UNIVERSITY OF CALIFORNIA SAN DIEGO

Towards a Unified Description of Ion Hydration: Analysis and Development of Data-Driven Many-Body Models of Halide Ions

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Chemistry

by

Alessandro Caruso

Committee in charge:

Professor Francesco Paesani, Chair
Professor Massimiliano Di Ventra
Professor John Weare
Professor Jerry Yang
Professor Joel Yuen-Zhou

2023

The Dissertation of Alessandro Caruso is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2023

DEDICATION

I dedicate the culmination of my efforts to my family and friends,
whose unwavering support has given me the strength to persevere.

EPIGRAPH

The mathematical problems involved are very difficult.

Lars Onsager and Nicholas N. T. Samaras

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I am deeply thankful of Professor Francesco Paesani for his continuous support. Through his invaluable guidance, I have been able to grow as a scientist and as a person. His dedication to fostering a stimulating research environment and his relentless pursuit of knowledge have been inspiring. His insights, expertise, and mentorship have been critical. I am sincerely grateful for his patience and understanding, and for his constant encouragement during the challenging moments in this journey. His faith in my abilities, even when I doubted myself, helped me stay focused and persevere.

I would like to thank all my friends and colleagues in the Paesani Research Group. I am thankful to Etienne Palos, Dr. Saswata Dasgupta, and past members Dr. Eleftherios Lambros and Dr. Sigbjørn L. Bore, with which I could always share a laugh and discuss science.

Chapter 1, in full, is a reprint of the material as it appears in "A "short blanket" dilemma for a state-of-the-art neural network potential for water: Reproducing experimental properties or the physics of the underlying many-body interactions?" Y. Zhai, A. Caruso, S. L. Bore, Z. Luo, and F. Paesani. In: *J. Chem. Phys.* 158.8 (2023), p. 084111. The dissertation author is the co-primary investigator and author of this paper.

Chapter 2, in full, is a reprint of the material as it appears in "Active learning of many-body configuration space: Application to the Cs^+–water MB-nrg potential energy function as a case study" Y. Zhai, A. Caruso, S. Gao, and F. Paesani. In: *J. Chem. Phys.* 152.14 (2020), p. 144103. The dissertation author is the co-primary investigator and author of this paper.

Chapter 3, in full, is a reprint of the material as it appears in "Data-driven many-body models enable a quantitative description of chloride hydration from clusters to bulk" A. Caruso, and F. Paesani. In: *J. Chem. Phys.* 155.6 (2021), p. 064502. The dissertation author is the primary investigator and author of this paper.

Chapter 4, in full, is a reprint of the material as it appears in "Accurate Modeling

of Bromide and Iodide Hydration with Data-Driven Many-Body Potentials" A. Caruso, X. Zhu, J. Fulton, F. Paesani. In: *J. Phys. Chem. B* 126.41 (2022), pp. 8266–8278. The dissertation author is the primary investigator and author of this paper.

Chapter 5, in full is currently being prepared for submission for publication of the material. E. Palos, A. Caruso, and F. Paesani. The dissertation author is the co-primary investigator and author of this material.

VITA

2016	Bachelor of Science, Università degli Studi di Roma "La Sapienza"
2018	Master of Science, Università degli Studi di Roma "La Sapienza"
2018-23	Graduate Teaching Assistant, University of California San Diego
2018-23	Graduate Student Researcher, University of California San Diego
2023	Doctor of Philosophy, University of California San Diego

PUBLICATIONS

Y. Zhai[†], **A. Caruso**[†], S. L. Bore[†], Z. Luo, and F. Paesani "A "short blanket" dilemma for a state-of-the-art neural network potential for water: Reproducing experimental properties or the physics of the underlying many-body interactions?". In: *J. Chem. Phys.* 158.8 (**2023**), p. 084111.

A. Caruso, X. Zhu, J. Fulton, F. Paesani "Accurate Modeling of Bromide and Iodide Hydration with Data-Driven Many-Body Potentials". In: *J. Phys. Chem. B* 126.41 (**2022**), pp. 8266–8278.

T. E. Gartner III, K. M. Hunter, E. Lambros, **A. Caruso**, M. Riera, G. R. Medders, A. Z. Panagiotopoulos, P. G. Debenedetti, and F. Paesani "Anomalies and Local Structure of Liquid Water from Boiling to the Supercooled Regime as Predicted by the Many-Body MB-pol Model". In: *J. Phys. Chem. Lett.* 13.16 (**2022**), pp. 3652–3658.

A. Caruso, and F. Paesani "Data-driven many-body models enable a quantitative description of chloride hydration from clusters to bulk". In: *J. Chem. Phys.* 155.6 (**2021**), p. 064502.

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". In: *J. Chem. Phys.* 152.14 (**2020**), p. 144103.

V. Migliorati, **A. Caruso**, and P. D'Angelo "Unraveling the hydration properties of the Ba^{2+} aqua ion: the interplay of quantum mechanics, molecular dynamics, and EXAFS spectroscopy". In: *Inorg. Chem.* 58.21 (**2019**), pp. 14551–14559.

[†]These authors contributed equally.

FIELDS OF STUDY

Major Field: Chemistry (Theoretical and Computational)

Studies in Theoretical and Computational Chemistry
Professor Francesco Paesani

ABSTRACT OF THE DISSERTATION

Towards a Unified Description of Ion Hydration: Analysis and Development of Data-Driven Many-Body Models of Halide Ions

by

Alessandro Caruso

Doctor of Philosophy in Chemistry

University of California San Diego, 2023

Professor Francesco Paesani, Chair

Ion hydration is a central topic of discussion in the scientific community, given the important role that solvated ions play in fields as biochemistry, electrochemistry, and environmental chemistry. Despite the large amount of experimental and theoretical studies that involve ions in aqueous solutions, a unified molecular description is still missing: existing models fail to accurately capture the intricate ion-water interactions, leading to inconsistencies between experimental and theoretical results. The surface propensity of halide ions at the air/water interface has been debated for over a century, and the recent introduction of new architectures for the description of molecular interactions

allows for the development of accurate and efficient models to describe their behavior. Advanced molecular modelling frameworks based on deep neural networks (DNNs) and many-body expansion (MBE) of the energy are explored and their limitations analysed. Data-driven MBE-based models as MB-nrg are able to provide chemical accuracy at great computational efficiency when compared to classical force fields (FFs), DFT-based *ab initio* molecular dynamics (AIMD), and modern DNN potentials. After developing an efficient active learning (AL) framework for the generation of comprehensive training sets, we proceed with the development of MB-nrg models for the description of the chloride, bromide, and iodide ions; these provide great accuracy in the description of both gas-phase clusters and bulk systems, by closely reproducing coupled cluster (CC) interaction energies and the experimental x-ray absorption spectra. Lastly, we address the errors associated with the use of semi-local exchange-correlation functionals in modeling hydrated ions within density functional theory (DFT): the recent introduction of density-corrected (DC) SCAN functionals provides a solution to the overdelocalization issue typically encountered in semi-local density functional approximations (DFAs), leading to considerable improvements in the energetics and structural features of hydrated ions. This dissertation presents a significant advancement in the understanding of ion hydration, showcasing novel methods for more precise theoretical predictions, and providing a solid foundation for future research in this challenging field.

Introduction

To gain an understanding of the fundamental processes occurring in aqueous systems, it is crucial to quantitatively characterize molecular driving forces and mechanisms that govern ion hydration. Such knowledge has far-reaching implications for various fields of science and engineering: ions often serve as intermediates in chemical reactions and catalytic processes,[1,2] and they play a central role in the stabilization of biomolecules[3–5] and the mediation of protein-protein interactions.[6,7] In atmospheric chemistry, hydrated ions they have been shown to take part in the growth process of cloud condensation nuclei,[8–11] and in electrochemistry, ionic solutions are a fundamental component of electrolytic cells, capacitors, and batteries.[12] The ubiquitous presence of hydrated ions and their central role in mediating fundamental processes has led to a strong motivation for characterizing them at a molecular level. Despite the significant experimental and theoretical efforts, understanding ion hydration at a microscopic level remains a challenge due to conflicting experimental results and the lack of accurate theoretical models.

In the gas phase at ambient conditions, Coulomb interactions prevent ions to exist as isolated species; however, when they are dissolved in water, they can be individually stabilized by the mediation of water molecules. The energetic contributions associated with ion-water interactions counteract the Coulomb attraction between oppositely charged ions, while entropic contributions arise from the reorganization of the hydrogen bond (H-bond) network of water due to the cavity the ion generates. In chemistry, ions are typically classified as "structure makers" or "structure breakers", based on their ability to facilitate or hinder the formation of H-bond networks around them.[13] Structure makers

are characterized by a small radius and high charge, which strengthens the surrounding H-bonds due to their strong interactions. In contrast, structure breakers have a small, diffuse charge that fails to counteract the disruptive effect of their size on the network of water molecules. However, this well-established idea has been recently challenged by various spectroscopic measurements, revealing the complexity of ion interactions.[14–18]

A significant example of the inadequacy of current models in representing the intricate interactions of hydrated ions is their distribution at the air/water interface. In 1910, Heydweiller discovered that the surface tension of salt solutions at varying concentrations was greater than that of pure water.[19] The increase in surface tension was found to be largely independent of the cation but varied significantly when comparing different anions, following an inverse Hofmeister series. Wagner initially proposed a theory that was later developed by Onsager and Samaras, suggesting that hydrated ions have different interfacial distributions due to interactions with their image charge, leading to surface depletion and different surface tensions.[20,21] This theory was later contradicted by measuring the electrostatic potential difference for solutions of halide salts; in fact, the anionic concentration (with the exception of fluoride) at the air/water interface was higher than expected.[22] Since then, subsequent analyses have been conducted using various experimental techniques, but these have produced conflicting results.[23–25]

In the most striking example, vibrational sum-frequency generation (vSFG) spectroscopy has been used in the study of the distribution of anions at the air/water interface, by exploiting its high sensitivity to changes in the water H-bond network. A first study using this technique has shown a considerable difference in the vSFG spectra of NaBr and NaI aqueous solutions with respect to pure water and to analogous solutions containing NaF and NaCl, suggesting an increased concentration of heavier halides at the interface.[23] A study that used a similar experimental setup, however, found a diminished concentration of the same ions at the interface, in clear opposition to the findings of the first work.[24]

From a theoretical perspective, molecular dynamics (MD) simulations have been

carried out for halide-water clusters using nonpolarizable[26,27] and polarizable[28–32] FFs. While nonpolarizable FFs depict the halide ion (chloride) away from the interface, polarizable FFs showed an increased concentration of all halides (excluding fluoride) preferentially at the interface, highlighting the importance of many-body effects for the description of hydrated halide ions. MD simulations of salt solutions and single-ion potentials of mean force (PMFs) calculated using polarizable FFs demonstrated that the concentration of halide ions at the air/water interface increases with ion size (from chloride to iodide, with fluoride being repelled from the interface).[33–37] An extended dielectric continuum (DC) theory predicted lower surface propensity of halide ions compared to previous theoretical findings.[38,39] Evidently, despite considerable efforts towards accurately characterizing ionic properties in aqueous systems, both experimental and theoretical results remain inconclusive. A unified ion hydration theory requires a quantitative description of the entirety of interactions at the molecular level. In the past decade, many-body potential energy functions (PEFs) have shown promise in describing aqueous systems, from small clusters to bulk solutions and interfaces.[40–44] Further research and development of these techniques may provide the necessary insight to resolve the complex behavior of hydrated ions in various environments.

Theoretical framework

Generally speaking, it is extremely challenging to model the behavior of chemical systems. Not only do we have to deal with the complex requirements of a quantum mechanical explanation of subatomic particles, but we also need to address the complexities of many-body systems consisting of particles that interact with each other.

Modern quantum chemistry methods are capable of providing a reliable description of the energetics of molecular systems. While explicitly correlated *ab initio* methods as CC[45,46] or configuration interaction (CI)[47] currently recover the most accurate energies and

structural features, their use is limited to systems with a handful of atoms, due to the large set of nonlinear equations to be solved and intermediate calculations required. Methods as second-order Møller-Plesset perturbation theory (MP2)[48,49] and density functional theory DFT[50–52] are widely used alternatives that provide balance between accuracy and computational efficiency. While these methods allow the treatment of systems comprising a few hundred atoms, the accuracy of the results is highly dependent on the choice of basis set and exchange-correlation functional, and errors in predicting energies can be in the order of 10 kcal/mol.[53] The computational cost of *ab initio* approaches makes it difficult or impossible to achieve chemical accuracy in the study of large molecular systems and long time-scale events.

From a theoretical standpoint, the total energy of a molecular system can be exactly decomposed into its n-body terms through the MBE of the energy as[54]

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

where the first term in Eq. (1), $v^{1\mathrm{B}}(i) = E(i) - E_{\mathrm{eq}}(i)$, is the deformation energy required to recover each monomer in the expansion from the equilibrium geometry. The remaining terms in the expansion are recursively defined as[54]

$$v^{\mathrm{nB}} = E_N - \sum_{i=1}^{N} v^{1\mathrm{B}}(i) - \sum_{i>j}^{N} v^{2\mathrm{B}}(i,j) - \sum_{i>j>k}^{N} v^{3\mathrm{B}}(i,j,k) - \ldots - v^{(\mathrm{n}-1)\mathrm{B}}. \tag{2}$$

Traditional FFs truncate Eq. (1) at the two-body (2B) term and account for many-body contributions implementing effective pairwise potentials.[55–66] The use of simple point-charge models, together with the efficient use of pairwise interactions, allowed molecular dynamics MD to establish itself as the workhorse in computational chemistry. In the modelling of aqueous systems, it was soon realized that pairwise potentials do not provide a faithful

representation of both gas and condensed phase properties.[67] Technical advancement in computing has enabled the development of more rigorous and reliable interaction potentials: state-of-the-art PEFs take advantage of efficient regression algorithms to shape large parametric models onto the multidimensional energy landscape of the underlying system. A first approach takes advantage of *ab initio* calculations to include short-range quantum mechanical effects by fitting large n-body polynomials within the MBE of the energy. More recently, the advancement in machine learning (ML) and the design of performant *ad hoc* computational architectures has allowed deep neural networks to emerge as a new approach to the development of modern PEFs. Despite their extensive use for modelling molecular systems,[68–92] current-state DNN potentials have recently been shown to provide limited accuracy and transferability across phases.[93]

Many-Body Models and the MB-nrg Framework

To model water-water interaction, several many-body models have been proposed in the last two decades, as CC-pol,[94] WHBB,[95] HBB2-pol,[96] and MB-pol;[97–99] these PEFs provide accurate prediction of the structural and dynamical properties, and the thermodynamics of water, from gas-phase to bulk and interfaces.[43,44,94–110] The MB-nrg framework has been built upon the MB-pol water model to describe generic molecular systems and ionic species.[111–116] Since the expansion of Eq. (1) converges rapidly for localized and non-metallic systems, the MBE of the energy is generally written, within the MB-nrg framework, as

$$E_N = V^{1\mathrm{B}} + V^{2\mathrm{B}} + V^{3\mathrm{B}} + V^{\mathrm{NB}}_{\mathrm{pol}}. \tag{3}$$

The first term on the right side of Eq. (3) is trivially the sum of the distortion energy of each monomer in the system, $V^{1\mathrm{B}} = \sum v^{1\mathrm{B}}(i)$. The remaining lower terms in the expansion

are approximated as

$$
\begin{aligned}
V^{\mathrm{2B}} &= \sum_{i>j}^{N} \varepsilon_{\mathrm{sr}}^{\mathrm{2B}}(i,j) + V_{\mathrm{elec}}^{\mathrm{2B}} + V_{\mathrm{disp}}^{\mathrm{2B}} \\
V^{\mathrm{3B}} &= \sum_{i>j>k}^{N} \varepsilon_{\mathrm{sr}}^{\mathrm{3B}}(i,j,k),
\end{aligned}
\tag{4}
$$

where $V_{\mathrm{elec}}^{\mathrm{2B}}$ represents the permanent electrostatic energy, which includes charge-charge, charge-dipole, and dipole-dipole interactions, and $V_{\mathrm{disp}}^{\mathrm{2B}}$ is the classical two-body dispersion, which employs Tang-Toennies damping functions and contains dispersion coefficients calculated using the exchange-hole dipole model (XDM) developed by Becke and Johnson.[107,108] $\varepsilon_{\mathrm{sr}}^{\mathrm{2B}}(i,j)$ and $\varepsilon_{\mathrm{sr}}^{\mathrm{3B}}(i,j,k)$ are correction terms, usually implemented in the form of permutationally invariant polynomials (PIPs), to model electronic quantum mechanical effects at short-range. Lastly, the $V_{\mathrm{pol}}^{\mathrm{NB}}$ represents the classical many-body polarization due to the remaining n-body terms in the expansion. Since the short-range correction terms become negligible as n increases, it usually suffices to include corrections up to the three-body term. The general MB-nrg framework allows to exploit the accuracy of *ab initio* electronic structure methods in the training of the PIPs by limiting expensive calculations to small gas-phase clusters: labels are often calculated at the coupled cluster level, with single, double, and perturbative triple excitations, i.e. CCSD(T), the "gold standard" for chemical accuracy.[117]

Summary

In Chapter 1, we present an in-depth analysis of the new generation of DNN-based PEFs.[93] As a proof of concept, water-water interaction is modeled through the state-of-the-art and award-winning DeePMD framework,[76,118,119] using MB-pol reference energies.[97,98,106] Usually, DNN potentials require training sets of condensed-phase configurations to allow for long-range effects to be taken into account; due to the large

number of molecules considered, DFT is usually employed, rather than CC methods, to retrieve reference energies. In principle, training on MB-pol labels allows to circumvent this limitation and to exploit the computational efficiency of DeePMD, extending MD simulations to larger systems or kinetically inaccessible states. The results presented show that, while training using bulk configurations provides accurate values of bulk properties, this accuracy is not transferred when considering vapor-liquid equilibrium (VLE) configurations; moreover, the predicted energy appears to be the result of error compensation in the single many-body energies of the system. Correcting for this behavior by extending the training set to comprise VLE or gas-phase cluster configurations inevitably introduces errors in the prediction of bulk configuration properties, suggesting that current DNN architectures are not suitable for a general and unified representation of molecular systems. The generation of representative training set plays a central role in the development of accurate potentials for MD simulations. As the number of particles of the system N increases, the $3N-3$ independent degrees of freedom do not allow to perform grid search methods for selecting new configurations and retrieve their labels. Scans along specific axes and normal mode sampling can mitigate the computational expense but introduce bias in the selection process. Chapter 2 goes over this delicate process, and introduces an AL framework for the selection of configurations relevant to the training of many-body models.[120] The new methodology shows great efficiency, selecting configurations that introduce the largest error during the training procedure. This work is lies the foundations for the development of halide-water PEFs. Chapter 3 introduces a general methodology for the development of halide-water MB-nrg potentials, accounting for MB correction terms up to the 3-body.[113] After generating chloride-water MB-nrg PEFs, we first presents MB analyses of gas-phase clusters, then the structural determination and EXAFS predictions of chloride-water solvation shells in bulk phase are compared with experiments. Our MB-nrg PEFs show a significant improve in accuracy with respect to other interaction potentials extensively employed in the field. Analogously, bromide- and iodide-water potentials are

introduced in Chapter 4.[114] Chapter 5 dives into the analysis of density-driven errors in the modelling of hydrated ions using semi-local exchange-correlation functionals to determine training labels through DFT. Building upon previous research,[121] DFT reference energies calculated with the strongly-constrained SCAN functional are used to train MB-DFT models of sodium- and chloride-water PEFs; these are compared with analogous PEFs built upon DC-SCAN reference energies that make use of the Hartree-Fock density to correct the overdelocalization typically experienced by semi-local DFAs. DC-DFT shows great improvements in the energetics and in the determination of the structural features of the hydrated ions. Lastly, Chapter 6 summarizes our findings and conclusions.

