Title

Data-Driven Mapping of Gas-Phase Quantum Calculations to General Force Field Lennard-Jones Parameters.

Permalink

https://escholarship.org/uc/item/9306p29f

Journal

Journal of Chemical Theory and Computation, 16(2)

ISSN

1549-9618

Authors

Kantonen, Sophie M
Muddana, Hari S
Schauperl, Michael
et al.

Publication Date

2020-02-11

DOI

10.1021/acs.jctc.9b00713

Peer reviewed

Published in final edited form as:
J Chem Theory Comput. 2020 February 11; 16(2): 1115–1127. doi:10.1021/acs.jctc.9b00713.

Data-Driven Mapping of Gas-Phase Quantum Calculations to General Force Field Lennard-Jones Parameters

Sophie M. Kantonen[1], **Hari Muddana**[1,2], **Michael Schauperl**[1], **Niel M. Henriksen**[1,3], **Lee-Ping Wang**[4], **Michael K. Gilson**[1,*]

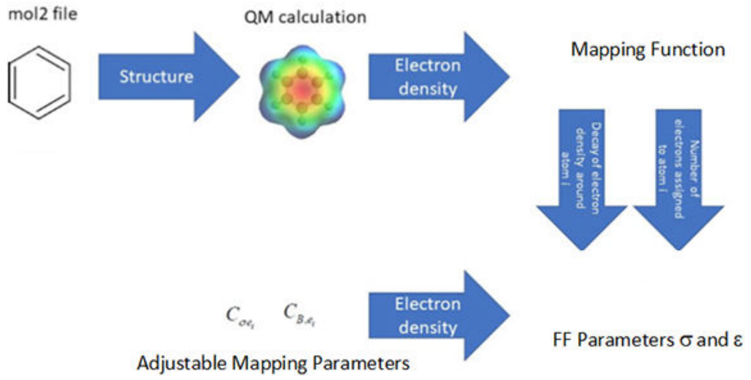
[1)]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, 9500 Gilman Drive, La Jolla, California 92093-0736, USA

[2)]OpenEye Scientific Software, Inc., 9 Bisbee Court, Suite D, Santa Fe, NM 87508, USA

[3)]AtomWise, Inc. 717 Market Street, Suite 800, San Francisco, California 94103, USA

[4)]Department of Chemistry, University of California Davis, One Shields Ave., Davis, CA, 95616, USA

Graphical Abstract


mol2 file
Structure
QM calculation
Electron density
Mapping Function
Decay of electron density around atom i
Number of electrons assigned to atom i
$C_{\sigma\varepsilon_i}$
C_{B,ε_i}
Electron density
Adjustable Mapping Parameters
FF Parameters σ and ε

1 Introduction

Molecular dynamics (MD) simulations are useful for a broad range of applications, including protein folding studies[1], drug discovery[2], and the determination of liquid structure and properties[3]. Various approaches have been taken to improve the ability of simulations to explore the thermodynamically relevant parts of configuration space, including hardware advancements[4–9] and more effective sampling algorithms[10–14]. While these efforts have dramatically improved our ability to generate well-converged results, errors persist in simulations, as highlighted for example, in the SAMPL series of blinded prediction exercises[15–21]. Therefore, attention is now turning again to the potential functions, or force fields (FF), as sources of error, as recently reviewed[22].

[*]Corresponding author.

Currently, most biomolecular simulations are still carried out with FFs having a simple functional form comprising harmonic bond-stretches and angle-bends, sinusoidal torsional terms, harmonic improper dihedrals, Lennard-Jones (LJ) interactions for van der Waals forces, and Coulombic interactions among atom-centered point charges for electrostatic interactions and hydrogen bonding[23]. This common functional form has the merit of being supported by many well-developed simulation packages and of affording great computational speed and therefore effective conformational sampling. From this starting point, a number of strategies may be adopted to improve the accuracy of the FF. One is to move to a functional form that captures the physics more faithfully and in more detail. For example, one may add terms that explicitly account for effects, such as electronic polarizability[24–27], that are accounted for only implicitly, at best, in the common functional form; or, one may substitute a more realistic form for an existing term, such as a Coulombic term that accounts for charge penetration[26–31]. Another strategy is to remain with today’s common functional form and instead look for parameters, such as atomic partial charges, torsional barriers, and Lennard-Jones well-depths and radii, that will lead to greater accuracy when used to compute experimental and/or quantum chemical reference data. This strategy may be pursued by using more, and more relevant, experimental data in the parameterization process. For example, data on host-guest binding thermodynamics have recently been used to guide parameter adjustment, leading to improved accuracy in calculated binding thermodynamics. Another approach to improving accuracy of force fields with the common functional form is to more comprehensively and systematically define[32] and adjust[33–38] the parameters so they more closely approach an optimal parameterization against a fixed dataset.

However, the large number of independently adjustable FF parameters in a typical FF can make full, multidimensional parameter optimization daunting. For example, the SMIRNOFF99Frost[39] FF, whose list of parameters has already been condensed through the replacement of atom-typing with direct chemical perception, still has 35 different Lennard-Jones types and hence 70 Lennard-Jones parameters. Simultaneous optimization of these parameters becomes particularly challenging when evaluating the objective function requires running time-consuming simulations, such as if one wishes to tune parameters against liquid-state properties. Even more extensive sampling of parameters will likely be needed if one moves from optimization to Bayesian sampling[40–42] in the parameter space. If either an optimization or a Bayes sampling algorithm misses a key sector of parameter space, the accuracy of the resulting parameterized FF will not be a measure of the quality achievable within the common functional form. This situation is problematic for at least two reasons. First, simulations using the incompletely optimized FF will not be as accurate as they could have been. Second, the resulting errors may provide misleading guidance regarding the need to move to a more complex functional form. Consequently, a methodology of fitting FF’s that use fewer adjustable parameters would in principle make the problem more tractable.

We therefore propose a step toward reducing the dimensionality of the adjustable parameter space of the common FF functional form. The basic idea (Figure 1) is to use a QM calculation on the molecule to be parameterized (or on a suitable fragment of a large molecule) to extract properties of the electron density that correlate with the FF parameters to be assigned—here the Lennard-Jones parameters of each atom in the molecule. We then set up a mathematical mapping from the electron density to the targeted FF parameters. The mapping is outfitted with a small set of adjustable parameters, and it is these mapping parameters, not the targeted FF parameters, that are subjected to optimization or sampling based on calculations of experimental observables. This approach reduces the number of adjustable parameters, because the FF parameters assigned to each atom in the molecule are largely controlled by the QM results. Only the mapping parameters are adjusted to maximize the agreement of simulated properties with experiment. Thus, the dimensionality of the optimization problem is greatly reduced. By the same token, this approach allows each atom in a molecule to have unique FF parameters without requiring a large number of atom types.

The present study aims to prove the principle of this approach by demonstrating that such a QM-to-FF mapping, trained to generate Lennard-Jones FF parameters that best replicate a small set of experimental liquid-state data, yields a competitive level of accuracy for an informative set of properties. It thus sets the stage for future applications aimed at building a comprehensive FF for general use. The current implementation builds on promising approaches from other groups[28,43–50], as it uses Slater orbitals to model the electron density associated with each atom in the molecule and extracts key correlates of the LJ parameters from these fits. The present FF-development approach is therefore termed the Slater-Derived Lennard-Jones (SDLJ) method. The following subsections detail the concepts and methodology, describe how the mapping parameters are adjusted against experimental liquid-state data for a small training set of compounds, and report on the quality of the results when the trained method is tested on a larger, non-overlapping set of compounds and additional observables. Implications and prospects are considered in the Discussion section.

2 Methods

2.1 Overview

The present method provides an approach to mapping from the electronic structure of a molecule, obtained from a quantum mechanical (QM) calculation, to suitable σ and ε Lennard-Jones parameters for each of its atoms. The mapping contains two adjustable parameters for each element (C, N, O, H), except that polar and nonpolar hydrogens have separate parameters. Here a polar hydrogen is defined simply as one directly bound to a nitrogen or oxygen atom. We separated polar and nonpolar hydrogens because initial studies showed that lumping all hydrogens made it impossible to reach a set of LJ parameters that would afford competitive accuracy (results not shown). We use the ForceBalance software[34] to optimize the 10 mapping parameters in order to minimize the error of simulated liquid state properties vs. experimental measurements. The trained mapping is then tested against the properties of a larger set of liquids, and the results are compared with those obtained with the pioneering and widely used GAFF force field[51]. For this proof-of-concept study, we focused on consequences of adjusting only Lennard-Jones parameters, thus the electrostatic interactions are modeled with atom-center partial charges assigned with the AM1/BCC method and all non-LJ force field terms were drawn from GAFF in accordance with established procedures.

The mapping from a molecule's electronic structure to σ_i and ε_i for each of its atoms, i, uses an atoms-in-molecules (AIM) approach similar, but not identical, to ones that have been published before[43,44]. We model the electron density around each atom *i* in terms of a Slater orbital, where the electron density decays exponentially with distance from the nucleus: $\rho_i(r) \propto \exp(-\beta_i r)$. The decay coefficient β_i associated with each atom *i* then is used to assign both the effective atomic polarizability and the effective ionization potential, both key quantities in the LJ interaction term. The following subsections detail each step summarized in this overview. The code used to go from a QM result to Lennard-Jones parameters are available at (https://github.com/SKantonen/PyBLJ). It is written in Python, and its inputs are a Lebedev grid file and a mol2 file to generate β_i coefficients for each atom in the given mol2 structure. The level of QM to be run can be manually set inside of the code, if so desired.

2.2 Electron density calculations

Each molecule to be studied was built with the open-source software Avogadro[52], and its structure was energy-minimized using the GAFF force-field. The Gaussian09[53] package was used to obtain the electron density at the CCSD/cc-pVTZ level of theory. CCSD was chosen as it is considered to be among the most accurate post Hartree-Fock methods for calculations on small molecules[54]. We found that the values of β changed by <1% when the size of the basis set was increased further. The electron density distribution around each atom was computed using spherical Lebedev grids (110 points, order 17)[55] and a uniform radial grid of 0.05 Å spacing out from 0–12 Å. The code used to generate these grids is built into the β parameter fitting code, using the aforementioned spacings and grid sizes.

2.3 Fitting electron-density decay coefficients (beta) to atoms-in-molecules.

While there exist multiple useful electron partitioning methods[56–58], we chose to use the Minimal Basis Iterative Stockholder (MBIS) method of Verstraelen[59]. The MBIS method partitions the total electron density of the molecule (Section 2.2) into atom-centered Slater orbitals centered on atoms *i*, in a manner that minimizes the Kullback-Liebler (KL) divergence[59] between the electron density distribution provided by a QM calculation $\rho(\boldsymbol{r})$, and the sum of the atomic densities. The MBIS method is attractive because it allows pro-densities to vary (allowing individual atoms to have unique parameters governing their pro-densities as opposed to predetermined pro-densities) and has already been successfully applied to force field development[46,59].

