JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

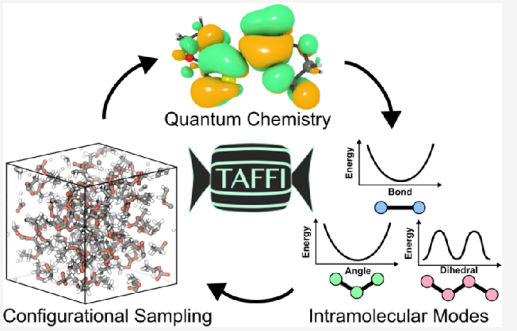
Topology Automated Force-Field Interactions (TAFFI): A Framework for Developing Transferable Force Fields

Bumjoon Seo, Zih-Yu Lin, Qiyuan Zhao, Michael A. Webb, and Brett M. Savoie*

Cite This: https://doi.org/10.1021/acs.jcim.1c00491 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Force-field development has undergone a revolution in the past decade with the proliferation of quantum chemistry based parametrizations and the introduction of machine learning approximations of the atomistic potential energy surface. Nevertheless, transferable force fields with broad coverage of organic chemical space remain necessary for applications in materials and chemical discovery where throughput, consistency, and computational cost are paramount. Here, we introduce a force-field development framework called Topology Automated Force-Field Interactions (TAFFI) for developing transferable force fields of varying complexity against an extensible database of quantum chemistry calculations. TAFFI formalizes the concept of atom typing and makes it the basis for generating systematic training data that maintains a one-to-one correspondence with force-field terms. This feature makes TAFFI arbitrarily extensible to new chemistries while maintaining internal consistency and transferability. As a demonstration of TAFFI, we have developed a fixed-charge force-field, TAFFI-gen, from scratch that includes coverage for common organic functional groups that is comparable to established transferable force fields. The performance of TAFFI-gen was benchmarked against OPLS and GAFF for reproducing several experimental properties of 87 organic liquids. The consistent performance of these force fields, despite their distinct origins, validates the TAFFI framework while also providing evidence of the representability limitations of fixed-charge force fields.



Downloaded via PURDUE UNIV on September 17, 2021 at 22:40:49 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

1. INTRODUCTION

Molecular dynamics (MD) simulations are a ubiquitous tool in contemporary materials and chemical characterization. The development of approximations to the atomistic potential energy surface (PES) has been central to extending MD simulations to address large systems, condensed phases, and long time scales.[1−7] Over the past several decades, many PES approximations (i.e., force-fields) have been implemented, spanning the gamut from relatively simple nonreactive, fixed-charged, and harmonic forms[8−15] to more recent and complex machine-learning based approximations.[16−25] Along this continuum there is an intrinsic trade-off between accuracy and complexity, with fixed-charge force fields being the most economical description but also exhibiting the most limited representability with respect to approximating the PES. Nevertheless, for specific force-field forms, it is still unclear to what extent representability limitations versus limited training data cause errors in the properties simulated by MD. This distinction is crucial because representability limitations are fundamental to the form of the force field,[26−29] whereas errors associated with training data or parametrization protocols can be redressed without increasing the computational cost or complexity of the force field.[30−37] It would thus be desirable to develop a framework capable of parametrizing force fields of varying complexity against common training data such that representability limitations could be established. In the current work, we demonstrate the implementation of such a framework to benchmark a new fixed-charged force-field from scratch, with the long-term goal of flexibly matching force-field complexity to the required accuracy of an MD simulation.

Apart from the specific form of the PES approximation, force fields are also distinguished by whether they are transferable across chemical species or only applicable to specific systems. The latter strategy is in principle more accurate and easier to implement, as transferability imposes additional requirements on the force field that may lead to accuracy trade-offs and also necessarily more training data. In a typical system-specific workflow, a user supplies one or more molecules that they want to simulate. A set of quantum chemistry calculations are performed to generate training data, and a one-off approximate force field is parametrized to the training data.[38−40] However, there are many applications, including molecular discovery and reactive systems, where transferable force fields with general

Received: May 3, 2021

© XXXX American Chemical Society

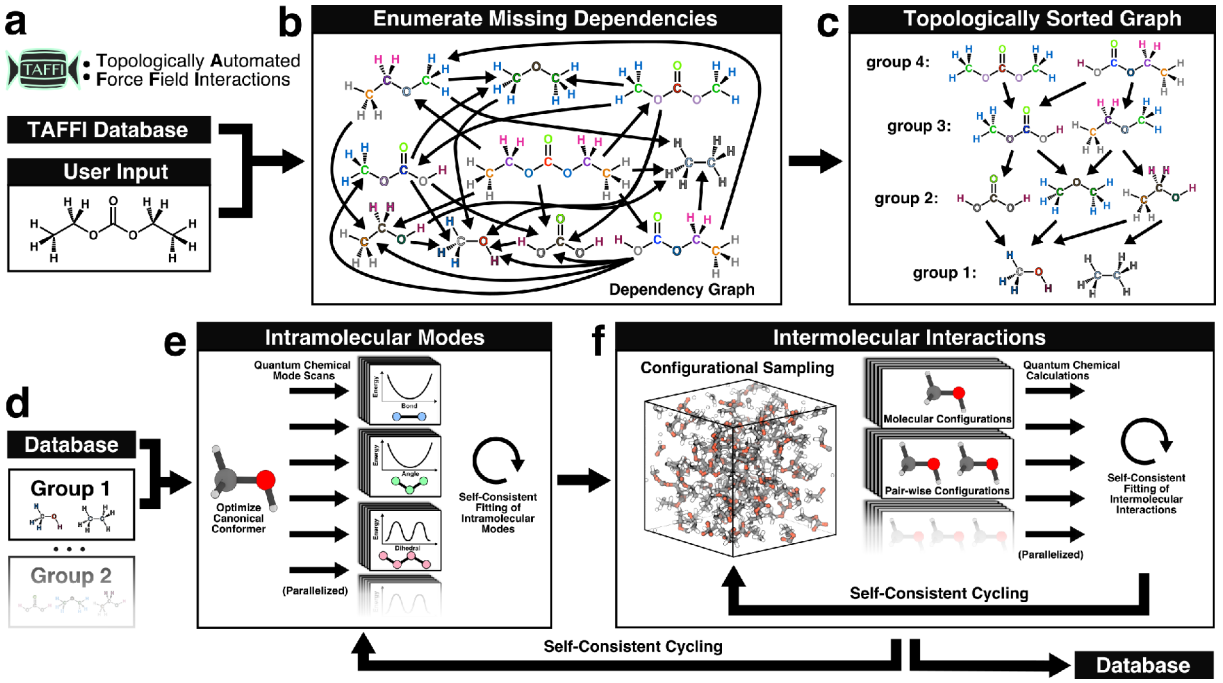


Figure 1. Chemical structure to simulation overview of the TAFFI methodology using diethyl carbonate as an example. (a) Topological criteria are used to determine the necessary parameters for the simulation and identify the missing parameters in the database. (b) An unsorted graph of the molecular dependencies for simulating diethyl carbonate. For simplicity only the dependencies associated with atom types (i.e., not bonds, angles, etc.) are shown. Arrows point toward dependencies, and unique atom types at a bond depth of two are distinctly colored. (c) TAFFI model compound rules produce directed acyclic dependency graphs that can always be linearized to sequentially organize calculations. (d) Hierarchical organization ensures that all dependencies exist prior to attempting the parametrization. (e) Intramolecular modes are parametrized using constrained mode scans from quantum chemistry. (f) Intermolecular interactions are parametrized using quantum chemical calculations on molecular configurations sampled from molecular dynamics. The TAFFI database is updated each cycle, and all quantum chemistry data are retained for future refitting and force-field extension.

applicability are clearly desirable due to the cost of parametrizing a force field from scratch every time a new molecule or material is encountered. Nevertheless, the on-the-fly parametrization concept is potentially still applicable to extending transferable force fields if the associated quantum chemistry data are stored and parametrizations of new molecules are performed in a backward-compatible fashion. This is the approach adopted in the force-field framework developed here.

The most mature transferable force fields are based on the concept of atom types, where the local bonding environment about each atom is used as the basis for transferring force-field terms across recurring bonding motifs. Atom typing reduces the number of parameters required to simulate new molecules, and the concept has precedence in thermodynamic increment theories going back to Pauling. Even in modern machine learning force fields, atom types are often latent variables that are learned during training.[20,23] However, the challenge for transferable force fields has always been with extending them to include coverage for new chemistries.[41−46] Among the specific challenges are generating training data for new chemistries that are consistent with the existing training corpus and performing new parametrizations with backward compatibility with the rest of the force field. For these reasons, the most popular transferable force fields with the largest chemical coverage are built on top of legacy force fields with decades of development (GAFF,[47,48] CGenFF,[49,50] and OPLS-AA[15,42,51,52]). Nevertheless, expanding the coverage of these force fields still typically involves retraining the whole force field. Although not yet fully realized, machine learning force fields present a parallel approach to achieving transferability by simply expanding training data to the point that de facto transferability is achieved. Among the ideas presented here is that these two approaches are not as incompatible as they seem. Specifically, the data generation problem for machine learning force fields is largely shared with the data generation problem for simpler force fields, and a framework that systematically expands a corpus of training data on the basis of new atom types would be advantageous regardless of the specific functional form used for the force field.

The current work addresses the challenges of producing arbitrarily extensible and transferable force fields based on quantum chemistry training data. The presented framework, topology automated force-field interactions (TAFFI),[53−55] accomplishes this by formalizing the concept of atom types using molecular graphs and defining a one-to-one correspondence between force-field parameters and the model compounds used to generate training data. These features are compatible with on-the-fly parametrization of new force-field parameters while maintaining self-consistency and backward compatibility. The result is an extensible force field supported by a continuously growing body of training data that can be fit to flexible force-field forms. In the current work, TAFFI is used to derive a fixed-charge force field (TAFFI-gen) for 87 organic molecules as a case study to illustrate the methodology and benchmark its performance. Additionally, over 2000 distinct force-field terms involving 270 unique atom types for TAFFI-gen are distributed with this work, including coverage for many common organic moieties. Condensed-phase simulation results using TAFFI-gen are compared with the GAFF and OPLS-AA

https://doi.org/10.1021/acs.jcim.1c00491
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

force-fields for the reproduction of a range of experimental liquid properties. The consistent performance of these force fields, despite their distinct origins, validates the TAFFI framework while also providing evidence of the representability limitations of fixed-charge force fields.

2. METHODS

2.1. Methodology Overview. An overview of the three stages of data generation and force-field parametrization within the TAFFI framework is provided here using diethyl carbonate as an example to guide the reader (Figure 1). A detailed description of each step is provided in the subsequent sections 2.2−2.4).

In stage 1 (Figure 1a−c), the atom types and modes associated with the user-supplied molecule(s) are determined (Figure 1a, section 2.2.1), and the model compounds necessary to parametrize any missing terms are generated (Figure 1b, section 2.2.2). Rules based on chemical topology are used for both of these steps to yield a unique dependency graph that can be sorted (Figure 1c, section 2.2.3) to schedule the parametrization calculations. Assuming no previous parameters exist, parametrization (i.e., stages 2 and 3) begins with simple molecules like ethane and methanol, which are at the base of the sorted dependency graph (Figure 1c, group 1), followed sequentially by larger molecules like ethanol, methoxyethane, and dimethyl carbonate. There is a one-to-one mapping between force-field terms and model compounds, such that each term is derived exclusively from the quantum chemistry training data of a single model compound. This one-to-one mapping ensures the extensibility (i.e., the ability to derive parameters for new functional groups) and backward compatibility (i.e., the ability to extend the force field without having to retrain the existing parameters) of force fields developed with TAFFI. At higher levels of the dependency graph, force-field parameters inherited from model compounds at lower levels are held fixed during parametrization.

In stage 2 (Figure 1e), the data generation and force-field parametrizations associated with intramolecular modes are performed. Each model compound is first initialized in a canonical conformation (section 2.3.1) then optimized at the target quantum chemistry level of theory. The optimized geometry is then used as an input for constrained mode scans of unique bonds, angles (section 2.3.2), and dihedrals (section 2.3.3). The intramolecular force-field modes associated with the model compounds are then parametrized to the quantum chemistry mode scans self-consistently with all other intramolecular parameters.

