Toward Data-Driven Many-Body Simulations of Biomolecules in Solution: N-Methyl Acetamide as a Proxy for the Protein Backbone

Ruihan Zhou,*,† Marc Riera,† and Francesco Paesani*,†,‡,¶,§

†*Department of Chemistry and Biochemistry, University of California San Diego,*
La Jolla, California 92093, United States

‡*Materials Science and Engineering, University of California San Diego,*
La Jolla, California 92093, United States

¶*Halicioğlu Data Science Institute, University of California San Diego,*
La Jolla, California 92093, United States

§*San Diego Supercomputer Center, University of California San Diego,*
La Jolla, California 92093, United States

E-mail: ruz088@ucsd.edu; fpaesani@ucsd.edu

Abstract

The development of molecular models with quantum-mechanical accuracy for predictive simulations of biomolecular systems has been a long standing goal in the field of computational biophysics and biochemistry. As a first step towards a transferable force field for biomolecules entirely derived from “first principles”, we introduce a data-driven many-body energy (MB-nrg) potential energy function (PEF) for N-methyl acetamide (NMA), a peptide bond capped by two methyl groups that is commonly used as a proxy for the protein backbone. The MB-nrg PEF is shown to accurately describe the energetics and structural properties of an isolated NMA molecule, including the normal modes of both *cis* and *trans* isomers and the energy variation along the isomerization path, as well as the multidimensional potential energy landscape of the NMA–H_2O dimer in the gas phase. Importantly, we show that the MB-nrg PEF is fully transferable, enabling molecular dynamics simulations of NMA in solution with quantum-mechanical accuracy. Comparisons with results obtained with a popular pairwise-additive force field for biomolecules and a classical polarizable PEF demonstrate the ability of the MB-nrg PEF to accurately represent many-body effects in NMA–H_2O interactions at both short and long distances, which is key to guaranteeing full transferability from the gas to the liquid phase.

INTRODUCTION

Water plays an active role in mediating most biological processes.[1] It is thus not surprising that significant effort has been made to identify possible correlations between the structure, dynamics, and function of biomolecules in solution and the properties of the surrounding water's hydrogen-bond network.[2] These efforts have led to the development of several fundamental concepts in biomolecular hydration, including hydrophobicity, crowding, and confinement.

Proteins are among the most important biomolecules. They are made up of unique sequences of amino acids, which determine both their structures and functions.[3] In these sequences, each amino acid is connected to its neighbors by peptide bonds.[4] Proteins play key roles in nearly all biological processes taking place in living systems, from catalyzing metabolic reactions[5] and replicating DNA[6] to responding to stimuli[7] and transporting molecules from one location to another.[8]

The hydration properties of a protein can have a considerable impact on how the protein folds into a three-dimensional shape and carries out its biological functions in solution.[9–11] For example, some proteins tend to form aggregates and self-assemble under specific thermodynamic conditions,[12–14] while other proteins, which lack hydrophobic cores, contain extensive regions that are intrinsically disordered.[15] Understanding how proteins interact with water at the molecular level is thus key to conceptualizing how proteins fold and function, which in turn provides insight into fundamental biological processes.[16,17]

Since the first computer simulation of a protein,[18] molecular dynamics (MD),[19] Monte Carlo (MC),[20] and coarse-grained (CG)[21] simulations have been extensively used to model the structural, thermodynamic, and dynamical properties of proteins. Recent advances in computer hardware and algorithms have made it possible to simulate larger and more complex proteins on increasingly long time scales,[22] often providing molecular insights that are difficult to obtain by other means. Although computer simulations play a central role in the study of biological systems,[23] there are still several challenges that must be overcome to improve their realism and predictive power, including the need for more accurate molecular models to represent the underlying multidimensional free-energy landscape[24] as well as more efficient algorithms to sample conformational changes that take place over long time scales.

Pairwise-additive force fields (FFs) have been the workhorse of molecular simulations of proteins since the introduction of the "consistent force field" (CFF).[25,26] Building upon these pioneering efforts, AMBER,[27] CHARMM,[28] OPLS,[29] and GROMOS[30] are currently among the most commonly used FFs in biomolecular simulations. These pairwise-additive FFs are based on relatively simple energy functions that aim to represent both the (intramolecular) distortion of a biomolecule and the (intermolecular) biomolecule–water and water–water interactions. In general, the energy functions adopted by pairwise-additive FFs are empirically parameterized to reproduce a set of experimental data, which are sometimes supplemented by data derived from quantum-mechanical calculations.

Although pairwise-additive FFs offer a computationally inexpensive representation of biomolecular systems, they are by construction unable to describe many-body effects, which drastically limits their transferability and predictive power.[24,31] To overcome these limitations, several polarizable FFs have been developed to capture many-body electrostatic interactions with classical expressions. Some polarizable FFs adopt the fluctuating charge approach, which determines many-body polarization by optimizing the total electrostatic energy with electronegativity equalization along different bonds.[32] A second method to account for many-body electrostatic interactions employs the Drude oscillator model as in the polarizable version of the CHARMM force field.[33,34] The Drude oscillator model attaches a virtual, partially charged, and massless site to each polarizable atom which, vibrating about the corresponding atom, effectively mimics the polarization of the electron cloud due to the surrounding electric field.[35–37] Other polarizable FFs, including AMOEBA[32,38] and the polarizable version of the AMBER force field,[39] explicitly account for induced dipoles on each (polarizable) molecule. Although relatively computationally more expensive than the Drude oscillator model, the induced-dipole model has been shown to be more accurate in reproducing the non-additive and directional nature of many-body polarization.[38]

Over the past decade, machine learning (ML) has gained popularity in molecular sciences, and, as a consequence, several neural network FFs derived from *ab initio* data have been reported in the literature.[40] In general terms, neural network FFs fall into two main categories, depending on whether they rely on predefined descriptors for the target molecular system or they learn the descriptors from the data during the training process.[41] ANI,[42,43] DeePMD,[44] Tensormol,[45] EANN,[46] BpopNN[47] and 4G-HDNNP[48] are examples of neural network FFs with predefined descriptors, which utilize parameters in a preset functional form for characterizing atomic geometries, typically decomposing the energies into atomic contributions.[49] Some neural network FFs represent the energy of a molecule as a sum of intrinsic bond energies such as BIM-NN,[50] or use an atomic-pairwise framework to represent molecular interactions, such as AP-Net.[51] PhysNet,[52] HIP-NN,[53] SchNet,[54,55] and AIMNet[56] are neural network FFs that learn a descriptor from structural information of the training systems without prior knowledge of the functional forms of the descriptor. Although neural network FFs represent a promising avenue to more realistic biomolecular simulations, their transferability across different phases and thermodynamic conditions remains challenging.[57]

In parallel with the development of neural network FFs, the last decade has witnessed the emergence of hybrid data-driven/physics-based potential energy functions (PEFs) that encode the correct physics of many-body interactions while using ML algorithms to "learn" short-range quantum-mechanical interactions directly from electronic structure data.[58] In this context, the many-body energy (MB-nrg) theoretical/computational framework,[59] which exploits the "nearsightedness" of electronic matter[60] to rigorously represent the energy of a given molecular system in terms of individual many-body contributions,[61] has successfully been applied to model the properties of various molecular systems across different phases, including water,[62–89] ammonia,[90] carbon dioxide,[91,92] methane,[93,94] halide and alkali-metal ions in water,,[95–104] nitrogen pentoxide in water,[105,106] and alkanes.[107] In this study, we present the first step towards the development of a general data-driven many-body FF for proteins by introducing an MB-nrg PEF for N-methyl acetamide (NMA) that is commonly used as a proxy of the protein backbone. Besides its relevance in the development of biomolecular FFs, NMA is also important because, given its high boiling point and low toxicity, it is used in several industrial processes.[108] In addition, NMA is used as a starting material in the synthesis of other chemicals, such as pharmaceuticals and polymers, as well as a non-toxic alternative to traditional solvents.[109,110] The hydration structure, solvation energetics, vibrational spectra, and isomerization of NMA have been studied, both experimentally and theoretically.[111–115] Recently, molecular models for the *trans* and *cis* isomers of gas-phase NMA have been developed using full and fragmented permutationally invariant polynomials (PIPs).[116–118]

Building upon the accuracy and predictive power of the MB-nrg PEFs, we introduce here a transferable MB-nrg PEF that accurately represents the structure and energetics of NMA, including the isomerization path, both in the gas phase and in solution. Given the generality and modularity of the MB-nrg PEFs, our results suggest that the MB-nrg theoretical/computational framework represents a viable approach to developing a comprehensive and transferable data-driven many-body force field that can accurately predict the structure and energetics of biomolecules across different phases.

METHODS

Many-body potential energy functions

The many-body expansion (MBE) of the energy allows for rigorously expressing the total energy of an N-body system as a sum of n-body energy contributions ($1 \leq n \leq N$) according to:[61]

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

Here, the 1-body energy, $\varepsilon^{1\mathrm{B}}(i)$, refers to the distortion energy of the ith monomer relative to its minimum-energy geometry, i.e., $\varepsilon^{1\mathrm{B}}(i) = E(i) - E_{eq}(i)$, where $E(i)$ and $E_{eq}(i)$ are the energies for the distorted and equilibrium geometries, respectively. For $n \geq 2$, the n-body (nB) energies are defined recursively through the following expression:

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

Since the MBE converges quickly for nonmetallic systems,[119–122] eq 1 provides a rigorous and efficient theoretical/computational framework for the development of full-dimensional PEFs where each n-body energy term is fitted to reproduce the corresponding reference values obtained from electronic structure calculations.[62–64,90–93,95,96,102,104,105,107,123–127] Examples of many-body PEFs derived from eq 1 are the Thole-type-model energy (TTM-nrg)[95,96] and many-body energy (MB-nrg)[95,96] PEFs for generic solutes in water. In both TTM-nrg and MB-nrg PEFs, water–water interactions are represented by MB-pol,[62–64] which has been shown to accurately predict the properties of water,[70] from small clusters in the gas phase[66–69,75–80,84,86] to liquid water,[65,72,81,87,89,128–130] the air/water interface,,[71,88,131–133] and ice.[73,74,82,83] Both MB-nrg and TTM-nrg PEFs are built upon an underlying many-body polarizable model supplemented by a 2-body dispersion energy term. The difference between the TTM-nrg and MB-nrg PEFs lays on the representation of short-range interactions, with the former using a sum of (2-body) Born-Mayer potentials,[134,135] and the latter using multidimensional PIPs at the 2-body level and, in some cases, at the 3-body level as well.[95,96] In this study, we introduce a slight variation of the MB-nrg framework in which the 2-body PIP is implemented on top of the 2-body Born-Mayer potentials.

Briefly, the TTM-nrg and MB-nrg NMA–H_2O PEFs approximate eq 1 as

$$E_N(r_1,..,r_N) = \sum_{i=1}^{N}\varepsilon^{1\mathrm{B}}(i) + \sum_{i>j}^{N}\varepsilon^{2\mathrm{B}}(i,j) + V_{\mathrm{pol}} \tag{3}$$

In both types of PEFs, the NMA 1-body energy, $\varepsilon^{1\mathrm{B}}$, describing the distortion of an NMA molecule from its equilibrium geometry, is represented by a PIP fitted to 1-body reference energies. A PIP of the 66 distances between all pairs of atoms in an NMA molecule, containing 30 first-degree monomials, 538 second-degree monomials, and 7471 third-degree monomials, was found to provide the optimal compromise between accuracy and computational efficiency. The point charges of the NMA atoms, which are used to calculate NMA–H_2O permanent electrostatic interactions (eq 4), were calculated for the equilibrium geometry of NMA using the Charge Model 5 (CM5) method[136] as implemented in Q-Chem 5.[137] The CM5 calculations were carried out with the hybrid, range-separated, meta-GGA ωB97M-V[138] functional in combination with the aug-cc-pVTZ basis set.[139] All NMA atoms were also assigned dipole polarizabilities, which are used to calculate NMA–H_2O polarization energy represented by $V_{\rm pol}$ in eq 3. The dipole polarizabilities were determined at the ωB97M-V/aug-cc-pVTZ level of theory according to the Exchange Dipole Moment (XDM) model[140–142] implemented in Q-Chem 5.[137]

The NMA–H_2O 2-body energy, $\varepsilon^{\rm 2B}$, is expressed as:

$$\varepsilon^{\rm 2B} = V_{\rm sr}^{\rm 2B} + V_{\rm elec} + V_{\rm disp} \tag{4}$$

In the TTM-nrg PEF, the short-range term, $V_{\rm sr}^{\rm 2B}$, describing 2-body NMA–H_2O repulsive interactions, is represented by a sum of pairwise Born-Mayer potentials between all atoms of the two molecules,[143]

$$V_{\rm sr}^{\rm 2B} = \sum_{\substack{\alpha \in {\rm NMA} \\ \beta \in {\rm water}}} A_{\alpha\beta} e^{-b_{\alpha\beta} R_{\alpha\beta}} \tag{5}$$

Here, α and β are atom indexes within the NMA and H_2O molecules, and $A_{\alpha\beta}$ and $b_{\alpha\beta}$ are fitting parameters.

In the MB-nrg PEF, $V_{\rm sr}^{\rm 2B}$ is represented by the same sum of pairwise Born-Mayer potentials adopted by the TTM-nrg PEF which is supplemented by a third-degree PIP, $V_{\rm PIP}^{\rm 2B}$, that is smoothly switched to zero as the NMA–H_2O distance becomes larger than a predefined cutoff set to 8.0 Å. It has been shown that $V_{\rm PIP}^{\rm 2B}$ effectively recovers quantum-mechanical short-range 2-body interactions (e.g., exchange-repulsion, charge transfer, and charge penetration) that arise from the overlap of the electron densities of the two molecules within the dimer.[144] Due to the relatively large number of atoms in an NMA molecule, the development of an accurate, yet computationally efficient, $V_{\rm PIP}^{\rm 2B}$

poses new challenges relative to other MB-nrg PEFs developed for small solutes in water. For example, the full third-degree NMA–H_2O PIP contains 36763 terms, which makes the associated computational cost effectively unaffordable for MD simulations of NMA in solution. In the present MB-nrg NMA–H_2O PEF, the number of terms in $V_{\mathrm{PIP}}^{\mathrm{2B}}$ was reduced from 36763 to 2209 by applying multiple filters which removed terms involving variables that were found to be irrelevant for the accurate representation of the NMA–H_2O 2-body potential energy landscape. In addition, $V_{\mathrm{PIP}}^{\mathrm{2B}}$ only includes the heavy atoms of NMA along with the hydrogen atom of the amide group. Additional filters were applied to remove all terms that only include intramolecular distances within the NMA monomer, since they are accurately described by the 1-body PIP. In the subsequent step, the distances of all intermolecular pairs were calculated, and all PIP terms of degree 3 or higher were removed if they contained a pair whose average distance is larger than 2.0 Åsince their interactions are accurately described by V_{pol}. A complete description of the filters applied in the development of $V_{\mathrm{PIP}}^{\mathrm{2B}}$ is reported in the Supporting Information.

In eq 4, V_{elec} describes permanent electrostatics that is calculated from Coulomb interactions between the point charges located on the NMA and water molecules within a dimer. As discussed in the original references,[62,63] the MB-pol point charges were fitted to reproduce the *ab initio* dipole moment of an isolated water molecule.[145] The last term of eq 4, V_{disp}, describes the 2-body dispersion energy and is expressed as:

$$V_{\mathrm{disp}} = -\sum_{\substack{\alpha \in \mathrm{NMA} \\ \beta \in \mathrm{water}}} f(\delta_{\alpha\beta} R_{\alpha\beta}) \frac{C_{6,\alpha\beta}}{R_{\alpha\beta}^6} \tag{6}$$

where α and β are atom indexes within the NMA and water molecules, $C_{6,\alpha\beta}$ is the associated dispersion coefficient, and $f(\delta_{\alpha\beta} R_{\alpha\beta})$ is the Tang-Toennies damping function[146] with damping parameter $\delta_{\alpha\beta}$:

$$f(\delta_{\alpha\beta}, R_{\alpha\beta}) = 1 - \exp(-\delta_{\alpha\beta} R_{\alpha\beta}) \sum_{n=0}^{6} \frac{(\delta_{\alpha\beta} R_{\alpha\beta})^n}{n!} \tag{7}$$

Similar to the MB-nrg PEFs for carbon dioxide[91,92] and methane,[93] $\delta_{\alpha\beta}$ was set to be equal to $b_{\alpha\beta}$

in eq 5, and all $C_{6,\alpha\beta}$ coefficients were calculated at the ωB97M-V/aug-cc-pVTZ level of theory using the XDM model[140–142] as implemented in Q-Chem 5.[137]

Finally, V_{pol} in eq 1 is a classical many-body polarization term based on a modified version of the Thole-type model that includes induced dipoles on all atoms as implemented in MB-pol[62,63] and previous MB-nrg PEFs.[90,91,93,95,96,105]

The 1-body and 2-body energy terms of the TTM-nrg and MB-nrg PEFs were fitted separately using the MB-Fit software[90] which minimizes the weighted sum of squared errors:

$$\chi^2 = \sum_{k \in \mathscr{S}} w_k \left[\varepsilon^{\text{nB}}(k) - \varepsilon^{\text{nB}}_{\text{ref}}(k)\right]^2 + \Gamma^2 \sum_l A_l^2 \tag{8}$$

Here, $\varepsilon^{\text{nB}}(k)$ and $\varepsilon^{\text{nB}}_{\text{ref}}(k)$ are the model (TTM-nrg or MB-nrg) and reference n-body energies, respectively, for the kth configuration of the corresponding n-body training sets. $\Gamma^2 \sum_l A_l^2$ is a regularization term,[147] favoring smaller linear fitting parameters A_l with magnitude determined by the regularization parameter Γ^2. The weights w_k were calculated to bias the fit in favor of low-energy configurations:

$$w_k = \left(\frac{\delta E}{\varepsilon^{\text{nB}}(k) - \varepsilon^{\text{nB}}_{\text{min}} + \delta E} \right)^2 \tag{9}$$

where $\varepsilon^{\text{nB}}_{\text{min}}$ is the minimum ε^{nB} in the corresponding n-body training set, and δE is a parameter that ensures less weighing on more distorted (i.e., high energy) configurations. All nonlinear fitting parameters entering the TTM-nrg and MB-nrg expressions for ε^{nB} were optimized using the simplex algorithm, where the linear parameters were determined at each step using ridge regression.[147] All technical details of the fitting procedure are discussed in ref 90.

The atomic charges and dipole polarizabilities, as well as Born-Mayer and dispersion coefficients for the TTM-nrg and MB-nrg NMA–H_2O PEFs are reported in the Supporting Information. Both TTM-nrg and MB-nrg NMA–H_2O PEFs are available in MBX.[148]

Training and test sets

Selecting appropriate training configurations is a nontrivial task in the development of data-driven PEFs. While a well-chosen training set is expected to maximize the coverage over the target n-body potential energy landscape, an “incomplete” training set may result in “holes” in the TTM-nrg and MB-nrg representations of the corresponding n-body energies, i.e., regions where the TTM-nrg and MB-nrg PEFs predict nonphysical n-body energies due to insufficient coverage in the corresponding n-body training sets. In the present study, the generation of the 1-body and 2-body training sets was performed with the MB-Fit software,[149] following the procedure described in ref 90.

The NMA 1-body training set contains 15000 configurations generated from normal-mode sampling carried out with the corresponding quantum distribution at 298.15 K (12500 configurations) and 5000 K (2500 configurations). Sampling at a high temperature allowed for generating higher molecular distortions that are necessary to guarantee “complete” coverage of the 1-body

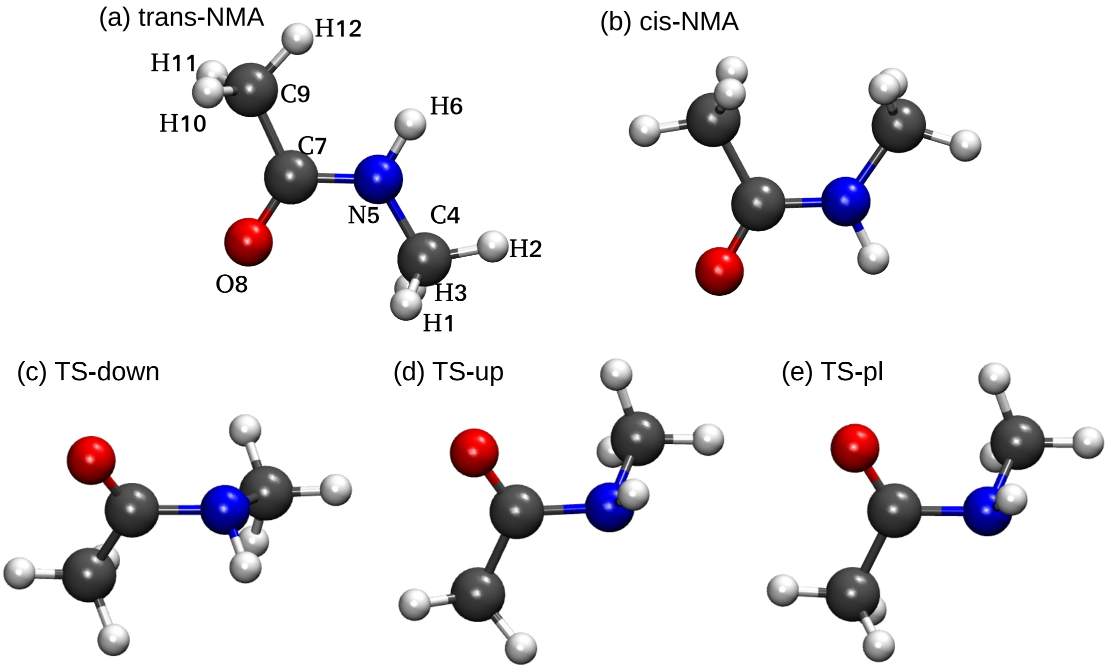

Figure 1: Geometries for the *tran*, *cis*, and three saddle-point structures of NMA reported in ref 150. The geometry optimizations were carried out at DF-MP2/AVQZ level of theory.

energy landscape. An additional set of 5000 configurations was generated from normal-mode sampling carried out at 298.15 K with the corresponding classical distribution. An equal number of configurations was generated from normal-mode sampling carried out for both the *trans* and *cis* structures of NMA (Figure 1a-b), as well as for the three saddle-point structures identified in ref 150 (Figure 1c-e).

An additional set of 13742 configurations extracted from metadynamics simulations[151] of an isolated NMA molecule was added to the 1-body training set. The metadynamics simulations were carried out with the PLUMED (v2.6) interface[152] for the LAMMPS (v2021.9.29) package,[153] using the Generalized AMBER Force Field (GAFF)[154,155] to describe the NMA molecule. The collective variable chosen for the metadynamics simulations was the C9-C7-N5-C4 dihedral angle that is part of the *cis-trans* isomerization path (see atom labels in Figure 1). In total, the final 1-body training set contains 33742 configurations. The corresponding 1-body test set contains 2000 configurations extracted from normal-mode sampling, and 3311 configurations extracted from metadynamics simulations.

The 2-body NMA–H_2O training set for both TTM-nrg and the MB-nrg PEFs consists of 27976 configurations. A first set of 2659 configurations was obtained from radial scans along interatomic distances between all atom pairs that have one atom located on the NMA molecule and one atom located on the water molecule. In these radial scans the two molecules were kept in their minimum-energy geometries and randomly rotated with respect to each other, while the interatomic distances were sampled using a logarithmic distribution favoring shorter distances. Some pairs were not included since they involve equivalent atoms, e.g., the three hydrogen atoms located on the same methyl group. Additional 6234 configurations were obtained from similar radial scans in which the geometries of the two molecules were distorted using different combinations of the corresponding normal modes. The last set of 19083 configurations was obtained from metadynamics simulations of a single NMA molecule in bulk water that were carried out in the isobaric-isothermal ensemble (NPT) at 298 K and 1 atm using LAMMPS[153] patched with PLUMED.[152] In these simulations, the GAFF force field was used to describe the NMA molecule and the water molecules were described by the TIP4P/2005 model.[156] Lorentz-Berthelot mixing rules were used to describe NMA–H_2O interactions.[157] The same dihedral angle used to generate the 1-body training set was used as the collective variable in the metadynamics simulations of an NMA molecule in bulk water. All NMA–H_2O pairs with an intermolecular distance between the two centers of mass shorter than 9.0 Å were extracted from the metadynamics simulations and subsequently screened using furthest point sampling.[158] The NMA–H_2O 2-body test set contains a total of 5961 configurations, including 260 configurations generated from radial scans with rigid NMA and H_2O molecules, 1184 configurations obtained from radial scans with distorted NMA and H_2O molecules, and 4517 configurations obtained from metadynamics simulations.

All electronic structure calculations of the 1-body and 2-body reference energies were carried out with MOLPRO 2019[159,160] using density fitting[161–163] second-order Møller-Plesset[164] (DF-MP2) perturbation theory in combination with the aug-cc-pVQZ (AVQZ) basis set.[139] The DF-MP2/aug-cc-pVQZ 2-body energies were corrected for the basis set superposition error using with the counterpoise method.[165]

Molecular dynamics simulations

MD simulations were carried out in the isothermal-isobaric (NPT: constant number of particles, pressure, and temperature) ensemble at 298.15 K and 1.0 atm for 1 NMA molecule and 277 water molecules (corresponding to $\sim$0.2 M concentration) in periodic boundary conditions. A time step of 0.5 fs was used to propagate the equations of motion according to the velocity-Verlet algorithm for 1 ns.[166] All MD simulations with the TTM-nrg and MB-nrg PEFs were carried out with LAMMPS[153] through the interface with the MBX software.[148] For comparison, MD simulations were also carried out with Amber22[167] using both the GAFF[154,155] and ff14SB[168] force fields for NMA combined with the TIP4P/2005 water model.[156] Specific details about the GAFF and ff14SB parameters for NMA are reported in the Supporting Information.

Results

NMA in the gas phase

Correlation plots between the DF-MP2/aug-cc-pVQZ reference values and the corresponding MB-nrg results for the 1-body energies of NMA (panel a), the 2-body energies for TTM-nrg NMA–H_2O PEF (panel b), and 2-body energies for MB-nrg NMA–H_2O PEF (panel c) are shown in Fig. 2. The corresponding root-mean-square deviations (RMSDs) are 0.2710 kcal/mol, 0.7826 kcal/mol, and 0.2690 kcal/mol, respectively. As discussed in the Methods section, the TTM-nrg PEF adopts the same representation of the 1-body energies as the MB-nrg PEF. Consistent with previous studies,[90,91,93,107] the MB-nrg PEF is able to accurately represent both 1-body and 2-body energies. In contrast, the limitations of a purely classical description of many-body interactions, which were already identified in refs. 93 and refs. 91, manifest in large deviations between the TTM-nrg 2-body energies and the corresponding DF-MP2/AVQZ reference data.

As shown in Figure 1, NMA can exist as either a *cis* or *trans* isomer, with the two configurations being related to each other through a rotation of the two terminal methyl groups about the peptide

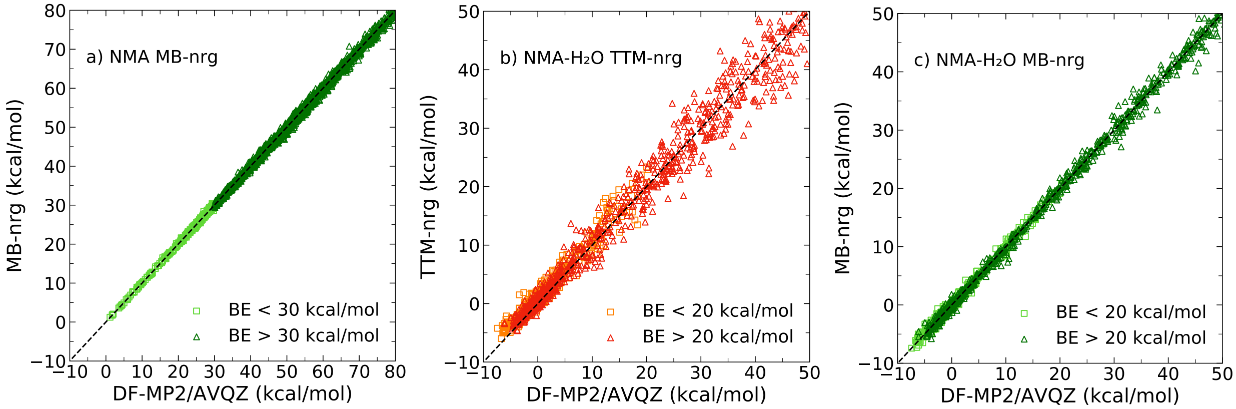

Figure 2: Correlation plots between the DF-MP2/AVQZ reference energies and the corresponding values calculated with the different PEFs for the 1-body and 2-body test sets. (a) NMA 1-body MB-nrg, (b) NMA-H_2O 2-body TTM-nrg, and (c) NMA-H_2O 2-body MB-nrg. In all panels, light and dark colors represent configuration with binding energies lower and higher than the corresponding cutoff values, which were set to 30 kcal/mol and 20 kcal/mol for the 1-body and 2-body configurations, respectively.

bond. To examine the ability of the MB-nrg PEF to correctly predict the structure of both isomers, the MB-nrg optimized geometries were compared with the corresponding DF-MP2/AVQZ reference structures, resulting in RMSDs of 0.016 Å and 0.008 Å for the *cis* and *trans* isomers, respectively. Detailed comparisons illustrating the accuracy of the MB-nrg PEF in reproducing the DF-MP2/AVQZ structural parameters for the two NMA isomers are reported in Tables S7 and S8 of the Supporting Information.

While the comparisons discussed above allow for assessing the accuracy of the MB-nrg PEF in describing the minimum-energy structures of both *cis* and *trans* isomers of NMA, they do not provide any information about the ability of the MB-nrg PEF to represent the overall morphology of the underlying multidimensional potential energy surface. To gain insights into the general shape of the NMA potential energy surface in the neighborhoods of the *cis* and *trans* minimum-energy structures, Figure 3 shows the comparisons between the reference DF-MP2/AVQZ harmonic frequencies for the *cis* and *trans* isomers and the corresponding values predicted by the MB-nrg PEF.

For both isomers, the MB-nrg PEF predicts harmonic frequencies within 20 cm^{-1} of the reference DF-MP2/AVQZ values with the exception of normal modes 1 and 14 of the *trans* isomer, and normal modes 15 and 28 of the *cis* isomer. The corresponding relative errors, defined as $\Delta\omega = (\omega_{MB-nrg} - \omega_{DF-MP2/AVQZ})/\omega_{DF-MP2/AVQZ}$, are below 3% for all normal modes except for the first four normal modes with the lowest frequencies. Since these four normal modes involve the collective motion of several atoms, the deviations with the reference DF-MP2/AVQZ values found in Figure 3 suggest that a PIP of degree higher than 3, which is currently used in the 1-body MB-nrg PEF, may be needed for a more accurate description of these low-frequency normal modes.

It should be noted that the MB-nrg PEF slightly underestimates the harmonic frequencies of normal mode 1 (N-C-H3 deformation), normal mode 3 (combination of C-N-C deformation and N-H out-of-plane bending), and normal mode 14 (combination of N-C-H3 rocking, C4-N stretching, and N-H in-plane bending), while slightly overestimating the harmonic frequencies of normal mode 6 (combination of C-N-C deformation, C-O in-plane bending, and C-C-H3 rocking) and mode 8 (combination of C-O in-plane bending and C-C stretching). These differences indicate that the MB-nrg PEF makes the carbonyl and amino sides of the amide bond slightly stiffer and softer, respectively, relative to reference DF-MP2/AVQZ values.

For normal modes with frequencies above 500 cm^{-1}, the second largest relative deviation (-1.67 %) from the reference DF-MP2/AVQZ values is found for normal mode 15 that corresponds to the amide III mode (combination of N-H in-plane bending and C7-N stretching). This deviation is a consequence of the stiffening and softening of the carbonyl and amino sides of the peptide

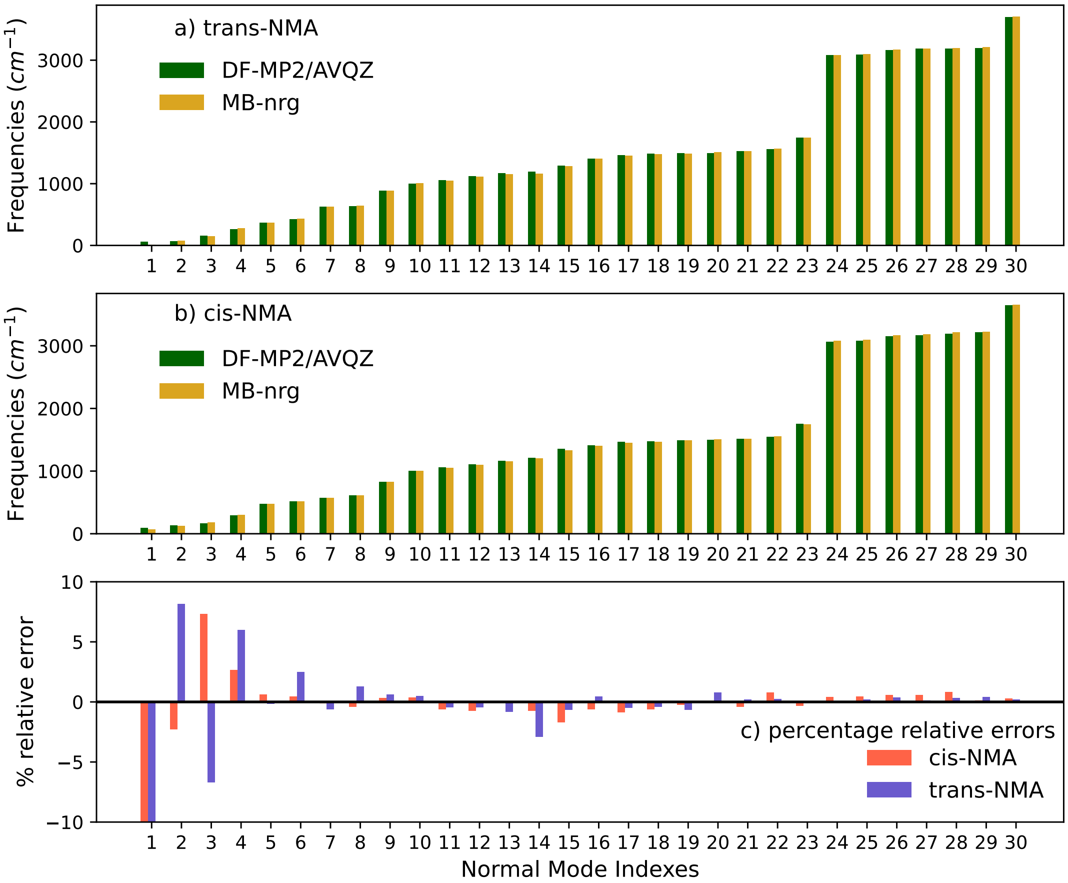

Figure 3: Comparisons between the DF-MP2/AVQZ reference harmonic frequencies (green) and the corresponding values calculated with the MB-nrg PEF (yellow) for the *cis* (a) and *trans* (b) isomers of NMA in the gas phase. Shown in c) are the percentage errors associated with the MB-nrg harmonic frequencies relative to the DF-MP2/AVQZ reference values for both isomers.

bond, respectively, as discussed above. The MB-nrg PEF accurately reproduces the harmonic frequencies of both amide I (normal mode 23) and amide II (normal mode 22) vibrations, which highlights the prospect for applying the MB-nrg PEFs[107] to simulations of vibrational spectra of biomolecules.[169,170]

To further demonstrate the ability of the MB-nrg PEF to correctly represent the overall multidimensional 1-body energy landscape of NMA, Figure 4 shows a comparison between the reference DF-MP2/AVQZ and MB-nrg energies along the isomerization pathway between the *cis* and *trans* configurations. The reference structures for this analysis were obtained from partial optimizations carried out at the DF-MP2/AVQZ level of theory in which the C9-C7-N5-C4 dihedral angle (Figure 1) was uniformly varied from the *cis* to the *trans* configuration while optimizing all other degrees of freedom. The comparison shown in Figure 4 demonstrates that the MB-nrg PEF is able to quantitatively reproduce the reference DF-MP2/AVQZ energies for configurations along the isomerization path that are not explicitly included in the training set, which provides evidence for the ability of the MB-nrg PEF to extrapolate outside the training set and thus accurately represent the global 1-body NMA multidimensional energy landscape.

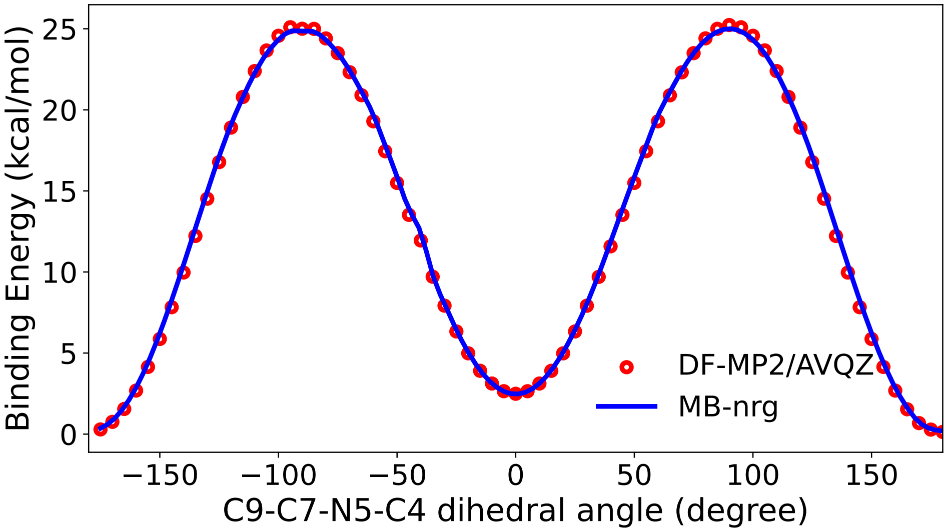

Figure 4: Energy variation along the *cis*–*trans* isomerization path of NMA described by the C9-C7-N5-C4 dihedral angle. The DF-MP2/AVQZ reference energies are shown as red circles and the corresponding MB-nrg values are shown as a blue line.

As discussed in Ref. 107, it should be emphasized that, by construction, the accuracy of any MB-nrg PEF is systematically improvable, e.g., by training higher-degree PIPs on larger training sets and/or using a higher level of electronic structure theory in the calculations of the reference energies and/or expanding the training sets. The incremental refinement of the present MB-nrg PEF for NMA, however, lies outside the scope of this study. Our primary objective instead is to demonstrate that the MB-nrg formalism can be extended to enable predictive simulations of biomolecules both in the gas phase and in solution.

NMA–H_2O dimer

Figure 5 shows the variation of the NMA–H_2O 2-body energy (i.e., interaction energy) along different radial scans for both the *cis* and *trans* isomers of NMA. Specifically, the panels in the left column (a, c, e, and g) show radial scans from the two methyl groups of NMA, while the panels in the right column (b, d, f, h) show radial scans from the carbonyl (b, f) and amino (d, h) groups. In all cases, the MB-nrg PEF accurately reproduces the DF-MP2/AVQZ energies. It should be noted that, although the deviations from the reference data become relatively larger at shorter distances, these configurations are less relevant in actual computer simulations given the associated interaction energy that is significantly larger than the thermal energy at ambient conditions ($k_BT = 0.59$ kcal/mol at 300 K, with k_B being Boltzmann's constant).

On the other hand, the performance of the TTM-nrg PEF clearly demonstrates the limitations of a classical representation of molecular interactions. In particular, TTM-nrg is unable to accurately reproduce the DF-MP2/AVQZ reference data at short distances, where the monomer's electron densities overlap, but is able to correctly describe long-range interactions, which primarily depend on many-body electrostatics and London dispersion forces.

NMA in liquid water

The last aspect that remains to be addressed is whether the high accuracy demonstrated by the MB-nrg PEF in representing the DF-MP2/AVQZ energy landscape of both an isolated NMA molecule

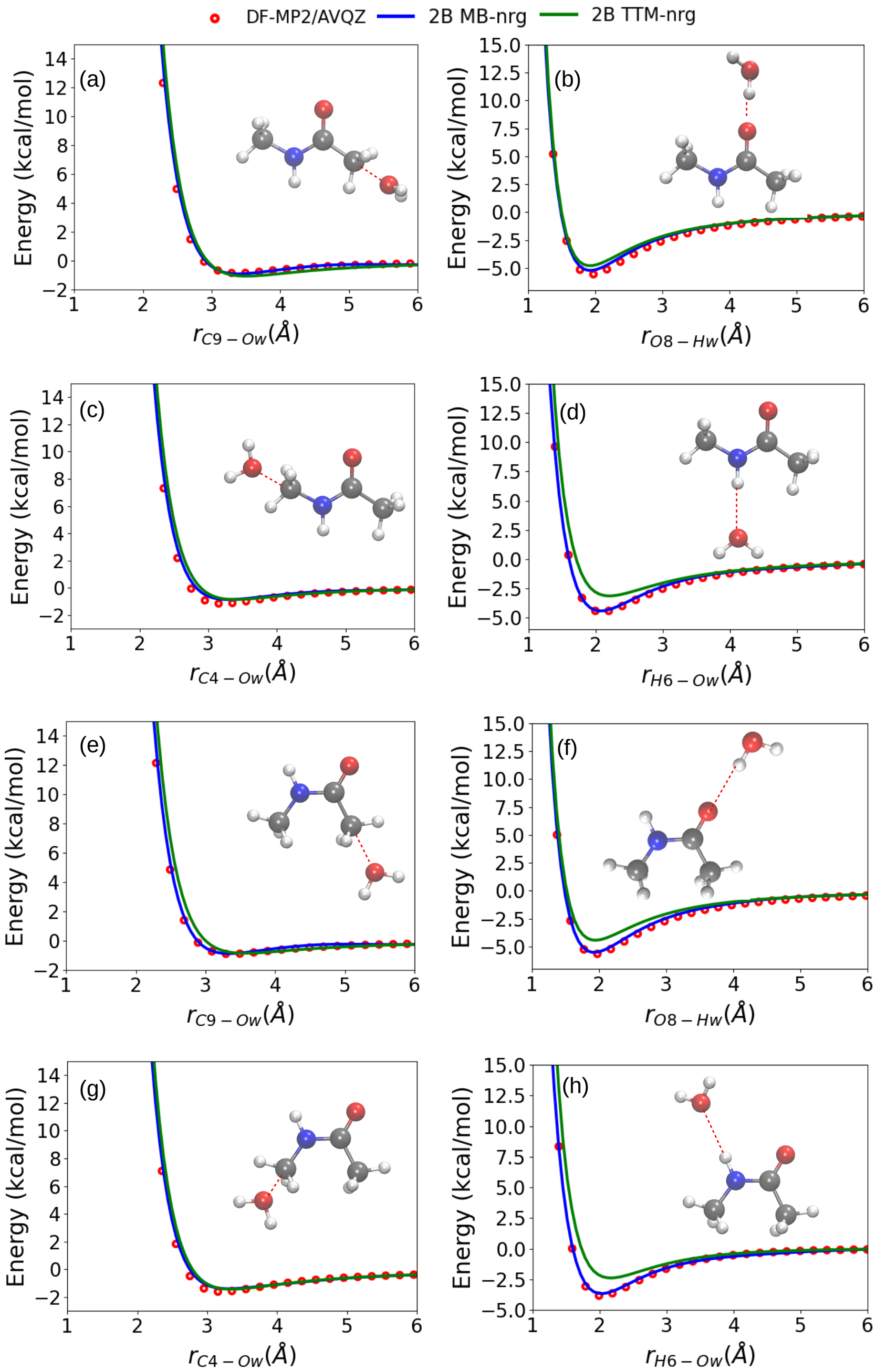

Figure 5: NMA–H_2O energy scans along several distances between the water molecule and different atoms of the *cis* (a-d) and *trans* (e-h) isomers of NMA. The DF-MP2/AVQZ reference energies are shown as red circles, while the corresponding MB-nrg and TTM-nrg values are shown as blue and green lines, respectively.

and an NMA–H_2O dimer in the gas phase is transferable to computer simulations of NMA in solution. To this end, we investigated the hydration structure of NMA as predicted by MD simulations carried out in periodic boundary conditions with both the TTM-nrg and MB-nrg PEFs at 298 K and 1 atm for a $\sim$0.2 M NMA solution (1 NMA molecule in a box of 277 water molecules). Besides the TTM-nrg and MB-nrg PEFs, our analyses also include results obtained from analogous MD simulations carried out with the empirical pairwise-additive ff14SB force field,[168] which is commonly used in biomolecular simulations, paired with the TIP4P-2005 water model.[156]

Fig. 6 shows the comparisons among the radial distribution functions (RDFs) between the oxygen atoms of the water molecules (Ow) and the H6, O8, and N5 atoms of NMA (see Fig. 1) calculated with the different models. Also shown are the RDFs obtained by applying the empirical potential structure refinement (EPSR) method to experimental data collected from neutron diffraction with isotopic substitution (NDIS) measurements carried out for a solution with a concentration of 1 mol of NMA per 15 mol of water. Overall, all models qualitatively reproduce the EPSR-based RDFs although some differences are noticeable. In particular, MB-nrg predicts more structured and somewhat tighter hydration shells, independently of the direction of approach and orientation of the water molecules. On the other hand, TTM-nrg predicts the least structured first hydration shell, which is particular evident in the case of the H6-Ow and O8-Hw RDFs. Interestingly, the empirical pairwise-additive potential ff14SB/TIP4P-2005 appears to resemble MB-nrg at short distances,

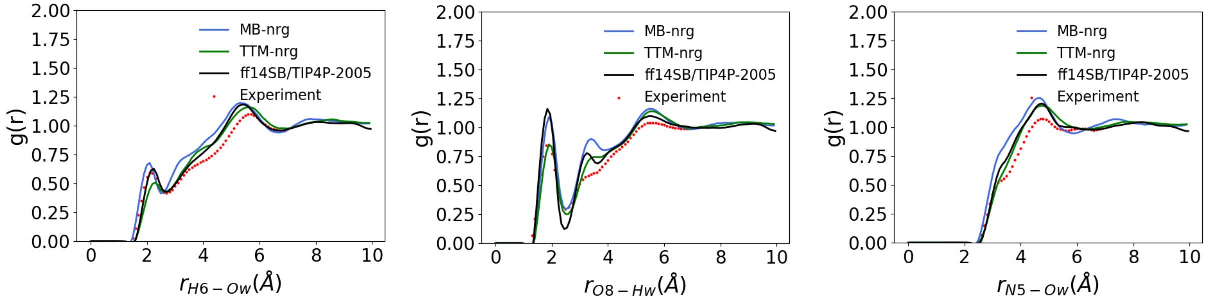

Figure 6: Radial distribution functions between different pairs of atoms located on the NMA and H_2O molecules calculated with the MB-nrg (blue) and TTM-nrg (green) PEFs, and the ff14SB/TIP4P-2005 force field (yellow). Also shown are the EPSR-based radial distribution functions derived from NDIS measurements.[171] See main text for details.

particularly within the first solvation shell of H6 and O8, but behaves more closely to TTM-nrg at larger distances. The hydrogen-bond lengths predicted by MB-nrg PEF are in agreement with values reported in the literature from hybrid quantum mechanics/molecular mechanics (QM/MM) simulations[172] and *ab initio* calculations.[173] Importantly, the TTM-nrg PEF underestimates the H6-Ow bond strength, which is consistent with the correlation plot reported in Fig. 2b, showing that the TTM-nrg PEF tends to underestimate the strength of attractive 2-body energies.

The differences between the EPSR-based and MB-nrg RDFs may be due to various reasons, including inaccuracies of the MB-nrg PEF which may be related to intrinsic inaccuracies of the DF-MP2/AVQZ training data and/or inaccuracies in the description of NMA–H_2O n-body energies with $n > 2$ which are only represented by many-body polarization in the current MB-nrg PEF. In this regard, it was shown explicitly representing 3-body energies with a 3-body PIP, in some cases, slightly improves the agreement with experimental measurements of extended X-ray absorption fine structure (EXAFS) spectra.[102,104,126,127] Another possible reason for the differences between the EPSR-based and MB-nrg RDFs may be related to the higher NMA concentration used in the experiments. In this regard, from the analyses of the NDIS measurements reported in Ref. 171 it was concluded that, at the concentration used in the experiments, NMA molecules can form hydrogen-bonded dimers and possibly chains. Since all MD simulations reported in this study were carried out at an NMA concentration that is $\sim$18 times lower than in the experiments of Ref. 171, the differences in the EPSR-based and MB-nrg RDFs can also possibly be due to the absence of NMA dimers or chains in the MD simulations. Both the role of 3-body interactions and NMA concentration will be the subject of future studies.

Conclusions

In this study, we introduced two data-driven many-body PEFs for NMA in water which were developed within the TTM-nrg and MB-nrg theoretical/computational frameworks. At the 1-body level, the TTM-nrg and MB-nrg PEFs adopt the same machine-learned PIP to represent the intramolecular distortion of an isolated NMA molecule. The 1-body training set includes configurations extracted from normal-mode sampling applied to minimum-energy and transition-state structures of NMA as well as configurations extracted from metadynamics simulations, which were then screened using furthest point sampling. The reference energies were calculated at the DF-MP2/AVQZ level of theory in combination with the aug-cc-pVQZ basis set. It was shown that the 1-body PEF accurately reproduces the DF-MP2/AVQZ multidimensional potential energy surface of an isolated NMA molecule, being able to correctly describe the energetics and normal modes of both *cis* and *trans* isomers as well as the isomerization path.

The 2-body NMA–H_2O training set was developed following the same protocol used for the 1-body training set. While the TTM-nrg PEF adopts conventional Born-Mayer potentials to describe repulsive NMA–H_2O interactions, the MB-nrg PEF adopts a machine-learned PIP that effectively represents short-range quantum-mechanical interactions. In order to guarantee fast computation of the 2-body PIP, we applied a specific filtering scheme that systematically removes the least relevant terms. From the analysis of various NMA–H_2O radial scans, we demonstrated that the MB-nrg PEF closely reproduces the DF-MP2/AVQZ reference interaction energies, independently of the direction of approach and orientation of the water molecule. In contrast, the TTM-nrg PEF was found to deviate appreciably from the DF-MP2/AVQZ reference interaction energies at short NMA–H_2O distances. These deviations can be traced back to the inability of classical-like functional forms (i.e., Born-Mayer potentials and many-body electrostatics) adopted by the TTM-nrg PEF to correctly describe quantum-mechanical interactions that arise from the overlap of the NMA and water electron densities at close range.

Finally, we analyzed the hydration structure of a dilute solution ($\sim$0.2 M) of NMA in water. Overall, both TTM-nrg and MB-nrg PEFs qualitatively reproduce the EPSR-based RDFs derived from neutron diffraction measurements. Interestingly, the MB-nrg PEF was found to predict more structured and somewhat more strongly bound hydration shells compared to the TTM-nrg PEF. These differences are in line with the conclusions drawn from the analysis of the NMA–H_2O dimer, indicating that the TTM-nrg PEF overall underestimates the strength of NMA–H_2O interactions. Some differences were found between the EPSR-based and MB-nrg RDFs. Since the experimental measurements were carried out for a much higher NMA concentration, these differences are likely due to the presence of hydrogen-bonded NMA dimers and possibly NMA chains is the experimental solution.

Future studies will focus on extending the MB-nrg theoretical/computational framework to the modeling generic peptides with the ultimate goal of developing a transferable, “first principles” data-driven many-body force field for biomolecules.

ASSOCIATED CONTENT

Supporting Information

Details on the 2-body NMA–H_2O permutationally invariant polynomial, list of the TTM-nrg and MB-nrg parameters, and additional comparisons of structural parameters and harmonic frequencies calculated with different NMA models.

ACKNOWLEDGEMENT

This research was supported by the Air Force Office of Scientific Research thorugh grant no. FA9550-20-1-0351 and the National Science Foundation through grant no. CHE-2102309. Computational resources were provided by the Department of Defense High Performance Computing Modernization Program (HPCMP), the Extreme Science and Engineering Discovery Environment (XSEDE), which was supported by the National Science Foundation grant no. ACI-1548562, and the Triton Shared Computing Cluster (TSCC) at the San Diego Supercomputer Center (SDSC).

Data availability

Any data generated and analyzed in this study are available from the authors upon request.

References

(1) Franks, F. *Water: A Matrix of Life*; Royal Society of Chemistry, 2000; Vol. 21.

(2) Ball, P. Water as an Active Constituent in Cell Biology. *Chem. Rev.* **2008**, *108*, 74–108.

(3) Petsko, G. A.; Ringe, D. *Protein Structure and Function*; New Science Press, 2004.

(4) Weiss, M. S.; Jabs, A.; Hilgenfeld, R. Peptide Bonds Revisited. *Nat. Struct. Biol.* **1998**, *5*, 676–676.

(5) Benkovic, S. J.; Hammes-Schiffer, S. A Perspective on Enzyme Catalysis. *Science* **2003**, *301*, 1196–1202.

(6) Forterre, P. The Origin of DNA Genomes and DNA Replication Proteins. *Curr. Opin. Microbiol.* **2002**, *5*, 525–532.

(7) Du, N.; Ye, F.; Sun, J.; Liu, K. Stimuli-Responsive Natural Proteins and Their Applications. *ChemBioChem* **2022**, *23*, e202100416.

(8) Griffith, J. K.; Baker, M. E.; Rouch, D. A.; Page, M. G.; Skurray, R. A.; Paulsen, I. T.; Chater, K. F.; Baldwin, S. A.; Henderson, P. J. Membrane Transport Proteins: Implications of Sequence Comparisons. *Curr. Opin. Cell Biol.* **1992**, *4*, 684–695.

(9) Laage, D.; Elsaesser, T.; Hynes, J. T. Water Dynamics in the Hydration Shells of Biomolecules. *Chem. Rev.s* **2017**, *117*, 10694–10725.

(10) Prabhu, N.; Sharp, K. Protein-Solvent Interactions. *Chem. Rev.* **2006**, *106*, 1616–1623.

(11) Bellissent-Funel, M.-C.; Hassanali, A.; Havenith, M.; Henchman, R.; Pohl, P.; Sterpone, F.; Van Der Spoel, D.; Xu, Y.; Garcia, A. E. Water Determines the Structure and Dynamics of Proteins. *Chem. Rev.* **2016**, *116*, 7673–7697.

(12) Chiti, F.; Dobson, C. M. Protein Misfolding, Functional Amyloid, and Human Disease. *Annu. Rev. Biochem.* **2006**, *75*, 333–366.

(13) Goldsbury, C.; Baxa, U.; Simon, M. N.; Steven, A. C.; Engel, A.; Wall, J. S.; Aebi, U.; Müller, S. A. Amyloid Structure and Assembly: Insights from Scanning Transmission Electron Microscopy. *J. Struct. Biol.* **2011**, *173*, 1–13.

(14) Wiltzius, J. J.; Sievers, S. A.; Sawaya, M. R.; Cascio, D.; Popov, D.; Riekel, C.; Eisenberg, D. Atomic Structure of the Cross-β Spine of Islet Amyloid Polypeptide (Amylin). *Protein Sci.* **2008**, *17*, 1467–1474.

(15) Awile, O.; Krisko, A.; Sbalzarini, I. F.; Zagrovic, B. Intrinsically Disordered Regions may Lower the Hydration Free Energy in Proteins: A Case Study of Nudix Hydrolase in the Bacterium Deinococcus Radiodurans. *PLoS Comput. Biol.* **2010**, *6*, e1000854.

(16) Bagchi, B. Water Dynamics in the Hydration Layer around Proteins and Micelles. *Chem. Rev.* **2005**, *105*, 3197–3219.

(17) Pal, S. K.; Zewail, A. H. Dynamics of Water in Biological Recognition. *Chem. Rev.* **2004**, *104*, 2099–2124.

(18) McCammon, J. A.; Gelin, B. R.; Karplus, M. Dynamics of Folded Proteins. *Nature* **1977**, *267*, 585–590.

(19) Karplus, M.; McCammon, J. A. Molecular Dynamics Simulations of Biomolecules. *Nat. Struct. Biol.* **2002**, *9*, 646–652.

(20) Vitalis, A.; Pappu, R. V. Methods for Monte Carlo Simulations of Biomacromolecules. *Annu. Rep. Comput. Chem.* **2009**, *5*, 49–76.

(21) Noid, W. G. Perspective: Coarse-Grained Models for Biomolecular Systems. *J. Chem. Phys.* **2013**, *139*, 09B201_1.

(22) Klepeis, J. L.; Lindorff-Larsen, K.; Dror, R. O.; Shaw, D. E. Long-timescale Molecular Dynamics Simulations of Protein Structure and Function. *Curr. Opin. Struct. Biol.* **2009**, *19*, 120–127.

(23) Saiz, L.; Klein, M. L. Computer Simulation Studies of Model Biological Membranes. *Acc. Chem. Res.* **2002**, *35*, 482–489.

(24) Ponder, J. W.; Case, D. A. Force Fields for Protein Simulations. *Adv. Protein Chem.* **2003**, *66*, 27–85.

(25) Lifson, S.; Warshel, A. Consistent Force Field for Calculations of Conformations, Vibrational Spectra, and Enthalpies of Cycloalkane and *n*-Alkane Molecules. *J. Chem. Phys.* **1968**, *49*, 5116–5129.

(26) Warshel, A.; Lifson, S. Consistent Force Field Calculations. II. Crystal Structures, Sublimation Energies, Molecular and Lattice Vibrations, Molecular Conformations, and Enthalpies of Alkanes. *J. Chem. Phys.* **1970**, *53*, 582–594.

(27) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.

(28) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *J. Comput. Chem.* **1983**, *4*, 187–217.

(29) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, *79*, 926–935.

(30) Scott, W. R.; Hünenberger, P. H.; Tironi, I. G.; Mark, A. E.; Billeter, S. R.; Fennen, J.; Torda, A. E.; Huber, T.; Krüger, P.; Van Gunsteren, W. F. The GROMOS Biomolecular Simulation Program Package. *J. Phys. Chem. A* **1999**, *103*, 3596–3607.

(31) MacKerell Jr, A. D. Empirical Force Fields for Biological Macro-Molecules: Overview and Issues. *J. Comput. Chem.* **2004**, *25*, 1584–1604.

(32) Ponder, J. W.; Wu, C.; Ren, P.; Pande, V. S.; Chodera, J. D.; Schnieders, M. J.; Haque, I.; Mobley, D. L.; Lambrecht, D. S.; DiStasio Jr, R. A., et al. Current Status of the AMOEBA Polarizable Force Field. *J. Phys. Chem. B* **2010**, *114*, 2549–2564.

(33) Vanommeslaeghe, K.; MacKerell Jr., A. CHARMM Additive and Polarizable Force Fields for Biophysics and Computer-Aided Drug Design. *Biochim. Biophys. Acta - Gen.* **2015**, *1850*, 861–871.

(34) Jana, M.; MacKerell Jr, A. D. CHARMM Drude Polarizable Force Field for Aldopentofuranoses and Methyl-aldopentofuranosides. *J. Phys. Chem. B* **2015**, *119*, 7846–7859.

(35) He, X.; Lopes, P. E.; MacKerell Jr, A. D. Polarizable Empirical Force Field for Acyclic Polyalcohols Based on the Classical Drude Oscillator. *Biopolymers* **2013**, *99*, 724–738.

(36) Patel, D. S.; He, X.; MacKerell Jr, A. D. Polarizable Empirical Force Field for Hexopyranose Monosaccharides Based on the Classical Drude Oscillator. *J. Phys. Chem. B* **2015**, *119*, 637–652.

(37) Anisimov, V. M.; Lamoureux, G.; Vorobyov, I. V.; Huang, N.; Roux, B.; MacKerell, A. D. Determination of Electrostatic Parameters for a Polarizable Force Field Based on the Classical Drude Oscillator. *J. Chem. Theory Comput.* **2005**, *1*, 153–168.

(38) Shi, Y.; Xia, Z.; Zhang, J.; Best, R.; Wu, C.; Ponder, J. W.; Ren, P. Polarizable Atomic Multipole-Based AMOEBA Force Field for Proteins. *J. Chem. Theory Comput.* **2013**, *9*, 4046–4063.

(39) Wang, Z.-X.; Zhang, W.; Wu, C.; Lei, H.; Cieplak, P.; Duan, Y. Strike a Balance: Optimization of Backbone Torsion Parameters of AMBER Polarizable Force Field for Simulations of Proteins and Peptides. *J. Comput. Chem.* **2006**, *27*, 781–790.

(40) Unke, O. T.; Chmiela, S.; Sauceda, H. E.; Gastegger, M.; Poltavsky, I.; Schütt, K. T.; Tkatchenko, A.; Müller, K.-R. Machine Learning Force Fields. *Chem. Rev.* **2021**, *121*, 10142–10186.

(41) Kocer, E.; Ko, T. W.; Behler, J. Neural Network Potentials: A Concise Overview of Methods. *Annu. Rev. Phys. Chem.* **2022**, *73*, 163–186.

(42) Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1: An Extensible Neural Network Potential with DFT Accuracy at Force Field Computational Cost. *Chem. Sci.* **2017**, *8*, 3192–3203.

(43) Devereux, C.; Smith, J. S.; Huddleston, K. K.; Barros, K.; Zubatyuk, R.; Isayev, O.; Roitberg, A. E. Extending the Applicability of the ANI Deep Learning Molecular Potential to Sulfur and Halogens. *J. Chem. Theory Comput.* **2020**, *16*, 4192–4202.

(44) Wang, H.; Zhang, L.; Han, J.; Weinan, E. DeePMD-kit: A Deep Learning Package for Many-Body Potential Energy Representation and Molecular Dynamics. *Comput. Phys. Commun.* **2018**, *228*, 178–184.

(45) Yao, K.; Herr, J. E.; Toth, D. W.; Mckintyre, R.; Parkhill, J. The TensorMol-0.1 Model Chemistry: a Neural Network Augmented with Long-Range Physics. *Chem. Sci.* **2018**, *9*, 2261–2269.

(46) Zhang, Y.; Hu, C.; Jiang, B. Embedded Atom Neural Network Potentials: Efficient and Accurate Machine Learning with a Physically Inspired Representation. *J. Phys. Chem. Lett.* **2019**, *10*, 4962–4967.

(47) Xie, X.; Persson, K. A.; Small, D. W. Incorporating Electronic Information into Machine Learning Potential Energy Surfaces via Approaching the Ground-State Electronic Energy as a Function of Atom-Based Electronic Populations. *J. Chem. Theory Comput.* **2020**, *16*, 4256–4270.

(48) Ko, T. W.; Finkler, J. A.; Goedecker, S.; Behler, J. A Fourth-Generation High-Dimensional Neural Network Potential with Accurate Electrostatics Including Non-Local Charge Transfer. *Nat. Commun.* **2021**, *12*, 398.

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

(50) Yao, K.; Herr, J. E.; Brown, S. N.; Parkhill, J. Intrinsic Bond Energies from a Bonds-in-Molecules Neural Network. *J. Phys. Chem. Lett.* **2017**, *8*, 2689–2694.

(51) Glick, Z. L.; Metcalf, D. P.; Koutsoukas, A.; Spronk, S. A.; Cheney, D. L.; Sherrill, C. D. AP-Net: An Atomic-Pairwise Neural Network for Smooth and Transferable Interaction Potentials. *J. Chem. Phys.* **2020**, *153*, 044112.

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

(53) Lubbers, N.; Smith, J. S.; Barros, K. Hierarchical Modeling of Molecular Energies using a Deep Neural Network. *J. Chem. Phys.* **2018**, *148*, 241715.

(54) Schütt, K.; Kindermans, P.-J.; Sauceda Felix, H. E.; Chmiela, S.; Tkatchenko, A.; Müller, K.-R. SchNet: A Continuous-Filter Convolutional Neural Network for Modeling Quantum Interactions. *Adv. Neural Inf. Process Syst.* **2017**, *30*.

(55) Schütt, K. T.; Arbabzadah, F.; Chmiela, S.; Müller, K. R.; Tkatchenko, A. Quantum-Chemical Insights from Deep Tensor Neural Networks. *Nat. Commun.* **2017**, *8*, 13890.

(56) Zubatyuk, R.; Smith, J. S.; Leszczynski, J.; Isayev, O. Accurate and Transferable Multitask Prediction of Chemical Properties with an Atoms-in-Molecules Neural Network. *Sci. Adv.* **2019**, *5*, eaav6490.

(57) Zhai, Y.; Caruso, A.; Bore, S. L.; Luo, Z.; Paesani, F. A “Short Blanket” Dilemma for a State-of-the-Art Neural Network Potential for Water: Reproducing Experimental Proper-

ties or the Physics of the Underlying Many-Body Interactions? *J. Chem. Phys.* **2023**, *158*, 084111.

(58) Paesani, F. Getting the Right Answers for the Right Reasons: Toward Predictive Molecular Simulations of Water with Many-Body Potential Energy Functions. *Acc. Chem. Res.* **2016**, *49*, 1844–1851.

(59) Lambros, E.; Dasgupta, S.; Palos, E.; Swee, S.; Hu, J.; Paesani, F. General Many-Body Framework for Data-Driven Potentials with Arbitrary Quantum Mechanical Accuracy: Water as a Case Study. *J. Chem. Theory Comput.* **2021**, *17*, 5635–5650.

(60) Prodan, E.; Kohn, W. Nearsightedness of Electronic Matter. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 11635–11638.

(61) Nesbet, R. Atomic Bethe-Goldstone Equations. *Adv. Chem. Phys.* **1969**, 1–34.

(62) Babin, V.; Leforestier, C.; Paesani, F. Development of a "First Principles" Water Potential with Flexible Monomers: Dimer Potential Energy Surface, VRT Spectrum, and Second Virial Coefficient. *J. Chem. Theory Comput.* **2013**, *9*, 5395–5403.

(63) Babin, V.; Medders, G. R.; Paesani, F. 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.* **2014**, *10*, 1599–1607.

(64) Medders, G. R.; Babin, V.; Paesani, F. Development of a "First-Principles" Water Potential with Flexible Monomers. III. Liquid Phase Properties. *J. Chem. Theory Comput.* **2014**, *10*, 2906–2910.

(65) Medders, G. R.; Paesani, F. Infrared and Raman Spectroscopy of Liquid Water through "First-Principles" Many-Body Molecular Dynamics. *J. Chem. Theory Comput.* **2015**, *11*, 1145–1154.

(66) Richardson, J. O.; Pérez, C.; Lobsiger, S.; Reid, A. A.; Temelso, B.; Shields, G. C.; Kisiel, Z.; Wales, D. J.; Pate, B. H.; Althorpe, S. C. Concerted Hydrogen-Bond Breaking by Quantum Tunneling in the Water Hexamer Prism. *Science* **2016**, *351*, 1310–1313.

(67) Cole, W. T.; Farrell, J. D.; Wales, D. J.; Saykally, R. J. Structure and Torsional Dynamics of the Water Octamer from THz Laser Spectroscopy Near 215 μm. *Science* **2016**, *352*, 1194–1197.

(68) Mallory, J. D.; Mandelshtam, V. A. Diffusion Monte Carlo Studies of MB-pol $(H_2O)_{2-6}$ and $(D_2O)_{2-6}$ clusters: Structures and Binding Energies. *J. Chem. Phys.* **2016**, *145*, 064308.

(69) Videla, P. E.; Rossky, P. J.; Laria, D. Communication: Isotopic Effects on Tunneling Motions in the Water Trimer. *J. Chem. Phys.* **2016**, *144*, 061101.

(70) Reddy, S. K.; Straight, S. C.; Bajaj, P.; Huy Pham, C.; Riera, M.; Moberg, D. R.; Morales, M. A.; Knight, C.; Götz, A. W.; Paesani, F. On the Accuracy of the MB-pol Many-Body Potential for Water: Interaction Energies, Vibrational Frequencies, and Classical Thermodynamic and Dynamical Properties from Clusters to Liquid Water and Ice. *J. Chem. Phys.* **2016**, *145*, 194504.

(71) Medders, G. R.; Paesani, F. Dissecting the Molecular Structure of the Air/Water Interface from Quantum Simulations of the Sum-Frequency Generation Spectrum. *J. Am. Chem. Soc.* **2016**, *138*, 3912–3919.

(72) Straight, S. C.; Paesani, F. Exploring Electrostatic Effects on the Hydrogen Bond Network of Liquid Water Through Many-Body Molecular Dynamics. *J. Phys. Chem. B* **2016**, *120*, 8539–8546.

(73) Pham, C. H.; Reddy, S. K.; Chen, K.; Knight, C.; Paesani, F. Many-Body Interactions in Ice. *J. Chem. Theory Comput.* **2017**, *13*, 1778–1784.

(74) Moberg, D. R.; Straight, S. C.; Knight, C.; Paesani, F. Molecular Origin of the Vibrational Structure of Ice I_h. *J. Phys. Chem. Lett.* **2017**, *8*, 2579–2583.

(75) Brown, S. E.; Götz, A. W.; Cheng, X.; Steele, R. P.; Mandelshtam, V. A.; Paesani, F. Monitoring Water Clusters "melt" Through Vibrational Spectroscopy. *J. Am. Chem. Soc.* **2017**, *139*, 7082–7088.

(76) Vaillant, C. L.; Cvitaš, M. T. Rotation-Tunneling Spectrum of the Water Dimer from Instanton Theory. *Phys. Chem. Chem. Phys.* **2018**, *20*, 26809–26813.

(77) Vaillant, C.; Wales, D.; Althorpe, S. Tunneling Splittings from Path-Integral Molecular Dynamics Using a Langevin Thermostat. *J. Chem. Phys.* **2018**, *148*, 234102.

(78) Schmidt, M.; Roy, P.-N. Path Integral Molecular Dynamic Simulation of Flexible Molecular Systems in Their Ground State: Application to the Water Dimer. *J. Chem. Phys.* **2018**, *148*, 124116.

(79) Bishop, K. P.; Roy, P.-N. 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.* **2018**, *148*, 102303.

(80) Videla, P. E.; Rossky, P. J.; Laria, D. Isotopic Equilibria in Aqueous Clusters at Low Temperatures: Insights from the MB-pol Many-Body Potential. *J. Chem. Phys.* **2018**, *148*, 084303.

(81) Hunter, K. M.; Shakib, F. A.; Paesani, F. Disentangling Coupling Effects in the Infrared Spectra of Liquid Water. *J. Phys. Chem. B* **2018**, *122*, 10754–10761.

(82) Moberg, D. R.; Sharp, P. J.; Paesani, F. Molecular-Level Interpretation of Vibrational Spectra of Ordered Ice Phases. *J. Phys. Chem. B* **2018**, *122*, 10572–10581.

(83) Moberg, D. R.; Becker, D.; Dierking, C. W.; Zurheide, F.; Bandow, B.; Buck, U.; Hudait, A.; Molinero, V.; Paesani, F.; Zeuch, T. The End of Ice I. *Proc. Natl. Acad. Sci. U.S.A.* **2019**, *116*, 24413–24419.

(84) Samala, N. R.; Agmon, N. Temperature Dependence of Intramolecular Vibrational Bands in Small Water Clusters. *J. Phys. Chem. B* **2019**, *123*, 9428–9442.

(85) Samala, N. R.; Agmon, N. Thermally Induced Hydrogen-Bond Rearrangements in Small Water Clusters and the Persistent Water Tetramer. *ACS omega* **2019**, *4*, 22581–22590.

(86) Cvitaš, M. T.; Richardson, J. O. Quantum Tunnelling Pathways of the Water Pentamer. *Phys. Chem. Chem. Phys.* **2020**, *22*, 1035–1044.

(87) Cruzeiro, V.; Wildman, A.; Li, X.; Paesani, F. Relationship Between Hydrogen-Bonding Motifs and the $1b_1$ Splitting in the X-ray Emission Spectrum of Liquid Water. *J. Phys. Chem. Lett.* **2021**, *12*, 3996–4002.

(88) Muniz, M. C.; Gartner III, T. E.; Riera, M.; Knight, C.; Yue, S.; Paesani, F.; Panagiotopoulos, A. Z. Vapor-Liquid Equilibrium of Water With the MB-pol Many-Body Potential. *The Journal of Chemical Physics* **2021**, *154*, 211103.

(89) Gartner III, T. E.; Hunter, K. M.; Lambros, E.; Caruso, A.; Riera, M.; Medders, G. R.; Panagiotopoulos, A. Z.; Debenedetti, P. G.; Paesani, F. Anomalies and Local Structure of Liquid Water from Boiling to the Supercooled Regime as Predicted by the Many-Body MB-pol Model. *J. Phys. Chem. Lett.* **2022**, *13*, 3652–3658.

(90) Bull-Vulpe, E. F.; Riera, M.; Götz, A. W.; Paesani, F. MB-Fit: Software Infrastructure for Data-Driven Many-Body Potential Energy Functions. *J. Chem. Phys.* **2021**, *155*, 124801.

(91) Riera, M.; Yeh, E. P.; Paesani, F. Data-Driven Many-Body Models for Molecular Fluids: CO_2/H_2O Mixtures as a Case Study. *J. Chem. Theory Comput.* **2020**, *16*, 2246–2257.

(92) Yue, S.; Riera, M.; Ghosh, R.; Panagiotopoulos, A. Z.; Paesani, F. Transferability of Data-Driven, Many-Body Models for CO_2 Simulations in the Vapor and Liquid Phases. *J. Chem. Phys.* **2022**, *156*, 104503.

(93) Riera, M.; Hirales, A.; Ghosh, R.; Paesani, F. Data-Driven Many-Body Models with Chemical Accuracy for CH_4/H_2O Mixtures. *J. Chem. Phys. B* **2020**, *124*, 11207–11221.

(94) Robinson, V. N.; Ghosh, R.; Egan, C. K.; Riera, M.; Knight, C.; Paesani, F.; Hassanali, A. The Behavior of Methane–Water Mixtures Under Elevated Pressures from Simulations Using Many-Body Potentials. *J. Chem. Phys.* **2022**, *156*, 194504.

(95) Bajaj, P.; Gotz, A. W.; Paesani, F. 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.* **2016**, *12*, 2698–2705.

(96) Riera, M.; Mardirossian, N.; Bajaj, P.; Götz, A. W.; Paesani, F. Toward Chemical Accuracy in the Description of Ion–Water Interactions through Many-Body Representations. Alkali-Water Dimer Potential Energy Surfaces. *J. Chem. Phys.* **2017**, *147*, 161715.

(97) Bajaj, P.; Wang, X.-G.; CarringtonJr, T.; Paesani, F. Vibrational Spectra of Halide-Water Dimers: Insights on Ion Hydration from Full-Dimensional Quantum Calculations on Many-Body Potential Energy Surfaces. *J. Chem. Phys.* **2017**, *148*, 102321.

(98) Riera, M.; Brown, S. E.; Paesani, F. 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* **2018**, *122*, 5811–5821.

(99) Bajaj, P.; Riera, M.; Lin, J. K.; Mendoza Montijo, Y. E.; Gazca, J.; Paesani, F. Halide Ion Microhydration: Structure, Energetics, and Spectroscopy of Small Halide–Water Clusters. *J. Phys. Chem. A* **2019**, *123*, 2843–2852.

(100) Bajaj, P.; Richardson, J. O.; Paesani, F. Ion-Mediated Hydrogen-Bond Rearrangement through Tunnelling in the Iodide–Dihydrate Complex. *Nat. Chem.* **2019**, *11*, 367.

(101) Bajaj, P.; Zhuang, D.; Paesani, F. Specific Ion Effects on Hydrogen-Bond Rearrangements in the Halide–Dihydrate Complexes. *J. Phys. Chem. Lett.* **2019**, *10*, 2823–2828.

(102) Zhuang, D.; Riera, M.; Schenter, G. K.; Fulton, J. L.; Paesani, F. Many-Body Effects Determine the Local Hydration Structure of Cs^+ in Solution. *J. Phys. Chem. Lett.* **2019**, *10*, 406–412.

(103) Riera, M.; Talbot, J. J.; Steele, R. P.; Paesani, F. Infrared Signatures of Isomer Selectivity and Symmetry Breaking in the $Cs^+(H_2O)_3$ Complex Using Many-Body Potential Energy Functions. *J. Chem. Phys.* **2020**, *153*, 044306.

(104) Caruso, A.; Paesani, F. Data-Driven Many-Body Models Enable a Quantitative Description of Chloride Hydration From Clusters to Bulk. *J. Chem. Phys.* **2021**, *155*, 064502.

(105) Cruzeiro, V. W. D.; Lambros, E.; Riera, M.; Roy, R.; Paesani, F.; Gotz, A. W. Highly Accurate Many-Body Potentials for Simulations of N_2O_5 in Water: Benchmarks, Development, and Validation. *J. Chem. Theory Comput.* **2021**, *17*, 3931–3945.

(106) Cruzeiro, V. W. D.; Galib, M.; Limmer, D. T.; Götz, A. W. Uptake of N_2O_5 by Aqueous Aerosol Unveiled Using Chemically Accurate Many-Body Potentials. *Nat. Commun.* **2022**, *13*, 1–7.

(107) Bull-Vulpe, E. F.; Riera, M.; Bore, S. L.; Paesani, F. Data-Driven Many-Body Potential Energy Functions for Generic Molecules: Linear Alkanes as a Proof-of-Concept Application. *J. Chem. Theory Comput.* **2022**,

(108) Li, Z.-L.; Zhong, F.-Y.; Zhou, L.-S.; Tian, Z.-Q.; Huang, K. Deep Eutectic Solvents Formed by N-Methylacetamide and Heterocyclic Weak Acids for Highly Efficient and Reversible Chemical Absorption of Ammonia. *Ind. Eng. Chem. Res.* **2020**, *59*, 2060–2067.

(109) Ciszewski, L.; Waykole, L.; Prashad, M.; Repić, O. A Practical Synthesis of 2-Arylamino-6-Alkylaminopurines from 2, 6-Dichloropurine. *Org. Process Res. Dev.* **2006**, *10*, 799–802.

(110) Akhter, M. S.; Alawi, S. M. Micellar Behaviour of Cetyltrimethylammonium Bromide in N-

methyl Acetamide—Alkanol and N, N-Dimethyl Acetamide—Alkanol Mixtures. *Colloids Surf. A Physicochem. Eng. Asp.* **2002**, *196*, 163–174.

(111) Buck, M.; Karplus, M. Hydrogen Bond Energetics: A Simulation and Statistical Analysis of N-Methyl Acetamide (NMA), Water, and Human Lysozyme. *J. Phys. Chem. B* **2001**, *105*, 11000–11015.

(112) Cieplak, P.; Kollman, P. On the Use of Electrostatic Potential Derived Charges in Molecular Mechanics Force Fields. The Relative Solvation Free Energy of *Cis*- and *Trans*-N-Methyl-Acetamide. *J. Comput. Chem.* **1991**, *12*, 1232–1236.

(113) Chang, W.; Wan, H.; Guan, G.; Yao, H. Isobaric Vapor–Liquid Equilibria for Water+Acetic Acid+(N-Methyl Pyrrolidone or N-Methyl Acetamide). *Fluid Ph. Equilibria.* **2006**, *242*, 204–209.

(114) Kaledin, A.; Bowman, J. Full Dimensional Quantum Calculations of Vibrational Energies of N-Methyl Acetamide. *J. Phys. Chem. A* **2007**, *111*, 5593–5598.

(115) Bradbury, E.; Elliott, A. The Infra-Red Spectrum of Crystalline N-Methyl Acetamide. *Spectrochim. Acta* **1963**, *19*, 995–1012.

(116) Qu, C.; Bowman, J. M. A Fragmented, Permutationally Invariant Polynomial Approach for Potential Energy Surfaces of Large Molecules: Application to N-Methyl Acetamide. *J. Chem. Phys.* **2019**, *150*, 141101.

(117) Nandi, A.; Qu, C.; Bowman, J. M. Full and Fragmented Permutationally Invariant Polynomial Potential Energy Surfaces for *Trans* and *cis* N-Methyl Acetamide and Isomerization Saddle Points. *J. Chem. Phys.* **2019**, *151*, 084306.

(118) Nandi, A.; Qu, C.; Houston, P. L.; Conte, R.; Bowman, J. M. 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.* **2021**, *154*, 051102.

(119) Hankins, D.; Moskowitz, J.; Stillinger, F. Water Molecule Interactions. *J. Chem. Phys.* **1970**, *53*, 4544–4554.

(120) Stoll, H. Correlation Energy of Diamond. *Phys. Rev. B* **1992**, *46*, 6700.

(121) Stoll, H. On the Correlation Energy of Graphite. *J. Chem. Phys.* **1992**, *97*, 8449–8454.

(122) Stoll, H. The Correlation Energy of Crystalline Silicon. *Chem. Phys. Lett.* **1992**, *191*, 548–552.

(123) Bukowski, R.; Szalewicz, K.; Groenenboom, G. C.; Van der Avoird, A. Predictions of the Properties of Water from First Principles. *Science* **2007**, *315*, 1249–1252.

(124) Wang, Y.; Huang, X.; Shepler, B. C.; Braams, B. J.; Bowman, J. M. Flexible Ab Initio Potential and Dipole Moment Surfaces for Water. I. Tests and Applications for Clusters Up to the 22-Mer. *J. Chem. Phys.* **2011**, *134*, 094509.

(125) Wang, Y.; Bowman, J. M. Ab Initio Potential and Dipole Moment Surfaces for Water. II. Local-Monomer Calculations of the Infrared Spectra of Water Clusters. *J. Chem. Phys.* **2011**, *134*, 154510.

(126) Caruso, A.; Zhu, X.; Fulton, J. L.; Paesani, F. Accurate Modeling of Bromide and Iodide Hydration with Data-Driven Many-Body Potentials. *J. Phys. Chem. B* **2022**, *126*, 8266–8278.

(127) Zhuang, D.; Riera, M.; Zhou, R.; Deary, A.; Paesani, F. Hydration Structure of Na+ and K+ Ions in Solution Predicted by Data-Driven Many-Body Potentials. *J. Phys. Chem. B* **2022**, *126*, 9349–9360.

(128) Reddy, S. K.; Moberg, D. R.; Straight, S. C.; Paesani, F. Temperature-Dependent Vibrational Spectra and Structure of Liquid Water from Classical and Quantum Simulations With the MB-pol Potential Energy Function. *J. Chem. Phys.* **2017**, *147*, 244504.

(129) Sun, Z.; Zheng, L.; Chen, M.; Klein, M. L.; Paesani, F.; Wu, X. Electron-Hole Theory of the Effect of Quantum Nuclei on the X-ray Absorption Spectra of Liquid Water. *Phys. Rev. Lett.* **2018**, *121*, 137401.

(130) Gaiduk, A. P.; Pham, T. A.; Govoni, M.; Paesani, F.; Galli, G. Electron Affinity of Liquid Water. *Nat. Commun.* **2018**, *9*, 1–6.

(131) Moberg, D. R.; Straight, S. C.; Paesani, F. Temperature Dependence of the Air/Water Interface Revealed by Polarization Sensitive Sum-Frequency Generation Spectroscopy. *J. Phys. Chem. B* **2018**, *122*, 4356–4365.

(132) Sun, S.; Tang, F.; Imoto, S.; Moberg, D. R.; Ohto, T.; Paesani, F.; Bonn, M.; Backus, E. H.; Nagata, Y. Orientational Distribution of Free OH Groups of Interfacial Water is Exponential. *Phys. Rev. Lett.* **2018**, *121*, 246101.

(133) Sengupta, S.; Moberg, D. R.; Paesani, F.; Tyrode, E. Neat Water–Vapor Interface: Proton Continuum and the Nonresonant Background. *J. Phys. Chem. Lett.* **2018**, *9*, 6744–6749.

(134) Arismendi-Arrieta, D. J.; Riera, M.; Bajaj, P.; Prosmiti, R.; Paesani, F. i-TTM Model for Ab Initio-Based Ion–Water Interaction Potentials. 1. Halide–Water Potential Energy Functions. *J. Phys. Chem. B* **2015**, *120*, 1822–1832.

(135) Riera, M.; Götz, A. W.; Paesani, F. The i-TTM Model for Ab Initio-Based Ion–Water Interaction Potentials. II. Alkali Metal Ion–Water Potential Energy Functions. *Phys. Chem. Chem. Phys.* **2016**, *18*, 30334–30343.

(136) Marenich, A. V.; Jerome, S. V.; Cramer, C. J.; Truhlar, D. G. 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.* **2012**, *8*, 527–541.

(137) Epifanovsky, E.; Gilbert, A. T.; Feng, X.; Lee, J.; Mao, Y.; Mardirossian, N.; Pokhilko, P.; White, A. F.; Coons, M. P.; Dempwolff, A. L., et al. Software for the Frontiers of Quantum

Chemistry: An Overview of Developments in the Q-Chem 5 Package. *J. Chem. Phys.* **2021**, *155*, 084801.

(138) Mardirossian, N.; Head-Gordon, M. ωB97M-V: A Combinatorially Optimized, Range-Separated Hybrid, Meta-GGA Density Functional with VV10 Nonlocal Correlation. *J. Chem. Phys.* **2016**, *144*, 214110.

(139) Dunning Jr, T. H. Gaussian Basis Sets for Use in Correlated Molecular Calculations. I. The Atoms Boron through Neon and Hydrogen. *J. Chem. Phys.* **1989**, *90*, 1007–1023.

(140) Becke, A. D.; Johnson, E. R. Exchange-Hole Dipole Moment and the Dispersion Interaction. *J. Chem. Phys.* **2005**, *122*, 154104.

(141) Johnson, E. R.; Becke, A. D. A Post-Hartree–Fock Model of Intermolecular Interactions. *J. Chem. Phys.* **2005**, *123*, 024101.

(142) Johnson, E. R.; Becke, A. D. A Post-Hartree-Fock Model of Intermolecular Interactions: Inclusion of Higher-Order Corrections. *J. Chem. Phys.* **2006**, *124*, 174104.

(143) Stone, A. *The Theory of Intermolecular Forces*; Oxford University Press, Oxford, 2013.

(144) Paesani, F. Water: Many-Body Potential from First Principles (From the Gas to the Liquid Phase). *Handbook of Materials Modeling: Methods: Theory and Modeling* **2018**, 1–25.

(145) Partridge, H.; Schwenke, D. W. The Determination of an Accurate Isotope Dependent Potential Energy Surface for Water From Extensive Ab Initio Calculations and Experimental Data. *J. Chem. Phys.* **1997**, *106*, 4618–4639.

(146) Tang, K.; Toennies, J. P. An Improved Simple Model for the Van Der Waals Potential Based on Universal Damping Functions for the Dispersion Coefficients. *J. Chem. Phys.* **1984**, *80*, 3726–3741.

(147) Hastie, T.; Tibshirani, R.; Friedman, J. H.; Friedman, J. H. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Springer, 2009; Vol. 2.

(148) MBX: A Many-Body Energy and Force Calculator for Data-Driven Many-Body Simulations. http://paesanigroup.ucsd.edu/software/mbx.html, Accessed: 2019-07-05.

(149) GitHub, MB-Fit: Software Infrastructure for Data-Driven Many-Body Potential Energy Functions. https://github.com/paesanilab/MB-Fit.

(150) Mantz, Y. A.; Branduardi, D.; Bussi, G.; Parrinello, M. Ensemble of Transition State Structures for the Cis-Trans Isomerization of N-Methylacetamide. *J. Phys. Chem. B* **2009**, *113*, 12521–12529.

(151) Laio, A.; Parrinello, M. Escaping Free-Energy Minima. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 12562–12566.

(152) Tribello, G. A.; Bonomi, M.; Branduardi, D.; Camilloni, C.; Bussi, G. PLUMED 2: New Feathers for an Old Bird. *Comput. Phys. Commun.* **2014**, *185*, 604–613.

(153) Thompson, A. P.; Aktulga, H. M.; Berger, R.; Bolintineanu, D. S.; Brown, W. M.; Crozier, P. S.; in 't Veld, P. J.; Kohlmeyer, A.; Moore, S. G.; Nguyen, T. D.; Shan, R.; Stevens, M. J.; Tranchida, J.; Trott, C.; Plimpton, S. J. LAMMPS – A Flexible Simulation Tool for Particle-Based Materials Modeling at the Atomic, Meso, and Continuum Scales. *Comput. Phys. Commun.* **2022**, *271*, 108171.

(154) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and Testing of a General AMBER Force Field. *J. Comp. Chem.* **2004**, *25*, 1157–1174.

(155) Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A. Automatic Atom Type and Bond Type Perception in Molecular Mechanical Calculations. *J. Mol. Graph. Model.* **2006**, *25*, 247–260.

(156) Abascal, J. L.; Vega, C. A General Purpose Model for the Condensed Phases of Water: TIP4P/2005. *J. Chem. Phys.* **2005**, *123*, 234505.

(157) Delhommelle, J.; Millié, P. Inadequacy of the Lorentz-Berthelot Combining Rules for Accurate Predictions of Equilibrium Properties by Molecular Simulation. *Mol. Phys.* **2001**, *99*, 619–625.

(158) Li, J.; Zhou, J.; Xiong, Y.; Chen, X.; Chakrabarti, C. An Adjustable Farthest Point Sampling Method for Approximately-Sorted Point Cloud Data. 2022 IEEE Workshop on Signal Processing Systems (SiPS). 2022; pp 1–6.

(159) Werner, H.-J.; Knowles, P. J.; Knizia, G.; Manby, F. R.; Schütz, M. Molpro: a General-Purpose Quantum Chemistry Program Package. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2012**, *2*, 242–253.

(160) Werner, H.-J.; Knowles, P. J.; Manby, F. R.; Black, J. A.; Doll, K.; Heßelmann, A.; Kats, D.; Köhn, A.; Korona, T.; Kreplin, D. A., et al. The Molpro Quantum Chemistry Package. *J. Chem. Phys.* **2020**, *152*, 144107.

(161) Whitten, J. L. Coulombic Potential Energy Integrals and Approximations. *J. Chem. Phys.* **1973**, *58*, 4496–4501.

(162) Dunlap, B. I.; Connolly, J.; Sabin, J. On Some Approximations in Applications of Xα Theory. *J. Chem. Phys.* **1979**, *71*, 3396–3402.

(163) Werner, H.-J.; Manby, F. R.; Knowles, P. J. Fast Linear Scaling Second-Order Møller-Plesset Perturbation Theory (MP2) Using Local and Density Fitting Approximations. *J. Chem. Phys.* **2003**, *118*, 8149–8160.

(164) Møller, C.; Plesset, M. S. Note on an Approximation Treatment for Many-Electron Systems. *Phys. Rev.* **1934**, *46*, 618.

(165) Boys, S. F.; Bernardi, F. The Calculation of Small Molecular Interactions by the Differences of Separate Total Energies. Some Procedures with Reduced Errors. *Mol. Phys.* **1970**, *19*, 553–566.

(166) Tuckerman, M. *Statistical Mechanics: Theory and Molecular Simulation*; Oxford university press, 2010.

(167) Case, D. A.; Duke, R. E.; Walker, R. C.; Skrynnikov, N. R.; Cheatham III, T. E.; Mikhailovskii, O.; Simmerling, C.; Xue, Y.; Roitberg, A.; Izmailov, S. A., et al. AMBER 22 Reference Manual. **2022**,

(168) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. FF14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from FF99SB. *J. Chem. Theory Comput.* **2015**, *11*, 3696–3713.

(169) Zhuang, W.; Hayashi, T.; Mukamel, S. Coherent Multidimensional Vibrational Spectroscopy of Biomolecules: Concepts, Simulations, and Challenges. *Angew. Chem. Int. Ed.* **2009**, *48*, 3750–3781.

(170) Woutersen, S.; Hamm, P. Nonlinear Two-Dimensional Vibrational Spectroscopy of Peptides. *J. Phys. Condens. Matter* **2002**, *14*, R1035.

(171) Di Gioacchino, M.; Bruni, F.; Ricci, M. A. N-methylacetamide Aqueous Solutions: A Neutron Diffraction Study. *J. Phys. Chem. B* **2019**, *123*, 1808–1814.

(172) Gao, J.; Freindorf, M. Hybrid Ab Initio QM/MM Simulation of N-Methylacetamide in Aqueous Solution. *J. Phys. Chem. A* **1997**, *101*, 3182–3188.

(173) Dannenberg, J. Enthalpies of Hydration of N-Methylacetamide by One, Two, and Three Waters and the Effect upon the CO Stretching Frequency. An Ab Initio DFT Study. *J. Phys. Chem. A* **2006**, *110*, 5798–5802.

For Table of Contents Only