Modeling molecular ensembles with gradient-domain machine learning force fields†

Alex M. Maldonado, Igor Poltavsky, Valentin Vassilev-Galindo, Alexandre Tkatchenko, and John A. Keith*

Gradient-domain machine learning (GDML) force fields have shown excellent accuracy, data efficiency, and applicability for molecules with hundreds of atoms, but the employed global descriptor limits transferability to ensembles of molecules. Many-body expansions (MBEs) should provide a rigorous procedure for size-transferable GDML by training models on fundamental $n$-body interactions. We developed many-body GDML (mbGDML) force fields for water, acetone, and methanol by training 1-, 2-, and 3-body models on only 1000 MP2/def2-TZVP calculations each. Our mbGDML force field includes intramolecular flexibility and intermolecular interactions, providing that the reference data properly describes these effects. We also compare this mbGDML approach to GAP, SchNet, and Ne_quiP potentials. Energy and force predictions of clusters containing up to 20 molecules are within 0.38 kcal/mol per monomer of reference supersystem calculations. This deviation partially arises from the restriction of the mbGDML model to 3-body interactions. Given these approximations, our automated mbGDML training schemes also resulted in fair agreement with reference radial distribution functions (RDFs) of bulk solvents. These results highlight mbGDML as valuable for modeling explicitly solvated systems with quantum-mechanical accuracy.

1 Introduction

Machine learning (ML) potentials and force fields have revolutionized atomistic modeling by facilitating larger and longer simulations crucial for modeling dynamic and kinetic properties. General-purpose ML potentials (e.g., ANI-2x, OrbNet Denali, AIQM1) model chemical (local) interactions and can be useful for subsets of chemical space. These approaches assist molecular screening but require enormous data sets of hundreds of thousands of structures. Alternatively, ML potentials can be tailored to specific systems to improve desired simulation reliability. This requires that models be retrained for each system, so training must involve minimal human involvement and computations to be practical.

Size transferability to hundreds of molecules is paramount for useful ML potentials. Most ML potentials rely on local descriptors or graph neural networks (GNNs) that partition total properties into atomic contributions. Local descriptors have been successful in numerous applications, but they inherently neglect or limit complicated non-local interactions by enforcing atomic radial cutoffs. Global descriptors (such as the Coulomb matrix and pairwise atomic distances) impose no such constraints and capture interactions at all scales. Still, they are usually restricted to the same number of atoms.

Gradient-domain ML (GDML) uses a global descriptor and has demonstrated remarkable success in many chemical applications with monomers or dimers. Moreover, GDML only needs energies and forces of approximately 1000 structures to accurately learn the potential energy surfaces of molecules and periodic materials. The global descriptor limits GDML to the same system it was trained on, whether a single molecule or a chemical reaction. Size-transferable GDML for molecular ensembles would provide rapidly trained force fields for high-quality molecular simulations involving solvents.

Many-body expansions (MBEs) should enable size-transferable GDML because systems with non-covalent clusters are naturally described in terms of $n$-body interactions. Data-driven, many-body potentials (e.g., MB-pol) have already been widely successful in modeling aqueous systems. This expansion is formally exact if all $N$-body interactions are accounted for with sufficient accuracy and precision. However, the expansion is typically truncated to the third order due to combinatorics. We expect training on fundamental $n$-body interactions found in clusters would extend GDML force fields to be useful for bulk liq-
uid simulations. Alternative approaches exist; for example, Gaussian Approximation Potential (GAP)\textsuperscript{[31,32]} was extended to liquid methane by decomposing energies into fundamental interactions (e.g., repulsion, dispersion, and electrostatic contributions) and different scales.\textsuperscript{[27]} This is another rigorous approach that requires considerable effort with large numbers of quantum chemical calculations.