In stage 3 (Figure 1f), the data generation and force-field parametrizations associated with intermolecular interactions are performed. Condensed-phase molecular dynamics are used to sample molecular and pairwise configurations of each model compound (section 1.1 in the SI). Quantum chemistry calculations of electrostatic potentials and interaction energies are performed on the molecular and pairwise configurations, respectively, and serve as the reference data for parametrizing the intermolecular force-field terms (sections 2.4.1−2.4.2). Finally, the intramolecular modes associated with the model compounds are refit to ensure self-consistency with the final intermolecular terms (e.g., partial charges and Lennard-Jones interactions).

Model compounds that are in the same group of the dependency graph are parametrized in parallel during stages 2 and 3. In the current example, the intramolecular and intermolecular terms for methanol and ethane would be derived first, followed by the compounds in group two (Figure 1c), and so forth, until all parameters are obtained that are necessary to simulate diethyl carbonate. The TAFFI database is updated at each step of the process to avoid redundant calculations when parametrizing new molecules. For example, the force-field terms associated with ethanol and ethane are at the base of the dependency graphs of many organic species, but they are only evaluated once and then stored for all future parametrizations.

2.2. Stage 1 - Organization of Calculations. Stage 1 of TAFFI consists of identifying the requisite force-field parameters (Figure 1a, section 2.2.1), generating model compounds for those parametrizations (Figure 1b, section 2.2.2), and ordering the parametrizations to ensure internal consistency (Figure 1c, section 2.2.3). Chemical topology (i.e., the molecular graph) plays a central role in stage 1 for automating the assignment and parametrization of the force field. The chemical topology can be expressed in a computationally useful form as an adjacency matrix, **A**, with dimensions equal to the number of atoms in the molecule and elements defined by

$$A_{ij} = \begin{cases} 1 & \text{if a bond exists between atom } i \text{ and atom } j \\ 0 & \text{if a bond doesn not exist between atom } i \text{ and atom } j \end{cases} \quad (1)$$

Chemical topology is used in stage 1 in three ways: (*i*) the definition of atom types, (*ii*) the definition of the model compounds, and (*iii*) for determining the molecular dependencies and order of calculations.

2.2.1. Definition of Atom Types. In TAFFI, the concept of an atom type is formalized based on the local molecular subgraph about each atom out to a specified number of bonded neighbors, *d*. In turn, bonds, angles, and dihedrals are uniquely defined based on the atom types involved in each mode. For the current work, a bond-depth $d = 2$ has been uniformly used for defining atom types. This choice enables TAFFI-gen to support a greater degree of chemical specificity than is present in other transferable force fields (e.g., a mixture of $d = 1$ and $d = 2$ types are common depending on the available experimental parametrization data) while still being usefully transferable.

Atom typing in TAFFI occurs via breadth-first searches of the molecular graph out to *d* bonds from the atom being typed. This procedure is seeded by querying the row of the adjacency matrix (eq 1) corresponding to the atom being typed and identifying the atoms bonded to it. This process is recursively applied $d - 1$ additional times by reseeding the search with the bonded atoms and excluding the atom seed from the previous generation to avoid backtracking. The subgraphs obtained in this way uniquely define the atom types in the molecule. TAFFI utilizes a string syntax for canonicalizing these subgraphs and expressing them in a machine-readable format. In this syntax, all numbers refer to atomic numbers (i.e., 1 corresponds to hydrogen, and 6 to carbon), open brackets ("[") designate bonds, and closed brackets ("]") designate the end of bonded groups (i.e., either the point at which *d* bonds is reached or at which a branch terminates). A bond is indicated between the atom directly following the open bracket, "[", and the first atom preceding the bracket that is not enclosed by a "]". The atom being typed is always designated first. For example, the atom type of the central carbon atom in ethanol is

https://doi.org/10.1021/acs.jcim.1c00491
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

encoded as [6[6[1][1][1]][8[1]][1][1]], where the first 6 refers to the central carbon atom itself. The [6[1][1][1]] refers to the bonded methyl. The [8[1]] refers to the bonded alcohol, and the final [1][1] specifies the two hydrogens directly bonded to the central carbon. To resolve the ambiguity associated with graph isomorphism, the ordering of branches within each atom type is determined by the mass of the bonded atoms, followed by the mass and number of next-nearest bonded atoms (similar to Cahn-Ingold-Prelog priority rules). Labels for unique bond, angle, and dihedral types are defined based on the atom types involved in each mode (e.g., "[6[6[1][1][1]][8[1]][1][1]] [1[6[8][6][1][1]]]" is the bond type associated with the C−H bond about the central carbon atom in ethanol).

2.2.2. Definition of Model Compounds. In TAFFI, all force-field parameters are derived from a set of algorithmically generated model compounds for which reference quantum chemistry data can be generated. For a given force-field term (e.g., a partial charge, bond type, angle type, etc.), the model compound is defined as the smallest acyclic molecule that both exhibits the required force-field term and conserves the Lewis structure of the associated atom types. For example, as shown in Figure 1b for d = 2, the model compound used to parametrize the partial charges of the terminal alkyl hydrogen, [1[6[6][1][1]]], is ethane, because ethane is the smallest molecule containing that atom type.

Starting with the target compound supplied by the user, these model compounds are generated in two steps. First, all atoms more than d bonds away from the targeted term are removed to form a preliminary compound. For atom types, bond types, and angles, this means truncating all atoms more than d bonds away from any atom involved in the targeted mode. For dihedrals, this means truncating all atoms more than d bonds away from the atoms defining the rotatable bond (i.e., the 2−3 atoms of the dihedral). Second, any undercoordinated atoms that result from this truncation are hydrogenated to a level that is consistent with the hybridization of the subgraph and necessary to form a valid Lewis structure. We emphasize that the resulting model compounds are independent of the specific user-supplied structure that initiated their generation. That is, each force-field term is parametrized using a unique model compound, and the user-supplied structures only play a role in identifying force-field terms in need of parametrization.

This definition of model compounds has two shortcomings that we note here but leave to future work to address. First, this definition leads to ambiguity in cases involving double bonds between nearest and next-nearest neighbors of the atoms associated with the force-field term (e.g., keto−enol tautomers). In these cases, double bonds with the highest bond energy are preferentially formed in the model compound.[55−57] For example, the model compound for the atom type [6[6[1][1][1]][6[8][6]][1][1]] is 2-butanone rather than 1-buten-2-ol (i.e., the ketone as opposed to the alcohol, consistent with the Erlenmeyer rule). This ambiguity could be addressed in the future by introducing bond orders into the atom types (e.g., using distinct symbols for double and triple bonds instead of specifying bonds generically with "[" and "]") such that distinct tautomers would be parametrized to distinct model compounds. Second, this definition leads to force-field terms associated with cyclic structures being derived from data for acyclic model compounds. We note that rings, and similarly conjugated groups, have intrinsically nonlocal contributions to their configurational energy that represents a challenge to the locality assumption of any force field based on atom types. This could be addressed in the future by using model compounds for rings and conjugated subunits that preserve these components, but this is outside of the scope of the current study.

It may happen that the model compounds exhibit new force-field terms that are distinct from the parent molecule. Thus, model compound generation is recursively performed for these new force-field terms until all model compounds have been generated for all unknown terms. Because each model compound is smaller than its parent, this recursion will eventually terminate with small model compounds containing approximately d non-hydrogen atoms. This procedure yields model compounds that are generally small and amenable to high-level quantum chemistry calculations. For example, 90% of model compounds generated in this study had six or fewer heavy atoms (the mode is four), and no model compound had greater than eight heavy atoms (Figure S1).

2.2.3. Definition of the Dependency Graph. The recursive generation of model compounds creates dependencies based on shared force-field terms. To account for these dependencies, it is necessary to order data generation and parametrizations (subsections 2.2−2.3) such that all force-field terms, besides those associated with a given model compound, have been obtained prior to performing each parametrization. These dependencies are enumerated during model compound generation and stored in a dependency graph. The dependency graph has nodes for all model compounds and directed connections between all dependent compounds (e.g., ethanol depends on ethane for the partial charges of atom type [1[6[6][1][1]]], but ethane does not depend on ethanol, Figure 1c). Prior to performing force-field parametrizations, a topological sort is applied to the dependency graph such that no dependencies exist within the same level of the sorted graph. Data generation and parametrization (stages 2 and 3) are then performed beginning with model compounds in the bottom level of this graph and working to the top (i.e., level 1 to level 4 in Figure 1c). This addresses the issue of force-field terms potentially being missing during parametrization because the terms at each level can be directly determined when all of the dependent terms in the lower levels of the dependency graph are known. The algorithm for model compound generation (section 2.2.2) in TAFFI has the important property that dependent model compounds are always identical to or smaller than their parent molecule. Consequently, the dependency graph for any molecule is directed and acyclic, and it is always possible to order calculations such that all dependencies exist at the time of parametrization.

2.2.4. Force-Field Expression. While the particular force-field expression used for fitting the data in the TAFFI database is flexible, this choice does guide which calculations are performed on the model compounds in the subsequent stages. For the current study, we employ the following fixed-charge functional form:

https://doi.org/10.1021/acs.jcim.1c00491
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

$$V_{FF} = \sum_{bonds} k_r(r - r_0)^2 + \sum_{angles} k_\theta(\theta - \theta_0)^2 + \sum_{dihedrals} \sum_{i=1}^{4} \frac{1}{2} V_i(1 + (-1)^{i+1}\cos(i\phi)) + \sum_{i>j} \left\{ \frac{q_i q_j e^2}{4\pi\epsilon_0 r_{ij}} + 4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right]\right\} \quad (2)$$

where k_r and r_0 are a bond-specific force constant and equilibrium displacement, respectively. k_θ and θ_0 are an angle-specific force constant and equilibrium angle, respectively. The V_i terms are dihedral-specific Fourier coefficients. r_{ij} are the interatomic separations. q_i are the atomic partial charges. e is the elementary charge, and ϵ_{ij} and σ_{ij} are the Lennard-Jones (LJ) parameters for each pairwise interaction. The summation for the Lennard-Jones and Coulomb potentials runs over all intermolecular atomic pairs and all intramolecular atomic pairs separated by more than three bonds (i.e., 1−4 intramolecular interactions are excluded). All dihedrals that rotate about double bonds are modeled as invertible harmonic modes by only using the $i = 2$ term in the dihedral expression. These functional forms are largely standardized in general force fields and are broadly implemented in existing MD packages, which makes them an obvious starting point for this initial benchmark study.

2.3. Stage 2 - Intramolecular Parameterizations. Stage 2 consists of generating reference quantum chemistry data and performing force-field parametrizations related to the intramolecular force-field parameters (sections 2.3.1−2.3.3). Parameterizing intramolecular modes is a prerequisite for generating reference data for intermolecular force-field terms in stage 3. Thus, stage 2 occurs first to yield a provisional force field, with the final intramolecular force-field terms refit after the stage 3 intermolecular terms.

2.3.1. Conformer Generation. The first step in generating reference quantum chemistry data for fitting intramolecular force-field terms is generating an optimized geometry for the model compounds. Here, all model compounds are initialized as the conformer with *trans* relationships for all backbone dihedrals (i.e., the all-*trans* conformer). This choice was motivated by the observation that the all-*trans* conformer is generally well-conditioned for converging sterically crowded dihedrals; however, more sophisticated conformer sampling schemes could be applied to generate a conformer closer to the global minimum.[58] The all-*trans* conformer is generated by (*i*) identifying the atoms belonging to the longest connected path in the molecular graph (i.e., the molecular backbone), (*ii*) aligning the backbone dihedrals in all-trans geometries, and (*iii*) repeating with the remaining branches of the molecular graph until all nonterminal dihedrals exhibit *trans* relationships. Since the all-*trans* designation leaves the conformation of terminal dihedrals ambiguous (e.g., the dihedral involving chlorine in 1-chlorobutane), the conformation of end groups is determined by explicitly generating and optimizing all end group conformers by steepest descent using the Universal Force Field (UFF),[59] then using the lowest energy conformer as the input structure for quantum chemical geometry optimization. This procedure yields a deterministic conformer and initial geometry for each model compound.