Chapter 1

A "short blanket" dilemma for a state-of-the-art neural network potential for water: Reproducing experimental properties or the underlying many-body physics?

1.1 Introduction

Molecular mechanics (MM) force fields (FFs)[122,123] have been the workhorse in computational chemistry since the early days of Monte Carlo (MC)[124] and molecular dynamics (MD) simulations.[125] Continued progress in hardware technologies,[126] accompanied by the development of more realistic representations of electrostatic interactions, has enabled not only molecular simulations of progressively larger systems but also the use of more rigorous polarizable FFs[127–131] that go beyond the pairwise additive approximation adopted by conventional fixed-charge FFs.[132–135]

At the same time, the development of efficient algorithms for correlated electronic structure methods, such as coupled cluster theory,[136–138] has enabled routine calculations of interaction energies for molecular clusters with chemical accuracy.[139–141] This has led to the emergence of a new class of analytical potentials that quantitatively reproduce each individual term of the many-body expansion (MBE) of the energy[142] calculated using corre-

lated electronic structure methods. When applied to aqueous systems[94–98,100–103,105–109] and molecular fluids,[143–145] these many-body (MB) potentials exhibit unprecedented accuracy, enabling predictive simulations from the gas to the condensed phases.[110] Concurrently, machine learning (ML) approaches have gained popularity in computational molecular sciences mainly due to the rapid evolution of GPU and TPU architectures.[146] In particular, potentials represented by deep neural networks (DNNs) derived from electronic structure data are routinely used to model various molecular systems, from clusters to liquids and materials.[68–92]

State-of-the-art MB and DNN approaches use regression algorithms to construct data-driven representations of the multidimensional energy landscape of the system of interest. This process involves generating representative training sets of reference data calculated at an appropriate level of theory. While MB potentials require tailored parametric functions for each term of the MBE, DNN potentials are usually trained on the total energy and forces of the entire system. By applying embedding schemes to construct low-dimensional descriptors of molecular environments, DNN potentials can compute the gradients required for the propagation of the equations of motion in MD simulations more efficiently than MB potentials.[118] On the other hand, since MB potentials only use information about small clusters, the corresponding training data can be calculated at a higher level of theory than DNN potentials. As a matter of fact, MB potentials are usually trained on reference energies calculated at the coupled cluster level of theory, including single, double, and perturbative triple excitations, i.e., CCSD(T), which is currently referred to as the "gold standard" for chemical accuracy.[117] Furthermore, by construction, the functional form of MB potentials allows for accurately representing all physical contributions to the interaction energies, including both short- and long-range many-body effects.[41,42,147]

To account for long-range interactions, DNN potentials are often trained on condensed-phase configurations, which allows for modeling long-range effects either implic-

itly, by effectively encoding long-range contributions into short-range representations, or explicitly, by adding effective electrostatic terms.[76,86,92,118,119,148] This implies that, due to the large number of molecules required to model condensed-phase systems, a lower level of theory than CCSD(T), usually density functional theory (DFT),[52] has to be used to retrieve the reference energies. In this context, it has recently been shown that the interplay between functional-driven and density-driven errors may impact the overall accuracy of DFT models and their transferability from gas-phase to condensed-phase systems.[121,149–153] By construction, these limitations also affect the ability of DNN potentials derived from DFT reference data to "extrapolate" to thermodynamic state points different from those used in the training process.[154]

In this work, we investigate the possibility of integrating the best features of MB potentials (i.e., accuracy and transferability) and DNN potentials (i.e., speed and ease to use) into a computational framework that can enable large-scale MD simulations with chemical accuracy. To this end, we focus on the molecular modeling of water as a prototypical system that has posed several challenges since the early days of MC and MD simulations[155,156] due to its rich phase behavior characterized by several anomalies.[157] As a representative state-of-the-art MB potential, we selected MB-pol[97,98,106] due to its demonstrated ability to correctly predict the properties of water across the entire phase diagram,[158] including gas-phase clusters,[159–161] liquid water,[162] the vapor-liquid interface,[163–165] and ice.[166–169] MB-pol has also recently been used to predict structural and thermodynamic properties of supercooled water down to 200 K at 1 atm, which were found to be in excellent agreement with experimental data that are available above 225 K.[43] However, due to the associated computational cost, the MD simulations with MB-pol reported in Ref. 43 were limited in terms of both system's size (up to 512 water molecules) and sampling time (up to ∼130 ns). The prospect of developing a fast DNN potential trained on MB-pol simulation data, which retains the same accuracy of MB-pol across the entire phase diagram, is thus particularly appealing. This will enable large-scale

simulations of water as a function of temperature and pressure, which will provide further insights into water's anomalous behavior and allow for full exploration of the so-called water's "no man's land" that has been proven difficult to probe experimentally.[170–175] To this end, we selected DeePMD[76,118,119] as a representative, state-of-the-art framework for developing a DNN potential of water trained on the MB-pol simulations of Ref. 43. DeePMD-based DNN potentials have already been used in MD simulations of various molecular systems, including water,[176–179] ionic liquids,[180] and metals,[181–183] and enabled MD simulations with up to 10 billion atoms.[184]

The article is organized as follows: In Section 1.2, we summarize the main features of the MB-pol potential (Section 1.2.1) and the DeePMD framework (Section 1.2.2). In Section 1.3.1, we first assess the ability of the DeePMD-based DNN potential to reproduce thermodynamic and structural properties of liquid water calculated with MB-pol from the boiling point down to deeply supercooled temperatures. We then use the DNN potential to characterize several vapor-liquid equilibrium properties as well as many-body dependent properties of gas-phase clusters. In Section 1.3.2, we introduce two other DeePMD-based potentials, DNN(VLE10) and DNN(VLE20), that are trained on an expanded training set that adds vapor-liquid configurations to the training set used to develop the DNN potential. The performance of both DNN(VLE10) and DNN(VLE20) potentials is assessed on the same structural and many-body-dependent properties used to assess the performance of the DNN potential. In Section 1.3.3, we introduce three DeePMD-based potentials, DNN(MB4), DNN(MB10), and DNN(MB20), that are trained to incorporate low-order many-body interactions, and assess their performance on the same structural and many-body dependent properties used in the assessment of the DNN potential. Lastly, in Section 1.4, we summarize our work and discuss possible future synergies between MB and DNN potentials.

1.2 Methods

1.2.1 MB-pol

Since the MB-pol potential of water has already been described in detail in the literature, we only overview here its salient features.[97,98,106] MB-pol was derived from the MBE that expresses the energy, E_N, of a system containing N (atomic or molecular) monomers as the sum of individual n-body energy contributions,

$$E_N(1,\ldots,N) = \sum_{i=1}^{N} \epsilon^{1\mathrm{B}}(i) + \sum_{i=1}^{N}\sum_{j>i}^{N} \epsilon^{2\mathrm{B}}(i,j) + \sum_{i=1}^{N}\sum_{j>i}^{N}\sum_{k>j>i}^{N} \epsilon^{3\mathrm{B}}(i,j,k) + \cdots + \epsilon^{\mathrm{NB}}(1,\ldots,N) \tag{1.1}$$

Here, $\epsilon^{1\mathrm{B}}$ represents the distortion energy of an isolated monomer, such that $\epsilon^{1\mathrm{B}}(i) = E(i) - E_{\mathrm{eq}}(i)$ where $E_{\mathrm{eq}}(i)$ is the energy of the i-th monomer in its equilibrium geometry. The n-body energies, ϵ^{nB}, are defined recursively for $1 < n \leq N$ by the expression

$$\begin{aligned}\epsilon^{\mathrm{nB}} = E_{\mathrm{n}}(1,\ldots,n) - \sum_{i=1}^{N} \epsilon^{1\mathrm{B}}(i) - \sum_{i=1}^{N}\sum_{i<j}^{N} \epsilon^{2\mathrm{B}}(i,j) - \sum_{i=1}^{N}\sum_{i<j}^{N}\sum_{i<j<k}^{N} \epsilon^{3\mathrm{B}}(i,j,k) - \cdots \\ \cdots - \sum_{i<j<k<\ldots}^{N} \epsilon^{(\mathrm{n}-1)\mathrm{B}}(i,j,k,\ldots,n-1)\end{aligned} \tag{1.2}$$

MB-pol approximates Eq. 1.2 as:

$$E_N(r_1,..,r_N) = \sum_{i=1}^{N} \epsilon^{1\mathrm{B}}(i) + \sum_{i>j}^{N} \epsilon^{2\mathrm{B}}(i,j) + \sum_{i>j>k}^{N} \epsilon^{3\mathrm{B}}(i,j,k) + E_{\mathrm{POL}} \tag{1.3}$$

The one-body term ($\epsilon^{1\mathrm{B}}$) is represented by the potential developed by Partridge and Schwenke.[185] The two-body term ($\epsilon^{2\mathrm{B}}$) describes four distinct contributions: permanent electrostatics, dispersion, 2B polarization, and 2B short-range interactions. The three-body term ($\epsilon^{3\mathrm{B}}$) describes two distinct contributions: 3B polarization and 3B short-range interactions. 2B and 3B short-range interactions are represented by 2B and 3B

permutationally invariant polynomial (PIPs)[186] that were fitted in order for ϵ_{2B} and ϵ_{3B} to reproduce 2B and 3B energies calculated at the CCSD(T) level of theory in the complete basis set (CBS) limit.[97,98] 2B and 3B polarization contributions are implicitly included in E_{POL} in Eq. 1.3 which represents classical many-body interactions at all orders through a polarization term. Further details of the MB-pol potential can be found in the original references.[97,98,106]

1.2.2 DeePMD

The DeePMD framework reads atomic positions and associated atom types as input features.[118] Neighbor information for each atom i is extracted from the input feature using a predefined cutoff radius (r_c) and stored as the coordinate difference of each ij atom pair into $\mathcal{R}^i \in \mathbb{R}^{N_i \times 3}$, where N_i is the number of neighboring atoms. Each local feature is then mapped onto generalized coordinates $\tilde{\mathcal{R}}^i$ as outlined in Ref. 119. A local embedding matrix, $\mathcal{G}^i$, is applied to each local feature $\mathcal{R}^i$ in order to ensure rotation and permutation symmetry while preserving translation symmetry. The resulting encoded feature matrix $\mathcal{D}^i \in \mathbb{R}^{M_1 \times M_2}$ takes the form

$$\mathcal{D}^i = (\mathcal{G}^{i1})^T \tilde{\mathcal{R}}^i (\tilde{\mathcal{R}}^i)^T \mathcal{G}^{i2} \tag{1.4}$$

and is passed to a fully-connected feed-forward DNN that maps it onto an "atomic energy" E_i.[119] The total energy E is then calculated as the sum of all E_i, while the atomic forces F and virials Ξ are calculated from the derivative of the DNN with respect to the corresponding atomic positions. The DNN parameters are optimized by minimizing the loss function:

$$L(p_e, p_f, p_\xi) = \frac{p_e}{N}\Delta E^2 + \frac{p_f}{3N}\Delta F^2 + \frac{p_\xi}{9N}\Delta \Xi^2. \tag{1.5}$$

where p_e, p_f, p_ξ are weighting factors, and ΔE, ΔF, and $\Delta \Xi$ are the prediction errors for the reference energy, force, and virial values, respectively. The weighting factors p_e, p_f, p_ξ

are adjusted as the training progresses in order to improve the quality of the fit.

The DeePMD-based DNN potentials presented in this study were developed with the Deep Potential Smooth Edition (DeepPot-SE),[119] following the procedure reported in Ref. 187, using 25, 50, and 100 neurons for the hidden embedding layers in the DeepPot-SE, while the submatrix of the embedding matrix uses 16 neurons. The distance cutoff was set to 6 Å, with a smoothing region of 0.5 Å. Each DNN potential is represented by a fully connected deep neural network with three layers of 240 neurons each.

Following Ref. 188, the training set for the DNN potential was constructed in an iterative fashion. Briefly, the training set comprises energies and forces of molecular configurations extracted from the MB-pol simulations of liquid water between 198 K and 368 K reported in Ref. 43 as well as additional configurations extracted from simulations carried out with three successive iterations of the DNN potential. The final training set includes 94770 configurations, each containing 256 molecules. All MB-pol reference data were computed using the MBX software package.[189] Additional details about the training set are discussed in the Supplementary Material. To account for variations in training, validation, and testing errors, four distinct potentials (hereafter referred to as seed 1, seed 2, seed 3, and seed 4, respectively) were trained using different random seeds. Only one of the four DNN potentials (seed 2) was then used in the MD simulations of liquid water and the vapor-liquid interface. Similarly, four distinct DeePMD-based potentials were trained on the expanded training sets containing vapor-liquid and cluster configurations as described in Sections 1.3.2 and 1.3.3, respectively. Fig. 1.1 shows the root-mean-square error (RMSE) curves of training and validation sets for the energies and forces per atom during the fitting process of the (seed 2) DNN potential. Overall, well-behaved learning curves are obtained for both quantities, with final errors of 0.01 kcal/mol and 1 kcal/mol Å on the energy and force validation errors, respectively. Similar errors have been reported for other state-of-the-art machine-learned potentials.[190]

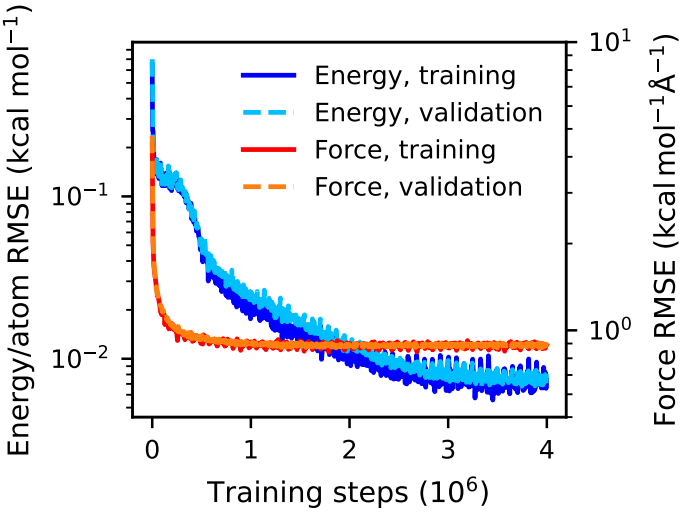

Figure 1.1. Variation of the DNN training and validation RMSEs per atom relative to the MB-pol values of the energy and force as a function of the number of training steps. For visual clarity, we show values averaged over 200 training steps.

1.2.3 Computational details

We performed two sets of MD simulations to determine the ability of the DNN potential to reproduce both bulk and interfacial properties of liquid water calculated with MB-pol. The first set of simulations was carried out for a box containing 256 water molecules in the isothermal-isobaric (NPT) ensemble at 1 atm and in the temperature range between 198 K to 368 K. The temperature was maintained using a global Nosé–Hoover thermostat chain of length 3 with a relaxation time of 0.05 ps, and the pressure was controlled by global Nosé–Hoover barostat with a relaxation time of 0.5 ps which was thermostatted by a Nosé–Hoover thermostat chain of length 3. At each temperature, the last frame of the MB-pol trajectories reported in Ref. 43 was used as the initial configuration for the NPT simulations with the DNN potential. The second set of simulations was carried out in the canonical (NVT) ensemble between 400 K and 575 K for a liquid slab of 512 water molecules in a box of dimensions 20 Å×20 Å×100 Å. The temperature was maintained using the same global Nosé–Hoover thermostat chain used in the NVT simulations. In

both NPT and NVT simulations, the velocity-Verlet algorithm was used to propagate the equations of motion with a time step of 0.5 fs according to Ref. 191. All simulations were carried out using the DeePMD-kit[187] plugin for LAMMPS.[192] A complete set of input files is available on GitHub (see Data Availability).