MBEs share characteristics with local descriptors but provide several key advantages. First, \(n\)-body interactions are more efficiently treated on a molecular basis instead of an atomic basis. Second, errors associated with MBE truncation can be corrected using long-range physical models.\textsuperscript{[28,29]} Third, these \(n\)-body contributions can be observed in relatively small clusters. Local descriptors require data on large clusters to achieve similar levels of size transferability.\textsuperscript{[30]} This opens the door for many-body GDML (mbGDML) force fields trained on high levels of theory that scale poorly with system size, such as CCSD(T). In addition, mbGDML naturally incorporates intramolecular/monomer flexibility, which is extremely challenging for analytical potentials.

Thus, mbGDML should provide size-transferable force fields trained on highly accurate quantum chemical methods. To evaluate this, we developed an automated framework in Python to facilitate training and application of mbGDML force fields (available at github.com/keithgroup/mbGDML). GDML force fields for water (\(\text{H}_2\text{O}\)), acetonitrile (\(\text{MeCN}\)), and methanol (\(\text{MeOH}\)) were trained on 1000 structures for 1-, 2-, and 3-body interactions. GAP and SchNet\textsuperscript{[33]} were also evaluated in this many-body framework. The size transferability of mbGDML was further assessed against a highly promising graph neural network, Neural Equivariant Interatomic Potentials (Neq[OF])\textsuperscript{[34]} Reference structures from the literature were used to benchmark energy and forces predictions. The following sections demonstrate mbGDML accuracies of 0.38 kcal/mol per monomer and 0.06 kcal/(mol Å) per atom for structures containing up to 20 monomers (120 atoms). Error cancellation dramatically improves relative energy predictions of mbGDML to less than 3 kcal/mol and achieves fair to excellent agreement with solvent radial distribution functions (RDFs).

2 Methods

The MBE represents the total system energy, \(E\), composed of \(N\) noncovalently connected (i.e., non-intersecting) fragments as the sum of \(n\)-body interaction energies.\textsuperscript{[32]}

\[
E = \sum_i E_i^{(1)} + \sum_{i<j} \Delta E_{ij}^{(2)} + \sum_{i<j<k} \Delta E_{ijk}^{(3)} + \cdots .
\]

Here, \(N\) is the number of monomers; \(i, j, k\) are monomer indices; \(E_i^{(1)}\) is the energy of monomer \(i\); and \(\Delta E_{ij}^{(n)}\) represents the \(n\)-body interaction energy contribution of the fragment containing monomers \(i, j, \ldots \) with lower order (< \(n\)) contributions removed. For example, the 2-body contribution of the fragment containing monomers \(i\) and \(j\) is

\[
\Delta E_{ij}^{(2)} = E_{ij}^{(2)} - E_i^{(1)} - E_j^{(1)},
\]

and the 3-body contribution with monomers \(i, j, k\) are

\[
\Delta E_{ijk}^{(3)} = E_{ijk}^{(3)} - \Delta E_{ij}^{(2)} - \Delta E_{jk}^{(2)} - E_i^{(1)} - E_j^{(1)} - E_k^{(1)}.
\]

Equation 1 is exact when all \(n\)-body contributions up to \(N\) are accounted for with exact accuracy and precision. This equation also holds for properties expressed as a derivative of energy (i.e., gradients).

Equation 1 represents the MBE framework where individual GDML force fields are trained on intramolecular (i.e., 1-body) and intermolecular (i.e., 2- and 3-body) energies and forces. Energies and forces were calculated with ORCA v4.2.0\textsuperscript{[35,36]} using second-order Möller–Plesset perturbation (MP2) theory\textsuperscript{[35]} the def2-TZVP basis set\textsuperscript{[37]} and the frozen core approximation. This level of theory was chosen for its efficiency and accuracy for noncovalent interactions, but future applications of mbGDML are recommended to use the highest levels of quantum chemical theory available for training data. Additional calculation details and discussion can be found in the Electronic Supplementary Information (ESI).

