nature computational science

Article

https://doi.org/10.1038/s43588-025-00790-0

Extending atomic decomposition and many-body representation with a chemistry-motivated approach to machine learning potentials

Received: 16 August 2024

Accepted: 13 March 2025

Published online: 14 April 2025

Check for updates

Qi Yu[1,2] ✉, Ruitao Ma[1], Chen Qu[3], Riccardo Conte[4], Apurba Nandi[5], Priyanka Pandey[6], Paul L. Houston[7], Dong H. Zhang[8] & Joel M. Bowman[6]

Most widely used machine learning potentials for condensed-phase applications rely on many-body permutationally invariant polynomial or atom-centered neural networks. However, these approaches face challenges in achieving chemical interpretability in atomistic energy decomposition and fully matching the computational efficiency of traditional force fields. Here we present a method that combines aspects of both approaches and balances accuracy and force-field-level speed. This method utilizes a monomer-centered representation, where the potential energy is decomposed into the sum of chemically meaningful monomeric energies. The structural descriptors of monomers are described by one-body and two-body effective interactions, enforced by appropriate sets of permutationally invariant polynomials as inputs to the feed-forward neural networks. Systematic assessments of models for gas-phase water trimer, liquid water, methane–water cluster and liquid carbon dioxide are performed. The improved accuracy, efficiency and flexibility of this method have promise for constructing accurate machine learning potentials and enabling large-scale quantum and classical simulations for complex molecular systems.

Computational simulations of molecular systems are essential for understanding complex processes in chemistry, biology and material sciences. A key challenge in both quantum and classical simulations is the extensive computations required for potential energy and force evaluations given molecular configurations. Direct ab initio calculations using accurate electronic-structure methods such as the 'gold standard' coupled cluster theory with single, double and perturbative triple excitations, CCSD(T)[1], are ideal. However, it quickly becomes prohibitive for systems with more than 15 atoms. Although density functional theory is widely used in ab initio molecular dynamics (MD) simulations due to its relative efficiency, its limited accuracy and still unfavorable computational scaling present challenges for long-time simulations of large and complex systems.

Over the past two decades, machine learning potentials (MLPs) have emerged as a promising approach to enable efficient and accurate computational simulations[2–24]. For high-dimensional systems with tens of thousands atoms, such as condensed-phase water, an atomistic representation of the potential is a popular choice[13]:

[1]Department of Chemistry, Fudan University, Shanghai, China. [2]Shanghai Innovation Institute, Shanghai, China. [3]Independent Researcher, Toronto, Ontario, Canada. [4]Dipartimento di Chimica, Università degli Studi di Milano, Milan, Italy. [5]Department of Physics and Materials Science, University of Luxembourg, Luxembourg City, Luxembourg. [6]Department of Chemistry and Cherry L. Emerson Center for Scientific Computation, Emory University, Atlanta, GA, USA. [7]Department of Chemistry and Chemical Biology, Cornell University, Ithaca, NY, USA. [8]State Key Laboratory of Molecular Reaction Dynamics, Dalian Institute of Chemical Physics, Chinese Academy of Sciences, Dalian, China. ✉e-mail: qi_yu@fudan.edu.cn

$$E_{\text{total}} = \sum_{i}^{N_{\text{atom}}} E_{i,\text{atomic}}, \quad (1)$$

where the total potential energy of the system E_{total} is decomposed as the sum of atomic local energies $E_{i,\text{atomic}}$ over all N_{atom} atoms, where i represents the index of each atom in the system. This representation has been widely applied in various MLPs. Typical examples include BPNN[13], SchNet[18], PhysNet[19], DeePMD[20] and EANN[21] and so on. Recent equivariant neural network (NN) potentials such as NequIP[22], MACE[23] and Allegro[24] also employ the atomistic representation.

Analogous to the difference between atomic and molecular orbital energies in electronic-structure theory, the concept of atomic local energy in these MLPs lacks effective chemical meaning[25,26], as the energy of the entire molecule, rather than that of individual atoms, is more relevant for capturing molecular structural signatures and environmental perturbations. Moreover, the physically undefined nature of atomic local energies may result in arbitrary assignments by modern NNs and compromises their transferability to different systems[26]. It is worth noting that machine learning force-field approaches, such as FFLUX[27–29], have been developed to obtain physically well-defined atomic properties using the quantum chemical topology framework[30,31] and Gaussian process regression. Another aspect of the atomistic representation of potential energy is related to computational scaling, where the cost scales linearly with the total number of atoms in the system. It remains an open question whether this scaling can be further improved to achieve greater efficiency while maintaining the same or higher level of accuracy.

Another approach to obtain MLPs for large molecular systems is the many-body representation, which has been widely reported in the literature since the 1980s[32–39]. Taking water potentials as examples, the most accurate ones, namely, MB-pol[40], q-AQUA[41] and q-AQUA-pol[42], use a many-body expansion for the total energy of N water monomers:

$$E_{\text{total}} = \sum_{i=1}^{N} E_{1-b}(i) + \sum_{i>j}^{N} E_{2-b}(i,j) + \sum_{i>j>k}^{N} E_{3-b}(i,j,k) + \sum_{i>j>k>l}^{N} E_{4-b}(i,j,k,l) + \cdots, \quad (2)$$

where i, j, k and l are indices of water monomers and each energetic term is obtained from training an MLP on the appropriate dataset. Specifically, the one-body (1–b) term E_{1-b} represents the potential for the isolated water monomer, often modeled using the spectroscopically accurate ab initio-based Partridge–Schwenke potential[43]. The two-body term E_{2-b} is an MLP fit to dimer interaction electronic energies, the three-body term E_{3-b} is an MLP fit to trimer interaction energies and the four-body term E_{4-b} is an MLP fit to tetramer interaction energies. This many-body formulation allows the use of permutationally invariant polynomial (PIP)-based methods, such as PIP[8], PIP-NN[9] and fundamental invariant (FI)-NN[11], to accurately describe the n-body interactions involving n molecules.

Despite recent successes in simulating water properties from the gas phase to the liquid phase using many-body MLPs[35,40–42,44], it is well-known that many-body representation suffers from the rapidly increasing number of three-body, four-body and higher-order terms. Consequently, long-time simulations of relevant molecular systems are often prohibitive.

In this work, we introduce a machine learning framework that combines the strengths of both methods while mitigating their weaknesses. This monomeric framework leverages a representation of the system's potential energy in terms of monomer energies instead of atomic local energies and employs molecular energies only at the one-body and two-body levels. Permutational invariance is enforced by using PIPs as inputs to NNs, describing a molecule's structure and environment. This critical aspect of our work echoes the use of PIPs[9,10] and later efficient FIs[11,45] as inputs to NNs in applications to gas-phase molecules. We term this approach MB-PIPNet. We demonstrate that this method achieves high accuracy across a variety of molecular systems, ranging from gas-phase clusters (for instance, water trimer and methane–water cluster) to condensed-phase systems (for instance, liquid water and carbon dioxide (CO_2)).

Our findings indicate that many-body interactions, such as three-body interactions, can be accurately described using only one-body and two-body PIP bases in the NN descriptor. This discovery substantially enhances the efficiency of our framework, enabling fast computational simulations of complex condensed-phase systems at the cost of conventional force fields. Our framework shows good performance in MD simulations of liquid water and achieves substantially better computational scaling compared with other atomistic machine learning models. Furthermore, our framework can be systematically extended to various types of molecular system. Related challenges and possible solutions are also discussed.

Results

Monomeric NN model

The proposed monomeric NN potential model, MB-PIPNet, is illustrated in Fig. 1. This framework relies on appropriate decomposition of the molecular system into N monomers. The total potential energy is then represented as the sum of the perturbed energy of each monomer, E_i, analogous to the atom-centered approach, such that

$$E_{\text{total}} = \sum_{i}^{N} E_i \quad (3)$$

The energy of each perturbed monomer is trained using a feed-forward NN model that utilizes specifically designed structural descriptors as the input layer. In detail, after decomposing the entire molecular system into N fragmental monomers, the Cartesian coordinates of monomer i are transformed into a self-structural descriptor, $\mathbf{G}_i(\text{self})$. We employ the widely used PIPs to construct this self-structural descriptor, which naturally ensures the necessary invariance to translation, rotation and permutation.

