Electronic Supplementary Material (ESI) for Digital Discovery.

Supplementary information: Modeling molecular ensembles with gradient-domain machine learning force fields

Alex M. Maldonado,[a] Igor Poltavsky,[b] Valentin Vassilev-Galindo,[bc] Alexandre Tkatchenko,*[b] and John A. Keith*[a]

[a] *Department of Chemical and Petroleum Engineering, University of Pittsburgh, Pittsburgh, Pennsylvania 15260, United States of America; E-mail: jakeith@pitt.edu*
[b] *Department of Physics and Materials Science, University of Luxembourg, L-1511 Luxembourg City, Luxembourg; E-mail: alexandre.tkatchenko@uni.lu*
[c] *IMDEA Materials Institute, C/Eric Kandel 2, 28906 Getafe, Madrid, Spain*

Contents

S1 Reproducibility **S2**
S1.1 Model chemistry . . . S2
S1.2 Sources of error . . . S2

S2 Configurational sampling **S3**

S3 Distance-based screening **S3**
S3.1 Optimizing the many-body cutoff . . . S4

S4 Data set curation **S6**

S5 Training **S6**
S5.1 GDML . . . S6
S5.1.1 Training statistics . . . S8
S5.2 SchNet and GAP . . . S9
S5.3 NequIP . . . S11

S6 Timings **S12**

S7 Isomer rankings **S13**
S7.1 Relative errors of mbGAP and mbSchNet . . . S13
S7.2 Absolute errors of mbGDML, mbGAP, and mbSchNet . . . S16
S7.3 Performance statistics . . . S19
S7.4 Effect of larger basis set . . . S19

S8 $(MeCN)_{16}$ prediction analysis **S20**
S8.1 GDML and SchNet feature space . . . S20
S8.2 Geometry descriptor . . . S20
S8.3 Trimer embeddings . . . S21
S8.4 Dimer embeddings . . . S22

S9 Molecular dynamics simulations S23
S9.1 Radial distribution functions S23

S1 Reproducibility

We provide all data, code, figures, and explanations in open-access repositories. Some data are archived separately to keep main repositories as small as possible.

- github.com/keithgroup/mbGDML - Foundational code for preparing, training, and analyzing many-body machine learning (mbML) models.
- github.com/aalexmmaldonado/reptar - Supporting code for data parsing and management.
- github.com/keithgroup/mbgdml-h2o-meoh-mecn - Main repository with scripts, data, and figures presented in this manuscript.
- github.com/keithgroup/mbgdml-h2o-meoh-mecn-engrads - All energy and gradient calculation output files.
- zenodo.org/record/7464581 - All trained GDML, SchNet, GAP, and NequIP models. Training logs are kept in the above main repository.
- zenodo.org/record/7112198 - ASE trajectories and submissions scripts for NVT MD simulations. A working copy of the data (exdir files) is kept in the main repository.

S1.1 Model chemistry

All *n*-body energies and forces were calculated in ORCA (v4.2.0)[1,2] with second-order Møller–Plesset perturbation (MP2) theory,[3] def2-TZVP basis set,[4] and the frozen core approximation. ORCA's tight self-consistent field (SCF) convergence criteria were used ($< 10^{-8}$ and $< 10^{-5}$ Eh change in total electronic energy and one-electron energy between two cycles, respectively). An integral screening threshold of 2.5×10^{-11} was also used.

MBE predictions must be benchmarked against supersystem calculations employing the same level of theory used to calculate *n*-body interactions. MP2/def2-TZVP with the same convergence criteria in ORCA was used to calculate energies and forces of various isomers of water,[5,6] acetonitrile,[7,8] and methanol.[9,10] The resolution of identity (RI) approximation[11] was used for 16 and 20mers from literature to reduce memory requirements. The RI approximation had minimal impact on the $(H_2O)_{16}$ results, which resulted in a 0.4 kcal mol^{-1} error and 0.01 kcal (mol Å)$^{-1}$ force RMSE.

S1.2 Sources of error

MBEs are known to suffer from basis set superposition error (BSSE), where basis functions of one molecule are used by others to lower the energy.[12–14] Many recommend the Boys-Bernardi "function counterpoise" (CP) correction where lower order contributions (e.g., monomers in a dimer) are calculated in the full cluster basis set. BSSE can also be reduced by using sufficiently large basis sets, extrapolating to the complete basis set (CBS) limit, or using explicitly correlated methods (e.g., F12).[15,16] Practical applications can easily implement these corrections; however, the focus here is on reproducing these data with ML potentials, not having the most accurate MBEs.

Data precision and SCF convergence criteria can also induce uncertainty in MBE predictions.[17–19] Computational chemistry output files must provide enough significant figures to calculate many-body interactions correctly. ORCA output files print energies down to 1×10^{-12} Eh (6.3×10^{-10} kcal mol^{-1}) and gradients to 1×10^{-8} Eh/Bohr [1.2×10^{-5} kcal (mol Å)$^{-1}$]. Richard et al.[17] presented a simple propagation-of-errors analysis where MBE uncertainty, dE, can be estimated by approximating the uncertainty in each subsystem calculation, δE, based on SCF energy convergence criteria. They state δE can be approximated by assuming that, in an SCF calculation using a 10^{-a} convergence criteria, the $a+1$ decimal digit is a random number. Using this approach with $a = 8$ (i.e., ORCA tight SCF convergence) results in an energy uncertainty of less than 0.01 kcal mol^{-1} for a 50mer—which is sufficient for our purposes here.

S2 Configurational sampling

A crucial aspect of training accurate ML potentials is curating data sets for 1-, 2-, and 3-body energies and forces. Accurate models require expansive sampling, typically involving global optimizations and lengthy molecular dynamics (MD) simulations driven by quantum chemical or classical methods. Data sets can quickly explode to thousands of structures, especially when n-body energies and forces are desired.

Initial spherical structures (radius of 10 Å) were generated using packmol[20] (v20.2.2). The number of monomers was determined by using the mass density of the solvent at 300 K. These structures contained 140 water, 48 acetonitrile, and 62 methanol molecules. A spherical, confining logfermi potential ($\beta = 6$; $T = 300$) was used to prevent dissociation and maintain the selected mass density during the simulation. A 4 ps simulation was used to equilibrate the system before the 1 ps production simulation (all using a 1 fs time step).

S3 Distance-based screening

Many methods have been proposed to reduce the amount of n-body contributions considered due to challenging combinatorics for MBEs on larger systems.[21,22] Distance-based screening is a straightforward technique that assumes the size of n-body contributions is inversely proportional to the distance between the monomers. If monomer distances are higher than some cutoff, it is ignored during MBE predictions. Here, we employed a distance-based size descriptor, L, of the sum of each monomer's center of mass, $\mathbf{CM}_i$, to the center of mass of the whole structure, $\mathbf{CM}$:

$$L = \sum_{i}^{N} \left\| \mathbf{CM}_i - \mathbf{CM} \right\| . \tag{1}$$

Note that $\| \ldots \|$ is the L2 norm. This is visually shown in Fig. S1.

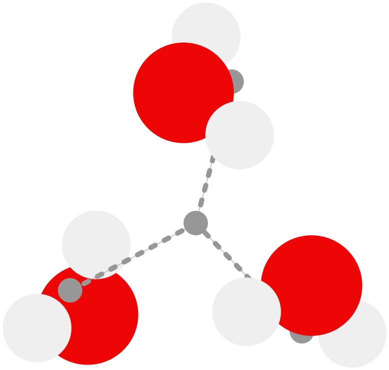


Figure S1: A visual representation of the distance-based size descriptor, *L*, for a typical water trimer. The center of masses of the monomers and structure are shown as gray circles. Dashed lines are the distances from the monomers' center of mass to the structure's center of mass.

S3.1 Optimizing the many-body cutoff

Determining cutoffs typically relies on empirical data such as the acceptable convergence of a large structure's *n*-body energy. We calculated 2-, 3-, and 4-body energies of a 16mer local minima for water,[6] acetonitrile,[8] and methanol[10] with respect to *L*. Geometry optimizations were not performed.

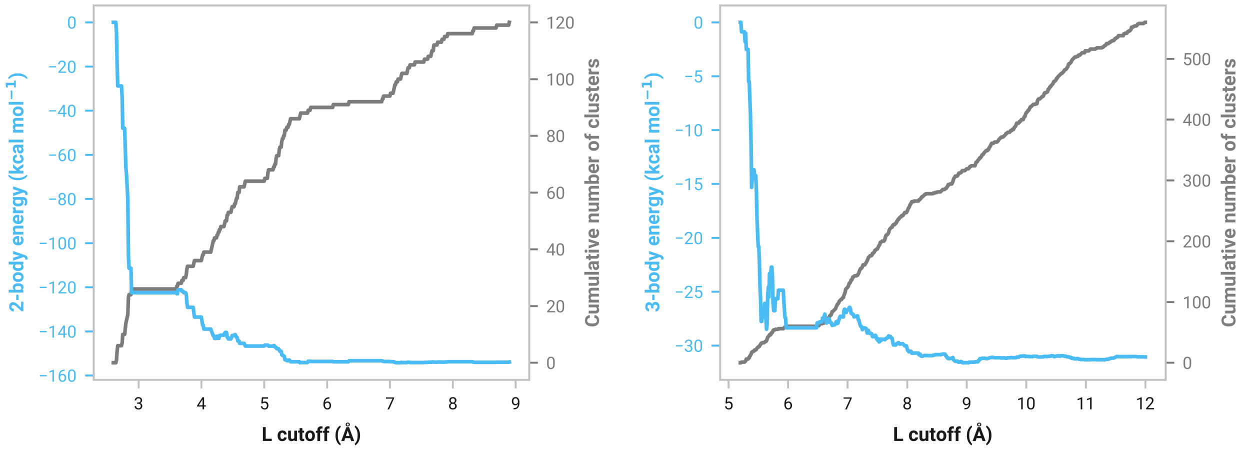
Table S1: *n*-body energies calculated with MP2/def2-TZVP of 16mers from literature in kcal mol^{-1}.

Structure	2-body	3-body	4-body
$(H_2O)_{16}$	−153.9	−31.1	−4.8
$(MeCN)_{16}$	−130.3	3.1	−0.8
$(MeOH)_{16}$	−117.8	−20.6	−2.2

Table S2: Many-body cutoff, *L*, is used for 2- and 3-body models in Å.

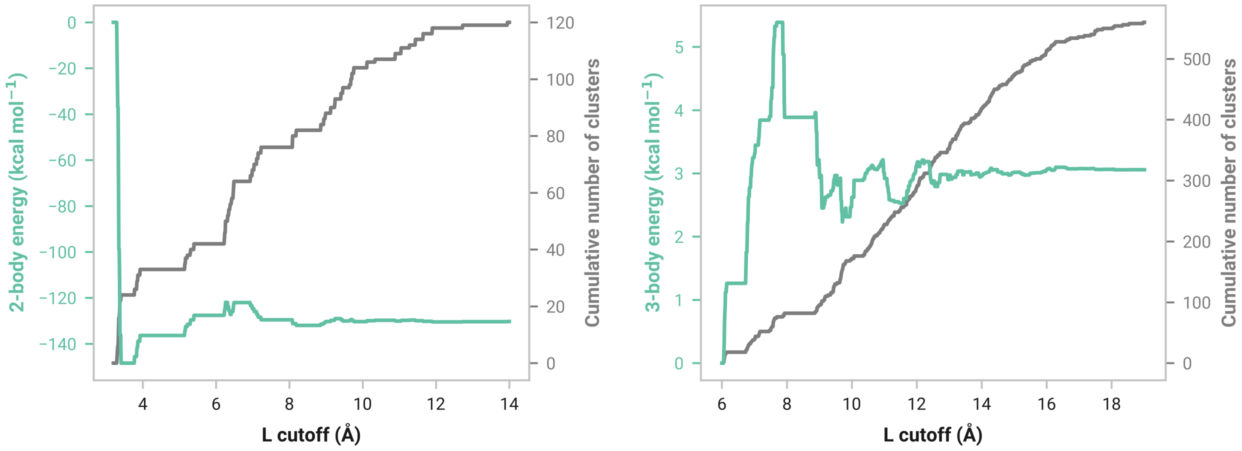
Solvent	2-body	3-body
H_2O	6	10
MeCN	9	17
MeOH	8	14

Table S1 shows total *n*-body energies for each structure and Fig. S2 demonstrates its dependence on the cutoff. The *L* cutoffs that balanced *n*-body accuracy and number of clusters are shown in Table S2 for 2- and 3-body models.



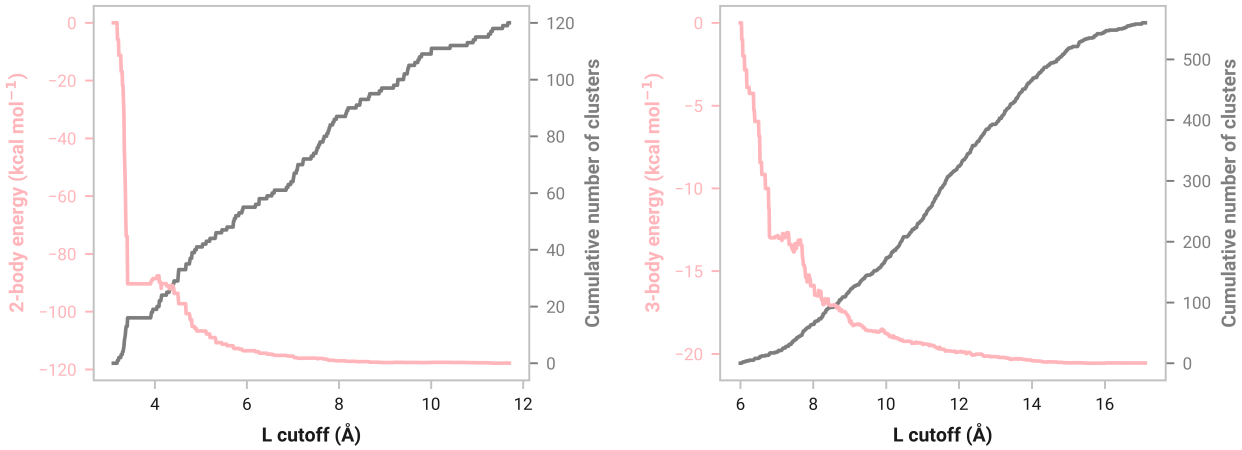
(a) Converged 2-body energy is -153.9 kcal mol^{-1}.   (b) Converged 3-body energy is -31.1 kcal mol^{-1}.

Figure S2: MP2/def2-TZVP *n*-body energies of $(H_2O)_{16}$ calculated with respect to *L*.



(a) Converged 2-body energy is -130.3 kcal mol^{-1}.   (b) Converged 3-body energy is 3.1 kcal mol^{-1}.

Figure S3: MP2/def2-TZVP *n*-body energies of $(MeCN)_{16}$ calculated with respect to *L*.



(a) Converged 2-body energy is -117.8 kcal mol^{-1}. (b) Converged 3-body energy is -20.6 kcal mol^{-1}.

Figure S4: MP2/def2-TZVP n-body energies of $(MeOH)_{16}$ calculated with respect to L.

S4 Data set curation

Trimer data sets were curated by randomly sampling 5000 structures from the production GFN2-xTB MD simulation while enforcing the L cutoff. Monomer and dimer data sets contained all unique structures composing the trimer data set. Data set sizes are shown in Table S3. While this introduces some sampling bias, it also minimizes the number of energy and gradient calculations.

The number of n-body structures grows substantially with the size of the supersystem.[21,22] We implemented a distance-based screening cutoff for all structures used to train mbGDML models to minimize predictions of structures with negligible n-body interactions. The "size" of a structure is computed by summing the distance between each monomer's center of mass to the center of mass of the whole structure. Any structure with a size larger than some predetermined cutoff is excluded from the training set and ignored during mbML predictions.

Table S3: The number of structures included in the n-body data sets.

Solvent	1-body	2-body	3-body
H_2O	14 027	11 124	5 000
MeCN	12 638	9 283	5 000
MeOH	13 006	10 169	5 000

S5 Training

S5.1 GDML

All GDML models were trained with the mbGDML Python package with physical symmetries (i.e., sGDML). We hereby drop the "s" from this point forward. GDML models for 1-, 2-, and 3-body interactions were trained using an iterative training procedure described in ref. 23. An initial model was trained on 200 data points by randomly sampling structures while attempting to preserve the data set energy distribution based on a histogram with bins determined by the Freedman-Diaconis rule.[24] A clustering algorithm was then used to distribute all structures into 50 groups based on geometric and energetic similarities. Force predictions of each group using the initial model were used to identify 100 representative structures with significant root-mean-squared error (RMSE) to include in the following training set. This procedure was repeated until a model trained on 1000 data points was obtained. Following this training procedure resulted in models that have more consistent performance (e.g., lower mean and maximum errors) across the data set.

Training GDML models primarily involves optimizing the kernel length scale, σ, by choosing the model with the lowest validation loss. The sGDML code uses force RMSE as the loss function. In some cases—especially during early models—only considering forces resulted in large energy errors. For example, Fig. S5 shows how optimizing sigma, σ, by just considering forces would result in rapidly rising energy errors. This is often not the case; the optimal hyperparameters usually minimize energy and force RMSE. To safely automate the learning procedure, we employed a weighted energy and force loss function was used to optimize GDML hyperparameters:

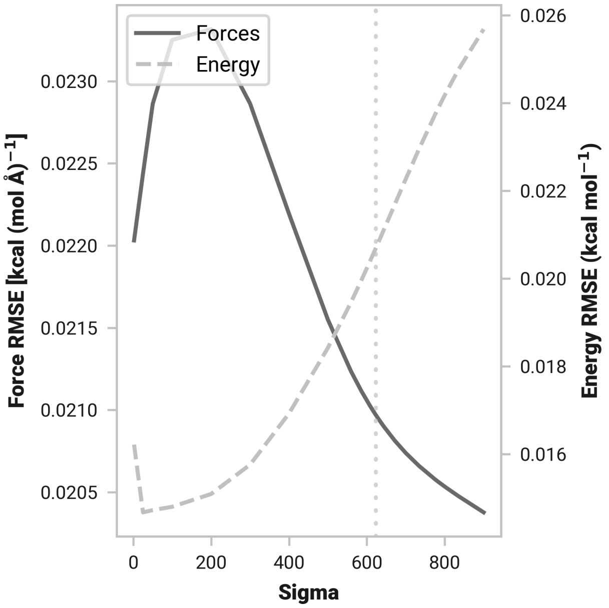


Figure S5: Force and energy validation RMSEs with respect to σ for the acetonitrile 3-body model with 300 training points. The optimal σ determined by Eqn 2 is marked by a vertical dashed line.

$$l = \frac{\rho}{Q}\left\|E - \hat{E}\right\|^2 + \frac{1}{n_{atoms}Q}\sum_{i=0}^{n_{atoms}}\left\|\mathbf{F_i} - \widehat{\mathbf{F}}_\mathbf{i}\right\|^2; \tag{2}$$

where ρ is a trade-off between energy and force errors, Q is the number of validation structures (2000 in our case), $\|\ldots\|$ is the L2 norm, and $\widehat{}$ is the model property prediction. Here, we used a ρ of 0.01, which places minimal importance on energy accuracy. The optimal σ for all final models happened to also have the lowest force RMSE.

Kernel length scale, σ, was optimized by a coarse grid search and then refined with Bayesian optimization.[25] Our σ values were not restricted to integers as in the sGDML code. Fig. S6 shows an example loss optimization curve of the water 2-body GDML model.

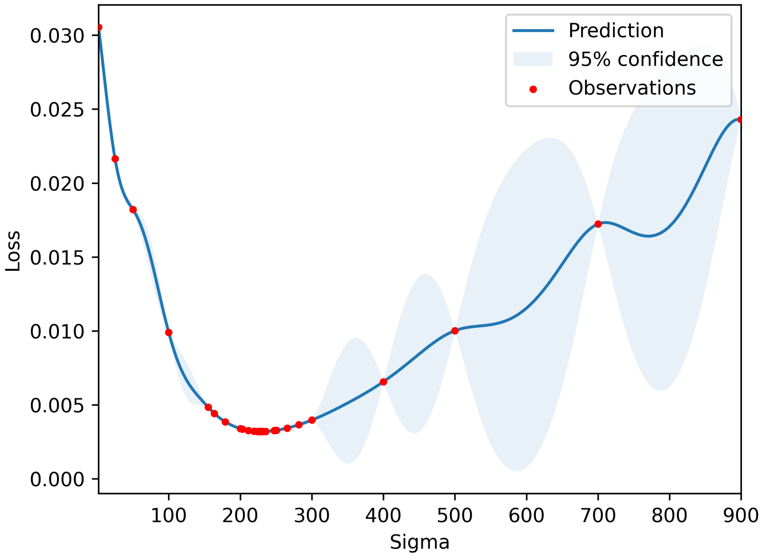


Figure S6: Example Bayesian optimization of the water 2-body GDML model when training on 1000 data points.

S5.1.1 Training statistics

Figures S7 and S8 show the energy and force RMSEs for the test set during training. Note that these data are not strict test sets. Models technically see these data when we select the worst-performing structures during the iterative training procedure. However, these test structures are not seen during hyperparameter optimization.

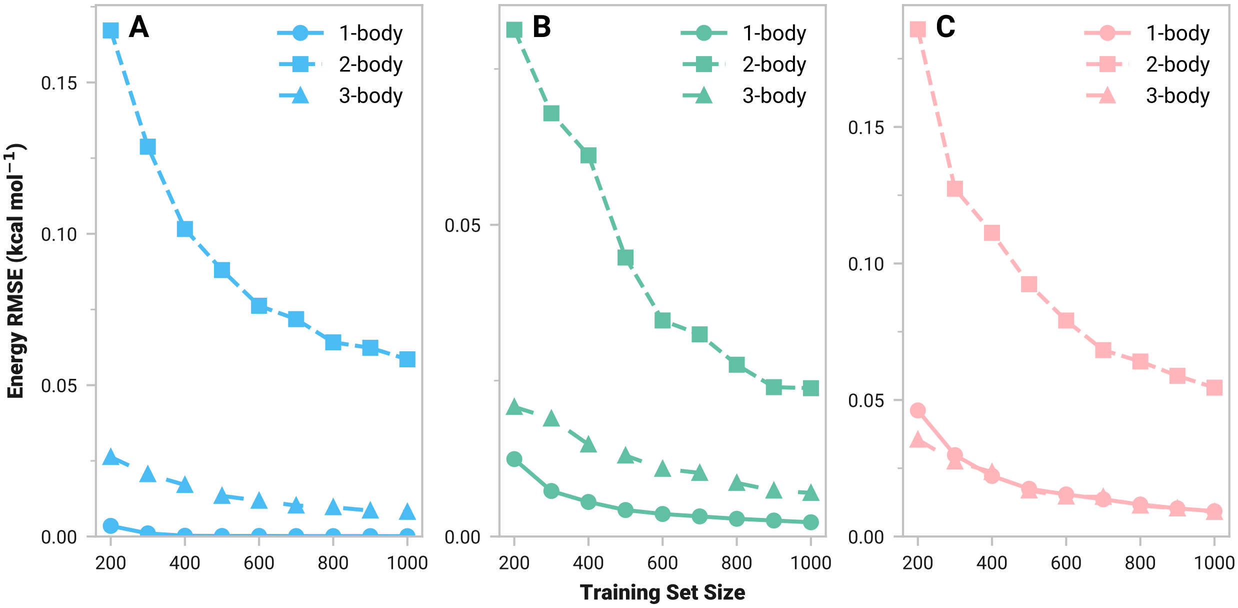


Figure S7: Model energy RMSEs (kcal mol^{-1}) of test sets during each training stage for (A) water, (B) acetonitrile, and (C) methanol.

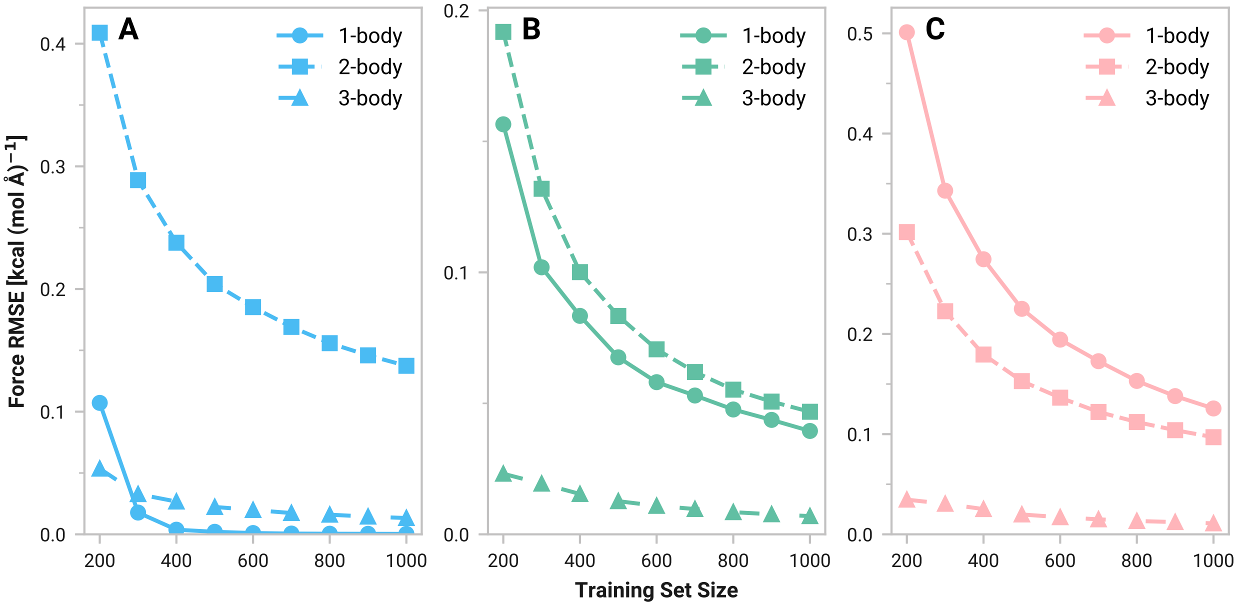


Figure S8: Model force RMSEs in kcal (mol Å)$^{-1}$ of test sets during each training stage for (A) water, (B) acetonitrile, and (C) methanol.

S5.2 SchNet and GAP

Both SchNet and GAP were trained on the same training set as GDML. A simple grid search was performed for optimal hyperparameters, which are listed below. Note that these SchNet and GAP models do not represent their highest performance. An iteratively trained approach should result

Table S4: Hyperparameters used for n-body SchNet models with 5 interaction blocks.

Solvent	n-body	Cutoff	Gaussians
H_2O	1	3	30
	2	10	25
	3	10	25
MeCN	1	5	50
	2	10	25
	3	10	25
MeOH	1	5	50
	2	10	25
	3	10	25

Table S5: Hyperparameters used for n-body GAP models with energy and force sigma being 0.001 and 0.01. The number of sparse points was set to 4000, but the actual number was typically smaller.

Solvent	n-body	n	l	cutoff	delta	zeta
H_2O	1	6	6	6	0.4	4
	2	12	6	9	0.4	4
	3	8	6	4	0.2	3
MeCN	1	12	8	4	0.2	4
	2	12	6	10	0.2	4
	3	6	6	6	0.2	4
MeOH	1	12	8	4	0.2	4
	2	16	6	12	0.1	3
	3	6	6	4	0.2	4

in superior models; however, attempts for the worst-performing models did not provide substantial improvement. For example, Fig. S9 shows the force MSE of an iteratively trained methanol 2-body GAP model. Overall, this iterative training procedure did not improve this case's force MSE.

Sometimes the loss function would increase (e.g., from 400 to 500 in Fig. S9), which seems counterproductive. Each iteration adds 100 structures from the worst-performing groups to the next training set. Often these structures come from a small portion of the data set. Thus, adding these structures will slightly increase the error of the more common points. Reducing the maximum force errors is the objective after each iteration.

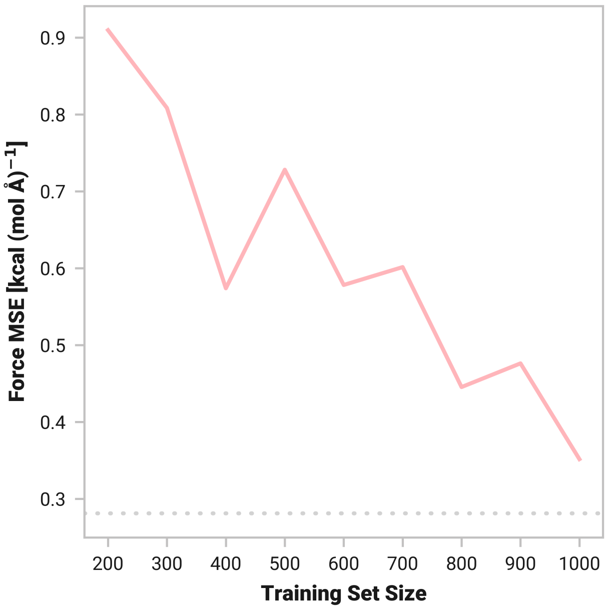


Figure S9: Force MSE in kcal (mol Å)$^{-1}$ for an iteratively trained GAP starting from 200 random structures on the entire data set (training, validation, and test). The horizontal dotted line is the model trained from the GDML training set.

S5.3 NequIP

For NequIP,[26] the objective was to learn total energies and forces to analyze size transferability. However, the largest training structures are limited to trimers for a fair comparison between the many-body ML potentials because 1- and 2-body structures are taken from the 3-body data set. Each solvent potential was trained on 1000 randomly selected trimers with a validation set of 2000 structures. Models used five interaction blocks, a feature multiplicity of 32, two radial layers with 64 hidden neurons, with the Adam optimizer and a learning rate of 0.01. Models with radial cutoffs of 4, 6, 8, and 10 Angstroms were trained, and their test errors are shown in Table S6. A cutoff of 8 Angstroms was selected based on a balance of accuracy and speed for larger structures.

Table S6: Energy and force MAEs for each NequIP model in kcal mol^{-1} and kcal (mol Å)$^{-1}$, respectively. Selected model values are shown in **bold**.

Solvent	Property	Cutoff			
		4	6	8	10
H_2O	Energy	0.3764	0.0592	**0.0235**	0.0236
	Force	0.4229	0.0934	**0.0599**	0.0594
MeCN	Energy	0.6582	0.2774	**0.1406**	0.1390
	Force	0.1241	0.0872	**0.0753**	0.0746
MeOH	Energy	0.2671	0.1046	**0.0588**	0.0592
	Force	0.1894	0.1242	**0.0983**	0.0984

S6 Timings

As mentioned in the manuscript, our objective was to develop a framework for a rapidly trained ML force field for arbitrarily large systems. We provide training timings for the many-body machine learning methods in Table S7. We advise the reader to be cautious when interpreting these training timings as they do not reflect best-case scenarios and could potentially be improved.

Table S7: Mean time (in seconds) to train a model on 1000 structures for a single set of hyperparameters (GDML and GAP) or epoch (SchNet and NequIP).

Solvent	N monomers	GDML	GAP	SchNet	NequIP
H_2O	1	9	82	116	
	2	36	1111	169	
	3	138	903	155	50
MeCN	1	33	2671	132	
	2	303	9549	161	
	3	722	4165	161	71
MeOH	1	31	2818	101	
	2	301	18016	182	
	3	667	2701	158	84

For example, direct timing comparisons between kernel methods (GDML and GAP) and neural networks (SchNet and NequIP) are nontrivial. GDML and GAP models are trained by selecting hyperparameters, computing parameters, then validating the model on a subset of structures. This is repeated until an optimal set of hyperparameters is found—around 20 iterations are used for GDML. GDML timings are almost entirely dependent on the number of atoms. GAP timings, however, mainly depend on the number of radial and angular basis functions and radial cutoff.

Timings for the neural network methods (SchNet and NequIP) are the average epoch time on a GPU. SchNet and NequIP required around 300 and 6500 epochs for highly accurate models, respectively. Neural network architecture and radial cutoffs are the primary influences on training time.

Prediction timings were computed by running a 1 ps NVE MD simulation on a randomly generated hexamer. Each simulation was initialized with the same atomic positions and velocities. Different methods were used to drive the MD simulation, including MP2. Table S8 shows the cumulative time to run the simulation.

Table S8: Time (in seconds) and speedup factor to run a 1 ps NVE MD simulation driven by various methods. Relative energy RMSEs with respect to starting structure are provided in kcal mol^{-1}. Both speedup and energy RMSEs are with respect to MP2/def2-TZVP values.

Solvent	Method	Time	Speedup	Energy RMSE
H_2O	MP2/def2-TZVP	31 884		
	RI-MP2/def2-TZVP	18 455	1.7	1.0
	MP2/def2-SVP	11 993	2.7	1.6
	mbGDML	271	117.7	1.5
	mbGAP	598	53.3	2.0
	mbSchNet	178	179.1	2.3
	GFN2-xTB	18	1 732.9	2.0
MeCN	MP2/def2-TZVP	1 409 317		
	RI-MP2/def2-TZVP	253 679	5.6	0.1
	MP2/def2-SVP	113 358	12.4	5.1
	mbGDML	405	3 476.0	2.6
	mbGAP	1 689	834.4	2.6
	mbSchNet	146	9 655.4	2.8
	GFN2-xTB	61	23 169.5	3.9
MeOH	MP2/def2-TZVP	430 192		
	RI-MP2/def2-TZVP	109 937	3.9	0.1
	MP2/def2-SVP	70 927	6.1	3.0
	mbGDML	624	689.2	2.5
	mbGAP	3 161	136.1	3.7
	mbSchNet	207	2 073.9	3.7
	GFN2-xTB	43	10 029.8	5.1

S7 Isomer rankings

Comparable small isomer rankings for mbGAP and mbSchNet are shown below. 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.

S7.1 Relative errors of mbGAP and mbSchNet

These figures are directly comparable to Fig. 1 in the main text, where the data are relative to the method's lowest energy structure. Predictions of structures with four or more fragments in the present MBE framework will have some neglected higher-order contributions. These contributions will affect absolute energy predictions but are nontrivial to compute without knowing the complete supersystem calculation (i.e., MP2 data). In practice, one would compute the relative isomer energies with respect to the lowest energy from that method. Thus, we opt to present the isomer rankings using this scheme. Furthermore, ML force fields (e.g., mbGDML) are trained on forces and reconstruct energy up to a constant defined for a given training set. Since our training data sets do not contain four and higher-order clusters, one can expect a constant shift in the energy predicted by mbGDML and reference calculations. This is a more fair representation of ML force fields' efficiency than the absolute energy differences, which unavoidably include a systematic error.

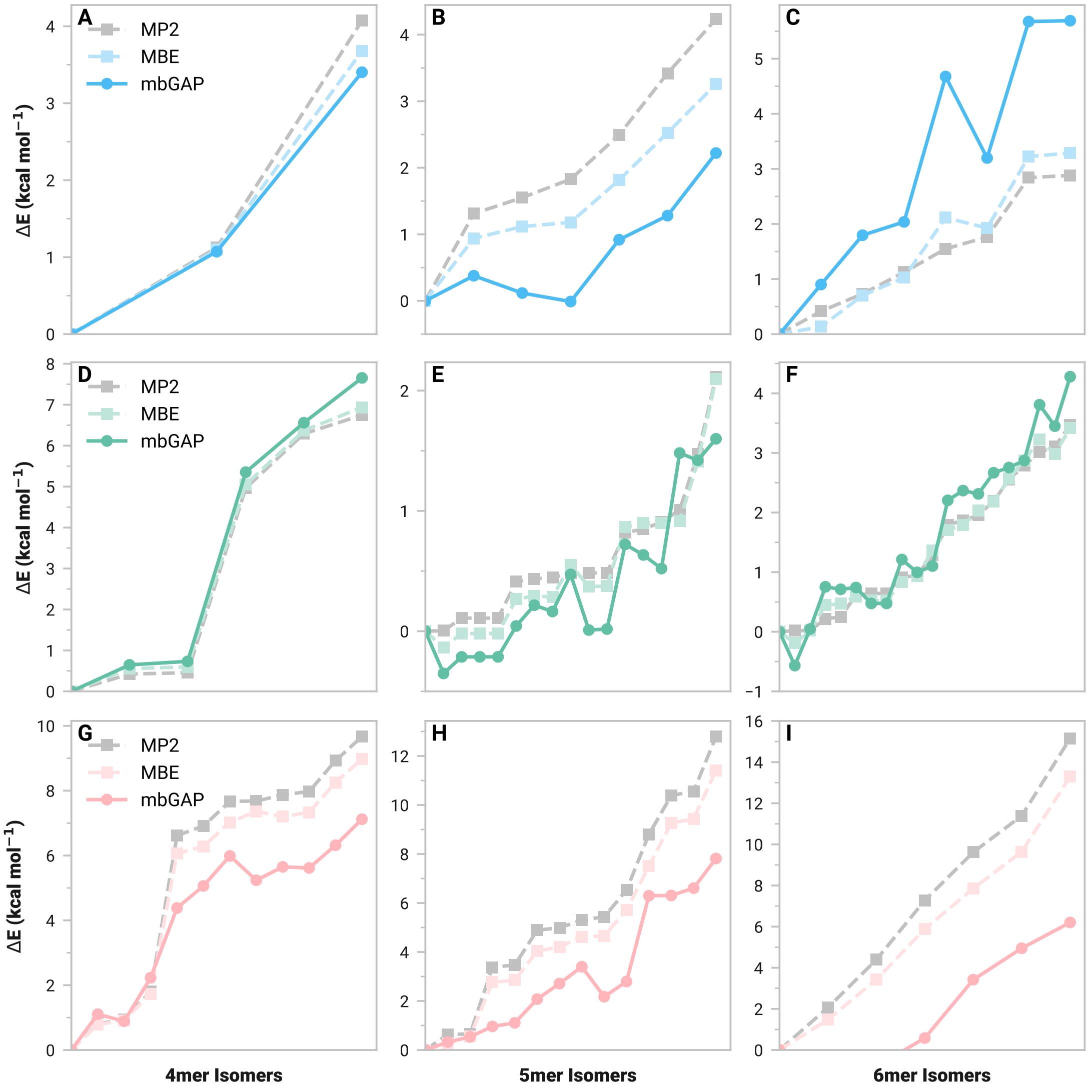


Figure S10: Relative energies (with respect to the method's lowest energy) of isomers containing four, five, and six monomers of (A-C) water, (D-F) acetonitrile, and (G-I) methanol. Dark-colored lines with circles are mbGAP predictions.

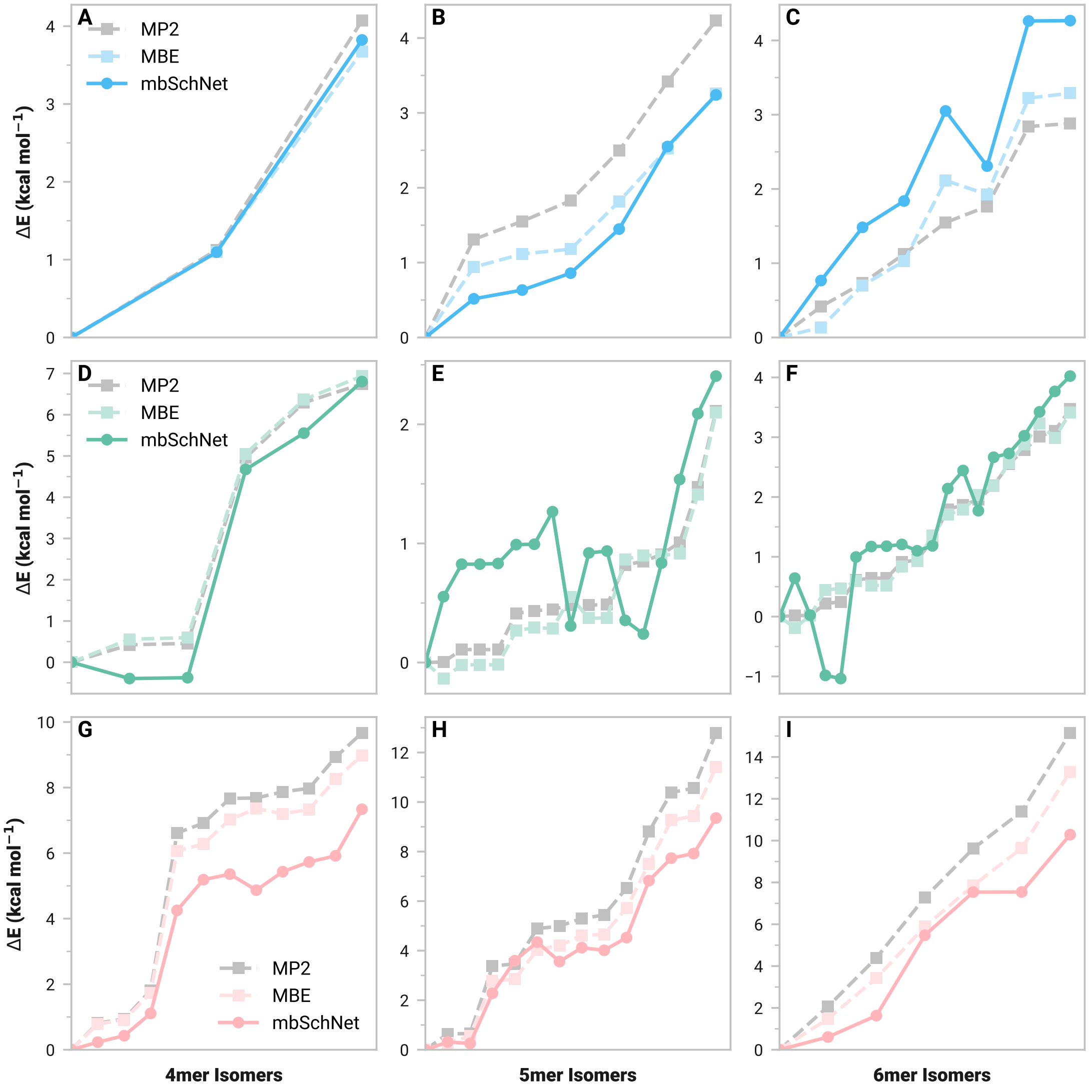


Figure S11: Relative energies (with respect to the method's lowest energy) of isomers containing four, five, and six monomers of (A-C) water, (D-F) acetonitrile, and (G-I) methanol. Dark-colored lines with circles are mbSchNet predictions.

S7.2 Absolute errors of mbGDML, mbGAP, and mbSchNet

The following figures plot relative energies relative to the lowest MP2/def2-TZVP energy. This alternative representation shows where the absolute energy prediction errors originate. For example, Fig. S12I show that MBE errors in the methanol 6mers are primarily from truncated higher-order contributions in the lowest energy structure.

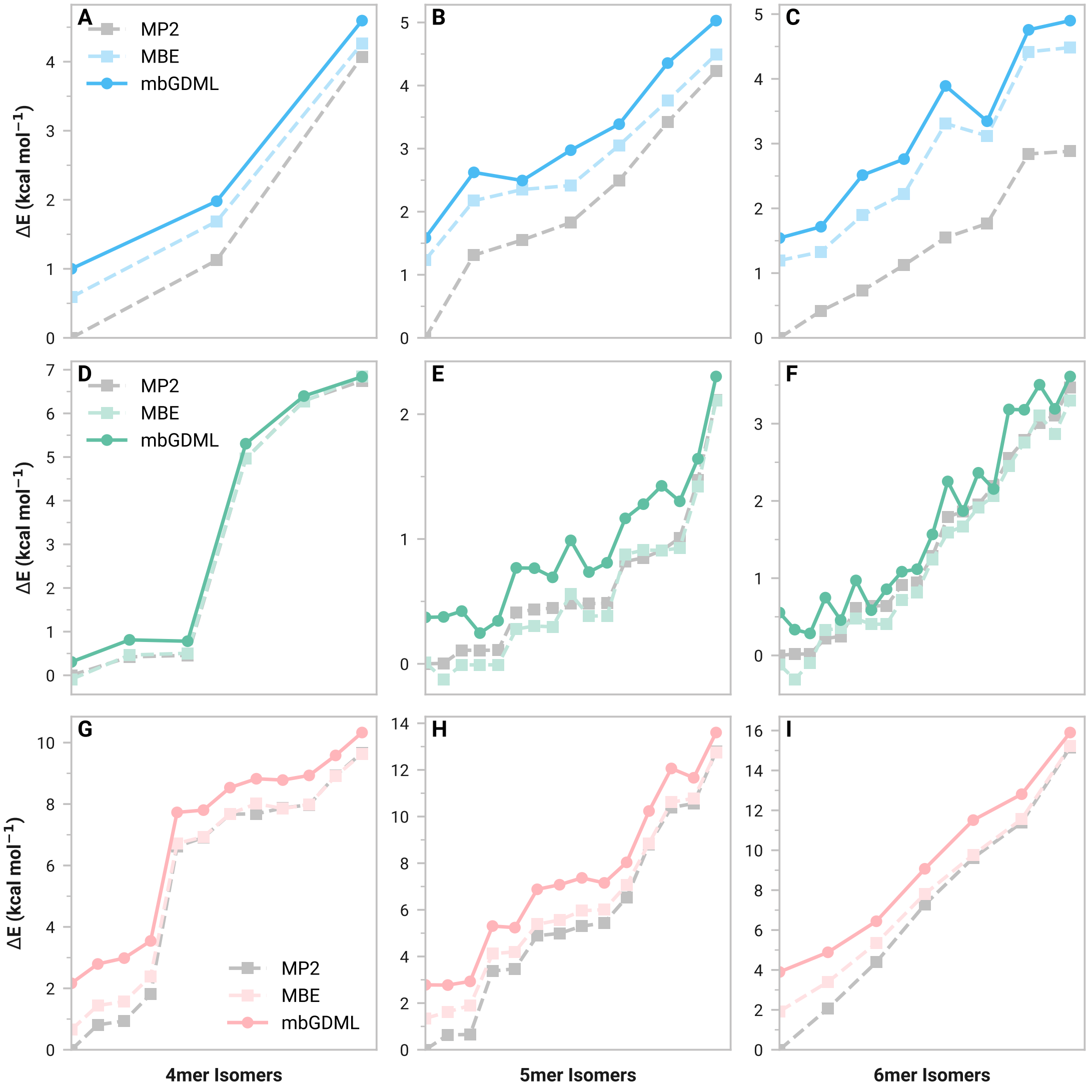


Figure S12: Relative energies (with respect to MP2 lowest energy) of isomers containing four, five, and six monomers of (A-C) water, (D-F) acetonitrile, and (G-I) methanol. Dark-colored lines with circles are mbGDML predictions.

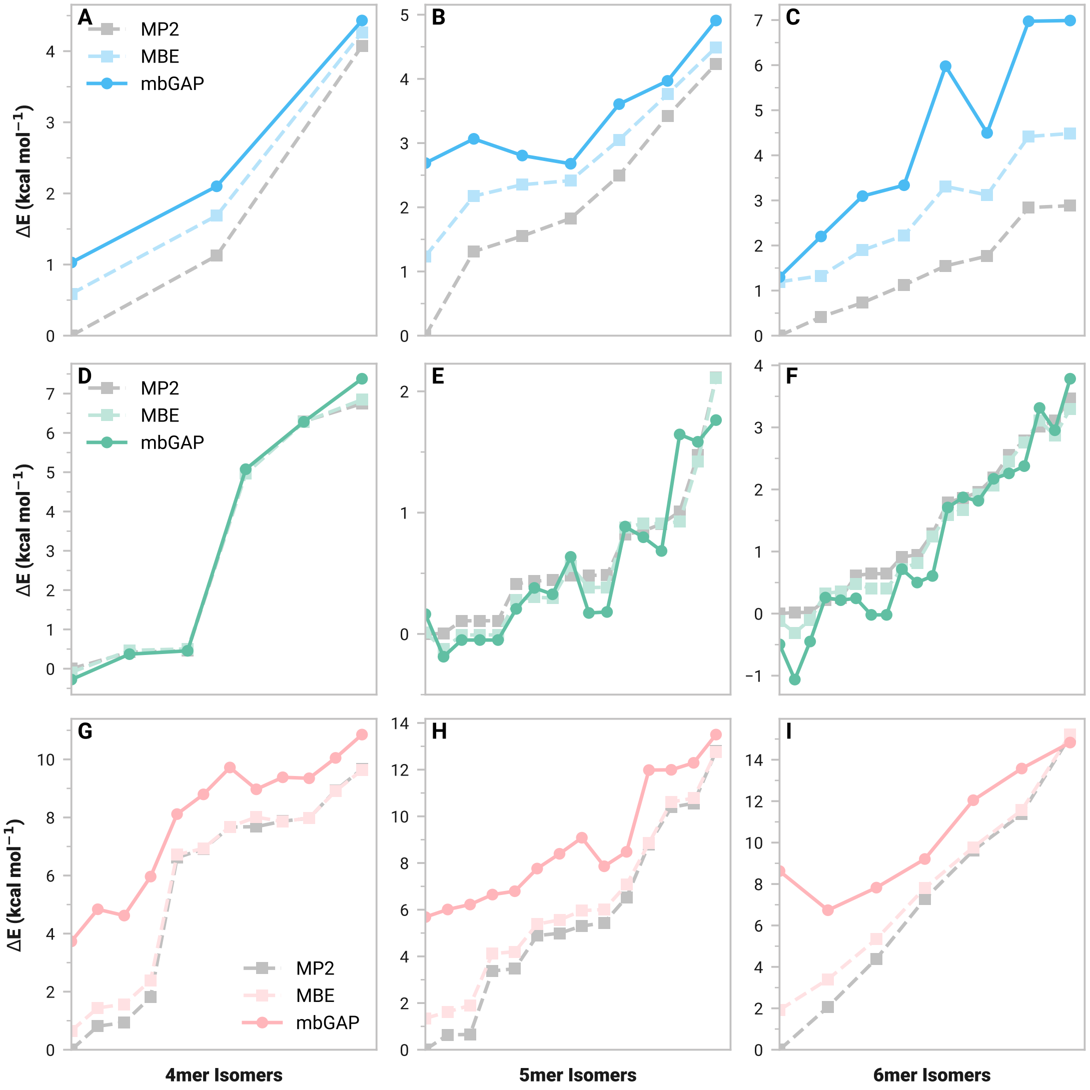


Figure S13: Relative energies (with respect to MP2 lowest energy) of isomers containing four, five, and six monomers of (A-C) water, (D-F) acetonitrile, and (G-I) methanol. Dark-colored lines with circles are mbGAP predictions.

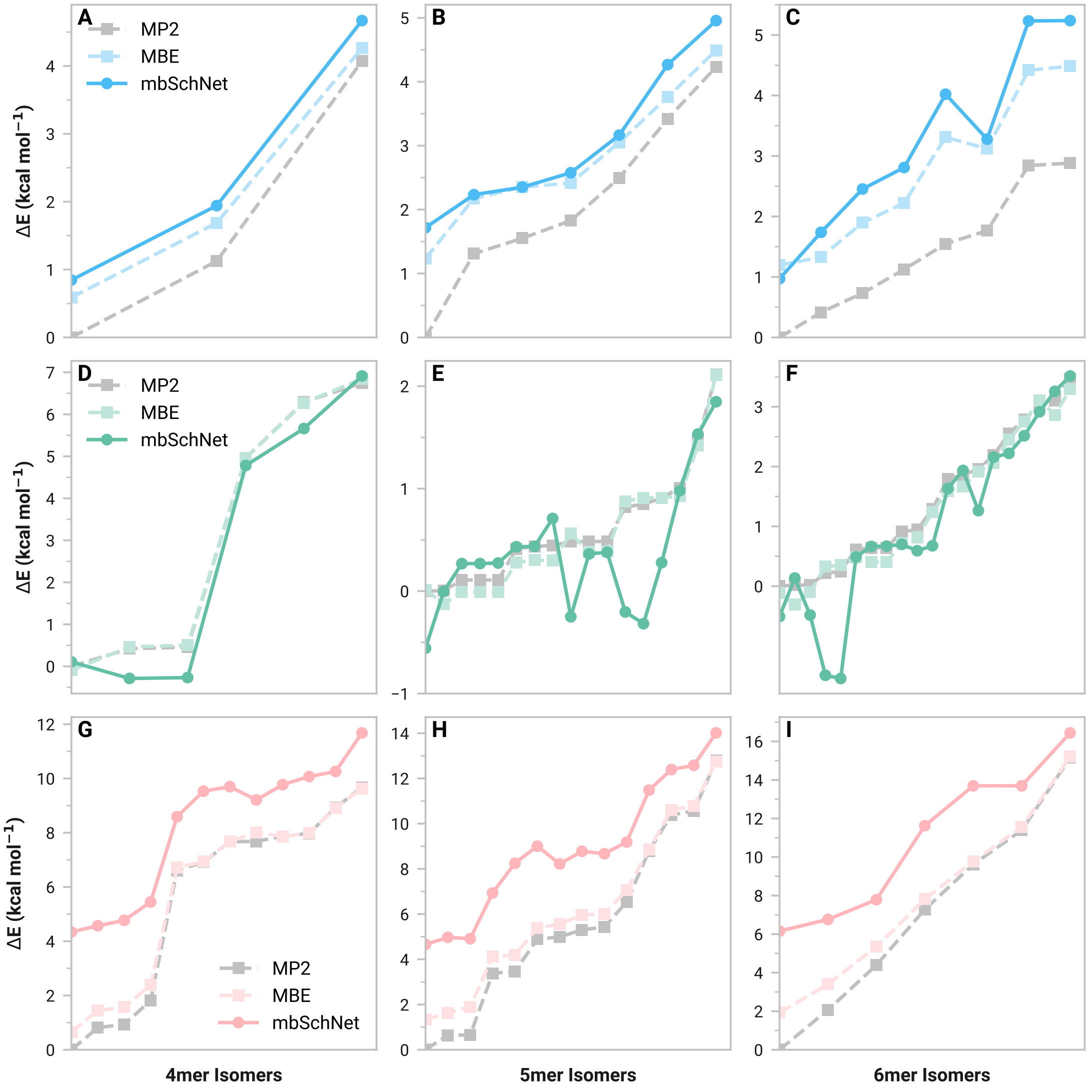


Figure S14: Relative energies (with respect to MP2 lowest energy) of isomers containing four, five, and six monomers of (A-C) water, (D-F) acetonitrile, and (G-I) methanol. Dark-colored lines with circles are mbSchNet predictions.

S7.3 Performance statistics

Table S9: MBE and mbML energy (kcal mol^{-1}) and force [kcal (mol Å)$^{-1}$] MAEs of isomers containing four to six monomers with respect to supersystem MP2/def2-TZVP calculations. Best values are shown in **bold**.

Solvent	Method	4mers		5mers		6mers	
		Energy	Force	Energy	Force	Energy	Force
H_2O	MBE	0.4482	0.2575	0.6635	0.3221	1.3333	0.5429
	mbGDML	0.7928	**0.5298**	1.0876	0.6274	**1.7650**	0.8681
	mbGAP	0.7879	0.6913	1.2695	0.8399	2.8829	1.2220
	mbSchNet	**0.7534**	0.6009	**0.9187**	**0.6055**	1.8043	**0.7738**
MeCN	MBE	0.0500	0.0094	0.0850	0.0161	0.1484	0.0234
	mbGDML	0.2603	**0.1044**	0.3173	0.1611	**0.2888**	**0.1782**
	mbGAP	**0.1806**	0.1841	**0.2008**	0.2183	0.3431	0.2573
	mbSchNet	0.4200	0.1411	0.3215	**0.1583**	0.3916	0.1808
MeOH	MBE	0.2531	0.0968	0.6010	0.1777	0.7333	0.1908
	mbGDML	**1.2598**	**0.7791**	**1.8054**	**0.9346**	**2.0891**	**0.8724**
	mbGAP	2.2924	1.1790	3.2052	1.4519	3.3708	1.4479
	mbSchNet	2.5887	1.0183	3.3021	1.2581	3.7513	1.2244

S7.4 Effect of larger basis set

The model chemistry used for the many-body data uses smaller basis sets than recommended for benchmark predictions using MBEs. As previously mentioned, our level of theory was selected for its balance of cost and accuracy. Table S10 demonstrates the expected accuracy improvement from using a much larger basis set: aug-cc-pVTZ. We certainly see improvement; however, this was not needed to demonstrate the effectiveness of mbGDML.

Table S10: Energy MAE (kcal mol^{-1}) of MBE with respect to MP2 calculations of various sized water isomers with the def2-TZVP and aug-cc-pVTZ basis sets.

Monomers	def2-TZVP	aug-cc-pVTZ
4	0.448	0.398
5	0.664	0.518
6	1.333	1.259

S8 (MeCN)$_{16}$ prediction analysis

S8.1 GDML and SchNet feature space

GDML is based on a kernel ridge estimator with the Matérn 5/2 kernel, $k_{5/2}(x_i,x_j)$,

$$k_{5/2}(x_i,x_j) = \left(1 + \frac{\sqrt{5}}{\sigma} d(x_i,x_j) \ + \frac{5}{3\sigma} d(x_i,x_j)^2\right) \exp\left(-\frac{\sqrt{5}}{\sigma} d(x_i,x_j)\right). \tag{3}$$

In GDML literature, σ is the kernel length-scale hyperparameter, and $d(x_i,x_j)$ is the Euclidean distance between x_i (the structure to predict) and x_j (a single training point). Note that these covariances are never explicitly computed during training and predictions; in practice, GDML uses the Hessian of the Matérn 5/2 kernel, $\mathrm{Hess}(k_{5/2})$. We use $k_{5/2}$ as our GDML feature space because of its straightforward interpretation for UMAP embedding. SchNet's feature space was the readout before the final dense atom-wise layer and pooling.

S8.2 Geometry descriptor

When visualizing the ML potential feature space, we found structures with high prediction errors clustered next to training data that were visually different. This could indicate that the ML potential interprets these structures are similar when they are not. A complete atomic position description is unnecessary since only general configurational differences are desired. Whatever descriptor is chosen should map to $\mathbb{R}$ for straightforward implementation in color maps.

We use an ad hoc, simple geometry descriptor of N acetonitrile molecules, g_N, as

$$g_N = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \frac{\theta_{ij}}{d_{ij}}. \tag{4}$$

The indices i and j represent one of the N molecules, and d_{ij} is the distance between their center of mass. We define a fictitious vector from the methyl carbon to the nitrogen to compute the angle between the two molecules, θ_{ij}. This angle is computed in the standard way with two vectors v_i and v_j,

$$\cos\theta_{ij} = \frac{v_i \cdot v_j}{||v_i|| \, ||v_j||}. \tag{5}$$

Thus, the geometry descriptor for our acetonitrile trimer would be

$$g_3 = \frac{\theta_{12}}{d_{12}} + \frac{\theta_{13}}{d_{13}} + \frac{\theta_{23}}{d_{23}}. \tag{6}$$

Conceptually, this is the cumulative ratio of rotation to translation of the first molecule moving into each position and back. For a dimer, however, this would only be one translation and rotation,

$$g_2 = \frac{\theta_{12}}{d_{12}}. \tag{7}$$

When using this geometry descriptor as the coloring scheme, we get the following UMAP embedding.

S8.3 Trimer embeddings

Below are the UMAP embeddings of the GDML and SchNet feature space for acetonitrile 3-body structures. GDML performs well with a sum of squared error of 0.291 (kcal mol^{-1})^2, whereas SchNet is 2.263 (kcal mol^{-1})^2.

Fig. S15 shows the SchNet feature space, a 2D embedding of trained and 3-body structures from $(MeCN)_{16}$ using UMAP.[27] A rather large cluster of points exists around (10, 2) with a decent overlap of test and train data. These structures all look similar to SchNet in feature space, according to UMAP. However, several test structures are isolated from training data, which should result in higher errors. Indeed, when we examine the testing data for energy error, the isolated structures generally have higher errors (Fig. S15A).

For example, some test structures embedded near training structures (4, 7) have high errors. This is surprising as SchNet should have learned similar structures. Further analysis reveals that while SchNet determines these structures to be similar in feature space, they are geometrically different. A simple, ad hoc geometry descriptor shows that all high-error structures are dissimilar to anything in the training set. SchNet has some difficulty with these structures, which results in a significant 16.1 kcal mol^{-1} error.

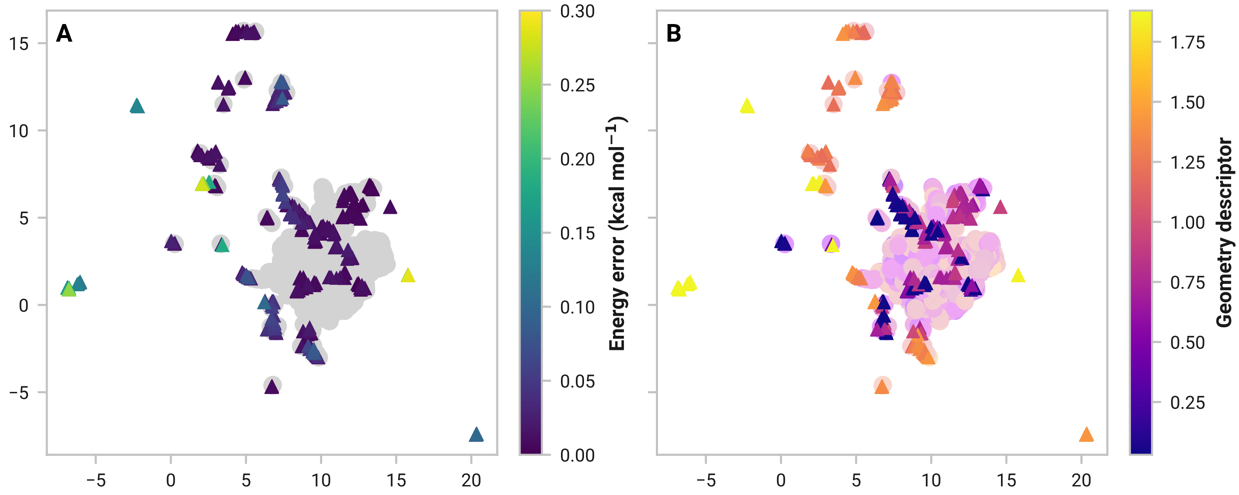


Figure S15: UMAP embeddings of acetonitrile, 3-body SchNet feature space of the training set (circles), and $(MeCN)_{16}$ structure (triangles). Points near each other are similar in high-dimensional feature space. (A) SchNet absolute prediction error of 3-body structures from $(MeCN)_{16}$. The maximum error is 0.289 kcal mol^{-1}, but the color scale is normalized to errors from all models. (B) Geometry descriptor of each structure. Similar values (i.e., colors) indicate similar geometries.

S8.4 Dimer embeddings

Below are the UMAP embeddings of the GDML and SchNet feature space for acetonitrile 2-body structures. GDML and SchNet perform well with squared errors of 0.04 (kcal mol^{-1})^2 and 0.08 (kcal mol^{-1})^2, respectively.

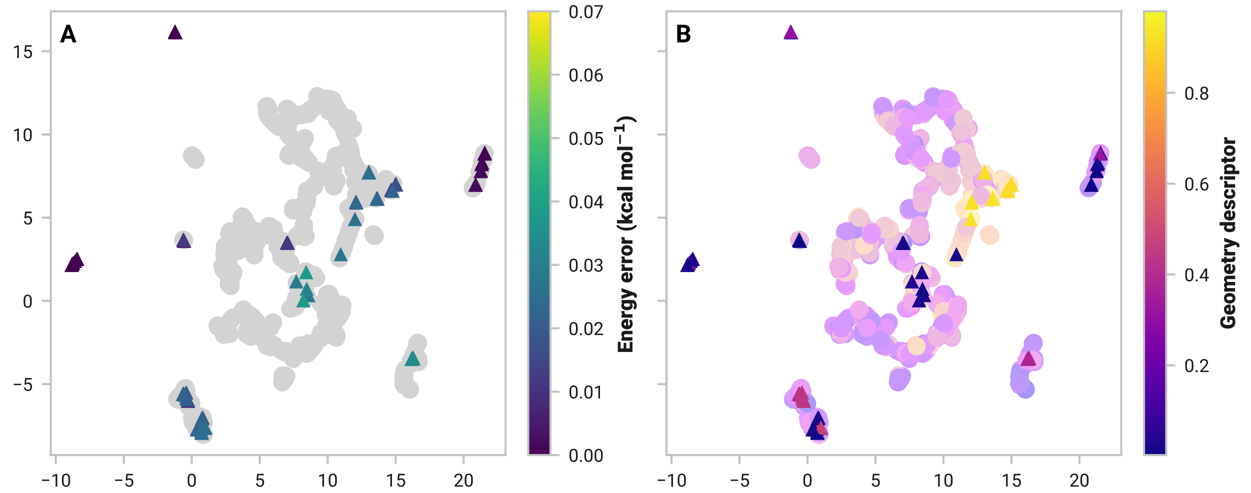


Figure S16: UMAP embeddings of acetonitrile, 2-body GDML feature space of the training set (circles), and $(MeCN)_{16}$ structure (triangles). Points near each other are similar in high-dimensional feature space. (A) GDML absolute prediction error of 3-body structures from $(MeCN)_{16}$. The maximum error is 0.038 kcal mol^{-1}, but the color scale is normalized to errors from all models. (B) Geometry descriptor of each structure. Similar values (i.e., colors) indicate similar geometries.

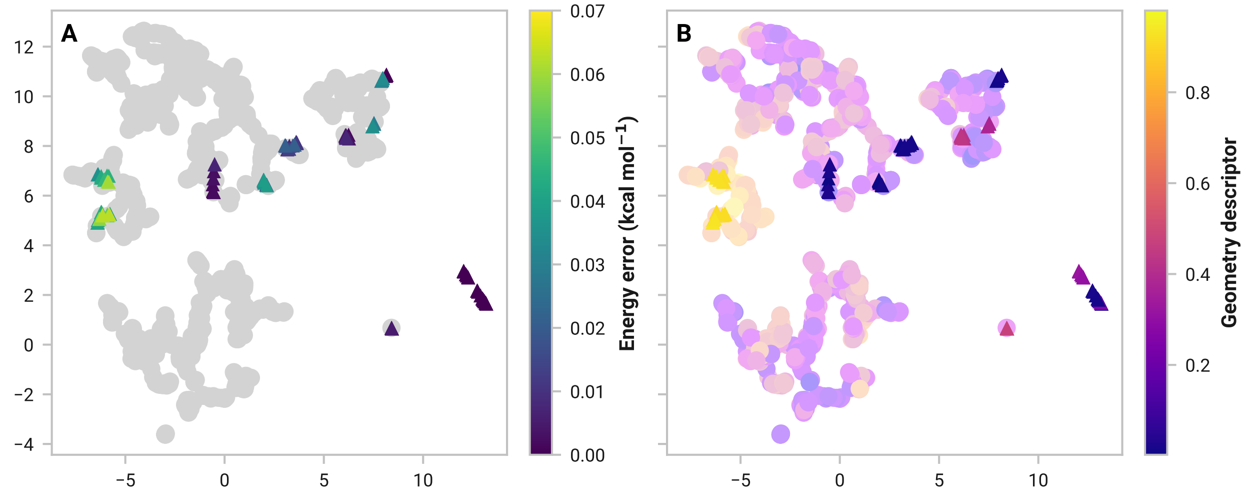


Figure S17: UMAP embeddings of acetonitrile, 2-body SchNet feature space of the training set (circles), and $(MeCN)_{16}$ structure (triangles). Points near each other are similar in high-dimensional feature space. (A) SchNet absolute prediction error of 3-body structures from $(MeCN)_{16}$. The maximum error is 0.063 kcal mol^{-1}, but the color scale is normalized to errors from all models. (B) Geometry descriptor of each structure. Similar values (i.e., colors) indicate similar geometries.

S9 Molecular dynamics simulations

Periodic MD simulations at constant volume and temperature (NVT) were performed in the atomic simulation environment (ASE; v3.22.1)[28] by using the minimum-image convention (MIC). After defining the periodic cell vectors, the MIC is applied to every n-body structure within the many-body and MIC cutoff before predictions are made.

Starting geometries were initialized with packmol (v20.2.2)[20] at 300 K mass density. Box lengths were 16 Å (137 molecules), 18 Å (67 molecules), and 16 Å (61 molecules) for water, acetonitrile, and methanol, respectively. Geometry optimizations were first done with the BGFS optimizer in ASE with a maximum force convergence criteria of 4.6 kcal (mol Å)$^{-1}$ and a maximum of 200 steps. Velocities were initialized at 100 K using the Maxwell-Boltzmann distribution. Temperature, set to 298.15 K, during the MD simulation was controlled with the Berendsen thermostat ($\tau_T = 0.1$ fs) as implemented in ASE. Coordinates were stored at each time step of 1 fs.

S9.1 Radial distribution functions

Radial distribution functions (RDFs), $g_{ab}(r)$, of atoms a and b were computed for water, acetonitrile, and methanol during the production region of the MD simulation. The start of production was determined with the `timeseries.detect_equilibration` function in the `pymbar` package.[29] This resulted in 5.4, 29.5, and 24.9 ps of sampling for water, acetonitrile, and methanol, respectively. Due to the large system size, less sampling was performed for the water MD simulation. For example, the $(MeOH)_{61}$ simulation took approximately 7.4 seconds/step on 12 cores, whereas $(H_2O)_{137}$ was 28.5 seconds/step on 24 cores.

Table S11 shows the difference between the mbGDML predicted RDF and references from the literature.[30–33] The number of nearest neighbors, n_{nearest}, was computed by integrating $g(r)$ up to the first minimum (i.e., first solvation shell). For acetonitrile, C_N represents the nitrile carbon. The H in the methanol RDFs is from the hydroxyl group.

Table S11: Computed radial distribution function properties with deviations from reference data are provided within parentheses.

Solvent	ab	r_{peak}	g_{ab}	n_{nearest}
H_2O	OO	2.75 (−0.04)	2.50 (0.00)	1.28 (0.06)
	OH	3.25 (−0.02)	1.57 (0.10)	0.62 (0.02)
	HH	2.35 (−0.08)	1.39 (0.05)	1.00 (0.08)
MeCN	NN	3.75 (−0.44)	1.18 (−0.17)	1.95 (1.44)[a]
	CN	3.35 (−0.03)	1.75 (−0.28)	2.35 (−0.04)
	C_NC_N	4.75 (0.23)	1.35 (−0.31)	3.33 (0.19)
MeOH	OO	2.85 (0.12)	2.17 (−1.03)	1.43 (−0.14)
	OH	1.95 (0.20)	1.76 (−0.99)	1.14 (0.31)
	HH	2.45 (0.08)	2.19 (−0.57)	2.15 (−0.20)

[a] The experimental reference does not exhibit a split peak; thus, the first solvation shell for the experimental reference is up to 6.29 Å instead of 4.85 Å for the mbGDML simulation.

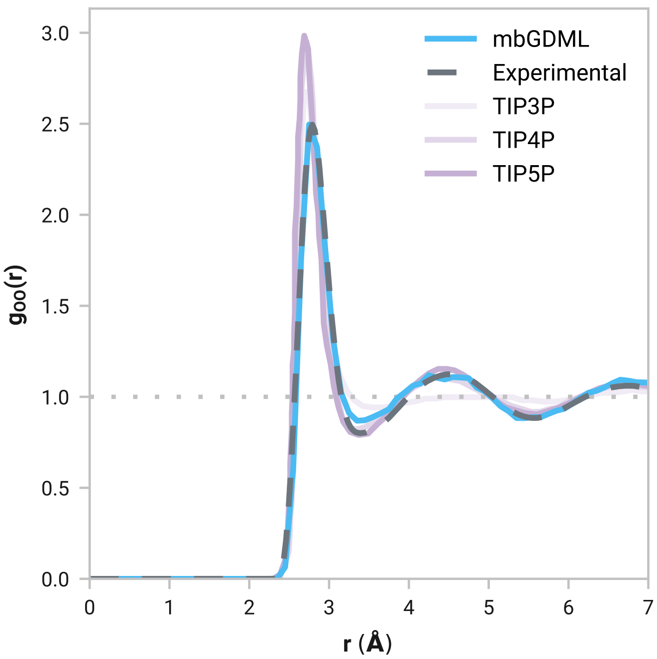


Figure S18: $g_{\mathrm{OO}}(r)$ RDF curve of water from mbGDML MD simulation. Comparisons are made against experimental[30] and classical[34] results.

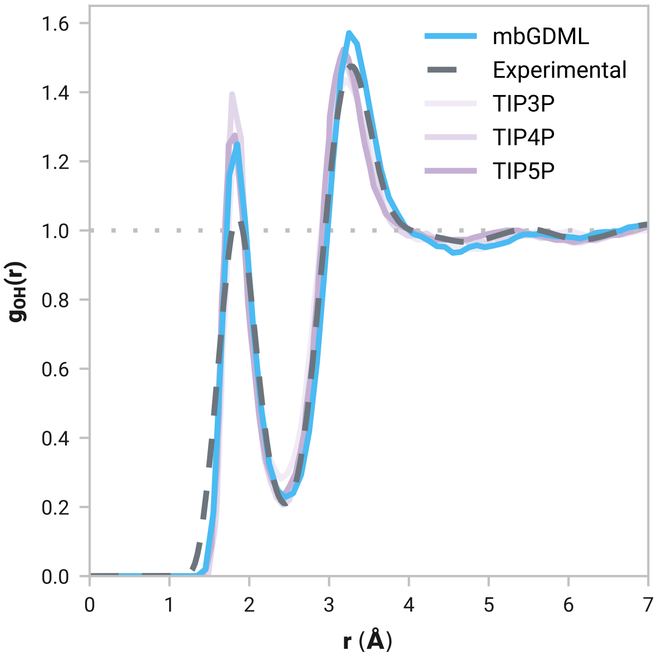


Figure S19: $g_{\mathrm{OH}}(r)$ RDF curve of water from mbGDML MD simulation. Comparisons are made against experimental[30] and classical[34] results.

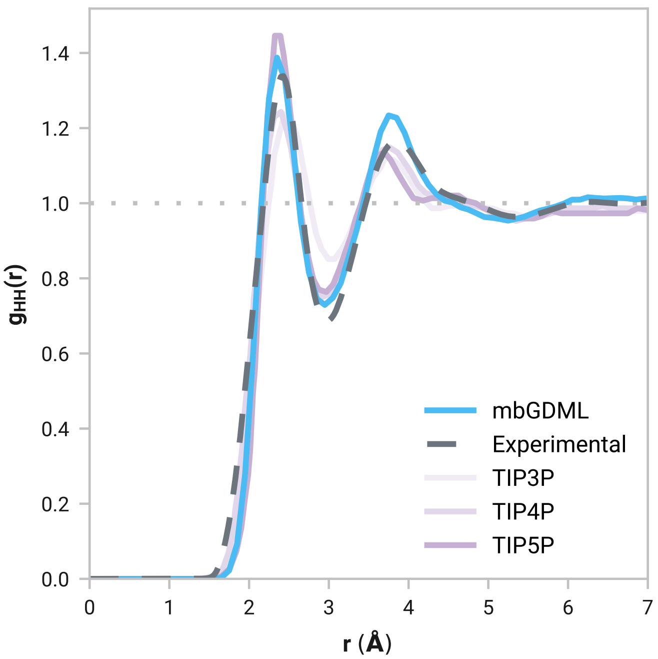


Figure S20: $g_{\mathrm{HH}}(r)$ RDF curve of water from mbGDML MD simulation. Comparisons are made against experimental[30] and classical[34] results.

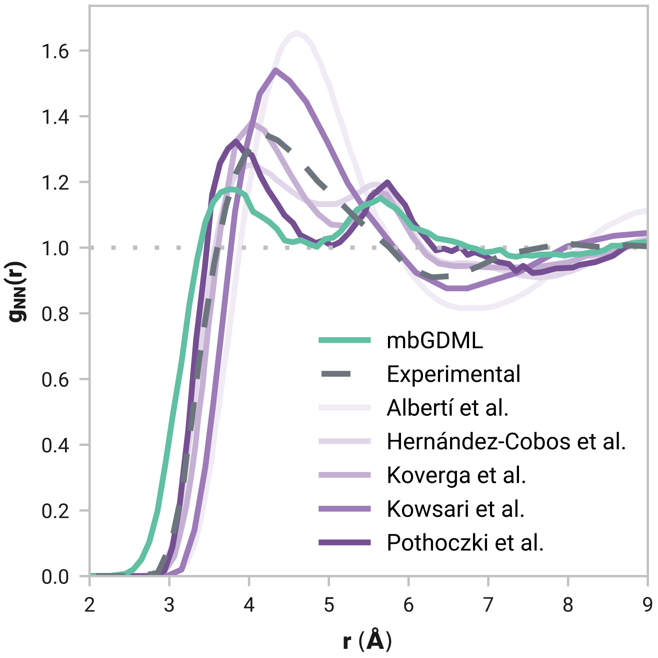


Figure S21: $g_{\mathrm{NN}}(r)$ RDF curve of acetonitrile from NVT simulations at 298.15 K driven by mbGDML. Comparisons are made against experimental[35] and classical[33,36–39] results.

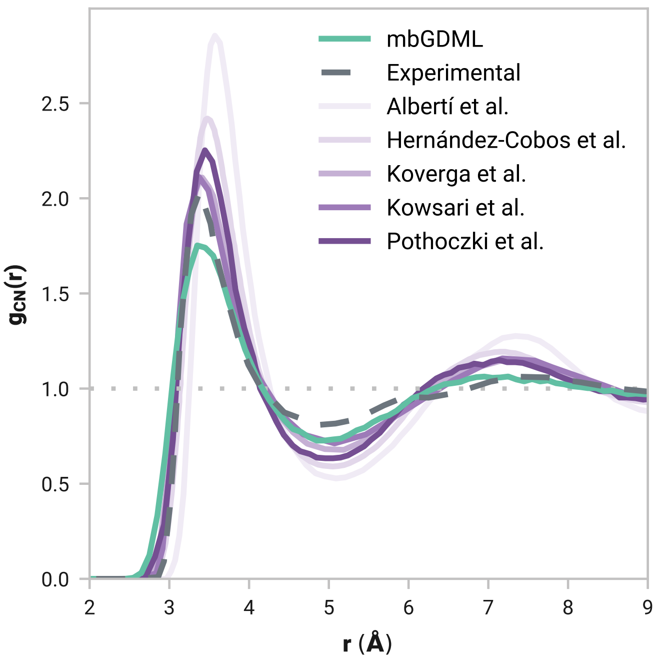


Figure S22: $g_{\mathrm{CN}}(r)$ RDF curve of acetonitrile from NVT simulations at 298.15 K driven by mbGDML. Comparisons are made against experimental[35] and classical[33,36–39] results.

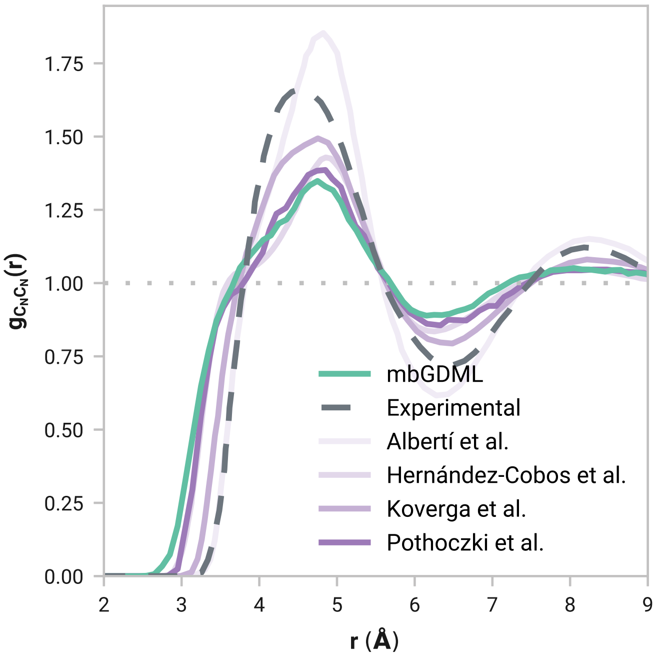


Figure S23: $g_{\mathrm{C_N C_N}}(r)$ RDF curve of acetonitrile from NVT simulations at 298.15 K driven by mbGDML. Comparisons are made against experimental[35] and classical[33,36,37,39] results.

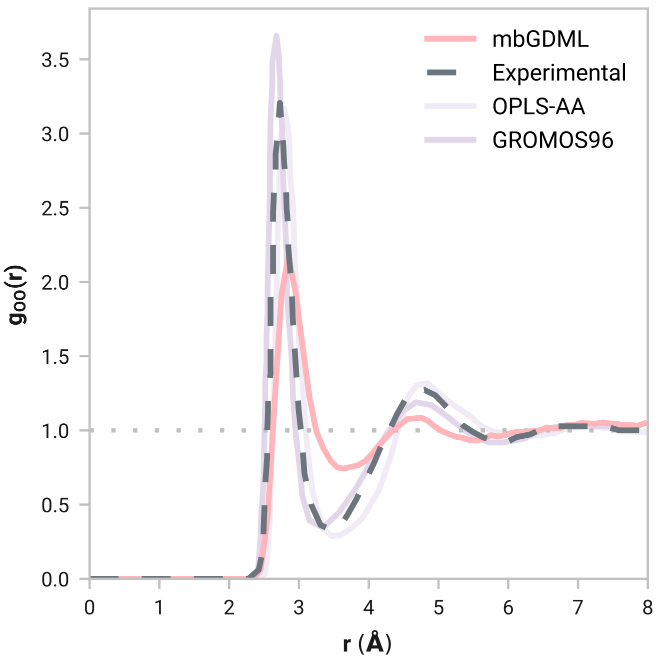


Figure S24: $g_{\mathrm{OO}}(r)$ RDF curve of methanol from NVT simulations at 298.15 K driven by mbGDML. Comparisons are made against experimental[31,32] and classical[40] results.

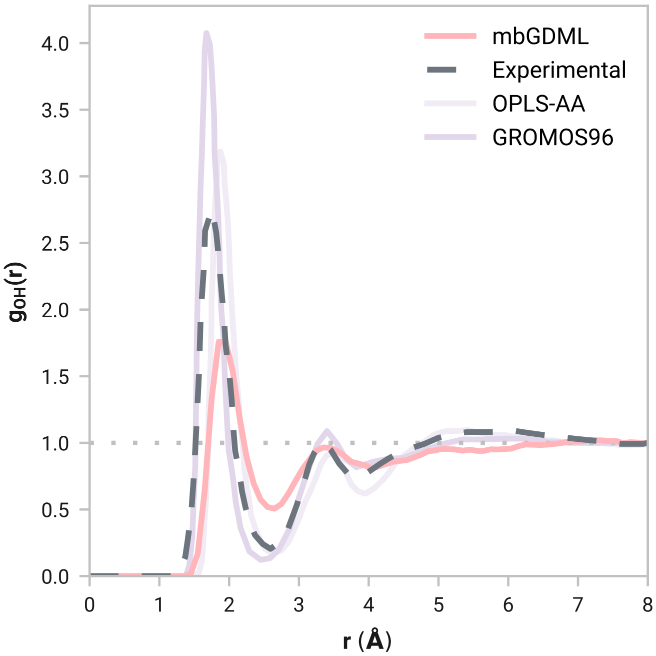


Figure S25: $g_{\mathrm{OH}}(r)$ RDF curve of methanol from NVT simulations at 298.15 K driven by mbGDML. Comparisons are made against experimental[31,32] and classical[40] results.

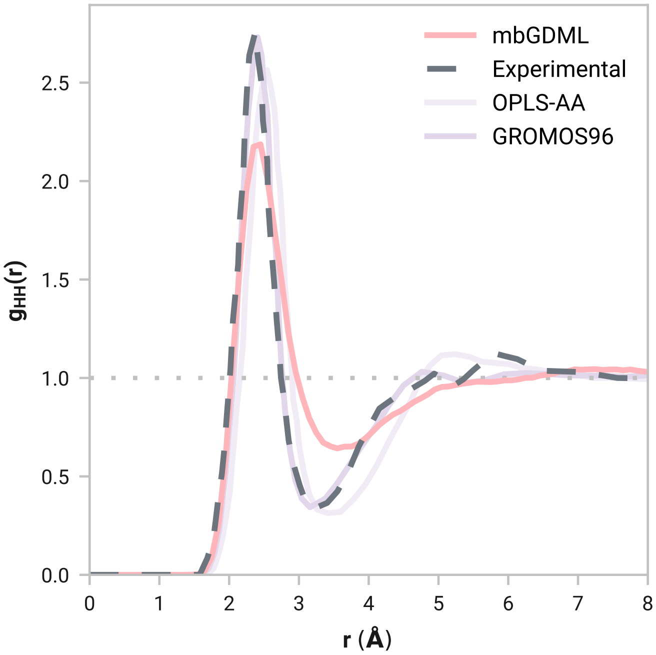


Figure S26: $g_{\mathrm{HH}}(r)$ RDF curve of methanol from NVT simulations at 298.15 K driven by mbGDML. Comparisons are made against experimental[31,32] and classical[40] results.

References

[1] F. Neese, "Software update: the ORCA program system, version 4.0", WIREs Comput. Mol. Sci. **8**, e1327 (2018).

[2] F. Neese, "The ORCA program system", WIREs Comput. Mol. Sci. **2**, 73–78 (2012).

[3] C. Møller and M. S. Plesset, "Note on an approximation treatment for many-electron systems", Phys. Rev. **46**, 618–622 (1934).

[4] F. Weigend and R. Ahlrichs, "Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy", Phys. Chem. Chem. Phys. **7**, 3297–3305 (2005).

[5] B. Temelso, K. A. Archer, and G. C. Shields, "Benchmark structures and binding energies of small water clusters with anharmonicity corrections", J. Phys. Chem. A **115**, 12034–12046 (2011).

[6] S. Yoo, E. Aprà, X. C. Zeng, and S. S. Xantheas, "High-level ab initio electronic structure calculations of water clusters $(H_2O)_{16}$ and $(H_2O)_{17}$: A new global minimum for $(H_2O)_{16}$", J. Phys. Chem. Lett. **1**, 3122–3127 (2010).

[7] A. Malloum, J. J. Fifen, and J. Conradie, "Binding energies and isomer distribution of neutral acetonitrile clusters", Int. J. Quantum Chem. **120**, e26221 (2020).

[8] K. Remya and C. H. Suresh, "Cooperativity and cluster growth patterns in acetonitrile: A DFT study", J. Comp. Chem. **35**, 910–922 (2014).

[9] S. L. Boyd and R. J. Boyd, “A density functional study of methanol clusters”, J. Chem. Theory Comput. **3**, 54–61 (2007).

[10] M. M. Pires and V. F. DeTuri, “Structural, energetic, and infrared spectra insights into methanol clusters $(CH_3OH)_n$, for n = 2–12, 16, 20. ONIOM as an efficient method of modeling large methanol clusters”, J. Chem. Theory and Comput. **3**, 1073–1082 (2007).

[11] F. Weigend, M. Häser, H. Patzelt, and R. Ahlrichs, “RI-MP2: optimized auxiliary basis sets and demonstration of efficiency”, Chem. Phys. Lett. **294**, 143–152 (1998).

[12] R. M. Richard, B. W. Bakr, and C. D. Sherrill, “Understanding the many-body basis set superposition error: Beyond Boys and Bernardi”, J. Chem. Theory Comput. **14**, 2386–2400 (2018).

[13] R. M. Richard, K. U. Lao, and J. M. Herbert, “Achieving the CCSD(T) basis-set limit in sizable molecular clusters: counterpoise corrections for the many-body expansion”, J. Phys. Chem. Lett. **4**, 2674–2680 (2013).

[14] J. F. Ouyang, M. W. Cvitkovic, and R. P. Bettens, “Trouble with the many-body expansion”, J. Chem. Theory Comput. **10**, 3699–3707 (2014).

[15] J. M. Herbert, “Fantasy versus reality in fragment-based quantum chemistry”, J. Chem. Phys. **151**, 170901 (2019).

[16] J. F. Ouyang and R. P. Bettens, “Many-body basis set superposition effect”, J. Chem. Theory Comput. **11**, 5132–5143 (2015).

[17] R. M. Richard, K. U. Lao, and J. M. Herbert, “Understanding the many-body expansion for large systems. I. Precision considerations”, J. Chem. Phys. **141**, 014108 (2014).

[18] K. U. Lao, K.-Y. Liu, R. M. Richard, and J. M. Herbert, “Understanding the many-body expansion for large systems. II. Accuracy considerations”, J. Chem. Phys. **144**, 164105 (2016).

[19] K.-Y. Liu and J. M. Herbert, “Understanding the many-body expansion for large systems. III. Critical role of four-body terms, counterpoise corrections, and cutoffs”, J. Chem. Phys. **147**, 161729 (2017).

[20] L. Martínez, R. Andrade, E. G. Birgin, and J. M. Martínez, “PACKMOL: A package for building initial configurations for molecular dynamics simulations”, J. Comp. Chem. **30**, 2157–2164 (2009).

[21] K.-Y. Liu and J. M. Herbert, “Energy-screened many-body expansion: A practical yet accurate fragmentation method for quantum chemistry”, J. Chem. Theory Comput. **16**, 475–487 (2019).

[22] E. E. Dahlke and D. G. Truhlar, “Electrostatically embedded many-body expansion for large systems, with applications to water clusters”, J. Chem. Theory Comput. **3**, 46–53 (2007).

[23] G. Fonseca, I. Poltavsky, V. Vassilev-Galindo, and A. Tkatchenko, “Improving molecular force fields across configurational space by combining supervised and unsupervised machine learning”, J. Chem. Phys. **154**, 124102 (2021).

[24] S. Chmiela, H. E. Sauceda, I. Poltavsky, K.-R. Müller, and A. Tkatchenko, “sGDML: Constructing accurate and data efficient molecular force fields using machine learning”, Comput. Phys. Commun. **240**, 38–45 (2019).

[25] F. Nogueira, *Bayesian Optimization: open source constrained global optimization tool for Python*, 2014.

[26] S. Batzner, A. Musaelian, L. Sun, M. Geiger, J. P. Mailoa, M. Kornbluth, N. Molinari, T. E. Smidt, and B. Kozinsky, “E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials”, Nat. Commun. **13**, 2453 (2022).

[27] L. McInnes, J. Healy, and J. Melville, “UMAP: uniform manifold approximation and projection for dimension reduction”, arXiv, preprint, arXiv:1802.03426 (2018).

[28] A. H. Larsen, J. J. Mortensen, J. Blomqvist, I. E. Castelli, R. Christensen, M. Dułak, J. Friis, M. N. Groves, B. Hammer, C. Hargus, et al., “The atomic simulation environment—a python library for working with atoms”, J. Phys.: Condens. Matter **29**, 273002 (2017).

[29] J. D. Chodera, “A simple method for automated equilibration detection in molecular simulations”, J. Chem. Theory Comput. **12**, 1799–1805 (2016).

[30] A. K. Soper, “The radial distribution functions of water as derived from radiation total scattering experiments: Is there anything we can say for sure?”, Int. Scholarly Res. Not. **2013**, 279463 (2013).

[31] T. Yamaguchi, K. Hidaka, and A. K. Soper, “The structure of liquid methanol revisited: a neutron diffraction experiment at -80 °c and +25 °c”, Mol. Phys. **96**, 1159–1168 (1999).

[32] T. Yamaguchi, K. Hidaka, and A. K. Soper, “Erratum: the structure of liquid methanol revisited: a neutron diffraction experiment at -80 °c and +25 °c”, Mol. Phys. **97**, 603–605 (1999).

[33] J. Hernández-Cobos, J. M. Martínez, R. R. Pappalardo, I. Ortega-Blake, and E. S. Marcos, “A general purpose acetonitrile interaction potential to describe its liquid, solid and gas phases”, J. Mol. Liq. **318**, 113975 (2020).

[34] M. W. Mahoney and W. L. Jorgensen, “A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions”, J. Chem. Phys. **112**, 8910–8922 (2000).

[35] E. K. Humphreys, P. K. Allan, R. J. Welbourn, T. G. Youngs, A. K. Soper, C. P. Grey, and S. M. Clarke, “A neutron diffraction study of the electrochemical double layer capacitor electrolyte tetrapropylammonium bromide in acetonitrile”, J. Phys. Chem. B **119**, 15320–15333 (2015).

[36] M. Albertí, A. Amat, F. De Angelis, and F. Pirani, “A model potential for acetonitrile: from small clusters to liquid”, J. Phys. Chem. B **117**, 7065–7076 (2013).

[37] V. A. Koverga, O. M. Korsun, O. N. Kalugin, B. A. Marekha, and A. Idrissi, “A new potential model for acetonitrile: Insight into the local structure organization”, J. Mol. Liq. **233**, 251–261 (2017).

[38] M. H. Kowsari and L. Tohidifar, “Systematic evaluation and refinement of existing all-atom force fields for the simulation of liquid acetonitrile”, J. Comput. Chem. **39**, 1843–1853 (2018).

[39] S. Pothoczki and L. Pusztai, “Intermolecular orientations in liquid acetonitrile: new insights based on diffraction measurements and all-atom simulations”, J. Mol. Liq. **225**, 160–166 (2017).

[40] K. Khasawneh, A. Obeidat, H. Abu-Ghazleh, R. Al-Salman, and M. Al-Ali, “Evaluation test of the most popular models of methanol using selected thermodynamic, dynamic and structural properties”, J. Mol. Liq. **296**, 111914 (2019).