2.3.2. Parameterization of Harmonic Modes. Bonds, angles, and dihedrals about double bonds are modeled here with harmonic forms (eq 2). In cases where a model compound has multiple resonance structures, if a dihedral has a double bond in any resonance structure, then it is modeled as a harmonic mode (e.g., all dihedrals in benzene are considered harmonic in TAFFI-gen). All harmonic modes are self-consistently fit to constrained quantum chemistry mode scans. Bond mode scans consist of compression and extension by 0.1 pm about the optimized bond length in steps of 0.02 pm. Angle mode scans consist of compression and expansion by 0.5° about the optimized angle in steps of 0.1°. Harmonic dihedral scans consist of compression and expansion by 0.5° about the optimized dihedral angle in steps of 0.1°. At each scan configuration, geometry optimizations are performed with the mode being parametrized constrained to a fixed value while optimizing all remaining degrees of freedom.

The harmonic modes associated with the model compounds are parametrized to minimize the following objective function:

$$\chi^2_{harm} = \sum_i \left(E_{QC,i} - \sum_{\nu_j \in local} V_{FF}(\nu_j) \right)^2 \quad (3)$$

where the index i runs over all scanned configurations. $E_{QC,i}$ is the single-point energy of configuration i relative to the minimum-energy configuration. The index j runs over all bonds, angles, and harmonic dihedrals that share an atom with the scanned mode (i.e., "local" modes), and $V_{FF}(\nu_j)$ is the force-field energy of mode ν_j in configuration i. The self-consistent fit over all local modes is performed because the force-field terms are generally not linearly independent. All fits are performed using the limited-memory Broyden−Fletcher−Goldfarb−Shanno algorithm with bound constraints (L-BFGS-B) to limit the fit variables to positive values. Initial guesses for the force constant and equilibrium displacement for each scanned mode are obtained by a linear least-squares fit to the quantum chemistry single-point energies with respect to the mode being fit. This procedure is repeated until reaching self-consistency among all intramolecular modes. During these fits, only the force-field terms associated with the model compound are parametrized, and any terms inherited from model compounds lower in the dependency graph are held fixed.

2.3.3. Parameterization of Flexible Dihedral Potentials. Dihedrals that rotate about single and triple bonds are modeled by TAFFI-gen with a truncated Fourier series. All flexible dihedrals are self-consistently fit to constrained quantum chemistry scans from [0, 2π) and [0,−2), in 5° steps, about each rotatable bond. During each quantum chemistry scan, the dihedral being parametrized is constrained to a fixed value while optimizing all remaining degrees of freedom. In the case where multiple dihedrals exist about the same bond, only the dihedral involving the heaviest atoms—or secondarily, the longest chain—is explicitly constrained during the scan. These scans are performed sequentially (i.e., the optimized geometry at each dihedral angle is used as the initial guess for the next step). Two scans are performed in opposite directions to mitigate the path-dependence of the scan (e.g., this can be important for sterically crowded dihedrals) and the lowest energy union of the two scans is used as reference data for the parametrization.

For parametrizing the constrained dihedrals, it is necessary to also perform dihedral scans using the force field so that the dihedrals can be fit to the residual energy difference between the two scans (eq 4). The geometries from the lowest energy

https://doi.org/10.1021/acs.jcim.1c00491
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

union of the quantum chemistry scans are used as the initial guess for constrained optimizations at the force-field level, except that all atoms more than d bonds away from the 2−3 atoms of the scanned dihedral are removed (these atoms are removed since they lie outside the subgraphs defining the dihedrals being parametrized). During these optimizations, the dihedrals being fit are constrained to the scanned value, but no other constraints/restraints are used. We note that using model compounds facilitates these scans, since typically only the dihedrals about the scanned bond are unknown at the time of the parametrization, which simplifies the situation compared to trying to perform force-field scans for large molecules with many simultaneously undetermined dihedrals. The force-field optimizations are performed in LAMMPS[60] using the Hessian-free truncated Newton algorithm.

The Fourier coefficients are fit to minimize the residual between the quantum chemistry and force-field potentials for the constrained dihedral rotation according to the following objective function:

$$\chi^2_{\text{Fourier}} = \sum_i \left(E_{\text{QC},i} - \sum_{\nu_j \notin \text{fit}} E_{\text{FF},i}(\nu_j) - \sum_{\nu_j \in \text{fit}} \sum_{k=1}^{4} \frac{1}{2} V_{j,k} \left(1 + (-1)^{k+1} \cos(k\phi_{i,j})\right) \right)^2 + \omega_{\text{L2}} N_{\text{fit}}^{-1} \sum_{i,j \in \text{fit}}^{N_{\text{fit}}} V_{i,j}^2 \tag{4}$$

where the index i runs over all scan configurations. $E_{\text{QC},i}$ is the single-point energy of the configuration. The second summation runs over all force-field terms that are not being fit (i.e., bonds, angles, unscanned dihedrals, electrostatics, and Lennard-Jones terms). The third summation runs over all dihedrals that share the scanned bond (i.e., $\nu_j \in$ fit). $V_{j,k}$ are the dihedral-specific force constants, and $\phi_{i,j}$ is the angle of dihedral j in configuration i. The last summation is an L2 regularization of the average magnitude of the dihedral fit coefficients that reduces overfitting to noisy data. ω_{L2} is set to 0.1% of the range of the fit values (i.e., the difference between $E_{\text{QC},i}$ and the second summation in eq 4). All fits are performed using the L-BFGS-B algorithm with bound constraints limiting the magnitude of the dihedral fit coefficients to 200% of the range of the fit potential.

During stage 2, the Lennard-Jones parameters and partial charges are not yet determined, so UFF parameters and approximate partial charges fit to the optimized geometry of the model compound (section 2.4.1) are used as an approximation. After stage 3, all intramolecular parameters are refit with updated partial charges and Lennard-Jones parameters using the same procedure.

2.4. Stage 3 - Intermolecular Parameterizations. Stage 3 consists of generating reference quantum chemistry data and performing force-field parametrizations related to the intermolecular force-field parameters (sections 2.4.1−2.4.2). Configurational sampling is critical for generating reference data for intermolecular terms, which requires stage 3 to occur after a preliminary force field is obtained from stage 2. After configurational sampling (section 1.1 in the SI), quantum chemistry calculations on molecular and pairwise configurations are used to parametrize the partial charges (section 2.4.1) and Lennard-Jones parameters (section 2.4.2), respectively.

2.4.1. Parameterization of Partial Charges. The electric potential calculated on a grid about each molecule in each sampled configuration (see section 1.1 in the SI) is used as reference data for the partial charge parametrization. The partial charges are fit to minimize the following objective function:

$$\chi^2_{\text{q}} = \sum_{\text{s}}^{N_{\text{samples}}} \left(\omega_{\text{pot}} N_{\text{pot}}^{-1} \sum_{i}^{N_{\text{pot}}} (V_{\text{QC},i} - V_{\text{FF},i})^2 + \omega_{\text{D}} \sum_{i}^{3} (D_{\text{QC},i} - D_{\text{FF},i})^2 + \omega_{\text{T}} \left(\sum_{i}^{N_{\text{atoms}}} q_i - q_{\text{T}} \right)^2 \right) \tag{5}$$

where the first summation (s) is over the sampled configurations. The second summation is over the squared deviations of the force-field description ($V_{\text{FF,i}}$) from the reference electric potential ($V_{\text{QC,i}}$) as calculated on the N_{pot} grid points. The third summation corresponds to the element-wise deviations of the force-field description ($D_{\text{FF,i}}$) from the reference molecular dipole (D_{QC}), and the fourth summation corresponds to deviations from the total molecular charge (q_{T}). ω_{pot}, ω_{D}, and ω_{T} are weighting coefficients for penalizing the electric potential, dipole, and total charge deviations, respectively. The s index is implied in all terms, but dropped for clarity. Partial charges (q_{i}) are fit using $\omega_{\text{pot}} = 1.0$, $\omega_{\text{D}} = 0.1$, an $\omega_{\text{T}} = 1.0$, specified in inverse atomic units. As implemented in ORCA v.4.1.2, the electric potential is calculated on a cubic grid with a grid spacing of 0.3 Å, and any grid points further than 2.8 Å from any atom or within the COSMO radius of any atom are discarded.[61−63]

The partial charges are fit in two steps. First, eq 5 is minimized while constraining polar atoms of an identical TAFFI atom type to have the same partial charge. Polar atoms are considered to be any non-hydrogen atoms besides carbon and hydrogen atoms that are not bonded to carbon. A second fit is then performed by minimizing eq 5 while holding the partial charges for the polar atom types constant and constraining all nonpolar atoms of the same type to have the same partial charge. This two step procedure is meant to improve the accuracy of the electric potential near the polar atoms and is similar to the original RESP algorithm[64] and recent variants.[65,66] This procedure differs from the original RESP algorithm in (*i*) the form of the objective function and (*ii*) the use of 200 configurations rather than a single configuration. We note that using model compounds tends to reduce the under-determined nature of grid-based partial charge fitting, since only a few atoms are being fit for any model compound and buried atoms are usually parametrized to separate model compounds from their bonded atoms. Fitting to multiple configurations also tends to reduce the magnitude of the partial charges. Together, these factors alleviated the need for the heuristic hyperbolic restraint used in RESP. Partial charge fits are performed using the BFGS algorithm.

2.4.2. Parameterization of Pairwise Interactions. Counterpoise corrected interaction energies (IE) of the sampled pairwise configurations (see section 1.1 in the SI) are used as reference data for the Lennard-Jones parametrization. The

https://doi.org/10.1021/acs.jcim.1c00491
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Lennard-Jones parameters are fit to minimize the following objective function:

$$\chi^2_{\mathrm{LJ}} = \omega_{\mathrm{IE}} N_{\mathrm{IE}}^{-1} \sum_{i}^{N_{\mathrm{IE}}} (\mathrm{IE}_{\mathrm{QC},i} - \mathrm{IE}_{\mathrm{FF},i})^2 + \omega_{\epsilon} N_{\epsilon}^{-1} \sum_{i}^{N_{\epsilon}} (\epsilon_{\mathrm{UFF},i} - \epsilon_{\mathrm{FF},i})^2 + \omega_{\sigma} N_{\sigma}^{-1} \sum_{i}^{N_{\sigma}} (\sigma_{\mathrm{UFF},i} - \sigma_{\mathrm{FF},i})^2 \quad (6)$$

where the first summation corresponds to squared deviations of the force-field interaction energy (IE_{FF}) from the counterpoise corrected interaction energy (IE_{QC}) over all N_{IE} pairwise samples. The second summation corresponds to the L2 regularization of the Lennard-Jones energy parameters ($\epsilon_{\mathrm{FF},i}$) with respect to the UFF reference values ($\epsilon_{\mathrm{UFF},i}$), and the third summation corresponds to the L2 regularization of the Lennard-Jones atomic radii ($\sigma_{\mathrm{FF},i}$) with respect to the UFF reference values ($\sigma_{\mathrm{UFF},i}$). The latter terms in the objective function are included to avoid extreme values in ϵ and σ that can occur when using only a least-squares objective function. The Lennard-Jones parameters are fit using ω_{IE} = 1. 0 mol/kcal, ω_{ϵ} = 1.0 mol/kcal, and ω_{σ} = 0.1 Å^{-1}. A comparison of the interaction energies calculated at the UFF level with the regularized and unregularized TAFFI-gen parameters (Figure S2) confirms that the regularization terms have only a small effect on the reproduction of the interaction energies. The interaction energies are calculated in the force-field description as the sum of intermolecular Lennard-Jones and electrostatic terms between the molecules in each configuration. The partial charges are held fixed during the fitting of the Lennard-Jones parameters. Any configurations with unstable interaction energies (i.e., $\mathrm{IE}_{\mathrm{QC}}$ > 0 kcal/mol) are excluded from the fit. The inclusion of larger molecular clusters for training was considered but decided against after observing the limited ability of the LJ potential to describe the distribution of dimer configurations (*vide infra*). For more sophisticated intermolecular potentials, the inclusion of larger clusters may be justified. Lennard-Jones fits are performed using the L-BFGS-B algorithm.

