MB-Fit: Software Infrastructure for Data-Driven Many-Body Potential Energy Functions

Ethan F. Bull-Vulpe,[1, a)] Marc Riera,[1, b)] Andreas W. Götz,[2, c)] and Francesco Paesani[1, 3, 2, d)]

[1)] *Department of Chemistry and Biochemistry, University of California San Diego,*
La Jolla, California 92093, USA

[2)] *San Diego Supercomputer Center, University of California San Diego,*
La Jolla, California 92093, USA

[3)] *Materials Science and Engineering, University of California San Diego,*
La Jolla, California 92093, USA

Many-body potential energy functions (MB-PEFs), which integrate data-driven representations of many-body short-range quantum mechanical interactions with physics-based representations of many-body polarization and long-range interactions, have recently been shown to provide high accuracy in the description of molecular interactions, from the gas to the condensed phase. Here, we present MB-Fit, a software infrastructure for the automated development of MB-PEFs for generic molecules within the TTM-nrg ("Thole-type model energy") and MB-nrg ("many-body energy") theoretical frameworks. Besides providing all the necessary computational tools for generating TTM-nrg and MB-nrg PEFs, MB-Fit provides a seamless interface with the MBX software, a many-body energy/force calculator for computer simulations. Given the demonstrated accuracy of the MB-PEFs, we believe that MB-Fit will enable routine, predictive computer simulations of generic (small) molecules in the gas, liquid, and solid phases, including, but not limited to, the modeling of isomeric equilibria in molecular clusters, solvation processes, molecular crystals, and phase diagrams.

[a)] These authors contributed equally; Electronic mail: ebullvul@ucsd.edu

[b)] These authors contributed equally; Electronic mail: mrierari@ucsd.edu

[c)] Electronic mail: agoetz@sdsc.edu

[d)] Electronic mail: fpaesani@ucsd.edu

I. INTRODUCTION

Molecular-level computer simulations, such as molecular dynamics (MD) and Monte Carlo (MC) simulations,[1,2] have become an indispensable tool in molecular sciences, providing fundamental insights into structural, thermodynamic, and dynamical properties of molecular systems, from materials to biomolecules, which are difficult (if not impossible) to obtain by other means.[3–8] However, the level of realism and the predictive power of any MD and MC simulation depends sensitively on the accuracy of the potential energy function (PEF) used to represent the multidimensional potential energy surface (PES) of the molecular system in question. In the early days of computer simulations, due to limited computational resources, the only effectively suitable PEFs were empirically parameterized force fields (FFs) that adopted relatively simple expressions to describe intramolecular distortions and purely pairwise additive functions to describe intermolecular interactions.[9,10] While more sophisticated (nonpolarizable and polarizable) FFs have been developed over the years and are still the most common PEFs used in MD and MC simulations,[11–15] in the last decade machine-learning (ML) models trained on electronic structure data have gained in popularity, enabling computer simulations with higher level of accuracy.[16–19]

Various types of ML PEFs have been proposed, including neural network potentials (NNPs),[20–29] Gaussian approximation potentials (GAPs),[30] moment tensor potentials (MTPs),[31] and spectral neighbor analysis potentials (SNAPs),[32] as well as PEFs based on the atomic cluster expansion,[33] graph networks, kernel ridge regression methods,[34] gradient-domain machine learning (GDML),[35] support vector machines (SVM),[36] permutationally invariant polynomials (PIPs),[37–42] and permutation invariant polynomial neural networks (PIP-NNs).[43–46] The interested reader is referred to several excellent reviews of ML-based PEFs which have recently appeared in the literature.[16,18,47–49] It is generally found that ML PEFs trained on gas-phase reference data provide highly accurate descriptions of individual molecules and small clusters of molecules[50,51] but are not necessarily able to describe condensed-phase systems.[52] On the other hand, ML PEFs trained on condensed-phase data are able to closely reproduce the corresponding *ab initio* simulations of liquid and solid phases but are not, in general, directly transferable to molecular clusters or air/solid and air/liquid interfaces.[53] In this context, it should be noted that since gas-phase training data are generated for molecular systems with a handful of atoms, they can be computed at relatively higher levels of theory, often coupled cluster with single, double, and perturbative triple excitations, i.e., CCSD(T), the “gold standard” for molecular interactions,[54] compared to training data for condensed-phase systems which are effectively limited to density functional theory (DFT) calculations.[17]

An alternative ML approach to the development of accurate multidimensional PEFs, which are transferable from the gas to the condensed phase, can be rigorously derived from the many-body expansion (MBE) of the energy.[55] These many-body PEFs (MB-PEFs) adopt a hybrid data-driven/physics-based scheme, where a data-driven model, which captures many-body (short-range) quantum-mechanical interactions arising from the overlap of the electron densities of individual molecules (e.g., Pauli repulsion, and charge transfer and penetration), is integrated with a physics-based model of many-body interactions, which are generally represented by classical many-body electrostatics.[56,57] A remarkable example of MB-PEFs is the MB-pol PEF for water.[58–60] MB-pol has been shown to correctly reproduce the properties of water,[56,61] from small gas-phase clusters[62–73] to liquid water,[74–80] the air/water interface,[81–85], and ice.[86–90] The hybrid data-driven/physics-based scheme originally developed for MB-pol was later extended to generic molecules through the introduction of two families of MB-PEFs, the TTM-nrg (for "Thole-type model energy") and MB-nrg (for "many-body energy") PEFs, for halide[91–93] and alkali-metal[94–96] ions in water, molecular fluids,[97,98] and small molecules in water.[99] When employed in computer simulations, the MB-nrg PEFs have been shown to consistently provide remarkable agreement with experimental data for both gas-phase and condensed-phase systems.[100–107]

Here, we present MB-Fit, a complete software infrastructure for the automated development of TTM-nrg and MB-nrg PEFs for generic molecules. Besides providing a complete array of computational tools for generating the necessary training and test sets, performing the required quantum mechanical (QM) calculations, and fitting the TTM-nrg and MB-nrg PEFs, MB-Fit is seamlessly integrated with the MBX software,[108] a many-body energy/force calculator in both finite and periodic boundary conditions, which enables computer simulations with both TTM-nrg and MB-nrg PEFs, currently supporting LAMMPS[109] and i-PI[110].

II. THEORY

A. Many-body potential energy functions

The total energy of a system containing N (atomic and/or molecular) monomers can be formally expressed as the sum of individual n-body energies, $\varepsilon^{\mathrm{nB}}$, from one-body (1B) to N-body (NB), which is known as the many-body expansion (MBE) of the energy:[55]

$$E_N(1,\ldots,N) = \sum_{i=1}^{N} \varepsilon^{\mathrm{1B}}(i) + \sum_{i<j}^{N} \varepsilon^{\mathrm{2B}}(i,j) + \sum_{i<j<k}^{N} \varepsilon^{\mathrm{3B}}(i,j,k) + \ldots + \varepsilon^{\mathrm{NB}}(1,\ldots,N), \qquad (1)$$

In Eq. 1, $\varepsilon^{\mathrm{1B}}(i)$, corresponds to the distortion energy of monomer i from the corresponding equilibrium geometry, i.e., $\varepsilon^{\mathrm{1B}}(i) = E(i) - E_{\mathrm{eq}}(i)$. It follows that $\varepsilon^{\mathrm{1B}} = 0$ for monoatomic monomers. A rearrangement of Eq. 1 allows all higher-order n-body energies, $\varepsilon^{\mathrm{nB}}(2,\ldots,n)$, to be defined recursively as

$$\begin{aligned}\varepsilon^{\mathrm{nB}}(1,\ldots,n) = \quad & E_n(1,\ldots,n) - \sum_{i} \varepsilon^{\mathrm{1B}}(i) - \sum_{i<j} \varepsilon^{\mathrm{2B}}(i,j) - \ldots \\ & - \sum_{i<j<k} \varepsilon^{\mathrm{3B}}(i,j,k) - \ldots - \varepsilon^{\mathrm{(n\text{-}1)B}}(1,\ldots,n-1)\end{aligned} \qquad (2)$$

Since the MBE converges quickly for non-metallic systems, Eq. 1 served as a rigorous theoretical framework for the development of TTM-nrg and MB-nrg PEFs for various aqueous systems and molecular fluids, which are fully transferable from the gas to the condensed phase. Specifically, building upon the MB-pol PEF for water, TTM-nrg and MB-nrg PEFs were introduced for halide[91,92] and alkali-metal ions[94,95] in water, water–carbon dioxide[97] and water–methane[98] mixtures, and small molecules in water.[99] When used in computer simulations, the TTM-nrg and MB-nrg PEFs consistently provide remarkable agreement with experimental data, which is effectively quantitative in the case of the MB-nrg PEFs.[97,98,100–105,107]

In the most general form, the TTM-nrg and MB-nrg PEFs approximate the MBE of Eq. 1 as

$$E_N(1,\ldots,N) = \sum_{i=1}^{N} V^{\mathrm{1B}}(i) + \sum_{i>j}^{N} V^{\mathrm{2B}}(i,j) + \sum_{i>j>k}^{N} V^{\mathrm{3B}}(i,j,k) + V_{\mathrm{pol}}(1,\ldots,N) \qquad (3)$$

where V^{1B}, V^{2B}, and V^{3B} are fitted to reproduce the corresponding reference values calculated at the desired QM level of theory, and V_{pol} is an implicit many-body polarization term representing induction interactions. In both TTM-nrg and MB-nrg PEFs, the 1B term is described by a set of PIPs,[37] i.e., $\varepsilon^{\mathrm{1B}} = V^{\mathrm{1B}}_{\mathrm{PIP}}$.

The 2B term of Eq. 3 is expressed as

$$V^{\mathrm{2B}} = V^{\mathrm{2B}}_{\mathrm{sr}} + V^{\mathrm{2B}}_{\mathrm{disp}} + V_{\mathrm{elec}} \qquad (4)$$

where $V^{\mathrm{2B}}_{\mathrm{sr}}$ describes short-range interactions between each pair of monomers, $V^{\mathrm{2B}}_{\mathrm{disp}}$ describes 2B dispersion energy, and V_{elec} describes permanent electrostatics. In the TTM-nrg PEFs, $V^{\mathrm{2B}}_{\mathrm{sr}}$ is represented by a sum of pairwise Born-Mayer functions between all pairs of atoms located on the two monomers (M1 and M2),[111]

$$V^{2\text{B}}_{\text{sr, TTM-nrg}} = \sum_{\substack{i\in M1\\ j\in M2}} A_{ij} e^{-b_{ij}R_{ij}} \tag{5}$$

Here, R_{ij} is the distance between atoms i and j on monomers M1 and M2, respectively, and A_{ij} and b_{ij} are fitting parameters. In the MB-nrg PEFs, $V^{2\text{B}}_{\text{sr}}$ is expressed in terms of a set of PIPs[37] that smoothly switch to zero as the distance between monomers M1 and M2 becomes larger than a predefined cutoff value,

$$V^{2\text{B}}_{\text{sr, MB-nrg}} = s_2\left(\frac{R_{11}-R_{\text{in}}}{R_{\text{out}}-R_{\text{in}}}\right) V^{2\text{B}}_{\text{PIP}}(M1, M2) \tag{6}$$

where R_{11} is the distance between the first atom of M1 and the first atom of M2. In Eq. 6, $s_2(t)$ is a switching function defined as

$$s_2(t) = \begin{cases} 1 & \text{if } t < 0 \\ 0.5 \times [1+\cos(\pi \text{t})] & \text{if } 0 \le t < 1 \\ 0 & \text{if } 1 \le t \end{cases} \tag{7}$$

By construction, $s_2(t)$ = 1 when $R_{11} \le R_{\text{in}}$ and $s_2(t)$ = 0 when $R_{11} \ge R_{\text{out}}$. Therefore, the values of the inner (R_{in}) and outer (R_{out}) cutoffs in Eq. 6 define the region over which $V^{2\text{B}}_{\text{sr, MB-nrg}}$ is slowly and continuously switched off. The values of R_{in} and R_{out} are chosen based on the distance at which the short-range component of $\varepsilon^{2\text{B}}$ is no longer required to accurately model the 2B QM energies. Since in the current version of MB-Fit the switching function is calculated based on the coordinates of the first atom of each monomer, it is recommended to define the atom ordering so that the most "central" atom in each monomer is listed first. It should be noted that, while this definition of R_{11} is well suited for small molecules, more general definitions of the switching distance between two monomers (e.g., the distance between the monomer's centers of mass) may be more appropriate for larger molecules and will be implemented in future releases of MB-Fit.

The 2B dispersion energy in both TTM-nrg and MB-nrg PEFs is represented by a sum of pairwise additive contributions,

$$V^{2\text{B}}_{\text{disp}} = \sum_{\substack{i\in M1\\ j\in M2}} -f(\delta_{ij}R_{ij})\frac{C_{6,ij}}{R^6_{ij}} \tag{8}$$

where R_{ij} is the distance between atoms i and j respectively on monomers M1 and M2, $C_{6,ij}$ is the corresponding dispersion coefficient derived from QM calculations, and $f(\delta_{ij}R_{ij})$ is the Tang-Toennies damping function,[112]

$$f(\delta_{ij}, R_{ij}) = 1 - \exp(-\delta_{ij}R_{ij}) \sum_{n=0}^{6} \frac{(\delta_{ij}R_{ij})^n}{n!} \tag{9}$$

where δ_{ij} is set equal to fitting parameter b_{ij} in Eq. 5.

While the 3B term in Eq. 1 is set to zero in the TTM-nrg PEFs, it is represented by a short-range term in the MB-nrg PEFs:

$$V^{3\text{B}} = V^{3\text{B}}_{\text{sr}} \tag{10}$$

As in the analogous 2B term, $V^{3\text{B}}_{\text{sr}}$ is represented by a PIP[37] over variables that are functions of the distances between all atoms of the three monomers of the trimer,

$$V^{3\text{B}}_{\text{sr}} = [s_3(t_{12})s_3(t_{13}) + s_3(t_{12})s_3(t_{23}) + s_3(t_{13})s_3(t_{23})]\, V^{3\text{B}}_{\text{PIP}}(M1, M2, M3) \tag{11}$$

Here, the sum of the three terms in the square bracket represents a compound switching function that smoothly goes to zero as any of the molecules moves apart from the other two. In Eq. 11, $s_3(t) = s_2(t)$ from Eq. 7, and

$$t_{mn} = \frac{R_{mn}}{R_{\text{cut}}} \tag{12}$$

where R_{mn} is the distance between the first atoms on monomers m and n, and R_{cut} is a predefined 3B cutoff chosen to disable the 3B short-range term at distances where its contribution is negligible.