3 Results and discussion

3.1 Small isomers

We evaluated mbGDML, mbGAP, and mbSchNet on tetratomers (4mers), pentamers (5mers), and hexamers (6mers) from the literature.\textsuperscript{[37–40]} These test structures have minimal higher-order (> 3-body) contributions that increase with the number of monomers. Furthermore, many-body ML (mbML) potentials considered here implement a distance-based cutoff for 2- and 3-body contributions (see the ESI for more details). Small clusters allow us to determine whether errors are from the underlying MBE framework or ML predictions.

ML potentials discussed here are trained on small data sets of only 1000 structures to showcase GDML data efficiency. Training sets were determined through an iterative training procedure to reduce the maximum model error.\textsuperscript{[40]} GAP and SchNet models were trained on the same training sets as GDML for a fair comparison. In theory, training sets could be tailored for GAP and SchNet to reduce errors; however, a cursory attempt did not substantially improve results. We reiterate that GAP and SchNet normally require substantially large training sets. In other words, GAP and SchNet potentials presented here are technically underfitted compared to standard practices. More information can be found in the ESI.

Fig. 1 shows relative isomer energies with respect to the lowest energy structure for MBE (light color) and mbGDML (dark color) methods. The ESI provides comparable figures for mbGAP and mbSchNet. Figures showing absolute energy predictions for these structures are also shown in the ESI and help determine where error cancellation comes into play. First, we discuss the inherent errors in MBE data versus supersystem MP2 data (gray). These MBE predictions generally capture the relative energy trends of water, acetonitrile, and methanol isomers. Water predictions showed increasing errors with system size, indicating the importance of higher-order contributions (as expected). Acetonitrile 5mers and 6mers (Fig. 1D-F) show small energy differences that are not monotonically increasing. This is likely due
Fig. 1: Relative energies of isomers containing four, five, and six monomers of (A-C) water, (D-F) acetonitrile, and (G-I) methanol. Gray dashed lines are the reference MP2/def2-TZVP calculations. Light-colored lines with squares are MBE predictions calculated with MP2/def2-TZVP with no distance-based cutoffs for 2- and 3-body predictions. Dark-colored lines with circles are mbGDML predictions. Different y-axis scales are used for each subplot to enhance visualization.

to challenging electrostatics and polarization from dipole-dipole interactions. Methanol isomer MBE predictions showed this same trend as water, but higher-energy structures now exhibit lower MBE errors. This suggests that higher-order contributions are crucial for stabilizing low-energy methanol structures. Basis set superposition error (BSSE) is known to be significant for MBEs, however, water isomer calculations with a larger basis set (aug-cc-pVTZ) only reduced the energy MAE by at most 0.15 kcal/mol.

We now discuss mbGDML data, which approximates the MBE potential energy surface. In general, mbGDML reasonably mimics MBE data, including innate errors made by the MBE framework, as seen in the acetonitrile 5mer and 6mer data. Note that fortuitous error cancellations of 2- and 3-body mbGDML predictions sometimes give the appearance of higher accuracy than MBE. mbGAP and mbSchNet potentials occasionally are better or worse than mbGDML; however, as previously mentioned, these methods generally require larger training sets and are likely to underperform. For example, Table 1 shows the MBE, mbGDML, mbGAP, and mbSchNet energy and force mean absolute errors (MAEs) with respect to supersystem MP2/def2-TZVP calculations for all 4-6mer structures considered here. All mbML models perform similarly for water and acetonitrile, but the methanol isomer errors for mbGAP and mbSchNet are nearly double that for mbGDML.

Previous studies also investigated mbML models for water. Nguyen et al. use permutationally invariant polynomials (PIPs), Behler-Parrinello neural networks (BPNNs), and GAP models for predicting water 1, 2-, and 3-body interactions. Their 2-body training set included 34,431 structures containing the global dimer minimum, saddle points, artificially compressed geometries, and geometries from path-integral molecular dynamics (PIMD) sim-
Table 1: Energy (kcal/mol) and force [kcal/(mol Å)] MAEs of 4-6mer predictions. MAEs on an energy/monomer and force/atom basis are shaded. Best ML potential values are bolded. Energy and forces are abbreviated as E and F, respectively.

<table>
<thead>
<tr>
<th>Method</th>
<th>H$_2$O</th>
<th>MeCN</th>
<th>MeOH</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>E</td>
<td>F</td>
<td>E</td>
</tr>
<tr>
<td>MBE</td>
<td>0.925</td>
<td>0.426</td>
<td>0.110</td>
</tr>
<tr>
<td>mbGDML</td>
<td>1.340</td>
<td>0.737</td>
<td>0.296</td>
</tr>
<tr>
<td>mbGAP</td>
<td>0.248</td>
<td>0.045</td>
<td>0.057</td>
</tr>
<tr>
<td>mbSchNet</td>
<td>2.185</td>
<td>0.690</td>
<td>0.368</td>
</tr>
<tr>
<td></td>
<td>0.237</td>
<td>0.043</td>
<td>0.070</td>
</tr>
</tbody>
</table>

Simulations using HBB2-pol. Their 3-body training set contained 10,001 structures from HBB2-pol MD and PIMD of small water clusters, liquid water, and ice phases. Table 2 shows 2- and 3-body interaction energy MAEs with their models and those calculated here. The PIP, BPNN, and GAP water models trained on large data sets achieved 2- and 3-body interaction energy MAEs on the order of 0.033–0.145 and 0.007–0.123 kcal/mol, respectively. Alternatively, our GDML force fields trained on only 1000 structures achieved MAEs of 0.030–0.047 and 0.041–0.093 kcal/mol for 2- and 3-body interaction energies. This shows that GDML models using small training sets can perform similarly to well-trained potentials requiring large training sets. We highlight that the GAP results from Ref. 42 demonstrate substantial accuracy improvements possible with more extensive training sets.

We reiterate that ML potential accuracy is intricately linked to reference data sets, but we specifically opted to show the promise of GDML with small training sets. In almost all cases, the water 2-body models prepared here outperformed those from Ref. 42 that used larger training sets. Presumably, our smaller data sets may contain structures that enhance the perceived accuracy of these models. The 3-body data exhibits the opposite trend, which can be attributed to data set quality. In general, additional sampling of configurational spaces would improve our mbML models; however, the objective here is to evaluate ML models that can be trained with minimal computational cost.

3.2 16mers

Predictions of medium-sized structures provide a straightforward test of size transferability. There are additional, albeit typically small, higher-order contributions in larger structures. Also, the n-body cutoffs are now in effect to reduce the number of computations. Table 3 shows energies and forces of hexadecamers (16mers) from the literature [44] computed with RI-MP2/def2-TZVP and compared against mbGDML, mbGAP, and mbSchNet results. The truncated MBE contributes a few kcal/mol errors depending on the system. For example, the MBE prediction of (H$_2$O)$_{16}$ results in a 3.3 kcal/mol error whereas (MeCN)$_{16}$ has only a 0.2 kcal/mol error. Missing higher-order contributions or basis set errors are the most likely causes. All mbML models performed similarly well with (H$_2$O)$_{16}$. Most errors originated from 3-body predictions, with error cancellation improving model performance.

3.2.1 Analysis of (MeCN)$_{16}$

We find that both mbSchNet and mbGAP models trained from smaller data sets have abnormally high errors for (MeCN)$_{16}$ and (MeOH)$_{16}$, respectively. In both cases, the 3-body model has substantial error accumulation. Cutoffs are not the issue because only −0.006 of the 16.1 kcal/mol error in mbSchNet’s prediction of (MeCN)$_{16}$ is from cutoffs implemented in the 2- and 3-body models. Prediction errors contribute the most; a massive −15.2 kcal/mol error comes from the 3-body SchNet model.

Assessing inadequacies of training data is more complicated. If 3-body structures from (MeCN)$_{16}$ are substantially different from the data sets, then the model may break down. To investigate this, we used dimensionality reduction to visualize high-dimensional similarity in 2D space. Similar structures in feature space should be clustered together and vice versa. Fig. 2 shows the GDML feature space, a 2D embedding of trained and 3-body structures from (MeCN)$_{16}$ using Uniform Manifold Approximation and Projection (UMAP). There is a si-
significant overlap between the GDML training set feature space and
the structures from (MeCN)$_{16}$. High overlap suggests that GDML
should have low prediction errors, which is the case. SchNet, on
the other hand, has several test structures isolated from training
data, resulting in higher errors (shown in the ESI).

