Extending the atomic decomposition and many-body representation, a chemistry-motivated monomer-centered approach for machine learning potentials

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

E-mail: qi_yu@fudan.edu.cn; zhangdh@dicp.ac.cn; jmbowma@emory.edu

[1]Department of Chemistry, Fudan University, Shanghai, 200438, P.R. China

[2]Independent Researcher, Toronto, Ontario M9B0E3, Canada

[3]Dipartimento di Chimica, Università degli Studi di Milano, via Golgi 19, 20133, Italy

[4]Department of Physics and Materials Science, University of Luxembourg, L-1511, Luxembourg City, Luxembourg.

[5]Department of Chemistry, Emory University and Cherry L. Emerson Center for Scientific Computation, Atlanta, Georgia, USA

[6]Department of Chemistry and Chemical Biology, Cornell University, Ithaca, New York, USA

[7]Department of Chemistry and Biochemistry, Georgia Institute of Technology, Atlanta, Georgia, USA

[8]State Key Laboratory of Molecular Reaction Dynamics, Dalian Institute of Chemical Physics, Chinese Academy of Sciences, Dalian, 116023, P.R. China

Abstract

Most widely used machine learned (ML) potentials for condensed phase applications rely on many-body permutationally invariant polynomial (PIP) or atom-centered neural networks (NN). However, these approaches still face challenges in achieving chemical interpretability in atomistic energy decomposition and fully matching the computational efficiency of traditional force fields. Here, we present a novel method that combines aspects of both approaches, and achieves state-of-the-art balance of 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. Without sophisticated neural network design, the structural descriptors of monomers are described by 1-body and 2-body effective interactions, enforced by appropriate sets of PIPs as inputs to the feed forward NN. We demonstrate the performance of this method through systematic assessments of models for gas-phase water trimer, liquid water, methane-water cluster and liquid CO_2. The high accuracy, fast speed, and flexibility of this method provide a new route for constructing accurate ML potentials and enabling large-scale quantum and classical simulations for complex molecular systems.

Introduction

Computational simulations of molecular systems are essential for understanding complex processes in chemistry, biology, and material sciences. A significant challenge in both quantum and classical simulations is the extensive computation 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), are ideal. However, it quickly becomes prohibitive for systems with more than 15 atoms. Although density functional theory (DFT) is widely used in *ab initio* molecular dynamics 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 (MLP) have emerged as a promising approach to enable efficient and accurate computational simulations.[1–23] For high-dimensional systems with tens of thousands atoms, such as condensed phase water, an atomistic representation of the potential is a popular choice:[12]

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

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 atoms. This Behler-Parrinello representation has been widely applied in various machine learning potentials. Typical examples include BPNN,[12] SchNet,[17] PhysNet,[18] DeePMD[19] and EANN[20] etc.. Recent equivariant neural network potentials such as NequIP,[21] MACE,[22] and Allegro[23] 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 ML potentials lacks effective chemical meaning,[24,25] 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 neural networks and compromises their transferability to different systems.[25] It is worth noting that ML force field approaches, such as FFLUX,[26–28] have been developed to obtain physically well-defined atomic properties using the quantum chemical topology framework[29,30] 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 machine learning potentials for large molecular systems is the many-body representation, which has been widely used in the literature since the 1980s.[31–38] Taking water potentials as examples, the most accurate ones, namely MB-pol,[39] q-AQUA,[40] and q-AQUA-pol,[41] 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 each term is obtained from training a machine learning potential (MLP) on the appropriate dataset. Specifically, the 1-body term represents the potential for the isolated water monomer, often modeled using the spectroscopically accurate *ab initio* based Partridge-Schwenke potential.[42] The 2-body term is an MLP fit to dimer interaction electronic energies, the 3-body term is an MLP fit to trimer interaction energies, and the 4-body term is an MLP fit to tetramer interaction energies. This many-body formulation allows the use of permutationally invariant polynomial based methods, such as PIP,[7] PIPNN,[8] and FINN,[10] 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,[34,39–41,43] it is well-known that many-body representation suffers from the rapidly increasing number of 3-body, 4-body, and higher-order terms. Con-

sequently, long-time simulations of relevant molecular systems are often prohibitive.

In this work, we introduce a novel machine learning framework that combines the strengths of both methods while mitigates their weaknesses. This new monomeric framework leverages a new representation of the system's potential energy in terms of monomer energies instead of atomic local energies and employs molecular energies only at the 1-body and 2-body levels. Permutational invariance is enforced by using PIPs as inputs to NNs, describing molecule's structure and environment. This critical aspect of our work echoes the use of PIPs[8,9] and later efficient Fundamental Invariants[10,44] as inputs to NNs in applications to gas phase molecules. We term this new approach MB-PIPNet. We demonstrate that this method achieves high accuracy across a variety of molecular systems, ranging from gas phase clusters (e.g., water trimer and methane-water cluster) to condensed phase systems (e.g., liquid water and CO_2).

Our findings indicate that many-body interactions, such as 3-body interactions, can be accurately described using only 1-body and 2-body PIP bases in the neural network descriptor. This discovery significantly enhances the efficiency of our framework, enabling fast computational simulations of complex condensed phase systems at the cost of conventional force field. Our framework exhibits excellent performance in molecular dynamics (MD) simulations of liquid water and achieves significantly better computational scaling compared to other atomistic machine learning models. Furthermore, our new framework can be systematically extended to various types of molecular systems. Related challenges and possible solutions are also discussed.

Results

Monomeric neural network model

The proposed monomeric neural network 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 perturbed energy of each monomer, E_i, analogous to the atom-centered approach, such that

$$E_{\text{total}} = \sum_{i}^{N} E_i \tag{3}$$

The energy of each perturbed monomer is trained using a feed-forward neural network 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. For instance, for a tetraatomic monomer, a set of symmetrized polynomials at order n can be generated:

$$\mathbf{G}_i(\text{self}) = \mathbf{P}(\mathbf{X}_i) = \hat{\mathbf{S}}[y_{12}^a y_{13}^b y_{14}^c y_{23}^d y_{24}^e y_{34}^f] \tag{4}$$

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 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 2-body PIPs as the environment descriptor for each monomer, $\mathbf{G}_i(\text{env})$:

$$\mathbf{G}_i(\text{env}) = \sum_{j=1}^{N_{\text{mol}} \in R_{\text{c}}} \mathbf{P}(\mathbf{X}_i, \mathbf{X}_j) f_c(\mathbf{X}_i, \mathbf{X}_j, R_{\text{c}}) \tag{5}$$

where N_{mol} represents the total number of surrounding monomers within a distance cutoff of R_{c}. $\mathbf{P}(\mathbf{X}_i, \mathbf{X}_j)$ is the corresponding 2-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 0 for the 2-body polynomials when the distance between two monomers exceeds R_c. Taking water as example, the 2-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 they approach zero asymptotically as the distance between the oxygen atoms increases.[40]

