Accurate calculation of many-body energies in water clusters using a classical geometry-dependent induction model

Kristina M. Herman,† Anthony J. Stone,*† and Sotiris S. Xantheas*¶†

†Department of Chemistry, University of Washington, Seattle, WA 98185, USA
‡Yusuf Hamied Department of Chemistry, University of Cambridge, CB2 1EW, UK
¶Advanced Computing, Mathematics and Data Division, Pacific Northwest National Laboratory, 902 Battelle Boulevard, P.O. Box 999, MSIN J7-10, Richland, WA 99352, USA

E-mail: ajs1@cam.ac.uk; sotiris.xantheas@pnnl.gov,xantheas@uw.edu

Abstract

We incorporate geometry-dependent distributed multipole and polarizability surfaces into an induction model that is used to describe the 3- and 4-body terms of the interaction between water molecules. The expansion is carried out up to hexadecapole with the multipoles distributed on the atom sites. Dipole-dipole, dipole-quadrupole, and quadrupole-quadrupole distributed polarizabilities are used to represent the response of the multipoles to an electric field. We compare the model against two large databases consisting of 43,844 3-body terms and 3,603 4-body terms obtained from high level ab initio calculations previously used to fit the MB-pol and q-AQUA interaction potentials. The classical induction model with no adjustable parameters reproduces the ab-initio 3- and 4-body terms contained in these two databases with a Root-Mean-Square-Error (RMSE) of 0.104/0.058 and a Mean-Absolute-Error (MAE) of 0.054/0.026 kcal/mol, respectively, results that are on a par with those obtained

1
by fitting the same data using tens of thousands of Permutationally Invariant Polynomials (PIPs). This demonstrates the accuracy of this physically motivated model in describing the 3- and 4-body terms in the interactions between water molecules with no adjustable parameters. The triple-dipole-dispersion energy was included in the 3-body energy and was found to be small but not quite negligible. The model represents a practical, efficient and transferable approach for obtaining accurate non-additive interactions for multi-component systems without the need of performing tens of thousands of high level electronic structure calculations and fitting them with tens of thousands of PIPs.

Main text

The notoriously strong many-body interactions in aqueous systems\textsuperscript{1–7} necessitate the use of complex and sophisticated force fields in order to reproduce their magnitude and variation with hydrogen bonding environment including those in the presence of solutes such as ions.\textsuperscript{8–28} Initially, most of the interaction potentials developed for water were pairwise additive, truncating the many-body expansion at the second order into an effective term that was designed to fold in the missing higher order terms. This was mainly achieved through the use of an enhanced static dipole moment. It was soon realized that an environment-dependent dipole moment, rather than an enhanced static one, was essential for ensuring the transferability of the potentials from small to larger water clusters (\(n=2\text{–}25\)).\textsuperscript{29} Furthermore, many-body effects have been shown to be critical in describing the structure of water\textsuperscript{30} and, in particular, its local structure near an ion.\textsuperscript{31} In order to capture the many-body nature of the interactions, a model must either be fit to high level \textit{ab initio} data to yield the 3-body (and higher) energy terms explicitly or it must be polarizable to account for the change in the charge distribution due to the field from other neighboring molecules, thus implicitly incorporating many-body effects into the model. In this paper, we incorporate geometry-dependent distributed multipole and polarizability surfaces into an induction
model for the 3- and 4-body terms of water and demonstrate that this physically motivated model describes these many-body terms with an accuracy that is on par with models fit to a large database of 43,844 3-body terms and 3,603 4-body terms obtained from high level \textit{ab initio} calculations.

A popular method of accurately representing many-body effects in classical potentials is through permutationally invariant polynomials (PIPs)\textsuperscript{32} fit to a large number (typically in the tens of thousands) of \textit{ab initio} calculations. This approach was pioneered by Bowman and co-workers in the WHBB\textsuperscript{15} and q-AQUA\textsuperscript{16} potentials, and has also been used extensively by Paesani and co-workers in the fitting of the MB-pol potential.\textsuperscript{13,14,33,34} These potentials were developed by correcting the deficiency in describing shorter-range interactions by fitting either a short-range correction or a stand-alone potential to extensive \textit{ab initio} data. More specifically, the WHBB potential was fit to roughly 40,000 MP2/aug-cc-pVTZ calculations to add a 3-body short-range correction to the TTM3-F potential,\textsuperscript{11} while the MB-pol potential was fit to >12,000 BSSE-corrected CCSD(T)/aug-cc-pVTZ calculations to add a 3-body short-range correction to the TTM4-F potential.\textsuperscript{12} Recently, the q-AQUA potential was fit to >45,000 BSSE-corrected CCSD(T)-F12a/aug-cc-pVTZ calculations to develop a standalone 3-body potential. The q-AQUA potential was also the first developed potential with a fitted explicit 4-body term, which was accomplished with over 3,000 CCSD(T)/aug-cc-pVTZ 4-body values. Inarguably, this approach has led to the construction of very accurate potential energy surfaces.\textsuperscript{29,35,36} However, there are a few notable disadvantages of this strategy. These consist of the high level of parameterization (hundreds to thousands of parameters) and the large number of very accurate \textit{ab initio} calculations required to develop the database that is then used to fit these potentials. As one extends to higher orders of the MBE (i.e., the 3- and 4-body terms), the \textit{ab initio} calculations for each geometry become more expensive due to the increase in system size and the number of associated geometries that are needed to ensure sufficient sampling of the conformational space also increases due to the surging of the degrees of freedom. Most importantly, incorporation of solutes requires that the process
be repeated for each many-body term containing the new species. That said, if a classical, transferable model could be used to model many-body effects (3-body and higher), it would leave only the 1- and 2-body terms to be either fit to \textit{ab initio} results (i.e. using PIPs) or calculated on-the-fly using high level \textit{ab initio} methods.