Thus, each atom, *i*, is assigned a “pro-density”, $\rho_{i(\boldsymbol{r})}$, of Slater form whose integral over all space gives the total number of electrons assigned to the atom, N_i. The Slater orbital of atom *i* is characterized by β_i, the spatial decay constant of the electron density[60]. This quantity is expected, on physical grounds, to correlate with both the size and the dispersion interactions of the atom[60,61], as detailed below.

As previously shown, the KL divergence is minimized by an iterative procedure, diagrammed in Figure 2. For a given iteration, *k*, of the MBIS method,

$$N_{i,k+1} = \int \rho(\boldsymbol{r}) \frac{\rho_{i,k}(\boldsymbol{r}, \beta_{i,k}, N_{i,k})}{\rho_{0,\mathrm{k}}(\boldsymbol{r})} d\boldsymbol{r} \tag{1}$$

$$\beta_{i,k+1} = 3N_{i,k}\int \rho(\boldsymbol{r})\frac{\rho_{i,k}\left(\boldsymbol{r},\beta_{i,k},N_{i,k}\right)}{\rho_{0,\mathrm{k}}(\boldsymbol{r})}|\boldsymbol{r}-\boldsymbol{R_i}|d\boldsymbol{r} \tag{2}$$

Here $\rho_{\mathrm{I,k}}(\mathbf{r})$ is the pro-density around atom *i* at iteration k; β_{ik} is the estimate of β_i for atom *i* at iteration *k*; $\rho_0(\mathbf{r})$ is the sum of all atomic pro-densities at position **r**; and $\boldsymbol{R}_i$ is the location of the nucleus of atom *i*. In practice, these integrals are estimated as sums over Lebedev grids, as noted above:

$$N_{i,k+1} = \sum_{g=1}^{N_p}\rho(\boldsymbol{r}_g)\frac{\rho_{i,k}\left(\boldsymbol{r_g},\beta_{i,k},N_{i,k}\right)}{\rho_0\left(\boldsymbol{r}_g\right)} \tag{3}$$

$$\beta_{i,k+1} = \sum_{g=1}^{N_p}\rho(\boldsymbol{r}_g)\frac{\rho_{i,k}\left(\boldsymbol{r}_g,\beta_{i,k},N_{i,k}\right)}{\rho_0\left(\boldsymbol{r}_g\right)}\left|\boldsymbol{r}_g-\boldsymbol{R_i}\right| \tag{4}$$

For the sums, the density at each grid point *g* is considered, with the sum being over all N_p grid points, each located at $\mathbf{r}_\mathrm{g}$. As per the MBIS method, an initial value, corresponding to iteration *k=1*, is chosen for all N_{i1} and β_{i1}, allowing for the determination of starting pro-densities via:

$$\rho_{i1}(r) = \frac{N_{i1}\beta_i^{3}}{8\pi}e^{\left(-\beta_{i1}|r-R_i|\right)} \tag{5}$$

These pro-densities are used to generate $N_{i,k}$ and $\beta_{i,k}$ for the next iteration (k=2), and the process is iterated until the changes in N and β between iterations *k* and *k+1* falls below some threshold, here a 0.05% absolute change. The initial values of these quantities are detailed in the following paragraph.

In accord with Verstraelen et al, we assigned and fit two pro-densities to each non-hydrogen atom. (Only one pro-density is used for each hydrogen atom.) The “core” pro-density captures electrons held close to the nucleus, while the “valence” one includes all other electrons, and the number of electrons in core versus valence is adjusted as part of the procedure. Thus, we allow two β’s and two N’s to be fit for each non-hydrogen atom, and we use only the valence β to generate LJ parameters. The rationale is that core electrons only contribute to non-bonded interactions at distances much too close to be relevant, mostly due to exchange repulsion[61]. The initial values of *N* are set to 2 for the core pro-density and the number of valence electrons of the element for the valence pro-density. Initial values of β_i for all atoms are set to 12 and 4 Å^{-1} for the core and valence orbitals, respectively. These values, which correspond to those obtained for a single nitrogen atom, suffice to generate convergent results in the iterative procedure just outlined. As expected, core occupancy was on average very near two electrons, but the addition of the core exponential allowed for a better fit of the valence exponential. Values of β typically converge relatively quickly, with the final value being insensitive to the starting value, as shown in Supplementary Figure 1.

2.4 Mapping QM results to Lennard-Jones Parameters

The Lennard-Jones model gives the van der Waals interaction energy between atoms *i* and *j* as

$$E_{LJ} = \quad \frac{A_{ij}}{r_{ij}{}^{12}} - \frac{B_{ij}}{r_{ij}{}^{6}} = 4\varepsilon_{ij}\left(\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right) \tag{6}$$

where A_{ij}, B_{ij}, σ_{ij} and ε_{ij} may be obtained from atomic "self" parameters σ_i, σ_j, $\boldsymbol{\varepsilon_i}$, and $\boldsymbol{\varepsilon_j}$ by mixing rules, such as $\sigma_{ij} = 0.5(\sigma_i + \sigma_j)$ and $\varepsilon_{ij} = \left(\varepsilon_i \varepsilon_j\right)^{\frac{1}{2}}$[62]. The next two subsections describe how σ_i and ϵ_i are assigned to each atom in a molecule, based on the QM electron densities.

2.4.1 Lennard-Jones sigma—

The parameter σ is essentially an atomic diameter, and β^{-1} is proportional to the expectation value of the distance of the electron density from the nucleus[59], so we write

$$\sigma_i = \quad \frac{C_{\sigma,\ e_i}}{\beta_i} \tag{7}$$

Here e_i is the element of atom *i*, and $C_{\sigma e_i}$ is the associated element-specific mapping parameter, which is adjusted with ForceBalance[34] (Section 2.5).

2.4.2 Lennard-Jones epsilon—

The coefficient of the dispersion component of the Lennard-Jones interaction (Eq 6) can be estimated from the London Equation[63]:

$$B_{ij} = \frac{3}{2}\frac{\eta_i \eta_j}{\eta_i + \eta_j}\alpha_i \alpha_j \tag{8}$$

where η and α are the ionization energy and polarizability, respectively, of the subscripted atoms. We follow Tkatchenko[43] in writing the homonuclear B coefficient in the form

$$B_i \propto \eta_i V_i{}^2 \tag{9}$$

where η_i and V_i are, respectively, the effective ionization energy and volume of atom *i*. (This assumes that the polarizability is simply proportional to volume; it is worth noting that Gould has argued for an element-specific scaling between these quantities[78].) As noted above, β^{-1} is related to the atomic radius, so $V_i^2 \propto \beta^{-6}$. For a single atom, the ionization energy is given by $\eta = \frac{\beta^2}{8}$[61], and we assume the same to be true for the effective ionization energy of an atom in a molecule. This is distinct from the Tkatchenko method, in that we allow each atom to have a unique effective ionization energy, which contributes to the determination of the LJ parameters. Hence, inserting the element-specific fitting parameter C_{B,e_i}, we have

$$B_i = \quad C_{Be_i}\eta_i \beta_i^{-6} = \frac{C_{Be_i}}{8}\beta_i^{-4} \tag{10}$$

Intuitively, the lower the value of β_i, and hence the more diffuse the electron density, the greater the dispersion coefficient. The element-specific fitting parameters C_{B,e_i} are adjusted with ForceBalance, as detailed in Section 2.5.

From the expressions above, we can also derive that

$$\varepsilon_i = \frac{C_{B,e_i}\eta_i}{4C_{\sigma,\mathrm{e_i}}6} = \frac{C_{B,e_i}\beta_i^2}{32C_{\sigma,\mathrm{e_i}}6} \tag{11}$$

Since ε corresponds to the depth of the LJ energy well, this says that a more diffuse electron density corresponds to a smaller well-depth. This trend reflects the fact that a more cdiffuse electron density also increases σ and thus cause the energy well to be at a greater distance where dispersion forces will be weaker. Note that some prior methods of deriving the dispersion term by AIM methods have neglected variations in the ionization energy[43–45] among atoms. From the present expression for ε, it is apparent that this neglect causes all atoms of a given element to have the same value of ε. In the present approach, different atoms of a given element can have different values of both σ and ε. These issues are further considered in the Discussion section.

2.5 Optimizing the elemental mapping parameters using ForceBalance.

We used ForceBalance[34], a software package that automatically adjusts fitting parameters using parametric gradients of simulated properties, to optimize the mapping parameters $C_{\sigma,e}$ and $C_{B,e}$ for the elements carbon, nitrogen, oxygen, and separately for polar and nonpolar hydrogen. These mapping parameters were adjusted so that simulations of seven pure organic liquids (Section 2.6) with the resulting LJ parameters would yield properties close to experiment. The adjusted mapping parameters were then tested by using simulations to compute the properties of 24 pure organic liquids outside the training set and comparing these with experiment. The test- and training-set molecules were chosen to be small and simple, so that simulations could be rapidly converged, while still representing significant chemical diversity. In addition, to test transferability, we included test-set compounds with functional groups absent from the training set. For this initial study, atom-centered partial charges were assigned with the AM1/BCC method[64], as noted above, and bonded parameters were drawn from GAFF[65], using the program Antechamber[66]. The iterative ForceBalance process was initiated from a set of mapping parameters that minimize the sum of squared deviations between the mapped LJ parameters and GAFF LJ parameters for the training set compounds. The stopping criterion for ForceBalance was essentially chosen manually, as the history of the objective function was evaluated to see if any meaningful improvements were being made. When the objective function had fallen significantly from its starting value and appeared to plateau, the ForceBalance program was halted and the parameters were evaluated.

For each training set compound, liquid phase simulations were performed with the AMBER molecular dynamics suite[67] to compute the heat of vaporization and density in the NPT ensemble at 298K and 1 atm. The Berendsen barostat and Langevin thermostat were used for all production simulations, and SHAKE was used to constrain all R-H bond-lengths. For each calculation, 1000 molecules were used in the simulation box and production simulations of 12 ns at 2 fs time steps were run. A cutoff of 8 Å was used for both the Particle Mesh Ewald and Lennard-Jones calculations. The mapping parameters were optimized over multiple ForceBalance iterations so that they produced LJ parameters that minimizes a regularized, weighted least-squares objective function computed from the squared deviations of the calculated observables and experimental reference data[34].