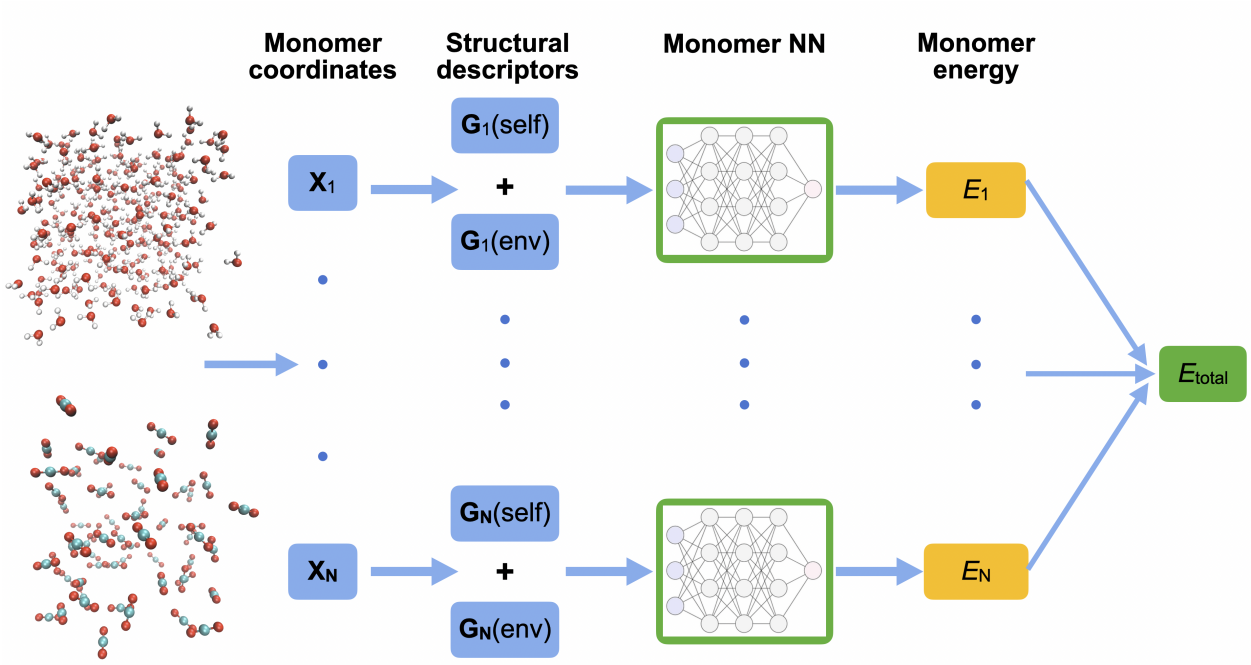

Figure 1: **Schematic of the MB-PIPNet architecture.** The coordinates of each fragmental monomer is first transferred to this monomer's self-structural descriptors $\mathbf{G}_i$(self), i.e., 1-body permutationally invariant polynomial (PIP) bases. The structural descriptors of each monomer's environment, $\mathbf{G}_i$(env), are generated by pair-wise monomer coordinates involving different monomers, i.e., 2-body PIP bases. The self- and environmental-descriptors of each monomer are combined as the input of the neural network 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.

The combination of $\mathbf{G}_i$(self) and $\mathbf{G}_i$(env) offers a systematic approach to describe the molecular response within a complex system. As will be explored below, using 1-b and 2-b PIPs as core components in these descriptors results in significantly more efficient computation compared to other ML methods.

Energetic properties of MB-PIPNet model for water trimer

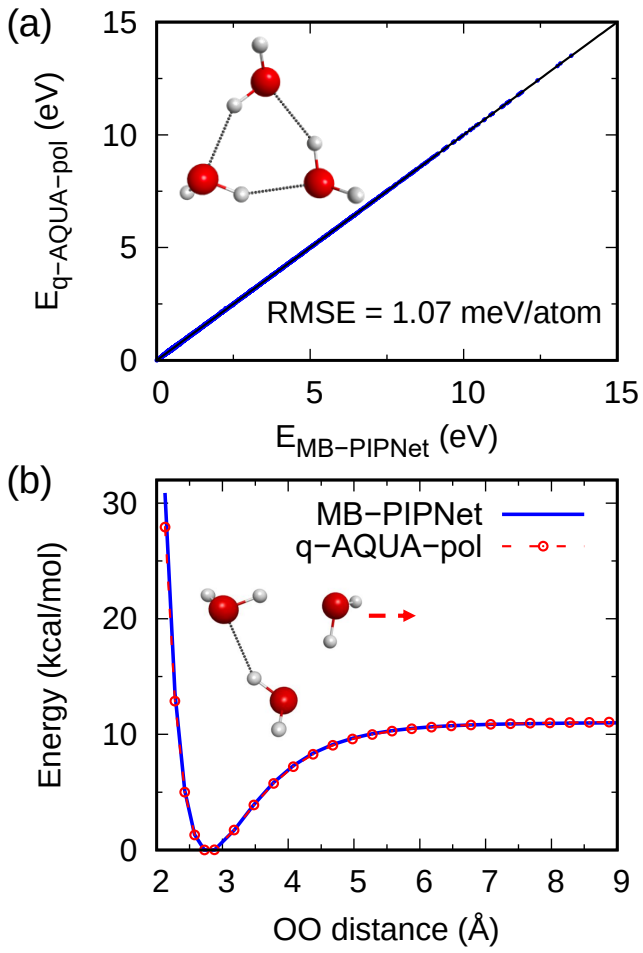

Figure 2: **Potential energy predictions from MB-PIPNet model of water trimer.** (a) Energy-energy correlation plot for MB-PIPNet model of water trimer with reference energies calculated using q-AQUA-pol. (b) Potential energy curve predicted by MB-PIPNet model with comparison to q-AQUA-pol reference data.

The use of only 1-body and 2-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/atom. As also shown

in Fig. 2(a), the corresponding MB-PIPNet model for water trimer exhibits high accuracy in predicting potential energies of a separate test dataset, with RMSE of 1.07 meV/atom. We also examined the accuracy of the MB-PIPNet potential in a representative potential energy cut. As shown in Fig. 2(b), excellent agreement with the q-AQUA-pol reference data is achieved. These single-point energy results provide direct evidence that using only 1-body and 2-body PIP bases in constructing structural descriptors, the trained MB-PIPNet potential is capable of handling the complex molecular environments beyond simple 2-body interactions.

Table 1: Harmonic frequencies and anharmonic diffusion Monte Carlo (DMC) zero point energy (ZPE) (in cm^{-1}) of water trimer from different methods.