Besides comparing the DNN and MB-pol radial distribution functions (RDFs), we also analyzed the ability of the DNN potential to describe the three-dimensional hydrogen-bond network by calculating the tetrahedral order parameter, q_{tet},[193]

$$q_{\mathrm{tet}} = 1 - \frac{3}{8} \cdot \sum_{j=1}^{3} \sum_{k=j+1}^{4} \left(\cos(\psi_{jk}) + \frac{1}{3} \right) \tag{1.6}$$

Here, ψ_{jk} is the angle between the oxygen of the central water molecule and the oxygen atoms of two neighboring water molecules within a cutoff of 3.5 Å. When $q_{\mathrm{tet}} = 1$, the water molecules are in a perfect tetrahedral arrangement, while $q_{\mathrm{tet}} = 0$ represents the ideal gas limit.

In addition to the MD simulations for liquid water and the vapor-liquid interface, we also performed many-body decomposition analyses for two different sets of cluster structures. The first set consists of the first eight low-lying energy isomers of the water hexamer (Fig. 1.2), with geometries taken from Ref. 158. The hexamer occupies a special place in the description of many-body interactions in water because it is the smallest cluster with low-lying isomers that display three-dimensional hydrogen-bonded arrangements similar to those found in liquid water and ice. The second set of clusters contains dimers and trimers extracted from the training sets used to fit the MB-pol 2B and 3B energy terms, respectively.[97,98]

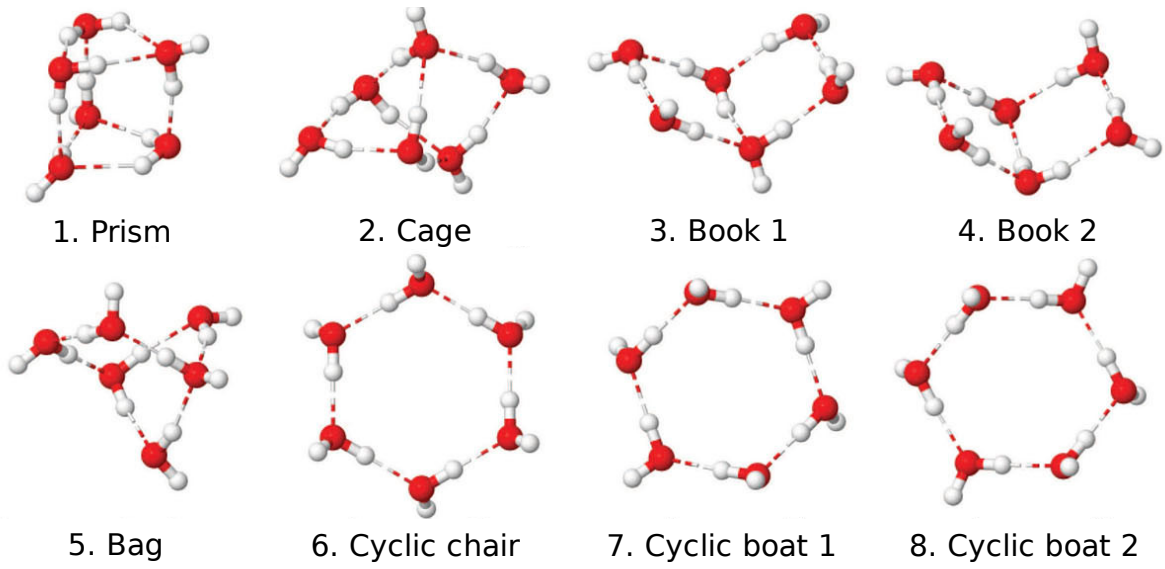

Figure 1.2. Structures of the first eight low-lying energy isomers of the water hexamer used in the analysis of interaction and many-body energies. The Cartesian coordinates of each isomer were taken from Ref. 158.

1.3 Results

1.3.1 DNN potential

As a first step in assessing the ability of the DNN potential to reproduce MB-pol, we analyze various properties of liquid water. In Fig. 1.3a, we show the temperature dependence of the liquid density from 198 K to 368 K. In general, the DNN potential reproduces the MB-pol results over the entire temperature range, predicting similar temperatures for the density maximum and minimum. A more quantitative analysis indicates that the DNN potential underestimates the MB-pol density by $\sim 1\%$ in the $220 - 290$ K range while it predicts a slightly denser liquid as the temperature approaches the boiling point. The DNN curve also displays a less negative slope for temperatures above $\sim$320 K, which suggests that it is relatively more difficult for the DNN potential to reproduce MB-pol as the liquid properties become more gas-like. To put the present comparison between the MB-pol and DNN potentials in context, we note that the density of

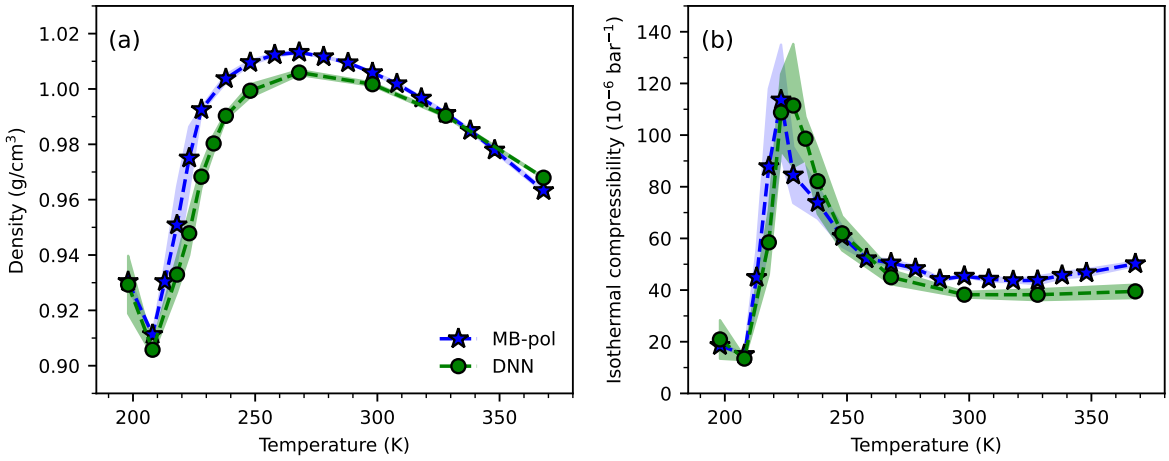

Figure 1.3. Temperature dependence of the density (a) and isothermal compressibility (b) calculated from the present NPT simulations carried out with the DNN potential at 1 atm (green) compared with the reference MB-pol values from Ref. 43 (blue). The associated shaded areas indicate 95% confidence intervals of the averages. Fig. S1 displays the time dependence of the densities calculated along the NPT trajectories carried out with the DNN potential at each temperature.

liquid water at 298 K predicted by an analogous DeePMD-based DNN potential trained on the SCAN functional was found to be ∼5% lower than the corresponding value calculated from the reference *ab initio* molecular dynamics (AIMD) simulations.[194]

The comparison between the DNN and MB-pol values for the isothermal compressibility as a function of temperature is shown in Fig. 1.3b. Similar to the density, the DNN values are in agreement with the MB-pol reference data, reproducing the steep increase of the isothermal compressibility below 250 K and predicting a maximum at ∼230 K, which is ∼7 K higher than the temperature predicted by MB-pol.[43] As in the case of the liquid density, Fig. 1.3b also indicates that the ability of the DNN potential to reproduce MB-pol somewhat deteriorates as the temperature approaches the boiling point. In particular, the DNN potential predicts a nearly constant value of the isothermal compressibility above 300 K, with no indication of a distinct minimum that is instead found in both experiment[172,195] and MB-pol simulations.[43,158]

A comparison between the structural properties of liquid water predicted by the

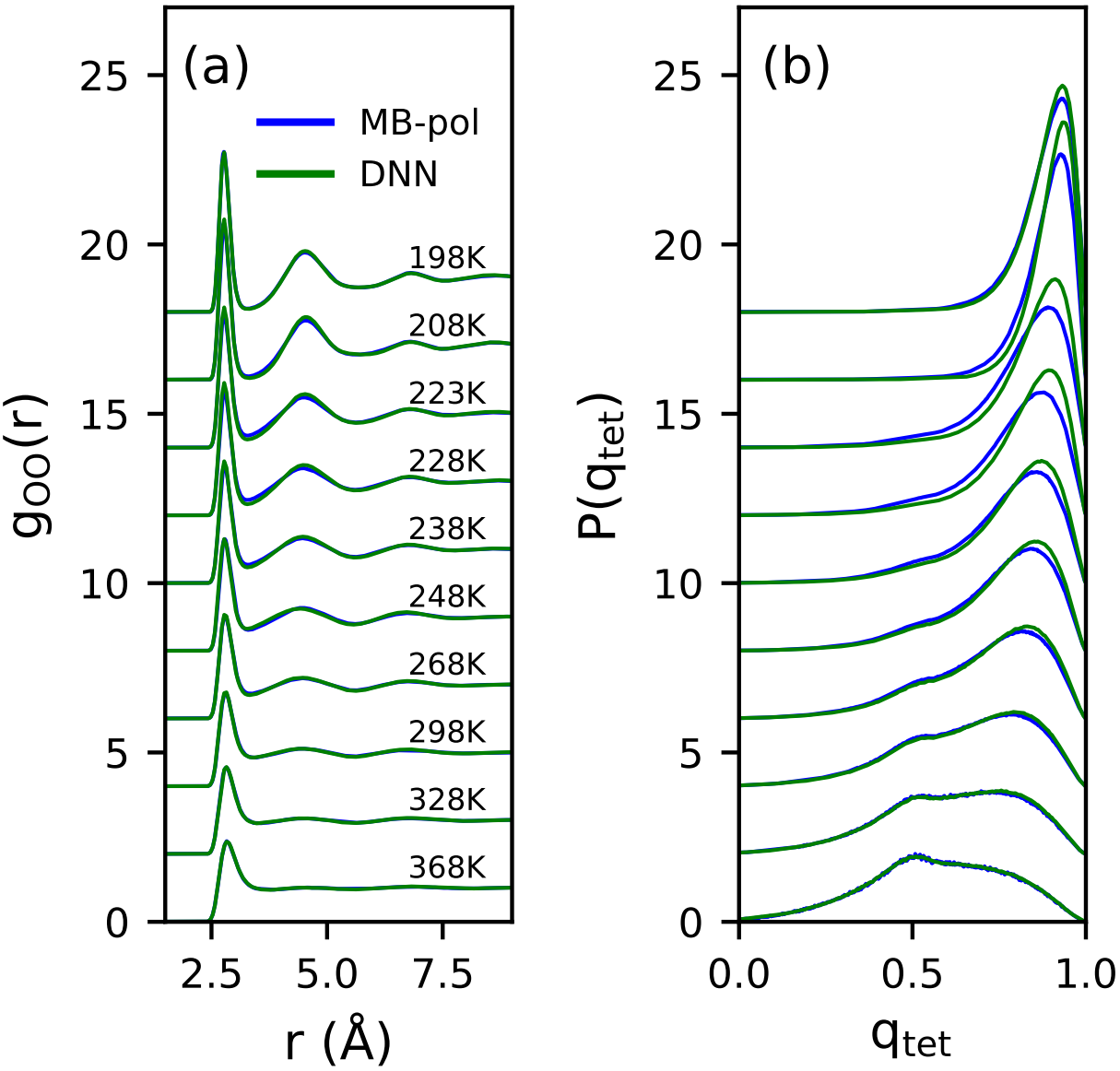

Figure 1.4. Oxygen-oxygen radial distribution functions (a) and tetrahedral order parameter distributions (b) calculated from the present NPT simulations carried out with the DNN potential at 1 atm (green) compared with the reference MB-pol values from Ref. 43 (blue).

DNN and MB-pol potentials between 198 K and 368 K at 1 atm is shown in Fig. 1.4. The oxygen-oxygen radial distribution functions (RDFs) calculated from the NPT simulations carried out with the two potentials (Fig. 1.4a) are nearly indistinguishable in the 238-368 K range. Small deviations are found at deeply supercooled temperatures, which become

more apparent when analyzing the corresponding distributions of the tetrahedral order parameter in Fig. 1.4b. These deviations appear to be consistent with the shift of the isothermal compressibility maximum to a slightly higher temperature predicted by the DNN potential (Fig. 1.3b).

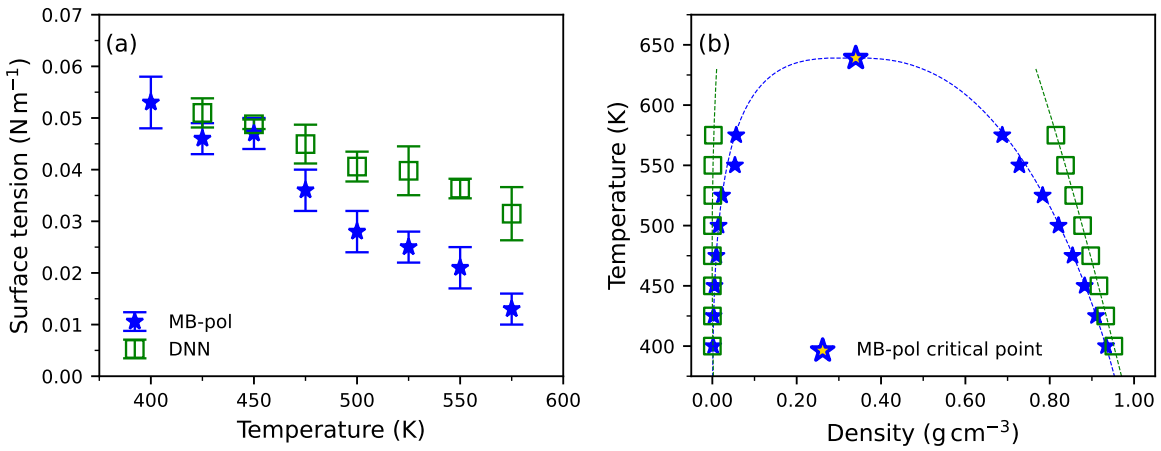

Figure 1.5. Surface tension (a) and vapor-liquid equilibrium densities (b) calculated from the present NVT simulations of a water slab carried out with the DNN potential (green) compared with the reference MB-pol values from Ref. 165 (blue).

Previous studies demonstrated that MB-pol correctly predicts structural, thermodynamic, and spectroscopic properties of the vapor-liquid interface, including the surface tension, vapor pressure, vapor and liquid densities,[158,165] as well as sum-frequency generation spectra.[163,164] To assess the ability of the DNN potential to reproduce properties that are not directly related to the MB-pol liquid configurations used during the training process, in Fig. 1.5, we analyze the surface tension and liquid-vapor equilibrium densities as a function of the temperature. These comparisons show that both surface tension and equilibrium densities predicted by the DNN potential deviate significantly from the corresponding MB-pol reference values as the temperature increases. Interestingly, while the liquid density predicted by the DNN potential decreases upon increasing temperature, in qualitative agreement with the expected physical behavior, the vapor density remains effectively constant over the entire temperature range. Following Ref. 165, we estimated

the critical temperature (T_c) and density (ρ_c) associated with DNN potential by fitting the vapor (ρ_v) and liquid (ρ_l) densities (Fig. 1.5) according to:

$$\frac{\rho_l + \rho_v}{2} = \rho_c + A(T_c - T) \tag{1.7}$$

$$\frac{\rho_l - \rho_v}{2} = \Delta\rho_0 \left(1 - T/T_c\right)^{\beta} \tag{1.8}$$

Here, A and $\Delta\rho_0$ are system-specific parameters to be adjusted in the fitting and $\beta = 0.326$ is the critical exponent of the three-dimensional Ising model.[196] The DNN potential predicts $T_c = 857 \pm 17$ K and $\rho_c = 0.302 \pm 0.002$ g cm^{-3}, which are significantly different from the MB-pol values of $T_c = 639 \pm 14$ K and $\rho_c = 0.34 \pm 0.03$ g cm^{-3}. As a reference, the corresponding experimental values are $T_c = 647$ K and $\rho_c = 0.32$ g cm^{-3}.[197]

In an attempt to rationalize the different performance of the DNN potential in reproducing bulk and interfacial properties calculated with MB-pol, we investigated the ability of the DNN potential to correctly describe many-body interactions. By construction, MB-pol quantitatively reproduces each term of the MBE (Eq. 1.1) calculated at the CCSD(T)/CBS level.[97,98] In this context, we have shown that a correct representation of each individual n-body contribution to the interaction energies is required in order for a water model to be both accurate and transferable across different thermodynamic state points.[129,150,154,198–200]

Following previous studies,[158,198] in Fig. 1.6 we present a many-body decomposition analysis of the interaction energies of the first eight low-lying energy isomers of the hexamer cluster (Fig. 1.2). As mentioned in the Introduction, among water clusters, the hexamer occupies a special place because it is the smallest cluster with low-lying isomers that exhibit three-dimensional structures reminiscent of hydrogen-bonding arrangements found in liquid water and ice. In addition, the large number of isomers with similar interaction energies makes the hexamer the ideal benchmarking system for determining the accuracy

