and 0.045, respectively, which places the PCM water results in between HF/6-31G* and HF/aug-cc-pV(T+d)Z (although closer to the latter than the former). Readers interested in examining the exact residuals for the linear regression models are referred to Figure S2 in the Supporting Information.

Distribution of Overpolarization. We next examined the fractional dipole enhancement of each method/basis set combination with respect to the experimental gas phase data. We did this by calculating the difference between the calculated and experimental dipole moments and then dividing by the experimental dipole moments. For this analysis we excluded molecules with zero dipole moment, as well as two outliers—carbon monoxide and nitrogen monoxide—for which the HF method predicts molecular dipoles that are

DOI: 10.1021/acs.jcim.9b00962
J. Chem. Inf. Model. 2020, 60, 249−258

both in the opposite direction of and much greater in magnitude than the experimentally measured dipoles. In the results that follow we have converted the fractional dipole enhancements into percentages.

We find that the median dipole moment enhancement of HF/6-31G* for this subset of molecules is 12.4%, but as suggested by the scatter plots in Figures 1a and 2a, there are some species with HF/6-31G* dipole moments that are >20% *lower* than gas phase and others that are >35% greater than gas phase (Figure 4a). The former are all hydrocarbon species (toluene, pentene, and propane) and are examples of the unphysical nature of some HF/6-31G* predictions. The latter are predominantly sulfur-containing molecules (e.g., sulfur dioxide and thiophene) and phosphine.

The combination of HF/aug-cc-pV(T+d)Z yields a slightly smaller median dipole moment enhancement of 11.0% (Figure 4b). The unphysically low dipole moments of the hydrocarbon species mentioned above are largely eliminated with the change of basis set, as is the very strong overpolarization of sulfur-containing species. The lone outlier in this regard is thiophene.

The combination of B3LYP/aug-cc-pV(T+d)Z with PCM benzene and PCM water yields median dipole moment enhancements of 18.5 and 36.8%, respectively (Figure 4c,d). In both cases, there are no species with calculated dipole moments lower than the gas phase experimental data. Moreover, consistent with our observations of the weighted residuals in Figure 3, the spread of the fractional dipole moment enhancement is quite small when PCM benzene is used; there are no molecules that are three or more bins away (i.e., ≥0.30) from the median in Figure 4c and only three molecules in total that are two bins away (i.e., ≥0.20 from the median). The PCM water result shows a greater spread of fractional dipole moment enhancements from the median that is again intermediate to the spreads observed with HF/6-31G* and HF/aug-cc-pV(T+d)Z (Figure 4d).

Together these data are consistent with the characteristic overpolarization estimates and weighted residuals we obtained with the linear regression model fits described previously. In general, HF/6-31G* gas phase calculations do not overpolarize molecular charge distributions as strongly as a more sophisticated method using PCM benzene, a nonpolar solvent of low dielectric constant. The results for the HF/aug-cc-pV(T+d)Z gas phase calculations demonstrate that there is an inherent overpolarization due to the HF method, whereas the inconsistency in the strength of the overpolarization across different molecules is due to the 6-31G* basis set.

Comparison of Dipole Moments Obtained with PCM Benzene and PCM Water. Having used PCMs of benzene and water as the polarizing media for the B3LYP/aug-cc-pV(T+d)Z calculations, we wanted to assess how the low dielectric implicit solvent results compared to the high dielectric implicit solvent results. In particular we were interested in how overpolarized the dipole moments obtained in PCM benzene would be compared to dipole moments obtained in PCM water. To do this we computed the point estimate and 95% CI for the parameter β in the following linear regression model:

$$\mu_{\text{benzene}} = (1 - \beta)\mu_{\text{gas}} + \beta\mu_{\text{water}}$$

where all of the dipole moments are calculated using the same method/basis set combination. We note that the above formula does not have the usual form of a linear regression model, but such a form can be obtained by rearranging it in the following manner:

$$\mu_{\text{benzene}} - \mu_{\text{gas}} = \beta(\mu_{\text{water}} - \mu_{\text{gas}})$$

Using the Theil−Sen estimator, we found that the point estimate for β was 0.484 with a 95% CI of [0.476, 0.488]. In other words, B3LYP/aug-cc-pV(T+d)Z together with PCM benzene yields molecular dipole moments that are nearly "halfway between" calculated gas phase and PCM water dipole moments. It is worth noting that dipole moments for the IPolQ[15] and IPolQ-Mod[43] charge models—to be discussed in greater detail in the next section—would be obtained by assuming that β is exactly 1/2. Therefore, as we will argue under Discussion and Conclusions, the dipole moments—and charge distributions—obtained with these electronic structure methods are likely "in the neighborhood of correct" for use in condensed phase simulations.

Comparison with Different Density Functional Theory (DFT) Method/Basis Set Combination. Finally, we wanted to understand to what extent our results might depend on our choice of DFT method and basis set. To that end, we tested the method/basis set combination of PW6B95/def2-TZVPD. The PW6B95 functional is a hybrid meta-GGA functional that is considerably newer than B3LYP and, along with SOGGA11-X,[44] was recently found to have the most accurate gas phase dipole moment predictions of any non-double hybrid functional relative to CCSD(T)/CBS values for a benchmark set of 152 molecular species.[42] Indeed, the accuracy of PW6B95 exceeded that of restricted MP2 in this previous study. The def2-TZVPD augmented basis set differs meaningfully from aug-cc-pV(T+d)Z in terms of the philosophy used in its construction and is more minimally augmented. Together with a PCM model of benzene, this combination yields dipole moments that are 1.178 times larger (95% CI [1.157, 1.207]) than experimental gas phase (Figure S3). This is very similar to the B3LYP/aug-cc-pV(T+d)Z result of 1.182, and the same is true for the PCM water results as well. It is interesting to note that two quite different combinations of DFT method and basis set yield such similar predictions for the characteristic overpolarization. Moreover, the variation of data with respect to the regression line (i.e., the weighted residuals) and the fractional overpolarization for this combination with both PCM benzene and PCM water are quite similar to the corresponding B3LYP/aug-cc-pV(T+d)Z results (Figures S4 and S5). These results lend credence to the notion that many modern density functionals in combination with suitable augmented basis sets could provide a reasonable foundation for these calculations.

DISCUSSION AND CONCLUSIONS

In this work we examined four different electronic structure method/basis set combinations to determine which might provide a reasonable and accurate foundation for the derivation of fixed-charge electrostatic models to be used in molecular mechanics force fields. To do this we looked at both the strength and consistency of each combination's overpolarization. We found that the HF/6-31G* combination, the basis for the majority of the AMBER and GAFF biomolecular force fields, does indeed yield molecular dipole moments that are—on average—greater than experimental gas phase dipole moments, but that its predictions have a large variance and likely contain systematic errors with respect to the elements or

DOI: 10.1021/acs.jcim.9b00962
J. Chem. Inf. Model. 2020, 60, 249−258

moieties contained in a given molecule (e.g., species containing sulfur were strongly overpolarized, while pure hydrocarbon species were underpolarized). The use of a larger, more advanced augmented triple-ζ basis set with the HF method substantially improved the consistency of the method, but did not appreciably change the strength of the overpolarization. Used in concert with a PCM model of benzene, the combinations of two relatively simple density functionals (B3LYP and PW6B95) with two different augmented triple-ζ basis sets (aug-cc-pV(T+d)Z and def2-TZVPD) generated consistently overpolarized dipole moments that were—in an average sense—approximately 2 times more overpolarized relative to experimental gas phase data compared to those generated by HF/6-31G*. The use of PCM water increases the overpolarization of these method/basis set combinations even more (to approximately 4 times that of HF/6-31G*). As we will describe in more detail below, we believe that a similar sort of method/basis set combination and the use of implicit (or explicit) solvent—of either moderate dielectric constant or "averaged" between gas phase and high dielectric constant—is strongly justified for charge derivation in future fixed-charge MM force fields, and indeed many practitioners in the force field development community have adopted similar strategies.

The AMBER ff03 force field represents perhaps the first major biomolecular force field application of the more advanced electronic structure methods used in this work.[45] Partial atomic charges for ff03 are obtained by performing a restrained electrostatic potential (RESP) fit to molecular electrostatic potentials calculated using B3LYP/cc-pVTZ and an $\varepsilon = 4$ implicit solvent. Perhaps the main shortcoming to this approach is lack of tight-*d* and diffuse functions in this basis set, both of which are necessary for obtaining accurate dipole moments. Nonetheless, this work proved to be a forerunner of an approach that has been adopted repeatedly since it was published. For example, the use of DFT methods with triple-ζ basis sets (or their plane-wave equivalent) and $\varepsilon = 4$ implicit solvent has been demonstrated in two recent works regarding training machine-learning models for partial atomic charges[46] and the derivation of partial atomic charges (and Lennard-Jones coefficients) using atoms-in-molecules (AIM) partitioning.[16] In the latter work, the authors argue that $\varepsilon = 4$ implicit solvent is not only a good "average" dielectric environment, but also represents a near-ideal dielectric constant for the cancellation of electronic distortion and polarization energies, thereby facilitating comparisons between free energies obtained from fixed-charge MD simulations and experiments.[16]

Taking into account the electronic distortion cost was also the motivator for the IPolQ charge derivation method, in which charges are computed by averaging the partial atomic charges obtained from RESP fits performed in gas phase and with explicit water molecules present using the combination of MP2/cc-pV(T+d)Z.[15] As demonstrated under Results, molecular dipole moments obtained using PCM benzene are likely a reasonable approximation to IPolQ-like approaches (e.g., IPolQ-Mod) that utilize implicit aqueous solvent.[43] It is worth noting, however, that these approaches may underpolarize some molecules by not accurately capturing effects such as hydrogen bonding with water that would be captured in the original IPolQ approach utilizing explicit solvent. A possible solution and variant on this idea has recently been proposed by Schauperl et al., who suggest allowing the linear combination of partial atomic charges obtained in gas phase and PCM water to be tuned to reproduce experimental data rather than having the coefficients fixed at 1/2 as in the IPolQ and IPolQ-Mod models.[47] In addition, this work demonstrates that gas phase molecular ESPs and dipole moments computed using PW6B95/aug-cc-pV(D+d)Z—a somewhat smaller double-ζ basis set than those used with the DFT methods in the present work—are accurate compared to theoretical reference calculations performed using the method/basis set combination of DSD-PBEP86-D3BJ[48,49]/aug-cc-pV(Q+d)Z. For the interested reader, we have performed a comparison of the method/basis set combinations used in this work to the same theoretical reference data in the Supporting Information (Figures S6 and S7). This comparison largely recapitulates both the qualitative and quantitative trends observed when the comparison is made to gas phase experimental data.

It is important to note that the use of DFT methods, larger basis sets, and PCM solvent does result in a greater computational cost. For *N*-methylacetamide, a molecule that is fairly typical of our benchmark set, the one-electron properties calculations in gas phase using B3LYP/aug-cc-pV(T+d)Z and PW6B95/def2-TZVPD are approximately 22 and 18 times more expensive than HF/6-31G*. The costs go up by approximately an order of magnitude with the use of implicit solvent, but the incorporation of modern domain decomposition approaches for implicit solvation[50] into Psi4 should reduce the cost of implicit solvent calculations to the same order of magnitude as the gas phase calculations. Nonetheless, the development of fast charge assignment methods (i.e., AM1-BCC-type approaches) that have been trained on these higher quality electronic structure methods would also be a useful contribution.[46]

In summary, our work leads us to recommend that ab initio charge derivation methods for future MM force fields be based on electronic structure calculations performed using DFT or other affordable post-HF methods, augmented basis sets of good quality, and implicit or explicit solvent that will systematically and reliably overpolarize the molecular charge distribution. It is important to note that any change of the underlying electrostatic model in a force field will require the reparametrization of the parameters governing van der Waals interactions (e.g., Lennard-Jones parameters). We believe, however, that such efforts will likely be worthwhile in terms of yielding nonbonded parameters that are more transferable because they will not have to compensate for moiety-specific systematic errors in the electrostatic description caused by the inconsistent overpolarization predictions of the sort that we have documented in this work. We hope to further explore the impact of different charge/electrostatic model derivation methods on overall force field accuracy in future work.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.9b00962.

Comparison of gas phase B3LYP/aug-cc-pV(T+d)Z dipole moments to experimental gas phase dipole moments; exact dipole moment residuals for all tested method/basis set combinations; comparison of PW6B95/def2-TZVPD dipole moments with PCM benzene and water to experimental gas phase dipole moments (including weighted residuals and fractional dipole moment enhancements); comparison of all tested method/basis set combination dipole moments to DSD-

DOI: 10.1021/acs.jcim.9b00962
J. Chem. Inf. Model. 2020, 60, 249−258

PBEP86-D3BJ/aug-cc-pV(Q+d)Z dipole moments (PDF)

Calculated and experimental dipole moments for all 47 molecules (CSV)

Optimized geometries for all 47 molecules in *xyz* format (ZIP)

AUTHOR INFORMATION

Corresponding Author

*E-mail: pnerenb@calstatela.edu.

ORCID

Michael Schauperl: 0000-0001-5648-8170

Paul S. Nerenberg: 0000-0002-9730-6983

Author Contributions

The work described in this manuscript was performed by all of the authors. The manuscript was written through the contributions of all authors.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

A.Z. and P.S.N. acknowledge the support of NASA Minority University Research and Education Project (MUREP) Institutional Research Opportunity Grant NNX15AQ06A. M.S. acknowledges support from the Austrian Science Fund (Erwin Schroedinger Fellowship J-4150).

REFERENCES

(1) Riniker, S. Fixed-Charge Atomistic Force Fields for Molecular Dynamics Simulations in the Condensed Phase: An Overview. *J. Chem. Inf. Model.* **2018**, *58*, 565−578.

(2) Nerenberg, P. S.; Head-Gordon, T. New Developments in Force Fields for Biomolecular Simulations. *Curr. Opin. Struct. Biol.* **2018**, *49*, 129−138.

(3) McCammon, J. A.; Gelin, B. R.; Karplus, M. Dynamics of Folded Proteins. *Nature* **1977**, *267*, 585−590.

(4) Hollingsworth, S. A.; Dror, R. O. Molecular Dynamics Simulation for All. *Neuron* **2018**, *99*, 1129−1143.

(5) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179−5197.

(6) Wang, J. M.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. a; Case, D. a. Development and Testing of a General Amber Force Field. *J. Comput. Chem.* **2004**, *25*, 1157−1174.

(7) Bayly, C. I.; Cieplak, P.; Cornell, W.; Kollman, P. A. A Well-Behaved Electrostatic Potential Based Method Using Charge Restraints for Deriving Atomic Charges: The RESP Model. *J. Phys. Chem.* **1993**, *97*, 10269−10280.

(8) Hehre, W. J.; Ditchfield, K.; Pople, J. A. Self-Consistent Molecular Orbital Methods. XII. Further Extensions of Gaussian-Type Basis Sets for Use in Molecular Orbital Studies of Organic Molecules. *J. Chem. Phys.* **1972**, *56*, 2257−2261.

(9) Hariharan, P. C.; Pople, J. A. The Influence of Polarization Functions on Molecular Orbital Hydrogenation Energies. *Theor. Chim. Acta* **1973**, *28*, 213−222.

(10) Francl, M. M.; Pietro, W. J.; Hehre, W. J.; Binkley, J. S.; Gordon, M. S.; DeFrees, D. J.; Pople, J. A. Self-Consistent Molecular Orbital Methods. XXIII. A Polarization-Type Basis Set for Second-Row Elements. *J. Chem. Phys.* **1982**, *77*, 3654−3665.

(11) Jakalian, A.; Bush, B. L.; Jack, D. B.; Bayly, C. I. Fast, Efficient Generation of High-Quality Atomic Charges. AM1-BCC Model: I. Method. *J. Comput. Chem.* **2000**, *21*, 132−146.

(12) Jakalian, A.; Jack, D. B.; Bayly, C. I. Fast, Efficient Generation of High-Quality Atomic Charges. AM1-BCC Model: II. Parameterization and Validation. *J. Comput. Chem.* **2002**, *23*, 1623−1641.

(13) Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. The Missing Term in Effective Pair Potentials. *J. Phys. Chem.* **1987**, *91*, 6269−6271.

(14) Karamertzanis, P. G.; Raiteri, P.; Galindo, A. The Use of Anisotropic Potentials in Modeling Water and Free Energies of Hydration. *J. Chem. Theory Comput.* **2010**, *6*, 1590−1607.

(15) Cerutti, D. S.; Rice, J. E.; Swope, W. C.; Case, D. A. Derivation of Fixed Partial Charges for Amino Acids Accommodating a Specific Water Model and Implicit Polarization. *J. Phys. Chem. B* **2013**, *117*, 2328−2338.

(16) Cole, D. J.; Vilseck, J. Z.; Tirado-Rives, J.; Payne, M. C.; Jorgensen, W. L. Biomolecular Force Field Parameterization via Atoms-in-Molecule Electron Density Partitioning. *J. Chem. Theory Comput.* **2016**, *12*, 2312−2323.

(17) Hickey, A. L.; Rowley, C. N. Benchmarking Quantum Chemical Methods for the Calculation of Molecular Dipole Moments and Polarizabilities. *J. Phys. Chem. A* **2014**, *118*, 3678−3687.

(18) *CRC Handbook of Chemistry and Physics*, 98th ed.; Haynes, W. M., Ed.; CRC Press: Boca Raton, FL, 2018.

(19) Thorne, L. R.; Suenram, R. D.; Lovas, F. J. Microwave Spectrum, Torsional Barrier, and Structure of BH3NH3. *J. Chem. Phys.* **1983**, *78*, 167−171.

(20) Johnson, R. D., III. NIST Computational Chemistry Comparison and Benchmark Database, NIST Standard Reference Database Number 101, Release 20, August 2019. http://cccbdb.nist.gov.

(21) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An Open Chemical Toolbox. *J. Cheminf.* **2011**, *3*, 33.

(22) Hanwell, M. D.; Curtis, D. E.; Lonie, D. C.; Vandermeersch, T.; Zurek, E.; Hutchison, G. R. Avogadro: An Advanced Semantic Chemical Editor, Visualization, and Analysis Platform. *J. Cheminf.* **2012**, *4*, 17.

(23) Parrish, R. M.; Burns, L. A.; Smith, D. G. A.; Simmonett, A. C.; DePrince, A. E.; Hohenstein, E. G.; Bozkaya, U.; Sokolov, A. Y.; Di Remigio, R.; Richard, R. M.; Gonthier, J. F.; James, A. M.; McAlexander, H. R.; Kumar, A.; Saitow, M.; Wang, X.; Pritchard, B. P.; Verma, P.; Schaefer, H. F.; Patkowski, K.; King, R. A.; Valeev, E. F.; Evangelista, F. A.; Turney, J. M.; Crawford, T. D.; Sherrill, C. D. Psi4 1.1: An Open-Source Electronic Structure Program Emphasizing Automation, Advanced Libraries, and Interoperability. *J. Chem. Theory Comput.* **2017**, *13*, 3185−3197.

(24) Becke, A. D. A New Mixing of Hartree-Fock and Local Density-Functional Theories. *J. Chem. Phys.* **1993**, *98*, 1372−1377.

(25) Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti Correlation-Energy Formula into a Functional of the Electron Density. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1988**, *37*, 785−789.

(26) Vosko, S. H.; Wilk, L.; Nusair, M. Accurate Spin-Dependent Electron Liquid Correlation Energies for Local Spin Density Calculations: A Critical Analysis. *Can. J. Phys.* **1980**, *58*, 1200−1211.

(27) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. Ab Initio Calculation of Vibrational Absorption and Circular Dichroism Spectra Using Density Functional Force Fields. *J. Phys. Chem.* **1994**, *98*, 11623−11627.

(28) Dunning, T. H. Gaussian Basis Sets for Use in Correlated Molecular Calculations. I. The Atoms Boron through Neon and Hydrogen. *J. Chem. Phys.* **1989**, *90*, 1007−1023.

(29) Woon, D. E.; Dunning, T. H. Gaussian Basis Sets for Use in Correlated Molecular Calculations. III. The Atoms Aluminum through Argon. *J. Chem. Phys.* **1993**, *98*, 1358−1371.

(30) Dunning, J.; Peterson, K. A.; Wilson, A. K. Gaussian Basis Sets for Use in Correlated Molecular Calculations. X. The Atoms Aluminum through Argon Revisited. *J. Chem. Phys.* **2001**, *114*, 9244−9253.

DOI: 10.1021/acs.jcim.9b00962
J. Chem. Inf. Model. 2020, 60, 249−258

(31) Zhao, Y.; Truhlar, D. G. Design of Density Functionals That Are Broadly Accurate for Thermochemistry, Thermochemical Kinetics, and Nonbonded Interactions. *J. Phys. Chem. A* **2005**, *109*, 5656−5667.
(32) Weigend, F.; Ahlrichs, R. Balanced Basis Sets of Split Valence, Triple Zeta Valence and Quadruple Zeta Valence Quality for H to Rn: Design and Assessment of Accuracy. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297−3305.
(33) Rappoport, D.; Furche, F. Property-Optimized Gaussian Basis Sets for Molecular Response Calculations. *J. Chem. Phys.* **2010**, *133*, 134105.
(34) Cancès, E.; Mennucci, B. New Applications of Integral Equations Methods for Solvation Continuum Models: Ionic Solutions and Liquid Crystals. *J. Math. Chem.* **1998**, *23*, 309−326.
(35) Barone, V.; Cossi, M. Quantum Calculation of Molecular Energies and Energy Gradients in Solution by a Conductor Solvent Model. *J. Phys. Chem. A* **1998**, *102*, 1995−2001.
(36) Tomasi, J.; Mennucci, B.; Cammi, R. Quantum Mechanical Continuum Solvation Models. *Chem. Rev.* **2005**, *105*, 2999−3093.
(37) Rappé, A. K.; Casewit, C. J.; Colwell, K. S.; Goddard, W. A.; Skiff, W. M. UFF, a Full Periodic Table Force Field for Molecular Mechanics and Molecular Dynamics Simulations. *J. Am. Chem. Soc.* **1992**, *114*, 10024−10035.
(38) Di Remigio, R.; Steindal, A. H.; Mozgawa, K.; Weijo, V.; Cao, H.; Frediani, L. PCMSolver: An Open-Source Library for Solvation Modeling. *Int. J. Quantum Chem.* **2019**, *119*, e25685.
(39) Theil, H. A Rank-Invariant Method of Linear and Polynomial Regression Analysis: Part I. *Proc. R. Neth. Acad. Sci.* **1950**, *53*, 386−392.
(40) Sen, P. K. Estimates of the Regression Coefficient Based on Kendall's Tau. *J. Am. Stat. Assoc.* **1968**, *63*, 1379−1389.
(41) Jones, E.; Oliphant, T.; Peterson, P. *SciPy: Open Source Scientific Tools for Python*; 2019. http://www.scipy.org/.
(42) Hait, D.; Head-Gordon, M. How Accurate Is Density Functional Theory at Predicting Dipole Moments? An Assessment Using a New Database of 200 Benchmark Values. *J. Chem. Theory Comput.* **2018**, *14*, 1969−1981.
(43) Mecklenfeld, A.; Raabe, G. Comparison of RESP and IPolQ-Mod Partial Charges for Solvation Free Energy Calculations of Various Solute/Solvent Pairs. *J. Chem. Theory Comput.* **2017**, *13*, 6266−6274.
(44) Peverati, R.; Truhlar, D. G. Communication: A Global Hybrid Generalized Gradient Approximation to the Exchange-Correlation Functional That Satisfies the Second-Order Density-Gradient Constraint and Has Broad Applicability in Chemistry. *J. Chem. Phys.* **2011**, *135*, 191102.
(45) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J.; Kollman, P. A Point-Charge Force Field for Molecular Mechanics Simulations of Proteins Based on Condensed-Phase Quantum Mechanical Calculations. *J. Comput. Chem.* **2003**, *24*, 1999−2012.
(46) Bleiziffer, P.; Schaller, K.; Riniker, S. Machine Learning of Partial Charges Derived from High-Quality Quantum-Mechanical Calculations. *J. Chem. Inf. Model.* **2018**, *58*, 579−590.
(47) Schauperl, M.; Nerenberg, P. S.; Jang, H.; Wang, L.-P.; Bayly, C. I.; Mobley, D. L.; Gilson, M. K. Force Field Partial Charges with Restrained Electrostatic Potential 2 (RESP2). 2019, *ChemRxiv*. Preprint. DOI: 10.26434/chemrxiv.10072799.v1.
(48) Kozuch, S.; Martin, J. M. L. DSD-PBEP86: In Search of the Best Double-Hybrid DFT with Spin-Component Scaled MP2 and Dispersion Corrections. *Phys. Chem. Chem. Phys.* **2011**, *13*, 20104−20107.
(49) Kozuch, S.; Martin, J. M. L. Spin-Component-Scaled Double Hybrids: An Extensive Search for the Best Fifth-Rung Functionals Blending DFT and Perturbation Theory. *J. Comput. Chem.* **2013**, *34*, 2327−2344.
(50) Stamm, B.; Lagardère, L.; Scalmani, G.; Gatto, P.; Cancès, E.; Piquemal, J. P.; Maday, Y.; Mennucci, B.; Lipparini, F. How to Make Continuum Solvation Incredibly Fast in a Few Simple Steps: A Practical Guide to the Domain Decomposition Paradigm for the Conductor-like Screening Model. *Int. J. Quantum Chem.* **2019**, *119*, e25669.

DOI: 10.1021/acs.jcim.9b00962
J. Chem. Inf. Model. 2020, 60, 249−258