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	1648.3	1662.2	1654.9
mode 14	1655.3	1665.3	1660.2
mode 15	1674.3	1684.0	1678.9
mode 16	3618.9	3621.1	3621.0
mode 17	3675.0	3681.2	3677.6
mode 18	3684.7	3689.3	3685.5
mode 19	3903.8	3907.2	3903.3
mode 20	3908.2	3910.2	3908.3
mode 21	3909.9	3914.3	3908.8
Harmonic ZPE			
Method	MB-PIPNet	q-AQUA-pol	*ab initio*[a]
	15997.7	16048.8	16017.7
Anharmonic DMC ZPE			
Method	MB-PIPNet	q-AQUA-pol[b]	WHBB[c]
	15593 ± 4	15616 ± 2	15587 ± 2

[a] CCSD(T)-F12a/aug-cc-pVTZ
[b] From Ref. 41
[c] From Ref. 45

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 to 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 PES, unconstrained DMC calculations were performed to determine the anharmonic zero-point energies (ZPE) of the water trimer using the trained MB-PIPNet potential. Notably, the rigorous "exact" 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 15593 ± 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 significantly 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] Of this dataset, 90% was randomly selected for training, with the remaining configurations used for testing. We employed the same 1-body and 2-body PIP bases as those used for the water trimer to generate the structural descriptors, $\mathbf{G}$(self) and $\mathbf{G}$(env). The RMSE errors on energy and force for the MB-PIPNet model, evaluated on the test set, are shown in Table 2. In comparison to 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[21] and MACE,[22] exhibit better performance, particularly for force predictions. These MPNN models typically involve tens of thousands of parameters, suggesting that the RMSE error of the MB-PIPNet

model could potentially be further reduced with more complicated NN structure and the incorporation of a message-passing mechanism.

Table 2: Root mean square errors (RMSE) for energy (meV/atom) and force (meV/Å) from different machine learning potentials trained on the same liquid water dataset from Ref. 46.

	BPNN[46]	EANN[48]	REANN[47]	NequIP[21]	MACE[22,49]	MB-PIPNet
Energy	2.33	2.1	0.8	0.93	0.63	1.19
Force	120	129	47	45	36.2	93.3

To obtain a 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 Zhai et al.[50] The training set consists of 75,874 different configurations from MD simulations of liquid water at various temperatures, using a cubic box of 256 water molecules under periodic boundary conditions. The total energy of each configuration was calculated using the MB-pol force field.[51] The training process over this extensive dataset converged quickly, as shown in Supplementary Fig. 1. The final training RMSE is only 0.29 meV/atom, which is notably smaller than the 0.39-0.44 meV/atom achieved using the DeePMD approach with the same dataset.[50] Fig. 3(a) 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/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 orbital energy against atomic orbital energy in electronic structure theory. Fig. 3(b) presents the 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 (P-S)[42] 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,

$$E_i(\text{q-AQUA}) = E_{1-b}(i) + \sum_{j}^{N} \frac{1}{2} E_{2-b}(i,j) + \sum_{j>k}^{N} \frac{1}{3} E_{3-b}(i,j,k) + \sum_{j>k>l}^{N} \frac{1}{4} E_{4-b}(i,j,k,l) \quad (6)$$

As seen, the P-S 1-body energies of 256 water molecules range from [0,5] kcal/mol, 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 range from [-20,-3] kcal/mol. In line with the q-AQUA results, the MB-PIPNet model provides reasonable predictions of monomer energies, with different water molecules display 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 1-body and 2-body PIP bases.

The trained MB-PIPNet model of liquid water was further employed in molecular dynamics simulations for bulk water properties using the i-PI software.[52] Fig. 3(c) shows the oxygen-oxygen (OO) radial distribution function (RDF) obtained from classical molecular dynamics 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. Fig. 4 presents additional OO RDFs results generated by the MB-PIPNet and MB-pol across a range of temperatures. Both models consistently demonstrate excellent

agreement with experimental data, highlighting the MB-PIPNet model's capability to accurately replicate the MB-pol force field in simulating the structural properties of water across different thermal conditions.