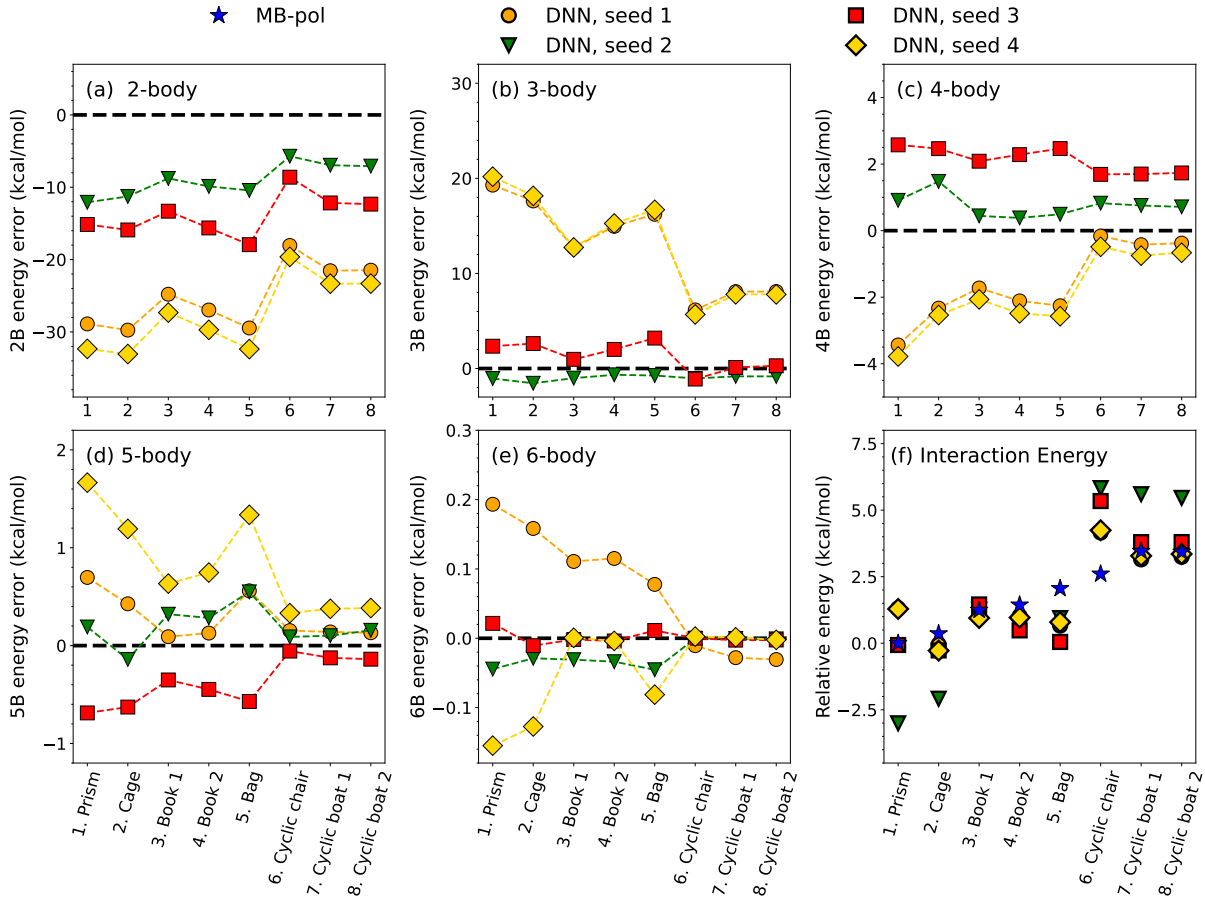

Figure 1.6. Many-body decomposition analysis for the eight low-lying energy isomers of the water hexamer (Fig. 1.2) calculated with four distinct DNN potentials trained on the same MB-pol training data using four different seeds to initialize the fitting process. Panels a) to e) show the errors associated with n-body energies ($n = 2 - 6$) calculated with the DNN potentials relative to the corresponding MB-pol values. Panel f) shows the errors associated with the interaction energies calculated with the DNN potentials relative to the corresponding MB-pol values. The DNN potential with seed 2 is used in the comparisons shown in Figs. 1.3-1.5.

of water models.[129] To provide a general perspective on DeePMD-based DNN potentials for water, in Fig. 1.6 we analyze the performance of four distinct DNN potentials trained on the same training set described in Section 1.2.2 but initialized using different random seeds, with seed 2 corresponding to the DNN potential used in Figs. 1.3-1.5.

All four DNN potentials provide statistically equivalent training, validation, and testing errors (see Tables S2 and S3 of the Supplementary Material). Fig. 1.6 shows that

none of the four DNN potentials is capable of correctly reproducing individual n-body energies ($n = 2 - 6$) calculated with MB-pol, with significantly large errors, on the order of $10 - 20$ kcal/mol, for 2- and 3-body contributions. Interestingly, these large errors compensate among different n-body contributions in such a way that, when the n-body energies are added together, they result in interaction energies that are in relatively better agreement with the MB-pol values than the individual n-body contributions. By definition, the interaction energies are calculated as the difference between the energy of the cluster and the sum of the 1-body energies of all six water molecules in the same distorted configurations as in the cluster. In this context, it should be noted that, besides MB-pol that reproduces the CCSD(T)/CBS reference energies of the hexamer isomers with chemical accuracy,[158] several modern polarizable force fields predict n-body and interaction energies of water clusters with significantly higher accuracy than the four DNN potentials examined here.[200] Direct comparisons between n-body and interaction energies calculated with the four distinct DNN potentials and the corresponding MB-pol reference values are shown in Fig. S3. Importantly, Fig. S3 shows that, besides displaying large errors, some of the DNN potentials (i.e., seed 1 and seed 4) also predict physically incorrect many-body contributions (e.g., positive 3-body contributions), which indicates that, in their conventional implementation, DeePMD-based DNN potentials are not able to correctly disentangle individual many-body contributions to the interaction energy of a given water system. Importantly, the inclusion of long-range effects through a classical electrostatic term does not improve the description of many-body energies as shown in Figs. S4 and S5 of the Supplementary Material. It should be noted that this behavior is not specific to DeePMD-based DNN potentials but appears to be common to other neural network potentials. For example, Figs. S6 and S7 of the Supplementary Material show that similar behavior is exhibited by Nequip-based potentials[92] trained on MB-pol. Interestingly, the Nequip-based potentials demonstrate superior accuracy in predicting the interaction energies of the water clusters, but also exhibit larger error compensation

among different n-body energies, with errors on 2- and 3-body energies being as large as $20 - 30$ kcal/mol.

1.3.2 DNN(VLE) potential

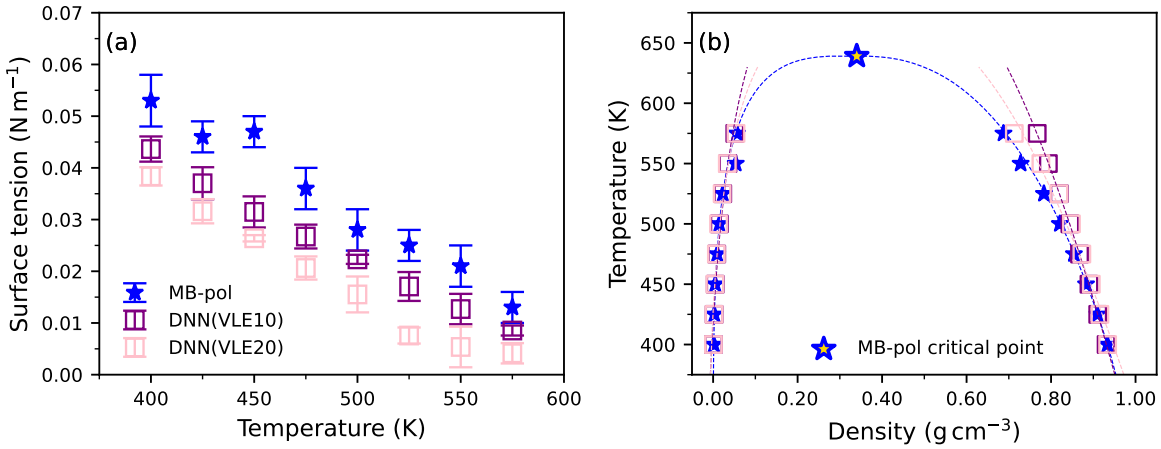

Figure 1.7. Surface tension (a) and vapor-liquid equilibrium densities (b) calculated from the present NVT simulations of a water slab carried out with the DNN(VLE10) (purple) and DNN(VLE20) (pink) compared with the reference MB-pol values from Ref. 165 (blue).

In an attempt to improve the performance of the DNN potential on vapor-liquid equilibrium properties, we used active learning to incorporate vapor-liquid configurations extracted from simulations carried out with the DNN potential in the temperature range between 268 K and 575 K. At the end of the active learning process, 2412 were added to the training set. The expanded training set was then used to train two potentials, DNN(VLE10) and DNN(VLE20), with a 10% and 20% probability of selecting vapor-liquid configurations during training, respectively. Fig. 1.7 shows that adding vapor-liquid configurations leads to more accurate predictions of both surface tension and vapor-liquid equilibrium densities. In particular, compared to the results obtained with the DNN potential, the surface tension predicted by both DNN(VLE10) and DNN(VLE20) shows the same temperature dependence as determined by MB-pol, although a systematic

deviation from the reference values is still observed at all temperatures. Similarly, adding vapor-liquid configurations to the training set improves the ability of the DNN(VLE10) and DNN(VLE20) potentials to describe the equilibrium densities of both vapor and liquid phases. While both DNN(VLE10) and DNN(VLE20) potentials quantitatively reproduce the MB-pol vapor densities over the entire temperature range, the predicted liquid densities, however, increasingly deviate from the MB-pol reference values as the temperature increases. As a consequence, the critical point is still overestimated by both potentials, with DNN(VLE10) predicting $T_c = 718 \pm 4$ K and $\rho_c = 0.359 \pm 0.003$ g cm^{-3} and DNN(VLE20) predicting $T_c = 674 \pm 6$ K and $\rho_c = 0.347 \pm 0.005$ g cm^{-3}. Adding vapor-liquid configurations to the training set was reported to enable simulations of "water along its liquid/vapor coexistence line with unprecedented precision".[80] Inspection of Fig. 3 of Ref. 80, however, indicates that relatively large deviations (similar to those found for DNN(VLE10) and DNN(VLE20) in Fig. 1.7) exist between the vapor and liquid densities predicted by the neural network potential used in the simulations and the corresponding reference RPBE-D3 values which, when extrapolated, lead to very different estimates for the critical point.[201]

To assess the ability of DNN(VLE10) and DNN(VLE20) to reproduce properties that do not directly depend on the coexistence between vapor and liquid phases, we examined the performance of both potentials on the same bulk and cluster properties used in Section 1.3.1 to determine the accuracy of the DNN potential. Fig. 1.8 shows the temperature dependence of the density and isothermal compressibility of liquid water predicted by DNN(VLE10) and DNN(VLE20). While both potentials are able to qualitatively reproduce the MB-pol trends, a comparison with the DNN results reported in Fig. 1.3 indicates that the inclusion of vapor-liquid configurations to the training set deteriorates the ability of the DeePMD-based potentials to reproduce the bulk properties. This is further confirmed by the analyses of the liquid density, RDFs, and q_{tet} distributions shown for the DNN(VLE20) potential in Figs. S17 and S18 of the Supplementary Material.

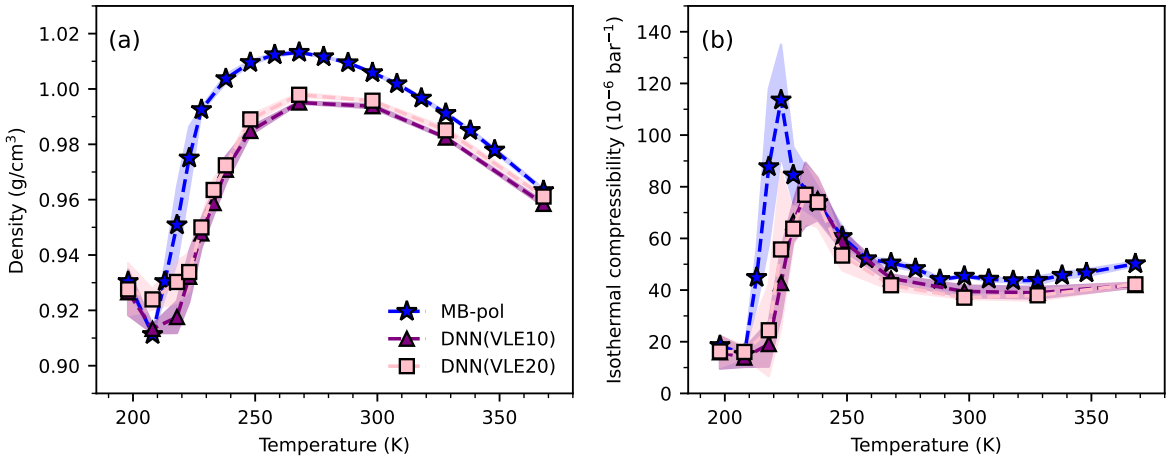

Figure 1.8. Temperature dependence of the density (a) and isothermal compressibility (b) calculated from NPT simulations carried out with the DNN(VLE10) (purple) and DNN(VLE20) (pink) potentials at 1 atm. Also shown for reference are the corresponding MB-pol values from Ref. 43 (blue). The associated shaded areas indicate 95% confidence intervals of the averages.

Finally, Fig. 1.9 reports the many-body decomposition analysis of the interaction energies of the hexamer isomers (Fig. 1.2) carried out with the DNN(VLE20) potential. The corresponding analysis carried out with DNN(VLE10) is reported in the Supplementary Material in Fig. S8. As for the DNN potential, we used four different seeds to develop four distinct DNN(VLE20) potentials that were trained on the expanded MB-pol training set containing vapor-liquid configurations. Seed 4 corresponds to the DNN(VLE20) potential used in the comparisons shown in Figs. 1.7 and 1.8. As in the case of DNN in Fig. 1.6, none of DNN(VLE20) potentials is able to correctly reproduce the reference MB-pol many-body energies, with errors that are on the order of $\sim$10 kcal/mol for 2-, 3-, and 4-body energies. Similar poor performance on the many-body decomposition analysis is exhibited by the DNN(VLE10) potential in Fig. S8 of the Supplementary Material. Analyses analogous to those shown in Fig. S3 for the DNN potential are reported in Figs. S9 and S10 for the DNN(VLE10) and DNN(VLE20) potentials, which lead to similar conclusions, i.e., both DNN(VLE10) and DNN(VLE20) predict physically incorrect many-body energies.

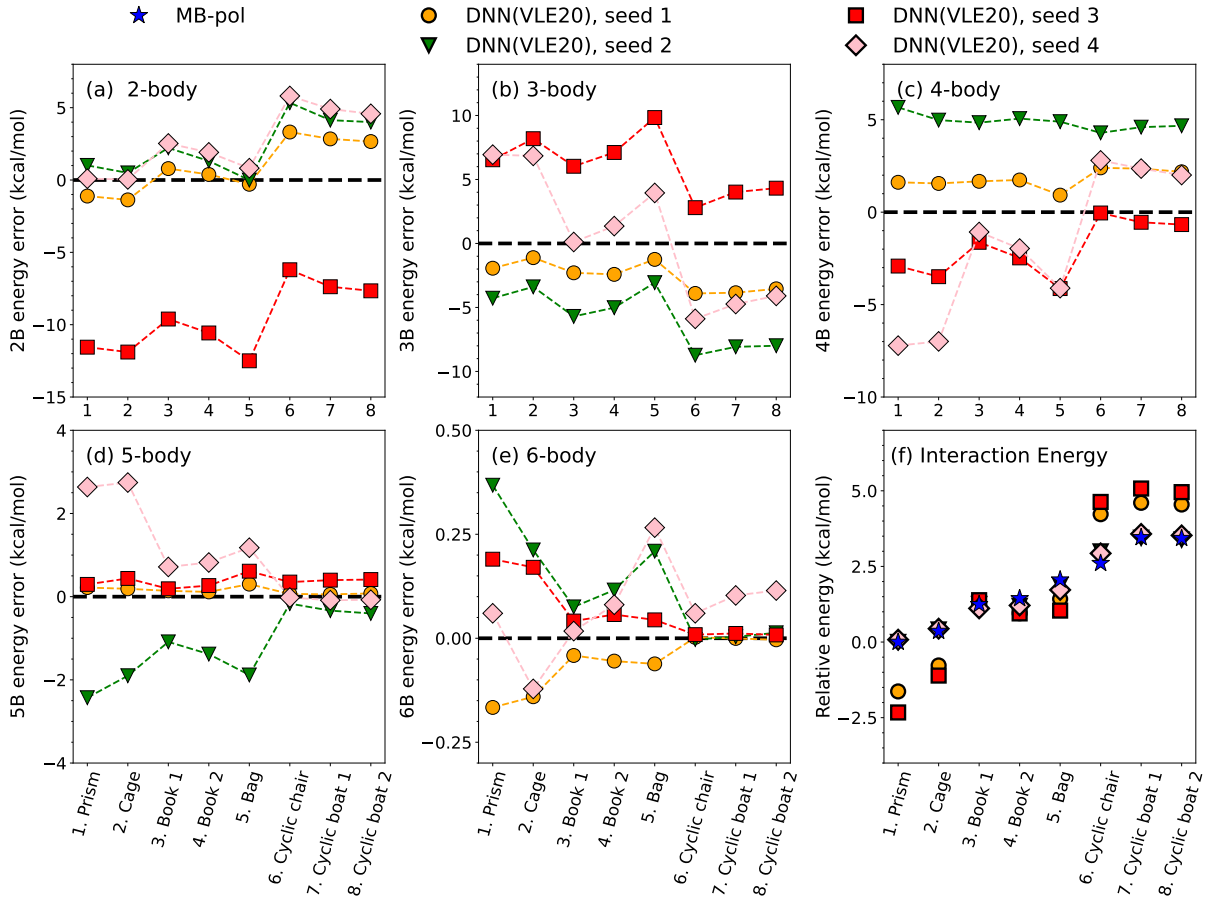

Figure 1.9. Many-body decomposition analysis for the eight low-lying energy isomers of the water hexamer (Fig. 1.2) calculated with four distinct DNN(VLE20) potentials that were trained on the expanded MB-pol training set containing vapor-liquid configurations using four different seeds to initialize the fitting process. Panels a) to e) show the errors associated with n-body energies ($n = 2 - 6$) calculated with the DNN(VLE20) potentials relative to the corresponding MB-pol values. Panel f) shows the errors associated with the interaction energies calculated with the DNN(VLE20) potentials relative to the corresponding MB-pol values. The DNN(VLE20) potential with seed 4 is used in the comparisons shown in Figs. 1.7 and 1.8.

The analyses presented in this section demonstrate that, while the description of vapor-liquid equilibrium properties can be improved by adding vapor-liquid configurations to the original DNN training set, this improvement is achieved at the cost of a less accurate representation of the bulk properties. Importantly, as in the case of the DNN potentials, the DNN(VLE10) and DNN(VLE20) potentials are unable to correctly capture the physics

of many-body interactions in water.