2.5. Data set Description. LAMMPS[60] and ORCA v.4.1.2[61] were used to perform the molecular dynamics simulations and quantum chemistry calculations, respectively, associated with reference data generation. All quantum chemistry calculations were performed at the ωB97X-D3[67,68]/def2-TZVP[69,70] level of theory for training the version of TAFFI-gen reported here.

To assess the performance of TAFFI-gen, we present a benchmark on the data set of small organic molecules introduced by Caleman et al. for GAFF and OPLS-AA.[71] The original MD-based predictions of liquid properties by Caleman included 147 molecules in their benchmark set. In the current study, we have excluded ring, nitro, and phosphate containing compounds, as they require a more sophisticated treatment of atom types and model compounds that is beyond the scope of the current work. After these exclusions, a total of 87 molecules at 146 distinct state points (i.e., multiple temperatures per molecule where included by Caleman et al.) are in the presented benchmark. A list of all benchmark compounds and individual property predictions is given in the Supporting Information of this work (Table S3).

Six properties were calculated from the MD trajectories: the density, enthalpy of vaporization, static dielectric constant, volumetric thermal expansion coefficient, isothermal compressibility, and quantum-corrected heat capacity at constant volume. Following the reference benchmark by Caleman, three types of MD simulations were performed to extract these properties. Gas phase simulations were run to obtain the expected potential energy per molecule in the gas phase for the enthalpy of vaporization calculation. Relatively long liquid phase simulations (i.e., 10 ns) in the NPT ensemble were run to compute all properties other than the heat capacity. Short liquid phase simulations (i.e., 100 ps) were run in the NVT ensemble with high sampling frequency to calculate the constant volume heat capacity using the two-phase method.[72,73] Details of the simulation and analysis methods are described in the SI. We note that the dielectric constants of methanoic acid have been omitted in analysis due to lack of convergence, which is revisited in the discussion. Besides this case, all available experimental data in ref 71 for the benchmark molecules have been included for comparison.

Finally, four error measures are reported for comparing the results for TAFFI-gen against experimental data and the other force fields (eqs 7−10). The mean absolute difference (MAD) is calculated as

$$\mathrm{MAD} = \frac{1}{N} \sum_{i}^{N} |x_{i,\mathrm{sim}} - x_{i,\mathrm{ref}}| \quad (7)$$

where N is the total number of data points, $x_{\mathrm{i,sim}}$ is the simulated value for each data point and $x_{i,\mathrm{ref}}$ is the corresponding reference value (DFT calculated value or experimental value). The mean signed difference (MSD) is calculated as

$$\mathrm{MSD} = \frac{1}{N} \sum_{i}^{N} (x_{i,\mathrm{sim}} - x_{i,\mathrm{ref}}) \quad (8)$$

with positive values indicating an average overestimation of the value by simulations. The mean absolute percent difference (MAPD) is calculated as

$$\mathrm{MAPD} = \frac{100}{N} \sum_{i}^{N} \frac{|x_{i,\mathrm{sim}} - x_{i,\mathrm{ref}}|}{x_{i,\mathrm{ref}}} \quad (9)$$

The mean signed percent difference (MSPD) is calculated as

$$\mathrm{MSPD} = \frac{100}{N} \sum_{i}^{N} \frac{x_{i,\mathrm{sim}} - x_{i,\mathrm{ref}}}{x_{i,\mathrm{ref}}} \quad (10)$$

We note that MAD and MSD are more sensitive to the large magnitude samples in the data set, whose deviations tend to be correspondingly larger than the small magnitude samples. MAPD and MSPD are more sensitive to the small magnitude samples in the data set.

The authors declare that the data supporting the findings of this study are available within the paper and its Supporting Information files. Python scripts that can be used to perform TAFFI atom typing, and TAFFI-gen parameter assignments are available through GitHub under the GNU GPL-3.0 License (https://github.com/bsavoie/TAFFI). A sample of the parametrization data has been uploaded to Zenodo with a persistent DOI (10.5281/zenodo.5164712).

3. RESULTS AND DISCUSSION

TAFFI-gen is parametrized to DFT reference data for small model compounds. Thus, the errors in TAFFI-gen predictions

https://doi.org/10.1021/acs.jcim.1c00491
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

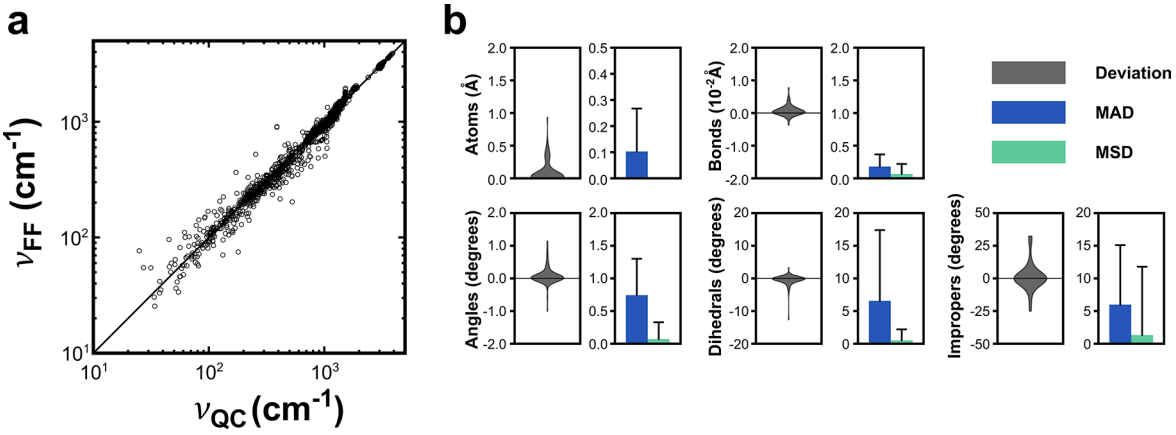


Figure 2. (a) Comparison of the TAFFI-gen and ωB97X-D3/def2-TZVP (DFT) normal-mode frequencies for the benchmark compounds. (b) The distributions of signed deviations ($x_{TAFFI} - x_{DFT}$) for selected structural features over all benchmarked compounds are shown in each violin plot. The distribution of atom deviations corresponds to the MAD in the atomic positions after alignment of the TAFFI-gen and DFT optimized structures. The other distributions correspond to the signed differences in the bond lengths, bending angles, dihedral angles, and improper angles in the optimized TAFFI-gen geometries and in the optimized DFT geometries. The mean and standard deviation of the mean absolute differences (MAD, blue) and mean signed differences (MSD, green) for each quantity calculated across all benchmark compounds are shown in the bar plots. Impropers are only included for 3-coordinate atoms.

can be decomposed into errors associated with the underlying DFT parametrization data and representability errors associated with the limited functional form of the force field. Regarding the first source of error, the dispersion-corrected range-separated functional used here is among the highest performing in benchmarks of conformational energetics and cluster interactions for organic species.[67,74−78] Nevertheless, even modern functionals have documented deficiencies for aqueous solutions and reaction barriers that would require higher fidelity training data for models of water or reactive force fields, which are beyond the present scope. Thus, for the current study, we acknowledge this potential source of error but consider it negligible in comparison with the representability errors associated with the simple functional form of the force field.

To quantify the magnitude of errors associated with the functional form of the force field, we have compared the TAFFI-gen predictions for normal modes and optimized geometries against DFT results for the benchmark compounds (Figure 2). Comparing the normal-mode frequencies provides a measure of the accuracy of forces in the force-field representation (Figure 2a). We observe a MAD of 52 cm^{-1} and MAPD of 6%, which is comparable to nontransferable quantum chemistry derived force fields using more complex forms.[39,40] In addition, the correlation coefficients between the DFT and force-field spectra (Pearson r = 0.377 ± 0.195 and Spearman ρ = 0.797 ± 0.098) exhibit similar magnitude to those recently reported between OPLS, GAFF, and experimental data.[79] These results suggest that in general TAFFI-gen exhibits accurate force behavior near equilibrium structures. Notably, the largest percent deviations are associated with low frequency modes (<1000 cm^{-1}), which is expected given the lack of explicit coupling between dihedral terms and the exclusion of improper modes in the current force field.

The predicted equilibrium structures of the benchmark compounds provide a second point of comparison between TAFFI-gen and the reference DFT level of theory (Figure 2b). These comparisons are performed by optimizing the compounds at the DFT and force-field levels starting from the same all-trans conformer, then aligning the structures via the Kabsh algorithm.[80,81] First, we observe that the deviations of atomic positions (MAD = 0.1 Å), bonds lengths (MAD = 0.002 Å), and bending angles (MAD = 0.7°) are all extremely small on a per molecule basis, which confirms the generally excellent agreement between TAFFI-gen and DFT for local structural features. Larger deviations are observed for proper dihedrals (MAD = 7°) and improper dihedrals (MAD = 6°). From the distribution of proper dihedral deviations, it is evident that these errors are driven by a small number of outliers that adopt distinct conformers at the TAFFI-gen level upon geometry optimization. In particular, terminal methyl groups proximate to esters and amides tend to twist relative to DFT predictions (Figure S4), which occurs for methyl acetate (dihedral MAD = 33°), methyl formate (36°), acetyl acetate (37°), N,N-dimethylacetamide (34°), *N*-methylformamide (36°), and *N*-methylacetamide (45°). In contrast, the errors in improper dihedrals appear to be systematic, with a relatively large standard deviation in MAD across the reference structures (5.99°). This is a consequence of not explicitly including improper modes in the force-field form. The errors in impropers are intuitively largest for planar conjugated units. For example, the largest error is exhibited by the improper defined about the carbonyl in 2,6-dimethylheptan-4-one, where TAFFI-gen exhibits an improper angle of 32° in contrast to 0° predicted by DFT. The optimized geometries for DFT and TAFFI-gen for the molecules with large MADs are compared in Figure S4. Although we have focused on the largest error cases to illustrate the limitations of the common force-field form employed here, the overall mean performance is nevertheless very accurate (Table 1). Namely, the overwhelming majority of structural features are quantitatively reproduced by TAFFI-gen, and the cases where incorrect conformers are stabilized are rare and isolated to the periphery of the molecules.

We note that the above comparison has been performed for the benchmark molecules and not for the model compounds actually used for TAFFI-gen parametrization. Figure S3 presents the analogous comparisons with DFT results for model compounds only, which show very similar deviations

https://doi.org/10.1021/acs.jcim.1c00491
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 1. Summary of TAFFI-gen Performance in Reproducing the DFT Normal Mode Frequencies and Structural Features of the 87 Molecules in the Benchmark Set

structure	MAD	MSD	N	molecules
normal modes (cm^{-1})	52.1	−14.7	2908	87
atoms (Å)	0.103	-[a]	1151	87
bonds (Å)	0.00181	0.000665	1064	87
angles (deg)	0.743	0.0714	1842	87
dihedrals (deg)	6.56	−0.520	1919	80
impropers (deg)	5.99	1.32	58	40

[a]Trivially zero due to structural alignment.

compared with the benchmark structures. The similar errors observed between these two cases provide evidence that the $d = 2$ atom typing of TAFFI-gen leads to excellent transferability between model compounds and larger molecules for structural features, while the limited representability of the force field is the main source of error with respect to the DFT reference data.

MD simulations of six liquid properties were performed to establish the performance of TAFFI-gen relative to OPLS-AA and GAFF in predicting experimental liquid properties (Figure 3). These properties include the density (ρ), heat of vaporization (ΔH_{vap}), static dielectric constant (ε), volumetric thermal expansion coefficient (α_P), isothermal compressibility (κ_T), and quantum-corrected heat capacity at constant volume (c_v) for the 87 molecules in the current benchmark. We note that among the liquid properties, ρ, ΔH_{vap}, and ε have historically been utilized as part of the OPLS-AA and GAFF parametrizations, whereas for TAFFI-gen these data are not utilized in any way and represent a test for the force field. Summary statistics across the benchmark are presented in Table 2, and the TAFFI-gen predictions for individual simulation conditions are presented in Table S3.