The ForceBalance objective function was described in previous work[34] and is briefly summarized here. It has a hierarchical structure with the top level given by the formula:

$$L_{\mathrm{tot}}(\mathbf{k}) = \sum_{T\,\in\,\mathrm{targets}} w_T L_T(\mathbf{k}) + w_{\mathrm{reg}}|\mathbf{k}|^2$$

where the total objective function L_{tot} depends on the optimization variables **k** and is equal to the sum of contributions from the parameterization targets L_T weighted by w_T, plus a Tikhonov regularization term weighted by w_{reg}. Each parameterization target is a weighted sum of contributions for one or more properties:

$$L_T(\mathbf{k}) = \sum_{J\,\in\,\mathrm{properties}} w_j^{(T)} L_j^{(T)}(\mathbf{k})$$

In this study, the target weights w_T and the property weights $w_j^{(T)}$ were set to unity for both properties used (density and heat of vaporization), allowing each to contribute equally. The term for each property $L_j^{(T)}(\mathbf{k})$ is given by a weighted and normalized sum over individual data points:

$$L_j^{(T)}(\mathbf{k}) = \frac{1}{\left(d_j^{(T)}\right)^2} \frac{\sum_{p\,\in\,\mathrm{points}} w_{jp}^{(T)} \left| y_{jp}^{(T)}(\mathbf{k}) - y_{jp,\,\mathrm{ref}}^{(T)} \right|^2}{\sum_{p\,\in\,\mathrm{points}} w_{jp}^{(T)}}$$

Where $y_{jp}^{(T)}$ and $y_{jp,\,\mathrm{ref}}^{(T)}$ are, respectively, the simulated and reference data point for property j and point p within target T. The quantity $d_j^{(T)}$ is a scaling factor used to normalize and remove physical units for property j, and has the same effect as an inverse square weight; we used values of 30 for density and 0.3 for heat of vaporization.

The optimization variables **k** are mapped to a set of physical parameters **K** by a exponential mapping as $K_\lambda = K_\lambda^{(0)}\exp[k_\lambda]$, where $K_\lambda^{(0)}$ represents the original parameter values. Under this exponential mapping, the physical parameters do not change sign from their original values. The regularization term may be expressed in terms of mathematical parameters as:

$$w_{\mathrm{reg}}|\mathbf{k}|^2 = w_{\mathrm{reg}} \sum_{\lambda\,\in\,\mathrm{params}} k_\lambda^2 = w_{\mathrm{reg}} \sum_{\lambda\,\in\,\mathrm{params}} \left(\ln \frac{K_\lambda}{K_\lambda^{(0)}}\right)^2$$

During the optimization, FB computes the gradients of simulated properties with respect to force field parameters, i.e. $\nabla_{\mathbf{k}} y_{jp}^{(T)}(\mathbf{k})$, and uses them to construct the gradient and approximate Hessian of L_{tot} in the parameter space. In this work, $y_{jp}^{(T)}(\mathbf{k})$ represent ensemble properties obtained from thermodynamic sampling, and $\nabla_{\mathbf{k}} y_{jp}^{(T)}(\mathbf{k})$ is computed using thermodynamic fluctuation formulas as described in previous work[68].

The overall computational workflow is diagrammed in Figure 3. Once the mapping parameters were optimized, they were used to generate LJ parameters for a larger test set consisting of 24 molecules, for which densities, heats of vaporization, and heat capacities were calculated and compared with experimental numbers. The uncertainties reported for the calculated properties were obtained by a previously described blocking method[69,70].

2.6 Training and test data

When optimizing and testing force field parameters against experimental observables, it is essential to use reliable data. Here, we obtained data from the ThermoML[71] archive provided by NIST, and used a separate compilation of liquid properties of organic molecules[72] as a cross check to ensure accurate numbers for both the training and test sets. For some values from the ThermoML archive, the values were taken as averages over multiple experimental sources. While experimental uncertainties are not typically provided for these data, the few compounds that do have uncertainties typically show them to be of the order of less than 1% standard deviation (for all properties examined), even when measurements that are nearly a century old[73] are included when calculating average and standard deviations. The experimental uncertainties for these compounds are reported as under ~5% for both heats of vaporization and densities. The training set comprised the following seven pure liquids: methanol, ethanol, aminoethanol, acetaldehyde, ethylamine, benzene, and acetonitrile. The test set comprises the 24 additional pure liquids listed in Table 3. In order to rigorously assess transferability of the fitted parameters, we chose test set compounds with functional groups not in the training set. Radial distribution functions were digitized from figures in various sources[74–76] to allow comparison of radial distribution functions calculated from simulations of both SDLJ and GAFF parameters. Radial distribution functions were computed for the present simulations with the cpptraj program.[77]

As the observation may be relevant for other studies that rely on pure liquid properties, it is perhaps worth also reporting that a few compounds initially included in the training or test sets were found to undergo very slow conformational interconversions, no matter what starting conformer was used, even in simulations as long as 10 ns. In particular, we observed few or no syn-anti conversions of the proton in the carboxylic acids formic and acetic acid, either in gas or liquid phases. Both QM calculations and standard force-fields point to a barrier between the two states of at least 6 kcal/mol[78], so we do not think this problem results from the particular parameters used here. Simulations of these molecules led to significant convergence problems in the thermodynamic properties, so these compounds were removed and are not present in this study. Similarly, short esters were removed for the same reason.

3 Results

This section first reports on the optimization of the ten elemental mapping parameters using ForceBalance and a small training set of molecules. Then the transferability of the resulting parameters is tested with a larger, nonoverlapping, set of 24 test molecules. The results are compared with experiment and with corresponding simulations using GAFF LJ parameters.

3.1 Optimization of elemental mapping parameters using ForceBalance

As detailed in Methods, an electronic structure calculation was run for each compound in the training set, and the MBIS method was used to compute β_i for each atom i in each compound. (Final β values for all atoms in each molecule in the training set are provided in Supplementary Table 4.) These quantities were then used with the expressions in Section 2.4, and after about 40 iterations of optimization of the mapping parameters using ForceBalance (Section 2.5), a large improvement in the ForceBalance objective function[34] was observed (Figure 4). The procedure led to modest additional improvement as it ran out to about 70 cycles. The values of the final Lennard-Jones parameters are provided in Table 2 (in the standard AMBER form $R_{min}/2$ and ε) and the densities and heats of vaporization of the training set molecules computed with the optimized parameters are compared with experiment in Table 1. Final values of the mapping parameters are provided in Supplementary Table 1. Additionally, select mol2 and frcmod files are provided in the GitHub repository for a few test set molecules.

It is expected that the value of β for each atom in a molecule, and hence the values of σ and ε assigned by this method, will depend to some degree on the conformation used for the QM calculation. To assess the sensitivity to conformation, we took several snapshots of butanol from a 300 K gas phase simulation and used the trained parameters to compute LJ parameters for all atoms in each conformation. As detailed in Supplementary Tables 2 and 3, the variations across conformations are small, with standard deviations in σ of at most 0.02 Å, and standard deviations in ε of at most 0.001 kcal/mol.

3.2 Test-set validation of optimized mapping parameters

The trained SDLJ method yields densities for the test set that agree with experiment about as well as those in the training set, and heats of vaporization with about double the relative mean unsigned error of the training set (Figure 6, Table 1, and Table 3). Importantly, the SDLJ parameters provide good agreement with experiment even for compounds with functional groups distinctly different from those in the training set. For example, SDLJ reproduces the properties of dioxolane and furan reasonably well, and indeed more closely than done by GAFF, although the training set includes no substituted phenyls or furans. Taken together, these observations suggest that the parameters were not overfitted. Overall, the new method yields accuracy on the test set similar to that obtained with GAFF LJ parameters (Table 3 and Figure 6). This is despite the fact that the SDLJ method has only ten fitting parameters and was trained on only 14 observables. In contrast, the test-set compounds span 16 GAFF atom types and thus include 32 LJ parameters that are, at least in principle, independently adjustable. However, it should also be noted that GAFF was not parameterized against the present training set of liquid properties.

At the same time, it is worth noting that some compounds show undesirably large errors when modeled with either SDLJ or GAFF. Examples include the density of formamide for SDLJ and especially GAFF, the heat of vaporization of propionitrile for SDLJ, and the heat of vaporization of o-xylene and butanol for GAFF. Further work is needed to assess whether such errors should be attributed to problems with the LJ parameters; problems with other parameters, such as partial charges; from limitation in the training set; or, perhaps, from limitations of the common functional form itself.

We further probed the reliability of the SDLJ-based FF by computing radial distribution functions (RDFs) and comparing them with available experimental[72–74] data. As shown in Figure 7, SDLJ tends to overestimate the sharpness of the first shell hydration peaks for H-bonding atoms in methanol and ethanol, while GAFF tends to underestimate these first-shell peaks. However, SDLJ does a better job of capturing longer ranged structure in these liquids, such as the subtle valley in the tail of the ethanol O-O interaction. It perhaps worth noting here that, unlike GAFF, SDLJ assigns polar hydrogens a nonzero radius. For methylamine and benzene (Figure 8), SDLJ does a somewhat better job than GAFF of capturing the overall shape and details of these less peaked RDFs. Overall, SDLJ does a reasonable job of capturing the fine structure of these liquids, even though the method was not trained on these data.

4 Discussion

This study has demonstrated the feasibility of constructing a physics-based, QM-to-FF mapping, which generates LJ parameters that yield pure liquid properties whose accuracy is similar to that of the well-accepted GAFF force field, despite having many fewer adjustable parameters. The transferability of the mapping is supported by the fact that good results were obtained for test set compounds having functional groups not represented in the training set. This work is founded on important prior advances in AIM analysis[57], dispersion interactions[43], and automated parameter optimization[34].

A key feature of our approach is the abandonment of atom-typing in the assignment of LJ parameters. Instead, under the present schema, each atom of a molecule is assigned unique LJ parameters based on the QM calculation. This is advantageous, as it largely side-steps the challenge of categorizing atoms according to their chemical environment. Indeed, although the atom type categorizations used in today’s FFs are useful and are well-motivated by chemical logic, it is not clear that they represent an optimal balance between parsimony and accuracy, and the requirement for atom-typing has been cited as a problematic aspect of FF parameterization[42,76]. Recently, this problem has been addressed with a demonstration that the typing itself, rather than just the parameters associated with a fixed set of types, can be sampled effectively[36]. Here, we have considered a second approach, one which does away entirely with LJ types. Further work is needed to ascertain which, if either, of these broad approaches will be most effective. An advantage of atom-typing is that it allows *ad hoc* adjustments that may at times be helpful. On the other hand, the present approach is advantageous in that it can facilitate comprehensive parameter optimization or Bayesian sampling by reducing the number of adjustable parameters, and removes the need to sample over the typing itself. It is perhaps encouraging that one of the leading methods of assigning atomic partial charges, namely RESP, similarly eschews typing and instead using a purely physics-based QM-to-FF mapping.