1.3.3 DNN(MB) potential

Since the inability of a water model to correctly represent many-body contributions to the underlying molecular interactions appears to be correlated with the lack of transferability of the model across different thermodynamic state points,[129] we investigated the possibility of "encoding" many-body effects in the DNN potentials within the DeePMD framework. To this end, we supplemented the original training set used for developing the DNN potentials discussed in Section 1.3.1 with a set of gas-phase clusters, including monomers, dimers, trimers, and tetramers which provides direct information about the low-order and most important terms (i.e., 1-body to 4-body terms) of the MBE in Eq. 1.1. We then used the expanded training set to train three different DeePMD-based potentials, referred to as DNN(MB4), DNN(MB10), and DNN(MB20), with 4%, 10%, and 20% probability of selecting gas-phase cluster configurations during the training process, respectively. Specific details about the composition of the extended training set are provided in Section S1 of the Supplementary Material.

Fig. 1.10 reports the same analysis reported in Fig. 1.6 for the DNN potential and shows the errors relative to the MB-pol reference values for n-body and interaction energies of the first eight isomers of the water hexamer (Fig. 1.2) calculated with four distinct versions of the DNN(MB20) potential which were trained on the same expanded training set but initialized with four distinct seeds. Seed 4 corresponds to the DNN(MB20) potential used in the comparisons shown in Figs. 11 and 12. Analogous analyses carried out with the DNN(MB4), and DNN(MB10) potentials are reported in Figs. S11 and S12 of the Supplementary Material, respectively. The addition of monomer, dimer, trimer, and tetramer configurations clearly allows the DNN(MB) potentials to become "aware" of the existence of distinct many-body contributions to the interactions energies, as demonstrated by the relatively smaller errors displayed by the DNN(MB20) 2-body, and 3-body energies

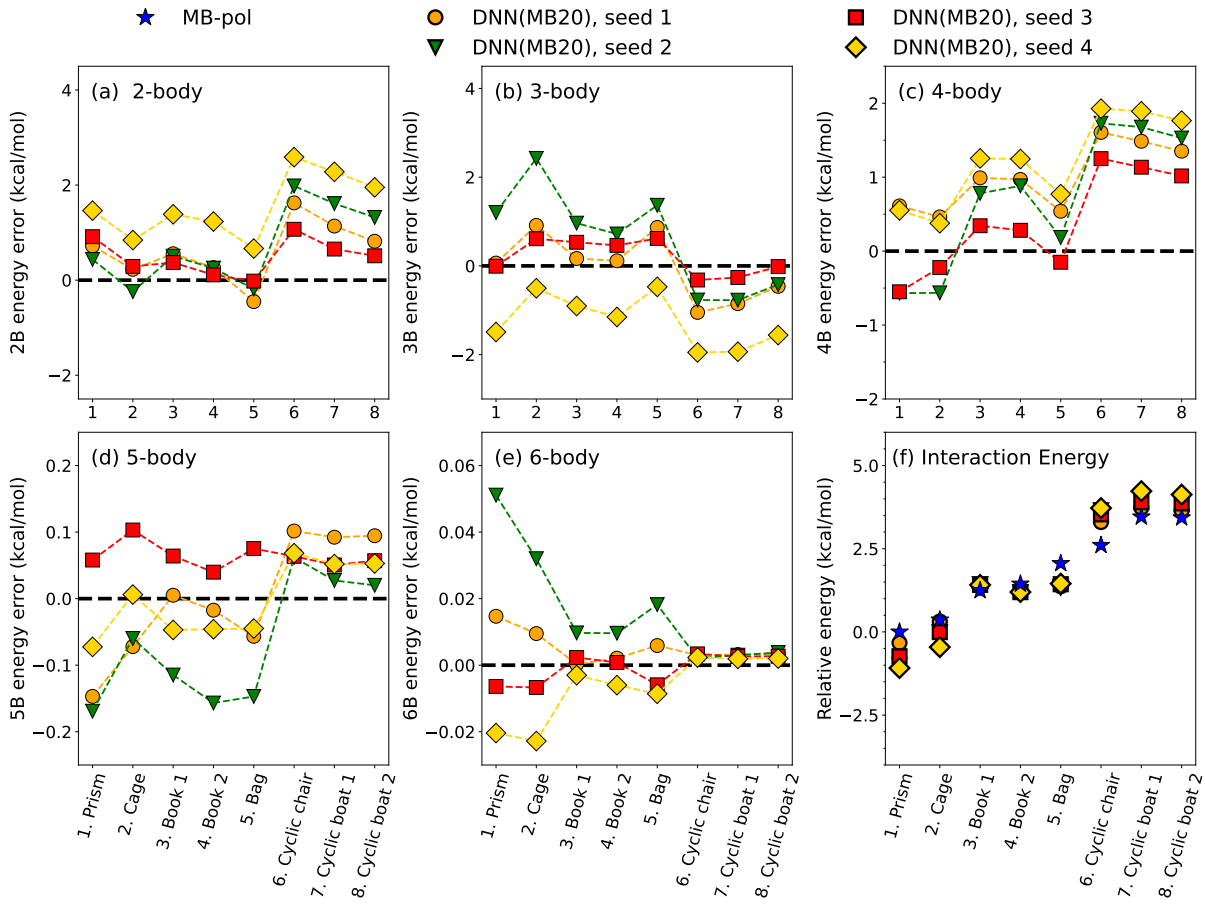

Figure 1.10. Many-body decomposition analysis for the eight low-lying energy isomers of the water hexamer (Fig. 1.2) calculated with four distinct DNN(MB20) potentials that were trained on the expanded MB-pol training set containing cluster configurations using four different seeds to initialize the fitting process. Panels a) to e) show the errors associated with n-body energies ($n = 2 - 6$) calculated with the DNN(MB20) potentials relative to the corresponding MB-pol values. Panel f) shows the errors associated with the interaction energies calculated with the DNN(MB20) potentials relative to the corresponding MB-pol values. The DNN(MB20) potential with seed 4 is used in the comparisons shown in Figs. 1.11 and 1.13.

compared to the corresponding errors associated with the DNN potentials in Fig. 1.6. Similar trends are observed in Figs. S13, S14, and S15 that show direct comparisons of individual many-body energies and interaction energies calculated with the DNN(MB4), DNN(MB10), and DNN(MB20) potentials, respectively. Additional analyses of the error distributions associated with 2-body and 3-body energies calculated for dimer and trimer

configurations of the cluster training set are reported in Fig. S16 and demonstrate that the DNN(MB4), DNN(MB10), and DNN(MB20) potentials significantly improve on the DNN potential in the ability to represent many-body interactions in water. It should, however, be noted that the DNN(MB) potentials still rely on significant error compensation among individual n-body energy contributions to minimize the error on the interaction energies of the hexamer isomers. The observed error compensation indicates that, as the DNN, DNN(VLE10), and DNN(VLE20) potentials, the DNN(MB) potentials are unable to "learn" that the interaction energy of a N-body system containing N water molecules is given by the sum of distinct n-body energy contributions (with $n = 2 - N$). To put things in perspective, the errors associated with the DNN(MB) predictions for each n-body energy contribution to the interaction energies of the hexamer isomers, in particular at the 2-body and 3-body levels, are still appreciably larger than those displayed by state-of-the-art polarizable force fields for water.[200]

Having demonstrated that extending the training set by adding monomer, dimer, trimer, and tetramer configurations allows the DNN(MB) potentials to recover a more balanced representation of many-body interactions, we next assess the ability of the DNN(MB4), DNN(MB10), and DNN(MB20) potentials to reproduce vapor-liquid equilibrium properties that were poorly predicted by the DNN potentials (Fig. 1.5). Fig. 1.11 shows that all three DNN(MB) potentials more closely reproduce the MB-pol trends for the surface tension and the equilibrium densities of both vapor and liquid phases over the entire temperature range examined in this study than the DNN potential. The critical parameters predicted by the DNN(MB4), DNN(MB10), and DNN(MB20) potential are $T_c = 655 \pm 2$ K and $\rho_c = 0.325 \pm 0.002$ g cm^{-3}, $T_c = 605 \pm 10$ K and $\rho_c = 0.32 \pm 0.01$ g cm^{-3}, and $T_c = 660 \pm 6$ K and $\rho_c = 0.338 \pm 0.005$ g cm^{-3}, respectively, which are in better agreement with the MB-pol values ($T_c = 639 \pm 14$ K and $\rho_c = 0.34 \pm 0.03$ g cm^{-3}) than the results obtained not only with the DNN potential but also with the DNN(VLE10) and DNN(VLE20) potentials. The structural differences at the vapor-liquid equilibrium

between DNN and DNN(MB20) are further highlighted in Fig. S17 which shows the density profiles predicted by the two potentials at different temperatures. In particular, the interface structure predicted by DNN(MB20) is significantly different from that predicted by DNN and in close agreement with the MB-pol results reported in Ref. 165.

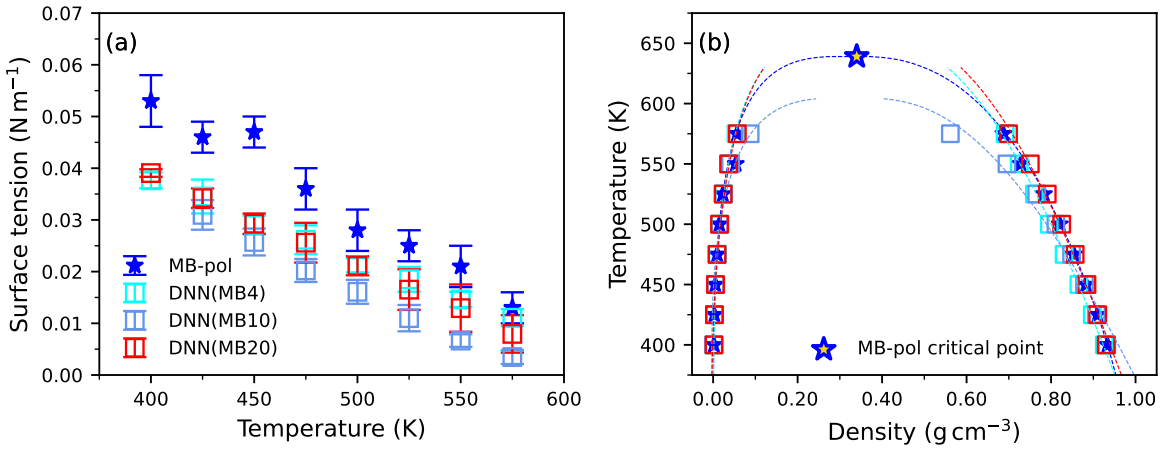

Figure 1.11. Surface tension (a) and vapor-liquid equilibrium densities (b) calculated from NVT simulations of a water slab carried out with the DNN(MB) potentials (cyan, light blue, red) compared with the reference MB-pol values from Ref. 165 (blue).

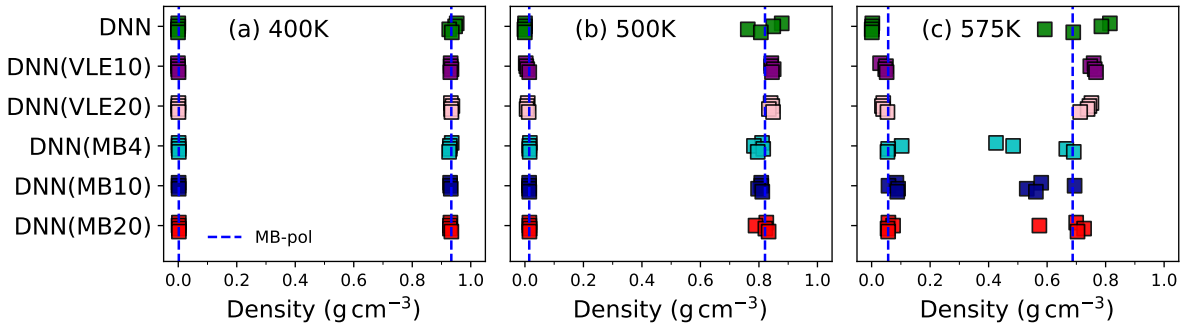

Figure 1.12. Vapor-liquid equilibrium densities calculated using the same four variants of each of the DNN, DNN(VLE), and DNN(MB) potentials used in the analyses of Fig. 1.6, 1.9, S8, 1.10, S11, and S12. The densities are compared with the reference MB-pol values from Ref. 165 (blue dashed line) at 400 K (panel a), 500 K (panel b), and 575 K (panel c).

Despite being able to provide more accurate estimates of the actual MB-pol critical point, Fig. 1.12 shows that the DNN(MB4), DNN(MB10), and DNN(MB20) potentials

display a higher degree of variability in their predictions of the liquid density at high temperatures than the DNN(VLE10) and DNN(VLE20) potentials. This high variability at high temperatures, which is also displayed by the DNN potentials, can be traced back to the lack of explicit vapor-liquid configurations in the corresponding training sets, which, in turn, highlights the difficulty for DeePMD-based potentials to be transferable across different phases over a wide range of thermodynamic conditions.

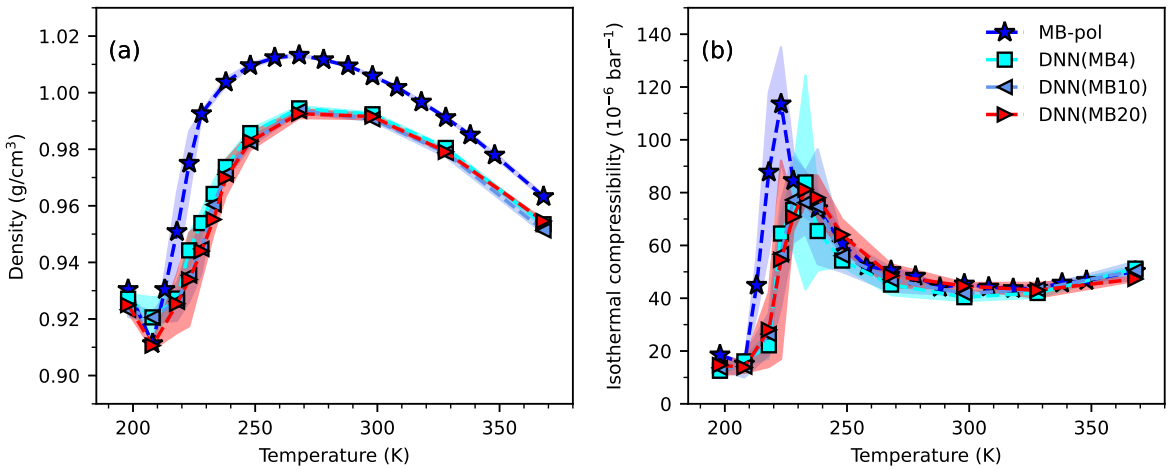

Figure 1.13. Temperature dependence of the density (a) and isothermal compressibility (b) calculated from NPT simulations carried out with the DNN(MB) potentials at 1 atm (cyan, light blue, red) compared with the reference MB-pol values from Ref. 43 (blue). The associated shaded areas indicate 95% confidence intervals of the averages. Fig. S2 show the density fluctuations along the NPT trajectories carried out with the DNN(MB20) potentials at each temperature.

The last question that remains to be addressed is whether the improved ability to represent many-body interactions and predict vapor-liquid equilibrium properties still allows the DNN(MB4), DNN(MB10), and DNN(MB20) potentials to accurately reproduce the liquid properties calculated with MB-pol. To this end, Fig. 1.13 shows comparisons between the temperature dependence of the density and isothermal compressibility calculated with the DNN(MB4), DNN(MB10), and DNN(MB20) potentials and the corresponding MB-pol reference values. The DNN(MB4), DNN(MB10), and DNN(MB20) potentials effectively predict indistinguishable (within statistical error) trends for both density and

isothermal compressibility, which are in qualitative agreement with the MB-pol reference values. Specifically, they correctly predict a minimum at ∼300 K in the isothermal compressibility which is instead absent in the DNN, DNN(VLE10), and DNN(VLE20) potentials. This suggests that, by being able to more accurately represent many-body interactions, the DNN(MB4), DNN(MB10), and DNN(MB20) potentials display a higher degree of transferability to the gas phase at ambient conditions. As in the case of the DNN(VLE10) and DNN(VLE20) potentials, the comparison between the results of Fig. 1.13 and Fig. 1.3, however, indicates that the addition of configurations different from bulk configurations (in this case, monomer, dimer, trimer, and tetramer configurations) to the training set overall deteriorates the ability of the DNN(MB4), DNN(MB10), and DNN(MB20) potentials to reproduce bulk properties calculated with MB-pol. Despite these differences, the liquid structure predicted by the DNN(MB4), DNN(MB10), and DNN(MB20) potentials is in close agreement with that of MB-pol as demonstrated by the comparisons of the RDFs and q_{tet} distributions calculated with DNN(MB20) that are shown in Fig. S18 of the Supplementary Material.

1.4 Conclusion

In this study, we analyzed the performance and degree of transferability of a DeePMD-based DNN potential for water trained on MB-pol reference configurations extracted from MD simulations of liquid water carried out from 198 K to 368 K at 1 atm. We found that the DNN potential is able to reliably reproduce structural and thermodynamic properties of liquid water as predicted by MB-pol from the boiling point down to deeply supercooled temperatures. However, while MB-pol exhibits remarkable accuracy from the gas to the condensed phase, the DNN potential does not share the same high level of transferability across phases. In particular, we found that the DNN potential is not able to accurately describe vapor-liquid equilibrium properties. More importantly,

a many-body decomposition analysis of the interaction energies of the hexamer isomers indicates that the DNN potential is not able to correctly "learn" many-body interactions and effectively relies on error compensation among individual many-body energy contributions to reproduce the interaction energy of a given N-body system containing N water molecules.