Not all structures with a high error are dissimilar in feature
space. Models should have learned similar structures and thus
should have performed well. A simple, ad hoc geometry descrip-
tor (discussed in the ESI) applied in Fig. 2B shows that all high-
error structures are dissimilar to anything in the data set. SchNet
has some difficulty with these structures, which results in a sub-
stantial 16.1 kcal/mol error. Many-body GAP’s 17.6 kcal/mol er-
ror in (MeOH)$_{16}$ is likely for the same reason. However, GAP
uses a local descriptor, making feature space more complicated to
analyze. Models under these circumstances were excluded from
further analyses (namely, mbSchNet for acetonitrile and mbGAP
for methanol).

3.3 20mers

Truncated higher-order contributions could be pertinent for ac-
curate absolute energies, as seen in the previous 16mer data. In
practice, relative energy accuracy is of primary importance. Yao
et al.\cite{Yao2019} trained mbML methanol potentials and analyzed their
performances on five (MeOH)$_{20}$ isomers. They used a Genera-
tive Adversarial Network (GAN) trained on RI-MP2/cc-pVTZ en-
ergies with the Coulomb descriptor. Training included 80% of
their data sets containing 844 800 monomers, 74 240 dimers, and
36 864 trimers.

Relative isomer energies of their methods are reported in Table
4. Their mbGAN potential accurately captures the isomer ranking
trend within 5 kcal/mol. We recomputed the energies using the
same model chemistry we used earlier (RI-MP2/def2-TZVP) and
show the results in Table 4. mbGDML and mbSchNet were within

Table 4: Relative energies (kcal/mol) of four (MeOH)$_{20}$ with
respect to the lowest energy structure (Isomer 0). Errors are
provided within the parentheses. Best ML potential values are
bolded.

<table>
<thead>
<tr>
<th>Method</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>RI-MP2$^a$</td>
<td>31.0</td>
<td>40.9</td>
<td>50.8</td>
<td>52.0</td>
</tr>
<tr>
<td>MBE$^a$</td>
<td>27.5</td>
<td>39.6</td>
<td>48.0</td>
<td>48.6</td>
</tr>
<tr>
<td>mbGAN$^b$</td>
<td>26.1</td>
<td>39.8</td>
<td>49.2</td>
<td>49.8</td>
</tr>
<tr>
<td>RI-MP2$^b$</td>
<td>28.5</td>
<td>38.9</td>
<td>49.9</td>
<td>49.5</td>
</tr>
<tr>
<td>mbGDML</td>
<td>29.1</td>
<td>38.9</td>
<td>51.2</td>
<td>48.4</td>
</tr>
<tr>
<td>mbSchNet</td>
<td>33.6</td>
<td>41.4</td>
<td>52.7</td>
<td>53.0</td>
</tr>
</tbody>
</table>

$^a$ RI-MP2/cc-pVTZ, MBE, and mbGAN (trained on 675 840 monomers, 59 392 dimers, and 29 491 trimers) data are from Ref. [48]

$^b$ RI-MP2/def2-TZVP data calculated here.

2.2 and 5.1 kcal/mol for all isomers considered in Ref. [48] and no-

3.4 Size transferability of local descriptors

As previously mentioned, many ML potentials use local descrip-
tors for size transferability. Recent developments of ML poten-
tials with local descriptors have involved GNNs.\cite{Battaglia2018, Neupaver2019} NequIP uses equivariant, continuous convolutions where edges connect every
atom within a cutoff radius.\cite{Neupaver2019} NequIP has achieved remarkable
accuracy and data efficiency on the MD17 data set, bulk water, formate dehydrogenation, and amorphous lithium phosphate.

Such models are inherently size transferable, but the accuracy is not typically studied when trained exclusively on small clusters. In theory, these potentials can train on the same data sets, but instead of n-body interactions, they would use total energies and forces. This would eliminate the need for an MBE framework. We trained NequIP on total energies and forces of 1000 trimers for water, acetonitrile, and methanol to assess this approach. Another 2000 trimers were used as a validation set.

We emphasize that this is an edge case of GNN potentials. If energies and forces of larger structures were readily available, these data would improve size transferability if they were included in the validation set. However, mbGDML models were never exposed to these larger clusters during training since the objective was to reproduce the 1-, 2-, and 3-body PES. Thus, training a NequIP on only trimers represents a straightforward comparison to mbGDML. These models were then tested against the identical isomers discussed above, with the results shown in Table 5.

While NequIP can be expectedly extrapolate to larger clusters, the errors are substantially higher than mbGDML. For example, the NequIP error on \((\text{H}_2\text{O})_{16}\) was more than 33 kcal/mol higher than mbGDML.

These static cluster results demonstrate that mbGDML is a reasonably accurate, size-transferable force field. The desired level of theory for reference data determines the MBE framework’s viability. Training on large clusters or bulk systems is likely more efficient if a lower scaling method is satisfactory. However, mbGDML becomes particularly useful when applications require force fields based on higher scaling methods. Recovering truncated higher-order contributions would also expect improve errors, but explicit 4-body interactions are rather challenging due to high demands on precision and combinatorics.\(^{49-51}\)

Electrostatic\(^{52}\) and more general quantum embedding approximations may be a practical route to avoid calculating higher-order contributions, but they are not considered here.

### 3.5 Molecular dynamics simulations

While accurate predictions of static clusters are essential, compelling applications for mbGDML would involve molecular simulations. Again, low energy and force errors are not conclusive of accurate molecular simulations,\(^{53}\) but experimentally measurable dynamic properties are an alternative and rigorous way to evaluate ML potentials. For example, the radial distribution function (RDF) is a vital bulk property that quantitatively defines liquid structure. Locations and intensities of peaks and valleys represent the solvation shells and liquid ordering. Accurately reproducing reference RDF curves is crucial for a practical size-transferable potential.

Periodic NVT simulations driven by mbGDML force fields were performed at 298.15 K for 10–30 ps in the atomic simulation environment (ASE)\(^{54}\) The minimum image convention was used with cubic boxes with lengths of 16 Å (137 molecules), 18 Å (67 molecules), and 16 Å (61 molecules) for water, acetonitrile, and methanol, respectively. Production trajectories were used to compute all possible RDFs. Some RDF curves are shown in Fig. 3. Water,\(^{55}\) acetonitrile,\(^{56}\) and methanol\(^{57-59}\) reference RDF curves are from neutron diffraction experiments. Results from classical MD simulations\(^{60-62}\) are also shown in Fig. 3. Note that classical references often include some fitting to empirical data,\(^{60-62}\) whereas mbGDML and others\(^{63}\) run calculations with no explicit empirical fitting. Individual figures of all RDF curves are shown in the ESI.

Dispersion and polarization are not always accurately treated with MP2 theory,\(^{66,67}\) and likewise, the underlying model chemistry (MP2/def2-TZVP) to train mbGDML force fields may not accurately reproduce experimental liquid properties. For example, MP2 yields excellent results for liquid water simulations when appropriate density corrections are employed,\(^{68-70}\) but these were not used in this study. To our knowledge, a thorough investigation has not been performed for liquid acetonitrile and methanol with MP2, so the agreement with experiments is more uncertain. Note that molecular simulations using Kohn-Sham density functional theory (DFT) results in comparable differences in RDFs shown in Fig. 3 depending on the exchange-correlation functional and dispersion treatment used.\(^{71-74}\)

The simulated RDFs with mbGDML fairly agree with the reference curves. In particular, the water \(g_{\text{OO}}(r)\) in Fig. 3 agrees remarkably well with experimental data. This is consistent with fragment-based \textit{ab initio} MD (AIMD) simulations.\(^{75,76}\) However, these AIMD simulations include higher-order contributions through electrostatic embedding. There are deviations in the \(g_{\text{OH}}(r)\) and \(g_{\text{HH}}(r)\) curves which may be partially due to the neglect of quantum nuclear effects.\(^{27}\)

In all cases, acetonitrile peaks from mbGDML are less intense than the reference curves. This indicates that the predicted liquid structure with mbGDML is less ordered than the deuterated neutron diffraction data.\(^{53}\) Notably, \(g_{\text{NN}}(r)\) has a wide peak with two distinct peaks that deviate from the experimental reference. However, classical RDFs from the literature can vary substantially. Some classical potentials\(^{61,77}\) result in a similar \(g_{\text{NN}}(r)\) shape while others\(^{50,62,63}\) better resemble the experimental reference.

### Table 5: Energy MAEs (kcal/mol) of various sized isomers for mbGDML (many-body global descriptor) and NequIP (local descriptor). MAEs on a per monomer basis are shaded. Best ML potential values are **bolded**.

<table>
<thead>
<tr>
<th>Solvent</th>
<th>Method</th>
<th>4mers</th>
<th>5mers</th>
<th>6mers</th>
<th>16mer</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>H(_2)O</strong></td>
<td>mbGDML</td>
<td>0.793</td>
<td>1.088</td>
<td>1.765</td>
<td><strong>4.013</strong></td>
</tr>
<tr>
<td></td>
<td>NequIP</td>
<td>0.198</td>
<td>0.218</td>
<td>0.294</td>
<td>0.251</td>
</tr>
<tr>
<td><strong>MeCN</strong></td>
<td>mbGDML</td>
<td>0.260</td>
<td>0.317</td>
<td>0.288</td>
<td>0.282</td>
</tr>
<tr>
<td></td>
<td>NequIP</td>
<td>0.065</td>
<td>0.063</td>
<td>0.048</td>
<td>0.018</td>
</tr>
<tr>
<td><strong>MeOH</strong></td>
<td>mbGDML</td>
<td><strong>1.260</strong></td>
<td>1.805</td>
<td><strong>2.089</strong></td>
<td><strong>5.561</strong></td>
</tr>
<tr>
<td></td>
<td>NequIP</td>
<td>0.342</td>
<td>0.374</td>
<td><strong>0.382</strong></td>
<td><strong>0.348</strong></td>
</tr>
</tbody>
</table>

Dispersion and polarization are not always accurately treated with MP2 theory,\(^{66,67}\) and likewise, the underlying model chemistry (MP2/def2-TZVP) to train mbGDML force fields may not accurately reproduce experimental liquid properties. For example, MP2 yields excellent results for liquid water simulations when appropriate density corrections are employed,\(^{68-70}\) but these were not used in this study. To our knowledge, a thorough investigation has not been performed for liquid acetonitrile and methanol with MP2, so the agreement with experiments is more uncertain. Note that molecular simulations using Kohn-Sham density functional theory (DFT) results in comparable differences in RDFs shown in Fig. 3 depending on the exchange-correlation functional and dispersion treatment used.\(^{71-74}\)

The simulated RDFs with mbGDML fairly agree with the reference curves. In particular, the water \(g_{\text{OO}}(r)\) in Fig. 3 agrees remarkably well with experimental data. This is consistent with fragment-based \textit{ab initio} MD (AIMD) simulations.\(^{75,76}\) However, these AIMD simulations include higher-order contributions through electrostatic embedding. There are deviations in the \(g_{\text{OH}}(r)\) and \(g_{\text{HH}}(r)\) curves which may be partially due to the neglect of quantum nuclear effects.\(^{27}\)

In all cases, acetonitrile peaks from mbGDML are less intense than the reference curves. This indicates that the predicted liquid structure with mbGDML is less ordered than the deuterated neutron diffraction data.\(^{53}\) Notably, \(g_{\text{NN}}(r)\) has a wide peak with two distinct peaks that deviate from the experimental reference. However, classical RDFs from the literature can vary substantially. Some classical potentials\(^{61,77}\) result in a similar \(g_{\text{NN}}(r)\) shape while others\(^{50,62,63}\) better resemble the experimental reference.
Methanol simulations appear more challenging for mbGDML. RDF peaks with respect to experimental data are less intense (same as acetonitrile). The shape of $g_{OO}(r)$, Fig. 3A, agrees well with the digitized experimental data. Classical simulations using GROMOS96 and OPLS/AA potentials have significantly more ordered liquid structure. For instance, their $g_{OH}(r)$ peaks are around 1.24 higher in intensity than mbGDML. While the $g_{OO}(r)$ is in good agreement with the experiment beyond 5 Å, the $g_{OH}(r)$ and $g_{HH}(r)$ curves are missing long-range liquid structure. Even though GDML employs a global descriptor, mbGDML is not capturing these long-range interactions. We suspect this is caused by truncations and cutoffs used in the MBE framework.

To summarize, even though the mbGDML models we use here have only included up to 3-body contributions, they generally predict the liquid structure of water, acetonitrile, and methanol well. Moreover, these force fields automatically include fully flexible molecules and perform no fitting to experimental properties. Further improvements could be made with more expansive training sets and higher-order contributions. For systems without classical parameters, mbGDML can be rapidly trained on relatively small amounts of data and provide valuable dynamical insights for explicitly solvated systems.

4 Conclusions

We have introduced a GDML-driven, many-body expansion framework that enables state-of-the-art size transferability toward molecular simulations of solvents. mbGDML force fields trained on only 1000 1-, 2-, and 3-body interactions accurately modeled small and medium isomers of water, acetonitrile, and methanol. Size-extrapolated predictions on static clusters of up to 20 monomers had energy errors of less than 0.38 kcal/mol per monomer for all three solvents. These results outperform NequIP trained on the same trimer data set by up to 34 kcal/mol for 16mers. Dynamic simulations of bulk systems using our mbGDML force fields provide semi-quantitative insights while avoiding expensive training data on bulk systems and fitting to experimental data. We stress that the computational cost and human involvement in training these mbGDML force fields are relatively small compared to that required for other modern ML potentials. Still, thorough efforts in configurational sampling and active learning would further improve model accuracies. We foresee promising applications of mbGDML for complicated, explicitly solvated systems where high levels of theory are desired.

Data availability

Code for preparing and using many-body ML potentials can be found at [github.com/keithgroup/mbGDML](https://github.com/keithgroup/mbGDML) (DOI: 10.5281/zenodo.6270373). All other code and data supporting this paper are available at [github.com/keithgroup/mbgdml-h2o-meoh-mecn](https://github.com/keithgroup/mbgdml-h2o-meoh-mecn) (DOI: 00.0000/00000000) and further detailed in the ESI.

Author contributions

Alex M. Maldonado: Conceptualization, Data Curation, Investigation, Methodology, Software, Validation, Visualization, Writing; Igor Poltavsky: Conceptualization, Methodology, Funding Acquisition, Supervision, Writing; Valentin Vassilev-Galindo: Conceptualization, Methodology, Supervision, Writing; Alexandre Tkatchenko: Conceptualization, Funding Acquisition, Supervision, Writing; John A. Keith: Conceptualization, Funding Acquisition, Supervision, Writing.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

A.M.M. and J.A.K. acknowledge support from the R. K. Mellon Foundation and the U.S. National Science Foundation (CBET-1653392, CBET-1705592, and CHE-1856460). This research was supported in part by the University of Pittsburgh Center for Research Computing through the resources provided. Specifically, this work used the H2P cluster, which is supported by NSF award number OAC-2117681. V.V.-G., I.P., and A.T. acknowledge support from the Luxembourg National Research Fund (FNR). This research was funded in whole, or partly, by the FNR grant reference C19/MS/13718694/QML-FLEX. For open access, the au-
Notes and references