The summary error statistics calculated across all systems for each force field illustrate a similar accuracy (and inaccuracy) of the three force fields for the various properties. Although some specific differences occur, which are discussed below, it is perhaps surprising that the mean performance is so consistent despite the distinct parametrization protocols and training data for the three force fields. For instance, all of the force fields exhibit relatively small errors for ρ and c_v, large systematic errors for ε (e.g., $R^2 < 30\%$ in all cases), and high correlation but large variances for ΔH_{vap}, α_P, and κ_T. These trends can be rationalized by the common functional form of these force fields. For instance, the Lennard-Jones potential is capable of recapitulating the molecular volume, which is the leading order contribution to density, but is an approximate description of van der Waals interactions which significantly contribute to ΔH_{vap}. Similarly, a fixed point-charge model is an aggressive simplification of electrostatic interactions, which explains the poor dielectric results in all cases, and also contributes to the high variances of the other fluctuation-based condensed phase properties. The heat capacity is well reproduced in all cases, which is also consistent with the generally accurate reproduction of local configurational energetics (i.e., bond, angle, and to a lesser degree dihedral terms) in these force fields. Thus, despite their independent reference data, the force fields exhibit similar average prediction behaviors that reflect the representability limitations of the functional form of the force field. The approximate treatment of intermolecular interactions, in particular, leads to shared trade-offs in reproducing thermodynamic properties. In the case of TAFFI-gen, the limitations of the LJ potential are clear when

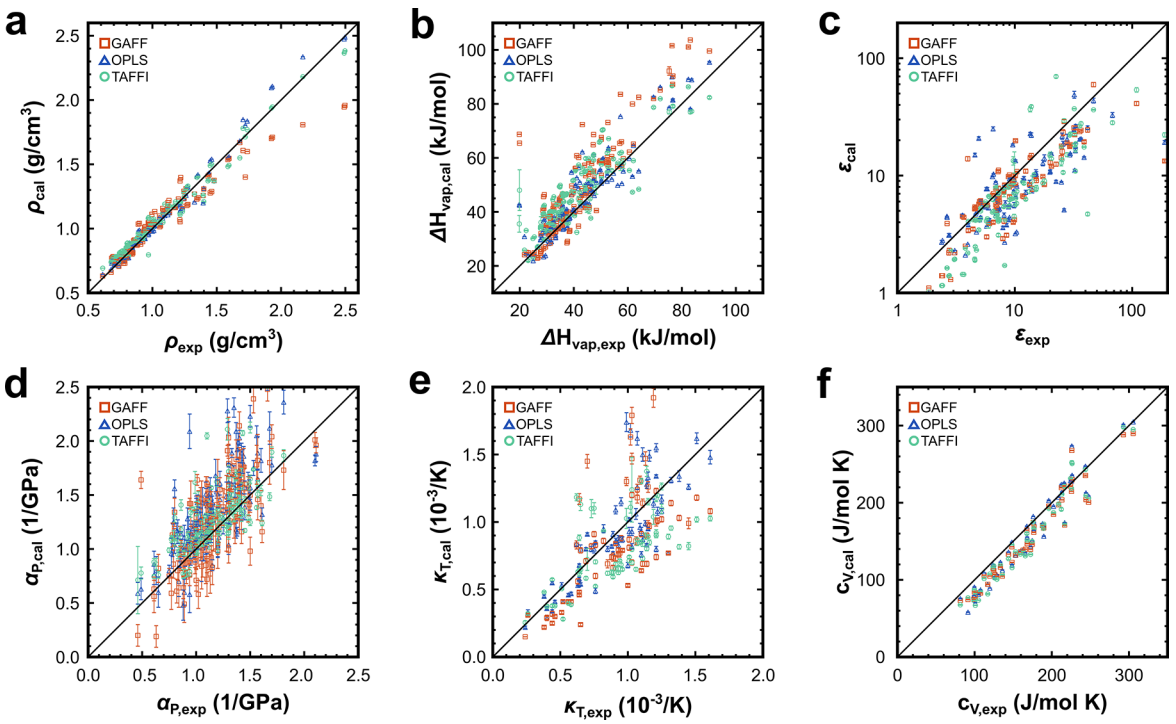


Figure 3. Comparisons of the experimental values for (a) densities, (b) enthalpies of vaporization, (c) static dielectric constants, (d) volumetric thermal expansion coefficients, (e) isothermal compressibilities, and (f) quantum-corrected heat capacities at constant volume with those predicted by GAFF (red), OPLS-AA (blue), and TAFFI-gen (green).

https://doi.org/10.1021/acs.jcim.1c00491
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 2. Comparison of the Errors in the Liquid Properties for the GAFF, OPLS-AA, and TAFFI-gen Force Fields[a]

force field	MAD[b]	MSD[b]	MAPD[c]	MSPD[c]	RMSD[b]	R^{2c}	N
ρ (g/cm^3)							
GAFF	0.0590	−0.0060	5.0970	0.8421	1.00	94.17	145
OPLS-AA	0.0311	0.0114	2.9424	1.2046	0.48	98.24	145
TAFFI-gen	0.0484	0.0231	5.0971	3.2570	0.58	97.94	145
ΔH_{vap} (kJ/mol)							
GAFF	7.7691	6.4625	19.7032	16.0226	11.10	78.69	143
OPLS-AA	4.3424	2.9003	11.2727	7.7738	6.18	87.17	143
TAFFI-gen	7.3489	5.9987	19.5204	16.9972	8.89	78.62	143
ε							
GAFF	6.1100	−4.9042	30.1701	−13.9654	19.90	29.84	97
OPLS-AA	6.9686	−4.7846	40.7308	−9.5976	18.67	25.60	103
TAFFI-gen	7.2708	−5.2487	37.8468	−25.1088	19.03	30.39	113
α_P (10^{-3}/K)							
GAFF	0.2411	0.1124	21.9688	9.5985	0.34	50.00	140
OPLS-AA	0.2528	0.1906	22.2424	16.9217	0.33	54.80	140
TAFFI-gen	0.1821	0.1308	16.5202	12.5512	0.27	58.43	140
κ_T (1/GPa)							
GAFF	0.2475	−0.0577	27.6643	−6.8676	0.31	43.49	73
OPLS-AA	0.1875	0.0273	20.3002	2.8656	0.29	52.02	73
TAFFI-gen	0.2584	−0.0811	27.5593	−5.7311	0.38	22.18	73
c_V (J/mol K)							
GAFF	17.7962	−15.4722	11.5785	−10.5375	21.01	93.89	50
OPLS-AA	16.5314	−12.0042	11.0421	−8.9901	20.48	93.51	50
TAFFI-gen	18.4626	−15.7177	12.3048	−11.1073	21.68	94.21	50

[a]The mean absolute difference (MAD), mean signed difference (MSD), mean absolute percent difference (MAPD), the mean signed percent difference (MSPD), the root mean square deviation (RMSD) from experimental values, and the correlation coefficient R^2 are reported. [b]In indicated units. [c]In units of %.

considering the error distributions in the interaction energies compared with the training data (Figure S2). Specifically, while TAFFI-gen exhibits excellent reproduction of the mean interaction energies (MSE of −0.09 kcal/mol for the model compound training data), the error residuals exhibit very long tails (kurtosis = 20.25), which is clear evidence of representability limitations associated with the pairwise fixed-charge form of the force field. Thus, these potentials should be considered averaged representations of a much more detailed orientational interaction energy.

Although our interpretation of the similar mean performance of the three force fields is that representability limitations dominate the general behaviors, this does not exclude some specific cases being the result of inaccurate parametrizations. For instance, the efforts of the Open Force-Field Consortium have highlighted many cases where additional accuracy can be squeezed from fixed-charge force fields by refining specific parameters.[15,46,82−85] Likewise, the fact that OPLS generally outperforms the other force fields illustrates that the specific force-field terms for TAFFI-gen might be improved by tuning the parametrization hyperparameters or supplementing the training data.

To facilitate a more fine-grained comparison between the force fields, the MAPD with respect to each liquid property is presented on a per functional group basis in Figure 4. Molecules were included in a category if they exhibited the specified functional group; thus, some molecules are included in multiple categories. We have also combined similar functional groups in some cases to avoid scarce or empty categories. We note that experimental data are not available for all compounds for all properties, thus the number of compounds in each category varies across properties, and bars have been omitted for cases where less than three data points were available. We note that a large number of distinct outliers are observable for GAFF that have previously been discussed by Caleman et al. and are thus not further remarked on here.

There are several informative outliers observed for all of the force fields that shed further light on representability limitations. For example, all of the force fields exhibit underestimated dielectric constants for the amides, which suggests the need for a more complex electrostatic description (e.g., via inclusion of off-site point charges or polarizable terms) to accurately account for the large molecular dipoles and strong hydrogen bonding associated with this functional group.[86] In the case of TAFFI-gen, poor reproduction of the planar amide geometry owing to the omission of improper dihedrals may also affect the dielectric results. Another noticeable trend is large overestimations of the volumetric thermal expansion coefficient and isothermal compressibility for the halides, which are mainly driven by small molecules with multiple halogens such as chloroform (>48/74% deviations, respectively), dichloro(fluoro)methane (>49%/n.a.), 1,1-dichloroethene (>43%/n.a.), and 1,1,2,2,-tetrochloroethane (>14/34%). There is also a trend for the heat capacity of halides to be underestimated (on average by >23%). It is known that halogens often exhibit anisotropic charge distributions with a positive electrostatic potential on the outermost part of the halogen, which cannot be accurately described using fixed-charge models.[87,88] On the basis of this understanding, various models have been developed for halides that include a virtual site with positive charge,[15,89−93] multipole electrostatics,[94,95] polarizability,[95−97] and angular-dependent LJ terms.[98] An additional trend is that the density and heat of

https://doi.org/10.1021/acs.jcim.1c00491
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

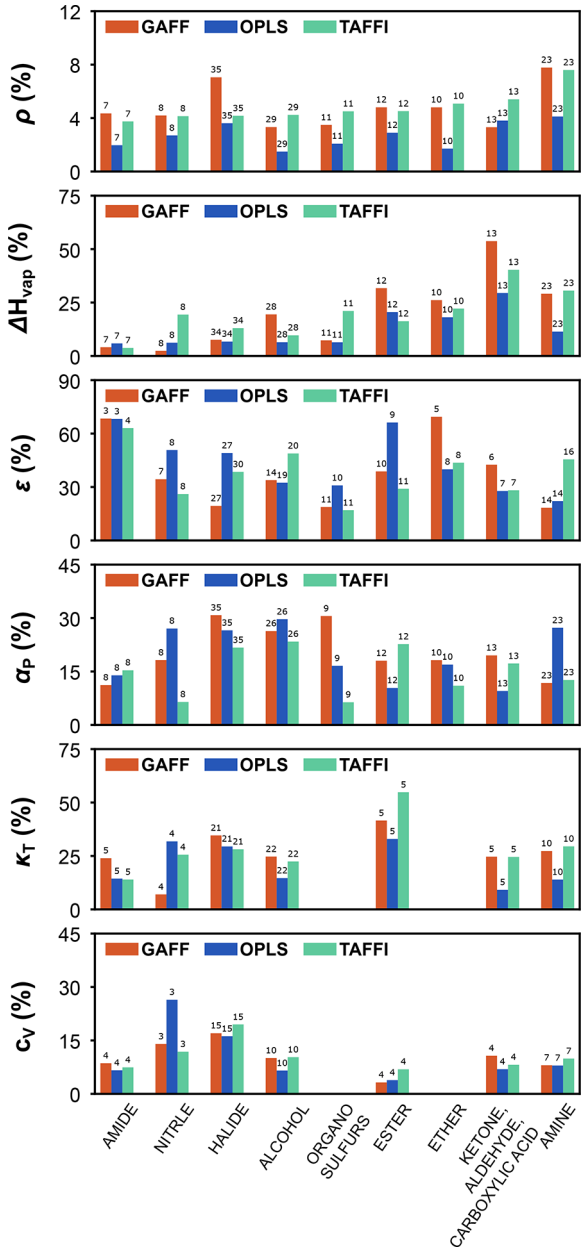


Figure 4. Mean absolute percent difference (MAPD) of the liquid properties for each functional group. The benchmark molecules are classified by the functional groups exhibited by each molecule. Each bar represents the average of the MAPD for all molecules belonging to each group. The numbers of molecules in each case are indicated above the bars, and properties with less than three values have been omitted. GAFF (red) and OPLS-AA (blue) data are from ref 71, whereas TAFFI-gen (green) data are from MD simulations performed in the current study.