Polarizable models, on the other hand, implicitly incorporate many-body effects. These potentials are typically founded on the multipole expansion, which expands the charge distribution of an atom or molecule in spherical harmonics. This expansion can yield multipoles at a single site or it can be distributed across multiple sites (often selected to be the atom sites) to improve the performance for shorter-range interactions.\textsuperscript{37–39} The interactions between multipoles are strictly pairwise. However, allowing these multipoles to change or polarize in response to the field of neighboring molecules gives rise to many-body effects. A multipolar description of the molecular charge distribution of the equilibrium geometry of water has already been implemented in force fields to varying degrees.\textsuperscript{8–12,17–21,23–26,40} However, it is expected that the distributed multipoles and polarizabilities change as a function of the intramolecular geometry. The necessity of representing this charge flux in potentials has been previously discussed by Dinur.\textsuperscript{41} The TTM2.1-F revision of the TTM2-F potential\textsuperscript{8} and the AMOEBA+CF\textsuperscript{22} potential incorporates, to some extent, the redistribution of charge density with intramolecular geometry by including a geometry-dependent correction to the atomic monopoles. Very recently, the SCME-f model\textsuperscript{42} introduced a single-site multipolar description of water which has a dipole and a quadrupole moment, both dependent on the intramolecular geometry. To the best of our knowledge, the use of a geometry-dependent distributed multipole (beyond monopole) and distributed polarizability surface in a classical potential, as the one presented in this study, has not been previously explored. However, it should be noted that dipole moment surfaces and dipole and quadrupole polarizability surfaces\textsuperscript{43} have been obtained in order to compute the infrared and Raman spectra of water, respectively. Notably, these surfaces, like SCME-f, yield a single molecular dipole moment or dipole-dipole polarizabilities rather than distributed multipoles and polarizabilities at each
atom site. By distributing the multipoles and polarizabilities across atom sites, a better representation of the charge redistribution can be achieved.\textsuperscript{37–39}

The success of a classical representation of the 3-body term with a detailed induction model, albeit based on moments and polarizabilities at the equilibrium geometry, has recently been demonstrated, including its transferability across different water-water-water (W-W-W) and ion-water-water (I-W-W) subsystems with only a single adjustable parameter.\textsuperscript{44} While it had been anticipated that induction energy would be the main contributor to many-body effects, this previous study demonstrated the quantitative accuracy of the induction model in doing so against accurate \textit{ab initio} benchmarks, hinting at the potential of this model to rival the accuracy of fitted 3-body terms using tens of thousands of PIPs. Furthermore, the minimal parameterization (a single damping parameter) and the naturally transferable nature of this route makes it particularly attractive. However, the previous model was limited to a multipolar and polarization description that was static in nature (only computed at the equilibrium monomer geometry). The description of distorted monomer geometries is crucial for modeling liquid water. For instance, during a molecular dynamics simulation the monomer geometries sampled can be far from the equilibrium geometry. Consequently the model must capture the variation of the distributed multipoles and polarizabilities with these highly distorted intramolecular geometries.

The goal of this work is to expand our previous induction model, which described the interactions between water molecules by implementing geometry-dependent distributed multipoles (dm) and polarizabilities (dp) and to extend it to the 4-body term. We have obtained distributed multipole and distributed polarizability surfaces as functions of the monomer intramolecular coordinates $\theta_{\text{HOH}}$, $R_{\text{ave}}$ (average of the two $R_{\text{OH}}$ distances), and $\Delta R$ (difference between the two $R_{\text{OH}}$ distances) (Figure 1) over a wide range of intramolecular geometries. The linearly interpolated surfaces are then used to re-examine the 3-body terms from a recent paper\textsuperscript{44} and demonstrate the improvement of the flexible water molecules in the 3-body description of (W-W-W) 3-body terms. The contribution from the triple-dipole-dispersion
interaction to the total 3-body interaction is also examined using the Axilrod-Teller-Muto potential.\textsuperscript{45,46} We also extend the developed classical induction model to the 4-body term in water. The results of this work demonstrate the practicality of the classical induction approach to accurately model the 3- and 4-body terms in water. More broadly, it demonstrates that an induction model is capable of achieving an accuracy that is on a par with potentials explicitly fitted to tens of thousands of \textit{ab initio} calculations with PIPs (containing up to tens of thousands of parameters). Most importantly, it also addresses the issue of transferability upon including an additional, different solvent molecule without the need to perform additional tens of thousands of high level electronic structure calculations for each system studied.