To improve the performance of the DNN potential on vapor-liquid equilibrium properties, we expanded the initial DNN training set of bulk configurations by adding configurations extracted from vapor-liquid equilibrium simulations carried out with MB-pol. While the new DNN(VLE10) and DNN(VLE20) potentials improve the description of the surface tension and equilibrium densities of both vapor and liquid phases, they predict less accurate bulk properties and are unable to correctly reproduce individual many-body contributions to the interaction energies.

In an attempt to explicitly encode many-body interactions onto the DNN potential, we also expanded the initial DNN training set by adding water monomer, dimer, trimer, and tetramer configurations, which provide direct information on the most important many-body contributions (i.e., 1-body to 4-body contributions) to the interaction energies in water systems. By improving the description of individual many-body contributions, the new DNN(MB4), DNN(MB10), and DNN(MB20) potentials are also able to reliably reproduce the vapor-liquid equilibrium properties predicted by MB-pol. We found, however, that all three potentials exhibit a high degree of variability in predicting the liquid density at high temperatures due to the lack of representative vapor-liquid configurations in their training sets, which limits their transferability over a wide range of thermodynamic conditions. Moreover, the improvement in the description of many-body interactions comes at the expense of a poorer representation of the liquid properties.

Although DeePMD-based potentials are intrinsically many-body in their functional form, our analyses show that they do not necessarily correctly represent the underlying many-body physics of the reference potentials. This suggests that some caution should be exercised when using DeePMD-based DNN potentials to predict thermodynamic properties

for state points that are not explicitly and thoroughly included in the training sets. Although our study focuses on water, similar behavior is likely to be found in DeePMD-based DNN potentials for other molecular systems, including aqueous solutions as well as molecular fluids and solids. In this context, we hope that our results can stimulate further developments of new training procedures and neural network architectures capable of correctly capturing the physics of many-body interactions in molecular systems.

With this caveat in mind, the computational efficiency provided by the DeePMD framework suggests that large-scale CCSD(T)-level MD simulations are possible by training DeePMD-based DNN potentials on data-driven many-body potentials derived from the MBE calculated at the CCSD(T) level of theory, such as MB-pol. However, for this to hold, the thermodynamic state points of interest in the DNN simulations must be adequately represented in the training sets generated using the reference data-driven many-body potentials. This suggests that a DNN potential trained on an extensive training set, including molecular configurations extracted from MB-pol simulations carried out over a wide range of thermodynamic conditions, is well suited for exploring the rich phase diagram of water,[44] particularly in the so-called "no man's land" region at low temperature, which has been proven difficult to probe experimentally.[172,174,175]

1.5 Acknowledgements

Chapter 1, in full, is a reprint of the material as it appears in "A "short blanket" dilemma for a state-of-the-art neural network potential for water: Reproducing experimental properties or the physics of the underlying many-body interactions?" Y. Zhai, A. Caruso, S. L. Bore, Z. Luo, and F. Paesani. In: *J. Chem. Phys.* 158.8 (2023), p. 084111. The dissertation author is the co-primary investigator and author of this paper.

We thank Maria Muniz, Athanassios Panagiotopoulos, and Vinicius Cruzeiro for stimulating discussions at the early stage of this research. This research was supported

by the Air Force Office of Scientific Research under award FA9550-20-1-0351 and used computational resources of the Department of Defense High Performance Computing Modernization Program (HPCMP) as well as the Triton Shared Computing Cluster (TSCC) at the San Diego Supercomputer Center (SDSC). This work was supported in part by the UC Southern California Hub, with funding from the UC National Laboratories division of the University of California Office of the President.

Chapter 2

Active learning of many-body configuration space: Application to the Cs^+–water MB-nrg potential energy function as a case study

2.1 Introduction

Computer simulations provide fundamental insights into the properties and behavior of molecular systems.[202–204] Since both accuracy and predictive ability of a molecular model are primarily limited by the computational cost associated with the model itself, developing cost-effective simulation approaches is key to studying increasingly more complex systems. It has recently become possible to perform molecular dynamics (MD) simulations of aqueous systems, from the gas to the condensed phase, retaining high accuracy in the description of the underlying molecular interactions.[129] This is achieved by employing many-body potential energy functions (PEFs) derived from high-level electronic structure data that are carried out on selected molecular configurations representative of the corresponding global many-body potential energy surfaces (PESs).[94–98,103,106–108] An optimal approach to the development of many-body PEFs would require identifying a minimal pool of configurations that can guarantee an accurate description of the system under exam and, at the same time, computation time is not lost on calculations on redundant configurations

describing similar regions of the many-body PES.

Efficient sampling of the configuration space is challenging due to the high dimensionality of the associated molecular configurations. In principle, a regular grid search would provide a homogeneous representation of all regions of the many-body PES. This approach, however, becomes unfeasible as the number of degrees of freedom increases. To reduce the size of the configuration space, it is common practice in the development of many-body PEFs to apply biases on the relative translations and rotations of the individual molecular species constituting the system under exam.[97,98,107,108] Although of practical use, this approach can lead to redundant training sets containing several molecular configurations representing similar regions of the target many-body PES. While algorithms designed to remove geometrically similar configurations exist, it is not guaranteed that screening based on structural similarity is sufficient for identifying only configurations necessary for a faithful description of the target many-body PES.

The success of machine learning (ML) in many areas of molecular sciences (e.g., see Refs. 68,71,74,76,205–209,209–215) makes it a promising tool for efficiently screening large pools of molecular configurations for the development of many-body PEFs. Most common ML approaches rely on supervised learning, which, however, requires large set of known labeled data to train a model capable to accurately predict the labels of previously unseen data.[216–218] Active learning (AL) provides a potential solution to the need for constructing beforehand large training sets by interactively generating training configurations at runtime. AL schemes are thus particularly appealing when using large training sets is prohibitively expensive either because of the high cost associated with determining the data labels or because of the high computational cost of the training stage.

In this study, we investigate the application of AL to generating representative training sets of molecular configurations necessary for the development of many-body PEFs, with a specific focus on two-body (2B) and three-body (3B) contributions to the Cs^+–water interaction energies. Our AL framework consists of a finite pool of molecular

configurations (i.e., $Cs^+(H_2O)$ dimers for the 2B pool and $Cs^+(H_2O)_2$ trimers for the 3B pool) whose energies are unknown, a training set with configurations selected from the pool, a predictive model (predictor) thirsting for the training set, and a learner that actively selects configurations from the pool. We assume that the size of the pool is beyond awareness of the learner and only a subset of the configurations (referred to as candidates) in the pool are available to the learner at each iteration. Through the application of our AL approach, we demonstrate that the size of the original pool of configurations used to develop the Cs^+–water MB-nrg PEF can be greatly reduced without compromising the accuracy with which the new MB-nrg PEFs describe Cs^+–water interactions, from small clusters to aqueous solutions.

2.2 Methods

2.2.1 MB-nrg potential energy functions

The total energy of a system containing N (atomic or molecular) monomers ("bodies"), can be rigorously expressed through the many-body expansion (MBE) of the energy,[219]

$$V_N = \sum_{i}^{N} V_i^{1B} + \sum_{i<j}^{N} V_{ij}^{2B} + \sum_{i<j<k}^{N} V_{ijk}^{3B} + \ldots + V^{NB} \tag{2.1}$$

where the V_i^{1B} corresponds to the energy required to distort monomer i from its equilibrium geometry. Therefore, $V^{1B}(i) = 0$ for atomic monomers, and $V^{1B}(i) = E(i) - E_{eq}(i)$ for molecular monomers, where $E(i)$ and $E_{eq}(i)$ are the energies of monomer i in distorted and equilibrium geometries, respectively. All higher n-body (nB) interaction terms (V^{nB}) in Eq. 2.2.1 are defined recursively through

$$V^{nB}(1,\ldots,n) = E_n(1,\ldots,n) - \sum_i V^{1B}(i) - \sum_{i<j} V^{2B}(i,j) - \ldots - \sum_{i<j<\cdots<n-1} V^{(n-1)B}(i,j,\ldots,(n-1)) \tag{2.2}$$

Within the MB-nrg framework, the water–water interactions are described by the MB-pol PEF,[97,98,106] which has been shown to correctly reproduce the properties of water[110,158] from small clusters in the gas phase,[159–161,220–229] to bulk water,[162,230–232] the air/water interface,[163,164,233,234] and ice.[166–168] The interactions between Cs^+ ions and water molecules are described through the MBE of Eq. 2.2.1. Specifically, the Cs^+–water MB-nrg PEF includes explicit 2B Cs^+–H_2O and 3B Cs^+–$(H_2O)_2$ terms, with all higher-order interactions being implicitly taken into account through a classical many-body polarization term.[108,235] The 2B term includes three contributions,

$$V^{2B} = V^{2B}_{short} + V^{2B}_{TTM} + V^{2B}_{disp} \tag{2.3}$$

where V^{2B}_{disp} is the 2B dispersion energy, and V^{2B}_{TTM} is the 2B classical polarization contribution described by a Thole-type model.[236] V^{2B}_{short} in Eq. 2.3 describes 2B short-range contributions represented by a 5th-degree permutationally invariant polynomial (PIP) in variables that are functions of the distances between the Cs^+ ion and each of the six sites of the MB-pol water molecule.[108]

Similarly, the 3B term of the Cs^+–water MB-nrg PEF includes two contributions,

$$V^{3B} = V^{3B}_{short} + V^{3B}_{TTM} \tag{2.4}$$

where V^{3B}_{TTM} is the 3B classical polarization contribution described by the same Thole-type model as in V^{2B}_{TTM}, and V^{3B}_{short} describes 3B short-range contributions that are represented

by a 4th-degree PIP in variables that are functions of the same distances as in V_{short}^{2B}.[235] The coefficients of both 2B and 3B PIPs were optimized using Tikhonov regression (also known as ridge regression)[237] to reproduce reference interaction energies obtained from high-level electronic structure calculations.

2.2.2 Interaction energies, fitting procedure, and MD simulations

The 2B and 3B reference energies were taken from Refs. 108 and 235 where MOLPRO (version 2015.1) was used to carry out electronic structure calculations at the coupled cluster level of theory using single, double and perturbative triple excitations, i.e., CCSD(T), the "gold standard" for chemical accuracy.[117] In Ref. 108, the 2B CCSD(T) energies were calculated in the complete basis set (CBS) limit that was achieved through a two-point extrapolation[138,139] between the values obtained with the correlation-consistent polarized valence triple zeta (aug-cc-pVTZ for H,O, and cc-pwCVTZ for Cs^+) and quadruple zeta (aug-cc-pVQZ for H,O, and cc-pwCVQZ for Cs^+) basis sets.[238–241] In Ref. 235, the 3B CCSD(T) energies were calculated using the aug-cc-pVTZ basis set for the O and H atoms, and the cc-pwCVTZ basis set for Cs^+, and were corrected for the basis set superposition error using the counterpoise method.[242] In both 2B and 3B energy calculations, the ECP46MDF pseudopotential was used for the core electrons of Cs^+.[243]

The original 2B training set consisted of $Cs^+(H_2O)$ dimer configurations generated on a uniform spherical grid, with the Cs^+–O distance in the 1.6 - 8 Å range.[108] For the present study, dimer configurations with interaction energies larger than 100 kcal/mol were removed since they were found to be not necessary for representing $Cs^+(H_2O)$ configurations sampled in MD simulations at ambient conditions. The 2B pool was then further reduced to 13525 dimer configurations after randomly removing 1547 configurations for the 2B test set.

Due to the larger number of degrees of freedom, the original 3B training set was

generated in Ref. 235 by extracting $Cs^+(H_2O)_2$ trimer configurations from MD simulations of a single Cs^+ ion in liquid water at 298.15 K. For the present study, the original 3B set of Ref. 235 was reduced to a 3B pool of 34441 configurations after randomly removing 4480 configurations for the 3B test set.

The MD simulations presented in Section 2.3.4 were carried out in the isobaric-isothermal (NPT) ensemble for a box containing a single Cs^+ ion and 277 H_2O molecules. The equations of motion were propagated using the velocity-Verlet algorithm with a timestep δt of 0.2 fs. The temperature of 298.15 K was controlled by Nosé-Hoover chains of 4 thermostats attached to each degree of freedom while the pressure of 1.0 atm was controlled following the algorithm described in Ref. 244. All MD simulations were carried out using an in-house software based on DL_POLY 2.0.[245]

2.2.3 Active learning

An AL framework based on uncertainty and error estimation was used to generate optimal 2B and 3B training sets with the goal of reducing the number of dimers and trimers necessary to develop Cs^+–water MB-nrg PEF, without compromising accuracy. The major difficulty faced by the active learner in generating optimal 2B and 3B training sets is represented by the need to determine the relevance of candidate dimer and trimer configurations before knowing the associated 2B and 3B energies.[246] It is apparent that the more accurate the active learner is, the more precise its assessment of a molecular configuration is. In addition, for efficiency purposes, the energy estimation made by the learner should be computationally inexpensive compared to the energy determination performed by the predictor.

In this context, Gaussian process regression (GPR) provides a general approach to assessing the relevance of a candidate configuration by accurately estimating the associated energy.[247] GPR implies a correlation between the unknown energies of the candidate configurations and the energies determined for configurations that are already in the

training sets. The correlation is expressed by the covariance matrix between known and unknown values of the energies, with the elements of the covariance matrix being calculated by a kernel function. GPR assumes that both known and unknown energies are distributed according to a multidimensional Gaussian distribution and then uses the covariance matrix to predict the conditional probability distribution of the unknown energies given the known energies. The ability of GPR to interpolate between known energy values makes it a good model for local uncertainty prediction. It should be noted that a similar approach is exploited by Gaussian Approximation Potential (GAP) models that have been developed to represent interatomic interactions.[248]

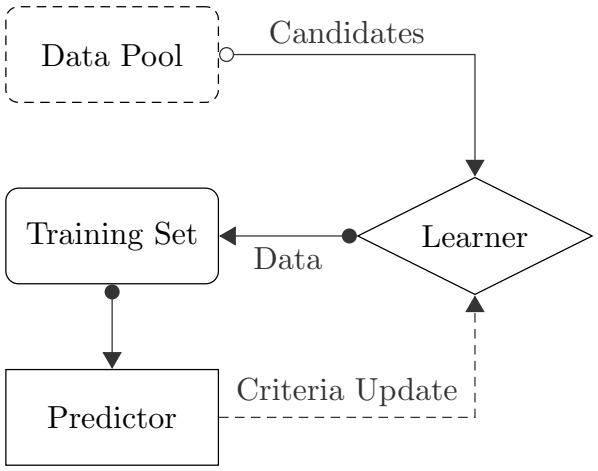

Figure 2.1. Schematic representation of the AL framework introduced in this study.

Our AL framework, shown in Fig. 2.1, consists of a pool of an unknown number of molecular configurations, corresponding to $Cs^+(H_2O)$ dimers for the 2B pool and $Cs^+(H_2O)_2$ trimers for the 3B pool, a predictor, and a learner that, based on feedback from the predictor, selects configurations from the pool and adds them to the training set. The complete AL protocol is summarized below:

- At each iteration t, the pool S sends a subset of configurations with unknown energies ($C_t = \{x_j\}_t \subseteq S$) to the learner as training set candidates.

- Depending on the iteration index t, a training set T_t is formed:
  - At $t = 0$, all configurations in C_0 are added to the training set T_0 and their actual energies are determined.
  - For $t > 0$:
    * The training set T_{t-1} from the previous iteration is divided into clusters $\{\tau_{t-1,k}\}$ containing a fixed number of molecular configurations, independent of the training set size.
    * A cluster label k_j is predicted for each candidate configuration x_j in C_t (i.e., each candidate configuration x_j is assigned to one of the clusters $\{\tau_{t-1,k}\}$).
    * The uncertainty ΔE_j on the energy of the candidate configuration x_j is estimated as the GPR variance calculated for the entire cluster τ_k, $k = k_j$.
    * The error Err_j on the energy of the candidate configuration x_j is defined as the average error associated with the energies predicted by the model for all the configurations in the cluster τ_k, $k = k_j$.
    * A selection probability $P_t(x_j)$, proportional to the weighted sum of the energy uncertainty and the energy error, is assigned to each candidate configuration x_j in C_t,

$$P_t(x_j) \propto [w_{\Delta E} * \Delta E_j + w_{Err} * Err_j] \tag{2.5}$$

    * A subset of configurations $\{\hat{x}_i\}_t \subseteq C_t$ is selected and, after determining the associated actual energies ϵ_i, added to the training set, $T_t = \{(\hat{x}_i, \epsilon_i)\}_t \cup T_{t-1}$.
- The model M is trained on the training set T_t.
- The errors associated with the energies predicted by the model for all configurations in the training set T_t are updated

- The cycle is stopped when the gradient of the test error becomes lower than a predefined value.

The division into clusters $\{\tau_{t-1,k}\}$ of equal size reduces the computational cost associated with GPR, which typically scales as $O(n^3)$.[247] Since a radial basis function (RBF) kernel, which is based on the L2 distance, is used to determine the similarity between two configurations, it follows that configurations close to the candidate configuration play a central role in the GPR process. The use of the RBF kernel function allows interpolation with GPR only between configurations that are in the same cluster as the candidate, which, in turn, helps reduce the computational cost without losing predictive accuracy. As shown in Eq. 2.5, the learner selects configurations based on the weighted sum of uncertainty and model error. This procedure ensures a balanced exploration of the configuration space, exploiting the decision-making process.