The self-structural descriptor, $\mathbf{G}_i(\text{self})$, effectively describes the energetic response of a monomer to changes in its own configuration. However, each monomer is subjected to a complex environment with extensive intermolecular interactions involving other molecules. Consequently, the potential energy of each monomer should be polarizable. To account for this, we use two-body PIPs as the environment descriptor for each monomer, $\mathbf{G}_i(\text{env})$. As detailed in the section on structural descriptors in Methods, this environment descriptor accounts for interactions with all surrounding monomers within a defined distance cut-off. The combination of $\mathbf{G}_i(\text{self})$ and $\mathbf{G}_i(\text{env})$ offers a systematic approach to describe the molecular response within a complex system. As will be explored next, using one-body and two-body PIPs as core components in these descriptors results in substantially more efficient computation compared with other machine learning methods.

Energetic properties of MB-PIPNet model for water trimer

The use of only one-body and two-body PIP bases as structural descriptors in our MB-PIPNet framework raises the question of whether MB-PIPNet can describe many-body interactions beyond two body. To address this, we first demonstrate MB-PIPNet's capability of capturing high-order interaction using the case of gas-phase water trimer. We trained an MB-PIPNet potential based on 45,812 trimer structures with energies calculated at the CCSD(T)-level q-AQUA-pol potential. This water trimer dataset spans a wide energy range of [0, 15] eV, and our final training root mean square error (RMSE) is only 1.08 meV per atom. As also shown in Extended Data Fig. 1a, the corresponding MB-PIPNet

https://doi.org/10.1038/s43588-025-00790-0

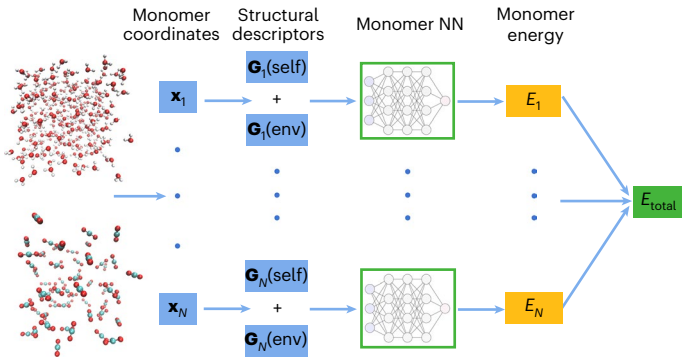


Fig. 1 | Schematic of the MB-PIPNet architecture. The coordinates of each fragmental monomer, $\mathbf{X}_i$, are first transferred to its self-structural descriptors $\mathbf{G}_i$(self), such as one-body PIP bases. The structural descriptors of each monomer's environment, $\mathbf{G}_i$(env), are generated by pairwise monomer coordinates involving different monomers, forming two-body PIP bases. The self- and environmental descriptors of each monomer are combined as the input of the NN and yield the effective monomeric potential energy, E_i. The final energy of the complicated molecular system, E_{total}, is the sum over monomeric energies of all fragmental monomers. Atom colors: H, white; C, cyan; O, red.

model for water trimer shows high accuracy in predicting potential energies of a separate test dataset, with an RMSE of 1.07 meV per atom. We also examined the accuracy of the MB-PIPNet potential in a representative potential energy cut. As shown in Extended Data Fig. 1b, good agreement with the q-AQUA-pol reference data is achieved. These single-point energy results provide direct evidence that using only one-body and two-body PIP bases in constructing structural descriptors, the trained MB-PIPNet potential is capable of handling the complex molecular environments beyond simple two-body interactions.

The accuracy of the trained MB-PIPNet potential for the water trimer is further verified through harmonic normal mode analysis and anharmonic diffusion Monte Carlo (DMC) calculations. As seen in Table 1, the MB-PIPNet potential provides accurate harmonic frequencies for the water trimer's global minimum structure, with deviations mostly smaller than 5 cm^{-1} compared with both q-AQUA-pol and CCSD(T)-F12a/aug-cc-pVTZ benchmark results. As a more stringent test of the accuracy and smoothness of the potential energy surface (PES), unconstrained DMC calculations were performed to determine the anharmonic zero-point energies (ZPEs) of the water trimer using the trained MB-PIPNet potential. Notably, the rigorous quantum DMC calculations provide the exact ZPE of the molecule and also serve as an effective tool for detecting 'holes' in the analytical PES. The trained MB-PIPNet potential was found to be 'hole' free and the calculated water trimer's ZPE is 15,593 ± 4 cm^{-1}, which agrees well with previous results using CCSD(T)-level PESs such as q-AQUA-pol and WHBB.

Liquid water with MB-PIPNet potential

Beyond gas-phase molecular clusters, it is crucial to assess the performance of the MB-PIPNet approach on condensed-phase systems, where each molecule is subject to a substantially more complex environment. To this end, we trained the MB-PIPNet model on a dataset consisting of 1,593 liquid water configurations, calculated at the revPBE0-D3 level of theory[46]. The RMSE on energy and force for the MB-PIPNet model, evaluated on the test set, are shown in Table 2. Compared with other MLPs, the MB-PIPNet model generally outperforms invariant atomistic MLPs, including BPNN and EANN. More sophisticated invariant and equivariant message-passing NN potentials, such as REANN[47], NequIP[22] and MACE[23], show better performance, particularly for force predictions. These MPNN models typically involve tens of thousands of parameters, suggesting that the RMSE of the MB-PIPNet model could potentially be further reduced with a more complicated NN structure and the incorporation of a message-passing mechanism.

Table 1 | Performance of MB-PIPNet model on vibrational analysis of water trimer

Harmonic frequency			
Method	**MB-PIPNet**	**q-AQUA-pol**	**Ab initio[a]**
Mode 1	155.4	160.4	154.5
Mode 2	174.7	176.5	178.6
Mode 3	188.9	188.3	185.7
Mode 4	195.7	194.3	191.7
Mode 5	219.7	220.1	220.2
Mode 6	229.1	237.5	228.3
Mode 7	329.2	337.0	332.4
Mode 8	351.9	350.8	346.4
Mode 9	438.4	437.3	437.1
Mode 10	555.4	562.1	558.8
Mode 11	648.9	653.8	656.8
Mode 12	829.6	844.6	846.5
Mode 13	1,648.3	1,662.2	1,654.9
Mode 14	1,655.3	1,665.3	1,660.2
Mode 15	1,674.3	1,684.0	1,678.9
Mode 16	3,618.9	3,621.1	3,621.0
Mode 17	3,675.0	3,681.2	3,677.6
Mode 18	3,684.7	3,689.3	3,685.5
Mode 19	3,903.8	3,907.2	3,903.3
Mode 20	3,908.2	3,910.2	3,908.3
Mode 21	3,909.9	3,914.3	3,908.8
Harmonic ZPE			
Method	**MB-PIPNet**	**q-AQUA-pol**	**Ab initio[a]**
	15,997.7	16,048.8	16,017.7
Anharmonic DMC ZPE			
Method	**MB-PIPNet**	**q-AQUA-pol[b]**	**WHBB[c]**
	15,593 ± 4	15,616 ± 2	15,587 ± 2

Harmonic frequencies and anharmonic DMC ZPE (in cm^{-1}) of water trimer from different methods. [a]CCSD(T)-F12a/aug-cc-pVTZ. [b]From ref. 42. [c]From ref. 67.

To obtain an MLP model for liquid water with higher accuracy than density functional theory, we trained another MB-PIPNet model of water using using reference data from ref. 48. The training set consists of 75,874 different configurations from MD simulations of liquid water at various temperatures, with total energy of each configuration calculated using the MB-pol force field[49]. The training process over this extensive dataset converged quickly, as shown in Supplementary Fig. 1. The final training RMSE is 0.29 meV per atom, which is notably smaller than the 0.39–0.44 meV per atom achieved using the DeePMD approach with the same dataset[48]. Figure 2a shows the performance of the MB-PIPNet model on energy predictions for a separate test dataset, showing good correlations with a small RMSE of 0.30 meV per atom. These energetic results verify the capability of the MB-PIPNet approach in handling complex and polarizable condensed-phase systems, such as liquid water, where many-body interactions play crucial roles in determining the corresponding physical and chemical properties.

Conventional Behler–Parrinello-type atomistic MLPs predict the atomic local energies of molecular systems. However, from a chemistry perspective, the energy of individual molecule is often of greater interest. A natural advantage of the MB-PIPNet model is its ability to directly predict the perturbed monomer energies of all individual molecules. This is analogous to the widely used concept of molecular

Table 2 | Comparison of model accuracies on the liquid-water dataset

	BPNN[46]	EANN[68]	REANN[47]	NequIP[22]	MACE[23,69]	MB-PIPNet
Energy	2.33	2.1	0.8	0.93	0.63	1.19
Force	120	129	47	45	36.2	93.3

RMSE for energy (meV per atom) and force (meV Å^{-1}) from different MLPs trained on the same dataset from ref. 46.

orbital energy against atomic orbital energy in electronic-structure theory. Figure 2b shows a scatter plot of the monomer energies of 256 water molecules from a representative liquid-water configuration using different methods. Given the coordinates of all water molecules, the Partridge–Schwenke[43] energy of each molecule is calculated using corresponding spectroscopically accurate water monomer potential. The q-AQUA energies are calculated through many-body expansion using our recently developed purely many-body PES for water, where the energy of each water molecule is calculated as

$$E_i(\text{q-AQUA}) = E_{1-b}(i) + \sum_{j}^{N} \frac{1}{2} E_{2-b}(i,j) + \sum_{j>k}^{N} \frac{1}{3} E_{3-b}(i,j,k) + \sum_{j>k>l}^{N} \frac{1}{4} E_{4-b}(i,j,k,l) \quad (4)$$

As seen, the Partridge–Schwenke one-body energies of 256 water molecules are in the range [0, 5] kcal mol^{-1}, indicating the distorted structures of these molecules in the liquid phase relative to the global minimum structure. When interactions among molecules are included, the water molecules are polarized and the corresponding q-AQUA monomer energies are in the range [−20, −3] kcal mol^{-1}. In line with the q-AQUA results, the MB-PIPNet model provides reasonable predictions of monomer energies, with different water molecules showing distinct perturbed energies due to their structural distortion and interactions with other molecules in the liquid phase. These observations provide additional evidence that the MB-PIPNet model reasonably describes the many-body interactions in complex molecular systems with structural descriptors constructed from only one-body and two-body PIP bases.

The trained MB-PIPNet model of liquid water was further employed in MD simulations for bulk water properties using the i-PI 2.0 software[50]. Figure 2c shows the oxygen–oxygen (OO) radial distribution function (RDF) obtained from classical MD simulations at 298 K, with the oxygen–hydrogen (OH) and hydrogen–hydrogen (HH) RDFs provided in Supplementary Fig. 2. As seen, the OO RDF obtained from the MB-PIPNet model agrees well with experimental data in terms of both peak positions and amplitudes. Extended Data Fig. 2 shows additional OO RDFs results generated by the MB-PIPNet and MB-pol across a range of temperatures. Both models consistently show agreement with experimental data, highlighting the capability of the MB-PIPNet model to accurately replicate the MB-pol force field in simulating the structural properties of water across different thermal conditions.

The oxygen–oxygen–oxygen (OOO) triplet angular distribution $P_{\mathrm{OOO}}(\theta)$ is another static property used to detect the tetrahedral orientational ordering of liquid water. We obtain $P_{\mathrm{OOO}}(\theta)$ by computing the angle formed by an oxygen atom of a water molecule and two of its oxygen neighbors, with the neighbors defined using a cut-off of 3.27 Å to yield an average OO coordination number of 4 (ref. 42). As shown in Fig. 2d, the distribution of $P_{\mathrm{OOO}}(\theta)$ from the MB-PIPNet model is in excellent agreement with experiment in terms of peak position, width and intensity.

Finally, the dynamic property of liquid water, specifically the self-diffusion coefficient D as a function of temperature, is investigated using MB-PIPNet model-based MD simulations. As seen in Table 3, the predicted D from the MB-PIPNet model agrees well with the experimental measurements across different temperatures. Similar performance is observed in previous MD simulations performed directly using MB-pol[51], although there are slight differences when compared with the MB-PIPNet results. It should be noted that the current self-diffusion coefficients were calculated using a simulation box of 256 water molecules. An increase in D is anticipated for an 'infinitely' large box, using the correction formula widely used in the literature and also our previous work[42,51].

Computational scaling with force-field-level cost

Thus far, we have demonstrated that the MB-PIPNet approach can accurately describe many-body interactions from gas-phase clusters to condensed-phase systems. The chemistry-motivated architecture of the MB-PIPNet method naturally provides detailed monomeric energies rather than conventional atomistic energies. Another appealing feature of the MB-PIPNet method is its favorable scaling and computational cost. This stems from two main aspects. The first one is associated with the use of PIPs as key components in structural descriptors. The generation of PIPs have been extensively verified to be systematic and efficient compared with other complicated ML descriptors[52]. Second, our MB-PIPNet framework employs a representation of the total energy as the sum of monomer energies. Consequently, the computational cost of the MB-PIPNet potential scales linearly with the number of molecules rather than the number of atoms, as is the case with most MLPs.

In Fig. 3, we compare the computational cost of the MB-PIPNet models with the q-TIP4P/F and TTM3-F force fields[53,54], as well as the DeePMD and REANN MLPs[47,48], for a single MD step of energy and gradient calculations across different sizes of liquid-water simulation boxes. In the single central processing unit (CPU) core simulations, we first verify the linear computational scaling of the MB-PIPNet approaches with respect to the number of water molecules. Moreover, with a larger two-body cut-off (R_c = 9 Å), the MB-PIPNet model shows computational efficiency comparable to the conventional polarizable water force field, TTM3-F, and is several times faster than the DeePMD potential. For simulations involving thousands of water molecules, the MB-PIPNet model with R_c = 9 Å substantially outperforms TTM3-F in speed, as it avoids the computationally expensive electrostatic Ewald summation. Similarly, the MB-PIPNet model with a shorter R_c = 6 Å, shows even faster performance, surpassing TTM3-F and approaching the speed of the non-polarizable q-TIP4P/F force field. As the system size increases, the computational cost of MB-PIPNet becomes almost identical to that of q-TIP4P/F. Furthermore, with a relatively short atomic environmental cut-off (R_c = 3 Å), the message-passing NN REANN exhibits computational performance comparable to the MB-PIPNet model with R_c = 9 Å. As the cut-off increases to R_c = 5.5 Å, the computational cost of REANN rises substantially. Notably, as demonstrated in ref. 55, sophisticated equivariant message-passing NNs such as MACE, when utilizing an Nvidia A100 graphics processing unit (GPU) with thousands of inner cores, can achieve computational speeds comparable to those of the MB-PIPNet model with R_c = 9 Å in a simulation box containing 512 water molecules. Remarkably, the MB-PIPNet model achieved this level of performance using only a single CPU core. Once extensive parallelization or GPU acceleration is applied, the MB-PIPNet approach is expected to achieve orders of magnitude improvements in computational efficiency than these MLPs. The computational performance of message-passing MLPs, such as REANN and MACE, can be substantially enhanced by choosing a smaller environmental cut-off R_c while maintaining reasonable model accuracy. This strategy for balancing accuracy and computational efficiency is also applicable to the MB-PIPNet model.

Performance on other molecular systems

The above two MB-PIPNet potentials are associated with gas-phase water trimer and liquid water. Here we further demonstrate the

https://doi.org/10.1038/s43588-025-00790-0

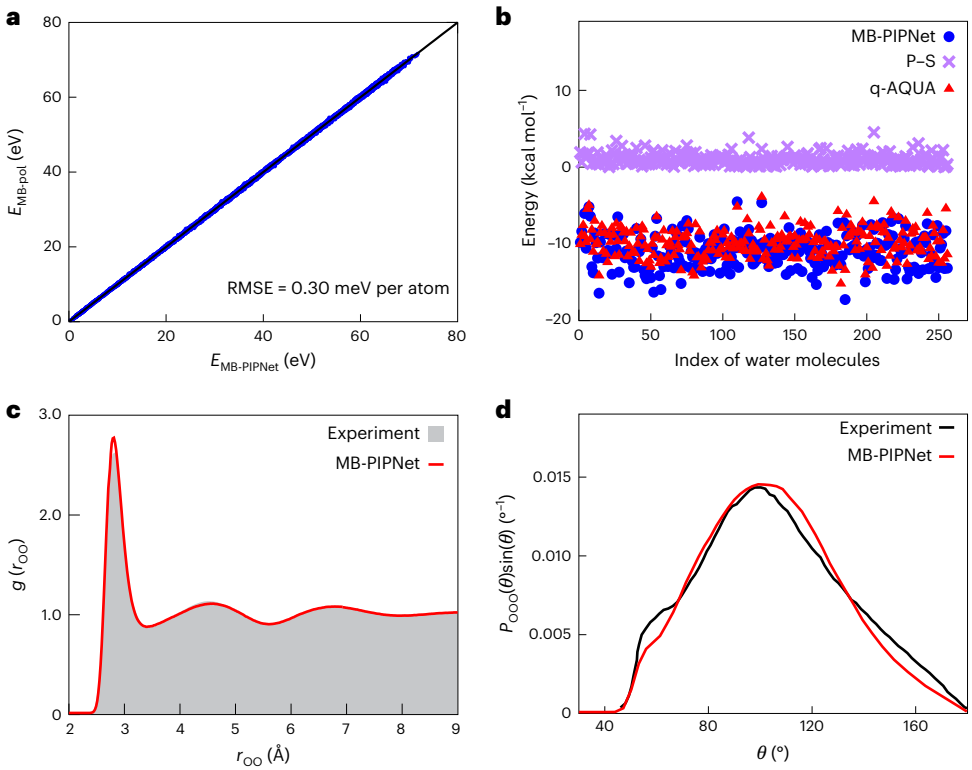


Fig. 2 | Potential energy predictions and MD simulation results of liquid water by MB-PIPNet model. a, Energy–energy correlation plot for MB-PIPNet model of liquid water with reference energies calculated using MB-pol. **b**, Scatter plot of monomer energies of 256 water molecules in a periodic cubic box predicted by the MB-PIPNet model, the Partridge–Schwenke (P–S) water monomer potential and the q-AQUA model. **c**, Oxygen–oxygen RDF, $g(r_{OO})$, as a function of O–O distance, r_{OO}, for liquid water at 298 K from classical MD simulations using MB-PIPNet model. The experimental data are taken from refs. 72,73. **d**, The oxygen–oxygen–oxygen triplet angular distribution functions of liquid water at 298 K predicted by MB-PIPNet model. The experimental data are taken from ref. 74. The triplet angular distribution functions shown here were normalized to $\int_0^{\pi} P_{OOO}(\theta)\sin(\theta)\mathrm{d}\theta$.

Table 3 | Dynamic property (self-diffusion coefficient) of liquid water by MB-PIPNet model

Temperature (K)	MB-PIPNet	MB-pol[a]	Experiment[b]
278	0.136±0.005	0.140	0.131
288	0.177±0.008	0.194	0.177
298	0.251±0.020	0.234	0.230
320	0.372±0.014	0.344	0.360

Self-diffusion coefficient D (Å^{2}ps^{-1}) of liquid water at different temperatures. [a]From ref. 51. [b]From refs. 70,71.

transferability of the MB-PIPNet approach to different molecular systems. Extended Data Fig. 3a illustrates the application of the MB-PIPNet model to methane–water clusters, CH_4 $(H_2O)_2$. The datasets were sourced from our previously developed many-body PES[56]. Utilizing the descriptors and network structures detailed in the section on training details in Methods, the RMSE for training and testing were determined to be 1.92 meV per atom and 1.90 meV per atom, respectively, which are reasonably small for gas-phase systems with energies up to 15 eV. As demonstrated in Supplementary Section 1, the computational cost for such a mixture system still scales linearly with the number of molecules, although distinct structural descriptors are used for different molecular types. Extended Data Fig. 3b presents another example of the MB-PIPNet model applied to liquid CO_2. Using only 2,687 BLYP-D3-level training data of configurations of 64 CO_2 molecules in a simulation box[57], the model achieved training and test RMSE values of 0.16 meV per atom and 0.26 meV per atom, respectively. This highlights the applicability of the MB-PIPNet method to various condensed-phase molecular systems. It is important to note that further assessments of model training and MD-based property calculations are necessary. In future work, we plan to generate additional training data and undertake more systematic MB-PIPNet potential training for a wide range of homogeneous and inhomogeneous systems.

The MB-PIPNet framework is applicable to a wide variety of molecules and molecular interactions. This versatility stems from our PIP library, which has been instrumental in developing more than 100 MLPs for different molecules[58]. Analogous to the behavior of FI-NN compared with PIP-NN, the FIs can also be employed as they are considered the minimal set of PIPs. The generation of FIs is systematic and has been frequently used in constructing high-dimensional MLPs. As noted in the recent work[12,59], an extension of FIs is anticipated for large systems with more than 20 atoms, or even 30 atoms.

Discussion

The MB-PIPNet framework discards the widely used atomic local energy decomposition method and represents the total potential energy as the sum of chemically meaningful monomer energies. Unlike the standard many-body expansion approach, in MB-PIPNet, the combination of one-body and two-body PIP-based descriptors provides a consistent way to incorporate the perturbation effect that results from the many-body interaction with other monomers. This treatment avoids the expensive computations of high-order terms in many-body representation and greatly improves the computational cost of MLP evaluation. These features of the MB-PIPNet approach open additional possibilities for performing computational simulations of complex systems such as molecular materials with first-principle accuracy but at the same speed as conventional force fields. To the best of our knowledge, such a balance between accuracy and efficiency has not yet been realized, even with an atomic MLP such as DeePMD.

https://doi.org/10.1038/s43588-025-00790-0

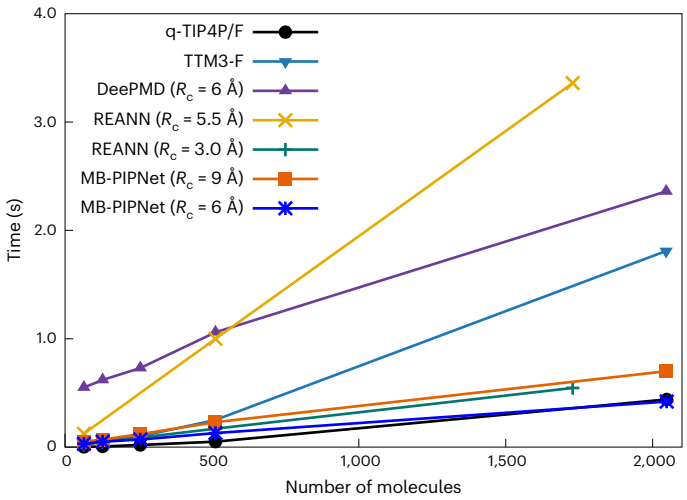


Fig. 3 | Computational cost of MB-PIPNet model. Computational time of single MD step (energy and force) versus number of water molecules in a periodic simulation box using different methods. All timing tests were performed using a single CPU core of the AMD EPYC 7002 processors. R_c is the atomic or monomeric environment cut-off applied in different MLPs.

When training MB-PIPNet potentials for condensed-phase systems such as liquid water, a distance cut-off is incorporated for the two-body PIPs of environmental descriptor **G**(env). Similar to other ML methods such as DeePMD, the current MB-PIPNet method lacks an explicit and robust description of the long-range effects. One possible direction is to incorporate a message-passing mechanism for generating structural descriptors[21,60,61]. Another remark on the limitations of MB-PIPNet framework pertains to the need for reasonable assignments of fragmented monomers in complex molecular systems. It can be naturally applied to systems such as gas-phase clusters and molecular liquids. However, additional efforts and further developments are required to extend the MB-PIPNet approach to reactive systems, large organic molecules, biomolecules or solid oxides.

For large organic molecules or biomolecules, the implementation of MB-PIPNet can be facilitated by: (1) improving the relevant PIP or FI theory to enable efficient computation of polynomials for systems with more than 20 atoms and (2) developing fragmentation theories to automatically divide large molecules into smaller, computationally manageable fragments[62]. For condensed-phase reactive systems or materials, such as solid oxides, atomistic machine learning methods are often more natural choices, and it is challenging to directly apply the MB-PIPNet approach. A potential direction could involve integrating the MB-PIPNet framework with atomistic machine learning approaches. For example, in the reaction of CO_2 with water to produce carbonic acid, the central reactive region could be modeled by an atomistic MLP, while the MB-PIPNet model could efficiently describe the non-reactive solvent water molecules, thereby accelerating the simulations. Such a hybrid framework involving the MB-PIPNet approach could also be applied to material–molecule systems, offering broad applications in catalyst design and material discovery.

The monomer-centered concept of our MB-PIPNet can be further applied in other machine learning approaches including invariant and equivariant NNs[19,22,23]. We hope that the proposed method will stimulate further development of MLPs in the wide fields of computational chemistry, physic, materials science and biology for classical and quantum simulations of complex systems with ab initio-level accuracy and conventional force-field cost.

Methods

Structural descriptors in MB-PIPNet approach

The PIPs are employed to generate the self-structural descriptor, $\mathbf{G}_i$(self), based on the Cartesian coordinates of monomer. For instance, for a tetraatomic monomer, a set of symmetrized polynomials at order n can be generated:

$$\mathbf{G}_i(\text{self}) = \mathbf{P}(\mathbf{X}_i) = \hat{\mathbf{S}}\left[y_{12}^a y_{13}^b y_{14}^c y_{23}^d y_{24}^e y_{34}^f\right] \quad (5)$$

where $\hat{\mathbf{S}}$ is the symmetrization operator that produces the appropriate sum of monomials with $a+b+c+d+e+f=n$. Each y_{ij} is the Morse-like variable in terms of internuclear distance r_{ij} between atom i and j, such that $y_{ij} = \exp(-r_{ij}/a_0)$ with a_0 as the hyperparameter.

The environmental descriptor $\mathbf{G}_i$(env) also employs the PIPs and is defined as:

$$\mathbf{G}_i(\text{env}) = \sum_{j=1}^{N_{\text{mol}}\in R_c} \mathbf{P}(\mathbf{X}_i, \mathbf{X}_j) f_c(\mathbf{X}_i, \mathbf{X}_j, R_c) \quad (6)$$

where N_{mol} represents the total number of surrounding monomers within a distance cut-off of R_c. $\mathbf{P}(\mathbf{X}_i, \mathbf{X}_j)$ is the corresponding two-body PIPs generated from Cartesian coordinates of monomer i and j. $f_c(\mathbf{X}_i, \mathbf{X}_j, R_c)$ is a switching function that ensures a smooth transition to zero for the two-body polynomials when the distance between two monomers exceeds R_c. Taking water as example, the two-body PIPs, $\mathbf{P}(\mathbf{X}_i, \mathbf{X}_j)$, are generated with 42 symmetry for the $H_2O\cdots H_2O$ pair. This includes permutational invariance for all four H atoms and both O atoms. These PIP bases are further purified to ensure that they approach zero asymptotically as the distance between the O atoms increases[41].

Reference datasets

For the water trimer, a total of 51,006 configurations were generated with energies calculated using the q-AQUA-pol potential. Among these, 45,332 configurations are trimer structures used in our previous three-body PES development in q-AQUA[41] and q-AQUA-pol[42] PESs. The remaining 5,674 structures were added by running DMC simulations using the initially trained MB-PIPNet models. The final dataset was randomly divided into a training dataset of 45,812 structures and a test dataset of 5,194 structures.

For liquid water, the first dataset is from ref. 46, which includes 1,593 liquid water configurations with each structure containing 64 water molecules. The energies were calculated at the revPBE0-D3 level of density functional theory. For the second dataset, we employed the one from ref. 48 where a final training dataset of 75,874 configurations and test dataset of 9,448 configurations were generated from MD simulations at various temperatures and pressure of 1 atm for a cubic box containing 256 molecules under periodic boundary conditions. The potential energy of each configuration was calculated from the MB-pol force field[49].

A total of 21,570 structures of $CH_4(H_2O)_2$ were obtained from our previous work[56] with energies calculated using developed many-body potential. Ninety percent of the dataset was used for training, resulting in 19,342 training and 2,228 testing configurations. The dataset for liquid CO_2 was directly obtained from ref. 57 where the configurations were obtained from MD simulations of bulk liquid states at temperature T = 220–300 K and pressure P = 100 bar for a system of 64 CO_2 molecules under periodic boundary condition. A total of 3,800 configurations at the BLYP-D3 level of theory were used, with 2,687 for training and 313 for testing.

Training details

For MB-PIPNet models of water and liquid CO_2 in this study, we employed sixth-order full-symmetry PIPs for the self-structural descriptor, **G**(self) and fourth-order full-symmetry two-body PIP bases for the environmental descriptor, **G**(env). Specifically, for both water trimer and liquid water, the input layer of the NN has a dimension of 188, with 49 for **G**(self) and 139 for **G**(env). For liquid CO_2, the input layer is 80, with 30 for **G**(self) and 50 for **G**(env). For $CH_4(H_2O)_2$, 80

PIPs with 41 symmetry and 30 PIPs with 21 symmetry are employed to describe $\mathbf{G}_i$(self) of CH_4 and H_2O, respectively. For the environmental descriptors, $\mathbf{G}_{CH_4}(\text{env})$ and $\mathbf{G}_{H_2O}(\text{env})$, 100 methane–water PIPs with 4,211 symmetry and 50 water–water PIPs with 42 symmetry are used. With this set-up, the input layers of the NN for CH_4 and H_2O are both 180. All the MB-PIPNet models utilize NNs with two hidden layers. The neuron structures for these layers are [30, 60] for the water trimer, [15, 30] for liquid water, [10, 20] for $CH_4(H_2O)_2$ and [10, 30] for liquid CO_2. The training of all MB-PIPNet models was realized using the Levenberg–Marquardt algorithm[63]. The training stopped when the learning rate dropped below 10^{-5}. The cut-off distances for the 3 systems were set as 9.0 Å.

DMC simulations of water trimer

We employed DMC calculations to rigorously determine the ground vibrational state wave function and the full-dimensional anharmonic ZPE. By linking the imaginary-time Schrödinger equation to the diffusion equation, we express the wave function evolution as

$$\frac{\partial\psi(\mathbf{x},\tau)}{\partial\tau}=\sum_{i=1}^{N}\frac{\hbar^2}{2m_i}\nabla_i^2\psi(\mathbf{x},\tau)-[V(\mathbf{x})-E_{\text{ref}}]\,\psi(\mathbf{x},\tau) \tag{7}$$

where $\psi(\mathbf{x},\tau)$ is the wave function, m_i is the mass of atom i and $V(x)$ is the potential energy of the molecule at configuration $\mathbf{x}$. The reference energy, E_{ref}, serves as the estimator of the ZPE[64], stabilizing the system to its ground state.

The DMC method utilizes $N(0)$ equally weighted Gaussian random walkers to sample the initial wave function and propagates equation (7) in imaginary time with a Gaussian distribution. During the diffusion process, the number of walkers is dynamically adjusted through

$$E_{\text{ref}}(\tau)=\langle V(\tau)\rangle-\alpha\frac{N(\tau)-N(0)}{N(0)} \tag{8}$$

where $E_{\text{ref}}(\tau)$, $N(\tau)$ and $\langle V(\tau)\rangle$ are the reference energy, the number of walkers and the averaged potential energy of all walkers at the time step τ, respectively.

For the water trimer system, DMC calculations were conducted with an imaginary time step $\Delta\tau = 5$ a.u. and $\alpha = 0.25$. Five independent DMC simulations were performed, each employing 40,000 random walkers. The diffusion equation was propagated for 40,000 steps, with the initial 2,000 steps allocated for equilibration.

MD simulations of liquid water

All classical MD simulations were conducted using i-PI 2.0 software[50], interfaced with the MB-PIPNet water potential. The simulation system includes 256 water molecules with the box dimensions adjusted to reproduce experimental liquid water densities at the corresponding temperatures. Simulations were performed in the canonical ensemble (NVT) with a Langevin thermostat. At each temperature, 3 independent trajectories were propagated for 1 ns with a time step of 0.5 fs.

The diffusion coefficient D is related to the slope of mean square displacement (MSD) relative to time through

$$D=\frac{1}{6}\lim_{t\to\infty}\frac{\mathrm{d}\langle\|\mathbf{r}(t)-\mathbf{r}(0)\|^2\rangle}{\mathrm{d}t} \tag{9}$$

where $\langle\|\mathbf{r}(t)-\mathbf{r}(0)\|^2\rangle$ is the MSD of the center of mass of each water molecule. The slope of MSD over time is obtained by applying a linear fit to the corresponding MSD curve.

Data availability

All data generated or analyzed during this study are available at https://doi.org/10.6084/m9.figshare.28510238.v1 (ref. 65). Source data are provided with this paper.

Code availability

The source codes and examples of the MB-PIPNet approach are available on Zenodo at https://doi.org/10.5281/zenodo.14954863 (ref. 66) and GitHub at (https://github.com/qiyuchem/MB-PIPNet and https://github.com/szquchen/MSA-2.0).

References

1. Bartlett, R. J. & Musiał, M. Coupled-cluster theory in quantum chemistry. *Rev. Mod. Phys.* **79**, 291–352 (2007).
2. Gkeka, P. et al. Machine learning force fields and coarse-grained variables in molecular dynamics: application to materials and biological systems. *J. Chem. Theory Comput.* **16**, 4757–4775 (2020).
3. Deringer, V. L., Caro, M. A. & Csányi, G. Machine learning interatomic potentials as emerging tools for materials science. *Adv. Mat.* **31**, 1902765 (2019).
4. Manzhos, S., Dawes, R. & Carrington, T. Neural network-based approaches for building high dimensional and quantum dynamics-friendly potential energy surfaces. *Int. J. Quantum Chem.* **115**, 1012–1020 (2014).
5. Manzhos, S. & Carrington Jr, T. Neural network potential energy surfaces for small molecules and reactions. *Chem. Rev.* **121**, 10187–10217 (2020).
6. Meuwly, M. Machine learning for chemical reactions. *Chem. Rev.* **121**, 10218–10239 (2021).
7. Braams, B. J. & Bowman, J. M. Permutationally invariant potential energy surfaces in high dimensionality. *Int. Rev. Phys. Chem.* **28**, 577–606 (2009).
8. Qu, C., Yu, Q. & Bowman, J. M. Permutationally invariant potential energy surfaces. *Annu. Rev. Phys. Chem.* **69**, 151–175 (2018).
9. Jiang, B. & Guo, H. Permutation invariant polynomial neural network approach to fitting potential energy surfaces. *J. Chem. Phys.* **139**, 054112 (2013).
10. Jiang, B., Li, J. & Guo, H. Potential energy surfaces from high fidelity fitting of ab initio points: the permutation invariant polynomial–neural network approach. *Int. Rev. Phys. Chem.* **35**, 479–506 (2016).
11. Shao, K., Chen, J., Zhao, Z. & Zhang, D. H. Fitting potential energy surfaces with fundamental invariant neural network. *J. Chem. Phys.* **145**, 071101 (2016).
12. Fu, B. & Zhang, D. H. Accurate fundamental invariant-neural network representation of ab initio potential energy surfaces. *Natl Sci. Rev.* **10**, nwad321 (2023).
13. Behler, J. & Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* **98**, 146401 (2007).
14. Behler, J. Four generations of high-dimensional neural network potentials. *Chem. Rev.* **121**, 10037–10072 (2021).
15. Chmiela, S., Sauceda, H. E., Müller, K.-R. & Tkatchenko, A. Towards exact molecular dynamics simulations with machine-learned force fields. *Nat. Commun.* **9**, 3887 (2018).
16. Bartók, A. P., Payne, M. C., Kondor, R. & Csányi, G. Gaussian approximation potentials: the accuracy of quantum mechanics, without the electrons. *Phys. Rev. Lett.* **104**, 136403 (2010).
17. Uteva, E., Graham, R. S., Wilkinson, R. D. & Wheatley, R. J. Interpolation of intermolecular potentials using Gaussian processes. *J. Chem. Phys.* **147**, 161706 (2017).
18. Schütt, K. T., Sauceda, H. E., Kindermans, P.-J., Tkatchenko, A. & Müller, K.-R. SchNet—a deep learning architecture for molecules and materials. *J. Chem. Phys.* **148**, 241722 (2018).
19. Unke, O. T. & Meuwly, M. PhysNet: a neural network for predicting energies, forces, dipole moments, and partial charges. *J. Chem. Theory Comput.* **15**, 3678–3693 (2019).

20. Zhang, L., Han, J., Wang, H., Car, R. & Weinan, E. Deep potential molecular dynamics: a scalable model with the accuracy of quantum mechanics. *Phys. Rev. Lett.* **120**, 143001 (2018).
21. Zhang, Y., Hu, C. & Jiang, B. Embedded atom neural network potentials: efficient and accurate machine learning with a physically inspired representation. *J. Phys. Chem. Lett.* **10**, 4962–4967 (2019).
22. Batzner, S. et al. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nat. Commun.* **13**, 2453 (2022).
23. Batatia, I., Kovacs, D. P., Simm, G., Ortner, C. & Csanyi, G. MACE: higher order equivariant message passing neural networks for fast and accurate force fields. *Adv. Neural Inf. Process. Syst.* **35**, 11423–11436 (2022).
24. Musaelian, A. et al. Learning local equivariant representations for large-scale atomistic dynamics. *Nat. Commun.* **14**, 579 (2023).
25. Heidar-Zadeh, F. et al. Information-theoretic approaches to atoms-in-molecules: Hirshfeld family of partitioning schemes. *J. Phys. Chem. A.* **122**, 4219–4245 (2017).
26. Uhlig, F., Tovey, S. & Holm, C. Emergence of accurate atomic energies from machine learned noble gas potentials. Preprint at https://arxiv.org/abs/2403.00377 (2024).
27. Konovalov, A., Symons, B. C. & Popelier, P. L. On the many-body nature of intramolecular forces in FFLUX and its implications. *J. Comput. Chem.* **42**, 107–116 (2021).
28. Symons, B. C. & Popelier, P. L. Application of quantum chemical topology force field FFLUX to condensed matter simulations: liquid water. *J. Chem. Theory Comput.* **18**, 5577–5588 (2022).
29. Manchev, Y. T. & Popelier, P. L. Modeling many-body interactions in water with Gaussian process regression. *J. Phys. Chem. A.* **128**, 9345–9351 (2024).
30. Bader, R. F. W. Atoms in molecules. *Acc. Chem. Res.* **18**, 9–15 (1985).
31. Popelier, P. L. A. *The Chemical Bond* 271–308 (John Wiley & Sons, 2014).
32. Gordon, M. S., Fedorov, D. G., Pruitt, S. R. & Slipchenko, L. V. Fragmentation methods: a route to accurate calculations on large systems. *Chem. Rev.* **112**, 632–672 (2012).
33. Hodges, M. P., Stone, A. J. & Xantheas, S. S. Contribution of many-body terms to the energy for small water clusters: a comparison of ab initio calculations and accurate model potentials. *J. Phys. Chem. A* **101**, 9163–9168 (1997).
34. Dahlke, E. E. & Truhlar, D. G. Electrostatically embedded many-body correlation energy, with applications to the calculation of accurate second-order Møller–Plesset perturbation theory energies for large water clusters. *J. Chem. Theory Comput.* **3**, 1342–1348 (2007).
35. Wang, Y. M., Shepler, B. C., Braams, B. J. & Bowman, J. M. Full-dimensional, ab initio potential energy and dipole moment surfaces for water. *J. Chem. Phys.* **131**, 054511 (2009).
36. Góra, U., Podeszwa, R., Cencek, W. & Szalewicz, K. Interaction energies of large clusters from many-body expansion. *J. Chem. Phys.* **135**, 224102 (2011).
37. Medders, G. R., Götz, A. W., Morales, M. A., Bajaj, P. & Paesani, F. On the representation of many-body interactions in water. *J. Chem. Phys.* **143**, 104102 (2015).
38. Yu, Q. & Bowman, J. M. VSCF/VCI vibrational spectroscopy of $H_7O_3^+$ and $H_9O_4^+$ using high-level, many-body potential energy surface and dipole moment surfaces. *J. Chem. Phys.* **146**, 121102 (2017).
39. Heindel, J. P. & Xantheas, S. S. The many-body expansion for aqueous systems revisited: I. Water–water interactions. *J. Chem. Theory Comput.* **16**, 6843–6855 (2020).
40. Zhu, X., Riera, M., Bull-Vulpe, E. F. & Paesani, F. MB-pol(2023): sub-chemical accuracy for water simulations from the gas to the liquid phase. *J. Chem. Theory Comput.* **19**, 3551–3556 (2023).
41. Yu, Q. et al. q-AQUA: a many-body CCSD(T) water potential, including 4-body interactions, demonstrates the quantum nature of water from clusters to the liquid phase. *J. Phys. Chem. Lett.* **13**, 5068–5074 (2022).
42. Qu, C. et al. Interfacing q-AQUA with a polarizable force field: the best of both worlds. *J. Chem. Theory Comput.* **19**, 3446–3459 (2023).
43. Partridge, H. & Schwenke, D. W. The determination of an accurate isotope dependent potential energy surface for water from extensive ab initio calculations and experimental data. *J. Chem. Phys.* **106**, 4618 (1997).
44. Zhu, Y.-C. et al. Torsional tunneling splitting in a water trimer. *J. Am. Chem. Soc.* **144**, 21356–21362 (2022).
45. Fu, B. & Zhang, D. H. Ab initio potential energy surfaces and quantum dynamics for polyatomic bimolecular reactions. *J. Chem. Theory Comput.* **14**, 2289–2303 (2018).
46. Cheng, B., Engel, E. A., Behler, J., Dellago, C. & Ceriotti, M. Ab initio thermodynamics of liquid and solid water. *Proc. Natl Acad. Sci. USA* **116**, 1110–1115 (2019).
47. Zhang, Y., Xia, J. & Jiang, B. REANN: a PyTorch-based end-to-end multi-functional deep neural network package for molecular, reactive, and periodic systems. *J. Chem. Phys.* **156**, 114801 (2022).
48. Zhai, Y., Caruso, A., Bore, S. L., Luo, Z. & Paesani, F. A 'short blanket' dilemma for a state-of-the-art neural network potential for water: reproducing experimental properties or the physics of the underlying many-body interactions? *J. Chem. Phys.* **158**, 084111 (2023).
49. Medders, G. R., Babin, V. & Paesani, F. Development of a 'first-principles' water potential with flexible monomers. III. Liquid phase properties. *J. Chem. Theory Comput.* **10**, 2906–2910 (2014).
50. Kapil, V. et al. i-PI 2.0: a universal force engine for advanced molecular simulations. *Comput. Phys. Commun.* **236**, 214–223 (2019).
51. Reddy, S. K. et al. On the accuracy of the MB-pol many-body potential for water: interaction energies, vibrational frequencies, and classical thermodynamic and dynamical properties from clusters to liquid water and ice. *J. Chem. Phys.* **145**, 194504 (2016).
52. Houston, P. L. et al. No headache for PIPs: a PIP potential for aspirin runs much faster and with similar precision than other machine-learned potentials. *J. Chem. Theory Comput.* **20**, 3008–3018 (2024).
53. Habershon, S., Markland, T. E. & Manolopoulos, D. E. Competing quantum effects in the dynamics of a flexible water model. *J. Chem. Phys.* **131**, 024501 (2009).
54. 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. *J. Chem. Phys.* **128**, 074506 (2008).
55. Cheng, B. Cartesian atomic cluster expansion for machine learning interatomic potentials. *npj Comput. Mater.* **10**, 157 (2024).
56. Conte, R., Qu, C. & Bowman, J. M. Permutationally invariant fitting of many-body, non-covalent interactions with application to three-body methane–water–water. *J. Chem. Theory Comput.* **11**, 1631–1638 (2015).
57. Mathur, R., Muniz, M. C., Yue, S., Car, R. & Panagiotopoulos, A. Z. First-principles-based machine learning models for phase behavior and transport properties of CO_2. *J. Phys. Chem. B* **127**, 4562–4569 (2023).
58. Houston, P. L. et al. PESPIP: software to fit complex molecular and many-body potential energy surfaces with permutationally invariant polynomials. *J. Chem. Phys.* **158**, 044109 (2023).

59. Chen, R., Shao, K., Fu, B. & Zhang, D. H. Fitting potential energy surfaces with fundamental invariant neural network. II. Generating fundamental invariants for molecular systems with up to ten atoms. *J. Chem. Phys.* **152**, 204307 (2020).
60. Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O. & Dahl, G. E. Neural message passing for quantum chemistry. In *International Conference on Machine Learning* (eds. Precup, D. & Teh, Y. W.) 1263–1272 (PMLR, 2017).
61. Schütt, K. T., Arbabzadah, F., Chmiela, S., Müller, K. R. & Tkatchenko, A. Quantum-chemical insights from deep tensor neural networks. *Nat. Commun.* **8**, 13890 (2017).
62. Qu, C. & Bowman, J. M. Communication: a fragmented, permutationally invariant polynomial approach for potential energy surfaces of large molecules: application to *N*-methyl acetamide. *J. Chem. Phys.* **150**, 141101 (2019).
63. Moré, J. J. The Levenberg–Marquardt algorithm: implementation and theory. In *Numerical Analysis: Proc. Biennial Conference held at Dundee, June 28–July 1, 1977* (ed. Watson, G. A.) 105–116 (Springer Berlin Heidelberg, 2006).
64. Anderson, J. B. A random-walk simulation of the Schrödinger equation: H_3^+. *J. Chem. Phys.* **63**, 1499–1503 (1975).
65. Yu, Q. et al. Data files for developing and testing MB-PIPNet models on water trimer, methane-water cluster, and liquid water. *figshare* https://doi.org/10.6084/m9.figshare.28510238.v1 (2025).
66. Yu, Q. Source code and example of MB-PIPNet approach. *Zenodo* https://doi.org/10.5281/zenodo.14954863 (2025).
67. Wang, Y. & Bowman, J. M. Rigorous calculation of dissociation energies (*D*) of the water trimer, $(H_2O)_3$ and $(D_2O)_3$. *J. Chem. Phys.* **135**, 131101 (2011).
68. Zhang, Y., Hu, C. & Jiang, B. Accelerating atomistic simulations with piecewise machine-learned ab initio potentials at a classical force field-like cost. *Phys. Chem. Chem. Phys.* **23**, 1815–1821 (2021).
69. Kovács, D. P., Batatia, I., Arany, E. S. & Csányi, G. Evaluation of the MACE force field architecture: from medicinal chemistry to materials science. *J. Chem. Phys.* **159**, 044118 (2023).
70. Mills, R. Self-diffusion in normal and heavy water in the range 1–45°. *J. Phys. Chem.* **77**, 685–688 (1973).
71. Holz, M., Heil, S. R. & Sacco, A. Temperature-dependent self-diffusion coefficients of water and six selected molecular liquids for calibration in accurate ^{1}H NMR PFG measurements. *Phys. Chem. Chem. Phys.* **2**, 4740–4742 (2000).
72. Skinner, L. B. et al. Benchmark oxygen–oxygen pair-distribution function of ambient water from X-ray diffraction measurements with a wide *Q*-range. *J. Chem. Phys.* **138**, 074506 (2013).
73. Skinner, L. B., Benmore, C. J., Neuefeind, J. C. & Parise, J. B. The structure of water around the compressibility minimum. *J. Chem. Phys.* **141**, 214507 (2014).
74. Soper, A. & Benmore, C. Quantum differences between heavy and light water. *Phys. Rev. Lett.* **101**, 065502 (2008).

Acknowledgements

Q.Y. and D.H.Z. acknowledge the support from National Natural Science Foundation of China (grant numbers 22473030 and 22288201). J.M.B. acknowledges support from NASA grant (80NSSC22K1167). R.C. thanks Università degli Studi di Milano for financial support under grant PSR2022_DIP_005_PI_RCONT.

Author contributions

Q.Y. conceived of the project, performed calculations and analyzed the data. R.M. performed timing tests. D.H.Z. and J.M.B. provided critical feedback. All authors discussed the results and contributed to writing the paper.

Competing interests

The authors declare no competing interests.

Additional information

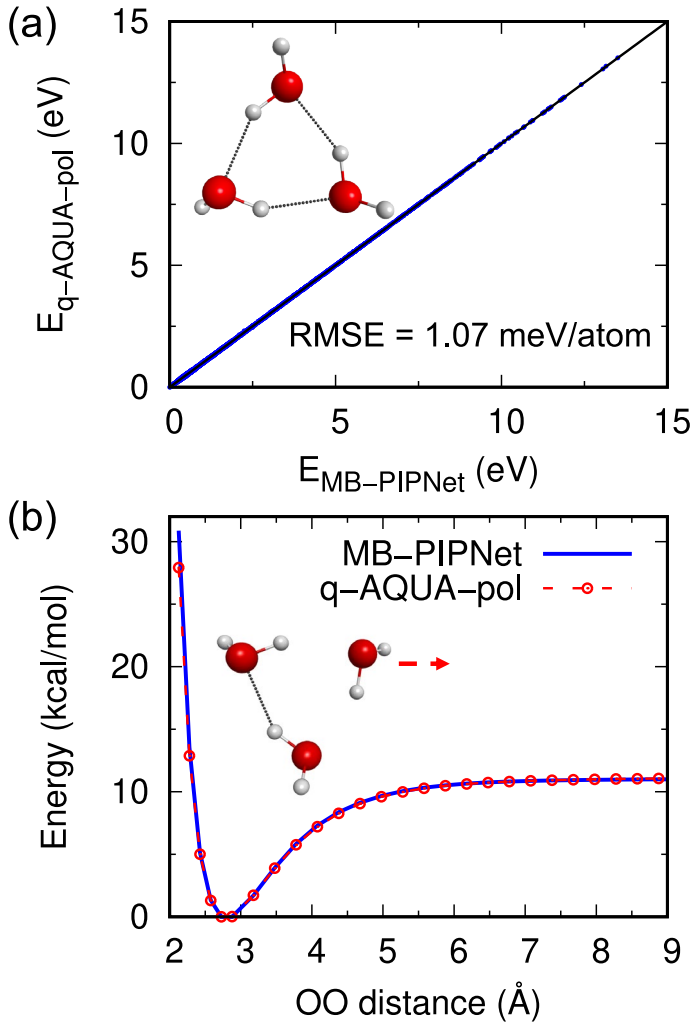
Extended data is available for this paper at https://doi.org/10.1038/s43588-025-00790-0.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43588-025-00790-0.

Correspondence and requests for materials should be addressed to Qi Yu.

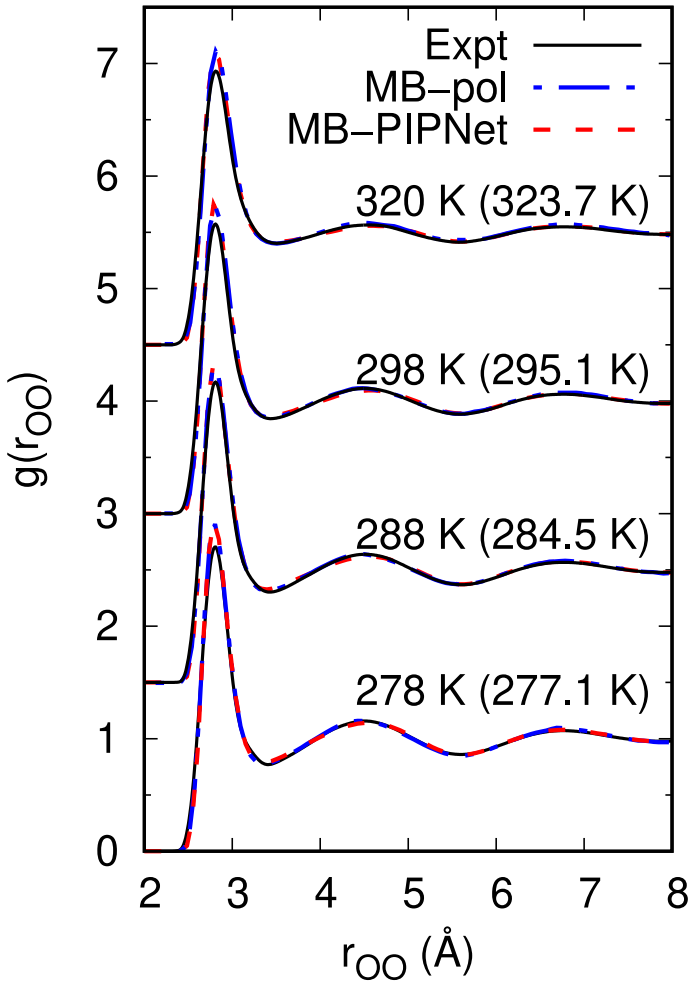
Peer review information *Nature Computational Science* thanks Bin Jiang and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Kaitlin McCardle, in collaboration with the *Nature Computational Science* team.

https://doi.org/10.1038/s43588-025-00790-0



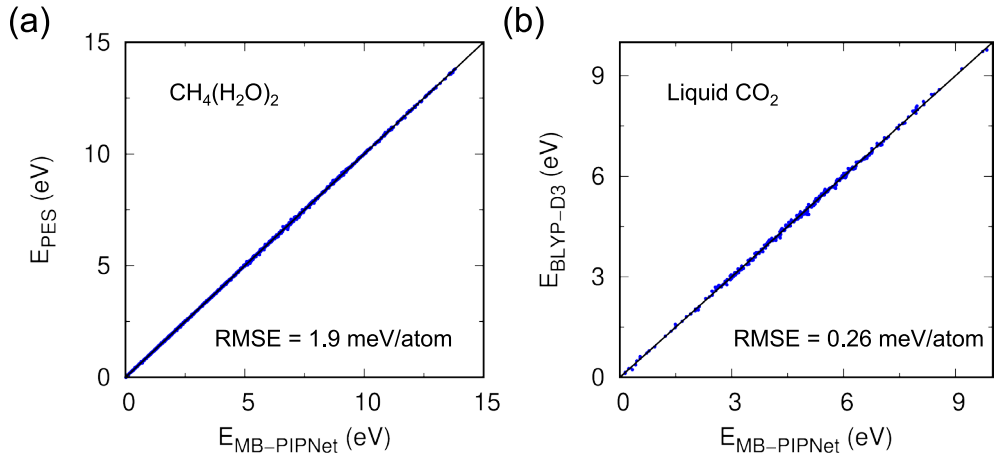
Extended Data Fig. 1 | Potential energy predictions from MB-PIPNet model of water trimer. (**a**) Energy-energy correlation plot for MB-PIPNet model of water trimer with reference energies calculated using q-AQUA-pol. (**b**) Potential energy curve predicted by MB-PIPNet model with comparison to q-AQUA-pol reference data. Atom colors: H-white, O-red.

https://doi.org/10.1038/s43588-025-00790-0



Extended Data Fig. 2 | Structural properties of liquid water at different temperatures predicted by MB-PIPNet model. OO radial distribution function from classical molecular dynamics simulations at different temperatures using MB-PIPNet model. The MB-pol data are taken from ref. 51. The experimental data are taken from ref. 71,72.

https://doi.org/10.1038/s43588-025-00790-0



Extended Data Fig. 3 | Performance of the MB-PIPNet model for methane-water clusters and liquid CO_2. (**a**) Correlation plots of test datasets for gas-phase $CH_4(H_2O)_2$ cluster with reference energies calculated using previously reported potential[56]. (**b**) Correlation plots of test datasets for liquid CO_2 with 64 molecules in simulation box with reference energies calculated at the BLYP-D3 level of theory[57].