The adjustment of the mapping parameters presumably allows them to capture or compensate for several issues in the primary QM calculation and the physical model used for the mapping. First, although the AIM concept is intuitively pleasing, is at best a physical approximation, so it is probably inevitable that some adjustment is needed. Second, even if an AIM analysis could provide flawless LJ parameters, adjustment would be needed to compensate for deficiencies in other FF terms, such as charge-charge interactions, for complexities that arise on going from gas to condensed phase, such as many-body effects, and for the neglect of nuclear quantum mechanics in typical classical simulations. By the same token, although gas-phase QM interaction data may be used to guide the adjustment of FF parameters, these interactions will inevitably change in subtle ways upon going into the condensed phase. We therefore chose to omit any gas phase QM data in the actual training of the mapping parameters, opting instead to use only condensed phase experimental observables.

Important recent studies have also used a tuned QM-to-FF mapping to assign LJ parameters without atom-typing[44,45]. The present approach is different in two key respects. The first difference is that we have modified only the LJ term and applied the GAFF and AM1/BCC partial charges without change, rather than simultaneously refitting bonded terms and adding off-atom partial charges. This approach makes it possible to isolate the effect of this one change on the accuracy of the FF, and also maintains the common functional form and thus compatibility with widely used simulation packages. It is worth noting that all such methods are expected to generate parameters that depend to some degree on the conformation of the molecule used in the QM calculations. In the present case, at least, initial testing shows only an encouragingly small dependence on conformation. It should nonetheless be noted that, if cases are encountered with the dependence on conformation is nontrivial, it should be possible to address these by averaging over thermodynamically accessible conformations and/or using molecular fragmentation approaches so that parameters can be assigned to relatively rigid molecular components.

The second difference is in the QM-to-FF mapping itself. In particular, prior studies have used the Tkatchenko-Scheffler (TS) approach[43], in which the AIM volume of each atom yields its AIM polarizability[77], which is used in turn to determine the coefficient of the dispersion interaction (B_i in our notation). Although the London dispersion interaction is determined by not only the polarizability but also the ionization energy[61], the TS approach assumes, in effect, that any effect of variations in the effective ionization energy across atoms on dispersion interactions is ultimately cancelled by other factors. As noted here (Section 2.4.2), this assumption causes all atoms of a given element (e.g. all carbons) to be assigned the identical value of epsilon, and thus, the same depth of the LJ energy well. Although this approach has led to methods of deriving non-bonded parameters which can provide good agreement with experiment[41,42,57], it runs counter to the empirical knowledge of typed force fields which allow both sigma and epsilon to vary for a given element. The present approach thus uses additional physical reasoning to extract not only the atomic volume but also an effective AIM ionization energy for each atom, and thus allows for variation in both sigma and epsilon across atoms of a given element.

As discussed above, the great potential benefit of the general approach taken here is the reduction in the dimensionality of the space of LJ parameters that need to be optimized or sampled. This is particularly important when evaluation of the objective function requires running simulations, as this causes each iteration to be computationally costly and thus increases the risk of missing key sectors of LJ parameter space. Thus, we envision applying the SDLJ approach within a broad FF optimization scheme, where the decreased dimensionality will make it easier to carry out a more thorough parameter optimization. We also anticipate using QM-to-FF mappings to generate other parameter types, such as torsional potentials. Additionally, we are working on a version of RESP charges which includes one adjustable parameter that scales the overall polarity of a molecule[81]. Together, these efforts could be combined in a coordinated optimization of an entire force-field with only a small set of parameters that need to be optimized against experimental observables. The relatively thorough sampling of parameter space this enables should not only lead to better parameters but also help to assess with greater confidence whether a proposed improvement in the functional form truly enables more accurate simulations than the starting functional form. Thus, the present approach is supportive of a systematic approach to advancing both the parameterization and the form of future force fields.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

MKG acknowledges funding from National Institute of General Medical Sciences (Grant number GM061300). LPW acknowledges funding from the National Institute of Allergy and Infectious Diseases (grant number AI130684). The contents of this paper are solely the responsibility of the authors and do not necessarily represent the official views of the NIH. MKG has an equity interest in and is a cofounder and scientific advisor of VeraChem LLC. MS acknowledges support of the Austrian Science Fund (FWF): Erwin Schroedinger fellowship J-4150.

References

1. Lindorff-Larsen K et al. Systematic Validation Of Protein Force Fields Against Experimental Data. PLOS ONE 7, E32131 (2012). [PubMed: 22384157]
2. Durrant JD & Mccammon JA Molecular Dynamics Simulations And Drug Discovery. BMC Biol. 9, 71 (2011). [PubMed: 22035460]
3. Kaminski G, Duffy EM, Matsui T & Jorgensen WL Free Energies Of Hydration And Pure Liquid Properties Of Hydrocarbons From The OPLS All-Atom Model. J. Phys. Chem 98, 13077–13082 (1994).
4. Lee T et al. GPU-Accelerated Molecular Dynamics And Free Energy Methods In Amber18: Performance Enhancements And New Features. J. Chem. Inf. Model 58, 10, 2043–2050 (2018) [PubMed: 30199633]
5. Salomon-Ferrer R, Götz AW, Poole D, Le Grand S & Walker RC Routine Microsecond Molecular Dynamics Simulations With AMBER On Gpus. 2. Explicit Solvent Particle Mesh Ewald. J. Chem. Theory Comput. 9, 3878–3888 (2013). [PubMed: 26592383]
6. Fine R, Dimmler G & Levinthal C FASTRUN: A Special Purpose, Hardwired Computer For Molecular Simulation. Proteins Struct. Funct. Bioinforma 11, 242–253 (1991).

7. Ito T Et al. A Special-Purpose Computer For Molecular Dynamics: GRAPE-2A. Proteins Struct. Funct. Bioinforma 20, 139–148 (1994).

8. Shaw DE. Anton 2: Raising The Bar For Performance And Programmability In A Special-Purpose Molecular Dynamics Supercomputer; SC14: International Conference For High Performance Computing, Networking, Storage And Analysis; 2014. 41–53.

9. Shaw DE et al. Millisecond-Scale Molecular Dynamics Simulations On Anton.SC ‘09 Article No. 39 (2009)

10. Barducci A, Bonomi M & Parrinello M Metadynamics. Wiley Interdiscip. Rev. Comput. Mol. Sci 1, 826–843 (2011).

11. Shirts MR & Chodera JD Statistically Optimal Analysis Of Samples From Multiple Equilibrium States. J. Chem. Phys 129, (2008).

12. Sugita Y & Okamoto Y Replica-Exchange Molecular Dynamics Method For Protein Folding. Chem. Phys. Lett 314, 141–151 (1999).

13. Miao Y, Feher VA & Mccammon JA Gaussian Accelerated Molecular Dynamics: Unconstrained Enhanced Sampling And Free Energy Calculation. J. Chem. Theory Comput. 11, 8, 3584–3595 (2015) [PubMed: 26300708]

14. Zheng L, Chen M & Yang W Random Walk In Orthogonal Space To Achieve Efficient Free-Energy Simulation Of Complex Systems. Proc. Natl. Acad. Sci 105, 20227–20232 (2008). [PubMed: 19075242]

15. Muddana HS, Fenley AT, Mobley DL & Gilson MK The SAMPL4 Host-Guest Blind Prediction Challenge: An Overview. J. Comput. Aided Mol. Des 28, 305–317 (2014). [PubMed: 24599514]

16. Muddana HS et al. Blind Prediction Of Host-Guest Binding Affinities: A New SAMPL3 Challenge. J. Comput. Aided Mol. Des 26, 475–487 (2012). [PubMed: 22366955]

17. Bosisio S, Mey ASJS & Michel J Blinded Predictions Of Host-Guest Standard Free Energies Of Binding In The SAMPL5 Challenge. J. Comput. Aided Mol. Des 31, 61–70 (2017). [PubMed: 27503495]

18. Yin J, Henriksen NM, Slochower DR & Gilson MK The SAMPL5 Host–Guest Challenge: Computing Binding Free Energies And Enthalpies From Explicit Solvent Simulations By The Attach-Pull-Release (APR) Method. J. Comput. Aided Mol. Des 31, 133–145 (2017). [PubMed: 27638809]

19. Geballe MT, et al. The SAMPL2 Blind Prediction Challenge: Introduction And Overview. J Comput Aided Mol Des. 24, 4, 259–79 (2010) [PubMed: 20455007]

20. Skillman AG SAMPL3: Blinded Prediction Of Host–Guest Binding Affinities, Hydration Free Energies, And Trypsin Inhibitors. J. Comput. Aided Mol. Des 26, 473–474 (2012). [PubMed: 22622621]

21. Nicholls A et al. Predicting Small-Molecule Solvation Free Energies: An Informal Blind Test For Computational Chemistry. J. Med. Chem 51, 769–779 (2008). [PubMed: 18215013]

22. Nerenberg PS & Head-Gordon T New Developments In Force Fields For Biomolecular Simulations. Curr. Opin. Struct. Biol 49, 129–138 (2018). [PubMed: 29477047]

23. Leach Andrew R. Molecular Modelling: Principles And Applications. Harlow, England: Prentice Hall, 2001.

24. Anisimov VM et al. Determination Of Electrostatic Parameters For A Polarizable Force Field Based On The Classical Drude Oscillator. J. Chem. Theory Comput. 1, 153–168 (2005). [PubMed: 26641126]

25. Maple JR et al. A Polarizable Force Field And Continuum Solvation Methodology For Modeling Of Protein-Ligand Interactions. J. Chem. Theory Comput 1, 694–715 (2005). [PubMed: 26641692]

26. Ponder JW et al. Current Status Of The AMOEBA Polarizable Force Field. J. Phys. Chem. B 114, 2549–2564 (2010). [PubMed: 20136072]

27. Wang J, Cieplak P, Luo R & Duan Y Development Of Polarizable Gaussian Model For Molecular Mechanical Calculations I: Atomic Polarizability Parameterization To Reproduce Ab Initio Anisotropy. J. Chem. Theory Comput. 15, 1146–1158 (2019). [PubMed: 30645118]

28. Van Vleet MJ, Misquitta AJ & Schmidt JR New Angles On Standard Force Fields: Toward A General Approach For Treating Atomic-Level Anisotropy. J. Chem. Theory Comput. 14, 2, 739–758 (2018) [PubMed: 29266931]

29. Freitag MA, Gordon MS, Jensen JH & Stevens WJ Evaluation Of Charge Penetration Between Distributed Multipolar Expansions. J. Chem. Phys 112, 7300–7306 (2000).