![Figure 1: The 3 coordinates describing the intramolecular geometry of a water molecule, used as a basis for the calculation of the distributed multipole and polarizability surfaces.](image)

The water monomer coordinate system ($\theta_{\text{HOH}}$, $R_{\text{ave}}$, $\Delta R$) used to evaluate the multipoles (up to hexadecapole) and polarizabilities (up to quadrupole-quadrupole) is shown in Figure 1. Details of the grid size, the process for selecting the range of values, and the level of theory used to compute the multipole and polarizability surfaces are described in the computational details section. The range of values of the monomer coordinates used for the calculation of the distributed multipole and polarizability surfaces, taken from Database A and a molecular dynamics simulation with the TTM2.1-F potential at $T=300K$, are
shown in the Supporting Information (Figure S1 and Table S1). Representations of the 3-dimensional surfaces depicting the value of each multipole or polarizability as a function of the \((θ_{\text{HOH}}, R_{\text{ave}}, ΔR)\) coordinates are included in Figures S2-S3 and are available via GitHub (https://github.com/kmherman/dm-dp-surfaces/H2O). In general, these surfaces show smooth, continuous changes in value as the intramolecular geometry changes. For the polarizability surfaces, there appears to be a more abrupt change in distributed polarizabilities for \(R_{\text{OH}} \geq 1.2\) Å. As a quantitative test of the linearly interpolated surfaces, 100 points were randomly selected from a uniform distribution of the 3-coordinates within the ranges of the scans. The \textit{ab initio} distributed multipoles and polarizabilities were computed and compared to the linearly interpolated value. The mean absolute errors (MAE) for each of the multipoles and polarizabilities for each atom site are listed in Table S2. Correlation plots of the multipoles and polarizabilities extracted from the interpolated surface against the calculated values from \textit{ab initio} are included in Figures S4-S5. This demonstrates the quantitative accuracy of the linearly interpolated surface in describing the distributed multipoles and polarizabilities for a very wide range of intramolecular geometries.

The interpolated multipole and polarizability surfaces allow for the description of the charge distribution and its polarizability at a wide range of intramolecular geometries. In the following, we will focus on benchmarking the induction model with a flexible multipolar/polarizability description on the following extensive datasets of 3- and 4-body energies, while comparing its performance to the version of the model based on static (equilibrium) distributed multipoles and polarizabilities.

**Database A: 12,260 3-body terms.** The original database comprising 12,347 trimers was developed by Babin et al. to fit the 3-body PIP in the MB-pol potential. These 3-body terms were evaluated at the CCSD(T) level of theory with the aug-cc-pVTZ basis set supplemented with midbond functions and were also corrected for basis set superposition error (BSSE). We have opted to remove trimer configurations from that database for which at least one of the following conditions is true:
1. the trimer contains a hydronium like molecule (three or more O-H distances ≤ 1.2 Å),

2. there exists an O⋯O distance that is ≤ 2.4 Å,

3. at least one of the intramolecular geometries is outside the range of the monomer scans used to obtain the surfaces, namely 60° ≤ θ_{HOH} ≤ 145°, 0.81 Å ≤ R_{ave} ≤ 1.21 Å and ∆R ≤ 0.40 Å.

We chose to remove trimers that are hydronium-like because, in these cases, it is not straightforward to divide the system into fragments (hydronium vs. water). Moreover the distributed multipoles and polarizabilities would be different if treated as hydronium rather than a neutral water molecule. The above criteria result in the removal of only 87 trimers from the original database of 12,347. Several representatives of these 87 trimers removed from the original dataset are shown in Figure S7.