Finally, V_{pol} in Eq. 1 describes induction energy and is represented by a classical many-body polarization term built upon a modified version of the Thole-type model originally introduced in Ref. 113.

B. Permutationally invariant polynomials

As discussed in the previous section, both the TTM-nrg and MB-nrg PEFs contain 1B terms represented by PIPs, with the MB-nrg PEFs also including explicit 2B and 3B PIP terms. These PIPs take the following general form:[37]

$$P(\xi_1, \xi_2, \ldots, \xi_N) = \sum_{l=0}^{L} A_l S[{\xi_1}^{a_{l1}}, {\xi_2}^{a_{l2}}, \ldots, {\xi_N}^{a_{lN}}] \tag{13}$$

Here, ξ_i is a variable defined as a function of the distance R_{jk} between sites j and k, which include both physical atoms and fictitious sites of the monomers contributing to the 1B, 2B, or 3B PIP.

N is the total number of such variables, L is the total number of monomials in the polynomial, A_l is a linear fitting parameter and coefficient for monomial l, and $S[\xi_1{}^{a_{l1}}, \xi_2{}^{a_{l2}}, \ldots, \xi_N{}^{a_{lN}}]$ is an operator that symmetrizes each monomial l to guarantee that the PIP is invariant with respect to permutations of equivalent sites. Each a_{li} passed to S indicates the degree of each ξ_i in monomial l. The theory behind the development and symmetrization process of the PIPs is detailed in Ref. 37. In the MB-nrg PEFs, the 2B and 3B PIPs have been shown to correct deficiencies intrinsic to classical representations (e.g., Born-Mayer and Lennard-Jones functions) of quantum-mechanical short-range interactions (e.g., Pauli repulsion, charge transfer and penetration) that arise from the overlap of the monomer's electron densities.[57,93,96]

In the MB-Fit software, four different functional forms are available for the variables ξ. Each form is a function of the distance R and one or two non-linear fitting parameters k and d_0:

$$\xi^{\exp}(R) = e^{-kR} \tag{14a}$$

$$\xi^{\exp 0}(R) = e^{-k(R-d_0)} \tag{14b}$$

$$\xi^{\mathrm{coul}}(R) = e^{-kR}/R \tag{14c}$$

$$\xi^{\mathrm{coul0}}(R) = e^{-k(R-d_0)}/R \tag{14d}$$

It is important to note that, while the functional forms with the d_0 parameter (Eqs. 14b and 14d) are usually able to more closely reproduce the QM training data, they may also lead to discontinuities in the representation of the target multidimensional PES which, in turn, may result in instabilities in MD and MC simulations that use MB-nrg PEFs containing these monomials.

C. Fitting procedure

Eqs. 3 and 4 show that both TTM-nrg and MB-nrg PEFs include terms describing 1B distortions ($V_{\mathrm{PIP}}^{\mathrm{1B}}$), 2B short-range interactions ($V_{\mathrm{sr}}^{\mathrm{2B}}$), 2B dispersion ($V_{\mathrm{disp}}^{\mathrm{2B}}$), permanent electrostatics (V_{elec}), and many-body polarization (V_{pol}). The MB-nrg PEFs also include an explicit term describing short-range 3B interactions ($V_{\mathrm{sr}}^{\mathrm{3B}}$ in Eq. 11). As discussed in section II A, $V_{\mathrm{disp}}^{\mathrm{2B}}$, V_{elec}, and V_{pol} are derived from QM calculations of dispersion coefficients, atomic charges, and atomic polarizabilities carried out for an isolated monomer. The remaining $V_{\mathrm{PIP}}^{\mathrm{1B}}$, $V_{\mathrm{sr}}^{\mathrm{2B}}$ and $V_{\mathrm{sr}}^{\mathrm{3B}}$ terms are fitted to reproduce reference 1B, 2B, and 3B energies calculated at the desired QM level of theory.

Specifically, all linear and nonlinear fitting parameters entering the expressions of $V_{\mathrm{PIP}}^{\mathrm{1B}}$, $V_{\mathrm{sr}}^{\mathrm{2B}}$ and $V_{\mathrm{sr}}^{\mathrm{3B}}$ are determined by minimizing the regularized weighted sum of squared residuals calculated for the corresponding training sets, $\mathscr{S}$,

$$\chi^2 = \sum_{k\in\mathscr{S}} w_k \left[V^{\mathrm{nB}}(k) - V^{\mathrm{nB}}_{\mathrm{ref}}(k)\right]^2 + \Gamma^2 \sum_l c_l^2 \tag{15}$$

where $V^{\mathrm{nB}}_{\mathrm{ref}}(k)$ and $V^{\mathrm{nB}}(k)$ are the reference QM and corresponding TTM-nrg or MB-nrg nB energies ($n = 1,2,3$) for the k^{th} configuration in the training set. The weights, w_k, are set to emphasize configurations with low energies,

$$w_k = \left(\frac{\mathrm{DE}}{\mathrm{E}_k - \mathrm{E}_{\mathrm{min}} + \mathrm{DE}}\right)^2 \tag{16}$$

where E_k is the energy of the k^{th} n-body (i.e. monomer, dimer, or trimer) configuration in the training set, and $\mathrm{E}_{\mathrm{min}}$ is the corresponding minimum energy. For $\varepsilon^{1\mathrm{B}}$ training sets, $\mathrm{E}_{\mathrm{min}}$ corresponds to the energy of the monomer's optimized geomety, while for $\varepsilon^{2\mathrm{B}}$ and $\varepsilon^{3\mathrm{B}}$, the binding energies of the minimum-energy dimer or trimer geometries are used. DE in Eq. 16 thus defines the range of favorably weighted energies, with $w_k = 0.25$ for $\mathrm{E}_k = \mathrm{DE}$ and $w_k = 1$ for $\mathrm{E}_k = \mathrm{E}_{\mathrm{min}}$. The regularization parameter, Γ, is introduced in order to reduce the variation of the linear fitting parameters (larger Γ values suppress any variation) without spoiling the overall accuracy of the fit (favored by smaller Γ values), contributing no more than 1% to χ^2. Given the small number of linear parameters, Γ is not necessary in fitting the TTM-nrg PEFs. In Eq. 15, the linear parameters, c_l, are obtained through singular value decomposition, while the simplex algorithm is used to optimize the nonlinear parameters.

III. SOFTWARE INFRASTRUCTURE

The MB-Fit software supports a number of features enabling the user to construct well-behaved TTM-nrg and MB-nrg PEFs for generic molecules by following a standardized workflow. Broadly, the steps in the workflow are as follows: 1) generate training and test sets, 2) set up and perform the required QM calculations for collecting the necessary training data, 3) optimize the linear and non-linear parameters entering the mathematical expressions of the TTM-nrg and MB-nrg PEFs, and 4) generate the TTM-nrg and MB-nrg PEF codes that are exported to MBX[108] for subsequent MD simulations with LAMMPS[109] or i-PI.[110] The features provided by MB-Fit for each step of the workflow shown in Fig. 1 are elaborated upon below. Optionally, some of the steps may be skipped if the user wishes to directly provide the necessary data (e.g., dispersion coefficients, atomic charges, and atomic polarizabilities), which may be acquired using software different from

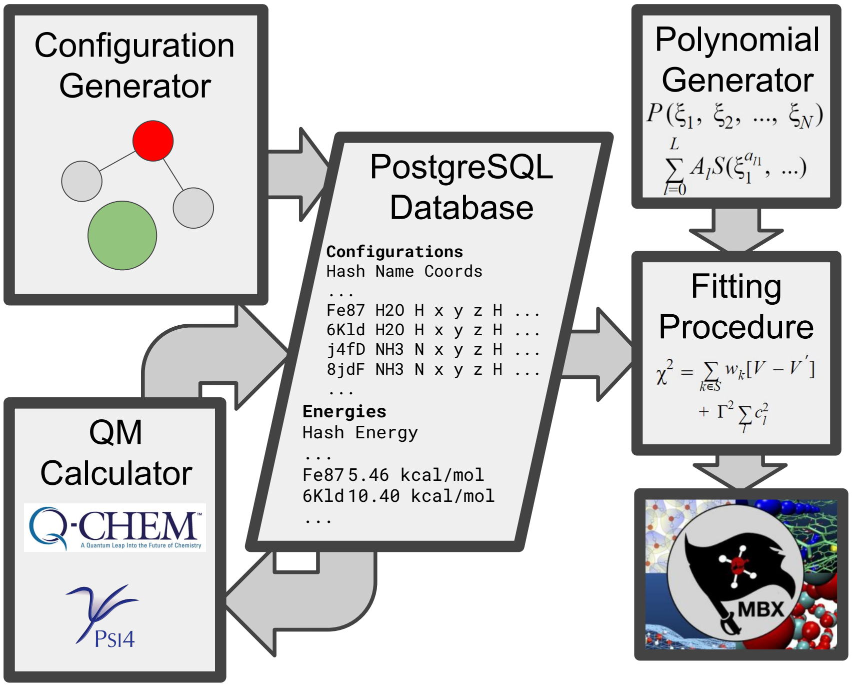

FIG. 1. MB-Fit drives generation of molecular configurations, quantum mechanical calculations, generation of permutationally invariant polynomials, parameter optimization, and export the final TTM-nrg and MB-nrg PEFs to the MBX[108] many-body energy/force calculator for molecular-level computer simulations. All data are stored in a central PostgreSQL database.

that currently supported by MB-Fit. MB-Fit is written in Python with a centralized postgreSQL database for storage of QM data. C++ is used in the codes for the parameter optimizations and energy evaluations of the TTM-nrg and MB-nrg PEFs. Maple[114] is employed for an optional factorization of the PIPs which allows for optimizing the run time of the final MB-PEFs.

It should be noted that, while the many-body formalism implemented in MB-Fit is completely general, its "off-the-shelf" application to large molecules (with more than ∼15-20 atoms) can become computationally expensive, both in terms of training and simulation.

A. Database

Storage of molecular configurations and associated nB energies is implemented using a PostgreSQL database.[115] This database can be either local or centralized, and allows simultaneous connections by multiple clients, facilitating collaboration. The basic unit of storage within the database is a molecular configuration, uniquely defined by a list of atoms and their coordinates alongside other molecular properties, such as net charge and spin multiplicity. Molecules are rotated into standard orientation and moved to their center of mass prior to insertion into the database in order to avoid repetition of configurations that differ only by rotations and/or translations. Each molecular configuration can be associated with one or more "models". A "model" is defined by the QM level of theory and the basis set to be used in the electronic structure calculations. Each configuration-model pair is associated with a number of electronic structure calculations required to obtain the corresponding nB energies. Tags can be assigned to groups of configuration-model pairs and later used to retrieve the corresponding nB energies. Generally, each training/test set is given a unique tag.

Database operations are implemented by server-side PostgreSQL functions and interfaced into the Python MB-Fit library using psycopg2.[116] Python interfaces are available to initialize the database, insert and retrieve configurations, generate input for electronic structure calculations, insert calculation results into the database, and generate training/test sets consisting of configurations and nB energies from the data stored in the database. Data insertion and retrieval is batched to allow a minimum of round-trips between the client and the server while avoiding transferring large amounts of data in a single payload. All operations run in average-case constant or linear time, thus removing any database access bottlenecks. Favorable runtime and scalability enable interactive retrieval of training set data consisting of tens of thousands of molecular configurations.

B. Training and test sets

As in the case of any ML PEF, training and test sets for the TTM-nrg and MB-nrg PEFs should provide a complete representation of the "physically relevant" low-energy regions of the target multidimensional PES which are explored in MD and MC simulations. At the same time, an adequate representation of high-energy configurations is also required for the TTM-nrg and MB-nrg PEFs in order to guarantee the absence of "holes" on the PES in regions where the PIPs, extrapolating from (incomplete) training sets, may predict unphysical energy values. To satisfy these requirements, the 1B training set is generated by sampling the harmonic distribution associated with the optimized structure of the monomer. Other local minima can also be sampled for complex molecules. Briefly, within the harmonic approximation, the canonical partition function for an N-atom molecule can be written as[117]

$$Z = \mathrm{Tr}\left(e^{-\beta\hat{H}}\right) = |\det(2\pi\mathbf{D})|^{-1}\int e^{-\frac{1}{2}(\mathbf{r}-\mathbf{q})^{\mathrm{T}}\mathbf{D}^{-1}(\mathbf{r}-\mathbf{q})}d\mathbf{r} \tag{17}$$

where

$$\hat{H} = \frac{1}{2}\nabla^{\mathrm{T}}\mathbf{M}^{-1}\nabla + \frac{1}{2}(\mathbf{r}-\mathbf{q})^{\mathrm{T}}\mathbf{K}(\mathbf{r}-\mathbf{q}) \tag{18}$$

is the Hamiltonian, with the minimum of the potential energy set equal to zero, $\beta = 1/k_{\mathrm{B}}T$ with k_{B} being Boltzmann's constant, $\mathbf{D}$ is the displacement-displacement correlation matrix (i.e., the distribution covariance matrix), $\mathbf{K}$ is the Hessian, $\mathbf{q}$ is the center of the Gaussian. The covariance matrix of the harmonic distribution is given by

$$\mathbf{D} = \mathbf{M}^{-1/2}d(\Omega)\mathbf{M}^{-1/2} \tag{19}$$

where $\mathbf{M}$ = diag(m_i) is the mass matrix, Ω = diag(ω_i) is the frequency matrix. For the classical harmonic partition function, the auxiliary function $d(\Omega)$ is defined as

$$d_{\mathrm{class}}(\omega_i;T) = \frac{k_{\mathrm{B}}T}{\omega_i^2} \tag{20}$$

while for the quantum harmonic partition function, it takes the following form

$$d_{\mathrm{quant}}(\omega_i;T) = \frac{\hbar}{2\omega_i}\coth\left(\frac{\hbar\omega_i}{2k_{\mathrm{B}}T}\right) \tag{21}$$

In Eqs. 20 and 21, $d_{\mathrm{class}}(\omega_i;T)$ and $d_{\mathrm{quant}}(\omega_i;T)$ describe the breadth of the corresponding harmonic distributions at temperature T along the *i*th normal mode.

MB-fit allows the user to generate the training and test sets by sampling molecular configurations using either the classical or the quantum harmonic distribution. Specifically, following Ref. 117, MB-fit samples a given (classical or quantum) harmonic distribution using inverse transform sampling that allows for sampling a normal distribution $\mathcal{N}(\mathbf{q},\mathbf{D})$, with mean $\mathbf{q}$ and covariance matrix $\mathbf{D}$, starting from an initial sequence of points uniform on $[0,1)^{3N}$. A transformation matrix is then constructed using the normal modes obtained from the mass-scaled Hessian which are obtained from the corresponding QM calculations. Since it was shown that effectively no differences are found when pseudorandom or quasirandom sequences are used to define the starting uniform distribution,[117] only the former is currently available in MB-Fit to generate the initial sequence of points. After transformation to $\mathcal{N}(\mathbf{q},\mathbf{D})$, these points correspond to unique molecular configurations that can thus be included in the training/test sets.

In sampling the classical and quantum harmonic distributions, the only free parameter to be chosen is the temperature, which effectively determines the range of molecular distortions that are included in the training/test sets. As discussed in Ref. 117, linear ($T_{i+1} - T_i$ = constant) or geometric (T_{i+1}/T_i = constant) temperature progressions can be used to efficiently sample both distributions. Ideally, the temperature range should be sufficiently wide to "excite" all normal modes of the monomer at the same time. Although this protocol often results in a maximum temperature that may be significantly higher than the temperature range usually explored in MD and MC simulations, highly distorted configurations generated at elevated temperatures guarantee that the 1B PIPs of the TTM-nrg and MB-nrg PEFs are well-behaved over a wide region of the configuration space. In this context, it should be noted that, since it primarily samples low-energy configurations, the classical harmonic distribution may lead to "holes" in the representation of $\varepsilon^{1\mathrm{B}}$ in regions of the multidimensional PES that are not properly represented in the training sets used to generate the corresponding PIPs. On the other hand, sampling exclusively with the quantum harmonic distribution may result in a sub-optimal representation of $\varepsilon^{1\mathrm{B}}$ in the minimum-energy regions of the multidimensional PES, especially for floppy molecules with several atoms. In these cases, it is thus recommended to supplement the training/test sets generated from sampling the classical harmonic distribution with configurations sampled with the corresponding quantum harmonic distribution.

It should also be noted that some complications may arise in generating training sets for "floppy" molecules since sampling high-frequency normal modes may cause low-frequency normal modes to break.[117] It was shown that this problem can be overcome by sampling each normal mode at a characteristic temperature directly related to its frequency, which effectively mimics the concept of the Einstein temperature introduced to model the heat capacity of crystals.[117] While this feature is currently not implemented in MB-Fit, it will become available in future releases.

For the training/test sets of the higher-body (>1B) terms of Eq. 1, MB-fit employs a general configuration generator that uses distance-based sampling with randomized rotations which is applicable to both rigid and flexible monomers. The latter can be sampled from the 1B training/test sets described above or re-generated as required with the normal-mode sampling algorithms. Configurations are automatically screened for inter-molecular atomic distances that fall below a predefined cutoff distance corresponding to a fraction (0.8) of the sum of the van der Waals radii of the two closest atoms on two monomers of the target n-body system. Molecules are randomly rotated using a quaternion-based algorithm.[118] However, the higher-body terms in Eq. 1 are, in general, associated with complicated, multidimensional energy landscapes. This implies that the random distance and rotation sampling may not always suffice for generating adequate 2B and 3B training sets for the MB-nrg PEFs. It is thus recommended to augment the 2B and 3B training sets by sampling from MD or MC simulations carried out under various temperature and pressure conditions.

Future releases of MB-Fit will include an active-learning approach to training set reduction which was shown to be effective in the development of representative training sets for ion–water MB-nrg PEFs.[107,119]

C. Quantum mechanical calculations

MB-Fit includes an interface that drives QM calculations in order to optimize molecular structures, perform normal-mode analysis, and compute molecular properties (e.g., atomic charges, atomic polarizabilities, and dispersion coefficients). MB-Fit supports running QM calculations locally or, alternatively, provides a job manager that generates short Python scripts for each nB energy calculation which can then be executed on HPC platforms or in a cloud or grid computing environment like Open Science Grid.[120] The details of job scheduling depend on the platform and are currently up to the user, but interfaces with common job schedulers will be included in future releases of MB-Fit. Once the calculations have completed, the job manager automatically parses QM data from output files and adds them to the database. The calculation of individual nB energies ($\varepsilon^{\mathrm{nB}}$ in Eq. 2) from QM outputs is fully automated. MB-Fit uses third-party software to perform QM calculations, currently supporting Q-Chem[121] and Psi4.[122] Extensions to other software are planned for future releases. Electronic structure calculations can be carried out at an arbitrary QM level of theory using an arbitrary basis set (among those available in Q-Chem and Psi4).

As described in Sec. II A, the TTM-nrg and MB-nrg PEFs adopt the same sets of atomic charges, atomic polarizabilities, and dispersion coefficients. While the user has complete freedom in selecting any method available to determine these quantities, it is recommended to calculate the atomic charges using the CM5 scheme,[123] and the atomic polarizabilities and dispersion coefficients using the exchange-hole dipole-moment model (XDM).[124–126] For the calculation of the individual nB energies MB-Fit provides the user with the option of correcting the BSSE using the counterpoise method.[127]

If users wish to use a "model" (i.e., a combination of QM method and basis set) not currently supported by Q-Chem or Psi4, or otherwise wishes to generate the necessary QM data in an alternative way, they are free to bypass MB-Fit and use another software for this step. The "model" and fitting procedure are independent of how the data is generated.

D. Implementation of the permutationally invariant polynomials

The C++ PIP evaluation functions and their analytical gradients are generated automatically with two main challenges in mind: 1) optimization of generation, and 2) optimization of evaluation. MB-fit adopts a dynamic programming algorithm to address the first challenge, while relying on optimization features provided by common C++ compilers (e.g., GCC[128] (GNU license) or ICPC[129] (Intel)) and Maple[114] to address the second challenge. Before compilation, Maple can optionally be applied to factorize the polynomials which reduces the number of floating point operations needed for evaluation.

Generation of PIPs up to arbitrary degree is supported, though polynomials of high degree may be excessively large for use in actual MD and MC simulations, depending on the available computational resources. Optional filtering of polynomials to exclude specific terms based on a number of factors (e.g., degree of certain variables, inter/intra-molecular character, etc.) is also available. A detailed description of the protocol used by MB-fit for the implementation of the PIPs is reported in the Supplementary Material.

E. Parameterization and training

As described in Sec. II A, both MB-PEFs adopt the same representation of the 1B term, $\varepsilon^{1\mathrm{B}}$, which is expressed by a PIP. In practice, for a given set of non-linear parameters entering the expressions of the corresponding monomials (Eqs. 14), which are obtained using the simplex algorithm, the linear coefficients of $V_{\mathrm{PIP}}^{1\mathrm{B}}$ are obtained from a linear least-squares fit (see section II C) by fitting

$$y_{\mathrm{ref}}(k) = \varepsilon_{\mathrm{ref}}^{1\mathrm{B}}(k) \tag{22}$$

where $\varepsilon_{\mathrm{ref}}^{1\mathrm{B}}(k)$ is the 1B QM reference energy for the k^{th} configuration in the 1B training set, with $\varepsilon_{\mathrm{ref}}^{1\mathrm{B}} = 0$ for the optimized geometry of the monomer.

In the case of the 2B energy, A_{ij}, b_{ij} and δ_{ij} of the the TTM-nrg PEFs in Eqs. 5 and 8 are fitting parameters, with $b_{ij} = \delta_{ij}$ by construction. For a given set of non-linear parameters (i.e., b_{ij}) from each simplex step, the linear parameter A_{ij} in Eq. 5 is obtained by fitting

$$y_{\mathrm{ref}}(k) = \varepsilon_{\mathrm{ref}}^{2\mathrm{B}}(k) - V_{\mathrm{elec}}(k) - V_{\mathrm{pol}}(k) - V_{\mathrm{disp}}^{2\mathrm{B}}(k) \tag{23}$$

where $\varepsilon_{\mathrm{ref}}^{2\mathrm{B}}(k)$ is the 2B QM reference energy for the k^{th} configuration in the 2B training set, with $V_{\mathrm{elec}}(k)$, $V_{\mathrm{pol}}(k)$, and $V^{\mathrm{2B}}_{\mathrm{disp}}(i)$ describing permanent electrostatics, polarization, and dispersion energy, respectively (Eqs. 3-8).

The MB-nrg PEFs use the same classical electrostatic model as the TTM-nrg PEFs, and include explicit representations of short-range 2B and 3B energies in terms of PIPs. In the case of the 2B energy (Eq. 4), MB-Fit provides the user with the option of fitting $V^{\mathrm{2B}}_{\mathrm{sr,\,MB\text{-}nrg}}$ in Eq. 6 with or without including $V^{\mathrm{2B}}_{\mathrm{sr,\,TTM\text{-}nrg}}$ in Eq. 5 as a baseline potential. For a given set of non-linear parameters entering the expressions of the corresponding monomials (Eqs. 14), the linear coefficients of $V^{\mathrm{2B}}_{\mathrm{PIP}}$ in Eq. 6 are thus obtained by fitting

$$y_{\mathrm{ref}}(k) = \varepsilon^{\mathrm{2B}}_{\mathrm{ref}}(k) - V_{\mathrm{elec}}(k) - V_{\mathrm{pol}}(k) - V^{\mathrm{2B}}_{\mathrm{disp}}(k) - \alpha V^{\mathrm{2B}}_{\mathrm{sr,\,TTM\text{-}nrg}}(k) \tag{24}$$

where $\alpha = 1$ when including the TTM-nrg Born-Mayer repulsion term $V^{\mathrm{2B}}_{\mathrm{sr,\,TTM\text{-}nrg}}$, otherwise $\alpha = 0$. Fitting of $V^{\mathrm{2B}}_{\mathrm{sr,\,MB\text{-}nrg}}$ over a baseline TTM-nrg potential is optional although recommended since it significantly reduces the probability of “holes” in $V^{\mathrm{2B}}_{\mathrm{PIP}}$.

In the case of the 3B energy (Eq. 10), for a given set of non-linear parameters entering the expressions of the corresponding monomials (Eqs. 14), the linear coefficients of $V^{\mathrm{3B}}_{\mathrm{PIP}}$ in Eq. 11 are obtained by fitting

$$y_{\mathrm{ref}}(k) = \varepsilon^{\mathrm{3B}}_{\mathrm{ref}}(k) - V_{\mathrm{pol}}(k) \tag{25}$$

where $\varepsilon^{\mathrm{3B}}_{\mathrm{ref}}(k)$ is the 3B QM reference energy for the kth configuration in the 3B training set.

F. Visualization and analysis tools

Once a TTM-nrg or MB-nrg PEF has been obtained, MB-Fit provides tools to retrieve the associated root-mean-square deviations (RMSDs) as well as the corresponding correlation plots for both training and test sets. Different customization options for the visualization are available, and the data used in the graphs are written as a data file for further inspection and visualization with external plotting programs. In the correlation plots, the reference QM nB energies are reported on the x-axis, and the corresponding TTM-nrg or MB-nrg values are reported on the y-axis.

G. Interface to MBX

MB-Fit provides an automated C++ code generator that enables the use of the TTM-nrg and MB-nrg PEFs in MBX, our many-body energy/force calculator.[108] Specifically, MB-Fit provides all the pieces of code that are needed by MBX and, if the location of the MBX software is provided by the user, it automatically adds them to MBX with no need of action by the user. MBX is currently interfaced with LAMMPS[109] and i-PI,[110] which thus allows the user to perform MD simulations with both TTM-nrg and MB-nrg PEFs in all common thermodynamic ensembles. Enhanced-sampling simulations and free-energy calculations are possible in both LAMMPS and i-PI through the interface with PLUMED.[130,131]

H. Availability and documentation

The MB-Fit software is freely available on Github.[132] Unit tests and regression tests ensure correctness of the software. Extensive documentation is provided in form of Jupyter notebooks that walk the user through the generation of TTM-nrg and MB-nrg PEFs, including details of the background theory. The user is strongly encouraged to refer to the Jupyter notebooks to get started building TTM-nrg and MB-nrg PEFs with MB-Fit.

IV. EXAMPLE: TTM-NRG AND MB-NRG PEFS FOR AMMONIA

Ammonia (NH_3) has been one of the most important industrial chemicals since the development of the Haber-Bosch process. NH_3 is widely used in the fertilizer and cleaning industries as well as in synthetic chemistry where it is the most common source of nitrogen.[133] Since the interactions between NH_3 molecules include all typical contributions (i.e., Pauli repulsion, permanent and induced electrostatics, hydrogen bonding, and dispersion), ammonia serves as an ideal test case to illustrate the workflow of the MB-fit software as well as the ability of MB-Fit to generate TTM-nrg and MB-nrg PEFs at an arbitrary QM level of theory which are fully transferable from the gas to the condensed phase. To this end, we present two sets of TTM-nrg and MB-nrg PEFs developed at the DF-FNO-CCSD(T)[134] and PBE0-D3(BJ)[135,136] levels of theory, respectively, combined with the aug-cc-pVTZ basis set.[137] Since these MB-PEFs primarily serve as a showcase for the MB-Fit ability to seamlessly generate transferable, many-body representations of molecular interactions and not for quantitative analyses of the properties of ammonia, the MB-nrg PEFs are constructed without including ε^{3B} in Eqs. 10 and 11.

The same 1B and 2B training and test sets were used for both sets of MB-PEFs, which allows for analyzing the relative accuracy of TTM-nrg and MB-nrg PEFs trained on two different

QM levels of theory, i.e., DF-FNO-CCSD(T) and PBE0-D3(BJ). Independent of the QM level of theory, the two sets of MB-PEFs share the same representations of V_{pol}, $V_{\mathrm{disp}}^{\mathrm{2B}}$, and V_{elec}. Specifically, the atomic charges were calculated with the CM5 method,[123] while the atomic polarizabilities and dispersion coefficients were calculated using the exchange-hole dipole-moment model (XDM).[124–126] Because it is currently not possible to perform CM5 and XDM calculations at the DF-FNO-CCSD(T) level of theory, in order to guarantee the same representations of V_{pol}, $V_{\mathrm{disp}}^{\mathrm{2B}}$, and V_{elec} for the two sets of TTM-nrg and MB-nrg PEFs, the CM5 and XDM calculations for all MB-PEFs were carried out with the ωB97M-V functional since it was shown to consistently provide the closest agreement with CCSD(T) data for various molecular interactions.[57,93,96,138–141] Both the CM5 and XDM calculations were carried out with Q-Chem v5.1[121] using the aug-cc-pVTZ[137] basis set. All 1B and 2B energies were calculated using Psi4[122] at the corresponding QM level of theory, including counterpoise correction for the BSSE.

A. One-body PEF

As discussed in Sec. II A, the TTM-nrg and MB-nrg PEFs adopt the same functional form for the 1B energies. The 1B configurations for the training and test sets were obtained from normal-mode sampling using a piece-wise distribution over temperature provided by MB-Fit. The distribution was constructed relative to the temperature (T_{max}) corresponding to the highest normal-mode frequency (ν_{max}) of the system (i.e., an isolated NH_3 molecules) as $T = \hbar \nu_{\mathrm{max}}/k_{\mathrm{B}}$. Specifically, 5%, 40%, 30%, 20%, and 5% of the total number of 1B configurations for the training (4098 configurations) and test (512 configurations) sets were generated from the corresponding classical harmonic distributions sampled at temperatures equal to $T_{\mathrm{max}}/100$, $T_{\mathrm{max}}/20$, $T_{\mathrm{max}}/10$, $T_{\mathrm{max}}/5$, and $T_{\mathrm{max}}/2$, respectively. Half of the configurations were generated based on the optimized minimum and umbrella inversion transition state structures, respectively. A fifth degree PIP, with six different exponential variables $\xi^{\mathrm{exp}}(R)$ corresponding to all the possible distances between pairs of atoms, containing 102 symmetrized terms, was fitted to the QM data for $\varepsilon^{\mathrm{1B}}$. The unweighted RMSDs of the 1B training set for configurations below 25 kcal/mol are 0.0967 and 0.0742 kcal/mol for two sets of MB-PEFs derived from DF-FNO-CCSD(T) and PBE0-D3(BJ) 1B energies, respectively. The correlation plots for the test sets shown in Fig. 2 demonstrate that the 1B PIPs are able to accurately reproduce the reference data over a wide range of 1B energies, independently of the QM level of theory. To assess the smoothness of the 1B PEFs, a relaxed scan along the umbrella motion

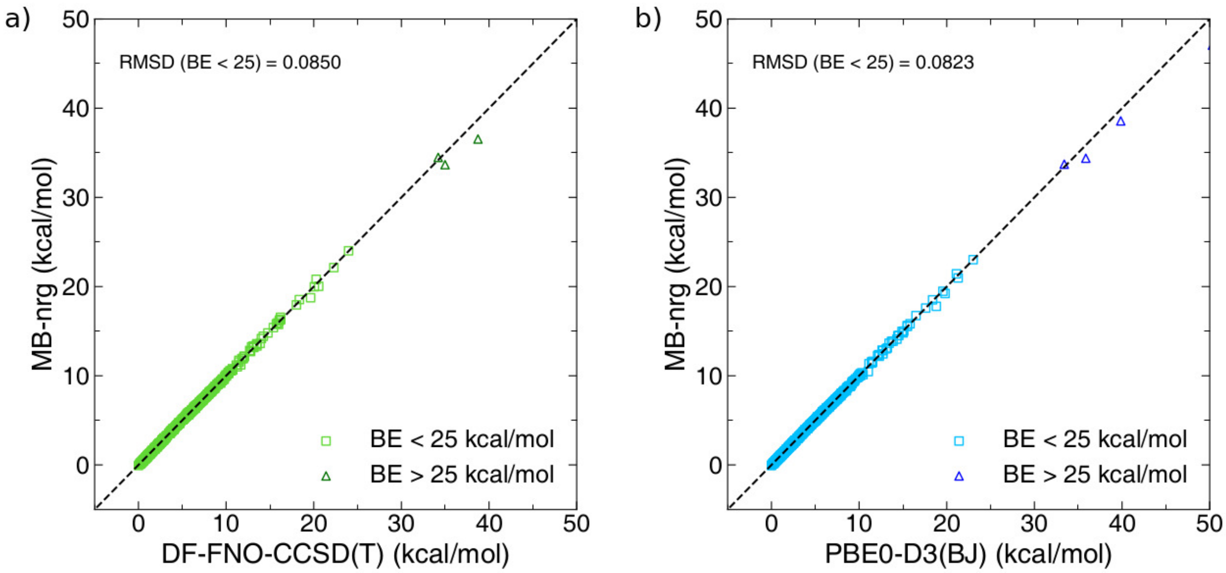

FIG. 2. Correlation plots between the reference 1B energies and the corresponding MB-nrg values calculated for the NH_3 test set. The reference 1B energies were calculated using DF-FNO-CCSD(T)/aug-cc-pVTZ (panel a) and PBE0-D3(BJ)/aug-cc-pVTZ (panel b). The TTM-nrg PEFs have the same 1B term as the corresponding MB-nrg PEFs. See main text for details.

of an isolated NH_3 molecule was performed at both DF-FNO-CCSD(T) and PBE0-D3(BJ) levels of theory. Fig. 3 shows that the 1B PEFs are able to quantitatively reproduce the corresponding reference data.

B. Two-body PEF

The 2B training (597 configurations) and test (200 configurations) sets for the TTM-nrg PEFs were obtained from scans along the distance between the two N atoms of the NH_3–NH_3 dimer, applying random rotations to the two molecules at each distance while keeping their geometries fixed at the corresponding optimized structure of an isolated NH_3 molecule. The RMSDs for the TTM-nrg training sets are 1.0424 and 1.0451 kcal/mol for the two MB-PEFs fitted to DF-FNO-CCSD(T) and PBE0-D3(BJ) 2B energies, respectively. The correlation plots for the test sets in Fig. 4 show that both TTM-nrg PEFs are able to semi-quantitatively reproduce the corresponding reference 2B energies, with an accuracy which is independent of the QM level of theory.

As shown in previous studies,[92,95,97–99] higher accuracy in modeling 2B energies is achieved by the MB-nrg PEFs that adopt short-range PIPs to effectively represent 2B quantum-mechanical in-

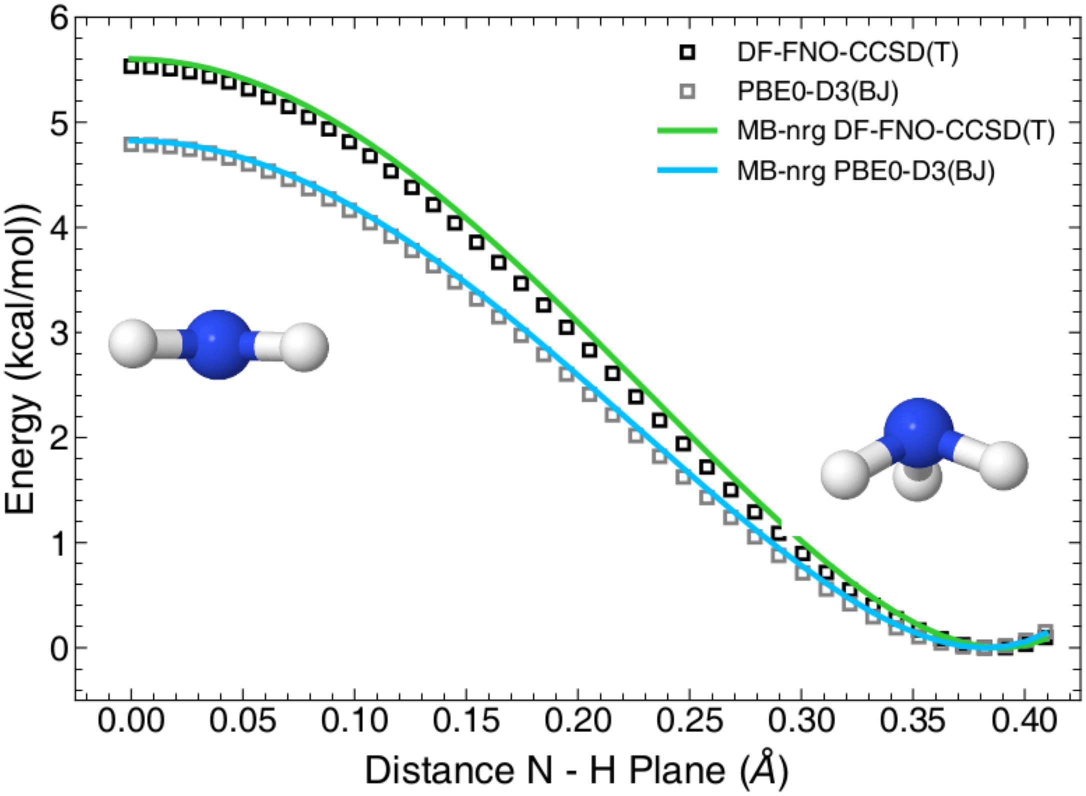

FIG. 3. Distortion energy calculated along a relaxed scan of the NH_3 umbrella motion as a function of the distance between the nitrogen atom and the plane defined by the three hydrogen atoms. Both DF-FNO-CCSD(T)/aug-cc-pVTZ (black) and PBE0-D3(BJ)/aug-cc-pVTZ (grey) reference 1B energies along with the corresponding MB-nrg values are shown. The TTM-nrg PEFs have the same 1B term as the corresponding MB-nrg PEFs. See main text for details.

teractions that arise from the overlap of the monomers' electron densities (e.g., Pauli repulsion, and charge transfer and penetration).[58,59,61,92,95,97,98,103,105,139] The 2B training (7261 configurations) and test (1449 configurations) sets for the MB-nrg PEFs were generated by including configurations extracted from three different sources. The first source was normal-mode sampling of the NH_3–NH_3 optimized dimer which was carried out adopting the same protocol described above for the 1B sets. The second source was scans along the N–N distance of the NH_3–NH_3 dimer with rigid NH_3 molecules as described above for the 2B TTM-nrg training and test sets. The third source was scans along the N–N distance of the NH_3–NH_3 dimer using distorted NH_3 configurations extracted from the 1B sets instead of rigid NH_3 molecules. The present MB-nrg PEFs for ammonia use 2B PIPs up to the fourth degree, which include 3, 23, 159, and 930 symmetrized

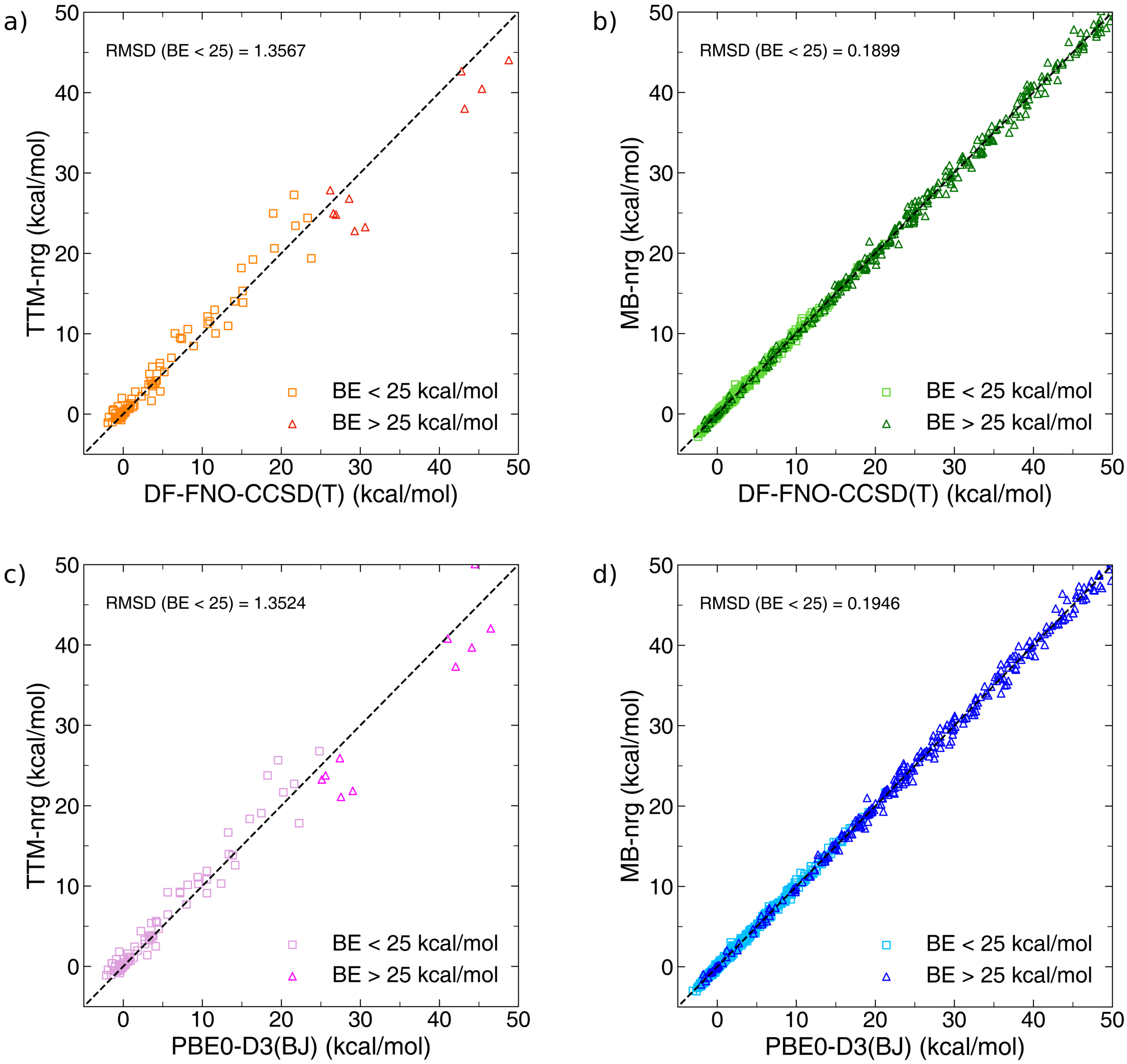

FIG. 4. Correlation plots between the reference 2B energies data and the corresponding TTM-nrg (panels a and c) and MB-nrg (panels b and d) values calculated for the NH_3–NH_3 2B test set. The reference 2B energies were calculated using DF-FNO-CCSD(T)/aug-cc-pVTZ (panels a and b) and PBE0-D3(BJ)/aug-cc-pVTZ (panels c and d). Also shown are the corresponding RMSDs. See main text for details.

terms of degree 1, 2, 3, and 4, respectively. The 2B PIPs contain 28 different variables corresponding to all possible intra- and inter-molecular distances between atoms of the NH_3–NH_3 dimer, all described with the $\xi^{\exp}(R)$ functional form. The unweighted RMSDs for the MB-nrg training sets are 0.1524 and 0.1542 kcal/mol for the two MB-nrg PEFs fitted to DF-FNO-CCSD(T) and PBE0-D3(BJ) 2B energies, respectively. Fig. 4 shows the correlation plots for the corresponding test sets

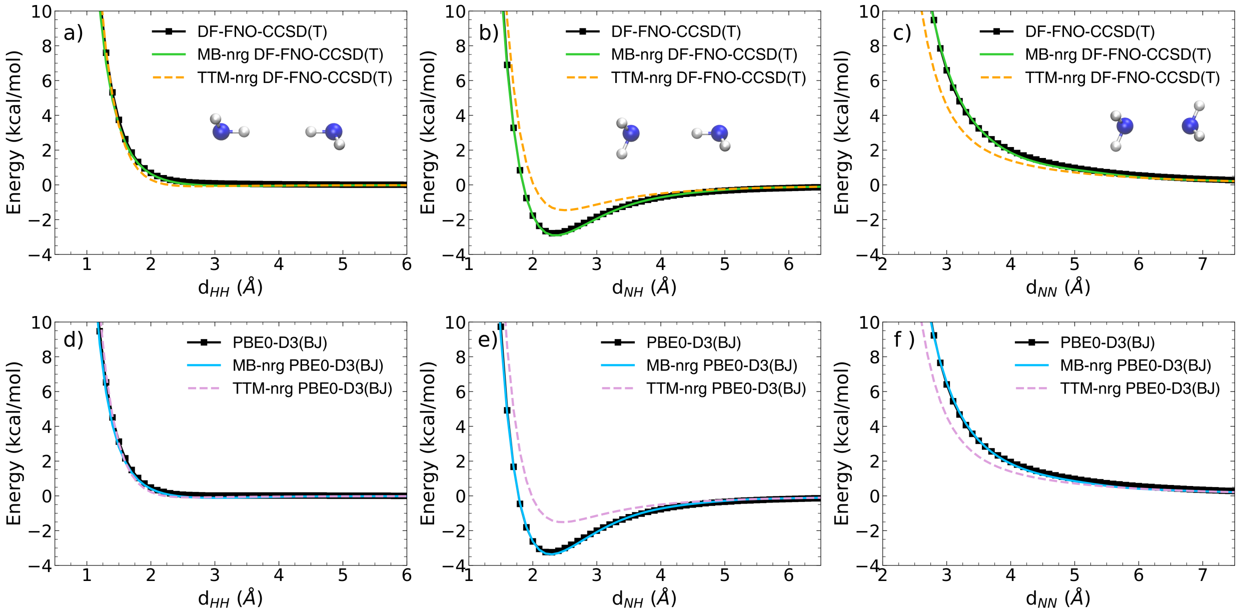

FIG. 5. Interaction energy scans along the H-H (panels a and d), H-N (panels b and e), and N-N (panels c and f) distances between the two monomers in the NH_3–NH_3 dimer. The reference interaction energies calculated using DF-FNO-CCSD(T)/aug-cc-pVTZ along with the corresponding TTM-nrg and MB-nrg values are shown in the top panels, while the reference interaction energies calculated using PBE0-D3(BJ)/aug-cc-pVTZ along with the corresponding TTM-nrg and MB-nrg values are shown in the bottom panels. See main text for details.

which, as previously found for other molecular systems,[97,98] demonstrate that the MB-nrg PEFs are able to quantitatively reproduce QM reference data, independently of the QM level of theory. It should be noted that the accuracy of the MB-nrg PEFs can be improved by increasing the degree of the PIPs and/or applying filters on terms containing variables that involve interatomic distances that are found to be less relevant for the representation of the underlying PES.

The smoothness of the 2B PEFs is assessed by performing three scans along the H-H, H-N, and N-N distances between the two monomers in the NH_3–NH_3 dimer. Fig. 5 shows the orientations of the monomers and performance of the four MB-PEFs on each of the scans relative to the QM reference data, providing further evidence for the ability of the MB-nrg PEFs to reproduce arbitrary QM reference data. On the other hand, the TTM-nrg PEFs display well known deficiencies that are common to all PEFs purely based on classical polarization.[142,143] This results in only a qualitative agreement between the TTM-nrg and QM 2B energies, with the TTM-nrg accuracy being particularly sensitive to the relative orientation of the two NH_3 molecules.

C. Many-body energies in ammonia clusters

While the analyses presented in Fig. 2 and 5 assess the ability of the TTM-nrg and MB-nrg PEFs to reproduce 1B and 2B energies that were the target of the training process, one of the greatest challenges for ML PEFs is to preserve the same accuracy for many-body energies and/or molecular systems that are not included in the training sets. Combining explicit data-driven representations of short-range low-order interactions with implicit (mean-field-like) many-body representations of high-order and long-range interactions, it has been shown that the MB-pol PEF[58–61,142] for water as well as the MB-nrg PEFs for ions in water[92,95,107,141] and various molecular fluids[97,98] are able to correctly reproduce each individual term of the MBE of Eq. 1. Relatively large deviations were instead observed for the TTM-nrg PEFs.[91,94,97,98]

To assess how the different many-body terms in the MBE for ammonia are represented by the TTM-nrg and MB-nrg PEFs, we performed many-body decompositions of the lowest-energy isomers of the $(NH_3)_n$ clusters, with $n = 2,3,4$. Fig. 6 shows the deviations ΔE per fragment associated with each nB term of the TTM-nrg and MB-nrg PEFs relative to the corresponding reference values (dashed line) which were calculated using the SAMBA approach.[144] Specifically, the 1B and 2B reference energies were calculated at the CCSD(T) level of theory using a two-point extrapolation between the aug-cc-pVTZ and aug-cc-pVQZ basis sets. The 3B reference energies were calculated at the CCSD(T)/aug-cc-pVTZ level of theory with a cluster counterpoise correction for the BSSE, while the 4B reference energies were calculated at the CCSD(T)/aug-cc-pVTZ level of theory. Also shown in Fig. 6 are the deviations calculated with DF-FNO-CCSD(T) and PBE0-D3(BJ) which allow for assessing the ability of the corresponding TTM-nrg and MB-nrg PEFs to reproduce the target energies as well as for quantifying the relative accuracy of the different models in reproducing the reference energies. The deviations are calculated as $\Delta E = \frac{1}{n}\sum_i^n \Delta E_i$, where n is the number of monomers, dimers, trimers, etc. in a given cluster, and ΔE_i is the the individual signed error of the i^{th} fragment. In summary, Fig. 6 shows the average error per monomer, dimer, trimer and tetramer for the three cluster sizes analyzed. While the 1B deviations from the CCSD(T)/SAMBA reference values are negligible for all models, significant errors are associated with the TTM-nrg representations of 2B energies. In contrast, the MB-nrg 2B energies closely reproduce the corresponding DF-FNO-CCSD(T) and PBE0-D3(BJ) target values for all three clusters. Importantly, while the 2B energies calculated with DF-FNO-CCSD(T) and the corresponding MB-nrg PEF are in close agreement with the CCSD(T)/SAMBA reference values,

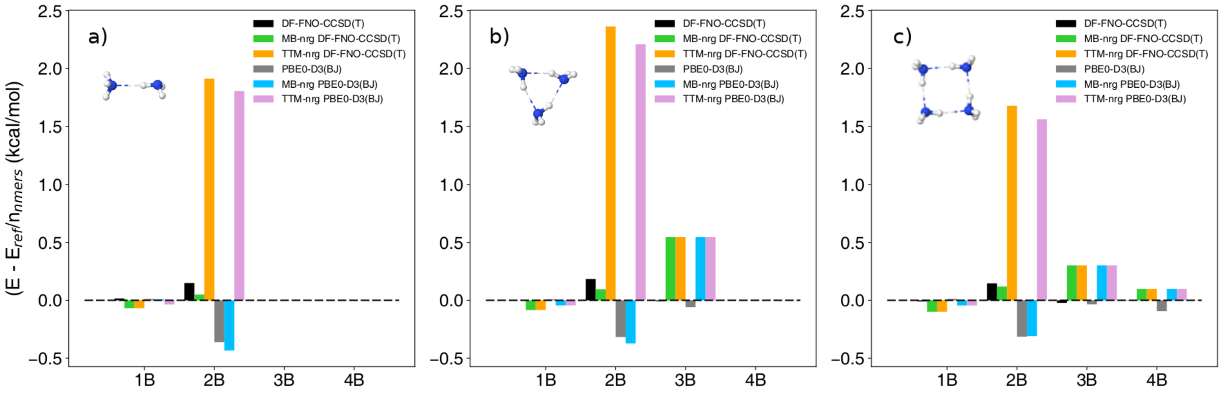

FIG. 6. Energy deviations per fragment from the reference CCSD(T)/SAMBA values for each individual many-body contribution to the interaction energies of $(NH_3)_n$ clusters, with n=2 (panel a), n=3 (panel b) and n=4 (panel c) calculated using DF-FNO-CCSD(T) and PBE0-D3(BJ), and the corresponding MB-nrg and TTM-nrg PEFs. See main text for details.

appreciable deviations are associated with 2B energies calculated using PBE0-D3(BJ) and the corresponding MB-nrg PEF. It should be noted that the relatively large deviations associated with the representation of 3B energies based on classical polarization which is adopted by the TTM-nrg and MB-nrg PEFs suggests that the inclusion of explicit 3B PIPs may be needed for a more quantitative description of 3B interactions. The analyses reported in Fig. 6 demonstrate that classical polarization is instead able to quantitatively reproduce the 4B energies.

D. Second virial coefficient

While the analysis of individual many-body energies allow for a general assessment of the ability of any PEF to describe the underlying molecular interactions, these quantities are not amenable to direct measurements. However, the interplay of many-body interactions directly determines the free-energy landscape that effectively determines structural, thermodynamic, and dynamical properties of any molecular system at finite temperature, which can be measured experimentally. Since these properties can be calculated using computer simulations and be directly related to the underlying molecular interactions using statistical mechanics principles, it follows that comparisons between measured and calculated properties provide an effective means to assess the ability of a PEF to realistically describe the molecular system of interest.

In this context, a direct probe of the overall 2B energy landscape is provided by the second virial coefficient, $B_2(T)$ given by

$$B_2(T) = -2\pi \int_0^\infty \left(\left\langle e^{-\frac{\varepsilon^{2B}(R)}{k_B T}} \right\rangle - 1 \right) R^2 dR \tag{26}$$

Here, ε^{2B} is the 2B energy of Eq. 1, k_B is Boltzmann's constant, R is the distance between the two monomer centers of mass in a given dimer configuration, and T is the temperature. We calculated $B_2(T)$ for ammonia by numerically solving the integral in Eq. 26 using the trapezoidal rule with an integration step $\Delta R = 0.05$ Å, and 120,000 dimer configurations generated via Monte Carlo sampling for each radial grid point. Fig. 7 shows the virial coefficient as a function of the temperature, calculated with both sets of TTM-nrg and MB-nrg PEFs trained on DF-FNO-CCSD(T) and PBE0-D3(BJ) data. These comparisons indicate that the TTM-nrg PEFs perform similarly, independently of the QM level of theory. This is in line with the analysis of the many-body energies presented in Fig. 5 which shows that, although PBE0-D3(BJ) predicts 2B energies that deviates appreciably from the CCSD(T) reference values, the functional form adopted by the TTM-nrg PEFs is too simple for quantitatively capturing these differences. On the other hand, these differ-

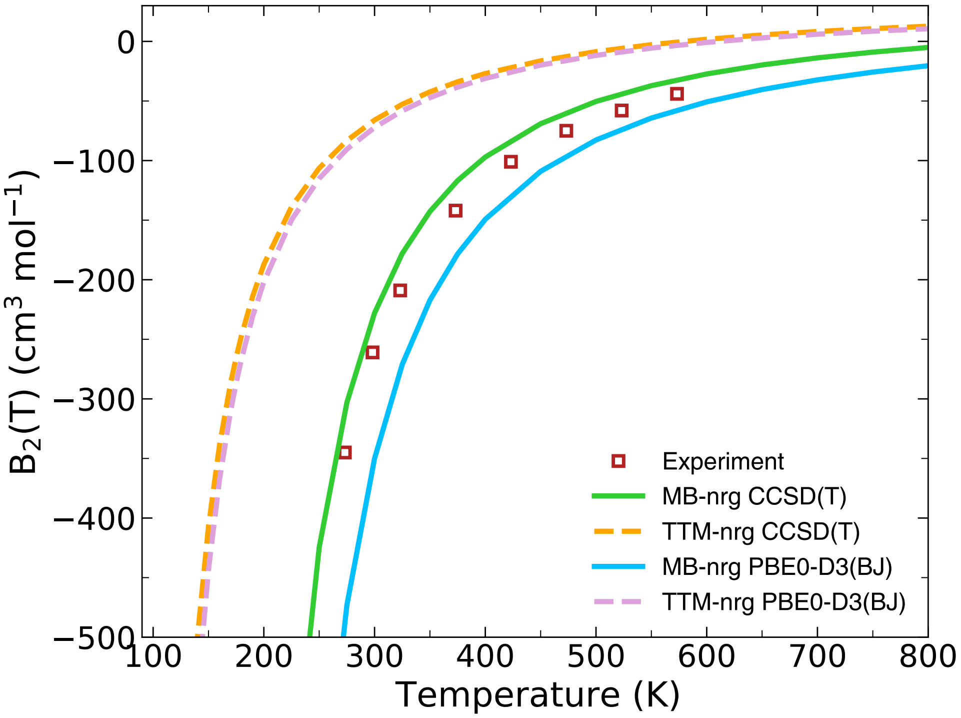

FIG. 7. Comparison between the experimental NH_3–NH_3 second virial coefficients and the corresponding values calculated with the TTM-nrg and MB-nrg PEFs trained on DF-FNO-CCSD(T) and PBE0-D3(BJ) data. The experimental data are from Ref. 145.

ences become apparent in the MB-nrg calculations of $B_2(T)$ which clearly show that the MB-nrg PEF trained on DF-FNO-CCSD(T) data closely reproduces the available experimental values over the entire temperature range. The larger $B_2(T)$ values predicted by both TTM-nrg PEFs can be traced back to the inability of these MB-PEFs to correctly reproduce the attractive region of the corresponding 2B energy landscape (Fig. 5). The lower $B_2(T)$ values obtained with the MB-nrg PEF trained on PBE0-D3(BJ) data instead are directly related to PBE0-D3(BJ) predicting overly attractive 2B energies as shown in Fig. 5.

E. Structure of liquid ammonia

Finally, we used classical MD simulations to investigate the structure of liquid ammonia. While the present MD simulations are not meant to provide a comprehensive analysis of the liquid properties, they can be used to assess the transferability of the TTM-nrg and MB-nrg PEFs from the gas to the condensed phase. To this purpose, classical MD simulations were carried out with the

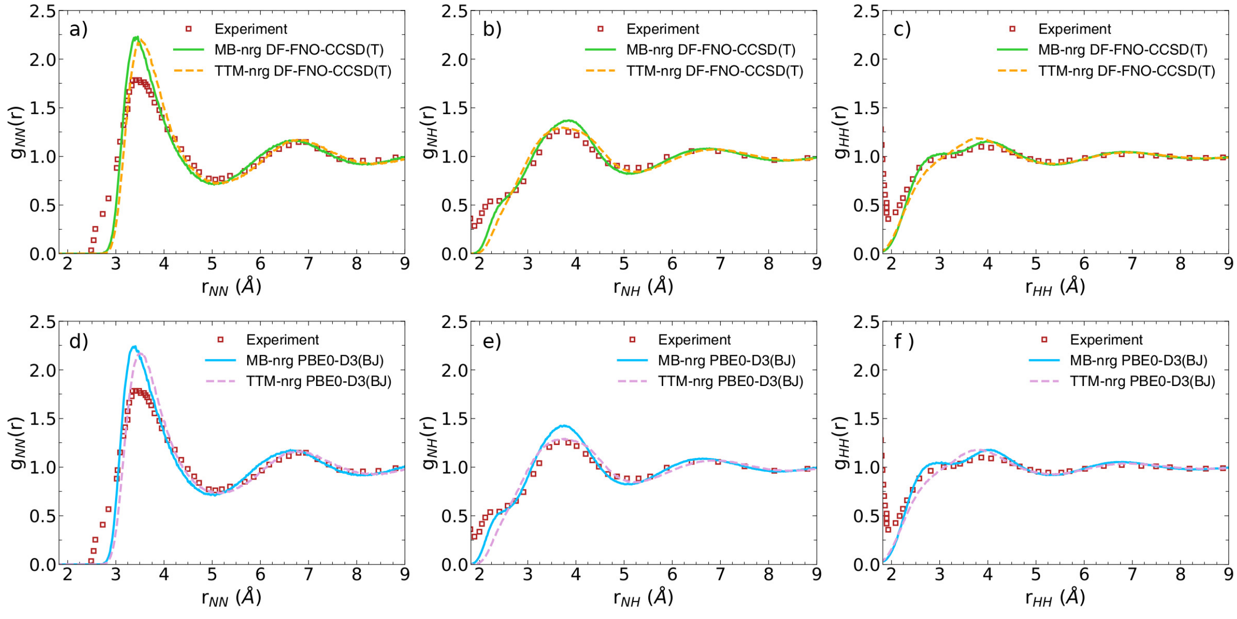

FIG. 8. Comparison between experimental and simulated nitrogen-nitrogen (g_{NN}, left panels), nitrogen-hydrogen (g_{NH}, central panels), and hydrogen-hydrogen (g_{HH}, right panels) radial distribution functions (RDFs) of liquid ammonia at 273 K. RDFs calculated using the MB-nrg and TTM-nrg PEFs trained on DF-FNO-CCSD(T) data are shown in the top panels, while those calculated with the corresponding MB-PEFs trained on PBE0-D3(BJ) data are shown in the bottom panels. The experimental RDFs are from Ref. 146.

MBX software[108] interfaced with i-PI.[110] All simulations were carried out in the canonical (NVT) ensamble at a temperature of 273 K using a cubic box of length 23.084 Å with 279 NH_3 ammonia molecules, which corresponds to the liquid density used in the experimental measurements of Ref. 146. The initial configuration was generated using Packmol,[147] and the system was equilibrated for 50 ps before a production run of 100 ps, using a time step of 0.2 fs. The temperature was controlled by a global Langevin thermostat.

Fig. 8 shows that all MB-PEFs provide nearly quantitative agreement with the experimental N-N, N-H, and H-H radial distribution functions (RDFs). In particular, the location of the different solvation shells is accurately predicted, while all MB-PEFs predict higher first peaks in the N–N and N–H RDFs. Various reasons may be responsible for these differences between the experimental and calculated RDFs: 1) intrinsic inaccuracies in the QM data used in the training of the TTM-nrg and MB-nrg PEFs, 2) inaccuracies in the TTM-nrg and MB-nrg representations of individual many-body terms of the MBE, 3) neglect of the explicit 3B term in the MB-nrg PEF, and 4) neglect of nuclear quantum effects. It is interesting to note that the differences that were apparent in the analysis of individual many-body energies of $(NH_3)_n$ clusters (Fig. 6) appear to be washed out in the MD simulations of the liquid phase, with all MB-PEFs performing similarly. In this context, it should be noted that our previous studies of various molecular fluids[97,98] show that the differences between different TTM-nrg and MB-nrg PEFs are somewhat suppressed in MD simulations carried out in the NVT ensemble but can lead to qualitatively different liquid structures and phase behavior when the simulations are performed in the isobaric-isothermal (NPT) ensemble. While these are certainly important aspects to investigate, they go beyond the scope of the present study and will be the subject of future studies.

V. CONCLUSIONS

We have introduced MB-Fit, an integrated software infrastructure that enables the automated development of fully transferable, data-driven MB-PEFs for generic molecules within the TTM-nrg and MB-nrg theoretical/computational frameworks. MB-Fit provides a complete array of tools to: 1) generate training and test sets for individual many-body energies, 2) set up and perform the required QM calculations of the necessary training data, 3) optimize both linear and non-linear parameters entering the mathematical expressions for the TTM-nrg and MB-nrg PEFs, and 4) generate the associated codes that are directly exported to the MBX energy/force calculator[108] that enables MD simulations with the TTM-nrg and MB-nrg PEFs using LAMMPS[109] and i-PI[110]. Given the demonstrated accuracy of the MB-pol PEF for water,[58–61] and the TTM-nrg and MB-nrg PEFs for ions in water[92,95,100–105,107] and various molecular fluids,[97,98] we believe that MB-Fit can open the door to routine, predictive computer simulations of small molecules in the gas, liquid, and solid phases, including, but not limited to, the modeling of molecular clusters, solvation structure and thermodynamics, heterogeneous processes at air/liquid and air/solid interfaces, molecular crystals, and phase diagrams.

VI. SUPPLEMENTARY MATERIAL

The supplementary material includes details on the implementation of the permutationally invariant polynomials in MB-Fit.

VII. ACKNOWLEDGEMENT

Different aspects of this work were supported by the National Science Foundation through grant no. ACI-1642336 (overall software development and implementation), the U.S. Department of Energy through grant no. DE-SC0019490 (generalization of the TTM-nrg and MB-nrg PEFs), and the Air Force Office of Scientific Research through grant no. FA9550-16-1-0327 (training/test set generation and PIP optimization). EBV was partially supported by a Halicioglu Data Science Institute Undergraduate Fellowship Award. This research used resources of the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by the National Science Foundation (Grant ACI-1548562),[148] the National Energy Research Scientific Computing Center (NERSC), which is supported by the Office of Science of the U.S. Department of Energy under Contract DE-AC02-05CH11231, as well as the Triton Shared Computing Cluster (TSCC) at the San Diego Supercomputer Center. We thank Sandra Brown for helpful discussions on normal mode sampling, Arun Kumar and Subhasis Dasgupta for helpful discussions on database design and Kaushik Ganapathy and Derek Tse for contributions to the initial software base.

VIII. DATA AVAILABILITY

Any data generated and analyzed in this study are available from the authors upon request. The TTM-nrg and MB-nrg PEFs used in this study are available in our open-access MBX software that can be downloaded from http://paesanigroup.ucsd.edu/software/mbx.html.

REFERENCES

[1] D. Frenkel and B. Smit, *Understanding Molecular Simulation: From Algorithms to Applications*, Vol. 1 (Elsevier, 2001).

[2] M. Tuckerman, *Statistical Mechanics: Theory and Molecular Simulation* (Oxford university press, 2010).

[3] W. F. Van Gunsteren and H. J. Berendsen, "Computer simulation of molecular dynamics: Methodology, applications, and perspectives in chemistry," Angew. Chem. Int. Ed. **29**, 992–1023 (1990).

[4] K. Binder, *Monte Carlo and Molecular Dynamics Simulations in Polymer Science* (Oxford University Press, 1995).

[5] A. Warshel, "Computer simulations of enzyme catalysis: Methods, progress, and insights," Annu. Rev. Biophys. Biomol. Struct. **32**, 425–443 (2003).

[6] M. Karplus and J. Kuriyan, "Molecular dynamics and protein function," Proc. Natl. Acad. Sci. U.S.A. **102**, 6679–6685 (2005).

[7] J. D. Durrant and J. A. McCammon, "Molecular dynamics simulations and drug discovery," BMC Biol. **9**, 1–9 (2011).

[8] K. Ohno, K. Esfarjani, and Y. Kawazoe, *Computational Materials Science: From Ab Initio to Monte Carlo Methods* (Springer, 2018).

[9] S. Lifson and A. Warshel, "Consistent force field for calculations of conformations, vibrational spectra, and enthalpies of cycloalkane and n-alkane molecules," J. Chem. Phys. **49**, 5116–5129 (1968).

[10] A. Warshel and S. Lifson, "Consistent force field calculations. II. Crystal structures, sublimation energies, molecular and lattice vibrations, molecular conformations, and enthalpies of alkanes," J. Chem. Phys. **53**, 582–594 (1970).

[11] A. K. Rappé, C. J. Casewit, K. Colwell, W. A. Goddard III, and W. M. Skiff, "UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations," J. Am. Chem. Soc. **114**, 10024–10035 (1992).

[12] T. A. Halgren and W. Damm, "Polarizable force fields," Curr. Opin. Struct. Biol. **11**, 236–242 (2001).

[13] A. D. MacKerell Jr, “Empirical force fields for biological macromolecules: Overview and issues,” J. Comput. Chem. **25**, 1584–1604 (2004).

[14] P. S. Nerenberg and T. Head-Gordon, “New developments in force fields for biomolecular simulations,” Curr. Opin. Struct. Biol. **49**, 129–138 (2018).

[15] J. A. Harrison, J. D. Schall, S. Maskey, P. T. Mikulski, M. T. Knippenberg, and B. H. Morrow, “Review of force fields and intermolecular potentials used in atomistic computational materials research,” Appl. Phys. Rev **5**, 031104 (2018).

[16] J. Behler, “Perspective: Machine learning potentials for atomistic simulations,” J. Chem. Phys. **145**, 170901 (2016).

[17] V. L. Deringer, M. A. Caro, and G. Csányi, “Machine learning interatomic potentials as emerging tools for materials science,” Adv. Mater. **31**, 1902765 (2019).

[18] F. Noé, A. Tkatchenko, K.-R. Müller, and C. Clementi, “Machine learning for molecular simulation,” Annu. Rev. Phys Chem. **71**, 361–390 (2020).

[19] T. Mueller, A. Hernandez, and C. Wang, “Machine learning for interatomic potential models,” J. Chem. Phys. **152**, 050902 (2020).

[20] T. B. Blank, S. D. Brown, A. W. Calhoun, and D. J. Doren, “Neural network models of potential energy surfaces,” J. Chem. Phys. **103**, 4129–4137 (1995).

[21] H. Gassner, M. Probst, A. Lauenstein, and K. Hermansson, “Representation of intermolecular potential functions by neural networks,” J. Phys. Chem. A **102**, 4596–4605 (1998).

[22] S. Lorenz, A. Groß, and M. Scheffler, “Representing high-dimensional potential-energy surfaces for reactions at surfaces by neural networks,” Chem. Phys. Lett. **395**, 210–215 (2004).

[23] S. Manzhos and T. Carrington Jr, “Using neural networks to represent potential surfaces as sums of products,” J. Chem. Phys. **125**, 194105 (2006).

[24] J. Behler and M. Parrinello, “Generalized neural-network representation of high-dimensional potential-energy surfaces,” Phys. Rev. Lett. **98**, 146401 (2007).

[25] S. A. Ghasemi, A. Hofstetter, S. Saha, and S. Goedecker, “Interatomic potentials for ionic systems with density functional accuracy based on charge densities obtained by a neural network,” Phys. Rev. B **92**, 045131 (2015).

[26] J. S. Smith, O. Isayev, and A. E. Roitberg, “ANI-1: An extensible neural network potential with dft accuracy at force field computational cost,” Chem. Sci. **8**, 3192–3203 (2017).

[27] K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, and K.-R. Müller, “Schnet – A deep learning architecture for molecules and materials,” J. Chem. Phys. **148**, 241722 (2018).

[28] L. Zhang, J. Han, H. Wang, R. Car, and E. Weinan, "Deep potential molecular dynamics: A scalable model with the accuracy of quantum mechanics," Phys. Rev. Lett. **120**, 143001 (2018).

[29] O. T. Unke and M. Meuwly, "Physnet: A neural network for predicting energies, forces, dipole moments, and partial charges," J. Chem. Theory Comput. **15**, 3678–3693 (2019).

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

[31] A. V. Shapeev, "Moment tensor potentials: A class of systematically improvable interatomic potentials," Multiscale Model. Simul. **14**, 1153–1173 (2016).

[32] A. P. Thompson, L. P. Swiler, C. R. Trott, S. M. Foiles, and G. J. Tucker, "Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials," J. Comput. Phys. **285**, 316–330 (2015).

[33] R. Drautz, "Atomic cluster expansion for accurate and transferable interatomic potentials," Phys. Rev. B **99**, 014104 (2019).

[34] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. Von Lilienfeld, "Fast and accurate modeling of molecular atomization energies with machine learning," Phys. Rev. Lett. **108**, 058301 (2012).

[35] S. Chmiela, A. Tkatchenko, H. E. Sauceda, I. Poltavsky, K. T. Schütt, and K.-R. Müller, "Machine learning of accurate energy-conserving molecular force fields," Sci. Adv. **3**, e1603015 (2017).

[36] A. Vitek, M. Stachon, P. Krömer, and V. Snáel, "Towards the modeling of atomic and molecular clusters energy by support vector regression," in *2013 5th International Conference on Intelligent Networking and Collaborative Systems* (IEEE, 2013) pp. 121–126.

[37] B. J. Braams and J. M. Bowman, "Permutationally invariant potential energy surfaces in high dimensionality," Int. Rev. Phys. Chem. **28**, 577–606 (2009).

[38] Z. Xie and J. M. Bowman, "Permutationally invariant polynomial basis for molecular energy surface fitting via monomial symmetrization," J. Chem. Theory Comput. **6**, 26–34 (2010).

[39] Y. Wang, X. Huang, B. C. Shepler, B. J. Braams, and J. M. Bowman, "Flexible, ab initio potential, and dipole moment surfaces for water. I. tests and applications for clusters up to the 22-mer," J. Chem. Phys. **134**, 094509 (2011).

[40] C. Qu, Q. Yu, and J. M. Bowman, "Permutationally invariant potential energy surfaces," Annu. Rev. Phys. Chem. **69**, 151–175 (2018).

[41] A. Nandi, C. Qu, and J. M. Bowman, "Full and fragmented permutationally invariant polynomial potential energy surfaces for trans and cis n-methyl acetamide and isomerization saddle

points," J. Chem. Phys. **151**, 084306 (2019).

[42] A. Nandi, C. Qu, P. L. Houston, R. Conte, and J. M. Bowman, "δ-machine learning for potential energy surfaces: A PIP approach to bring a DFT-based pes to CCSD(T) level of theory," J. Chem. Phys. **154**, 051102 (2021).

[43] B. Jiang and H. Guo, "Permutation invariant polynomial neural network approach to fitting potential energy surfaces," J. Chem. Phys. **139**, 054112 (2013).

[44] J. Li, B. Jiang, and H. Guo, "Permutation invariant polynomial neural network approach to fitting potential energy surfaces. II. Four-atom systems," J. Chem. Phys. **139**, 204103 (2013).

[45] B. Jiang and H. Guo, "Permutation invariant polynomial neural network approach to fitting potential energy surfaces. III. Molecule-surface interactions," J. Chem. Phys. **141**, 034109 (2014).

[46] C. Xie, X. Zhu, D. R. Yarkony, and H. Guo, "Permutation invariant polynomial neural network approach to fitting potential energy surfaces. IV. Coupled diabatic potential energy matrices," J. Chem. Phys. **149**, 144107 (2018).

[47] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, and A. Walsh, "Machine learning for molecular and materials science," Nature **559**, 547–555 (2018).

[48] J. Behler, "Four generations of high-dimensional neural network potentials," Chem. Rev. (2021) , https://doi.org/10.1021/acs.chemrev.0c00868.

[49] S. Manzhos and T. Carrington Jr, "Neural network potential energy surfaces for small molecules and reactions," Chem. Rev. (2021) , https://doi.org/10.1021/acs.chemrev.0c00665.

[50] T. Morawietz and J. Behler, "A density-functional theory-based neural network potential for water clusters including van der Waals corrections," J. Phys. Chem. A **117**, 7356–7366 (2013).

[51] C. Schran, J. Behler, and D. Marx, "Automated fitting of neural network potentials at coupled cluster accuracy: Protonated water clusters as testing ground," J. Chem. Theory Comput. **16**, 88–99 (2019).

[52] D. Rosenberger, J. S. Smith, and A. E. Garcia, "Modeling of peptides with classical and novel machine learning force fields: A comparison," J. Phys. Chem. B **125**, 3598–3612 (2021).

[53] S. Yue, M. C. Muniz, M. F. Calegari Andrade, L. Zhang, R. Car, and A. Z. Panagiotopoulos, "When do short-range atomistic machine-learning models fall short?" J. Chem. Phys. **154**, 034111 (2021).

[54] J. Rezac and P. Hobza, "Describing noncovalent interactions beyond the common approximations: How accurate is the "gold standard," CCSD(T) at the complete basis set limit?" J. Chem. Theory Comput. **9**, 2151–2155 (2013).

[55]D. Hankins, J. Moskowitz, and F. Stillinger, "Water molecule interactions," J. Chem. Phys. **53**, 4544–4554 (1970).

[56]F. Paesani, "Getting the right answers for the right reasons: Toward predictive molecular simulations of water with many-body potential energy functions," Acc. Chem. Res. **49**, 1844–1851 (2016).

[57]F. Paesani, "Water: Many-body potential from first principles (from the gas to the liquid phase)," Handbook of Materials Modeling: Methods: Theory and Modeling , 635–660 (2020).

[58]V. Babin, C. Leforestier, and F. Paesani, "Development of a "first principles" water potential with flexible monomers: Dimer potential energy surface, VRT spectrum, and second virial coefficient," J. Chem. Theory Comput. **9**, 5395–5403 (2013).

[59]V. Babin, G. R. Medders, and F. Paesani, "Development of a "first principles" water potential with flexible monomers. II: Trimer potential energy surface, third virial coefficient, and small clusters," J. Chem. Theory Comput. **10**, 1599–1607 (2014).

[60]G. R. Medders, V. Babin, and F. Paesani, "Development of a "first-principles" water potential with flexible monomers. III. Liquid phase properties," J. Chem. Theory Comput. **10**, 2906–2910 (2014).

[61]S. K. Reddy, S. C. Straight, P. Bajaj, C. Huy Pham, M. Riera, D. R. Moberg, M. A. Morales, C. Knight, A. W. Götz, and F. Paesani, "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).

[62]J. O. Richardson, C. Pérez, S. Lobsiger, A. A. Reid, B. Temelso, G. C. Shields, Z. Kisiel, D. J. Wales, B. H. Pate, and S. C. Althorpe, "Concerted hydrogen-bond breaking by quantum tunneling in the water hexamer prism," Science **351**, 1310–1313 (2016).

[63]W. T. Cole, J. D. Farrell, D. J. Wales, and R. J. Saykally, "Structure and torsional dynamics of the water octamer from thz laser spectroscopy near 215 μm," Science **352**, 1194–1197 (2016).

[64]J. D. Mallory and V. A. Mandelshtam, "Diffusion Monte Carlo studies of MB-pol $(H_2O)_{2-6}$ and $(D_2O)_{2-6}$ clusters: Structures and binding energies," J. Chem. Phys. **145**, 064308 (2016).

[65]P. E. Videla, P. J. Rossky, and D. Laria, "Communication: Isotopic effects on tunneling motions in the water trimer," J. Chem. Phys. **144**, 061101 (2016).

[66]S. E. Brown, A. W. Götz, X. Cheng, R. P. Steele, V. A. Mandelshtam, and F. Paesani, "Monitoring water clusters "melt" through vibrational spectroscopy," J. Am. Chem. Soc. **139**, 7082–7088 (2017).

[67] C. L. Vaillant and M. T. Cvitaš, “Rotation-tunneling spectrum of the water dimer from instanton theory,” Phys. Chem. Chem. Phys. **20**, 26809–26813 (2018).

[68] C. Vaillant, D. Wales, and S. Althorpe, “Tunneling splittings from path-integral molecular dynamics using a Langevin thermostat,” J. Chem. Phys. **148**, 234102 (2018).

[69] M. Schmidt and P.-N. Roy, “Path integral molecular dynamic simulation of flexible molecular systems in their ground state: Application to the water dimer,” J. Chem. Phys. **148**, 124116 (2018).

[70] K. P. Bishop and P.-N. Roy, “Quantum mechanical free energy profiles with post-quantization restraints: Binding free energy of the water dimer over a broad range of temperatures,” J. Chem. Phys. **148**, 102303 (2018).

[71] P. E. Videla, P. J. Rossky, and D. Laria, “Isotopic equilibria in aqueous clusters at low temperatures: Insights from the MB-pol many-body potential,” J. Chem. Phys. **148**, 084303 (2018).

[72] N. R. Samala and N. Agmon, “Temperature dependence of intramolecular vibrational bands in small water clusters,” J. Phys. Chem. B **123**, 9428–9442 (2019).

[73] M. T. Cvitaš and J. O. Richardson, “Quantum tunnelling pathways of the water pentamer,” Phys. Chem. Chem. Phys. **22**, 1035–1044 (2020).

[74] G. R. Medders and F. Paesani, “Infrared and Raman spectroscopy of liquid water through “first-principles” many-body molecular dynamics,” J. Chem. Theory Comput. **11**, 1145–1154 (2015).

[75] S. C. Straight and F. Paesani, “Exploring electrostatic effects on the hydrogen bond network of liquid water through many-body molecular dynamics,” J. Phys. Chem. B **120**, 8539–8546 (2016).

[76] S. K. Reddy, D. R. Moberg, S. C. Straight, and F. Paesani, “Temperature-dependent vibrational spectra and structure of liquid water from classical and quantum simulations with the MB-pol potential energy function,” J. Chem. Phys. **147**, 244504 (2017).

[77] K. M. Hunter, F. A. Shakib, and F. Paesani, “Disentangling coupling effects in the infrared spectra of liquid water,” J. Phys. Chem. B **122**, 10754–10761 (2018).

[78] Z. Sun, L. Zheng, M. Chen, M. L. Klein, F. Paesani, and X. Wu, “Electron-hole theory of the effect of quantum nuclei on the X-ray absorption spectra of liquid water,” Phys. Rev. Lett. **121**, 137401 (2018).

[79] A. P. Gaiduk, T. A. Pham, M. Govoni, F. Paesani, and G. Galli, “Electron affinity of liquid water,” Nat. Commun. **9**, 1–6 (2018).

[80] V. Cruzeiro, A. Wildman, X. Li, and F. Paesani, "Relationship between hydrogen-bonding motifs and the $1b_1$ splitting in the X-ray emission spectrum of liquid water," J. Phys. Chem. Lett. **12**, 3996–4002 (2021).

[81] G. R. Medders and F. Paesani, "Dissecting the molecular structure of the air/water interface from quantum simulations of the sum-frequency generation spectrum," J. Am. Chem. Soc. **138**, 3912–3919 (2016).

[82] D. R. Moberg, S. C. Straight, and F. Paesani, "Temperature dependence of the air/water interface revealed by polarization sensitive sum-frequency generation spectroscopy," J. Phys. Chem. B **122**, 4356–4365 (2018).

[83] S. Sun, F. Tang, S. Imoto, D. R. Moberg, T. Ohto, F. Paesani, M. Bonn, E. H. Backus, and Y. Nagata, "Orientational distribution of free OH groups of interfacial water is exponential," Phys. Rev. Lett. **121**, 246101 (2018).

[84] S. Sengupta, D. R. Moberg, F. Paesani, and E. Tyrode, "Neat water–vapor interface: Proton continuum and the nonresonant background," J. Phys. Chem. Lett. **9**, 6744–6749 (2018).

[85] M. C. Muniz, T. E. Gartner III, M. Riera, C. Knight, S. Yue, F. Paesani, and A. Z. Panagiotopoulos, "Vapor-liquid equilibrium of water with the MB-pol many-body potential," J. Chem. Phys. **154**, 211103 (2021).

[86] C. H. Pham, S. K. Reddy, K. Chen, C. Knight, and F. Paesani, "Many-body interactions in ice," J. Chem. Theory Comput. **13**, 1778–1784 (2017).

[87] D. R. Moberg, S. C. Straight, C. Knight, and F. Paesani, "Molecular origin of the vibrational structure of ice I_h," J. Phys. Chem. Lett. **8**, 2579–2583 (2017).

[88] D. R. Moberg, P. J. Sharp, and F. Paesani, "Molecular-level interpretation of vibrational spectra of ordered ice phases," J. Phys. Chem. B **122**, 10572–10581 (2018).

[89] D. R. Moberg, D. Becker, C. W. Dierking, F. Zurheide, B. Bandow, U. Buck, A. Hudait, V. Molinero, F. Paesani, and T. Zeuch, "The end of ice I," Proc. Natl. Acad. Sci. U.S.A. **116**, 24413–24419 (2019).

[90] L. del Rosso, M. Celli, D. Colognesi, S. Rudic, N. J. English, and L. Ulivi, "Density of phonon states in cubic ice I_c," ChemRxiv , https://doi.org/10.26434/chemrxiv.14769987.v1 (2021).

[91] D. J. Arismendi-Arrieta, M. Riera, P. Bajaj, R. Prosmiti, and F. Paesani, "i-TTM model for ab initio-based ion–water interaction potentials. 1. Halide–water potential energy functions," J. Phys. Chem. B **120**, 1822–1832 (2015).

[92] P. Bajaj, A. W. Götz, and F. Paesani, "Toward chemical accuracy in the description of ion–water interactions through many-body representations. I. Halide–water dimer potential energy surfaces," J. Chem. Theory Comput. **12**, 2698–2705 (2016).

[93] B. B. Bizzarro, C. K. Egan, and F. Paesani, "Nature of halide–water interactions: Insights from many-body representations and density functional theory," J. Chem. Theory Comput. **15**, 2983–2995 (2019).

[94] M. Riera, A. W. Götz, and F. Paesani, "The i-TTM model for ab initio-based ion–water interaction potentials. II. Alkali metal ion–water potential energy functions," Phys. Chem. Chem. Phys. **18**, 30334–30343 (2016).

[95] M. Riera, N. Mardirossian, P. Bajaj, A. W. Götz, and F. Paesani, "Toward chemical accuracy in the description of ion–water interactions through many-body representations. Alkali-water dimer potential energy surfaces," J. Chem. Phys. **147**, 161715 (2017).

[96] C. K. Egan, B. B. Bizzarro, M. Riera, and F. Paesani, "Nature of alkali ion–water interactions: Insights from many-body representations and density functional theory. II," J. Chem. Theory Comput. **16**, 3055–3072 (2020).

[97] M. Riera, E. P. Yeh, and F. Paesani, "Data-driven many-body models for molecular fluids: CO_2/H_2O mixtures as a case study," J. Chem. Theory Comput. **16**, 2246–2257 (2020).

[98] M. Riera, A. Hirales, R. Ghosh, and F. Paesani, "Data-driven many-body models with chemical accuracy for CH_4/H_2O mixtures," J. Chem. Phys. B **124**, 11207–11221 (2020).

[99] V. W. D. Cruzeiro, E. Lambros, M. Riera, R. Roy, F. Paesani, and A. W. Götz, "Highly accurate many-body potentials for simulations of N_2O_5 in water: Benchmarks, development, and validation," J. Chem. Theory Comput. , https://doi.org/10.1021/acs.jctc.1c00069 (2020).

[100] P. Bajaj, X.-G. Wang, T. Carrington Jr, and F. Paesani, "Vibrational spectra of halide–water dimers: Insights on ion hydration from full-dimensional quantum calculations on many-body potential energy surfaces," J. Chem. Phys. **148**, 102321 (2018).

[101] M. Riera, S. E. Brown, and F. Paesani, "Isomeric equilibria, nuclear quantum effects, and vibrational spectra of $M^+(H_2o)_{n=1-3}$ clusters, with M = Li, Na, K, Rb, and Cs, through many-body representations," J. Phys. Chem. A **122**, 5811–5821 (2018).

[102] P. Bajaj, M. Riera, J. K. Lin, Y. E. Mendoza Montijo, J. Gazca, and F. Paesani, "Halide ion microhydration: Structure, energetics, and spectroscopy of small halide–water clusters," J. Phys. Chem. A **123**, 2843–2852 (2019).

[103] P. Bajaj, J. O. Richardson, and F. Paesani, “Ion-mediated hydrogen-bond rearrangement through tunnelling in the iodide–dihydrate complex,” Nat. Chem. **11**, 367 (2019).

[104] P. Bajaj, D. Zhuang, and F. Paesani, “Specific ion effects on hydrogen-bond rearrangements in the halide–dihydrate complexes,” J. Phys. Chem. Lett. **10**, 2823–2828 (2019).

[105] D. Zhuang, M. Riera, G. K. Schenter, J. L. Fulton, and F. Paesani, “Many-body effects determine the local hydration structure of Cs^+ in solution,” J. Phys. Chem. Lett. **10**, 406–412 (2019).

[106] M. Riera, J. J. Talbot, R. P. Steele, and F. Paesani, “Infrared signatures of isomer selectivity and symmetry breaking in the $Cs^+(H_2O)_3$ complex using many-body potential energy functions,” J. Chem. Phys **153**, 044306 (2020).

[107] A. Caruso and F. Paesani, “Data-driven many-body models enable a quantitative description of chloride hydration from clusters to bulk,” ChemRxiv , 10.26434/chemrxiv.14755449.v1 (2021).

[108] “MBX: A many-body energy and force calculator,” `http://paesanigroup.ucsd.edu/software/mbx.html`.

[109] “LAMMPS molecular dynamics simulator,” `http://lammps.sandia.gov`.

[110] “i-PI: A universal force engine,” `http://ipi-code.org`.

[111] A. J. Stone, *The Theory of Intermolecular Forces* (Oxford University Press, Oxford, 2013).

[112] K. Tang and J. P. Toennies, “An improved simple model for the van der waals potential based on universal damping functions for the dispersion coefficients,” J. Chem. Phys. **80**, 3726–3741 (1984).

[113] C. Burnham, D. Anick, P. Mankoo, and G. Reiter, “The vibrational proton potential in bulk liquid water and ice,” J. Chem. Phys. **128**, 154519 (2008).

[114] Maplesoft Development Team, “Maple 2019 Maplesoft, a division of Waterloo Maple Inc., Waterloo, Ontario,” https://www.maplesoft.com.

[115] PostgreSQL Development Team, “PostgreSQL,” version 11.3.0, https://www.postgresql.org.

[116] T. P. Team, “Postgresql driver for python – psycopg,” https://www.psycopg.org/.

[117] S. E. Brown, “From ab initio data to high-dimensional potential energy surfaces: A critical overview and assessment of the development of permutationally invariant polynomial potential energy surfaces for single molecules,” J. Chem. Phys. **151**, 194111 (2019).

[118] K. Shoemake, “Uniform random rotations,” in *Graphics Gems III*, edited by D. Kirk (Academic Press, Inc, 1992).

[119] Y. Zhai, A. Caruso, S. Gao, and F. Paesani, "Active learning of many-body configuration space: Application to the Cs^+–water MB-nrg potential energy function as a case study," J. Chem. Phys. **152**, 144103 (2020).

[120] R. Pordes, D. Petravick, B. Kramer, D. Olson, M. Livny, A. Roy, P. Avery, K. Blackburn, T. Wenaus, F. Würthwein, I. Foster, R. Gardner, M. Wilde, A. Blatecky, J. McGee, and R. Quick, "The open science grid," J. Phys. Conf. Ser. **78**, 012057 (2007).

[121] Y. Shao, Z. Gan, E. Epifanovsky, A. T. B. Gilbert, M. Wormit, J. Kussmann, A. W. Lange, A. Behn, J. Deng, X. Feng, D. Ghosh, M. Goldey, P. R. Horn, L. D. Jacobson, I. Kaliman, R. Z. Khaliullin, T. Kús, A. Landau, J. Liu, E. I. Proynov, Y. M. Rhee, R. M. Richard, M. A. Rohrdanz, R. P. Steele, E. J. Sundstrom, H. L. Woodcock III, P. M. Zimmerman, D. Zuev, B. Albrecht, E. Alguire, B. Austin, G. J. O. Beran, Y. A. Bernard, E. Berquist, K. Brandhorst, K. B. Bravaya, S. T. Brown, D. Casanova, C.-M. Chang, Y. Chen, S. H. Chien, K. D. Closser, D. L. Crittenden, M. Diedenhofen, R. A. DiStasio Jr., H. Dop, A. D. Dutoi, R. G. Edgar, S. Fatehi, L. Fusti-Molnar, A. Ghysels, A. Golubeva-Zadorozhnaya, J. Gomes, M. W. D. Hanson-Heine, P. H. P. Harbach, A. W. Hauser, E. G. Hohenstein, Z. C. Holden, T.-C. Jagau, H. Ji, B. Kaduk, K. Khistyaev, J. Kim, J. Kim, R. A. King, P. Klunzinger, D. Kosenkov, T. Kowalczyk, C. M. Krauter, K. U. Lao, A. Laurent, K. V. Lawler, S. V. Levchenko, C. Y. Lin, F. Liu, E. Livshits, R. C. Lochan, A. Luenser, P. Manohar, S. F. Manzer, S.-P. Mao, N. Mardirossian, A. V. Marenich, S. A. Maurer, N. J. Mayhall, C. M. Oana, R. Olivares-Amaya, D. P. O'Neill, J. A. Parkhill, T. M. Perrine, R. Peverati, P. A. Pieniazek, A. Prociuk, D. R. Rehn, E. Rosta, N. J. Russ, N. Sergueev, S. M. Sharada, S. Sharmaa, D. W. Small, A. Sodt, T. Stein, D. Stück, Y.-C. Su, A. J. W. Thom, T. Tsuchimochi, L. Vogt, O. Vydrov, T. Wang, M. A. Watson, J. Wenzel, A. White, C. F. Williams, V. Vanovschi, S. Yeganeh, S. R. Yost, Z.-Q. You, I. Y. Zhang, X. Zhang, Y. Zhou, B. R. Brooks, G. K. L. Chan, D. M. Chipman, C. J. Cramer, W. A. Goddard III, M. S. Gordon, W. J. Hehre, A. Klamt, H. F. Schaefer III, M. W. Schmidt, C. D. Sherrill, D. G. Truhlar, A. Warshel, X. Xua, A. Aspuru-Guzik, R. Baer, A. T. Bell, N. A. Besley, J.-D. Chai, A. Dreuw, B. D. Dunietz, T. R. Furlani, S. R. Gwaltney, C.-P. Hsu, Y. Jung, J. Kong, D. S. Lambrecht, W. Liang, C. Ochsenfeld, V. A. Rassolov, L. V. Slipchenko, J. E. Subotnik, T. Van Voorhis, J. M. Herbert, A. I. Krylov, P. M. W. Gill, and M. Head-Gordon, "Advances in molecular quantum chemistry contained in the Q-Chem 4 program package," Mol. Phys. **113**, 184–215 (2015).

[122] D. G. Smith, L. A. Burns, A. C. Simmonett, R. M. Parrish, M. C. Schieber, R. Galvelis, P. Kraus, H. Kruse, R. Di Remigio, A. Alenaizan, A. M. James, S. Lehtola, J. P. Misiewicz, M. Scheurer, R. A. Shaw, J. B. Schriber, Y. Xie, Z. L. Glick, D. A. Sirianni, J. S. O'Brien, J. M. Waldrop, A. Kumar, E. G. Hohenstein, B. P. Pritchard, B. R. Brooks, H. F. Schaefer III, A. Y. Sokolov, K. Patkowski, A. E. DePrince III, U. Bozkaya, R. A. King, F. A. Evangelista, T. D. Turney, Justin M. Crawford, and C. D. Sherrill, "PSI4 1.4: Open-source software for high-throughput quantum chemistry," J. Chem. Phys. **152**, 184108 (2020).

[123] A. V. Marenich, S. V. Jerome, C. J. Cramer, and D. G. Truhlar, "Charge model 5: An extension of hirshfeld population analysis for the accurate description of molecular interactions in gaseous and condensed phases," J. Chem. Theory Comput. **8**, 527–541 (2012).

[124] A. D. Becke and E. R. Johnson, "Exchange-hole dipole moment and the dispersion interaction," J. Chem. Phys. **122**, 154104 (2005).

[125] E. R. Johnson and A. D. Becke, "A post-Hartree–Fock model of intermolecular interactions," J. Chem. Phys. **123**, 024101 (2005).

[126] E. R. Johnson and A. D. Becke, "A post-Hartree-Fock model of intermolecular interactions: Inclusion of higher-order corrections," J. Chem. Phys. **124**, 174104 (2006).

[127] S. F. Boys and F. Bernardi, "The calculation of small molecular interactions by the differences of separate total energies. some procedures with reduced errors," Mol. Phys. **19**, 553–566 (1970).

[128] Free Software Foundation, Inc., "GCC, the GNU compiler collection," https://gcc.gnu.org.

[129] Intel Corporation, "Intel Compiler for C++," https://software.intel.com.

[130] M. Bonomi, D. Branduardi, G. Bussi, C. Camilloni, D. Provasi, P. Raiteri, D. Donadio, F. Marinelli, F. Pietrucci, R. A. Broglia, and M. Parrinello, "PLUMED: A portable plugin for free-energy calculations with molecular dynamics," Comput. Phys. Commuun. **180**, 1961–1972 (2009).

[131] G. A. Tribello, M. Bonomi, D. Branduardi, C. Camilloni, and G. Bussi, "PLUMED 2: New feathers for an old bird," Comput. Phys. Commun. **185**, 604–613 (2014).

[132] GitHub, "MB-Fit: Software infrastructure for data-driven many-body potential energy functions," https://github.com/paesanilab/MB-Fit.

[133] D. R. MacFarlane, P. V. Cherepanov, J. Choi, B. H. Suryanto, R. Y. Hodgetts, J. M. Bakker, F. M. F. Vallana, and A. N. Simonov, "A roadmap to the ammonia economy," Joule **4**, 1186–1205 (2020).

[134] A. E. DePrince III and C. D. Sherrill, “Accuracy and efficiency of coupled-cluster theory using density fitting/Cholesky decomposition, frozen natural orbitals, and at t_1-transformed hamiltonian,” J. Chem. Theory Comput. **9**, 2687–2696 (2013).

[135] C. Adamo and V. Barone, “Toward reliable density functional methods without adjustable parameters: The PBE0 model,” J. Chem. Phys. **110**, 6158–6170 (1999).

[136] S. Grimme, J. Antony, S. Ehrlich, and H. Krieg, “A consistent and accurate *ab initio* parameterization of density functional dispersion correction (DFT-D) for the 94 elements H–Pu,” J. Comp. Phys. **132**, 154104:1–19 (2010).

[137] T. H. Dunning Jr, “Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen,” J. Chem. Phys. **90**, 1007–1023 (1989).

[138] C. K. Egan and F. Paesani, “Assessing many-body effects of water self-ions. I: $OH^{-}(H_2O)_n$ clusters,” J. Chem. Theory Comput. **14**, 1982–1997 (2018).

[139] M. Riera, E. Lambros, T. T. Nguyen, A. W. Götz, and F. Paesani, “Low-order many-body interactions determine the local structure of liquid water,” Chem. Sci. **10**, 8211–8218 (2019).

[140] C. K. Egan and F. Paesani, “Assessing many-body effects of water self-ions. II: $H_3O^{+}(H_2O)_n$ clusters,” J. Chem. Theory Comput. **15**, 4816–4833 (2019).

[141] F. Paesani, P. Bajaj, and M. Riera, “Chemical accuracy in modeling halide ion hydration from many-body representations,” Adv. Phys. X **4**, 1631212 (2019).

[142] G. R. Medders, A. W. Götz, M. A. Morales, P. Bajaj, and F. Paesani, “On the representation of many-body interactions in water,” J. Chem. Phys. **143**, 104102 (2015).

[143] E. Lambros and F. Paesani, “How good are polarizable and flexible models for water: Insights from a many-body perspective,” J. Chem. Phys. **153**, 060901 (2020).

[144] U. Góra, R. Podeszwa, W. Cencek, and K. Szalewicz, “Interaction energies of large clusters from many-body expansion,” J. Chem. Phys. **135**, 224102 (2011).

[145] J. Hirschfelder, F. McClure, and I. Weeks, “Second virial coefficients and the forces between complex molecules,” J. Chem. Phys. **10**, 201–214 (1942).

[146] M. Ricci, M. Nardone, F. Ricci, C. Andreani, and A. Soper, “Microscopic structure of low temperature liquid ammonia: A neutron diffraction experiment,” J. Chem. Phys. **102**, 7650–7655 (1995).

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

[148] J. Towns, T. Cockerill, M. Dahan, I. Foster, K. Gaither, A. Grimshaw, V. Hazlewood, S. Lathrop, D. Lifka, G. D. Peterson, R. Roskies, J. R. Scott, and N. Wilkins-Diehr, “XSEDE: Accelerating scientific discovery,” Comput. Sci. Eng. **16**, 62–74 (2014).