30. Gootz TD, Subashi TA & Lindner DL Simple Spectrophotometric Assay For Measuring Protein Binding Of Penem Antibiotics To Human Serum. Antimicrob. Agents Chemother. 32, 159–163 (1988). [PubMed: 3364940]

31. Rackers JA et al. An Optimized Charge Penetration Model For Use With The AMOEBA Force Field. Phys. Chem. Chem. Phys 19, 276–291 (2016). [PubMed: 27901142]

32. Zanette C et al. Toward Learned Chemical Perception Of Force Field Typing Rules. J. Chem. Theory Comput. 15, 1, 402–423 (2018) [PubMed: 30512951]

33. Betz RM & Walker RC Paramfit: Automated Optimization Of Force Field Parameters For Molecular Dynamics Simulations. J. Comput. Chem 36, 79–87 (2015). [PubMed: 25413259]

34. Wang L-P, Martinez TJ & Pande VS Building Force Fields: An Automatic, Systematic, And Reproducible Approach. J. Phys. Chem. Lett 5, 1885–1891 (2014). [PubMed: 26273869]

35. Sen G et al. Comparing Optimization Strategies For Force Field Parameterization. Https://Arxiv.Org/Abs/1812.00326. (2018)

36. Brommer P & Gähler F Potfit: Effective Potentials From Ab-Initio Data. Model. Simul. Mater. Sci. Eng 15, 295–304 (2007).

37. Hülsmann M et al. Optimizing Molecular Models Through Force-Field Parameterization Via The Efficient Combination Of Modular Program Packages In Foundations Of Molecular Modeling And Simulation: Select Papers From FOMMS 2015 (Eds. Snurr RQ, Adjiman CS & Kofke DA) 53–77 (Springer, 2016).

38. Faller R, Schmitz H, Biermann O & Müller-Plathe F Automatic Parameterization Of Force Fields For Liquids By Simplex Optimization. J. Comput. Chem 20, 1009–1017 (1999).

39. Mobley DL et al. Escaping Atom Types In Force Fields Using Direct Chemical Perception. J. Chem. Theory Comput. 14, 6076–6092 (2018). [PubMed: 30351006]

40. Ge Y & Voelz VA Model Selection Using Biceps: A Bayesian Approach For Force Field Validation And Parameterization. J. Phys. Chem. B 122, 5610–5622 (2018). [PubMed: 29518328]

41. Wu S, Angelikopoulos P, Papadimitriou C, Moser R & Koumoutsakos P A Hierarchical Bayesian Framework For Force Field Selection In Molecular Dynamics Simulations. Philos. Trans. R. Soc. Math. Phys. Eng. Sci 374, 20150032 (2016).

42. Dutta R, Brotzakis ZF & Mira A Bayesian Calibration Of Force-Fields From Experimental Data: TIP4P Water. J. Chem. Phys 149, 154110 (2018). [PubMed: 30342443]

43. Tkatchenko A & Scheffler M Accurate Molecular Van Der Waals Interactions From Ground-State Electron Density And Free-Atom Reference Data. Phys. Rev. Lett 102, 073005 (2009). [PubMed: 19257665]

44. Cole DJ, Vilseck JZ, Tirado-Rives J, Payne MC & Jorgensen WL Biomolecular Force Field Parameterization Via Atoms-In-Molecule Electron Density Partitioning. J. Chem. Theory Comput. 12, 2312–2323 (2016). [PubMed: 27057643]

45. Horton JT, Allen AEA, Dodda LS & Cole DJ Qubekit: Automating The Derivation Of Force Field Parameters From Quantum Mechanics. J. Chem. Inf. Model 59, 1366–1381 (2019). [PubMed: 30742438]

46. Vandenbrande S, Waroquier M, Speybroeck V & Verstraelen T The Monomer Electron Density Force Field (MEDFF): A Physically Inspired Model For Noncovalent Interactions. J. Chem. Theory Comput. 13, (2016).

47. Swart M & Duijnen PTV Atomic Radii In Molecules For Use In A Polarizable Force Field. Int. J. Quantum Chem. 111, 1763–1772 (2011).

48. Visscher KM & Geerke DP Deriving Force-Field Parameters From First Principles Using A Polarizable And Higher Order Dispersion Model. J. Chem. Theory Comput. 15, 1875–1883 (2019). [PubMed: 30763086]

49. Pérez De La Luz A, Aguilar-Pineda JA, Méndez-Bermúdez JG & Alejandre J Force Field Parametrization From The Hirshfeld Molecular Electronic Density. J. Chem. Theory Comput. 14, 5949–5958 (2018). [PubMed: 30278120]

50. Mei Y et al. Numerical Study On The Partitioning Of The Molecular Polarizability Into Fluctuating Charge And Induced Atomic Dipole Contributions. J. Phys. Chem. A 119, 5865–5882 (2015). [PubMed: 25945749]

51. Wang J, Wolf RM, Caldwell JW, Kollman PA & Case DA Development And Testing Of A General Amber Force Field. J. Comput. Chem 25, 1157–1174 (2004). [PubMed: 15116359]

52. Hanwell MD et al. Avogadro: An Advanced Semantic Chemical Editor, Visualization, And Analysis Platform. J. Cheminformatics 4, 17 (2012).

53. Frisch M et al. Gaussian 09, Revision B.01. Gaussian 09 Revis. B01 Gaussian Inc Wallingford CT (2009).

54. Kummel H A Biography Of The Coupled Cluster Method Recent Profess In Many Body Theories. Proceedings Of The 11th International Conference. 17 5311–5325 (2003).

55. Lebedev VI Values Of The Nodes And Weights Of Ninth To Seventeenth Order Gauss-Markov Quadrature Formulae Invariant Under The Octahedron Group With Inversion. USSR Comput. Math. Math. Phys 15, 44–51 (1975).

56. Introducing DDEC6 Atomic Population Analysis: Part 1. Charge Partitioning Theory And Methodology. RSC Adv. 6, 47771–47801 (2016)

57. Heidar-Zadeh F, Ayers PW How Pervasive Is The Hirshfeld Partitioning? J. Chem. Phys 142, 044107 (2015) [PubMed: 25637969]

58. Verstraelen T et al. Hirshfeld-E Partitioning: AIM Charges With An Improved Trade-Off Between Robustness And Accurate Electrostatics J. Chem. Theory Comput. 9, 5, 2221–2225 (2013) [PubMed: 26583716]

59. Verstraelen T et al. Minimal Basis Iterative Stockholder: Atoms In Molecules For Force-Field Development. J. Chem. Theory Comput. 12, 3894–3912 (2016). [PubMed: 27385073]

60. Hoffmann-Ostenhof M & Hoffmann-Ostenhof T ‘Schrodinger Inequalities” And Asymptotic Behavior Of The Electron Density Of Atoms And Molecules. Phys. Rev. A 16, 1782–1785 (1977).

61. Van Vleet MJ, Misquitta AJ, Stone AJ & Schmidt JR Beyond Born-Mayer: Improved Models For Short-Range Repulsion In Ab Initio Force Fields. J. Chem. Theory Comput. 12, 3851–3870 (2016). [PubMed: 27337546]

62. Ueber Die Anwendung Des Satzes Vom Virial In Der Kinetischen Theorie Der Gase - Lorentz - 1881 - Annalen Der Physik - Wiley Online Library Https://Onlinelibrary.Wiley.Com/Doi/Abs/10.1002/Andp.18812480110.

63. Eisenschitz R & London F Über Das Verhältnis Der Van Der Waalsschen Kräfte Zu Den Homöopolaren Bindungskräften. Z. Für Phys. 60, 491–527 (1930).

64. Jakalian A, Jack D & Bayly C Fast, Efficient Generation Of High Quality Atomic Charges. AM1 - BCC Model: II. Parameterization And Validation. J. Comput. Chem 23, 1623–41 (2002). [PubMed: 12395429]

65. Wang J, Wolf RM, Caldwell JW, Kollman PA & Case DA Development And Testing Of A General Amber Force Field. J. Comput. Chem 25, 1157–1174 (2004). [PubMed: 15116359]

66. Wang J, Wang W, Kollman PA & Case DA Automatic Atom Type And Bond Type Perception In Molecular Mechanical Calculations. J. Mol. Graph. Model 25, 247–260 (2006). [PubMed: 16458552]

67. Salomon-Ferrer R, Case DA & Walker RC An Overview Of The Amber Biomolecular Simulation Package. Wiley Interdiscip. Rev. Comput. Mol. Sci 3, 198–210 (2013).

68. Wang L-P et al. Systematic Improvement Of A Classical Molecular Model Of Water. J. Phys. Chem. B 117, 9956–9972 (2013). [PubMed: 23750713]

69. Flyvbjerg H & Petersen HG Error Estimates On Averages Of Correlated Data. J. Chem. Phys 91, 461–466 (1989).

70. Henriksen NM, Fenley AT & Gilson MK Computational Calorimetry: High-Precision Calculation Of Host–Guest Binding Thermodynamics. J. Chem. Theory Comput. 11, 4377–4394 (2015). [PubMed: 26523125]

71. Frenkel M et al. Thermomlan XML-Based Approach For Storage And Exchange Of Experimental And Critically Evaluated Thermophysical And Thermochemical Property Data. 1. Experimental Data. J. Chem. Eng. Data 48, 2–13 (2003).

72. Enthalpies Of Vaporization Of Organic Compounds: A Critical Review And Data Compilation. (Blackwell Scientific, 1985).

73. Mathews JH The Accurate Measurement Of Heats Of Vaporization Of Liquids J. Am. Chem. Soc 48, 3, 562–576 (1926)

74. Figueroa-Gerstenmaier S, Giudice S, Cavallo L & Milano G A Molecular Model For H2 Interactions In Aliphatic And Aromatic Hydrocarbons. Phys. Chem. Chem. Phys. PCCP 11, 3935–42 (2009). [PubMed: 19440622]

75. Kosztolányi T, Bakó I & Palinkas G Hydrogen Bonding In Liquid Methanol, Methylamine, And Methanethiol Studied By Molecular-Dynamics Simulations. J. Chem. Phys 118, 4546–4555 (2003).

76. Saiz L, Padró JA & Guàrdia E Structure And Dynamics Of Liquid Ethanol. J. Phys. Chem. B 101, 78–86 (1997).

77. PTRAJ And CPPTRAJ: Software For Processing And Analysis Of Molecular Dynamics Trajectory Data. - Pubmed - NCBI. Https://Www.Ncbi.Nlm.Nih.Gov/Pubmed/26583988.

78. Lim VT, et al. Assessing The Conformational Equilibrium Of Carboxylic Acid Via Quantum Mechanical And Molecular Dynamics Studies On Acetic Acid | Journal Of Chemical Information And Modeling. J. Chem. Inf. Model 59, 5, 1957–1964 (2019). [PubMed: 30742770]

79. Grimme S A General Quantum Mechanically Derived Force Field (QMDFF) For Molecules And Condensed Phase Simulations. J. Chem. Theory Comput. 10, 4497–4514 (2014). [PubMed: 26588146]

80. Brinck T, Murray JS & Politzer P Polarizability And Volume. J. Chem. Phys 98, 4305–4306 (1993).

81. Schauperl M et al. Force Field Partial Charges With Restrained Electrostatic Potential 2 (RESP2). (2019) Doi:10.26434/Chemrxiv.10072799.V1.