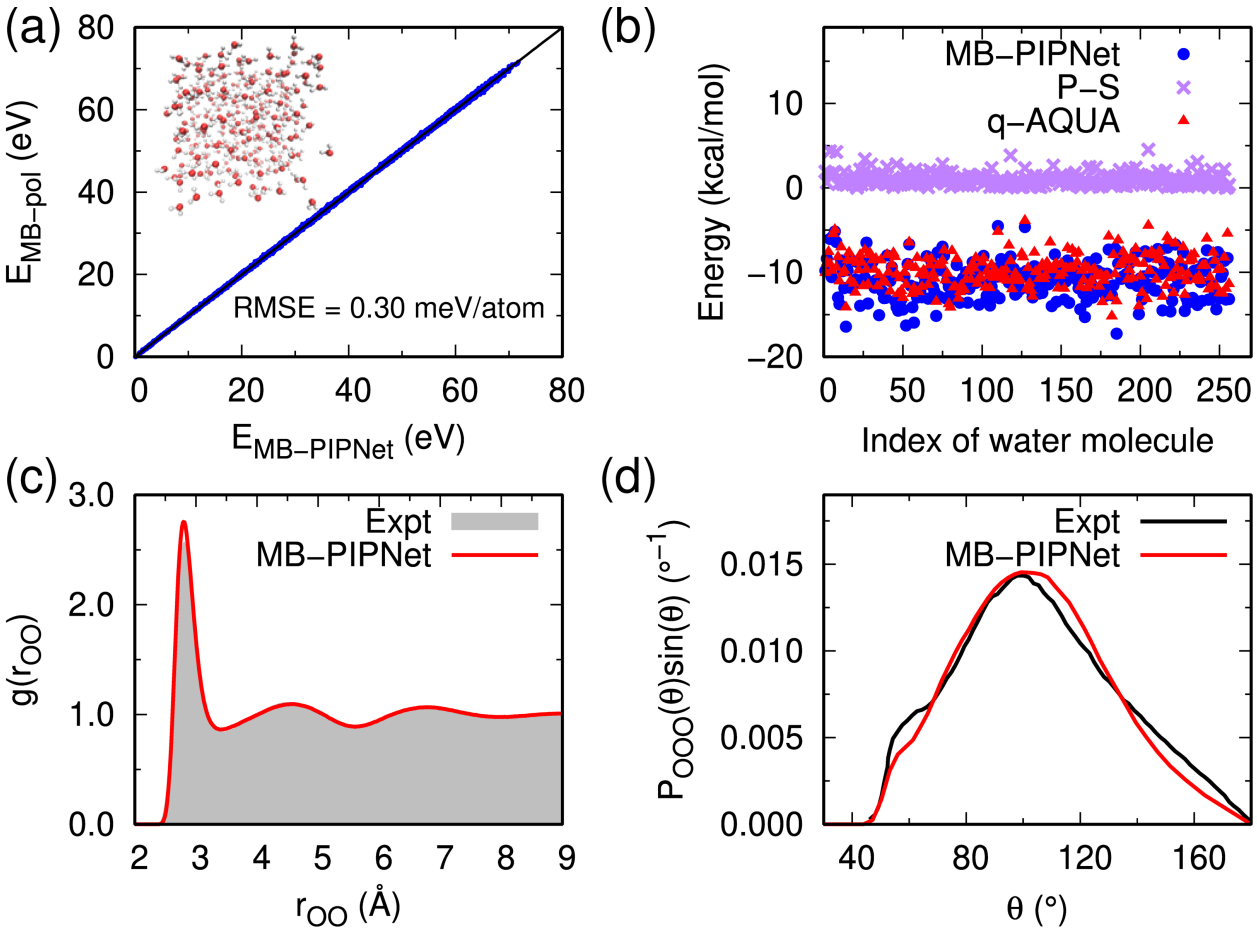

Figure 3: **Potential energy predictions and molecular dynamics simulation results of liquid water by MB-PIPNet model.** (a) Energy-energy correlation plot for MB-PIPNet model of liquid water with reference energies calculated using MB-pol. (b) Scatter plot of monomer energies of 256 water molecules in a periodic cubic box predicted by MB-PIPNet model, Partridge-Schwenke (P-S) water monomer potential, and q-AQUA model. (c) OO radial distribution function for liquid water at 298 K from classical MD simulations using MB-PIPNet model. The experimental data are taken from Ref. 53,54. (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. 55. The triplet angular distribution functions shown here were normalized to $\int_0^\pi P_{OOO}(\theta)\sin(\theta)d\theta$.

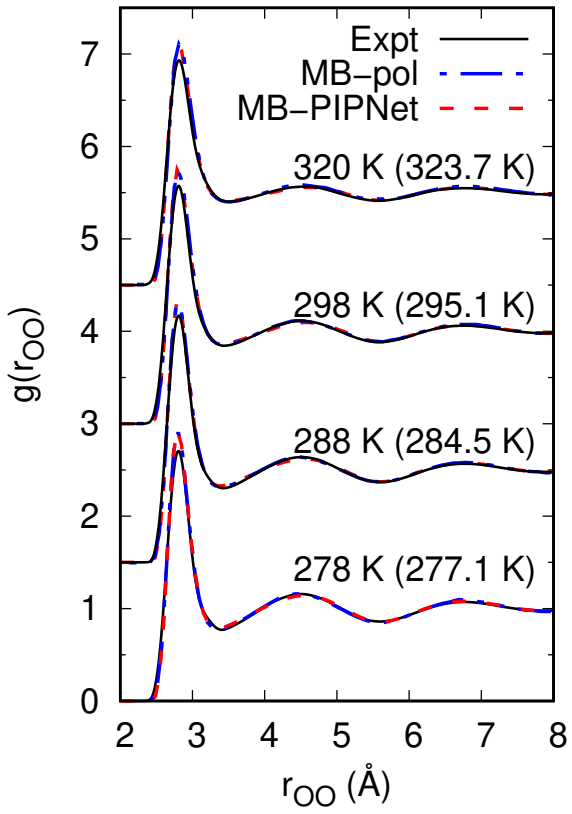

Figure 4: **Structural properties of liquid water at different temperatures predicted by MB-PIPNet model.** OO radial distribution function from classical molecular dynamics simulations at different temperatures using MB-PIPNet model. The MB-pol data are taken from Ref. 56. The experimental data are taken from Ref. 53,54

The oxygen-oxygen-oxygen triplet angular distribution $P_{OOO}(\theta)$ is another static property used to detect the tetrahedral orientational ordering of liquid water induced by the H-bonded network. We obtain $P_{OOO}(\theta)$ 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 cutoff of 3.27 Å to yield an average oxygen-oxygen coordination number of around 4.[41] As shown in Fig. 3(d), the distribution of $P_{OOO}(\theta)$ 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.

The self-diffusion coefficient D is calculated based on the slope of mean square displacements over time. 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,[56] although there are slight differences when compared to 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.[41,56]

Table 3: Self-diffusion coefficient D (Å^2/ps) of liquid water at different temperatures

Temperature (K)	MB-PIPNet	MB-pol[a]	Expt.[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

[a] from Ref. 56
[b] from Ref. 57 and 58

Before illustrating the computational efficiency of the MB-PIPNet approach, we investigate the effect of cut-off distance, R_c, in constructing the environment descriptor $\mathbf{G}$(env) in Eq. 5. In addition to the MB-PIPNet model trained using $R_c = 9$ Å, we also trained a MB-PIPNet model with $R_c = 15$ Å, termed as MB-PIPNet-long. During the model training phase, we did not observe a significant decrease in the training error for MB-PIPNet-long compared to the short-range model with $R_c = 9$ Å. Comparisons of the OO radial distribution functions between the two models are shown in Supplementary Fig. 5, where no significant differences are observed. This suggests the 2-body cut-off distance of 9 Å is sufficient to incorporate most interactions in liquid water and long-range interactions are implicitly included during the training process. A more careful assessment of the long-range interaction in the MB-PIPNet model is required and is subjected to our future study.

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 permutationally invariant polynomials as key components in structural descriptors. The generation of PIPs have been extensively verified to be systematic and efficient compared to other complicated ML descriptors.[59] Second, our MB-PIPNet framework employs a novel 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. 5, we compare the computational cost of the MB-PIPNet models with the q-TIP4P/F and TTM3-F force fields,[60,61] as well as the DeePMD and REANN MLPs,[47,50] for a single MD step of energy and gradient calculations across different sizes of liquid water simulation boxes. In the single 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 2-body cutoff (R_c=9 Å), the MB-PIPNet model demonstrates 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 significantly 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-b becomes almost identical to that of q-TIP4P/F. Furthermore, with a relatively short atomic environmental cutoff ($R_c = 3$

Å), the message-passing neural network REANN exhibits computational performance comparable to the MB-PIPNet model with R_c=9 Å. As the cutoff increases to R_c=5.5 Å, the computational cost of REANN rises significantly. Notably, as demonstrated in Ref. 62, sophisticated equivariant message-passing neural networks such as MACE, when utilizing an Nvidia A100 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.

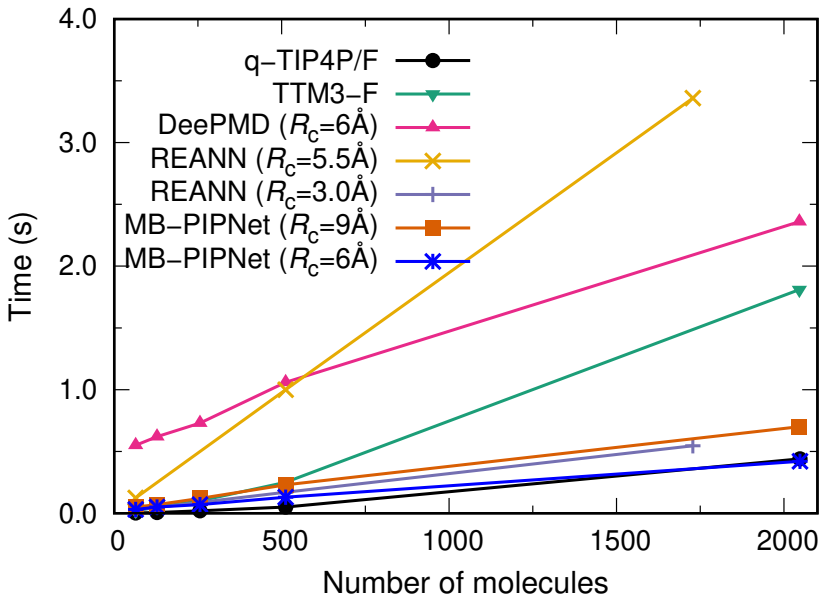

Figure 5: **Computational cost of MB-PIPNet model.** Computational time of single molecular dynamics step (energy and force) versus number of water molecules in a periodic simulation box using different methods. All timing tests were performed using a single CPU core of the AMD EPYC 7002 processors. R_c is the atomic or monomeric environment cutoff applied in different MLPs.