**Database B: 43,844 3-body terms.** The full 3-body dataset compiled by Bowman and co-workers\(^ {16} \) comprises 45,332 three-body terms evaluated at the CCSD(T)-F12a/aug-cc-pVTZ level of theory and corrected for BSSE. This dataset was used to fit the recently developed q-AQUA potential\(^ {16} \) and also encompasses the datasets that were previously used to train the MB-pol\(^ {14} \) and WHBB\(^ {15} \) potentials, respectively. That is, database A is a subset of database B. This dataset contains very accurate energies for an extensive sampling of the configurational space with the maximum O⋯O distance being 9.5 Å. Using the same criteria described above for Database A, we removed 1,488 trimers from this dataset, leaving 43,844 (out of 45,332) trimer configurations to benchmark our proposed induction model against. The majority (1,401) of the trimers removed contained at least one molecule whose intramolecular geometry was outside the range of our scans. The remaining 87 trimers contained a molecule deemed "hydronium-like" based on their O⋯O or O⋯H distances. Several representatives of the trimers removed from the original dataset are shown in Figure S6.

**Database C: 3,603 4-body terms.** Database C comprises 3,603 water tetramer structures and their corresponding CCSD(T)/aug-cc-pVTZ 4-body terms. The full dataset of
3,692 4-body terms was used to fit a 4-body PIP for the q-AQUA potential. However, we removed 29 tetramers that have at least one molecule with an intramolecular geometry outside the range of our scan and 60 tetramers due to small O⋯O or O⋯H values following the same criteria used for the 3-body term; see Figure S6 for examples of the 89 water tetramer structures that were removed from the original dataset.

Figure 2: Contributions from 3-body dispersion (kcal/mol) to the total 3-body energies for the 43,844 trimers in Database B estimated with the Axilrod–Teller–Muto (ATM) potential.

We also consider contributions from 3-body dispersion, estimated from the Axilrod-Teller-Muto (ATM) potential. Figure 2 shows the distribution of the individual 3-body dispersion energies for each of the 43,844 trimers in Database B. This demonstrates that the contribution from 3-body dispersion is typically quite small in magnitude and usually repulsive. The maximum 3-body dispersion energy, albeit for a very small number of trimers in Database B, is found to be ∼0.2 kcal/mol (cf. Figure 2).

The top two panels of Figure 3 show the correlation plots comparing the accuracy of the “static dm/dp” (distributed multipoles and polarizabilities at the equilibrium geometry, left panel) vs. the “flexible dm/dp” (geometry-dependent distributed multipoles and polarizabilities, right top panel) versions of the induction model on Database B. The static model yields
Figure 3: Comparison of the 3-body energies (Database B, kcal/mol) of the classical model ($E_{\text{model}}$) including induction and dispersion using static multipoles and polarizabilities at the equilibrium geometry (top left) and the geometry-dependent multipoles and polarizabilities (top right) vs. the ab-initio results. The bottom panel shows the difference between these two versions of the model as a function of the total deformation energy (kcal/mol) of the trimers.

an RMSE of 0.152 kcal/mol while the flexible model yields 0.104 kcal/mol (see also Table 1), indicating a clear improvement with the geometry-dependent multipoles and polarizabilities. Comparing the static and flexible versions of the model more directly, the bottom panel of Figure 3 shows the difference in the calculated energy using the equilibrium multipoles and polarizabilities (at the distorted geometries) versus the flexible multipoles and polarizabilities from the interpolated surface as a function of the total trimer deformation energy, which was computed with the Partridge-Schwenke potential energy surface. As one might expect, the use of the geometry-dependent multipoles and polarizabilities is more critical for the trimers.
whose constituent monomers are more distorted from the equilibrium geometry. However, it is sometimes the case that both representations yield similar 3-body energies, due to a cancellation of errors. Nevertheless, the maximum difference between the computed 3-body terms is $\sim 5$ kcal/mol, which is quite substantial given the range of 3-body energies in the database (roughly -8 to 6 kcal/mol). That said, the inclusion of geometry-dependent multipoles and polarizabilities is paramount in improving the accuracy of the induction model for the 3-body term. Additionally, while the contribution from the 3-body dispersion is very small compared to the range of induction energies (-8 to 6 kcal/mol), incorporating it in the classical model improves the agreement with the \textit{ab initio} benchmarks. More specifically, the RMSE decreases from 0.089 to 0.084 kcal/mol on Database A and 0.106 to 0.104 on Database B (see Table 1). Given that the 3-body dispersion estimates from the ATM potential are very fast to evaluate since they take into account just the pairwise distances and angles between oxygen atoms, it is worthwhile to include it for quantitative accuracy.

Table 1: The root mean square error (RMSE) and mean absolute error (MAE, in parentheses) in kcal/mol of the classical model for the 3-body terms of databases A and B compared to potentials that were fit to PIPs.

<table>
<thead>
<tr>
<th>Database A</th>
<th>Database B</th>
<th>Number of fitted parameters</th>
<th>Size of training set</th>
</tr>
</thead>
<tbody>
<tr>
<td>12,260 trimers</td>
<td>43,844 trimers</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Induction only</td>
<td></td>
<td>0</td>
<td>n/a</td>
</tr>
<tr>
<td>Static dm/dp</td>
<td>0.140 (0.076)</td>
<td>0.152 (0.074)</td>
<td>0</td>
</tr>
<tr>
<td>Flexible dm/dp</td>
<td>0.089 (0.050)</td>
<td>0.106 (0.056)</td>
<td>0</td>
</tr>
<tr>
<td>Induction + Dispersion</td>
<td></td>
<td>0</td>
<td>n/a</td>
</tr>
<tr>
<td>Static dm/dp</td>
<td>0.142 (0.077)</td>
<td>0.154 (0.074)</td>
<td>0</td>
</tr>
<tr>
<td>Flexible dm/dp</td>
<td>0.084 (0.048)</td>
<td>0.104 (0.054)</td>
<td>0</td>
</tr>
<tr>
<td>Potentials fit to PIPs</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HBB2-pol</td>
<td>0.158$^a$</td>
<td>n/a</td>
<td>132</td>
</tr>
<tr>
<td>WHBB5</td>
<td>0.104$^a$</td>
<td>n/a</td>
<td>1,380</td>
</tr>
<tr>
<td>MB-pol</td>
<td>0.028$^b$</td>
<td>0.116 (0.035)</td>
<td>1,173</td>
</tr>
<tr>
<td>q-AQUA</td>
<td>n/a</td>
<td>0.026$^c$, 0.032$^d$</td>
<td>14,412</td>
</tr>
</tbody>
</table>

$^a$ Reported RMSE on entire 12,347 trimers in the MB-pol training set.$^{14}$

$^b$ Reported RMSE fitting error on training set.$^{14}$

$^c$ Reported RMSE fitting error on short-range portion of the q-AQUA training set.$^{16}$

$^d$ Reported RMSE fitting error on long-range portion of the q-AQUA training set.$^{16}$
It should be emphasized that the classical model based on geometry dependent multipoles and polarizabilities has no adjustable parameters. All values used in the potential (multipoles, polarizabilities, damping parameters, $C_9$ coefficients) are derived \textit{ab initio}. The root mean squared errors (RMSEs) of the model are 0.084 and 0.104 kcal/mol on Databases A and B (cf. Table 1), respectively, demonstrating its transferability over a wide range of trimer configurations. Table 1 organizes the RMSEs and MAEs for the classical model (this work) against other interaction potentials (q-AQUA, \textsuperscript{16} MB-pol, \textsuperscript{13,14} HBB2-pol, \textsuperscript{48} WHBB5\textsuperscript{15}) that were fitted using PIPs to \textit{ab initio} trimer calculations. On Database A, the classical model (induction+dispersion) yields a lower RMSE than HBB2-pol and WHBB5 (note that the RMSE of HBB2-pol and WHBB5 potentials was reported on the entire dataset of 12,347 trimers while the classical model on Dataset A, in which we removed 87 trimers from the original dataset for reasons stated above). The reported excellent accuracy for MB-pol on Database A (left top panel of Fig. 4) is in reality a training/fitting error as the entire database A was used for the fitting of that model using 1,173 parameters (1,163 linear and 10 non-linear). However, that potential performs worst for the larger Database B, as can be seen from Figure 4 (top right panel) and Table 1. For the larger number of trimers in Database B, the RMSD/MAE of MB-pol increases by $\sim4.1x$ / $\sim1.8x$ with respect to the fitting errors for Database A (0.116/0.035 kcal/mol for Database B vs. 0.028/0.019 kcal/mol for Database A). It is interesting that the TTM2.1-F potential (bottom panels of Figure 4), which uses geometry-dependent distributed monopoles and inducible point dipoles (5 parameters total) and is fitted to $\sim45$ \textit{ab-initio} dimer and no trimer points produces RMSD/MAE values of 0.210/0.110 kcal/mol, respectively. This is to be compared with the corresponding RMSD/MAE values of 0.104/0.054 kcal/mol for the classical induction + dispersion model based on geometry-dependent multipoles and polarizabilities suggesting that it outperforms both MB-pol and TTM2.1-F on Database B, which contains database A (MB-pol training set). Importantly, these results highlight the natural transferability of the classical model to a very wide range of trimer configurations.
Figure 4: Comparison of the 3-body energy (kcal/mol) calculated using the MB-pol (top panels) and TTM2.1-F (bottom panels) potentials against the corresponding benchmark values ($E_{\text{CCSD(T)}}$) for Database A (left) and Database B (right).