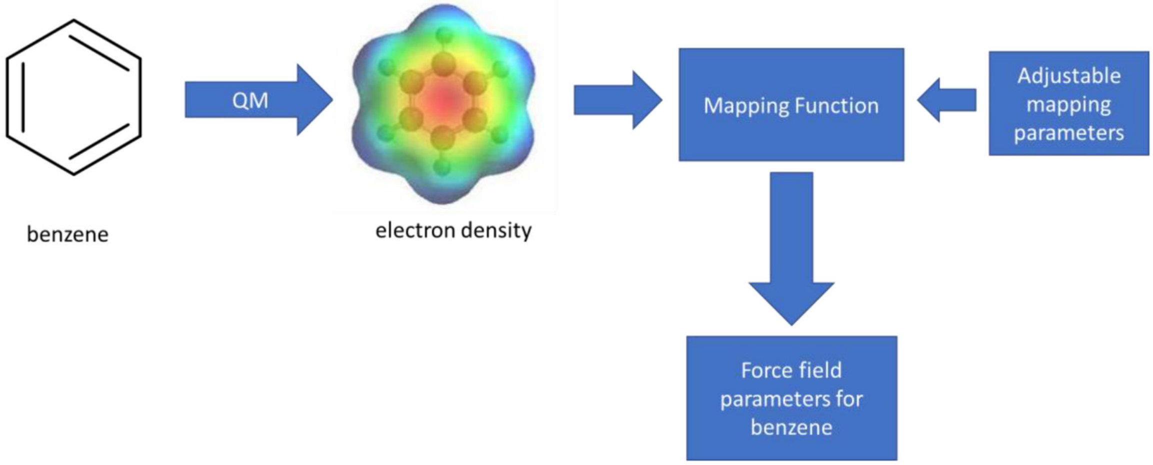

benzene
QM
electron density
Mapping Function
Adjustable mapping parameters
Force field parameters for benzene

Figure 1.

Generic scheme for use of a parameterized mapping from QM results to FF parameters, with benzene as an example.

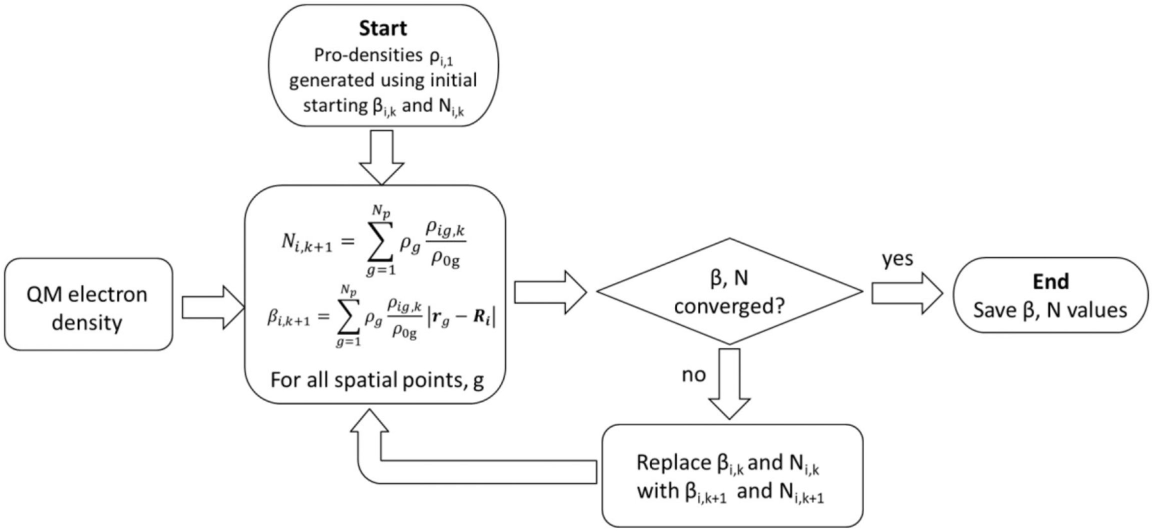

Start
Pro-densities $\rho_{i,1}$ generated using initial starting $\beta_{i,k}$ and $N_{i,k}$
QM electron density
$N_{i,k+1} = \sum_{g=1}^{N_p} \rho_g \frac{\rho_{ig,k}}{\rho_{0g}}$
$\beta_{i,k+1} = \sum_{g=1}^{N_p} \rho_g \frac{\rho_{ig,k}}{\rho_{0g}} |\boldsymbol{r}_g - \boldsymbol{R}_i|$
For all spatial points, g
β, N converged?
yes
End
Save β, N values
no
Replace $\beta_{i,k}$ and $N_{i,k}$ with $\beta_{i,k+1}$ and $N_{i,k+1}$

Figure 2:

Schematic of the iterative stockholder method used to fit β and N for each atom-in-molecule i to the molecular electron density from a QM calculation. Symbols are defined in the main text.

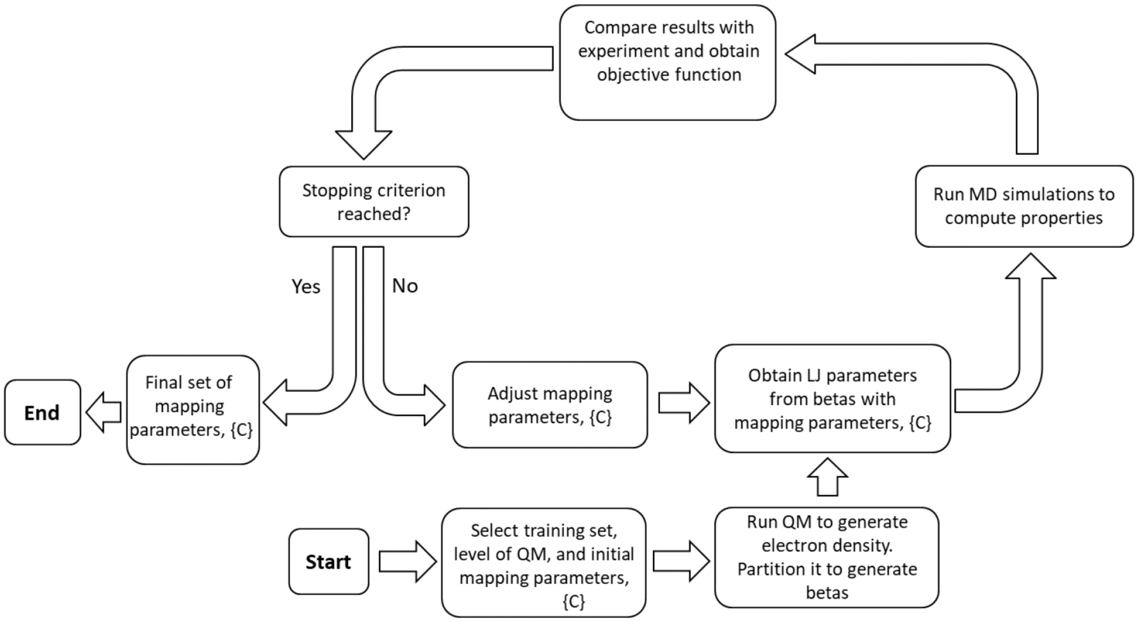

Compare results with experiment and obtain objective function
Stopping criterion reached?
Run MD simulations to compute properties
Yes
No
End
Final set of mapping parameters, {C}
Adjust mapping parameters, {C}
Obtain LJ parameters from betas with mapping parameters, {C}
Start
Select training set, level of QM, and initial mapping parameters, {C}
Run QM to generate electron density. Partition it to generate betas

Figure 3:

Optimization of the ten QM-to-FF mapping parameters. Quantum mechanical calculations are carried out on training set molecules to compute beta for each atom. These values are mapped to LJ parameters using the mapping parameters. Simulations are carried out to compare simulation results to experiment, and the parameters are iteratively updated based on the gradient of the objective function in parameter space. Once the mapping parameters are optimized, they are saved and used to generate LJ parameters for other molecules.

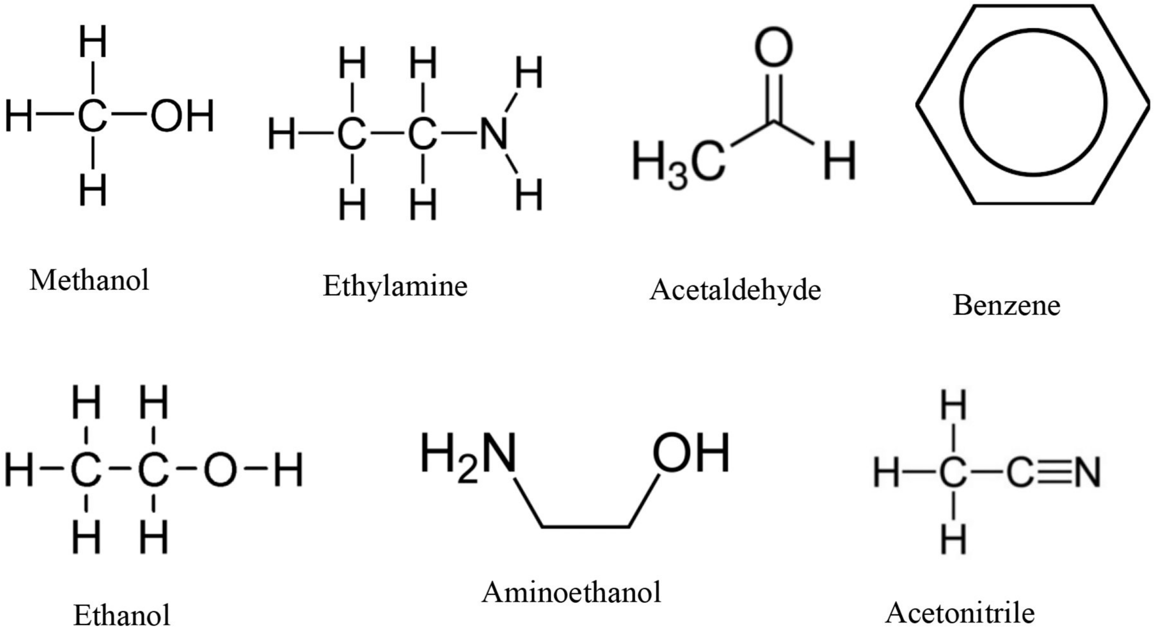


Figure 4:
Molecules used to train the elemental mapping parameters.

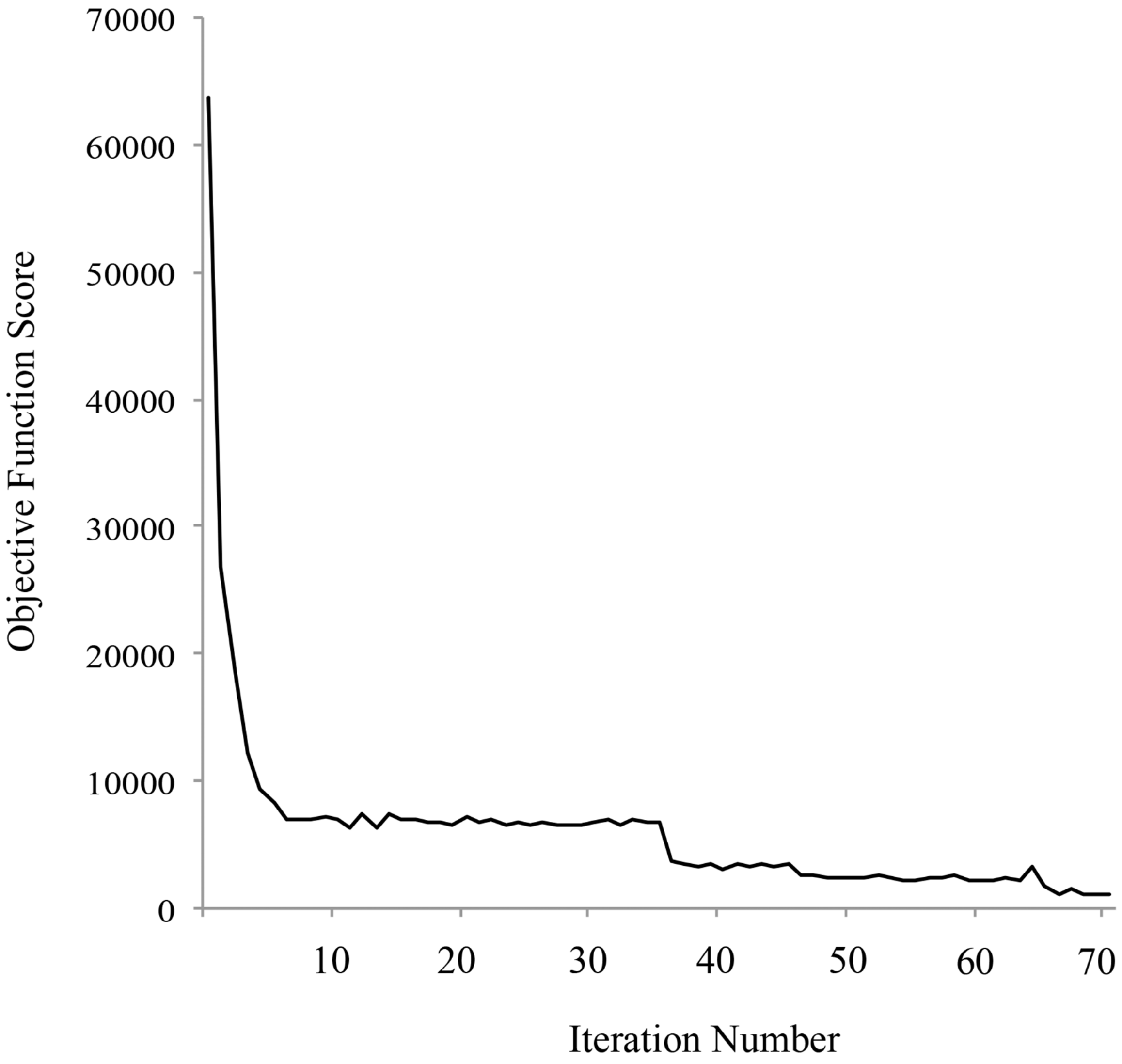


Figure 5:

History of objective function of training set over entirety of the optimization run using ForceBalance.

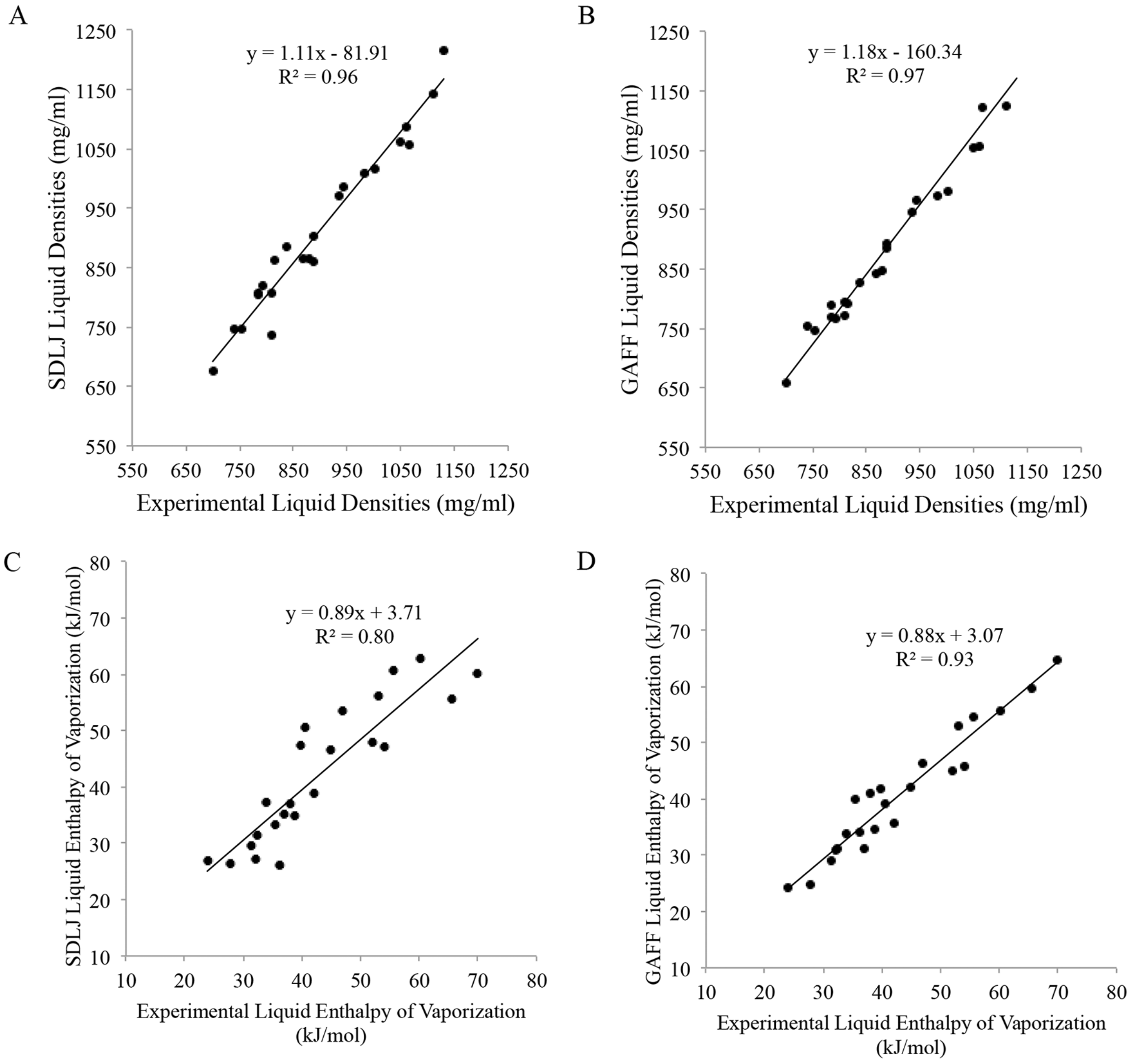


Figure 6:

Scatter plots of experimental values of density and heat of vaporization versus SDLJ or GAFF simulated results. (A) Experimental and SDLJ densities. (B) Experimental and GAFF densities. (C) Experimental and SDLJ heats of vaporization. (D) Experimental and GAFF heats of vaporization. Linear regression results are provided on each panel.

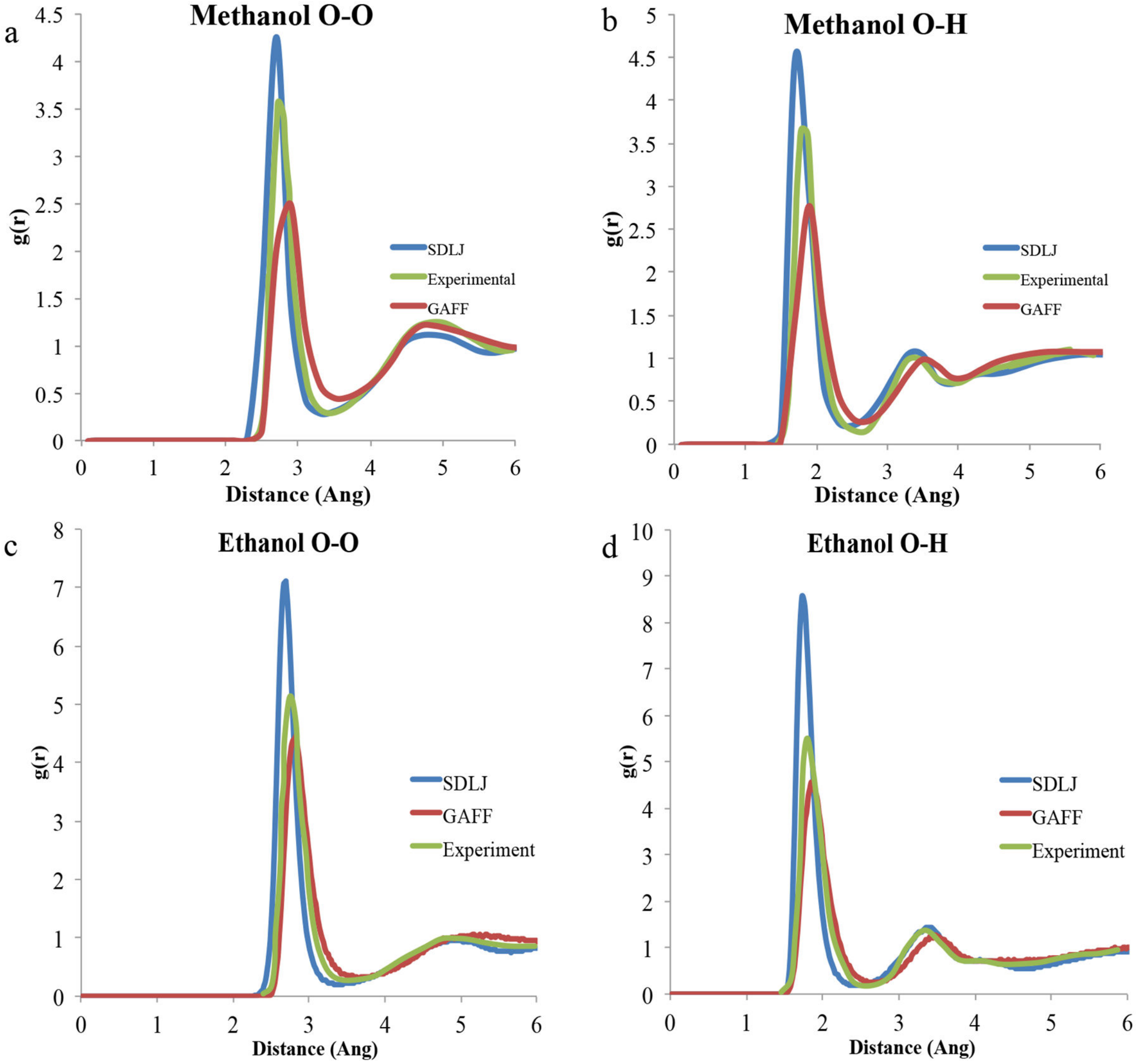


Figure 7:

Radial distribution functions of various pair interactions in neat methanol and ethanol.

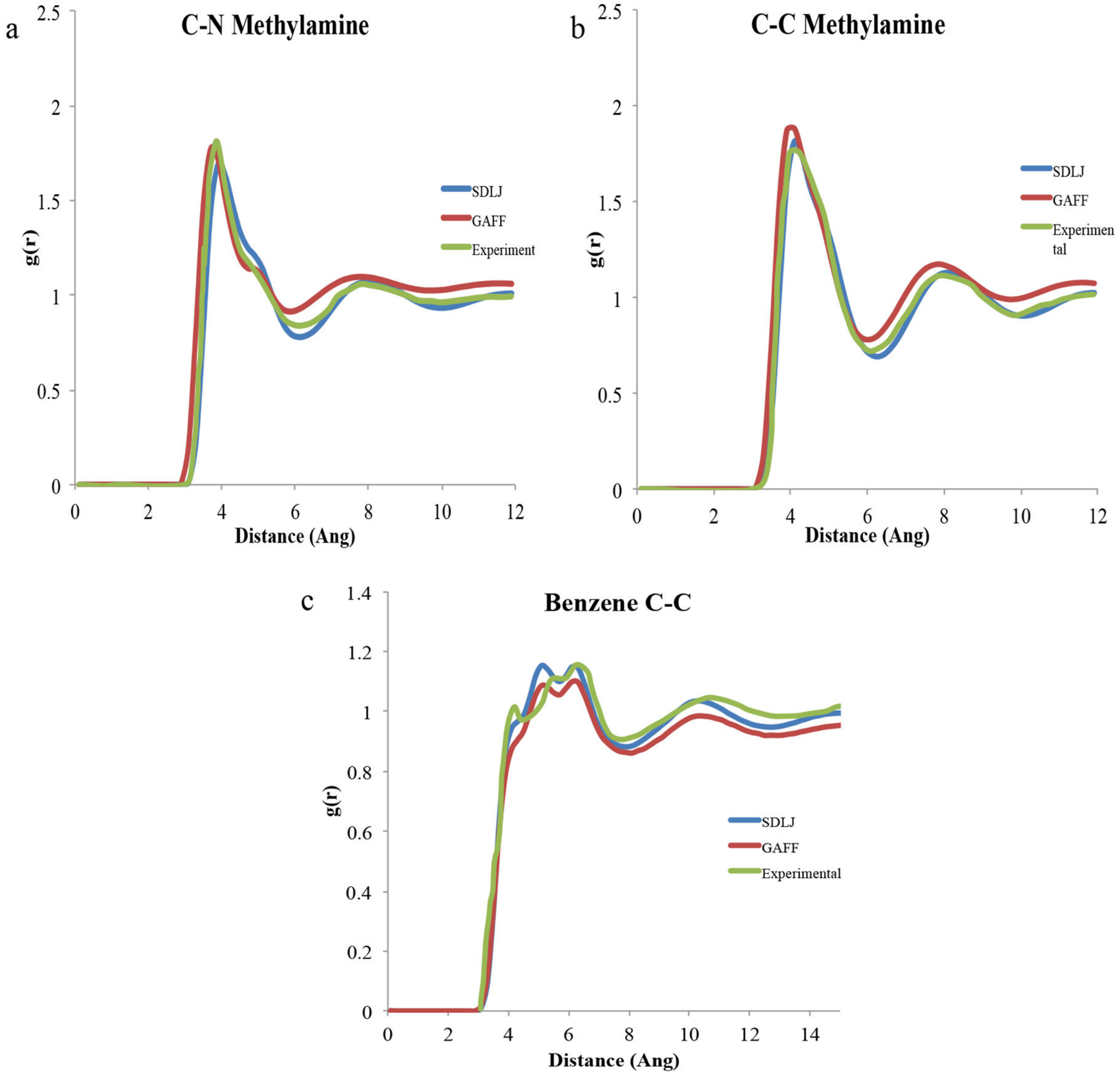


Figure 8:

Radial distribution functions of various pair interactions in neat methylamine and benzene.

Table 1.

Final results for training set molecules. Percent mean unsigned errors (MUE) are computed as $100\sum_{i}^{N_{data}}\frac{|C_i - E_i|}{C_i}$ where N_{data} is the number of data (compounds), and C_i and E_i are the computed and experimental quantities, respectively, for compound *i*. Uncertainties were obtained by the blocking method (see Methods).

	Density (mg/ml)			**ΔH_{Vap} (kJ/mol)**		
Compound	**Experimental**	**Computed**	**±**	**Experimental**	**Computed**	**±**
Methanol	784	762.5	0.50	31.3	36.1	0.28
Ethanol	789	784.0	0.55	42.3	42.3	0.68
Aminoethanol	1011	1024.8	0.34	58.0	60.2	0.71
Ethylamine	688	709.1	0.21	29.0	29.3	0.51
Acetaldehyde	784	813.0	0.41	26.1	28.3	0.50
Benzene	876	880.9	0.53	33.9	32.6	0.52
Acetonitrile	786	747.3	0.40	33.4	31.8	0.52
MUE		**2.4%**			**5.0%**	

Table 2:

Values of Rmin/2 and ε for training set molecules generated from the final optimized values of the elemental mapping parameters.

Molecule	Atom	$R_{min}/2$ (Å)	ε (kcal/mol)
Methanol			
	c	1.936	0.079
	h-c	1.448	0.025
	o	1.486	0.149
	h-o	0.776	0.028
Ethanol			
	c1	1.916	0.081
	c2	2.012	0.074
	h-c1	1.439	0.025
	h-c2	1.423	0.025
	o	1.488	0.148
	h-o	0.779	0.028
Aminoethanol			
	c-n	1.946	0.079
	c-o	1.939	0.079
	o	1.497	0.146
	n	1.958	0.304
	h-n	0.782	0.027
	h-o	0.755	0.029
	h-c-n	1.412	0.026
	h-c-o	1.447	0.025
Ethylamine			
	c1	2.010	0.074
	c2	1.928	0.080
	h-c1	1.426	0.025
	h-c2	1.435	0.025
	n	1.957	0.304
	h-n	0.802	0.026
Acetaldehyde			
	c1	1.999	0.075
	c2	1.843	0.088
	h-c1	1.419	0.026
	h-c2	1.423	0.025
	o	1.478	0.150
Benzene			
	c	1.972	0.076
	h-c	1.368	0.028
Acetonitrile			

Molecule	**Atom**	**$R_{min}/2$ (Å)**	**ε (kcal/mol)**
	c-c	1.986	0.075
	c-n	1.845	0.088
	h-c-n	1.374	0.027
	n	1.891	0.326

Table 3.

Test-set results for SDLJ method and GAFF. See Table 1 for definitions. Uncertainties were obtained by the blocking method (see Methods).

	Density (mg/ml)					ΔH_{VAP} (kJ/mol)				
Compound	**Exp.**	**SDLJ**	**±**	**GAFF**	**±**	**Exp.**	**SDLJ**	**±**	**GAFF**	**±**
Propyl acetate	888	903.8	0.29	892.3	0.44	38	37.1	0.59	41.1	0.62
Acetone	784	805.5	0.34	769.0	0.28	31.3	29.5	0.44	29.0	0.45
THF	889	859.9	0.42	884.5	0.32	32.2	27.1	0.51	30.9	0.56
DMF	944	987.5	0.27	965.7	0.54	46.8	53.6	0.51	46.4	0.57
Propionitrile	792	818.6	0.48	767.4	0.29	36.3	26.1	0.41	34.0	0.42
Methylamine	700	677.3	0.29	658.9	0.51	23.8	26.8	0.38	24.2	0.31
Butylamine	740	745.6	0.24	753.6	0.45	34	37.3	0.62	33.9	0.59
Butanol	810	735.7	0.49	794.8	0.29	52	48.0	0.71	45.0	0.68
Isopropanol	786	808.2	0.29	790.0	0.35	45	46.5	0.43	42.2	0.59
Glycol	1110	1143.8	0.47	1125.2	0.39	65.6	55.7	0.55	59.7	0.43
Phenyl-2-propanone	1001	1015.6	0.41	981.7	0.25	55.5	60.6	0.68	54.6	0.73
Furan	936	971.0	0.42	946.4	0.41	27.7	26.4	0.46	24.9	0.36
Formamide	1130	1214.5	0.39	1260.9	0.96	60.2	62.9	0.37	55.7	0.37
Propenal	839	884.9	0.43	828.6	0.34	32.3	31.4	0.34	31.1	0.41
Dioxolane	1065	1057.3	0.40	1123.6	0.41	35.5	33.3	0.50	40.0	0.44
Propenoic Acid	1050	1063.0	0.44	1055.5	0.32	53.1	56.2	0.59	52.9	0.55
Toluene	867	865.5	0.32	843.3	0.34	37	35.1	0.52	31.1	0.54
1,3-propanediol	1060	1086.3	1.72	1056.4	0.54	70	60.1	0.59	64.8	1.09
3-Pentanone	809	807.6	0.38	772.7	0.31	38.7	34.9	0.66	34.5	0.54
o-xylene	880	864.8	0.28	848.1	0.61	42	38.9	0.64	35.8	2.85
Pyridine	982	1009.8	0.49	974.8	0.40	40.5	50.5	0.51	39.2	0.42
Pentylamine	755	746	0.39	745.9	0.36	39.7	47.5	0.62	41.8	0.56
3-Pentanol	815	863.2	0.38	793.0	0.29	54	47.0	0.70	45.9	0.63
%MUE		**3.0%**		**2.5%**			**11%**		**7.8%**	
MSE		**13.8**		**−0.1**			**−0.8**		**−2.2**	