vaporization predictions for amines are better predicted by OPLS compared with GAFF and TAFFI. This performance is attributed to the additional use of *ab initio* calculated hydrogen-bond strengths during OPLS parametrization for these terms[99] in the interest of reproducing experimental solvation free energies.[100]

A distinct outlier for TAFFI-gen is diethyl carbonate, which exhibits a large density underestimation in comparison with the experiment (MSPD = −18%; this is the outlier visible in Figure 3a at $\rho_{exp} \sim 0.9$). This is the only carbonate in the benchmark, and carbonates are unique in that they are the only benchmarked functional group that extends beyond the $d = 2$ graph specificity explored here for TAFFI-gen. In particular, the $d = 2$ model compound for the backbone oxygens (atom type [8[6[6][1][1]][6[8][8]]]) is ethoxyformic acid, which fails to preserve the carbonate structure. The other benchmarked properties of diethyl carbonate are also relatively poorly reproduced (ΔH_{vap}, −21%; ε, −19%; α_P, 86%; κ_T, 200%; c_v, −10%), which we attribute to the poor congruence between the model compounds and the target carbonate. This is further confirmed by an experiment where we reparameterized the diethyl carbonate LJ force-field terms for the ether oxygens and the carbonate carbon with ethyl methyl carbonate, which preserves the carbonate. In this case, the errors in comparison with the experiment are much smaller (ρ, −1%; ΔH_{vap}, 4%; ε, −2%; α_P, 34%; κ_T, 25%; c_v, −4%). This is a revealing example of how a fixed graph specificity (i.e., $d = 2$ in the current study) can lead to nonsystematic errors when applied to large functional groups.

Methanoic acid is also a distinct outlier for all of the force fields. This system exhibits long correlation times for the system dipoles, which have been previously established to originate from strong dimer interactions.[71,101] For TAFFI-gen, the dipole correlation decay could not be converged even with longer 50 ns trajectories (not shown). Additionally, the overestimation of the heat of vaporization for the ketone, aldehyde, and carboxylic acid group is disproportionately affected by methanoic acid (>110% deviation), where the other outliers are relatively minor [1-methoxy-2-(2-methoxyethoxy)ethane (>30%) and pentane-2,4-dione (>35%)]. Excluding methanoic acid from the group for heat of vaporization results in MAPD values similar to those of other oxygen-containing functional groups (GAFF, 20.24%; OPLS-AA, 13.94%; and TAFFI, 27.58%). The ΔH_{vap} for methanoic acid is overestimated by approximately a factor of two, which may also be caused by the formation of dimers in the gas phase under experimental conditions. This is an illustrative case of how fixing the force-field complexity does not lead to systematic errors across distinct chemistries. To achieve a target accuracy for a given set of properties, it is possible to simplify the force field in some cases, while it is necessary to add complexity in others. The development of more sophisticated models for hydrogen bonding in methanoic acid indirectly substantiates this point.[101−105]

As noted by Caleman et al., there are also cases where the simulation conditions may exacerbate prediction errors in comparison with the experiment. For example, the benchmarks for some alcohols and amines, including propane-1,2,3-triol and (2-hydroxyethoxy)ethan-2-ol, and ethane-1,2-diamine, are performed near their melting point. This results in highly viscous liquids at the simulation temperatures (Table S3) and likely exacerbates errors in the fluctuation-derived properties that are not representative of simulations further away from the phase transition.

4. CONCLUSIONS

It would be useful to have a force-field framework that could bridge simple fixed-charge force fields on the one hand and complex machine learning force fields on the other. The present work takes the first step in this direction by establishing a parametrization framework (TAFFI) based on an extensible quantum chemistry data set that can be used to fit transferable force fields of varying complexity. With the

https://doi.org/10.1021/acs.jcim.1c00491
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

TAFFI framework, we have formalized the concept of atom typing and made it the basis for generating systematic training data that maintains a one-to-one correspondence with force-field terms. This feature makes TAFFI arbitrarily extensible to new chemistries while maintaining internal consistency and transferability. As a demonstration of TAFFI, we have developed a fixed-charge force field, TAFFI-gen, from scratch that includes coverage for many common organic moieties. The performance of TAFFI-gen was benchmarked against OPLS-AA and GAFF for reproducing several experimental properties of 87 organic liquids. The comparable accuracy between TAFFI-gen and existing force fields in this benchmark is quite encouraging in light of the decades of optimization the existing force fields have undergone and their use of experimental data. Nevertheless, a major conclusion from this case study is that the similar qualitative behaviors of these force fields reflect the representability limitations of their simple functional form in approximating the atomistic PES. In particular, similar trade-offs and inaccuracies are observed in all of the force fields, which motivates a more sophisticated treatment of intermolecular interactions.

We have been careful to document the shortcomings of TAFFI-gen, since our long-term goal is not to simply make the best fixed-charge force field but to develop a data-driven means of matching force-field complexity to simulation targets. For instance, amide and halogen containing molecules exhibited among the largest deviations in TAFFI-gen for various liquid properties. Although it would be possible to introduce *ad hoc* corrections to the LJ parameters and partial charges associated with these functional groups, it would come at the expense of increasing errors in reproducing the interaction energies in the training data, and thus would likely lead to uncontrolled errors in other liquid properties. Such *ad hoc* corrections are what we want to avoid with TAFFI. From our perspective, a better pathway forward is to systematically increase the complexity of specific force-field terms based on well-defined error metrics. For example, selectively adding lone-pair sites or Drude particles to specific functional groups could foreseeably be done in a data-driven manner to improve the accuracy of a specific property without introducing *ad hoc* corrections. A similar strategy could apply to more complex intramolecular terms, like improper dihedrals or coupled modes. Applying impropers to most amides is clearly justified based on the large deviations from the training data, while for other moieties the deviations are minor and the terms can be neglected. Likewise, we observed that carbonates require larger model compounds than other functional groups, which motivates potentially treating distinct functional groups at variable levels of graph specificity (i.e., in contrast to the fixed $d = 2$ specificity used here for all benchmarks). Within the context of the TAFFI framework, such comparative retraining against shared training data is possible while retaining transferability and on-the-fly extensibility. Additionally, the systematic expansion of training data based on the occurrence of new atom types is also a promising basis for training transferable ML force fields for organic chemistry.

The current study is limited to liquid simulations of neutral noncyclic organic species, but several extensions to other classes of molecules and force-field forms are obvious and underway. Because TAFFI is based solely on quantum chemistry data, it can be extended to ionic and radical species that have limited coverage in existing experimentally based force fields. The extension to ions and radicals will require a more general treatment of formal charges in the atom types and model compounds than has been presented here. We have also noted that cyclic molecules and large conjugated groups fundamentally challenge the locality assumption implicit in the use of atom types. A workable near-term solution is to parametrize such systems whole and later use the data generated in this way to establish general ring and conjugation corrections. With respect to extending TAFFI to support the parametrization of more complex force fields, it will be necessary to augment the calculations currently performed on model compounds to include properties like atomic polarizability, heat of formation, and bond-dissociation energies that would justify more complex parametrizations. The small model compounds used by TAFFI for generating reference data is an advantage in this respect, as higher levels of theory and more extensive characterizations can be afforded while pursuing broad coverage of organic chemical space.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.1c00491.

Implementation details for molecular dynamics simulations, structural comparisons for model compounds, interaction energy histograms, and summary of simulated properties (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Brett M. Savoie − *Davidson School of Chemical Engineering, Purdue University, West Lafayette, Indiana 47906, United States;* orcid.org/0000-0002-7039-4039; Email: bsavoie@purdue.edu

Authors

Bumjoon Seo − *Davidson School of Chemical Engineering, Purdue University, West Lafayette, Indiana 47906, United States*

Zih-Yu Lin − *Davidson School of Chemical Engineering, Purdue University, West Lafayette, Indiana 47906, United States*

Qiyuan Zhao − *Davidson School of Chemical Engineering, Purdue University, West Lafayette, Indiana 47906, United States;* orcid.org/0000-0003-3228-8160

Michael A. Webb − *Department of Chemical and Biological Engineering, Princeton University, Princeton, New Jersey 08540, United States;* orcid.org/0000-0002-7420-4474

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.1c00491

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The work performed by B.S. and B.M.S. was made possible by the National Science Foundation (NSF) Division of Chemical, Bioengineering, Environmental, and Transport Systems (CBET) through support provided by the Electrochemical Systems Program (grant number, 2045887-CBET; Program Manager, Dr. Carol Read). Acknowledgment is made to the Donors of the American Chemical Society Petroleum Research Fund for support of the work by Z.-Y.L. The work of Q.Z. was

made possible through support of the Purdue Process Safety and Assurance Center. M.A.W. acknowledges support from Princeton University. B.M.S. also acknowledges partial support through the Dreyfus Program for Machine Learning in the Chemical Sciences and Engineering. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant no. ACI-1548562. Simulations were performed on the Comet supercomputer at the University of California, San Diego, under allocation no. TG-CHE190014.