To better understand this difference, Figure 5 shows the cumulative RMSE as a function of the total deformation energy, $E_{\text{deformation}}$, of the trimer, evaluated using the Partidge-Schwenke potential energy surface. We observe that the MB-pol potential performs very accurately for small $E_{\text{deformation}}$ values. However, it produces errors for highly distorted trimer configurations that are slightly larger than the flexible classical induction + dispersion model. The accuracy of the classical model on subsets of Database B separated by the maximum of the total deformation energy of the trimers, shown in Figure S7, justifies that most outliers consist of quite deformed configurations. While it is true that the highly
Figure 5: The cumulative RMSE (kcal/mol) as a function of the total deformation energy of the trimer \( E_{\text{deformation}}, \) kcal/mol).

distorted, high energy structures are less relevant for many applications, it highlights the critical need for very extensive sampling to build a training set to train the PIPs, which is a nontrivial task. On the other hand, the classical model is naturally transferable to more distorted intramolecular geometries with a clear improvement when geometry-dependent distributed multipoles and polarizabilities \( \text{(classical model - flexible)} \) are used, performing extremely well even for high-energy (large deformation) structures. However, this can be a challenge for PIPs because, as with most supervised machine learning techniques, the accuracy of the functions in regions outside their training data is not guaranteed.

Having established the success of the classical model for the 3-body interactions in water clusters, we now turn our attention to the 4-body term. Figure 6 depicts a correlation plot demonstrating the performance of the classical model on Database C. Table 2 lists the RMSE and MAE values for the induction model \( \text{(this work)} \) against the q-AQUA potential, which is the only potential to date that includes an explicitly fitted 4-body term.\(^{16} \) Surprisingly, we find that the model with static multipoles and polarizabilities yields a slightly better RMSE
Table 2: The root mean square error (RMSE) and mean absolute error (MAE, in parentheses) in kcal/mol of the classical model for the 4-body terms of Database C and a comparison with the reported fitting error of the q-AQUA potential.

<table>
<thead>
<tr>
<th></th>
<th>Database C</th>
<th>Fitted parameters</th>
<th>Size of training set</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Induction only</strong></td>
<td>Static dm/dp</td>
<td>0.036 (0.020)</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Flexible dm/dp</td>
<td><strong>0.058 (0.026)</strong></td>
<td>n/a</td>
</tr>
<tr>
<td><strong>Potentials fit to PIPs</strong></td>
<td>q-AQUA</td>
<td>0.021(^a)</td>
<td>200</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3,692</td>
</tr>
</tbody>
</table>

\(^a\) Reported RMSE fitting error on entire 3,692 tetramers for the q-AQUA 4-body training set.\(^{16}\)

than the flexible model. The largest difference is in the strongly attractive 4-body terms, which we find is over-estimated by the model with flexible moments. Again, it is difficult to compare the accuracy of the model to q-AQUA since that model is fitted to Database C using \(~200\) parameters and we have removed 89/3,692 tetramers from the original database used to fit q-AQUA. However, we anticipate that the induction model will be quite transferable to a wide range of configurations, as observed for the 3-body term. Furthermore, the induction model is naturally capable of computing higher-order contributions (5-body and higher) to the induction energy through the same formalism.

In conclusion, this work demonstrates the success of a classical model in representing 3-body and 4-body interactions for a wide range of configurations and intramolecular geometries. In general, a flexible (geometry-dependent) representation of the distributed multipoles and polarizabilities significantly improves the accuracy of the model and naturally becomes more important as intramolecular geometries become more distorted from their equilibrium geometry. To the best of our knowledge, the role of a geometry-dependent multipole and polarizability surface of the 3- and 4-body term has not been investigated previously. In addition to the utility of geometry-dependent multipoles and polarizabilities, we find that multipoles up to the quadrupole moment and polarizabilities up to quadrupole-quadrupole are necessary to achieve an accuracy on a par with potentials fitted to tens of thousands of high level \textit{ab initio} calculations. In our opinion, the lack of geometry-dependent multi-
Figure 6: Comparison of the 4-body energy (kcal/mol) calculated using the classical induction model \( E_{\text{model}} \) against the benchmark CCSD(T) values \( E_{\text{CCSD(T)}} \) in Database C.