As demonstrated above, our MB-PIPNet approach achieves a lower training RMSE than the invariant MLPs such as DeepMD using the same dataset while maintaining force field-level efficiency. Additionally, the MB-PIPNet potential is highly parallelizable, owing to its

use of 1-body and 2-body PIPs for constructing structural descriptors and the formulation of monomeric energies. These features make it well-suited for microsecond-long simulations with first-principles accuracy of complex molecular systems. For instance, a recent study probing the liquid-liquid transition in supercooled water required long-term molecular dynamics (MD) simulations, spanning several years of GPU computational time, using the DeepMD-based potential.[63] With our MB-PIPNet approach, it could be expected that at the same or higher level of accuracy, better statistical significant investigations of the unique properties of water could be conducted with significantly lower computational cost. Furthermore, in quantum mechanics/molecular mechanics (QM/MM) simulations of biomolecular systems, the MM region typically comprises tens of thousands of atoms—often 10,000 to 20,000 water molecules are required to properly solvate the biomolecule.[64] Instead of relying on conventional force fields, our MB-PIPNet approach opens new possibilities for biomolecular simulations at fully *ab initio* levels of accuracy. While it is not yet feasible to perform *ab initio* calculations for energy and forces on the entire system for training the model, as we will discuss later, the MB-PIPNet framework can be integrated with other strategies, such as many-body expansion, to construct accurate global PES for complex systems.

Performance on other molecular systems and outlook

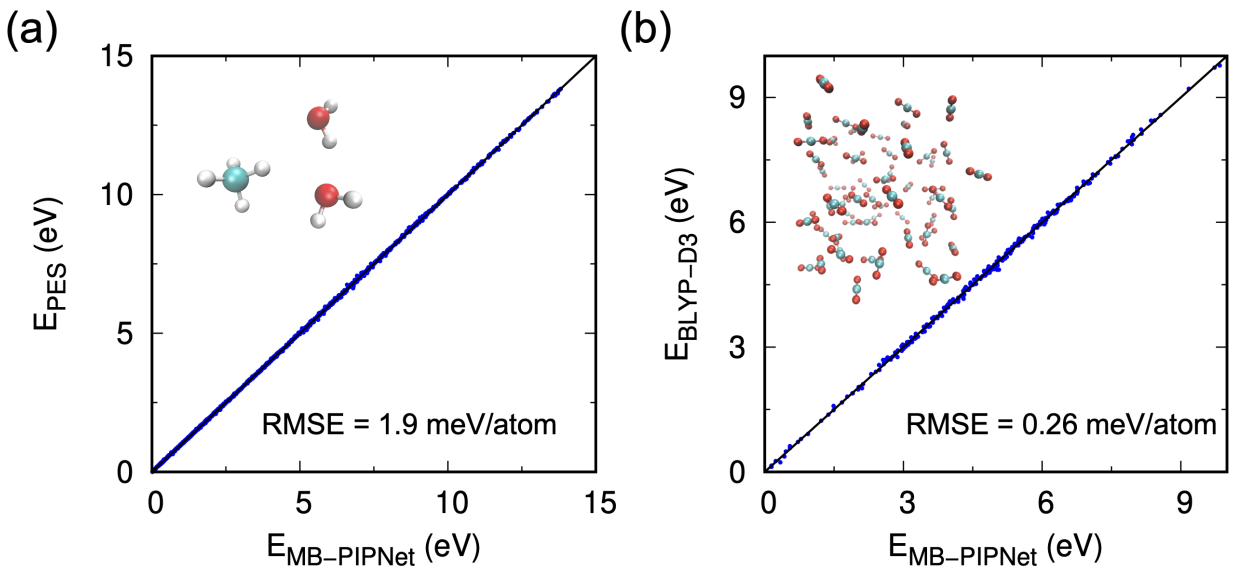

Figure 6: **Performance of the MB-PIPNet model for methane-water clusters and liquid CO2.** (a) Correlation plots of test datasets for gas-phase $CH_4(H_2O)_2$ cluster with reference energies calculated using previously reported potential.[65] (b) Correlation plots of test datasets for liquid CO2 with 64 molecules in simulation box with reference energies calculated at the BLYP-D3 level of theory.

The above two MB-PIPNet potentials are associated with gas-phase water trimer and liquid water. Here, we further demonstrate the transferability of the MB-PIPNet approach to different molecular systems. Fig. 6(a) illustrates the application of the MB-PIPNet model to methane-water clusters, $CH_4(H_2O)_2$. A total of 19,342 training structures and 2,228 test structures, along with their associated energies, were sourced from our previously developed many-body PES.[65] Utilizing the descriptors and network structures detailed in the Methods section, the RMSE for training and testing were determined to be 1.92 meV/atom and 1.90 meV/atom, respectively, which are reasonably small for gas-phase systems with energies up to 15 eV. As demonstrated in Supplementary Note 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. Fig. 6(b) 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,[66] the model achieved training and test RMSE values of 0.16 meV/atom and 0.26 meV/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 60 MLPs for different molecules.[67] Our PIP library can be directly interfaced with our MB-PIPNet approach to generate 1-body and 2-body terms as structural descriptors. However, the number of PIPs increases significantly with the number of atoms and polynomial orders, which often limits its application to systems with fewer than 15 atoms. Analogous to the behavior of FI-NN compared to PIP-NN, the fundamental invariant polynomials 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,[11,68] an extension of FIs is anticipated for large systems with more than 20 atoms, or even 30 atoms. Extensive test employing FIs in our MB-PIPNet framework, in terms of accuracy and additional efficiency, is under investigation.

Discussion

In this study, we present a novel framework for developing machine learning potentials for both gas-phase and condensed phase molecular systems. This framework, MB-PIPNet, discards the widely used atomic local energy decomposition method and avoids the expensive computations of high-order terms in many-body representation. By representing the total potential energy as the sum of chemically meaningful monomeric energies, the MB-PIPNet

framework accurately describes monomer structural distortion, environmental polarizations, and the final potential. It should be stressed that, unlike the standard many-body expansion approach, in MB-PIPNet, the combination of 1-body and 2-body PIP-based descriptors provides a consistent way to incorporate the perturbation effect of each monomer, where the perturbation results from the many-body interaction with other monomers. The MB-PIPNet framework greatly improves the computational cost of MLP evaluation, as the final computation scales linearly with the number of molecules instead of atoms. These features of the MB-PIPNet approach open new 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 atomic MLP like DeePMD.