REFERENCES

(1) Jorgensen, W. L.; Tirado-Rives, J. Potential Energy Functions for Atomic-Level Simulations of Water and Organic and Biomolecular Systems. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 6665−6670.
(2) Huang, J.; MacKerell, A. D., Jr. Force Field Development and Simulations of Intrinsically Disordered Proteins. *Curr. Opin. Struct. Biol.* **2018**, *48*, 40−48.
(3) Nerenberg, P. S.; Head-Gordon, T. New Developments in Force Fields for Biomolecular Simulations. *Curr. Opin. Struct. Biol.* **2018**, *49*, 129−138.
(4) Lemkul, J. A.; Huang, J.; Roux, B.; MacKerell, A. D., Jr An Empirical Polarizable Force Field Based on the Classical Drude Oscillator Model: Development History and Recent Applications. *Chem. Rev.* **2016**, *116*, 4983−5013.
(5) Liang, T.; Shin, Y. K.; Cheng, Y.-T.; Yilmaz, D. E.; Vishnu, K. G.; Verners, O.; Zou, C.; Phillpot, S. R.; Sinnott, S. B.; Van Duin, A. C. Reactive Potentials for Advanced Atomistic Simulations. *Annu. Rev. Mater. Res.* **2013**, *43*, 109−129.
(6) Xu, P.; Guidez, E. B.; Bertoni, C.; Gordon, M. S. Perspective: Ab Initio Force Field Methods Derived From Quantum Mechanics. *J. Chem. Phys.* **2018**, *148*, 090901.
(7) Riniker, S. Fixed-Charge Atomistic Force Fields for Molecular Dynamics Simulations in the Condensed Phase: An Overview. *J. Chem. Inf. Model.* **2018**, *58*, 565−578.
(8) Cornell, W.; Cieplak, P.; Bayly, C.; Gould, I.; Merz, K.; Ferguson, D.; Spellmeyer, D.; Fox, T.; Caldwell, J.; Kollman, P. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179−5197.
(9) Debiec, K.; Cerutti, D.; Baker, L.; Gronenborn, A.; Case, D.; Chong, L. Further Along the Road Less Traveled: AMBER Ff15ipq, an Original Protein Force Field Built on a Self-Consistent Physical Model. *J. Chem. Theory Comput.* **2016**, *12*, 3926−3947.
(10) MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B* **1998**, *102*, 3586−3616.
(11) Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; de Groot, B. L; Grubmuller, H.; MacKerell, A. D CHARMM36m: An Improved Force Field for Folded and Intrinsically Disordered Proteins. *Nat. Methods* **2017**, *14*, 71−73.
(12) Daura, X.; Mark, A.; van Gunsteren, W. Parametrization of Aliphatic CHn United Atoms of GROMOS96 Force Field. *J. Comput. Chem.* **1998**, *19*, 535−547.
(13) Horta, B.; Merz, P.; Fuchs, P.; Dolenc, J.; Riniker, S.; Hünenberger, P. A GROMOS-Compatible Force Field for Small Organic Molecules in the Condensed Phase: The 2016H66 Parameter Set. *J. Chem. Theory Comput.* **2016**, *12*, 3825−3850.
(14) Jorgensen, W. L.; Tirado-Rives, J. The OPLS [Optimized Potentials for Liquid Simulations] Potential Functions for Proteins, Energy Minimizations for Crystals of Cyclic Peptides and Crambin. *J. Am. Chem. Soc.* **1988**, *110*, 1657−1666.
(15) Harder, E.; Damm, W.; Maple, J.; Wu, C.; Reboul, M.; Xiang, J.; Wang, L.; Lupyan, D.; Dahlgren, M.; Knight, J.; Kaus, J.; Cerutti, D.; Krilov, G.; Jorgensen, W.; Abel, R.; Friesner, R. OPLS3: AForce Field Providing Broad Coverage of Drug-Like Small Molecules and Proteins. *J. Chem. Theory Comput.* **2016**, *12*, 281−296.
(16) Behler, J.; Parrinello, M. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Phys. Rev. Lett.* **2007**, *98*, 583−4.
(17) Bartók, A. P.; Payne, M. C.; Kondor, R.; Csányi, G. Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, without the Electrons. *Phys. Rev. Lett.* **2010**, *104*, 136403.
(18) Artrith, N.; Urban, A. An Implementation of Artificial Neural-Network Potentials for Atomistic Materials Simulations: Performance for TiO2. *Comput. Mater. Sci.* **2016**, *114*, 135−150.
(19) Khorshidi, A.; Peterson, A. A. Amp: A Modular Approach to Machine Learning in Atomistic Simulations. *Comput. Phys. Commun.* **2016**, *207*, 310−324.
(20) Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1: An Extensible Neural Network Potential with DFT Accuracy at Force Field Computational Cost. *Chem. Sci.* **2017**, *8*, 3192−3203.
(21) Zhang, L.; Han, J.; Wang, H.; Car, R.; Weinan, E. Deep Potential Molecular Dynamics: A Scalable Model with the Accuracy of Quantum Mechanics. *Phys. Rev. Lett.* **2018**, *120*, 143001.
(22) Yao, K.; Herr, J. E.; Toth, D. W.; Mckintyre, R.; Parkhill, J. The TensorMol-0.1 Model Chemistry: A Neural Network Augmented with Long-Range Physics. *Chem. Sci.* **2018**, *9*, 2261−2269.
(23) Schütt, K. T.; Arbabzadah, F.; Chmiela, S.; Müller, K. R.; Tkatchenko, A. Quantum-Chemical Insights From Deep Tensor Neural Networks. *Nat. Commun.* **2017**, *8*, 1−8.
(24) Schütt, K. T.; Sauceda, H. E.; Kindermans, P.-J.; Tkatchenko, A.; Müller, K.-R. SchNet−A Deep Learning Architecture for Molecules and Materials. *J. Chem. Phys.* **2018**, *148*, 241722.
(25) Chmiela, S.; Sauceda, H. E.; Müller, K.-R.; Tkatchenko, A. Towards Exact Molecular Dynamics Simulations with Machine-Learned Force Fields. *Nat. Commun.* **2018**, *9*, 1−10.
(26) Anisimov, V.; Lamoureux, G.; Vorobyov, I.; Huang, N.; Roux, B.; MacKerell, A. Determination of Electrostatic Parameters for a Polarizable Force Field Based on the Classical Drude Oscillator. *J. Chem. Theory Comput.* **2005**, *1*, 153−168.
(27) Lemkul, J.; Huang, J.; Roux, B.; Mackerell, A. An Empirical Polarizable Force Field Based on the Classical Drude Oscillator Model: Development History and Recent Applications. *Chem. Rev.* **2016**, *116*, 4983−5013.
(28) McDaniel, J.; Schmidt, J. Physically-Motivated Force Fields From Symmetry-Adapted Perturbation Theory. *J. Phys. Chem. A* **2013**, *117*, 2053−2066.
(29) McDaniel, J.; Choi, E.; Son, C.; Schmidt, J.; Yethiraj, A. Conformational and Dynamic Properties of Poly (Ethylene Oxide) in an Ionic Liquid: Development and Implementation of a First-Principles Force Field. *J. Phys. Chem. B* **2016**, *120*, 231−243.
(30) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J.; Kollman, P. A Point-Charge Force Field for Molecular Mechanics Simulations of Proteins Based on Condensed-Phase Quantum Mechanical Calculations. *J. Comput. Chem.* **2003**, *24*, 1999−2012.
(31) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. Comparison of Multiple Amber Force Fields and Development of Improved Protein Backbone Parameters. *Proteins: Struct., Funct., Genet.* **2006**, *65*, 712−725.
(32) Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E. Improved Side-Chain Torsion Potentials for the Amber Ff99SB Protein Force Field. *Proteins: Struct., Funct., Genet.* **2010**, *78*, 1950−1958.
(33) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. Ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters From Ff99SB. *J. Chem. Theory Comput.* **2015**, *11*, 3696−3713.

https://doi.org/10.1021/acs.jcim.1c00491
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

(34) 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*, 2151−2165.
(35) Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. Evaluation and Reparametrization of the OPLS-AA Force Field for Proteins via Comparison with Accurate Quantum Chemical Calculations on Peptides. *J. Phys. Chem. B* **2001**, *105*, 6474−6487.
(36) Mackerell, A. D., Jr; Feig, M.; Brooks, C. L., III Extending the Treatment of Backbone Energetics in Protein Force Fields: Limitations of Gas-Phase Quantum Mechanics in Reproducing Protein Conformational Distributions in Molecular Dynamics Simulations. *J. Comput. Chem.* **2004**, *25*, 1400−1415.
(37) Best, R. B.; Zhu, X.; Shim, J.; Lopes, P. E.; Mittal, J.; Feig, M.; MacKerell, A. D., Jr Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone ϕ, ψ and Side-Chain χ1 and χ2 Dihedral Angles. *J. Chem. Theory Comput.* **2012**, *8*, 3257−3273.
(38) Cacelli, I.; Prampolini, G. Parametrization and Validation of Intramolecular Force Fields Derived From DFT Calculations. *J. Chem. Theory Comput.* **2007**, *3*, 1803−1817.
(39) Horton, J. T.; Allen, A. E.; Dodda, L. S.; Cole, D. J. QUBEKit: Automating the Derivation of Force Field Parameters From Quantum Mechanics. *J. Chem. Inf. Model.* **2019**, *59*, 1366−1381.
(40) Grimme, S. A General Quantum Mechanically Derived Force Field (QMDFF) for Molecules and Condensed Phase Simulations. *J. Chem. Theory Comput.* **2014**, *10*, 4497−4514.
(41) Vanommeslaeghe, K.; Raman, E.; MacKerell, A. Automation of the CHARMM General Force Field (CGenFF) II: Assignment of Bonded Parameters and Partial Atomic Charges. *J. Chem. Inf. Model.* **2012**, *52*, 3155−3168.
(42) Shivakumar, D.; Harder, E.; Damm, W.; Friesner, R.; Sherman, W. Improving the Prediction of Absolute Solvation Free Energies Using the Next Generation OPLS Force Field. *J. Chem. Theory Comput.* **2012**, *8*, 2553−2558.
(43) Boyd, N.; Wilson, M. Optimization of the GAFF Force Field to Describe Liquid Crystal Molecules: The Path to a Dramatic Improvement in Transition Temperature Predictions. *Phys. Chem. Chem. Phys.* **2015**, *17*, 24851−24865.
(44) Doherty, B.; Zhong, X.; Gathiaka, S.; Li, B.; Acevedo, O. Revisiting OPLS Force Field Parameters for Ionic Liquid Simulations. *J. Chem. Theory Comput.* **2017**, *13*, 6131−6145.
(45) Jin, Z.; Yang, C.; Cao, F.; Li, F.; Jing, Z.; Chen, L.; Shen, Z.; Xin, L.; Tong, S.; Sun, H. Hierarchical Atom Type Definitions and Extensible All-Atom Force Fields. *J. Comput. Chem.* **2016**, *37*, 653−664.
(46) Mobley, D. L.; Bannan, C. C.; Rizzi, A.; Bayly, C. I.; Chodera, J. D.; Lim, V. T.; Lim, N. M.; Beauchamp, K. A.; Slochower, D. R.; Shirts, M. R.; Gilson, M. K.; Eastman, P. K. Escaping Atom Types in Force Fields Using Direct Chemical Perception. *J. Chem. Theory Comput.* **2018**, *14*, 6076−6092.
(47) Wang, J.; Wolf, R.; Caldwell, J.; Kollman, P.; Case, D. Development and Testing of a General Amber Force Field. *J. Comput. Chem.* **2004**, *25*, 1157−1174.
(48) Case, D.; Betz, R.; Cerutti, D.; Cheatham, T.; III, I; Darden, T.; Duke, R.; Giese, T.; Gohlke, H.; Goetz, A.; Homeyer, N.; Izadi, S.; Janowski, P.; Kaus, J.; Kovalenko, A.; Lee, T.; LeGrand, S.; Li, P.; Lin, C.; Luchko, T.; Luo, R.; Madej, B.; Mermelstein, D.; Merz, K.; Monard, G.; Nguyen, H.; Nguyen, H.; Omelyan, I.; Onufriev, A.; Roe, D.; Roitberg, A.; Sagui, C.; Simmerling, C.; Botello-Smith, W.; Swails, J.; Walker, R.; Wang, J.; Wolf, R.; Wu, X.; Xiao, L.; Kollman, P. *Amber 2016*; University of California: San Francisco, 2016.
(49) Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; Mackerell, A. CHARMM General Force Field: A Force Field for Drug-Like Molecules Compatible with the CHARMM All-Atom Additive Biological Force Fields. *J. Comput. Chem.* **2009**, *31*, 671−690.
(50) Mayne, C.; Saam, J.; Schulten, K.; Tajkhorshid, E.; Gumbart, J. Rapid Parameterization of Small Molecules Using the Force Field Toolkit. *J. Comput. Chem.* **2013**, *34*, 2757−2770.
(51) Jorgensen, W.; Maxwell, D.; Tirado-Rives, J. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.* **1996**, *118*, 11225−11236.
(52) Roos, K.; Wu, C.; Damm, W.; Reboul, M.; Stevenson, J. M.; Lu, C.; Dahlgren, M. K.; Mondal, S.; Chen, W.; Wang, L.; Abel, R.; Friesner, R. A.; Harder, E. D. OPLS3e: Extending Force Field Coverage for Drug-Like Small Molecules. *J. Chem. Theory Comput.* **2019**, *15*, 1863−1874.
(53) Savoie, B. M.; Webb, M. A.; Miller, T. F., III Enhancing Cation Diffusion and Suppressing Anion Diffusion via Lewis-Acidic Polymer Electrolytes. *J. Phys. Chem. Lett.* **2017**, *8*, 641−646.
(54) Khot, A.; Shiring, S. B.; Savoie, B. M. Evidence of Information Limitations in Coarse-Grained Models. *J. Chem. Phys.* **2019**, *151*, 244105.
(55) Zhao, Q.; Savoie, B. M. Self-Consistent Component Increment Theory for Predicting Enthalpy of Formation. *J. Chem. Inf. Model.* **2020**, *60*, 2199−2207.
(56) Sanderson, R. T. Electronegativity and Bond Energy. *J. Am. Chem. Soc.* **1983**, *105*, 2259−2261.
(57) Sanderson, R. *Chemical Bonds and Bonds Energy*; Academic Press: New York, 1976.
(58) Pracht, P.; Bohle, F.; Grimme, S. Automated Exploration of the Low-Energy Chemical Space with Fast Quantum Chemical Methods. *Phys. Chem. Chem. Phys.* **2020**, *22*, 7169−7192.
(59) Rappe, A.; Casewit, C.; Colwell, K.; Goddard, W.; Skiff, W. UFF, a Full Periodic Table Force Field for Molecular Mechanics and Molecular Dynamics Simulations. *J. Am. Chem. Soc.* **1992**, *114*, 10024−10035.
(60) Plimpton, S. Fast Parallel Algorithms for Short-Range Molecular Dynamics. *J. Comput. Phys.* **1995**, *117*, 1−19.
(61) Neese, F. The ORCA Program System. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2012**, *2*, 73−78.
(62) Klamt, A.; Schüürmann, G. COSMO: ANew Approach to Dielectric Screening in Solvents with Explicit Expressions for the Screening Energy and Its Gradient. *J. Chem. Soc., Perkin Trans. 2* **1993**, *2*, 799−805.
(63) Breneman, C. M.; Wiberg, K. B. Determining Atom-Centered Monopoles From Molecular Electrostatic Potentials. The Need for High Sampling Density in Formamide Conformational Analysis. *J. Comput. Chem.* **1990**, *11*, 361−373.
(64) Bayly, C. I.; Cieplak, P.; Cornell, W.; Kollman, P. A. A Well-Behaved Electrostatic Potential Based Method Using Charge Restraints for Deriving Atomic Charges: The RESP Model. *J. Phys. Chem.* **1993**, *97*, 10269−10280.
(65) Janecek, M.; Kuhrova, P.; Mlynsky, V.; Otyepka, M.; Sponer, J.; Banas, P. W-Resp: Well-Restrained Electrostatic Potential-Derived Charges. Revisiting the Charge Derivation Model. *J. Chem. Theory Comput.* **2021**, *17*, 3495−3509.
(66) Schauperl, M.; Nerenberg, P. S.; Jang, H.; Wang, L.-P.; Bayly, C. I.; Mobley, D. L.; Gilson, M. K. Non-Bonded Force Field Model with Advanced Restrained Electrostatic Potential Charges (RESP2). *Commum. Chem.i* **2020**, *3*, 1−11.
(67) Lin, Y.-S.; Li, G.-D.; Mao, S.-P.; Chai, J.-D. Long-Range Corrected Hybrid Density Functionals with Improved Dispersion Corrections. *J. Chem. Theory Comput.* **2013**, *9*, 263−272.
(68) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A Consistent and Accurate Ab Initio Parametrization of Density Functional Dispersion Correction (DFT-D) for the 94 Elements H-Pu. *J. Chem. Phys.* **2010**, *132*, 154104.
(69) Schäfer, A.; Horn, H.; Ahlrichs, R. Fully Optimized Contracted Gaussian Basis Sets for Atoms Li to Kr. *J. Chem. Phys.* **1992**, *97*, 2571−2577.
(70) Weigend, F.; Ahlrichs, R. Balanced Basis Sets of Split Valence, Triple Zeta Valence and Quadruple Zeta Valence Quality for H to Rn:

https://doi.org/10.1021/acs.jcim.1c00491
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Design and Assessment of Accuracy. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297−3305.
(71) 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*, 61−74.
(72) Berens, P. H.; Mackay, D. H. J.; White, G. M.; Wilson, K. R. Thermodynamics and Quantum Corrections From Molecular Dynamics for Liquid Water. *J. Chem. Phys.* **1983**, *79*, 2375−2389.
(73) Pascal, T. A.; Lin, S.-T.; Goddard, W. A., III Thermodynamics of Liquids: Standard Molar Entropies and Heat Capacities of Common Solvents From 2PT Molecular Dynamics. *Phys. Chem. Chem. Phys.* **2011**, *13*, 169−181.
(74) Chen, M.; Ko, H.-Y.; Remsing, R. C.; Calegari Andrade, M. F.; Santra, B.; Sun, Z.; Selloni, A.; Car, R.; Klein, M. L.; Perdew, J. P.; Wu, X. Ab Initio Theory and Modeling of Water. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114*, 10846−10851.
(75) Yao, Y.; Kanai, Y. Free Energy Profile of Nacl in Water: First-Principles Molecular Dynamics with Scan and ωb97x-v Exchange−Correlation Functionals. *J. Chem. Theory Comput.* **2018**, *14*, 884−893.
(76) Seeger, Z. L.; Izgorodina, E. I. A Systematic Study of DFT Performance for Geometry Optimizations of Ionic Liquid Clusters. *J. Chem. Theory Comput.* **2020**, *16*, 6735−6753.
(77) Sure, R.; Grimme, S. Comprehensive Benchmark of Association (Free) Energies of Realistic Host−Guest Complexes. *J. Chem. Theory Comput.* **2015**, *11*, 3785−3801.
(78) Lao, K. U.; Schäffer, R.; Jansen, G.; Herbert, J. M. Accurate Description of Intermolecular Interactions Involving Ions Using Symmetry-Adapted Perturbation Theory. *J. Chem. Theory Comput.* **2015**, *11*, 2473−2486.
(79) Henschel, H.; Andersson, A. T.; Jespers, W.; Mehdi Ghahremanpour, M.; van der Spoel, D. Theoretical Infrared Spectra: Quantitative Similarity Measures and Force Fields. *J. Chem. Theory Comput.* **2020**, *16*, 3307−3315.
(80) Kabsch, W. A Solution for the Best Rotation to Relate Two Sets of Vectors. *Acta Crystallogr., Sect. A: Cryst. Phys., Diffr., Theor. Gen. Crystallogr.* **1976**, *32*, 922−923.
(81) Kabsch, W. A Discussion of the Solution for the Best Rotation to Relate Two Sets of Vectors. *Acta Crystallogr., Sect. A: Cryst. Phys., Diffr., Theor. Gen. Crystallogr.* **1978**, *34*, 827−828.
(82) Fennell, C. J.; Wymer, K. L.; Mobley, D. L. A Fixed-Charge Model for Alcohol Polarization in the Condensed Phase, and Its Role in Small Molecule Hydration. *J. Phys. Chem. B* **2014**, *118*, 6438−6446.
(83) Mobley, D. L.; Bayly, C. I.; Cooper, M. D.; Shirts, M. R.; Dill, K. A. Small Molecule Hydration Free Energies in Explicit Solvent: An Extensive Test of Fixed-Charge Atomistic Simulations. *J. Chem. Theory Comput.* **2009**, *5*, 350−358.
(84) Sun, H.; Jin, Z.; Yang, C.; Akkermans, R. L.; Robertson, S. H.; Spenley, N. A.; Miller, S.; Todd, S. M. COMPASS II: Extended Coverage for Polymer and Drug-Like Molecule Databases. *J. Mol. Model.* **2016**, *22*, 47.
(85) Kramer, C.; Spinn, A.; Liedl, K. R. Charge Anisotropy: Where Atomic Multipoles Matter Most. *J. Chem. Theory Comput.* **2014**, *10*, 4488−4496.
(86) Harder, E.; Anisimov, V. M.; Whitfield, T.; MacKerell, A. D.; Roux, B. Understanding the Dielectric Properties of Liquid Amides From a Polarizable Force Field. *J. Phys. Chem. B* **2008**, *112*, 3509−3521.
(87) Murray, J. S.; Lane, P.; Clark, T.; Politzer, P. σ-Hole Bonding: Molecules Containing Group VI Atoms. *J. Mol. Model.* **2007**, *13*, 1033−1038.
(88) Clark, T.; Hennemann, M.; Murray, J. S.; Politzer, P. Halogen Bonding: The σ-Hole. *J. Mol. Model.* **2007**, *13*, 291−296.
(89) Ibrahim, M. A. Molecular Mechanical Study of Halogen Bonding in Drug Discovery. *J. Comput. Chem.* **2011**, *32*, 2564−2574.
(90) Rendine, S.; Pieraccini, S.; Forni, A.; Sironi, M. Halogen Bonding in Ligand−Receptor Systems in the Framework of Classical Force Fields. *Phys. Chem. Chem. Phys.* **2011**, *13*, 19508−19516.
(91) Kolář, M.; Hobza, P. On Extension of the Current Biomolecular Empirical Force Field for the Description of Halogen Bonds. *J. Chem. Theory Comput.* **2012**, *8*, 1325−1333.
(92) Jorgensen, W. L.; Schyman, P. Treatment of Halogen Bonding in the OPLS-AA Force Field: Application to Potent Anti-Hiv Agents. *J. Chem. Theory Comput.* **2012**, *8*, 3895−3901.
(93) Soteras Gutiérrez, I.; Lin, F.-Y.; Vanommeslaeghe, K.; Lemkul, J. A.; Armacost, K. A.; Brooks, C. L., III; MacKerell, A. D., Jr Parametrization of Halogen Bonds in the CHARMM General Force Field: Improved Treatment of Ligand−Protein Interactions. *Bioorg. Med. Chem.* **2016**, *24*, 4812−4825.
(94) Bereau, T.; Kramer, C.; Meuwly, M. Leveraging Symmetries of Static Atomic Multipole Electrostatics in Molecular Dynamics Simulations. *J. Chem. Theory Comput.* **2013**, *9*, 5450−5459.
(95) Mu, X.; Wang, Q.; Wang, L.-P.; Fried, S. D.; Piquemal, J.-P.; Dalby, K. N.; Ren, P. Modeling Organochlorine Compounds and the σ-Hole Effect Using a Polarizable Multipole Force Field. *J. Phys. Chem. B* **2014**, *118*, 6456−6465.
(96) Du, L.; Gao, J.; Bi, F.; Wang, L.; Liu, C. A Polarizable Ellipsoidal Force Field for Halogen Bonds. *J. Comput. Chem.* **2013**, *34*, 2032−2040.
(97) Lin, F.-Y.; MacKerell, A. D., Jr Polarizable Empirical Force Field for Halogen-Containing Compounds Based on the Classical Drude Oscillator. *J. Chem. Theory Comput.* **2018**, *14*, 1083−1098.
(98) Carter, M.; Rappé, A. K.; Ho, P. S. Scalable Anisotropic Shape and Electrostatic Models for Biological Bromine Halogen Bonds. *J. Chem. Theory Comput.* **2012**, *8*, 2461−2473.
(99) Rizzo, R. C.; Jorgensen, W. L. OPLS All-Atom Model for Amines: Resolution of the Amine Hydration Problem. *J. Am. Chem. Soc.* **1999**, *121*, 4827−4836.
(100) Kashefolgheta, S.; Oliveira, M. P.; Rieder, S. R.; Horta, B. A.; Acree, W. E., Jr; Hünenberger, P. H. Evaluating Classical Force Fields Against Experimental Cross-Solvation Free Energies. *J. Chem. Theory Comput.* **2020**, *16*, 7556−7580.
(101) Jedlovszky, P.; Turi, L. A New Five-Site Pair Potential for Formic Acid in Liquid Simulations. *J. Phys. Chem. A* **1997**, *101*, 2662−2665.
(102) Qian, W.; Krimm, S. Electrostatic Model for the Interaction Force Constants of the Formic Acid Dimer. *J. Phys. Chem. A* **1998**, *102*, 659−667.
(103) Hermida Ramón, J. M.; Rıos, M. A. A New Intermolecular Polarizable Potential for Cis-Formic Acid. Introduction of Many-Body Interactions in Condensed Phases. *Chem. Phys.* **1999**, *250*, 155−169.
(104) Roszak, S.; Gee, R. H.; Balasubramanian, K.; Fried, L. E. New Theoretical Insight Into the Interactions and Properties of Formic Acid: Development of a Quantum-Based Pair Potential for Formic Acid. *J. Chem. Phys.* **2005**, *123*, 144702.
(105) Schnabel, T.; Cortada, M.; Vrabec, J.; Lago, S.; Hasse, H. Molecular Model for Formic Acid Adjusted to Vapor−Liquid Equilibria. *Chem. Phys. Lett.* **2007**, *435*, 268−272.

https://doi.org/10.1021/acs.jcim.1c00491
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX