JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

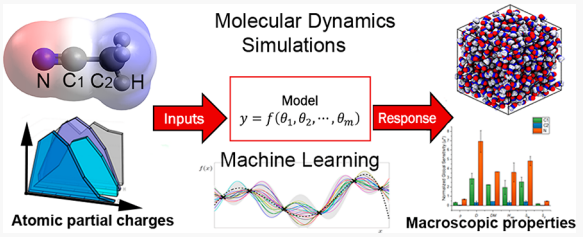
Uncertainty Quantification and Sensitivity Analysis of Partial Charges on Macroscopic Solvent Properties in Molecular Dynamics Simulations with a Machine Learning Model

James S. Peerless, Albert L. Kwansa, Branden S. Hawkins, Ralph C. Smith, and Yaroslava G. Yingling*

Cite This: J. Chem. Inf. Model. 2021, 61, 1745−1761 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: The molecular dynamics (MD) simulation technique is among the most broadly used computational methods to investigate atomistic phenomena in a variety of chemical and biological systems. One of the most common (and most uncertain) parametrization steps in MD simulations of soft materials is the assignment of partial charges to atoms. Here, we apply uncertainty quantification and sensitivity analysis calculations to assess the uncertainty associated with partial charge assignment in the context of MD simulations of an organic solvent. Our results indicate that the effect of partial charge variance on bulk properties, such as solubility parameters, diffusivity, dipole moment, and density, measured from MD simulations is significant; however, measured properties are observed to be less sensitive to partial charges of less accessible (or buried) atoms. Diffusivity, for example, exhibits a global sensitivity of up to 22×10^{-5} cm^2/s per electron charge on some acetonitrile atoms. We then demonstrate that machine learning techniques, such as Gaussian process regression (GPR), can be effective and rapid tools for uncertainty quantification of MD simulations. We show that the formulation and application of an efficient GPR surrogate model for the prediction of responses effectively reduces the computational time of additional sample points from hours to milliseconds. This study provides a much-needed context for the effect that partial charge uncertainty has on MD-derived material properties to illustrate the benefit of considering partial charges as distributions rather than point-values. To aid in this treatment, this work then demonstrates methods for rapid characterization of resulting sensitivity in MD simulations.

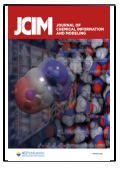


1. INTRODUCTION

The molecular dynamics (MD) method is one of the most commonly applied simulation techniques to achieve atomistic-scale assessment of chemical and biological systems. Classical MD allows for simulations of systems on the order of ten million atoms and up to microsecond time scales,[1] several orders of magnitude beyond any quantum-mechanical simulation technique. State-of-the-art MD software leverages advanced graphics processing unit (GPU) architectures for efficient parallelization of calculations making larger and longer all-atom MD simulations less computationally daunting.[2,3] Moreover, MD software, both open-source and proprietary packages alike, continues to become more well-documented and user-friendly, opening up the possibilities of MD simulation applications to a variety of problems by a wider scientific audience. However, the assumptions made in the formulation and parametrization of classical MD experiments require considerable due diligence from the user. Indeed, the basis for the classical MD method assumes Newtonian physical relations are adequate for reproducing the phenomena of interest. Further, the adoption of a chosen force field (the function from which all interatomic forces are derived) involves intrinsic assumptions, such as the representation of chemical bonds as harmonic springs along with the assignment of force constants (e.g., estimated by normal-mode analysis of vibrational spectroscopic data[4]). As of yet, there is no one-size-fits-all universal force field; MD users must judiciously choose a force field (or method of developing their own) that can reliably reproduce the structure or physicochemical property in which they are interested.

Typically, atomistic MD force field parameters are implemented as scalar values with no associated uncertainty. For example, in classical AMBER/CHARMM-type force fields, the potential interatomic energy is calculated as a sum of bonded and nonbonded terms. Bonded terms include parameters for equilibrium bond lengths, angles, and dihedral angles with associated force constants. Nonbonded terms usually include a pairwise calculation for van der Waals energy, often represented by a Lennard-Jones function, and electrostatic energy which is dependent on the atomic partial charges. Whereas all other parameters of a given force field typically

Received: October 15, 2020
Published: March 17, 2021



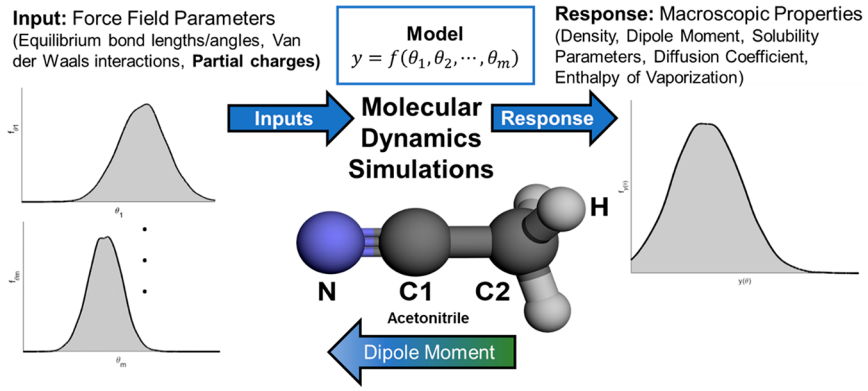


Figure 1. Diagram of process for uncertainty quantification and acetonitrile molecule with atom labels used in this work.

depend on the atom type alone (e.g., the same carbon−carbon bond distance is applied to all bonded sp^3 carbon pairs), the determination of partial charges is molecule-specific and is often a considerable portion of force field parametrization since partial charges are highly dependent on the surrounding chemical environment of the atom in question along with other factors such as molecular conformation and orientation. Moreover, the use of set point charges in classical MD introduces further uncertainty into the assignment of atomic partial charges.

Calculations of atomic partial charges typically involve two steps: a quantum-mechanical (QM) calculation to determine the molecular electrostatic potential (MEP) and an algorithm to collapse the MEP onto atomic point charges. The QM calculation can be performed at various levels of theory, such as Hartree−Fock (HF) or Møller−Plesset second-order perturbation (MP2), with a selected QM basis set. Conversion from the high-fidelity MEP to atomic charges is usually a less involved calculation but not without its own set of decisions.[5] A common method is the restrained electrostatic potential (RESP) fitting of atomic charges which optimizes point charges to reproduce the MEP while restraining "buried" (i.e., less solvent-accessible) atoms to low charge. This is done because the charges of such atoms (e.g., methyl carbon) are shown to be poorly determined when fitting to the MEP; otherwise stated, the reproduction of the MEP is less sensitive to the charges of buried atoms.[6−8] The RESP method is usually employed in conjunction with the AMBER family of force fields investigated in this study[9] but is by no means the only method in wide adoption.[5] However, it is known that the choice of QM and fitting algorithm will directly affect the macroscopic properties measured from MD simulation, such as diffusion.[8,10,11] Moreover, many software and standard approaches provide little guidance to reasoning for careful parametrization and fitting of partial charges to macroscopic properties (apart from density).

With the variation of the calculation of partial charge parameters, it is critical to understand (and ideally, quantify) the error that this variation may impart on calculated macroscopic properties. Uncertainty quantification (UQ) is the mathematical discipline of characterizing the uncertainties associated with a given model prediction.[12] At the root of UQ studies focused on a specific model, $y = f(x, q)$, is often the assumption that the parameters contained in the vector q are random variables with an associated distribution. The parameter uncertainty is then propagated through the model, f, to quantify the uncertainty associated with the model response, y. Techniques for uncertainty quantification allow not only estimates of the overall uncertainty associated with y, but also rigorous sensitivity analysis to attribute uncertainty to specific parameters. UQ experiments often use this treatment for efficient parametrization and parameter selection to improve model accuracy as well.[12]

There have been significant efforts to quantify response uncertainty within MD formalism as well as optimize parameters in force fields dating back to the 1990s. Wong and Rabitz performed a preliminary sensitivity analysis and principal component analysis (PCA) for free energy calculations and identified the most sensitive parameters of the potential energy function in amino acid simulations.[13] Zhu and Wong performed sensitivity analysis and parameter optimization on popular polarizable and nonpolarizable water models used in MD, suggesting new force field parameters via UQ.[14,15] Work by Frederiksen and co-workers presented an approach employing Bayesian ensembles to quantify uncertainty associated with metallic interatomic potential parameters, ultimately producing force field parameters for molybdenum with error estimates for simulated mechanical properties.[16] Cailliez and Pernot provided a rigorous framework for force field parametrization and uncertainty prediction with argon simulations.[17] Similarly, Angelikopoulos et al. put forth another method for parameter prediction and error estimation using efficient Bayesian uncertainty quantification, also demonstrated on argon parameters[18] followed by application to nanoscale flow simulations.[19] Rizzi and co-workers performed extremely detailed uncertainty quantification for error propagation[20] and parameter estimation[21] for the TIP4P water model, as well as parameters in ionic flow simulations[22,23] and diffusion in metallic alloys.[24] The work on parameter estimation for TIP4P water was further improved upon by Jacobson et al. for multiple macroscopic properties.[25] Significant work has also been performed on the uncertainty quantification of specific measurements from MD, such as the time-dependent diffusion coefficient,[26] glass transition temperature,[27] shear viscosity,[28] or solvation energies.[29] Additional research studies have sought to perform UQ and optimization of force field parameters for alkanes,[30] graphene,[31] and reactive[32] force fields. Furthermore, studies have applied principles of UQ and Bayesian statistics to rapid and automated force field parametrization[33] or potential energy calculation within MD itself.[34] However, there is still a lack of UQ studies pertaining specifically to the uncertainty of partial charge assignment despite the lack of a universal protocol and the well-known effect that different partial charges can have on the results of MD simulations.[8,10,11]

https://doi.org/10.1021/acs.jcim.0c01204
J. Chem. Inf. Model. 2021, 61, 1745−1761

In this work, we focus only on the uncertainty associated with partial charge assignment, as this is most often encountered by MD users even when using well-established force fields. By assuming an input distribution of partial charges for a common organic solvent, we performed hundreds of short simulations to calculate macroscopic properties. We quantified the uncertainty associated with these properties as a result of partial charge uncertainty and additionally provide local and global sensitivity estimates for each atom type. We also present the results of the formulation of a surrogate model to predict MD responses from partial charge assignment using Gaussian process regression (GPR).[35,36] The resulting GPR surrogate reduces the required simulation time from days to seconds, allowing further exploration of the parameter space. Our results highlight the need for careful assignment of partial charges in liquid-phase MD simulations yet also provide specific guidelines on the responses that are more likely to be highly sensitive to charge assignment technique.

2. METHODS

The following summarizes the simulation, analysis, and statistical methods employed in this work. Additional information on the construction of parameter distributions used in this study is provided as well.

2.1. Simulation Methods. In this study, the model of interest is molecular dynamics of the common organic solvent acetonitrile (ACN). This solvent has a few benefits that make it advantageous for this study. First of all, ACN has only four symmetrically independent atoms, henceforth referred to as C1, C2, N, and H (see Figure 1). Given that all partial charges must sum to zero (i.e., $q_{C1} + q_{C2} + q_N + 3q_H = 0$) for an uncharged molecule, we have three degrees of freedom in the atomic partial charges. This limits the dimensionality of the parameter space thus reducing computational load exponentially.

Additionally, the rigidity and simple geometry of the molecule reduce the amount of variation in bulk properties due to angle and dihedral energies. Although other force field parameters (i.e., bonded and nonbonded parameters other than partial charges) will undoubtedly effect responses, the angles and dihedrals in ACN are relatively well-understood and noncompetitive. The polarity of the molecule provides an interesting opportunity as well, as the simple reduction of partial charge character to the overall dipole moment provides a directly correlated response that may test our hypotheses. A more complicated dipole−charge relationship or a molecule with a very weak intrinsic dipole would not allow this straightforward calculation as a sanity check of our analysis and statistical manipulations. The existence of a buried atom, C2, also allows the investigation of partial charges of this type on bulk properties. Typically, buried atomic charges carry the highest uncertainty when point charges are fitted from *ab initio*-calculated ESP data.

Finally, the ubiquity of ACN as a solvent in organic synthesis makes it the subject of various *ab initio* and molecular dynamics studies in which partial charges were calculated, dating back to 1982.[37−42] This allows for rich source data from which to base our assumptions as to the uncertainty associated with the partial charge parameters. Details on these sources and how they were used to formulate parameter distributions are discussed in section 2.3.

All simulations were performed using the AMBER 16 molecular dynamics package.[43] All simulation responses were calculated from 50 ns constant-temperature, constant-pressure (NPT) simulations of 693 ACN molecules at 300 K and 1 atm with a 2 fs time step and a cutoff of 8 Å. The Berendsen/weak-coupling thermostat and Berendsen barostat (τ_T = 1.0 ps, τ_P = 1.0 ps) were used for temperature and pressure regulation, respectively. Periodic boundary conditions were applied in all directions, and the particle-mesh Ewald (PME) algorithm was applied for long-range electrostatic interactions. The SHAKE algorithm was used to restrain bonds with hydrogen and allow the 2 fs time step. All simulations used the general AMBER force field (GAFF)[9] for all bonded and Lennard-Jones parameters.

A simulation size of 693 molecules corresponds to 4158 atoms in a cubic box of approximately 40 Å for each dimension. The choice of 50 ns was made based on accurate calculation of the self-diffusion coefficient (D). The calculation of D is discussed in detail in section 2.2.2 and its effect on simulation size is further explored in section S1 of the Supporting Information (SI).

To serve as initial coordinates for subsequent simulations, an initial seed simulation was constructed from 693 energy-minimized ACN molecules in a cubic box with partial charges calculated from the R.E.D. Server Development[44] restrained electrostatic potential (RESP) method (discussed further in section 2.3). This box was then energy minimized, heated to 300 K under constant-temperature, constant-volume (NVT) conditions at a rate of 3 K/ps, equilibrated at 300 K under NVT for 0.3 ns, NPT-equilibrated at 1 atm and 300 K for 0.5 ns, and NPT-equilibrated at 1 atm and 300 K with the SHAKE algorithm applied for 1 ns. Finally, this system underwent a 50 ns production simulation as described above. For all the simulations used thereafter, a randomly chosen frame from this seed simulation was used as initial coordinates. The partial charges were modified as needed, and a 1 ns equilibration run (NPT, 1 atm, 300 K) was performed prior to the 50 ns data-collection run from which properties are calculated.

2.2. Quantities of Interest. For quantities of interest to serve as responses (y) in the study, six well-known macroscopic (or bulk) solvent properties that can be readily calculated from single-component simulations of this size were chosen: density (ρ), self-diffusion coefficient (D), molecular dipole moment (DM), enthalpy of vaporization (H_{vap}), and the dispersion (δ_d) and polar-plus-hydrogen-bonding (δ_{h+p}) Hansen solubility parameters (HSPs). Reference values for these bulk properties of ACN are shown below in Table 1 along with their symbols and the units used in this work.

The HSPs are often used in solubility calculations and represent the cohesive energy density due to dispersion (or van der Waals) forces and electrostatic forces, respectively.[49] Although typically three parameters (δ_d, δ_p, δ_h), the polar

Table 1. Symbols, Reference Values, and Units for the Bulk Properties Calculated from Simulations in This Work[a]

y	description	ref value (y^0)	units	ref
ρ	density (20 °C)	0.786	g/cm^3	45
D	self-diffusion coefficient	3.38	10^{-5} cm^2/s	46
DM	dipole moment	3.92	Debye	47
H_{vap}	enthalpy of vaporization	7.89	kcal/mol	48
δ_d	dispersion HSP	15.3	MPa$^{1/2}$	49
δ_{h+p}	polar + H-bond HSP	19.0	MPa$^{1/2}$	49

[a]Unless noted otherwise, all properties measured at 25 °C and 1 atm.

https://doi.org/10.1021/acs.jcim.0c01204
J. Chem. Inf. Model. 2021, 61, 1745−1761

and hydrogen-bonding terms cannot be effectively separated from simulation measurements and are thus combined ($\delta_{h+p}^2 = \delta_h^2 + \delta_p^2$). The HSPs are related to the Hildebrand solubility parameter (δ_T) and H_{vap} by

$$\delta_d^2 + \delta_p^2 + \delta_h^2 = \delta_d^2 + \delta_{h+p}^2 = \delta_T^2 = \frac{H_{vap} - RT}{V_m}$$

where R is the molar gas constant, T is temperature, and V_m is the molar volume. Hence, the HSPs are directly related to the enthalpy of vaporization at a given density and temperature.

Bulk properties were derived from the all-atom molecular dynamics simulation trajectories via an in-house computer script that employs AmberTools17's[50] cpptraj trajectory analysis program.[51] Details of the calculation of each property are described below.

2.2.1. Density. Density (ρ) was calculated as the total mass of all molecules divided by the system volume:

$$\rho = \frac{1}{V}\sum_{i=1}^{N_{molecules}} \frac{MW}{N_A}$$

where V = volume, MW = molar mass, and N_A = Avogadro's constant.

2.2.2. Self-Diffusion Coefficient. The self-diffusion coefficient (D) was derived from the linear regression of mean-square displacement (MSD) vs time data. If Y = MSD and X = time, then the linear fit of $Y(X)$ is represented by

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

where X is an independent variable, $\hat{Y}$ is the estimator of MSD(t), $\hat{\beta}_0$ is the estimator of the y-intercept, and $\hat{\beta}_1$ is the estimator of the slope.[52] The self-diffusion coefficient (D) is estimated by[53]

$$\hat{D} = \hat{\beta}_1/(2n)$$

and the associated standard error (SE) was calculated as

$$SE_{\hat{D}} = SE_{\hat{\beta}_1}/(2n)$$

where n = number of spatial dimensions (e.g., 3, in this case), the SE of the slope estimator $(SE_{\hat{\beta}_1}) = S_{Y|X}/(SD_X\sqrt{N-1})$, the root-mean-square error $(S_{Y|X}) = \sqrt{\frac{1}{N-2}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2}$, the sample standard deviation of X $(SD_X) = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2}$, and N = sample size.[52] Furthermore, block averaging was used to average estimates of D from multiple time blocks across the trajectory rather than calculating a single estimate of D from the entire trajectory. The mean estimate of D ($\bar{D}$), mean SE of D ($\overline{SE}_D$), and SD of D (SD_D) were calculated as

$$\bar{D} = \frac{1}{N_{blocks}}\sum_{i=1}^{N_{blocks}} D_i$$

$$\overline{SE}_D = \frac{1}{N_{blocks}}\sqrt{\sum_{i=1}^{N_{blocks}} SE_{\hat{D}_i}^2}$$

$$SD_D = \sqrt{\frac{1}{N_{blocks}-1}\sum_{i=1}^{N_{blocks}}(D_i - \bar{D})^2}$$

This method of averaging values from multiple blocks of a simulation is akin to the recommended practice of measuring diffusivity from multiple independent simulations (MIS) provided the relaxation time of the system is sufficiently short.[54] In these calculations, N_{blocks} = 10 for each 50 ns simulation. Details regarding why this value for N_{blocks} was chosen for the accurate representation of self-diffusivity is discussed in section S1 of the Supporting Information.

2.2.3. Molecular Dipole Moment. The molecular dipole moment (DM) was calculated as the summation over all atoms of the product of an atom's partial charge and the atom's position vector shifted to the center of mass of the molecule:

$$\mu = |\vec{\mu}| = |\sum_{i=1}^{N_{atoms}} q_i(\vec{r_i} - \vec{r_c})|$$

where q_i = charge of atom i, $\vec{r_i}$ = position vector of atom i, $\vec{r_c}$ = position vector of the molecule's center of mass (based on eq 7 from the work of Buckingham[55]).

2.2.4. Enthalpy of Vaporization. The enthalpy of vaporization (H_{vap}) was derived from the total intermolecular nonbonded energy as follows:

$$H_{vap} = (E_{LJ} + E_{Coulombic})/N_{molecules} + RT$$

where E_{LJ} = intermolecular Lennard-Jones energy (van der Waals attraction + Pauli exclusion), $E_{Coulombic}$ = intermolecular Coulombic energy, $N_{molecules}$ = number of molecules, R = molar gas constant, and T = temperature (based on eq 7 from the work of Wang and Hou[45]). More rigorously, this would be symbolized as ΔH_{vap}, as it is a relative rather than an absolute measurement, but we employ the simpler H_{vap} notation in this work.

2.2.5. Hansen Solubility Parameters. Hansen solubility parameters (HSPs) were predicted in representation of the dispersion HSP (δ_d) and another HSP (δ_{h+p}) representing a combination of the H-bond HSP (δ_h) and the polar HSP (δ_p).

The quantities δ_d and δ_{h+p} were derived from the intermolecular Lennard-Jones energy and the intermolecular Coulombic energy, respectively, as

$$\delta_d = \sqrt{E_{LJ}/(VN_A)}$$

$$\delta_{h+p} = \sqrt{E_{Coulombic}/(VN_A)}$$

where E_{LJ} = intermolecular Lennard-Jones energy, $E_{Coulombic}$ = intermolecular Coulombic energy, V = volume, and N_A = Avogadro's constant (based on eq 4 from the work of Belmares et al.[56]). The algorithm to obtain the intermolecular energies for the bulk solvent systems involved looping over each solvent molecule in the system, calculating the intermolecular energy quantities (Lennard-Jones and Coulombic) between that solvent molecule and all other solvent molecules, summing the intermolecular energy values obtained after looping over all molecules, and then dividing the total energy values by two to correct for the duplicate interactions included in the summation.

2.3. Reference Parameters. In order to quantify the uncertainty of measured responses associated with partial charge parameters, it is necessary to infer the distribution of these parameters to be treated as random variables. Although estimates of uncertainty associated with *ab initio* calculations exist,[57] it is more beneficial from an MD practitioner's standpoint to take into account the overall uncertainty

https://doi.org/10.1021/acs.jcim.0c01204
J. Chem. Inf. Model. 2021, 61, 1745−1761

Table 2. Partial Charges and Their Sources for ACN Used to Infer Parameter Distributions[a]

source	q_N	q_{C1}	q_{C2}	q_H
Cabaleiro-Lago (HF/6-311+G*)[38]	−0.532	0.481	−0.479	0.177
Cabaleiro-Lago (MP2/6-311+G*)[38]	−0.494	0.475	−0.552	0.190
Grabuleda (HF/6-31G*)[39]	−0.490	0.382	−0.2376	0.115
Grabuleda (HF/6-311+G*)[39]	−0.532	0.481	−0.479	0.177
Nikitin (MP2/6-311++G(3df,3p))[40]	−0.5126	0.4917	−0.5503	0.1904
Caleman (OPLS)[41]	−0.56	0.46	−0.08	0.06
Caleman (GAFF)[41]	−0.5168	0.4484	−0.4008	0.1564
Koverga (1)[42]	−0.475	0.305	0.185	−0.005
Koverga (2)[42]	−0.475	0.305	0.182	−0.004
RESP-A1 (HF/6-31G*)	−0.4936	0.3924	−0.2543	0.1185
AM1-BCC	−0.3758	0.2087	−0.045	0.0707
sampling distribution	$\sim N(-0.4961, 0.0477^2)$	$\sim N(0.4068, 0.09^2)$	$\sim U(-0.6, 0.2)$	$\sim U(-0.025, 0.2)$

[a]Details on how these charges were calculated are provided in the text.

associated with assigning partial charges. This includes not only uncertainty associated with *ab initio* calculations of electrostatic potential but also with the algorithm chosen to reduce those calculations to point charges necessary for classical force fields. Moreover, there are numerous approximations applied to the various flavors of *ab initio* calculation that may not be straightforward in application for effective calculation. There also exist several semiempirical methods that have been developed to efficiently calculate partial charges from atom types and various corrections to reproduce more rigorous *ab initio* results without the arduous or, for larger compounds, intractable first-principles calculation.[58−63] For all of these potential decisions, what *ab initio* calculation methods to employ, the mapping of a higher-level calculation to point charges, or the choice to use a semiempirical method, there exists no clear standard or decision methodology that is universally accepted in the MD community. Hence, the parameter distributions for ACN partial charges are inferred from verified values available in the literature, as well as those calculated from methods typically combined with AMBER force fields. The values on which the resulting distributions are based thus consider the uncertainty surrounding all levels of partial charge calculation, aside from egregious user error.

The sources and values of nine literature partial charge sets, along with the two sets calculated for this work and the assumed sampling distributions are shown in Table 2. The literature sources applied a variety of methods for the determination of atomic partial charges; yet, all were validated by different means. In the work by Cabaleiro-Lago and Rios, the partial charges were calculated to reproduce the MEP calculated from MP2 and HF calculations (both with a 6-311+G* basis set) under the constraint that the resulting molecular dipole moment matches that of the corresponding *ab initio* calculation.[38] The resulting potential function using these point charges effectively reproduced ACN dimer interactions calculated from MP2 calculations. Grabuleda, Jaime, and Kollman similarly compared results in reproducing macroscopic properties by HF calculations using both the 6-31G* and 6-311+G* basis sets for construction of the MEP followed by RESP fitting.[39] The charges from the 6-31G* calculation produced a more accurate dipole moment, while 6-311+G* better reproduced experimental density and enthalpy of vaporization values. Nikitin and Lyubartsev presented a new all-atom model of ACN in which charges were calculated from the MP2 method with the larger basis set of 6-311++G(3df,3p), resulting in more accurate measurements of molecular dipole moment prior to further modifications of the nonelectrostatic components of the potential model.[40] In work by Caleman and co-workers, a benchmark of 146 organic solvents (including ACN) was performed using both the OPLS/AA and GAFF force fields.[41] For the OPLS/AA molecules, partial charges were assigned based on atom types in the OPLS/AA force field. GAFF charges, conversely, were applied via HF calculation with the 6-311G** basis set and the Merz−Singh−Kollman scheme,[61] thus differing slightly from other sets employing similar basis sets for HF calculation. Interestingly, OPLS/AA performed better at reproducing experimental measurements for a variety of properties over the entire set of solvents, though ACN was an outlier for both force fields. The most recent and divergent set of partial charges comes from work by Koverga et al. in which Car−Parrinello *ab initio* MD simulations of ACN were performed employing the B-LYP&TM/90Ry level of theory with and without van der Waals corrections.[42] The resulting model accurately reproduced high-fidelity electronic properties of ACN.

The two sets of partial charges performed for this work utilize common techniques used for partial charge assignment with AMBER-type force fields. The RESP-A1 charge set employed R.E.D. Server Development[44] to perform HF calculations with the 6-31G* basis set and fit partial charges via the RESP algorithm, originally developed by Bayly et al.[6] Specifically, for RESP-A1, two stages are employed: (1) fitting with a hyperbolic restraint penalty function for all heavy (non-hydrogen) atoms using a restraint weight, qwt, of 0.0005 au and (2) fitting with hyperbolic restraints only for nonpolar heavy atoms (methyl and methylene carbon atoms) with a stronger qwt of 0.001 au while equivalencing hydrogen atoms and fixing charges on atoms within polar groups; this 2-stage fitting scheme has been denoted by Bayly et al. as "wk.fr/st.eq" referring to the weaker and then stronger restraint weight and the absence and then presence of charge equivalencing. This method of producing partial charges from *ab initio* calculations has been widely adopted in the development of AMBER force fields,[7,8,10] including GAFF.[9] There are plenty of other options for the RESP charge parametrization provided by R.E.D. Server Development, and the resulting charge sets from those calculations are discussed in the SI (see section S2). They are not included in the general construction of parameter distributions as not to overly weight R.E.D.-produced charge sets, but the data in section S2 suggests negligible change to the parameter distributions.

https://doi.org/10.1021/acs.jcim.0c01204
J. Chem. Inf. Model. 2021, 61, 1745−1761

Finally, the AM1-BCC charge set utilizes the Austin model 1-bond charge correction (AM1-BCC) semiempirical method[58,59] of applying partial charges as implemented in AmberTools17.[50] This is a common method for applying partial charges to large biomolecules for which *ab initio* calculations may be intractable. The method employs semiempirical QM calculations for the efficient assignment of partial charges that perform generally well with AMBER-type force fields. It should be noted that this method is not typically applied to small molecules (such as ACN) that may be easily parametrized via a more rigorous method (e.g., RESP fitting), but AM1-BCC is included in the parameter distribution as it is generally accepted for use in partial charge assignment in AMBER simulations.

Figure 2 below shows histograms of the 11 charge sets described above, along with kernel density estimates (KDEs) and the sampling distributions described in the last row of Table 2. From this illustration, it may be more clear why an uninformed prior (or uniform distribution) was chosen for the highly variable charges on C2 and H. Moreover, the distributions lend credence to the earlier assertion that the charges on buried atoms are generally the most variable depending on the method of charge assignment used, as values for C2 partial charge range from −0.6 to 0.2 e (electron units or elementary charge).

2.4. Local Sensitivity Analysis (Finite Differences). The local sensitivity analysis in this work follows the finite differences algorithm described in ref 12 section 7.3.1. In short, the method analyzes the sensitivity of a model to small parameter perturbations around a nominal parameter vector, q^0, by constructing the sensitivity matrix, $\chi(q^0)$, from responses at these perturbations:

$$\chi_{ij}(q^0) = \frac{\partial f_i}{\partial q_j}(q^0), \qquad q_j \in q^0$$

Thus, if the parameter vector $q^0 = [q^0_{C1}, q^0_{C2}, q^0_N]$, then the elements associated with the partial charge on N would be

$$\frac{\partial f}{\partial q_N} = \frac{f([q^0_{C1}, q^0_{C2}, q^0_N + h_N]) - f([q^0_{C1}, q^0_{C2}, q^0_N])}{h_N}$$

where h_N is a small perturbation of the partial charge on N. Thus, the sensitivity matrix $\chi^{n\times p}$ is constructed for p parameters and n observations of model $f(q)$. The resulting χ matrix and the residual vector (R) can then be used to estimate the elemental variance, σ_e^2, and the $p \times p$ covaraince matrix, V, by

$$\sigma_e^2 = \frac{1}{n-p} R^T R, \qquad R = f(q^0) - \mathbb{E}[f(q^0)]$$

$$V = \sigma_e^2 (\chi^T \chi)^{-1}$$

Additionally, the mean simulation variance, $\bar{\sigma}_{sim}$, is calculated to estimate the average variance observed from a single simulation:

$$\bar{\sigma}_{sim} = \frac{1}{N} \sum_i^N \sigma_i$$

where N is the total number of simulations and σ_i is the root-mean-square (RMS) fluctuation of the response of interest observed during the simulation, which is often reported as uncertainty estimates in single-simulation calculations.

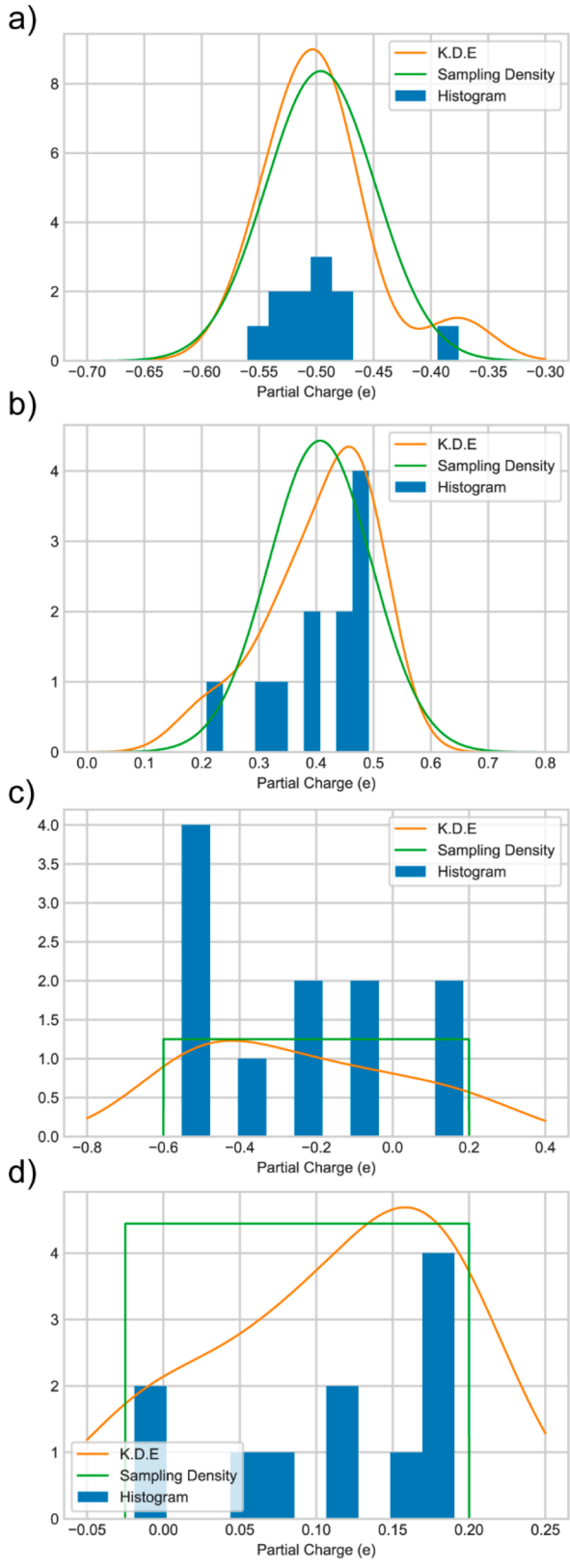


Figure 2. Histograms, kernel density estimates (KDEs) and assumed sampling distributions of the parameter sets shown in Table 2 for (a) N, (b) C1, (c) C2, and (d) H.

2.5. Global Sensitivity Analysis (Morris Screening). Global sensitivity estimates were calculated using the Morris screening approach as described in section 15.2 of ref 12. In

https://doi.org/10.1021/acs.jcim.0c01204
J. Chem. Inf. Model. 2021, 61, 1745−1761

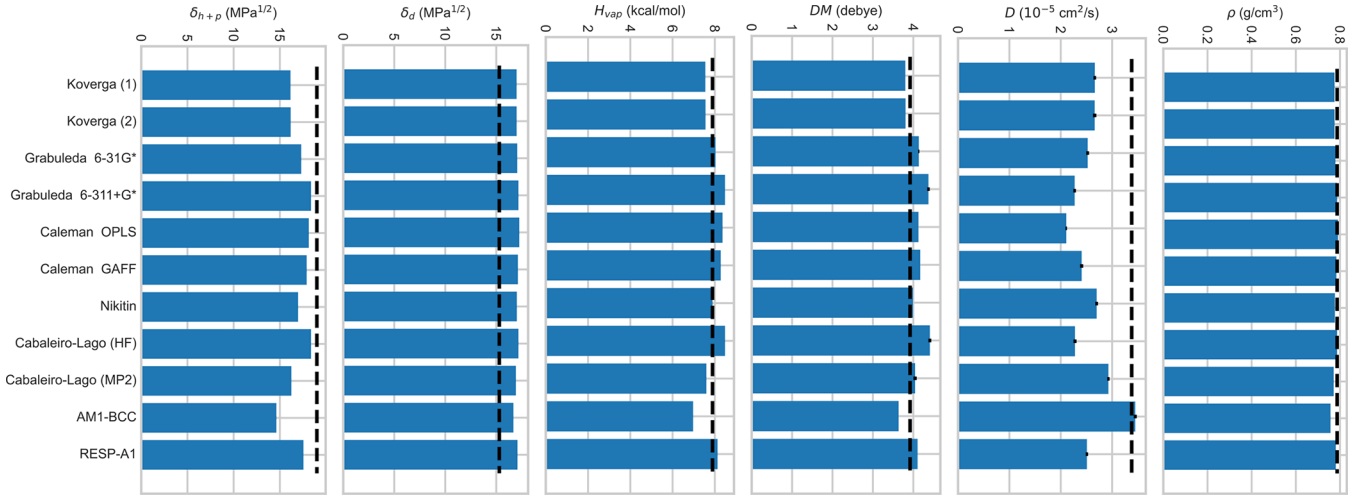


Figure 3. Mean responses measured from 10 simulations using initial charge sets shown in Table 2. Error bars indicate standard deviations (not visible if a standard deviation is below 1 pixel in height). Dashed lines indicate reference values from Table 1.

short, the method estimates the global sensitivity measurements, μ_i^*, and their associated variance, σ_i^2, for each parameter $q_i \in q$, which are defined by

$$\mu_i^* = \frac{1}{r}\sum_{j=1}^{r} |d_i^j(q)|$$

$$\sigma_i^2 = \frac{1}{r-1}\sum_{j=1}^{r} (d_i^j(q) - \mu_i)^2, \qquad \mu_i = \frac{1}{r}\sum_{j=1}^{r} d_i^j(q)$$

where

$$d_i^j = \frac{f(q^j + \Delta e_i) - f(q^j)}{\Delta}$$

is the elementary effect for parameter i and sample j. Δ is the predefined stepsize, and e_i is the ith row of the identity matrix. Note that σ_i is more correctly termed the Morris interaction index, given the relatively large stepsize employed in this work, yet it is a coarse approximation of the standard deviation.[12]

The Morris sampling strategy was also employed to efficiently explore the parameter space reducing the required responses from $2pr$ to $(p + 1)r$ samples by utilizing neighboring measurements. The algorithm for producing sample points is explained in more detail in ref 12, but to produce r parameter points q^*, the parameter distributions described in section 2.3 were transformed to $U(0, 1)$ prior to sampling.

2.6. Surrogate Model Formulation with Gaussian Process Regression. Gaussian process regression (GPR)[35,36] was performed as implemented in the scikit-learn[64] python package. GPR was selected as a method of surrogate formulation as it is efficient at interpolating relatively low-dimensional observations while providing reliable uncertainty estimates of prediction. During training, response RMS error from single simulations was passed as noise (α). To quantify the prediction performance of the GPR surrogate, two indicators were used. First, the log marginal likelihood (LML) of the prediction as calculated from the hyperparameter optimization was recorded. Also, leave one out (LOO) cross-validation[31] was performed in which the model is retrained on all points in the training set but one. The response for this point is predicted, the error recorded, and the process is repeated for all points in the training set. The resulting errors are summarized and reported as the LOO RMS error.

3. RESULTS AND DISCUSSION

We present the results of this study in the following order. First, we show the calculated responses from the reference parameter sets shown in Table 2. This provides context for experiments, yet the low sample size and sparse distribution disallows more rigorous UQ and sensitivity analysis. We then performed a naïve sampling of the parameter space, followed by estimates of local and global sensitivities. Finally, we present the results of applying a Gaussian process machine learning model to predict responses directly from partial charges for greatly increased sampling efficiency.

3.1. Reference Responses. In an effort to provide context to future measurements, we first established a baseline of responses from the initial parameter sets discussed in section 2.3. Ten simulations were performed as described in section 2.1 at each parameter set displayed in Table 2. The resulting responses were then averaged and displayed in Figure 3 with reference values for comparison.

From the plots in Figure 3, one can notice that no charge set exactly reproduces the experimental reference values for all responses. Moreover, for some responses, all charge sets overestimate or underestimate experimental values, likely due to the choice of force field, thermostat (especially in the case of diffusivity), or other intrinsic biases in the implementation of MD. It is here that we must highlight that we do not attempt to estimate optimal charge parameters in this work; doing so would ignore the effects of the other parameters involved in the simulation technique. Instead, we look to quantify the uncertainty brought about by partial charges only and focus little on the reproduction of experimental measurements. UQ experiments aimed at parametrization based on the reproduction of experiment are dependent on the associated cost function (i.e., the relative weights of measurement variance from reference values) and should take into account as many sources of uncertainty (and how they covary) as feasible to avoid cancellation of errors.

The relative variance of responses can be observed from Figure 3 in a qualitative sense. For example, one may readily observe that diffusivity values vary widely whereas density

https://doi.org/10.1021/acs.jcim.0c01204
J. Chem. Inf. Model. 2021, 61, 1745−1761

values and dispersion HSPs are more consistent between charge sets. This result may be expected as it stands to reason that density and dispersion HSPs are directly tied to the van der Waals parameters kept constant through all simulations. Further quantitative comparison, however, is reserved as we are addressing an admittedly sparse input distribution (see histograms in Figure 2). Instead, we perform a more quantitative analysis on a denser sampling of the assumed parameter space.

3.2. Naïve Sampling. We more fully explored the partial charge parameter space for ACN by naively sampling 500 parameter sets from the input distributions described in section 2.3. However, the sampling of these parameters cannot be done completely independently. Doing so would (a) ignore any possible covariance between parameters and (b) violate the condition that all charges must add up to zero. By investigating the correlation plots between partial charge parameters reported in Table 2 (see Figure 4), the sparse

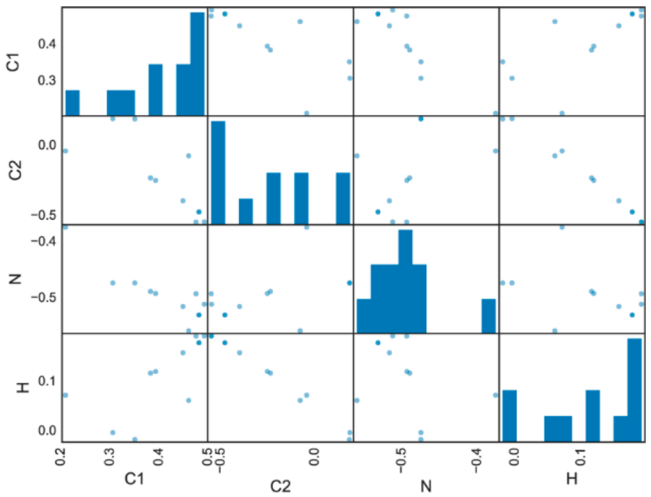


Figure 4. Correlation plot of reference partial charge data shown in Table 2. Histograms of partial charges shown on the diagonal.

distributions complicate the ability to formulate well-defined joint distributions between parameters. However, in the case of the C2 and H charges, there is a very clear negative correlation.

Therefore, we henceforth pulled from the sampling densities of C1, C2, and N (see Figure 2) and set the corresponding H charges to compensate and reach a total net charge of zero. Thus, we assume the model formulation:

$$y = f(q);\ q = [q_{C1},\ q_{C2},\ q_N]$$

where $f(q)$ is our MD simulation. The partial charge on H, q_H, is then calculated after sampling from

$$q_H = \frac{0 - q_{C1} - q_{C2} - q_N}{3}$$

After sampling from the assumed distributions, the 500 simulations were performed and analyzed as described in the methods section. The partial charge parameters and measured responses from all 500 simulations are shown in the large correlation plot in Figure 5. This plot provides a massive amount of qualitative information on the parameter and response distributions and how they relate to each other. Plots in the top-left quadrant confirm that we have indeed sampled independently from the parameter distributions described in section 2.3, as the histograms match the assumed distribution and the off-diagonal correlation plots show no covariance between parameters.

The plots in the lower-right quadrant show the response histograms and how the responses correlate with each other. Vertical lines on the histograms note the approximate location of the experimental reference values. It is interesting to note that the average response values from all the naïve sampling simulations come reasonably close to experimental values for all responses other than the HSPs, though this is discussed in more quantitative detail in the summary below.

What is also interesting is the way in which the responses are single-valued and correlate well with each other, as displayed by the tight distributions in the off-diagonal plots. It is helpful in this case to observe the response correlations in the dipole moment (DM) column as this is the most straightforward response calculation and is most directly related to atomic partial charges. In other words, the only controlled-variable differences between simulations are directly encompassed by the dipole moment response. It is expected that the other responses would then be single-valued with respect to dipole moment if they were to differ at all and thus single-valued with respect to each other. In observing the DM responses with respect to other responses in Figure 5, it is notable that at low dipole moment values, the pairwise responses exhibit less of this single-valued nature, which indicates stochastic forces become more dominant than the electrostatic forces represented in the dipole moment response.

The response covariance is as one might expect; at more extreme dipole moments, the density and intermolecular interaction energies (represented by H_{vap}, δ_d, and δ_{h+p}) all increase, while the diffusivity decreases, indicating less molecular motion. Hence, the responses all act in a relatively predictable fashion and can be treated similarly going forward.

Most germane to consideration of partial charge parametrization, however, are the correlation plots displayed in the top-right and lower-left of Figure 5. These plots indicate how the responses correlate with partial charge values, thus providing a qualitative indication of sensitivity. It may be comforting to observe, given the previous discussion on partial charge assignment methods and the broad distribution associated with the C2 atom, that the responses show no clear correlation with the charge on the buried C2 atom. This indicates that, despite the broad and uninformed prior applied to the parameter (C2 charge), there is little effect on the resulting measured properties. For the other two parameters (C1 and N charges), there does appear to be a clear correlation between responses and parameters but quantitative assessment will require additional sensitivity analysis.

A summary of the responses from naïve sampling experiments over the entire parameter space is displayed in Figure 6. Figure 6a displays the mean response and its relation to the reference responses shown in Table 1. This data can be thought of as the relative accuracy of the responses from a simulation with average partial charges (as defined by the distributions in section 2.3). The results mirror the observations from Figure 3 above. We observe that there is a systematic underestimation of diffusivity, which is likely due to other factors in the simulation (e.g., thermostat[65] and system-size effects[66]). The density, dipole moment, and enthalpy of vaporization, however, are within 2−3% of the reference values.

https://doi.org/10.1021/acs.jcim.0c01204
J. Chem. Inf. Model. 2021, 61, 1745−1761

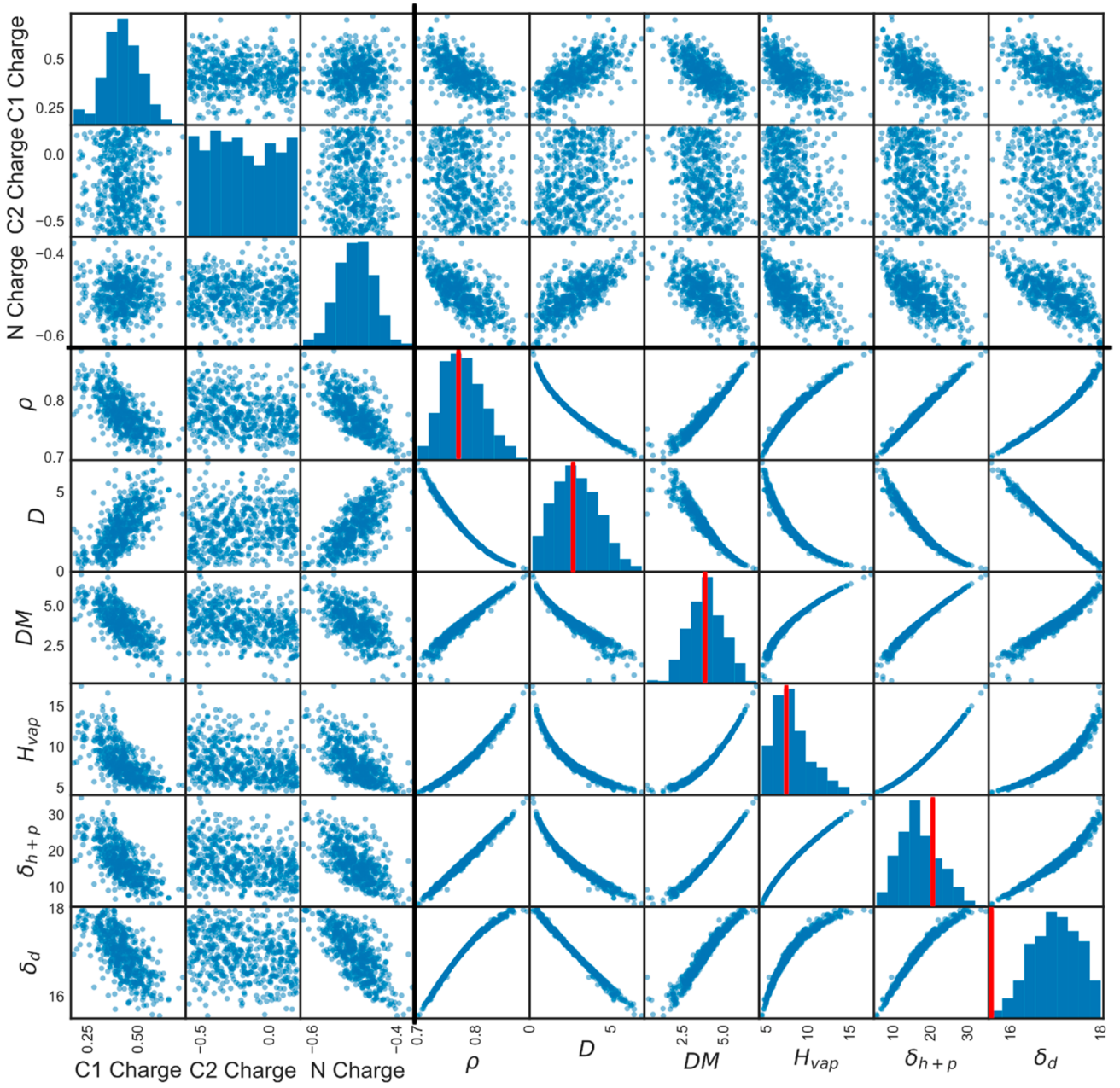


Figure 5. Correlation plots of parameters and responses from 500 naïve sampling simulations with histograms on the diagonal. Bold lines delineate between parameters and responses. Vertical red lines indicate reference responses from Table 1. Note that the responses are mostly single-valued and correlate well with each other, suggesting the reliance of these responses on dipole moment. Additionally, note that δ_d and δ_{h+p} are over- and underestimated in aggregate, indicating a cancellation of errors. (Note that the complete range of δ_d is much smaller compared to that of δ_{h+p}.)

This is not in any way to suggest that the best method of partial charge assignment is to survey the literature for all possible methods and partial charge values, perform representative simulations at these values, and take the average response. Doing so would not only be terribly inefficient and open to bias but would also ignore interactions with other sources of uncertainty within MD simulations. This is exemplified by the HSP responses in Figure 6a; δ_{h+p} is underestimated, whereas δ_d is overestimated, on average. However, these are both components of the enthalpy of vaporization, H_{vap}, which shows little difference from reference values. This may be due to an effective cancellation of errors, i.e., in force field parametrization (which often takes H_{vap} into account), the higher van der Waals forces may be compensating for lower electrostatic interactions. Indeed, one of the benefits of the RESP method, recommended for use with GAFF, is its slight overestimation of molecular dipole moment (over gas-phase values), which provides a dipole moment in closer proximity to that in a condensed phase and thus meshes well with the TIP3P water model typically used in conjunction with AMBER-type force fields.[7,8]

What is also likely, however, is the intrinsic error associated with heuristics applied to the formulation of a classical MD force field itself. In the nonbonded interaction term, the van der Waals and electrostatic (or Coulombic) interactions are neatly separated to aid in calculation, parametrization, and general tractability. This, of course, is not an exact model of interatomic interactions and thus we may be observing an

https://doi.org/10.1021/acs.jcim.0c01204
J. Chem. Inf. Model. 2021, 61, 1745−1761

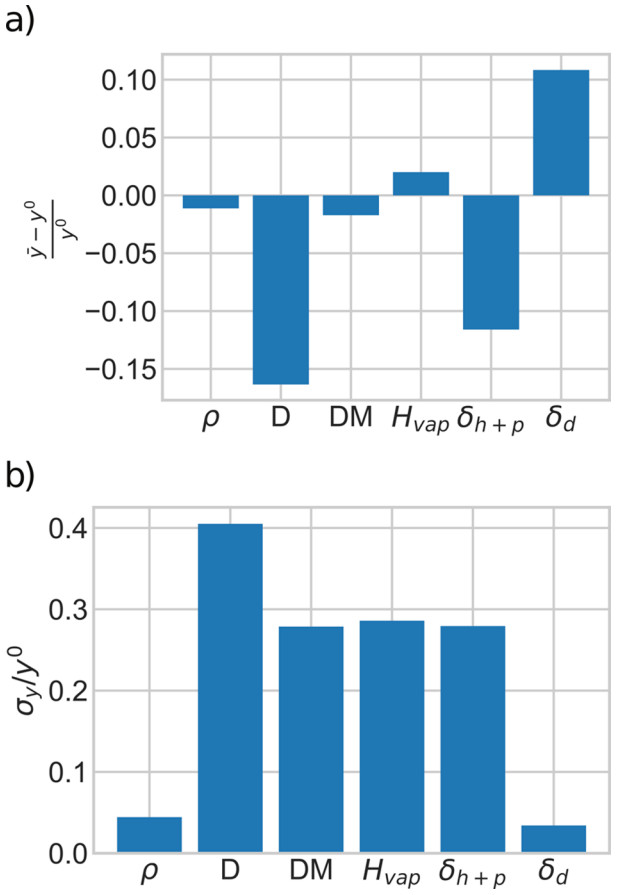


Figure 6. Summary of naive sampling simulations. (a) Difference of mean response ($\bar{y}$) from reference response (y^0, from Table 1) divided by y^0. (b) Standard deviation of responses (σ_y) divided by y^0.

artifact of this necessary assumption. Yet in both cases, this serves to reiterate the fact that partial charges and force fields are not wholly independent of each other. Thus, if one looks to apply partial charges, the most "accurate" (in terms of reproducing high-fidelity *ab initio* calculations) may not be the most useful in conjunction with a given force field.

Figure 6b displays the normalized standard deviation of the responses throughout the parameter space as sampled in this experiment. This can be interpreted as a very basic quantification of sensitivity. Responses less associated with Coulombic interactions such as density and dispersion HSP show little variance, while diffusivity varies by over 40% on average. As the density and dispersion HSP show very little variance, it is suggested that the variance of H_{vap} is primarily due to the electrostatic interactions summarized by δ_{h+p}. This is further supported by the very tight $H_{vap} - \delta_{h+p}$ correlation plot seen in Figure 5.

3.3. Local Sensitivity. A more quantitative measurement of sensitivity in the vicinity of nominal parameters was performed using the finite differences approach outlined in section 2.4. The same q vector as that used for the naïve sampling experiment was employed, and RESP-A1 values were used as nominal parameters:

$$q^0 = [q^0_{C1}, q^0_{C2}, q^0_N] = [0.3924, -0.2543, -0.4936]$$

with stepsize $h = q^0 \times 10^{-4}$. Smaller values of h are often used in local sensitivity experiments, but such small changes in partial charge can lead to rounding errors in simulation. For each perturbation as well as for the nominal parameters, 30 simulations were performed to resolve elemental response variance.

The resulting sensitivity matrices, χ, are illustrated in Figure 7. (The raw responses from the finite difference simulations are provided in section S3, Figure S4, of the Supporting Information.) Points indicate sensitivity values for each

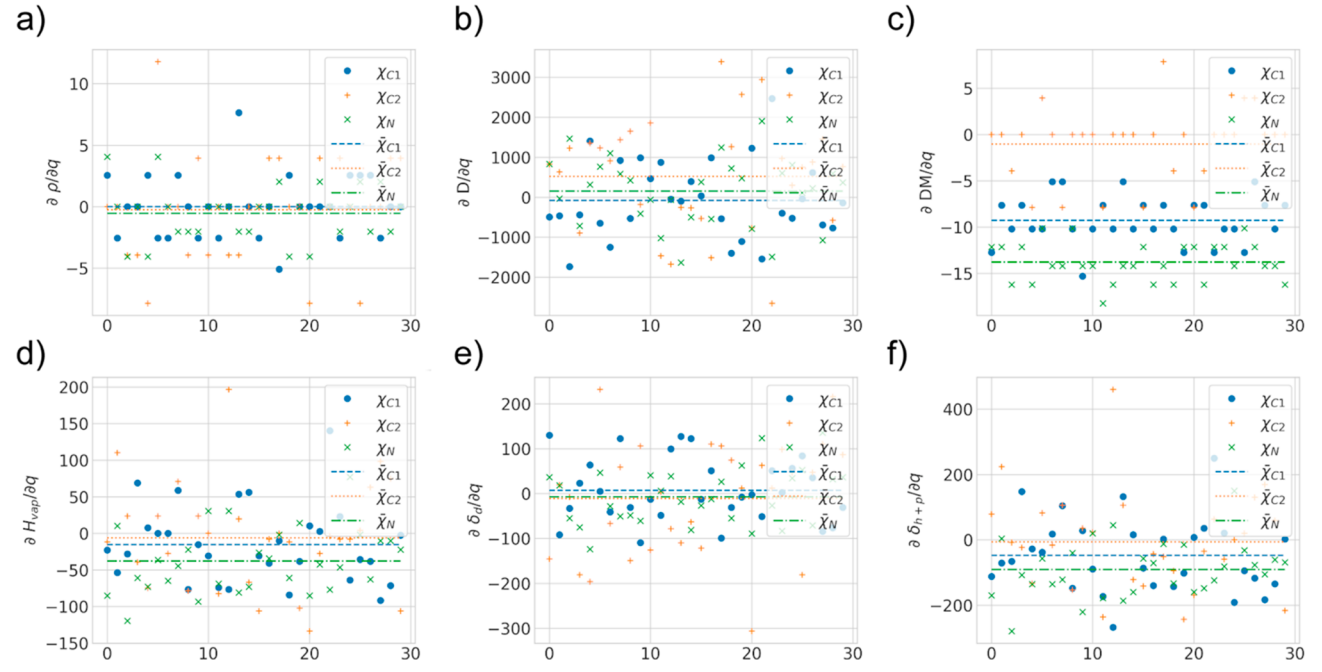


Figure 7. Components of sensitivity matrices (χ) from finite differences experiments for (a) density, (b) self-diffusivity, (c) dipole moment, (d) enthalpy of vaporization, (e) dispersion HSP, and (f) polar-plus-hydrogen-bonding HSP. Individual simulation sensitivity measurements plotted on x-axes. Dashed lines represent mean sensitivity measurements for each parameter ($\bar{\chi}_i$).

https://doi.org/10.1021/acs.jcim.0c01204
J. Chem. Inf. Model. 2021, 61, 1745−1761

parameter calculated from single simulations, whereas dashed lines indicate the average parameter sensitivity ($\bar{\chi}_i$) over all 30 simulations.

Many of the observations from the naïve sampling experiment results in Figure 6b are reproduced. For example, both ρ and δ_d show little to no sensitivity to any of the parameters, as indicated by values of $\bar{\chi}$ close to zero (Figure 7a and e, respectively). Yet the plots of χ provide additional insight into the sensitivity of responses to specific variables. This is most clear in the case of *DM* (see Figure 7c) where there is relatively low variance in responses at each perturbation. It can be seen that there is little to no change to *DM* due to changes in the charge on C2 (i.e., *DM* is effectively insensitive to q_{C2}). Comparatively, *DM* is most sensitive to the charge on N, followed by the charge on C1. Both H_{vap} and δ_{h+p} reproduce this trend, though the elemental variances are particularly large, as indicated by the overlapping of points of different types on the plots in Figure 7d and f. Average sensitivity values for *D* show the opposite trend, yet the elemental variances exhibited in Figure 7b are even larger, which further clouds the picture of parametric sensitivity of this response.

The root of these elemental variances are shown in comparison to the average single-simulation errors in Table 3. In all cases but *D*, the root of the elemental variance is

Table 3. Root of Elemental Variance (σ_e) and Average Single-Simulation RMS Error ($\bar{\sigma}_{sim}$) Calculated for Finite Difference Experiment As Described in Section 2.4

y	y^0	σ_e	$\bar{\sigma}_{sim}$	units
ρ	0.786	8.1×10^{-5}	3.3×10^{-3}	g/cm^3
D	3.38	3.0×10^{-2}	1.5×10^{-3}	10^{-5} cm^2/s
DM	3.92	7.0×10^{-5}	2.8×10^{-3}	Debye
H_{vap}	7.89	1.2×10^{-3}	6.7×10^{-2}	kcal/mol
δ_d	15.3	2.3×10^{-3}	0.10	MPa$^{1/2}$
δ_{h+p}	19.0	2.3×10^{-3}	0.14	MPa$^{1/2}$

orders of magnitude below the average RMS error of a single simulation. This suggests that the random variance that takes place in a single simulation is greater than the variance imposed by the partial charge variations of this magnitude. Hence, these local sensitivity values do not provide adequate certainty for perturbations of this size.

Hence, a more global sensitivity study was conducted to fully explore the parameter space and get a quantitative measurement of sensitivity for these values with large elemental variance. However, it is notable that this experiment produced very little sensitivity of ρ and δ_d to all parameters, and sensitivity of *DM* being largest for N followed by C1 and finally C2. Given the covariance between *DM* and the other responses illustrated in Figure 5, it can be hypothesized that the other responses for which elemental variances were large would follow a similar trend. Additionally, the diagonal elements of the covariance matrices, *V*, showed insignificant values indicating no parameter covariance (see section S3, Table S1, in the Supporting Information). Analysis of the variance between identical simulations also suggests that ten simulations, rather than 30, provides response estimates with effectively the same precision (see section S4, Table S2, in the Supporting Information). Hence, in global sensitivity analysis going forward, ten simulations at each parameter set are performed to estimate reliable responses.

3.4. Global Sensitivity. To get better estimates of parameter sensitivity, we looked to employ a global sensitivity approach. Unlike local sensitivity metrics, which estimate sensitivity in small step sizes about a nominal value, global sensitivity metrics seek to quantify sensitivity throughout the parameter space. In this way, elemental variances in small parameter perturbations can be overcome.

We performed a detailed global sensitivity analysis using the Morris screening approach as described in section 2.5. The same parameter vector q was used as in the naïve sampling and local sensitivity experiments, and the same prior distributions as those discussed in section 2.3 were used. We employed a sample size of $r = 10$ and a stepsize of $\Delta = 2/3$. This is a relatively large stepsize in order to maintain efficiency, yet we confirm later that the results of this study are reproducible at a smaller stepsize. The variance between simulations at $n = 10$ suggests a sample size of $r = 10$ to be sufficient to capture sensitivity. Note that this step is applied after transformation of parameter distributions to $U(0, 1)$ and thus represents stepping over 2/3 of the entire parameter space. Based on our variance calculations from the local sensitivity calculations (see section S4, Table S2, in the Supporting Information), we performed $n = 10$ simulations at each sample point and calculated the mean response. Thus, a total of $(p + 1)rn = 400$ simulations were necessary for this analysis.

The resulting global sensitivity measurements, μ^*, are shown graphically in Figure 8. By normalizing the global sensitivity values by the reference responses, y^0, a comparison of relative sensitivity between responses can be more easily made (see Figure 8b).

In many ways, the results in Figure 8b reiterate with much greater certainty the observations from previous estimates of sensitivity. As hypothesized above, the parametric sensitivities follow the same trend as was observed for *DM*, as the sensitivity is greatest to the N partial charge, followed by C1, and then by C2, for all responses. Responses of ρ and δ_d are largely insensitive to changes in partial charges, which may be attributed to their association with van der Waals parameters.

However, the global sensitivity measurement allows for a more detailed description of parameter variance effects on responses that were clouded by elemental variance in previous experiments. Diffusivity, for example, is clearly shown as the most sensitive response to partial charge parameters, though the measurement variance belies the Morris interaction index (σ_i) associated with *D* calculation. Similarly, sensitivity measurements for H_{vap} and δ_{h+p} show significant associated variance as represented by the Morris interaction index. Conversely, *DM* sensitivity measurements show relatively low Morris interaction indices, owing to the direct and straightforward dependence of dipole moment on partial charges. Yet even for responses that show significant Morris interaction indices, they are low enough to provide direct comparison of sensitivity values.

3.5. Sampling Method. After the estimation of sensitivity measurements associated with the partial charges on C1, C2, and N, there were still some questions remaining in regard to the formulation of the experiment itself. As described in section 3.2, the input partial charge data was too sparse to reliably formulate joint distributions, so charge compensation on the hydrogen atoms was employed to maintain a net neutral charge based on the strong correlation between C2 and H in the reference data (see Figure 4). However, there is a concern that the insensitivity to the C2 atom observed in all responses

https://doi.org/10.1021/acs.jcim.0c01204
J. Chem. Inf. Model. 2021, 61, 1745−1761

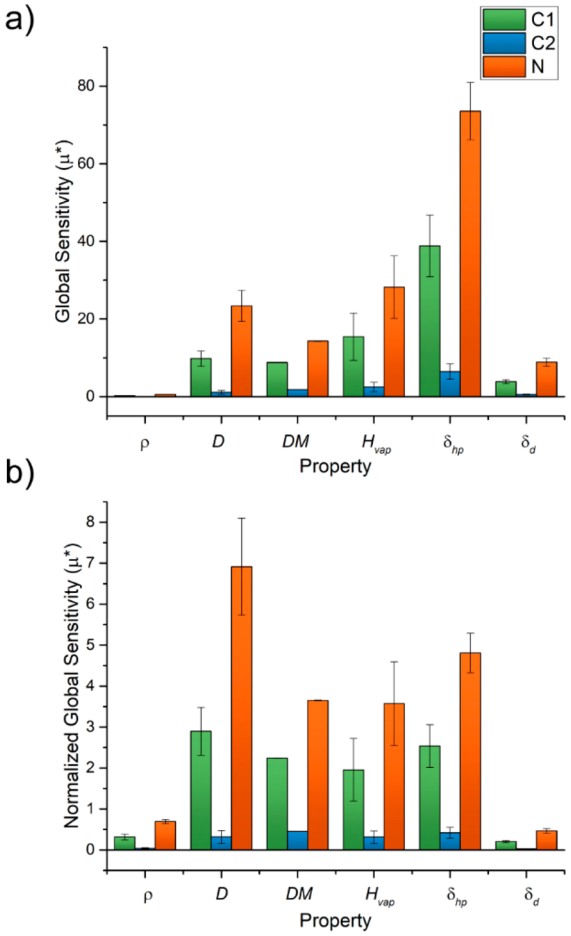


Figure 8. Global sensitivity measurements (μ^*) displayed as both (a) absolute values (in typical response units/e) and (b) normalized by reference responses (y^0) from Table 1 for comparison between responses (e^{-1}). Error bars indicate the Morris interaction index, σ_i, as calculated in section 2.5.

(see Figure 8) could be an artifact of the sampling method. Given that all changes to net molecular charge are compensated by the charge on H atoms, it is trivial to observe that changes to the charge on C2 with C1 and N charges held constant will not significantly affect the overall dipole moment (see Figure 1). As strong correlation was observed between responses and *DM* (see Figure 5), this artifact could conceivably be carried over to other responses.

To address this possibility, the same global sensitivity strategy employed in section 3.4 was repeated with the three other possible charge schemes, i.e., $q = [q_H, q_{C2}, q_N]$, $q = [q_H, q_{C1}, q_N]$, and $q = [q_H, q_{C1}, q_{C2}]$. For q_H, we employed the uniform sampling density shown in the last row of Table 2 and Figure 2d. All other constants in the global sensitivity calculation were retained ($r = 10$, $\Delta = 2/3$, $n = 10$ simulations per sample point). The resulting normalized global sensitivity measurements are displayed in Figure 9.

The first observation from the sensitivity measurements is the extremely large Morris interaction index observed for sensitivity measurements where H and C2 charges were sampled independently in the same experiment (see Figure 9a and c). Hence, it appears that our choice of compensation on the hydrogen atoms is a useful method of reducing the sensitivity measurement uncertainty by effectively decoupling phenomena caused by H charges and C2 charges. Indeed, the

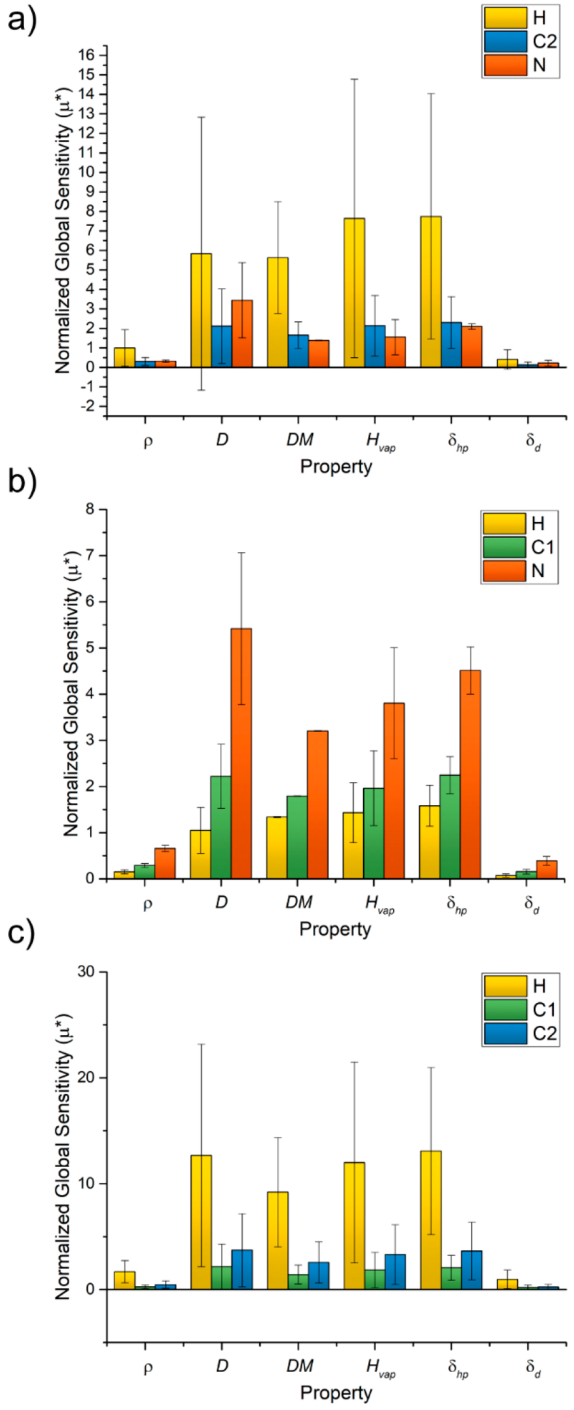


Figure 9. Normalized global sensitivity measurements (μ^*) from other sampling methods, q = (a) $[q_H, q_{C2}, q_N]$, (b) $[q_H, q_{C1}, q_N]$, (c) $[q_H, q_{C1}, q_{C2}]$. Error bars indicate the associated Morris interaction index, σ_i, calculated as in section 2.5.

independent sampling from C2 and H distributions simultaneously confuses the parametric global sensitivity calculation.

In the second case, where $q = [q_H, q_{C1}, q_N]$, the charge on C2 is now used for compensation and the H charge is sampled from the associated parameter distribution. With this sampling method, we observe a similar trend in response sensitivity of parameters in that the sensitivity is greatest to the N charge, followed by C1, and then H. In comparison to the initial global sensitivity results in section 3.4, however, the responses are more sensitive to changes in the H charge than changes in the

https://doi.org/10.1021/acs.jcim.0c01204
J. Chem. Inf. Model. 2021, 61, 1745−1761

C2 charge. Indeed, sensitivities of responses to the charge on H approach and are sometimes within error of those to the charge on N. This is a result of the multiplicity of H atoms; a change of 0.1 e to the charge on a single H atom effects all three H atoms equally, thus corresponding to a 0.3 e total charge modification. Hence, we observe the sensitivity to H charges being approximately three times that of C2 charges in the original global sensitivity scheme.

The results observed for the third case, where $q = [q_H, q_{C1}, q_{C2}]$, can again be explained by considering the balance atom, N. The increased sensitivity observed for C2 relative to C1 can be attributed to the effect that charge redistribution has on the overall dipole. By transferring charge from the closer C1 to N, the overall effect on the dipole moment of the molecule is less extreme than the transfer from C2 to N on the other side of the molecule. Hence, even the buried atom C2 has an effect on ultimate properties, yet those are dwarfed by the sensitivity attributed to H.

3.6. Surrogate Model Formulation via Gaussian Process Regression. As can be gleaned from our discussion of global sensitivity analysis procedures, uncertainty quantification and sensitivity analysis of MD simulations can involve large numbers of simulations and potentially massive simulation times. In our case, the 400 simulations necessary for global sensitivity analysis involve approximately 5.5 h per simulation and correspond to over 90 days of total, sequential computational time on a modern GPU. Depending on available resources, the total computational time can be reduced to some degree by executing multiple simulations simultaneously; however, the computational time will increase exponentially as one adds more parameters or degrees of freedom, not to mention additional resources necessary to explore more complex materials that require longer or larger simulations.

In cases where models are computationally expensive, it is often useful to develop an efficient surrogate model that estimates the responses from the high-fidelity model. Gaussian process regression (GPR) is an attractive method for developing surrogate models as they are extremely efficient at interpolating low-dimensional models ($p < 100$) with continuous response surfaces and can provide Bayesian estimates of uncertainties associated with predictions.[35] In our case, we have only three parameters ($p = 3$), and previous sampling experiences suggest that responses continuously change with parameter changes (i.e., there are no second-order jumps in responses at specific parameter sets).

Thus, we trained a GPR surrogate model on random subsets of the data from the naïve sampling experiment described in section 3.2. It should be noted that this training set is ideal for GPR, as it fully explores the parameter space of interest. GPR is not recommended for extrapolation to predict responses outside the parameter training set.[12,35] In training the GPR surrogate model, a primary decision is the choice of kernel function or the basis covariance function for training points. In the optimization of our GPR surrogate model, we explored relatively simple kernels to avoid overfitting. Additionally, we have some knowledge that the underlying function is neither periodic nor exponential in nature. Thus, we explored the radial basis function (RBF), Matern, and the rational quadratic (RQ) kernels. More information on these kernels and their implementation can be found in chapter 4 of ref 35.

For all responses, these three kernels were tested, along with the software default (a nonoptimized RBF kernel) and the kernels with an optimized constant kernel multiplier to add offsets. The kernels were trained on random subsets of simulation responses from the naïve sampling experiment in section 3.2 of sample size n = 10, 20, 50, 100, 250, and 500. Representative plots of the GPR effectiveness, as quantified by the LOO prediction error and the log marginal likelihood, are shown in Figure 10 for the H_{vap} response. Similar plots are shown in section S5 of the Supporting Information for all other responses. From this data, the rational quadratic kernel with a constant kernel multiplier ($k*RQ$) was employed for the GPR surrogate model as it performed the best at both minimizing LOO prediction error and maximizing log marginal likelihood.

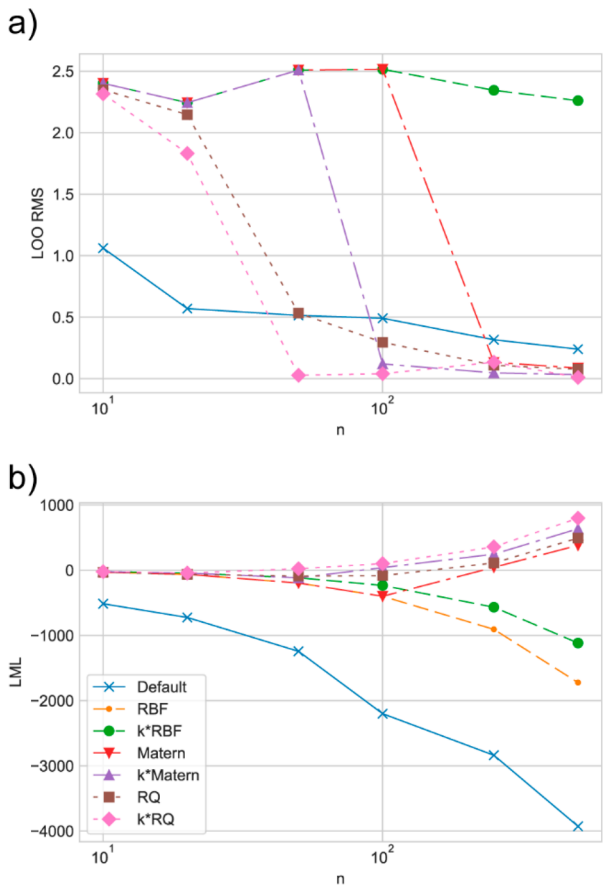


Figure 10. (a) Leave one out (LOO) RMS error and (b) log marginal likelihood (LML) from Gaussian process regression models to predict enthalpy of vaporization (H_{vap}) with various kernels as a function of the number of random simulations from the naïve sampling experiment used as a training set. Similar plots for all other responses are provided in section S5 of the SI (Figures S5 and S6).

In the following experiments, the GPR was trained on all 500 parameter and response points from the naïve sampling experiment in section 3.2 with the $k*RQ$ kernel. After training, the surrogate model was validated by performing the same global sensitivity analysis experiment as in section 3.4 but using the GPR surrogate to predict responses in lieu of running actual simulations. Hence, the experiment could be conducted in a matter of seconds rather than the days of computational time necessary for running actual simulations. The resulting sensitivity measurements, shown in Figure 11a, are nearly identical to those presented in the high-fidelity experiment (see Figure 8b) with only slightly larger measurement uncertainties

https://doi.org/10.1021/acs.jcim.0c01204
J. Chem. Inf. Model. 2021, 61, 1745−1761

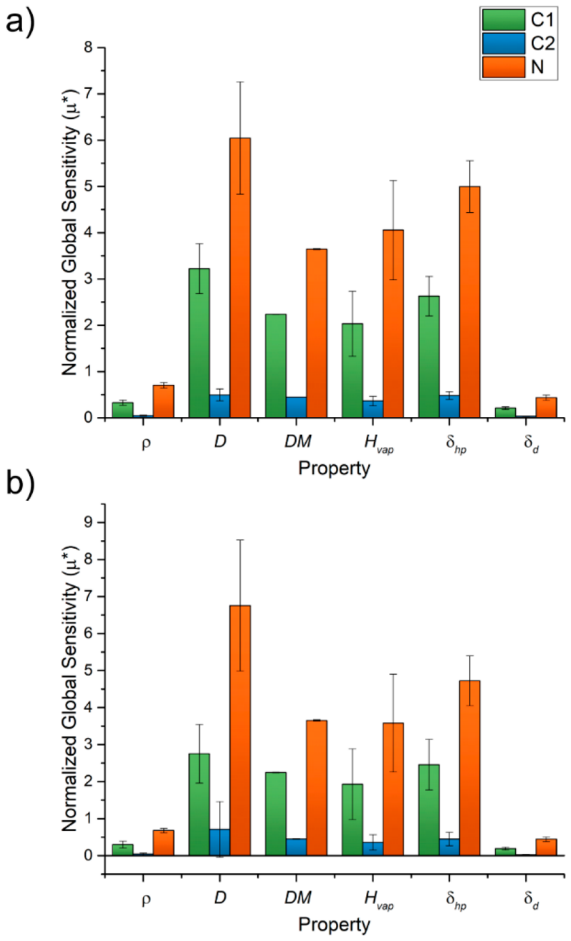


Figure 11. Results from Morris screening experiments performed with GPR surrogate model for property prediction. (a) Reproduction of original global sensitivity analysis with GPR-predicted responses. (b) Results of global sensitivity with a smaller stepsize, $\Delta = 1/50$.

due to the increased uncertainty associated with GPR predictions.

This ability to predict MD responses in seconds rather than hours allows us to make changes to our sensitivity analysis methodology and rapidly observe results. For example, we looked to explore the effect of stepsize (Δ) on the global sensitivity approach applied above. We observed in section 3.3 that partial charge changes on the order of 10^{-5} were too small to effectively calculate sensitivity, yet the previous stepsize of $\Delta = 2/3$ employed in global sensitivity may be contributing to some of the uncertainty observed in sensitivity measurements. The global sensitivity experiment was repeated with a smaller stepsize of $\Delta = 1/50$, yet the results were not significantly different (see Figure 11b) from the larger stepsize experiment.

4. CONCLUSION

In this work, we present the results of a rigorous uncertainty quantification and sensitivity analysis of the effect of atomic partial charges on MD simulations for organic solvent properties. By surveying the literature and performing popular methods of partial charge calculation, we discovered a range of potential atomic partial charges for the relatively simple case of ACN. From this distribution of partial charge parameters, we were able to observe a range of macroscopic solvent properties calculated from MD. We conducted a local sensitivity analysis experiment but found the elemental variance of some bulk properties to be too large to rigorously determine parametric sensitivity. A global sensitivity analysis, however, utilizing a Morris screening approach allowed us to effectively estimate sensitivity measurements for each atomic charge in the molecule. We also developed a predictive surrogate model using Gaussian process regression (GPR) for further exploration of global sensitivity parametrization, allowing us to validate our results with exponentially faster computation speed.

The results provide significant insight into the nature of partial charges and their potential effect on MD calculations. First, we see that some bulk properties (ρ, δ_d) are relatively insensitive to partial charges. This result is somewhat intuitive, as these properties are directly related to nonelectrostatic parameters, yet it sheds further light on the necessity of observing multiple quantities of interest during force field parametrization or validation. Simply reproducing the experimental density in experiment does not provide validation of electrostatic parametrization.

We also observe that the differences in partial charge calculation methods can create uncertainty in the parametrization procedure. It can be difficult for MD practitioners to determine *a priori* which method of charge assignment is the most useful for their experiment. Yet, we note in our calculations that QM accuracy may not be the most effective metric for partial charge quality—rather, users should consider the entire parameter set and employ the partial charge set that is compatible with the force field of their choosing. There may be a cancellation of errors or other forms of compensation occurring. As with force fields and interatomic potentials in general, there is no universal technique that will apply to all MD experiments. Hence, one must determine a validation procedure that provides sufficient evidence that the phenomenon of interest is reproduced to an acceptable level of certainty.

We also illustrate that machine learning can be an effective tool for uncertainty quantification of MD simulations. Often computationally expensive calculations eschew rigorous UQ analysis. Yet, we observe that the application of surrogate model formulation via GPR can provide rapid interpolation of simulation results, provided effective training. The recent explosion in interest in machine learning application to computational simulation[67−70] has given rise to many fascinating debates on how these advanced algorithms may enhance simulation methodology while maintaining scientific rigor. As these algorithms become more and more well recognized, their careful use as rapid interpolators for surrogate model calculations to better define parametric sensitivities and validation is certainly a promising partnership between machine learning and traditional chemical calculations.[31]

Finally, we note the overall sensitivity of bulk properties calculated from MD simulations is not to be ignored. Our analysis shows that partial charge variance introduced by various popular calculation methods can result in significant changes to macroscopic properties such as enthalpy of vaporization and self-diffusivity. However, it is important to note that low-accessibility (or buried) atoms, typically associated with the most partial charge uncertainty, fortunately had the lowest effect on macroscopic properties. We believe these results are likely to be found in other systems where Coulombic interactions are important to the quantities of interest (i.e., less so in low-polarity environments).

https://doi.org/10.1021/acs.jcim.0c01204
J. Chem. Inf. Model. 2021, 61, 1745−1761

Moving forward, we believe that the techniques and principles reported here are of great importance to the MD community. Partial charge assignment is one of the most often applied parametrization steps in MD calculations and is critical to ultimate results. It is thus crucial to understand the associated uncertainties and how they may affect the quantities of interest. It is the view of the authors that partial charge derivation methods would provide estimates of their uncertainty such that MD practitioners are more aware of the potential effect. Of course, these parameters do not exist independently of the remainder of the interatomic potential; the compatibility of the charge method and broader force field must be considered.

Further, the application of machine learning-based surrogate models can be a boon to more rigorous experiments going forward. Validation has been an important but difficult step in MD simulation since its invention, and the ability of readily available algorithms to streamline this process is indeed promising. While the generalization of the specific results to MD simulations more broadly may be somewhat speculative as applications of MD vary so widely, the procedure of sensitivity analysis via trained surrogate can be reapplied to specific cases as part of the validation step of MD simulations where valuable, e.g., when parameter distributions may be assumed, partial charge uncertainty is high (especially on exposed atoms), and polarity is likely important to the quantity of interest.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.0c01204.

Diffusivity calculations and diffusion experiment plots, partial charges from R.E.D. Server Development, additional final difference results for density, self-diffusivity, dipole moment, enthalpy of vaporization, HSP, response variance in identical parameters, and GPR kernel test results (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Yaroslava G. Yingling − *Department of Materials Science and Engineering, North Carolina State University, Raleigh, North Carolina 27695, United States of America;* orcid.org/0000-0002-8557-9992; Email: yara_yingling@ncsu.edu

Authors

James S. Peerless − *Department of Materials Science and Engineering, North Carolina State University, Raleigh, North Carolina 27695, United States of America*

Albert L. Kwansa − *Department of Materials Science and Engineering, North Carolina State University, Raleigh, North Carolina 27695, United States of America*

Branden S. Hawkins − *Department of Materials Science and Engineering, North Carolina State University, Raleigh, North Carolina 27695, United States of America*

Ralph C. Smith − *Department of Mathematics, North Carolina State University, Raleigh, North Carolina 27695, United States of America*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.0c01204

Funding

The authors acknowledge the funding provided by the National Science Foundation (CMMI-1727603 and CMMI-1763025) and the NSF Research Traineeship on Data-Enabled Science and Engineering of Atomic Structures (DGE-1633587).

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors would like to acknowledge the National Science Foundation Research Traineeship on Data-enabled Science and Engineering of Atomic Structures (SEAS) for financially supporting this work (DGE-1633587). The first author would also like to thank Nina J. B. Milliken for fruitful discussions on many aspects of this work. The first author also thanks Patxi Fernandez-Zelaia for motivating work on Gaussian process regression.

■ REFERENCES

(1) Zhao, G.; Perilla, J. R.; Yufenyuy, E. L.; Meng, X.; Chen, B.; Ning, J.; Ahn, J.; Gronenborn, A. M.; Schulten, K.; Aiken, C.; Zhang, P. Mature HIV-1 Capsid Structure by Cryo-Electron Microscopy and All-Atom Molecular Dynamics. *Nature* **2013**, *497*, 643.

(2) Gao, Y.; Iqbal, S.; Zhang, P.; Qiu, M. Performance and Power Analysis of High-Density Multi-GPGPU Architectures: A Preliminary Case Study. *2015 IEEE 17th International Conference on High Performance Computing and Communications, 2015 IEEE 7th International Symposium on Cyberspace Safety and Security, and 2015 IEEE 12th International Conference on Embedded Software and Systems* **2015**, 66−71.

(3) Nash, J. A.; Singh, A.; Li, N. K.; Yingling, Y. G. Characterization of Nucleic Acid Compaction with Histone-Mimic Nanoparticles through All-Atom Molecular Dynamics. *ACS Nano* **2015**, *9* (12), 12374−12382.

(4) 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 2nd Generation Force-Field for the Simulation of Proteins, Nucleic-Acids, and Organic-Molecules. *J. Am. Chem. Soc.* **1995**, *117* (19), 5179−5197.

(5) Sigfridsson, E.; Ryde, U. Comparison of Methods for Deriving Atomic Charges from the Electrostatic Potential and Moments. *J. Comput. Chem.* **1998**, *19* (4), 377−395.

(6) 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* (40), 10269−10280.

(7) Cieplak, P.; Cornell, W. D.; Bayly, C.; Kollman, P. A. Application of the Multimolecule and Multiconformational RESP Methodology to Biopolymers: Charge Derivation for DNA, RNA, and Proteins. *J. Comput. Chem.* **1995**, *16* (11), 1357−1377.

(8) Fox, T.; Kollman, P. A. Application of the RESP Methodology in the Parametrization of Organic Solvents. *J. Phys. Chem. B* **1998**, *102* (41), 8070−8079.

(9) Wang, J.; 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* (9), 1157−1174.

(10) Wang, J.; Cieplak, P.; Kollman, P. A. How Well Does a Restrained Electrostatic Potential (RESP) Model Perform in Calculating Conformational Energies of Organic and Biological Molecules? *J. Comput. Chem.* **2000**, *21* (12), 1049−1074.

(11) Gross, K. C.; Seybold, P. G.; Hadad, C. M. Comparison of Different Atomic Charge Schemes for Predicting PKa Variations in Substituted Anilines and Phenols. *Int. J. Quantum Chem.* **2002**, *90* (1), 445−458.

(12) Smith, R. C. *Uncertainty Quantification: Theory, Implementation, and Applications*; SIAM: Philadelphia, 2014.

(13) Wong, C. F.; Rabitz, H. Sensitivity Analysis and Principal Component Analysis in Free Energy Calculations. *J. Phys. Chem.* **1991**, *95* (24), 9628−9630.
(14) Zhu, S.; Wong, C. F. Sensitivity Analysis of Water Thermodynamics. *J. Chem. Phys.* **1993**, *98* (11), 8892−8899.
(15) Zhu, S.-B.; Wong, C. F. Sensitivity Analysis of a Polarizable Water Model. *J. Phys. Chem.* **1994**, *98* (17), 4695−4701.
(16) Frederiksen, S. L.; Jacobsen, K. W.; Brown, K. S.; Sethna, J. P. Bayesian Ensemble Approach to Error Estimation of Interatomic Potentials. *Phys. Rev. Lett.* **2004**, *93* (16), 165501.
(17) Cailliez, F.; Pernot, P. Statistical Approaches to Forcefield Calibration and Prediction Uncertainty in Molecular Simulation. *J. Chem. Phys.* **2011**, *134* (5), No. 054124.
(18) Angelikopoulos, P.; Papadimitriou, C.; Koumoutsakos, P. Bayesian Uncertainty Quantification and Propagation in Molecular Dynamics Simulations: A High Performance Computing Framework. *J. Chem. Phys.* **2012**, *137* (14), 144103.
(19) Angelikopoulos, P.; Papadimitriou, C.; Koumoutsakos, P. Data Driven, Predictive Molecular Dynamics for Nanoscale Flow Simulations under Uncertainty. *J. Phys. Chem. B* **2013**, *117* (47), 14808−14816.
(20) Rizzi, F.; Najm, H. N.; Debusschere, B. J.; Sargsyan, K.; Salloum, M.; Adalsteinsson, H.; Knio, O. M. Uncertainty Quantification in MD Simulations. Part I: Forward Propagation. *Multiscale Model. Simul.* **2012**, *10* (4), 1428−1459.
(21) Rizzi, F.; Najm, H. N.; Debusschere, B. J.; Sargsyan, K.; Salloum, M.; Adalsteinsson, H.; Knio, O. M. Uncertainty Quantification in MD Simulations. Part II: Bayesian Inference of Force-Field Parameters. *Multiscale Model. Simul.* **2012**, *10* (4), 1460−1492.
(22) Rizzi, F.; Jones, R. E.; Debusschere, B. J.; Knio, O. M. Uncertainty Quantification in MD Simulations of Concentration Driven Ionic Flow through a Silica Nanopore. I. Sensitivity to Physical Parameters of the Pore. *J. Chem. Phys.* **2013**, *138* (19), 194104.
(23) Rizzi, F.; Jones, R. E.; Debusschere, B. J.; Knio, O. M. Uncertainty Quantification in MD Simulations of Concentration Driven Ionic Flow through a Silica Nanopore. II. Uncertain Potential Parameters. *J. Chem. Phys.* **2013**, *138* (19), 194105.
(24) Rizzi, F.; Salloum, M.; Marzouk, Y. M.; Xu, R.-G.; Falk, M. L.; Weihs, T. P.; Fritz, G.; Knio, O. M. Bayesian Inference of Atomic Diffusivity in a Binary Ni/Al System Based on Molecular Dynamics. *Multiscale Model. Simul.* **2011**, *9* (1), 486−512.
(25) Jacobson, L. C.; Kirby, R. M.; Molinero, V. How Short Is Too Short for the Interactions of a Water Potential? Exploring the Parameter Space of a Coarse-Grained Water Model Using Uncertainty Quantification. *J. Phys. Chem. B* **2014**, *118* (28), 8190−8202.
(26) Kim, C.; Borodin, O.; Karniadakis, G. E. Quantification of Sampling Uncertainty for Molecular Dynamics Simulation: Time-Dependent Diffusion Coefficient in Simple Fluids. *J. Comput. Phys.* **2015**, *302*, 485−508.
(27) Patrone, P. N.; Dienstfrey, A.; Browning, A. R.; Tucker, S.; Christensen, S. Uncertainty Quantification in Molecular Dynamics Studies of the Glass Transition Temperature. *Polymer* **2016**, *87*, 246−259.
(28) Kim, K.-S.; Han, M. H.; Kim, C.; Li, Z.; Karniadakis, G. E.; Lee, E. K. Nature of Intrinsic Uncertainties in Equilibrium Molecular Dynamics Estimation of Shear Viscosity for Simple and Complex Fluids. *J. Chem. Phys.* **2018**, *149* (4), No. 044510.
(29) Yang, X.; Lei, H.; Gao, P.; Thomas, D. G.; Mobley, D. L.; Baker, N. A. Atomic Radius and Charge Parameter Uncertainty in Biomolecular Solvation Energy Calculations. *J. Chem. Theory Comput.* **2018**, *14* (2), 759−767.
(30) Messerly, R. A.; Knotts, T. A.; Wilding, W. V. Uncertainty Quantification and Propagation of Errors of the Lennard-Jones 12−6 Parameters for n -Alkanes. *J. Chem. Phys.* **2017**, *146* (19), 194110.
(31) Dhaliwal, G.; Nair, P. B.; Singh, C. V. Uncertainty Analysis and Estimation of Robust AIREBO Parameters for Graphene. *Carbon* **2019**, *142*, 300−310.
(32) Mishra, A.; Hong, S.; Rajak, P.; Sheng, C.; Nomura, K.; Kalia, R. K.; Nakano, A.; Vashishta, P. Multiobjective Genetic Training and Uncertainty Quantification of Reactive Force Fields. *npj Comput. Mater.* **2018**, *4* (1), 42.
(33) Hülsmann, M.; Reith, D. SpaGrOW—A Derivative-Free Optimization Scheme for Intermolecular Force Field Parameters Based on Sparse Grid Methods. *Entropy* **2013**, *15* (12), 3640−3687.
(34) Tran, A. V.; Wang, Y. Reliable Molecular Dynamics: Uncertainty Quantification Using Interval Analysis in Molecular Dynamics Simulation. *Comput. Mater. Sci.* **2017**, *127*, 141−160.
(35) Rasmussen, C. E.; Williams, C. K. I. *Gaussian Processes for Machine Learning*; MIT Press, 2006.
(36) Santner, T. J.; Williams, B. J.; Notz, W. I. *The Design and Analysis of Computer Experiments* **2003**, DOI: 10.1007/978-1-4757-3799-8.
(37) Hurle, R. L.; Woolf, L. A. Self-Diffusion in Liquid Acetonitrile under Pressure. *J. Chem. Soc. Faraday Trans. 1 Phys. Chem. Condens. Phases* **1982**, *78* (7), 2233.
(38) Cabaleiro-Lago, E. M.; Rios, M. A. A Potential Function for Intermolecular Interaction in the Acetonitrile Dimer Constructed from Ab Initio Data. *J. Phys. Chem. A* **1997**, *101* (44), 8327−8334.
(39) Grabuleda, X.; Jaime, C.; Kollman, P. A. Molecular Dynamics Simulation Studies of Liquid Acetonitrile: New Six-Site Model. *J. Comput. Chem.* **2000**, *21* (10), 901−908.
(40) Nikitin, A. M.; Lyubartsev, A. P. New Six-Site Acetonitrile Model for Simulations of Liquid Acetonitrile and Its Aqueous Mixtures. *J. Comput. Chem.* **2007**, *28* (12), 2020−2026.
(41) Caleman, C.; van Maaren, P. J.; Hong, M.; Hub, J. S.; Costa, L. T.; van der Spoel, D. Force Field Benchmark of Organic Liquids: Density, Enthalpy of Vaporization, Heat Capacities, Surface Tension, Isothermal Compressibility, Volumetric Expansion Coefficient, and Dielectric Constant. *J. Chem. Theory Comput.* **2012**, *8* (1), 61−74.
(42) Koverga, V. A.; Korsun, O. M.; Kalugin, O. N.; Marekha, B. A.; Idrissi, A. A New Potential Model for Acetonitrile: Insight into the Local Structure Organization. *J. Mol. Liq.* **2017**, *233*, 251−261.
(43) Case, D. A.; Betz, R. M.; Botello-Smith, W.; Cerutti, D. S.; Cheatham, T. E., II; Darden, T. A.; Duke, R. E.; Giese, T. J.; Gohlke, H.; Goetz, A. W.; Homeyer, N.; Janowski, P.; Kaus, J.; Kovalenko, A.; Lee, T.S.; LeGrand, S.; Li, P.; Lin, C.; Luchko, T.; Luo, R.; Madej, B.; Mermelstein, D.; Merz, K.M.; Monard, G.; Nguyen, H.; Nguyen, H.T.; Omelyan, I.; Onufriev, A.; Roe, D.R.; Roitberg, A.; Sagui, C.; Simmerling, C. L. et al. *AMBER 2016*; University of California, San Francisco, 2016.
(44) Vanquelef, E.; Simon, S.; Marquant, G.; Garcia, E.; Klimerak, G.; Delepine, J. C.; Cieplak, P.; Dupradeau, F.-Y. R.E.D. Server: A Web Service for Deriving RESP and ESP Charges and Building Force Field Libraries for New Molecules and Molecular Fragments. *Nucleic Acids Res.* **2011**, *39*, W511−W517.
(45) Wang, J.; Hou, T. Application of Molecular Dynamics Simulations in Molecular Property Prediction. 1. Density and Heat of Vaporization. *J. Chem. Theory Comput.* **2011**, *7* (7), 2151−2165.
(46) Awan, M. A.; Dymond, J. H. Transport Properties of Nonelectrolyte Liquid Mixtures. XI. Mutual Diffusion Coefficients for Toluene + n-Hexane and Toluene + Acetonitrile at Temperatures from 273 to 348 K and at Pressures up to 25 MPa. *Int. J. Thermophys.* **2001**, *22*, 679.
(47) Nelson, R.; Lide, D.; Maryott, A. *Selected Values of Electric Dipole Moments for Molecules in the Gas Phase*; National Standard Reference Data System, 1967.
(48) Antosik, M.; Galka, M.; Malanowski, S. K. Vapor-Liquid Equilibrium for Acetonitrile + Propanenitrile and 1-Pentanamine + 1-Methoxy-2-Propanol. *J. Chem. Eng. Data* **2004**, *49*, 11.
(49) Hansen, C. *Hansen Solubility Parameters: A User's Handbook*, 2nd ed.; CRC Press: Boca Raton, FL, USA, 2007.
(50) Case, D. A.; Cerutti, D. S.; Cheatham, T. E., II; Darden, T. A.; Duke, R. E.; Giese, T. J.; Gohlke, H.; Goetz, A. W.; Greene, D.; Homeyer, N.; Izadi, S.; Kovalenko, A.; Lee, T. S.; LeGrand, S.; Li, P.; Lin, C.; Liu, J.; Luchko, T.; Luo, R.; Mermelstein, D.; Merz, K. M.; Monard, G.; Nguyen, H.; Omelyan, I.; Onufriev, A.; Pan, F.; Qi, R.;

https://doi.org/10.1021/acs.jcim.0c01204
J. Chem. Inf. Model. 2021, 61, 1745−1761

Roe, D. R.; Roitberg, A.; Sagui, C.; Simmerling, C. L.; Botello-Smith, W. M.; Swails, J.; Walker, R. C.; Wang, J.; Wolf, R. M.; Wu, X.; Xiao, L.; York, D.; Kollman, P. A. *AMBER 2017*; University of California, San Francisco, 2017.
(51) Roe, D. R.; Cheatham, T. E., III PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J. Chem. Theory Comput.* **2013**, *9* (7), 3084−3095.
(52) Kleinbaum, D. G.; Kupper, L. L.; Muller, K. E.; Nizam, A. *Applied Regression Analysis and Other Multivariable Methods*, 3rd ed.; Brooks/Cole: Pacific Grove, CA, 1998.
(53) Broadbent, T. A. A.; Einstein, A.; Hertz, H.; Dryden, H. L.; Murnaghan, F. P.; Bateman, H. Investigations on the Theory of the Brownian Movement. *Math. Gaz.* **1957**, *41* (337), 231.
(54) Pranami, G.; Lamm, M. H. Estimating Error in Diffusion Coefficients Derived from Molecular Dynamics Simulations. *J. Chem. Theory Comput.* **2015**, *11* (10), 4586−4592.
(55) Buckingham, A. D. Molecular Quadrupole Moments. *Q. Rev., Chem. Soc.* **1959**, *13* (3), 183.
(56) Belmares, M.; Blanco, M.; Goddard, W. A.; Ross, R. B.; Caldwell, G.; Chou, S. H.; Pham, J.; Olofson, P. M.; Thomas, C. Hildebrand and Hansen Solubility Parameters from Molecular Dynamics with Applications to Electronic Nose Polymer Sensors. *J. Comput. Chem.* **2004**, *25* (15), 1814−1826.
(57) Hanke, F. Sensitivity Analysis and Uncertainty Calculation for Dispersion Corrected Density Functional Theory. *J. Comput. Chem.* **2011**, *32* (7), 1424−1430.
(58) 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* (2), 132−146.
(59) 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* (16), 1623−1641.
(60) Rocha, G. B.; Freire, R. O.; Simas, A. M.; Stewart, J. J. P. RM1: A Reparameterization of AM1 for H, C, N, O, P, S, F, Cl, Br, and I. *J. Comput. Chem.* **2006**, *27* (10), 1101−1111.
(61) Besler, B. H.; Merz, K. M.; Kollman, P. A. Atomic Charges Derived from Semiempirical Methods. *J. Comput. Chem.* **1990**, *11* (4), 431−439.
(62) Mulliken, R. S. Electronic Population Analysis on LCAO−MO Molecular Wave Functions. I. *J. Chem. Phys.* **1955**, *23* (10), 1833−1840.
(63) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. Development and Use of Quantum Mechanical Molecular Models. 76. AM1: A New General Purpose Quantum Mechanical Molecular Model. *J. Am. Chem. Soc.* **1985**, *107* (13), 3902−3909.
(64) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.
(65) Basconi, J. E.; Shirts, M. R. Effects of Temperature Control Algorithms on Transport Properties and Kinetics in Molecular Dynamics Simulations. *J. Chem. Theory Comput.* **2013**, *9* (7), 2887−2899.
(66) Yeh, I.-C.; Hummer, G. System-Size Dependence of Diffusion Coefficients and Viscosities from Molecular Dynamics Simulations with Periodic Boundary Conditions. *J. Phys. Chem. B* **2004**, *108* (40), 15873−15879.
(67) Peerless, J. S.; Milliken, N. J. B.; Oweida, T. J.; Manning, M. D.; Yingling, Y. G. Soft Matter Informatics: Current Progress and Challenges. *Adv. Theory Simulations* **2018**, *0* (0), 1800129.
(68) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Quantum Chemistry Structures and Properties of 134 Kilo Molecules. *Sci. Data* **2014**, *1*, 140022.
(69) Sultan, M. M.; Kiss, G.; Shukla, D.; Pande, V. S. Automatic Selection of Order Parameters in the Analysis of Large Scale Molecular Dynamics Simulations. *J. Chem. Theory Comput.* **2014**, *10* (12), 5217−5223.
(70) Ferguson, A. L. Machine Learning and Data Science in Soft Materials Engineering. *J. Phys.: Condens. Matter* **2018**, *30* (4), No. 043002.

https://doi.org/10.1021/acs.jcim.0c01204
J. Chem. Inf. Model. 2021, 61, 1745−1761