The performance of MB-PIPNet approach has been systematically illustrated in quantum and classical simulations, i.e. quantum DMC and classical MD, of various molecular systems such as gas-phase water trimer and liquid water. As mentioned, generating molecular structural descriptors using PIPs or FIs is systematic and efficient, while ensuring rotational, translational, and permutational invariance. When training MB-PIPNet potentials for condensed phase systems like liquid water, a distance cutoff is incorporated for the 2-body PIPs of environmental descriptor $\mathbf{G}$(env). Our tentative tests did not find significant differences between models with different cutoff thresholds. However, like 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.[20,69,70]

Another advantage of the MB-PIPNet model is its flexible accuracy for the final PES. As shown in our examples and similar to other MLP approaches, the MB-PIPNet method can be directly used to train the total potential energy of the system. However the final accuracy of these MLPs heavily relies on the level of electronic structure theory, usually density functional theory, for generating the training data. This poses challenges in systematically elevating

the accuracy of the MLP to methods like CCSD(T) as CC calculations are often prohibitive for larger systems.[71,72] Conventional many-body representation of the potential is ideal for developing CC-level potential for condensed phase system like liquid water, but suffers from costly computation of high-order terms. Our MB-PIPNet framework offers new possibility for generating CC-level potential through a combination of many-body expansion. For example, in the case of liquid water, conventional 1-body and 2-body interactions can directly employ our recently developed high-accuracy CCSD(T)-level PESs in q-AQUA potential. For 3-body and higher-body interactions, the MB-PIPNet approach could directly fit the total $>$ 2-body energy contribution calculated from the q-AQUA PES. Maintaining the accuracy of the final model, the combination of MB-PIPNet and many-body expansion is highly efficient without extensive calculations of higher-order terms and requires no additional structural descriptors for MB-PIPNet, as both methods reply on the same set of 1-b and 2-b PIP bases.

A final remark on our MB-PIPNet framework pertains to the need for reasonable assignments of fragmented monomers in complex molecular systems. It can be naturally applied to systems like 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.[73] For condensed-phase reactive systems or materials, such as solid oxides, atomistic ML 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 ML 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 ML potential, while the MB-PIPNet model could efficiently describe the nonreactive solvent water molecules, thereby accelerating the

simulations. Such hybrid framework involving 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 neural networks.[18,21,22] We hope that the proposed new 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

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[40] and q-AQUA-pol[41] PESs. The remaining 5,674 structures were added by running diffusion Monte Carlo 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 data set of 5,194 structures.

For liquid water, the first dataset is from Cheng et al.[46] which includes 1593 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 Zhai et al.[50] We refer readers to the original publication for more details. Briefly, a final training data set of 75,874 configurations and test dataset of 9,448 configurations were generated from MD simulations at different 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 MB-pol force field.[51]

A total of 21,570 structures of $CH_4(H_2O)_2$ were obtained from our previous work[65] with

energies calculated using developed many-body potential. 90% of the dataset is used for training, resulting in 19,342 training and 2,228 testing configurations. The dataset for liquid CO_2 was directly obtained from Mathur et al.[66] where the configurations were obtained from MD simulations of bulk liquid states at T = 220-300 K and P = 100 bar for a system of 64 CO_2 molecules under periodic boundary condition. A total of 3,800 configurations at BLYP-D3 level of theory were used, with 2,687 for training and 313 for test.

Training details

For all MB-PIPNet models in this study, we employed 6th-order full-symmetry permutationally invariant polynomials for the self-structural descriptor, $\mathbf{G}$(self), and 4-th order full-symmetry 2-body PIP bases for the environmental descriptor, $\mathbf{G}$(env). Specifically, for both water trimer and liquid water, the input layer of NN has a dimension of 188, with 49 for $\mathbf{G}$(self) and 139 for $\mathbf{G}$(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, $\mathbf{G}_{CH_4}$(env) and $\mathbf{G}_{H_2O}$(env), 100 methane-water PIPs with 4211 symmetry and 50 water-water PIPs with 42 symmetry are used. With this setup, the input layers of NN for CH_4 and H_2O are both 180. For liquid CO_2, the input layer is 80, with 30 for $\mathbf{G}$(self) and 50 for $\mathbf{G}$(env). 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 respectively. The training of all MB-PIPNet models was realized using the Levenberg-Marquardt algorithm.[74] The training stopped when the learning rate dropped below 10^{-5}. The cutoff distances for the three systems were set as 9.0 Å.

Diffusion Monte Carlo simulations of water trimer

We performed diffusion Monte Carlo (DMC) calculations which is considered as the standard approach to rigorously calculate the ground vibrational state wave function and the

anharmonic ZPE in full dimensionality. This is based on the similarity between the diffusion equation and the imaginary-time Schrödinger equation with an energy shift E_{ref}

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

The reference energy E_{ref} in Eq. 7 is used to stabilize the diffusion system in its ground state, serving as an estimator for the zero-point energy.[75] We employed the unbiased, unconstrained implementation of DMC using $3N$ Cartesians.[76] In this method, the DMC calculation begins with an initial guess of the ground-state wave function, represented by a population of $N(0)$ equally weighted Gaussian random walkers. These walkers then diffuse randomly in imaginary time according to a Gaussian distribution. The population is controlled by a birth-death processes.[76] To maintain the number of random walkers around the initial value $N(0)$, E_{ref} is adjusted at the end of each time step according to

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

where $N(\tau)$ is the number of walkers at the time step τ, α is a feedback parameter, typically around 0.1, and $\langle V(\tau) \rangle$ represents the average potential energy of all of the walkers at that step. Finally the average of the E_{ref} provides an estimate of the ZPE.

In this study, the DMC calculations were carried out for water trimer with the imaginary time step $\Delta\tau = 5$ a.u. and $\alpha = 0.1$. Five independent DMC calculations were performed. In each DMC calculation, the number of walkers is 40 000, and these walkers are equilibrated for 5000 time steps followed by 40 000 propagation steps. The statistical uncertainty is estimated as the standard deviation of the 5 DMC runs for the same system.

Molecular dynamics simulations of liquid water

We interfaced the MB-PIPNet water potential with the i-PI software[52] to enable classical molecular dynamics simulations to be performed for bulk water. The canonical ensemble

(NVT) MD simulations were conducted with 256 water molecules in a periodically replicated simulation box with the experimental density set to be that at corresponding temperatures. For the classical MD simulation, at each temperature, we ran three independent 1 ns trajectories with time step of 0.25 fs. The static and dynamical properties were calculated as an average over the three trajectories.

The self diffusion coefficient, D, of liquid water can be calculated from:

$$D = \frac{1}{3}\int_0^\infty \langle \mathbf{v}(0)\cdot\mathbf{v}(t)\rangle dt = \frac{1}{6}\lim_{t\to\infty}\frac{d\langle \parallel \mathbf{r}(t)-\mathbf{r}(0)\parallel^2\rangle}{dt} \tag{9}$$

where $\langle \parallel \mathbf{r}(t)-\mathbf{r}(0)\parallel^2\rangle$ is the mean square displacement (MSD). For each trajectory, we used the center of mass of each water molecule to calculate the MSDs and conducted linear fits to obtain the slope of the MSD curve. The self-diffusion constant D is simply 1/6 of the MSD slope and the final values reported in the main text are from the averaged values over different trajectories.

References

(1) Gkeka, P.; Stoltz, G.; Barati Farimani, A.; Belkacemi, Z.; Ceriotti, M.; Chodera, J. D.; Dinner, A. R.; Ferguson, A. L.; Maillet, J.-B.; Minoux, H.; others Machine learning force fields and coarse-grained variables in molecular dynamics: application to materials and biological systems. *J. Chem. Theory Comput.* **2020**, *16*, 4757–4775.

(2) Deringer, V. L.; Caro, M. A.; Csányi, G. Machine learning interatomic potentials as emerging tools for materials science. *Adv. Mat.* **2019**, *31*, 1902765.

(3) 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.* **2014**, *115*, 1012–1020.

(4) Manzhos, S.; Carrington Jr, T. Neural network potential energy surfaces for small molecules and reactions. *Chem. Rev.* **2020**, *121*, 10187–10217.

(5) Meuwly, M. Machine learning for chemical reactions. *Chem. Rev.* **2021**, *121*, 10218–10239.

(6) Braams, B. J.; Bowman, J. M. Permutationally Invariant Potential Energy Surfaces in High Dimensionality. *Int. Rev. Phys. Chem.* **2009**, *28*, 577–606.

(7) Qu, C.; Yu, Q.; Bowman, J. M. Permutationally Invariant Potential Energy Surfaces. *Annu. Rev. Phys. Chem.* **2018**, *69*, 151–175.

(8) Jiang, B.; Guo, H. Permutation invariant polynomial neural network approach to fitting potential energy surfaces. *J. Chem. Phys.* **2013**, *139*, 054112.

(9) 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.* **2016**, *35*, 479–506.

(10) Shao, K.; Chen, J.; Zhao, Z.; Zhang, D. H. Communication: Fitting potential energy surfaces with fundamental invariant neural network. *J. Chem. Phys.* **2016**, *145*, 071101.

(11) Fu, B.; Zhang, D. H. Accurate fundamental invariant-neural network representation of ab initio potential energy surfaces. *Natl. Sci. Rev.* **2023**, *10*, nwad321.

(12) Behler, J.; Parrinello, M. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Phys. Rev. Lett.* **2007**, *98*, 146401.

(13) Behler, J. Four generations of high-dimensional neural network potentials. *Chem. Rev.* **2021**, *121*, 10037–10072.

(14) Chmiela, S.; Sauceda, H. E.; Müller, K.-R.; Tkatchenko, A. Towards exact molecular dynamics simulations with machine-learned force fields. *Nat. Commun.* **2018**, *9*, 3887.

(15) Bartók, A. P.; Payne, M. C.; Kondor, R.; Csányi, G. Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. *Phys. Rev. Lett.* **2010**, *104*, 136403.

(16) Uteva, E.; Graham, R. S.; Wilkinson, R. D.; Wheatley, R. J. Interpolation of intermolecular potentials using Gaussian processes. *J. Chem. Phys.* **2017**, *147*, 161706.

(17) 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.* **2018**, *148*, 241722.

(18) Unke, O. T.; Meuwly, M. PhysNet: A Neural Network for Predicting Energies, Forces, Dipole Moments, and Partial Charges. *J. Chem. Theory Comput.* **2019**, *15*, 3678–3693.

(19) 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.* **2018**, *120*, 143001.

(20) 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.* **2019**, *10*, 4962–4967.

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

(22) 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. Advances in Neural Information Processing Systems. 2022; pp 11423–11436.

(23) Musaelian, A.; Batzner, S.; Johansson, A.; Sun, L.; Owen, C. J.; Kornbluth, M.; Kozin-

sky, B. Learning local equivariant representations for large-scale atomistic dynamics. *Nat. Commun.* **2023**, *14*, 579.

(24) Heidar-Zadeh, F.; Ayers, P. W.; Verstraelen, T.; Vinogradov, I.; Vöhringer-Martinez, E.; Bultinck, P. Information-theoretic approaches to atoms-in-molecules: Hirshfeld family of partitioning schemes. *J. Phys. Chem. A* **2017**, *122*, 4219–4245.

(25) Uhlig, F.; Tovey, S.; Holm, C. Emergence of Accurate Atomic Energies from Machine Learned Noble Gas Potentials. 2024; `https://arxiv.org/abs/2403.00377`.

(26) Konovalov, A.; Symons, B. C.; Popelier, P. L. On the many-body nature of intramolecular forces in FFLUX and its implications. *J. Comput. Chem.* **2021**, *42*, 107–116.

(27) Symons, B. C.; Popelier, P. L. Application of quantum chemical topology force field FFLUX to condensed matter simulations: Liquid water. *J. Chem. Theory Comput.* **2022**, *18*, 5577–5588.

(28) Manchev, Y. T.; Popelier, P. L. Modeling Many-Body Interactions in Water with Gaussian Process Regression. *J. Phys. Chem. A* **2024**, *128*, 9345–9351.

(29) Bader, R. F. W. Atoms in molecules. *Acc. Chem. Res.* **1985**, *18*, 9–15.

(30) Popelier, P. L. A. *The Chemical Bond*; John Wiley & Sons, Ltd, 2014; Chapter 8, pp 271–308.

(31) Gordon, M. S.; Fedorov, D. G.; Pruitt, S. R.; Slipchenko, L. V. Fragmentation methods: A route to accurate calculations on large systems. *Chem. Rev.* **2012**, *112*, 632–672.

(32) 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* **1997**, *101*, 9163–9168.

(33) 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.* **2007**, *3*, 1342–1348.

(34) 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.* **2009**, *131*, 054511.

(35) Góra, U.; Podeszwa, R.; Cencek, W.; Szalewicz, K. Interaction energies of large clusters from many-body expansion. *J. Chem. Phys.* **2011**, *135*, 224102.

(36) 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.* **2015**, *143*, 104102.

(37) Yu, Q.; Bowman, J. M. Communication: VSCF/VCI vibrational spectroscopy of H7O3+ and H9O4+ using high-level, many-body potential energy surface and dipole moment surfaces. *J. Chem. Phys.* **2017**, *146*, 121102.

(38) Heindel, J. P.; Xantheas, S. S. The many-body expansion for aqueous systems revisited: I. Water–water interactions. *J. Chem. Theory Comput.* **2020**, *16*, 6843–6855.

(39) 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* **2023**, *19*, 3551–3556.

(40) Yu, Q.; Qu, C.; Houston, P. L.; Conte, R.; Nandi, A.; Bowman, J. M. 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. Letts.* **2022**, *13*, 5068–5074.

(41) Qu, C.; Yu, Q.; Houston, P. L.; Conte, R.; Nandi, A.; Bowman, J. M. Interfacing q-AQUA with a Polarizable Force Field: The Best of Both Worlds. *J. Chem. Theory Comput* **2023**, *19*, 3446–3459.

(42) 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.* **1997**, *106*, 4618.

(43) Zhu, Y.-C.; Yang, S.; Zeng, J.-X.; Fang, W.; Jiang, L.; Zhang, D. H.; Li, X.-Z. Torsional Tunneling Splitting in a Water Trimer. *J. Am. Chem. Soc.* **2022**, *144*, 21356–21362.

(44) Fu, B.; Zhang, D. H. Ab initio potential energy surfaces and quantum dynamics for polyatomic bimolecular reactions. *J. Chem. Theory Comput.* **2018**, *14*, 2289–2303.

(45) Wang, Y.; Bowman, J. M. Communication: Rigorous calculation of dissociation energies (D) of the water trimer, (H2O)3 and (D2O)3. *J. Chem. Phys.* **2011**, *135*, 131101.

(46) Cheng, B.; Engel, E. A.; Behler, J.; Dellago, C.; Ceriotti, M. Ab initio thermodynamics of liquid and solid water. *Proc. Natl. Acad. Sci.* **2019**, *116*, 1110–1115.

(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.* **2022**, *156*, 114801.

(48) 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.* **2021**, *23*, 1815–1821.

(49) 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.* **2023**, *159*, 044118.

(50) 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.* **2023**, *158*, 084111.

(51) 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.* **2014**, *10*, 2906–2910.

(52) Kapil, V. et al. i-PI 2.0: A Universal Force Engine for Advanced Molecular Simulations. *Comput. Phys. Commun.* **2019**, *236*, 214–223.

(53) Skinner, L. B.; Huang, C.; Schlesinger, D.; Pettersson, L. G. M.; Nilsson, A.; Benmore, C. J. Benchmark Oxygen-oxygen Pair-distribution Function of Ambient Water from X-ray Diffraction Measurements with a Wide Q-range. *J. Chem. Phys.* **2013**, *138*, 074506.

(54) Skinner, L. B.; Benmore, C. J.; Neuefeind, J. C.; Parise, J. B. The Structure of Water Around the Compressibility Minimum. *J. Chem. Phys.* **2014**, *141*, 214507.

(55) Soper, A.; Benmore, C. Quantum Differences Between Heavy and Light water. *Phys. Rev. Letts.* **2008**, *101*, 065502.

(56) Reddy, S. K.; Straight, S. C.; Bajaj, P.; Huy Pham, C.; Riera, M.; Moberg, D. R.; Morales, M. A.; Knight, C.; Götz, A. W.; Paesani, F. 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.* **2016**, *145*, 194504.

(57) Mills, R. Self-diffusion in Normal and Heavy Water in the Range 1-45°. *J. Phys. Chem.* **1973**, *77*, 685–688.

(58) Holz, M.; Heil, S. R.; Sacco, A. Temperature-dependent Self-diffusion Coefficients of Water and Six Selected Molecular Liquids for Calibration in Accurate 1H NMR PFG Measurements. *Phys. Chem. Chem. Phys.* **2000**, *2*, 4740–4742.

(59) Houston, P. L.; Qu, C.; Yu, Q.; Pandey, P.; Conte, R.; Nandi, A.; Bowman, J. M. 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.* **2024**, *20*, 3008–3018.

(60) Habershon, S.; Markland, T. E.; Manolopoulos, D. E. Competing quantum effects in the dynamics of a flexible water model. *J. Chem. Phys.* **2009**, *131*, 024501.

(61) 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.* **2008**, *128*, 074506.

(62) Cheng, B. Cartesian atomic cluster expansion for machine learning interatomic potentials. *npj Comput Mater.* **2024**, *10*, 157.

(63) Sciortino, F.; Zhai, Y.; Bore, S. L.; Paesani, F. Pinpointing the Location of the Elusive Liquid-Liquid Critical Point in Water. *ChemRxiv.* **2024**, doi:10.26434/chemrxiv–2024–dqqws.

(64) Senn, H. M.; Thiel, W. QM/MM methods for biomolecular systems. *Angew. Chem. Int. Ed.* **2009**, *48*, 1198–1229.

(65) 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.* **2015**, *11*, 1631–1638.

(66) 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 CO2. *J. Phys. Chem. B* **2023**, *127*, 4562–4569.

(67) Houston, P. L.; Qu, C.; Yu, Q.; Conte, R.; Nandi, A.; Li, J. K.; Bowman, J. M. PESPIP: Software to fit complex molecular and many-body potential energy surfaces with permutationally invariant polynomials. *J. Chem. Phys.* **2023**, *158*, 044109.

(68) 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.* **2020**, *152*, 204307.

(69) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. Neural message passing for quantum chemistry. International conference on machine learning. 2017; pp 1263–1272.

(70) Schütt, K. T.; Arbabzadah, F.; Chmiela, S.; Müller, K. R.; Tkatchenko, A. Quantum-chemical insights from deep tensor neural networks. *Nat. Commun.* **2017**, *8*, 13890.

(71) Daru, J.; Forbert, H.; Behler, J.; Marx, D. Coupled Cluster Molecular Dynamics of Condensed Phase Systems Enabled by Machine Learning Potentials: Liquid Water Benchmark. *Phys. Rev. Lett.* **2022**, *129*, 226001.

(72) Chen, M. S.; Lee, J.; Ye, H.-Z.; Berkelbach, T. C.; Reichman, D. R.; Markland, T. E. Data-Efficient Machine Learning Potentials from Transfer Learning of Periodic Correlated Electronic Structure Methods: Liquid Water at AFQMC, CCSD, and CCSD(T) Accuracy. *J. Chem. Theory Comput.* **2023**, *19*, 4510–4519.

(73) 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.* **2019**, *150*, 141101.

(74) Moré, J. J. The Levenberg-Marquardt algorithm: implementation and theory. Numerical analysis: proceedings of the biennial Conference held at Dundee, June 28–July 1, 1977. 2006; pp 105–116.

(75) Anderson, J. B. A Random-walk Simulation of the Schrödinger Equation: H_3^+. *J. Chem. Phys.* **1975**, *63*, 1499–1503.

(76) Kosztin, I.; Faber, B.; Schulten, K. Introduction to the Diffusion Monte Carlo Method. *Am. J. Phys.* **1996**, *64*, 633–644.

Data availability

The data generated and used in this study are available at upon request to the authors.

Code availability

The in-house program of the MB-PIPNet approach is available (https://github.com/qiyuchem/MB-PIPNet). The Monomial Symmetrization Approach (MSA) software to generate permutationally invariant polynomials are available (https://github.com/szquchen/MSA-2.0). The i-PI program used to perform the MD simulations in this work is available (https://github.com/i-pi/i-pi).

Acknowledgment

Q.Y. and D.H.Z. acknowledge the support from National Natural Science Foundation of China (grant no. 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 the project, performed calculations, and analyzed the data. R.M. performed timing tests. All authors contributed to writing the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available.