poles and polarizabilities and the absence of dipole-quadrupole and quadrupole-quadrupole polarizabilities have hindered the ability of existing polarizable potentials to describe the many-body terms with high accuracy. Furthermore, we find that incorporating estimated contributions from 3-body dispersion, which are typically very small and repulsive for the systems considered in this study, marginally improves the accuracy of the classical model with respect to \textit{ab initio}. Importantly, we show that the classical model produces 3- and 4-body energies with an accuracy that is on a par with the quantitative accuracy of many-body potentials fitted to tens of thousands of \textit{ab-initio} calculations using hundreds to tens of thousands of parameters. The quantitative accuracy of the classical model demonstrates that induction (and dispersion for the 3-body) is indeed sufficient to describe accurately the many-body interactions in water. The classical model introduced in this work is naturally transferable to a wide range of highly distorted configurations, an attribute that is
not guaranteed with fitted PIPs. In addition, this approach is transferable to new chemical systems in a natural and straightforward manner. More specifically, it would only be necessary to derive distributed multipoles and polarizabilities from the monomer wavefunction of additional solvent molecules which is quite fast for an individual monomer and needs to be carried out only once. If the molecule is suspected to undergo large changes in intramolecular geometry, it is recommended to develop a geometry-dependent description of the charge distribution and polarizability. For systems with only a few degrees of freedom (like water), this can be achieved by performing multiple monomer calculations at a range of intramolecular geometries. For systems with more degrees of freedom, it might be useful to follow a recently developed approach that uses a neural network to predict distributed multipoles for a wide range of systems and geometries.\textsuperscript{49} Furthermore, with this approach, one need not be concerned with building a very large and sufficiently diverse training set from costly \textit{ab-initio} calculations as required for fitting using PIPs. Nevertheless, we do acknowledge the usefulness of being able to benchmark our classical model against existing benchmark databases consisting of tens of thousands of high level \textit{ab initio} points. We hope that the use of an accurate classical model based on induction and dispersion to describe many-body effects in hydrogen bonded systems can simplify the process of developing highly accurate potentials for multi-component systems by eliminating the need to perform computationally expensive high level \textit{ab initio} calculations.

**Computational details**

We consider two contributions to the 3-body energy from induction and dispersion, respectively. As for the 4-body energy, the only contribution considered in this study is from the induction energy, which is is computed as previously described.\textsuperscript{44} The induced moment $t$ at site $a$ on molecule $A$ due to the multipole $u$ on site $b$ of molecule $B$ is evaluated self-
consistently using the following expression:

\[ \Delta Q_a^t = - \sum_{B \neq A} \alpha_{tt'}^{aa'} T_{tu}^{ab}(Q_b^t + \Delta Q_u^b) \]  

(1)

where \( \alpha_{tt'}^{aa'} \) is the polarizability of the multipole and \( T_{tu}^{ab} \) is the interaction tensor. Once the induced moments are converged, the induction energy of molecule \( A \) is calculated as:

\[ E_{\text{ind}}^A = \frac{1}{2} f_{TT} \sum_{B \neq A} \Delta Q_t^a T_{tu}^{ab} Q_u^b \]  

(2)

where \( f_{TT} \) is the Tang-Toennies damping function:

\[ f_{TT}(R) = 1 - \sum_{k=0}^{6} \frac{(\beta R)^k}{k!} e^{-\beta R} \]  

(3)

In our previous work, we chose \( \beta \) to be a single adjustable parameter. In this work, we set \( \beta \) as the Born-Mayer coefficient

\[ \beta = -0.31(r_{vdW}^1 + r_{vdW}^2) + 3.43 \]  

(4)

where \( r_{vdW} \) is the van der Waals radius of each of the pairs of atoms (from CCSD, O: 3.07 bohr, H: 2.63 bohr). Because of this, the decay of the damping function with respect to interatomic distance is slightly different for each pair of sites. The above choice yields \( \beta \) values of 1.66, 1.53, and 1.80 for the O····H, O····O, and H····H pairs of sites, respectively.

In our previous work, we found that a single damping parameter of \( \beta=1.60 \) worked quite well for a subset of Database A, which contained trimers with intramolecular geometries close to the equilibrium ones. That value of \( \beta \) used in the previous study is very close to the calculated value of 1.66 used for the O····O in the current study. A single damping parameter worked sufficiently well previously, because most of the trimers examined in that earlier study did not have two oxygen atoms or two hydrogen atoms very close to one another. However,
in applying this model to the full (more extensive) dataset B, we find it more appropriate to use the above 3 values of $\beta$ for the different pairs of sites instead.

The multipole expansion was truncated at the hexadecapole (H) with the respective multipoles distributed on the atom sites of H$_2$O. Dipole-dipole (D–D), dipole-quadrupole (D–Q), and quadrupole-quadrupole (Q–Q) distributed polarizabilities are used to represent the polarizabilities of the multipoles in the presence of an electric field. Tables comparing the accuracy of the induction model with different multipolar (M, MD, MDQ, MDQO, MDQOH) and polarizability (D–D, D–Q, Q–Q) expansions are included in the SI (Tables S3 and S4). While it has been established that a multipolar representation up to the hexadecapole was necessary to obtain converged molecule dipole moments in ice Ih,$^5$ we find essentially no change in the 3- and 4-body terms by extending the model beyond the quadrupole moments. However, inclusion of D–D, D–Q, and Q–Q polarizabilities (RMSE: 0.083 kcal/mol) significantly improves the accuracy of the model over using only D–D (RMSE: 0.122 kcal/mol) or D–D and D–Q polarizabilities (RMSE: 0.238 kcal/mol). Rather than using the distributed multipoles and polarizabilities computed at the equilibrium geometry (and placed at the equilibrium configuration), we explore geometry-dependent distributed multipole and polarizability representations of the water monomers. These distributed multipole and polarizability surfaces were computed at a 3-dimensional grid of points for H$_2$O. The $\theta_{\text{HOH}}$ coordinate was scanned every 5° from 60–145°, the $R_{\text{ave}}$ coordinate was scanned every 0.05 Å from 0.81–1.21 Å, and the $\Delta R$ coordinate was scanned every 0.04 Å from 0.00–0.40 Å (see Figure 1 for the definition of these coordinates). The total number of grid points for which the multipoles and polarizabilities were computed amounted to 1,782. The multipoles at each point were derived using CamCasp6.0$^5$ from the HF/aug-cc-pVTZ$^54$ wavefunction calculated using Psi4.$^5$ The 3-dimensional grid was linearly interpolated to yield distributed multipole and polarizability surfaces for each of the atom sites of H$_2$O.

The sensitivity of the distributed multipoles and polarizabilities with respect to level of theory (PBE0,$^5$ HF, MP2) and basis set (aug-cc-pVTZ and aug-cc-pVQZ)$^5$ is examined
more closely in the SI (Tables S6-S7). Increasing the basis set size from aug-cc-pVTZ to aug-cc-pVQZ leads to small differences in the resulting distributed multipoles and polarizabilities. The PBE0 and MP2 multipoles and polarizabilities are in excellent agreement with noticeable differences from those of HF. Table S7 compares the performance of the static induction model on Database A and C using the equilibrium geometry multipoles and polarizabilities derived from the PBE0/aVTZ, PBE0/aVQZ, HF/aVTZ, HF/aVQZ, MP2/aVTZ, and MP2/aVQZ wavefunctions. We have chosen to use the distributed multipoles and polarizabilities derived at the HF/aug-cc-pVTZ level of theory because we find better correspondence between the accuracy of the 3-body term and the 4-body terms using these values when compared to the PBE0/aug-cc-pVTZ ones (see Figures S8-S10).

The Axilrod-Teller-Muto (ATM) \textsuperscript{45,46} 3-body dispersion energy was evaluated using the computed $C_9$ coefficient at each trimer configuration as:

$$E_{\text{disp}}^{ijk} = C_9^{ijk} \frac{(1 + 3 \cos \phi_i \cos \phi_j \cos \phi_k)}{R_{ij}^3 R_{ik}^3 R_{jk}^3}$$

where $R_{ij}$, $R_{ik}$, and $R_{jk}$ are the pairwise distances between the indicated molecules and $\phi_i$, $\phi_j$, and $\phi_k$ are the angles between $R_{ij}$ and $R_{ik}$, $R_{ij}$ and $R_{jk}$, and $R_{ik}$ and $R_{jk}$ vectors, respectively.

The Ow-Ow-Ow $C_9$ triple-dipole-dispersion coefficient of 303.67 hartree bohr\textsuperscript{9} was calculated from the isotropic dipole-dipole polarizabilities at imaginary frequencies for the water molecule, using a single site at the oxygen atom. This is close to the value of 308.2 evaluated from pseudo-spectral dipole oscillator strength distributions by Margoliash \textit{et al.}\textsuperscript{57}

The MB-pol calculations were performed with the MBX distribution of the code.\textsuperscript{13,14,33,34,58} The TTM2.1-F\textsuperscript{9} calculations were performed using the open-source distribution from the Pacific Northwest National Laboratory (available upon request).
Acknowledgement

This work was supported from the Center for Scalable Predictive methods for Excitations and Correlated phenomena (SPEC), which is funded by the U.S. Department of Energy, Office of Science, Basic Energy Sciences, Chemical Sciences, Geosciences and Biosciences Division as part of the Computational Chemical Sciences (CCS) program under FWP 70942 at Pacific Northwest National Laboratory (PNNL), a multi-program national laboratory operated for DOE by Battelle. This research also used resources of the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

Supporting Information Available

Figures motivating the range of the 3D intramolecular scan, tables outlining the ranges of multipoles/polarizabilities, figures and tables demonstrating the success of the linear interpolation, plots showing representations of the multipoles and polarizability surfaces, tables comparing static multipoles and polarizabilities derived from different levels of theory and basis sets, comparisons of the accuracy of the models truncating the multipoles and polarizabilities at different orders, examples of trimers and tetramers removed from the databases

References


(3) Xantheas, S. Significance of higher-order many-body interaction energy terms in wa-


(11) Fanourgakis, G. S.; Xantheas, S. S. Development of transferable interaction potentials for water. V. Extension of the flexible, polarizable, Thole-type model potential (TTM3-
F, v. 3.0) to describe the vibrational spectra of water clusters and liquid water. *Journal of Chemical Physics* **2008**, *128*, 074506.


(53) Misquitta, A. J.; Stone, A. J. CamCASP: a program for studying intermolecular interactions and for the calculation of molecular properties in distributed form. 2019.