Different reduction methods that exploit either molecular features[249] or model diversity[250] have recently been proposed. While some are based on correlation estimation as in our AL framework, approaches solely based on molecular features lack the adaptability that arises from the constant feedback of the fitting model. In contrast, our AL framework improves its ability to select new structures as the process advances: a small subset (5%) of candidates is selected and added to the training set to improve the reliability of the learner, at each iteration. As our AL framework, approaches based on model diversity, such as the query by committee (QBC) methods of Ref. 250, share similar advantages over feature-based approaches. However, our AL framework differs from the QBC method of Ref. 250 in three main aspects. 1) Our AL framework does not assume any knowledge about the initial pool. In contrast, since the inclusion criterion used in Ref. 250 is chosen empirically, the resulting AL approach is system dependent. 2) Since the candidates in Ref. 250 are selected based on a preset value of the inclusion criterion, a balanced exploration of the configuration space is not guaranteed. Our AL framework instead

assigns a probability to each configuration in the pool. This implies that there always exists the possibility to select low-probability candidates, which, consequently, guarantee a balanced exploration of the configuration space. 3) While the AL approach used in Ref. 250 only relies on the standard deviation calculated using the predictor model, our AL framework exploits both the training error calculated using the predictor model (i.e., the MB-nrg PEF) and the uncertainty calculated using the Gaussian process regression, which results in a performance improvement of the overall AL framework. In this context, it should be noted that, although our AL framework improves upon reduction methods that exploit either molecular features[249] or model diversity[250], a perfect training set reduction may still not be achieved due to the practical impossibility of achieving a perfect balance between exploration of completely new configurations and exploitation of configurations already in the training sets.

In this study, we used the KMeans module available in the Scikit-learn Python package, *version 0.21.3*, to cluster both the $Cs^+(H_2O)$ dimers and the $Cs^+(H_2O)_2$ trimers in the corresponding 2B and 3B training sets and the cluster size was fixed at 50 configurations. For GPR we used the class *GaussianProcessRegressor* and the *RBF* kernel available in the same Python package.

Both GPR and KMeans require a vector representation of the 2B and 3B structures in the high-dimensional configuration space. For this purpose, we used the many-body tensor representation (MBTR) of atomic environments.[251] MBTR defines a structural descriptor that is easily computable and well suited to calculate the kernels for both GPR and KMeans. The MBTR descriptor is constructed by storing the terms of the Coulomb matrix[205] associated with each pair of the N_e chemical elements constituting the molecular system of interest into an $N_e \times N_e \times d$ tensor, where d is the largest number of unique

pairs of the same two chemical elements. The MBTR descriptor thus takes the form

$$f_k(x, \boldsymbol{z}) = \sum_{\boldsymbol{i}}^{N_a} w_k(\boldsymbol{i}) \mathcal{D}(x, g_k(\boldsymbol{i})) \prod_{j=1}^{k} C_{z_j, Z_{i_j}}, \tag{2.6}$$

where $\boldsymbol{z} \in \mathbf{N}^k$ are atomic numbers, $\boldsymbol{i} = (i_1, \ldots, i_k) \in \{1, \ldots, N_a\}$ are index tuples, k runs over the number of atoms, $\mathcal{D}$ is a broadening function, C is the element correlation matrix, and g_k is a function that assigns a scalar to the k atoms based on a k-body physical feature. The MBTR descriptor is then discretized and rearranged in the form of a vector.

We used the Python package *qmmlpack* for the vector representation of the 2B and 3B configurations in their respective training sets. The broadening function $\mathcal{D}$ was chosen to be the normal distribution with $k = 2, 3$. The inverse of the distance, r^{-1}, and the angle, θ, were used as g_k for $k = 2, 3$, respectively. The number of bins and the width of the normal distribution were tuned to guarantee the efficiency of the MBTR calculations, without compromising accuracy.

2.3 Results

The results of our AL framework are presented in the following three subsections. First, we discuss the learning curves for the 2B and 3B energies, and comparisons are made between our AL framework and a generic approach based on a random selection (RS) of molecular configurations. Second, we introduce sketch-maps[252] of different 2B and 3B training sets generated through our AL framework and discuss the corresponding distributions of 2B and 3B energies. Third, we analyze the interaction and many-body energies of small water clusters as well as the Cs^+-oxygen radial distribution functions (RDFs) of liquid water calculated using different 2B and 3B training sets generated through our AL framework.

2.3.1 Learning curves of 2B and 3B energies

Figs. 2.2 and 2.3 show the learning curves for the 2B Cs^+–H_2O and 3B Cs^+–$(H_2O)_2$ energies, respectively, calculated for the training (left panels) and test (right panels) sets as a function of the training set size. Learning curves obtained using both our AL framework (blue) and RS approach (magenta) are shown.

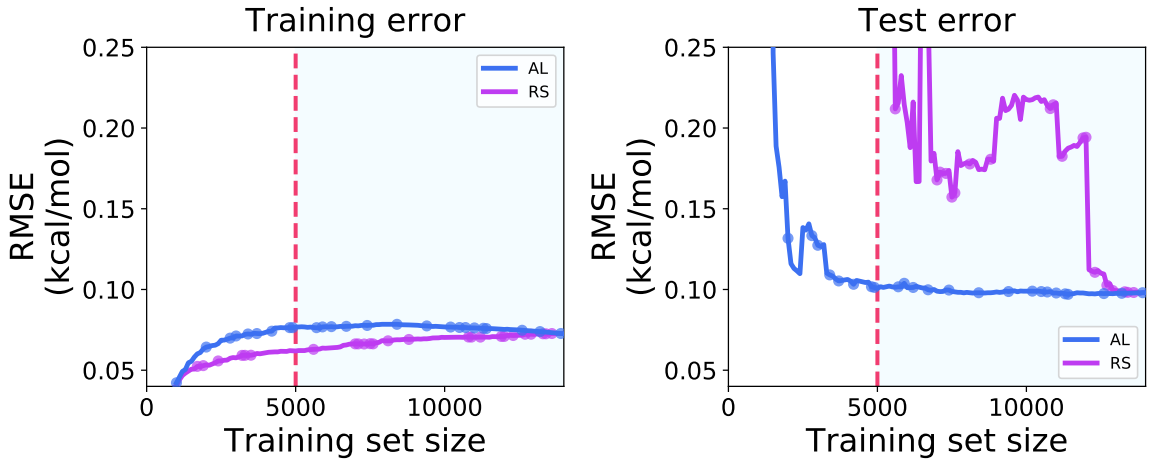

Figure 2.2. RMSEs (in kcal/mol) associated with the 2B training (left) and test (right) sets displayed as a function of the training set size. Blue and magenta curves correspond to AL and RS learning curves, respectively. The dashed line indicates the optimal training set size as determined in this study.

The training root-mean-square errors (RMSEs) associated with the RS approach increase monotonically as a function of the training set size for both 2B and 3B energies while the corresponding AL curves display steeper increases for smaller training sets, reach a maximum, and then decrease. The test RMSEs show different trends, with the curves obtained with our AL framework displaying a significantly faster decrease as a function of the training set size. Since our AL framework specifically targets configurations with higher uncertainties and neighborhood training errors, these configurations are selected more frequently by the learner and added to the training set. It follows that the configurations that are left in the pool after each iteration are associated with progressively smaller uncertainties and training errors. This implies that, when added to the training sets in

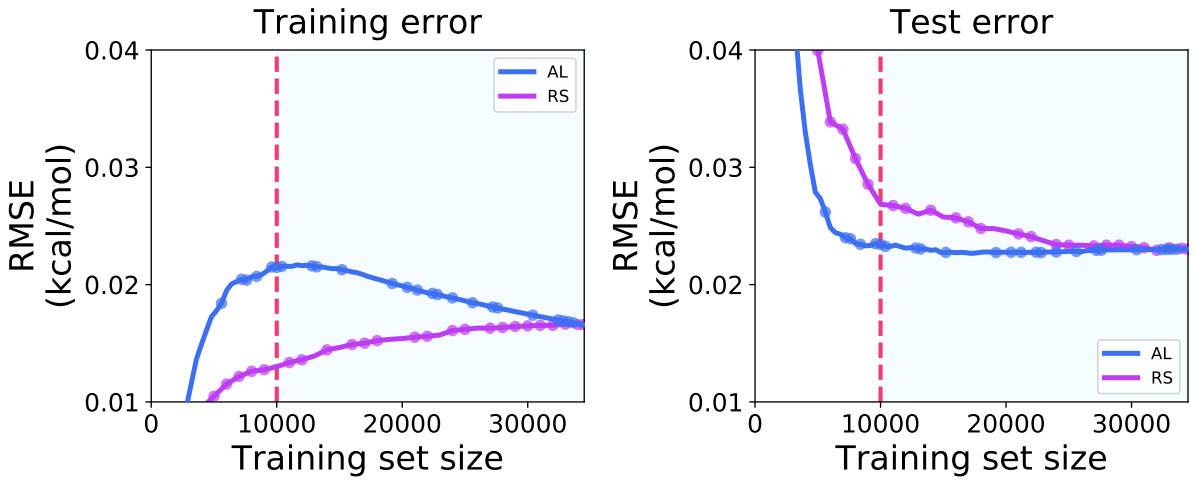

Figure 2.3. RMSEs (in kcal/mol) associated with the 3B training (left) and test (right) sets displayed as a function of the training set size. Blue and magenta curves correspond to AL and RS learning curves, respectively. The dashed line indicates the optimal training set size as determined in this study.

subsequent iterations, these configurations necessarily lead to a decrease of the training RMSEs and only negligible variations in the test RMSEs as shown in in Figs. 2.2 and 2.3.

As a general rule, the simultaneous stabilization or decrease of the training error and the stabilization of the test error are good indicators of the convergence of the learning process.[253,254] Therefore, based on the analysis of both training and test RMSEs obtained with our AL framework, cutoff values for the training set size could be chosen. The optimal numbers of configurations in the 2B and 3B training sets for the Cs^+–water MB-nrg PEFs were determined to be 5000 $Cs^+(H_2O)$ dimers and 10000 $Cs^+(H_2O)_2$ trimers, respectively.

2.3.2 Sketch-maps

Sketch-maps have been shown to be useful tools for representing high-dimensional configuration spaces with lower-dimensional projections that are easily interpretable in terms of well-defined structural features.[252,255,256]

To provide structural insights into the composition of the 2B and 3B training sets, with varying sizes, obtained with our AL framework, MBTR was used to generate

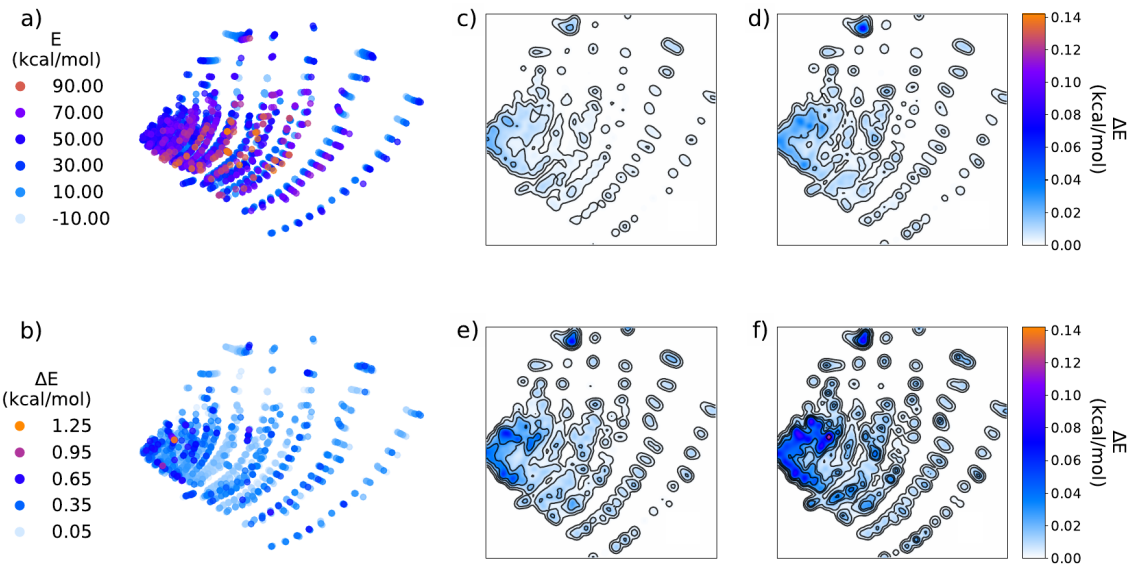

Figure 2.4. Sketch-maps of the 2B configurations. The map in in a) represents the reference CCSD(T) energies while the map in b) represents the difference, ΔE, between the reference CCSD(T) energies and the energies predicted by the MB-nrf PEF trained on the full pool of 2B configurations. The maps in c) to f) represent the difference, ΔE, between the energies predicted by the MB-nrg PEF trained on of the full training set and the corresponding values predicted by MB-nrg PEFs trained on the reduced-size training sets of 10000, 8000, 6000, and 4000 configurations generated using the AL framework, respectively.

the sketch-maps shown in Figs. 2.4 and 2.5, respectively. Panel a) of Fig. 2.4 is a representation of the entire 2B pool projected onto a 2-dimensional space. Each point on the map corresponds to a $Cs^+(H_2O)$ dimer configuration and the associated color indicates the corresponding CCSD(T) reference 2B energy. Since the 2B pool was generated on a grid by varying the Cs^+-O distance and distorting the water molecule, these features are reflected in the resulting sketch-map where points cluster together, in an orderly fashion.

Panel b) of Fig. 2.4 shows a sketch-map of the energy differences between the reference 2B energies and the corresponding values predicted by the MB-nrg PEF trained on the full 2B pool (13525 configurations). This comparison shows that the the MB-nrg PEF provides an accurate description of the overall 2B energy landscape, with deviations larger than 0.5 kcal/mol only found for $Cs^+(H_2O)$ dimers with associated binding energies

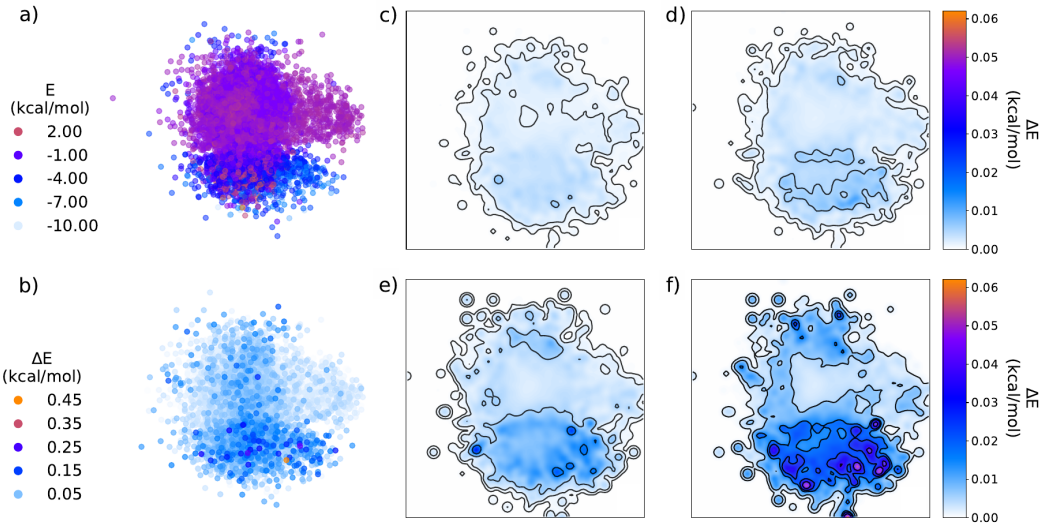

Figure 2.5. Sketch-maps of the 3B configurations. The map in in a) represents the reference CCSD(T) energies while the map in b) represents the difference, ΔE, between the reference CCSD(T) energies and the energies predicted by the MB-nrf PEF trained on the full pool of 2B configurations. The maps in c) to f) represent the difference, ΔE, between the energies predicted by the MB-nrg PEF trained on of the full training set and the corresponding values predicted by MB-nrg PEFs trained on the reduced-size training sets of 20000, 15000, 10000, and 5000 configurations generated using the AL framework, respectively.

larger than 40 kcal/mol, and deviations on the order of 0.04 kcal/mol for $Cs^+(H_2O)$ dimer configurations with lower binding energies (less than 40 kcal/mol). It should be noted that dimer configurations with larger binding energies are unlikely to be visited in MD simulations at ambient conditions and are included in the 2B training sets to guarantee that the PIPs representing short-range interactions within the the MB-nrg PEF are well-behaved at short Cs^+–water distances. Panels c-f) show sketch-maps of the differences between 2B energies predicted by the MB-nrg PEF trained on the full 2B pool and the corresponding values predicted by MB-nrg PEFs trained on progressively smaller training sets containing 10000, 8000, 6000, 4000 configurations generated using our AL framework. As expected, systematically reducing the training set size introduces progressively larger errors, with training sets with fewer than 4000 dimer configurations leading to overfitting. This analysis shows that our AL framework allows for significantly

reducing the original 2B Cs^+–H_2O training set without compromising the overall accuracy of the resulting MB-nrg PEF. In this context, it should be noted that the areas of the sketch-maps in panels c-f) that display larger deviations from the original MB-nrg PEF of Ref. 108, as the training set size decreases, correspond to dimer configurations for which the original MB-nrg PEF also shows larger deviations from the CCSD(T) reference data (panel b).

Similar conclusions can be drawn from the analysis of the sketch-maps of the 3B training sets shown in Fig. 2.5. Since the original 3B pool was generated by extracting $Cs^+(H_2O)_2$ trimers from MD simulations of a single Cs^+ ion in liquid water, the resulting sketch-map (panel a) displays a more uniform distribution in the 2-dimensional space compared to the corresponding sketch-map obtained for the 2B pool. Depending on the associated CCSD(T) reference 3B energies, trimer configurations broadly cluster in two areas, with the "dividing surface" being between -5.0 and -3.0 kcal/mol; this is highlighted by the sudden change in color in panel a). Also in this case, the original MB-nrg PEF closely reproduces the CCSD(T) reference 3B energies over the entire configuration space of the 3B pool, as shown in panel b). As for the 2B energies, progressively smaller training sets of 20000, 15000, 10000, 5000 configurations, generated using our AL framework and analyzed through the sketch-maps shown in panels c-f), lead to progressively larger deviations from the original MB-nrg PEF. It should be noted that, on average, the deviations remain smaller than 0.06 kcal/mol even for the smallest training set (5000 trimer configurations).

2.3.3 Clusters analysis

To assess the relative accuracy of the various training sets generated using our AL framework and determine how the associated differences in the representation of 2B and 3B energies affect the ability of the resulting MB-nrg PEFs to reproduce the properties of water from the gas to the condensed phase, deviations from the reference 2B and 3B energies of low-lying isomers of the $Cs^+(H_2O)_{n=1-3}$ clusters are analyzed in Fig. 2.6.

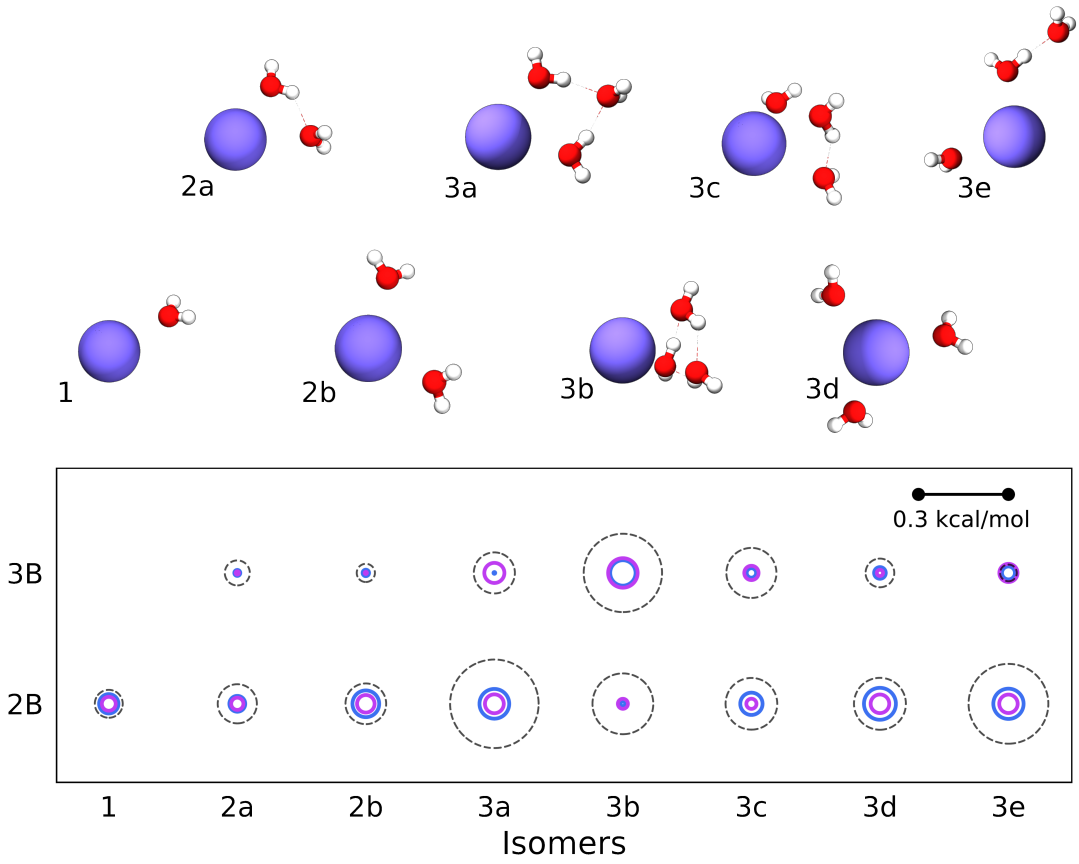

Figure 2.6. Schematic representation of the errors associated with the 2B and 3B energies of low-lying isomers of $Cs^+(H_2O)_{n=1-3}$ clusters. The dashed black circles represent the difference between the reference CCSD(T) energies and the corresponding values obtained with the MB-nrg PEF trained on the full 2B and 3B pools. The other solid circles represent the differences between the energies predicted by the MB-nrg PEF trained on the full 2B and 3B pools and the corresponding values predicted by MB-nrg PEFs trained on 4000 2B configurations and 5000 3B configurations, with blue and magenta corresponding to the AL and RS training sets, respectively.

This analysis is carried out for several MB-nrg PEFs generated using the minimal 2B and 3B training sets shown in Figs. 2.4 and 2.5 of 4000 dimer configurations and 5000 trimer configurations, respectively. Also shown for comparison are the deviations obtained with the same training sets generated from random selection. In all cases, the differences between the 2B and 3B energies predicted by the different MB-nrg PEFs are comparable for all clusters, and often smaller than the corresponding differences between the original

MB-nrg PEF fitted to the full 2B and 3B training sets and the CCSD(T) reference data. This analysis thus indicates that the reduction of the training set sizes does not affect the ability of the resulting MB-nrg PEFs to correctly represent 2B and 3B energies in small water clusters. It should be noted that this is true for both families of MB-nrg PEFs derived from training sets generated through AL and RS. This similarity can be attributed to the intrinsic low dimensionality of the $Cs^+(H_2O)$ dimers and $Cs^+(H_2O)_2$ trimers that make up the corresponding 2B and 3B training sets, which allowed for extensive sampling of the relevant configurations for the development of the original training sets in Refs. 108 and 235 . However, while no appreciable differences exist in the performance of the two sets of MB-nrg PEFs, AL clearly provides a more efficient approach to the selection of the training set sizes as demonstrated by the significantly higher variability associated with the learning curves obtained with the RS approach. The efficiency of the AL framework thus becomes particular important when, differently from the present case of the Cs^+–water MB-nrg PEF, no prior information on training sets is provided. This aspect of our AL framework will be the subject of a forthcoming study.

2.3.4 Radial distribution functions

To investigate the effects of training set reduction on modeling the properties of bulk solutions, the Cs^+–O RDFs calculated using different MB-nrg PEFs obtained from fits to different combinations of 2B and 3B training sets generated using AL (left panels) and RS (right panels) are analyzed in Figs. 2.7 and 2.8. The effects of the 2B training set is first assessed in Fig. 2.7 by analyzing the performance of five MB-nrg PEFs generated by fitting the 2B term to 2B training sets of various sizes (full, 10000, 8000, 6000, and 4000 dimer configurations) while fitting the 3B term to the full 3B training set for training the 3B term (34441 trimer configurations). The resulting RDFs calculated from MD simulations with the resulting MB-nrg PEFs generated from both AL and RS training sets are shown in the top left and right panels of Fig.2.7, respectively. As discussed in

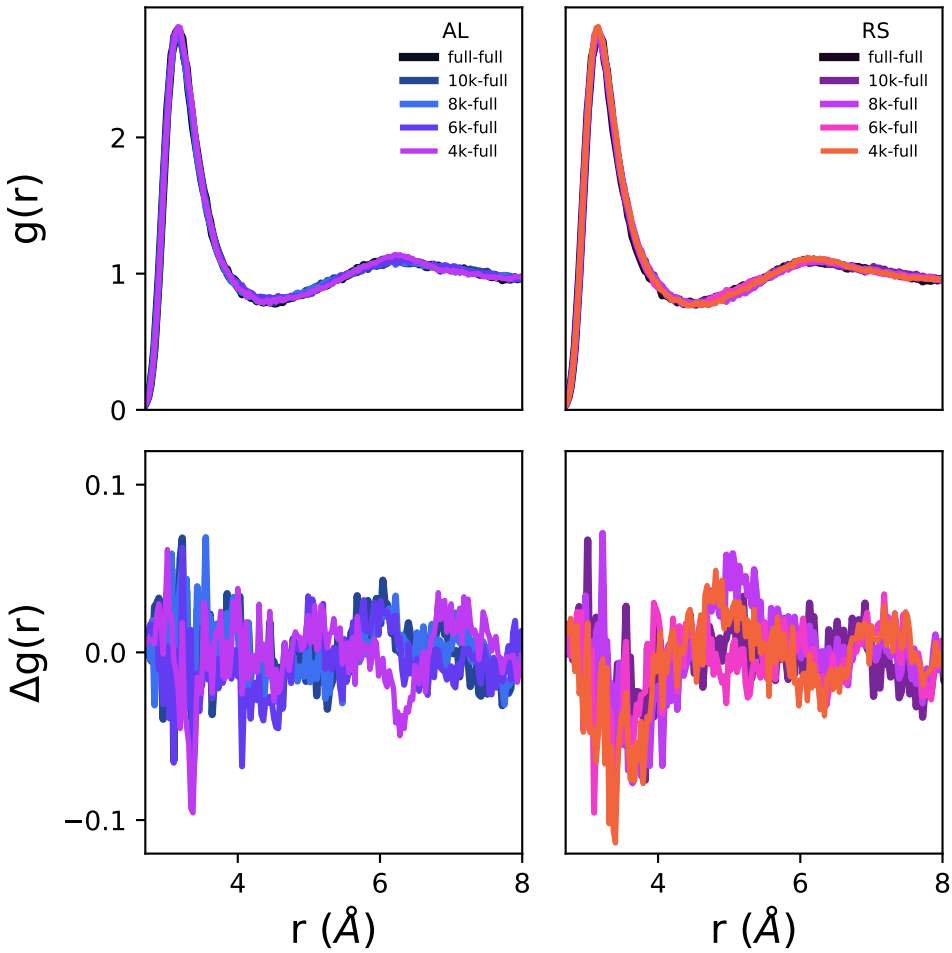

Figure 2.7. Top panels: Cs^+-O RDFs calculated from MD simulations with MB-nrg PEFs trained on progressively smaller 2B training sets (in the range of 10000-4000 dimer configurations) generated through AL (left) and RS (right), and the full 3B pool. Bottom panels: Differences between the RDF calculated with the MB-nrg PEF trained on the full 2B and 3B pool and the corresponding RDFs calculated with MB-nrg PEFs trained on the reduced-size AL (left) and RS (right) 2B training sets, and the full 3B pool.

more detail in Ref. 235, the Cs^+–O RDF displays a narrow peak, corresponding to the first hydration shell, at 3.16 Å, and a broader peak, corresponding to the second hydration shell, at ∼6 Å. No appreciable differences are found between the RDFs obtained using MB-nrg PEFs with progressively smaller 2B training sets. This is further evidenced by the curves shown in the bottom panels of Fig. 2.7 representing the differences between the RDFs calculated with each of the MB-nrg PEFs trained on reduced 2B training sets and

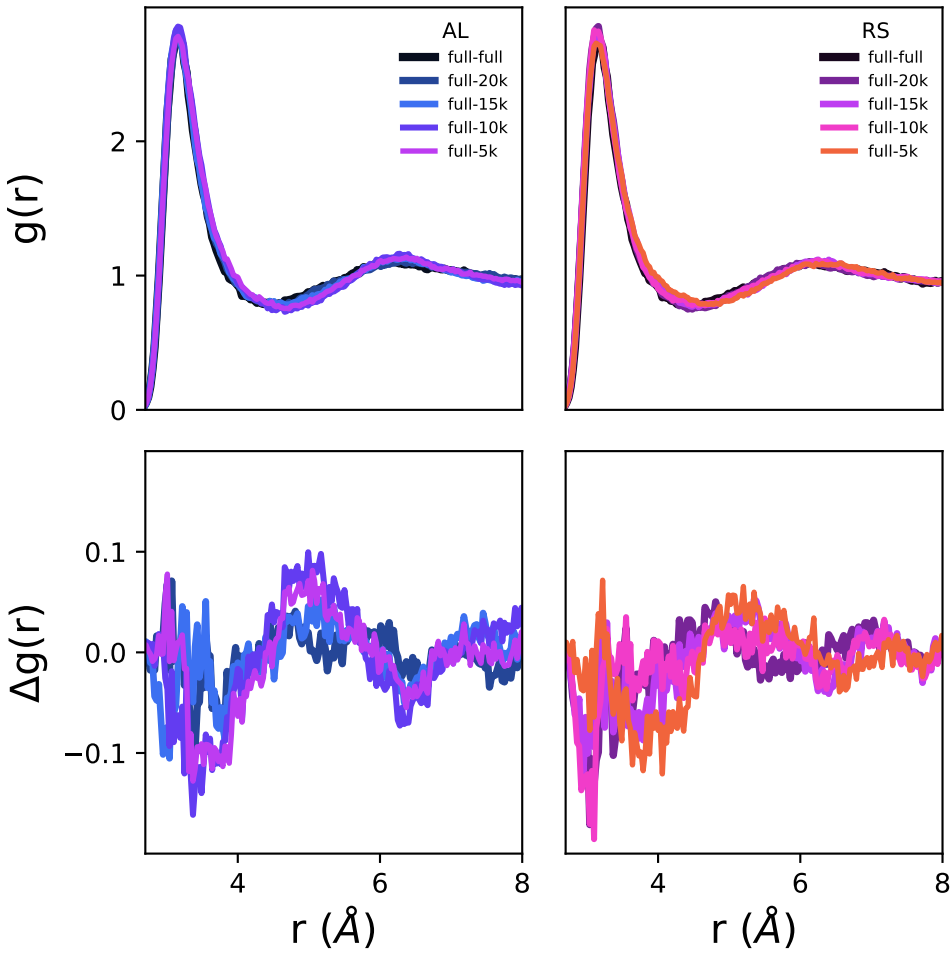

Figure 2.8. Top panels: Cs^+-O RDFs calculated from MD simulations with MB-nrg PEFs trained on progressively smaller 3B training sets (in the range of 20000-5000 trimer configurations) generated through AL (left) and RS (right), and the full 2B training set. Bottom panels: Differences between the RDF calculated with the MB-nrg PEF trained on the full 2B and 3B pool and the corresponding RDFs calculated with MB-nrg PEFs trained on the reduced-size AL (left) and RS (right) 3B training sets, and the full 2B pool.

the RDF calculated with the MB-nrg PEF trained on the full 2B training set.

Similarly, the effects of the 3B training set size reduction are investigated in Fig. 2.8 through the analysis of five MB-nrg PEFs generated by fitting the 3B term to 3B training sets of various sizes (full, 20000, 15000, 10000, and 5000 trimer configurations) while fitting to the 2B term to the full 2B training set. In this case, reducing the 3B training set size to less than 10000 trimer configurations results in small differences in the Cs^+–water RDF for

distances larger than 5.0 Å, which lead to a shift of the second hydration shell to slightly larger distances. However, as shown in the bottom panels of Fig. 2.8, these differences are barely noticeable and do not lead to any qualitative change in the hydration structure of Cs^+ in liquid water.

Overall, the analysis of both cluster and bulk properties demonstrates that the application of our AL framework to the original pools of 2B and 3B configurations of Refs. 162 and 235, respectively, leads to significantly smaller training sets, without loss of accuracy, which, in turn, largely reduces the cost associated with the development of CCSD(T)-level MB-nrg PEFs.

2.4 Conclusions

In this study, we introduced an AL framework for generating representative training sets needed for the development of MB-nrg PEFs.[107,108] Our AL framework is based on uncertainty and error estimation, and consists of a pool of an unknown number of molecular configurations, a predictor, and a learner that, based on feedback from the predictor, selects configurations from the pool and adds them to the training set. The selection process relies on Gaussian process regression and clustering of the configurations in the training set, which allows for efficiently identifying the most relevant configurations needed to accurately represent the target many-body PES.

The application of our AL framework to the development of a Cs^+–water MB-nrg PEF chosen as a case study led to significantly smaller training sets than those found necessary for the development of the original MB-nrg PEF. Analyses of the interaction and many-body energies calculated for small $Cs^+(H_2O)_n$ cluster as well as the Cs^+-oxygen RDF calculated from MD simulations of a single Cs^+ ion in water demonstrate that the new MB-nrg PEFs derived from the reduced-size training sets generated through AL display the same accuracy of the original MB-nrg PEF derived from the full 2B and 3B

pools.[108,235]

Given the computational cost associated with reference CCSD(T) calculations of individual many-body energies, our AL framework is particularly well-suited to the development of many-body PEFs, with chemical and spectroscopic accuracy, which can then be used in MD simulations of the target molecular system, from the gas to the condensed phase.

2.5 Acknowledgements

Chapter 2, in full, is a reprint of the material as it appears in "Active learning of many-body configuration space: Application to the Cs^+–water MB-nrg potential energy function as a case study" Y. Zhai, A. Caruso, S. Gao, and F. Paesani. In: *J. Chem. Phys.* 152.14 (2020), p. 144103. The dissertation author is the co-primary investigator and author of this paper.

We thank Andreas Götz for helpful discussions on the selection of optimal training sets, Matthias Rupp for guidance on using the MBTR descriptor, and Giulio Imbalzano and Michele Ceriotti for suggestions on generating of the 2B and 3B sketch-maps. This research was supported by the National Science Foundation through grant no. ACI-1642336. All calculations and simulations used resources of the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by the National Science Foundation through grant no. ACI-1053575, under allocation TG-CHE110009, and the Triton Shared Computing Cluster (TSCC) at the San Diego Supercomputer Center (SDSC).

Chapter 3

Data-Driven Many-Body Models Enable a Quantitative Description of Chloride Hydration from Clusters to Bulk

3.1 Introduction

A molecular-level characterization of the hydration properties of charged species from small clusters to bulk solutions and interfaces is key to understanding many physicochemical processes. Notably, hydrated ions are found as stabilizing species for biomolecules [3–5] and are involved in catalytic and transport processes. [1,2,257–261] It is hard to overstate the importance of ions in electrochemical processes, as charged species are a necessary component of electrolytic and galvanic cells. [12] Hydrated ions have also been shown to take part in the growth process of cloud condensation nuclei. [8–11]

While ions are ubiquitous in natural and industrial processes, a predictive understanding of the driving forces that determine the molecular properties of ions in aqueous solutions is still missing. For instance, it is known that, depending on their intrinsic electronic structure, ions can either strengthen or weaken the structure of the surrounding water hydrogen-bonding (H-bonding) network. To describe this effect, Hofmeister's original classification of ions according to their ability to modulate protein solubility [262]

subsequently led to the classification of ions as “structure makers" and “structure breakers".[13] Broadly speaking, “structure makers" are small and highly charged ions that are strongly hydrated and are, therefore, believed to strengthen the H-bonds between the surrounding water molecules. On the other hand, “structure breakers" are large and usually monovalent ions that, because of their size, disrupt the structure of the water H-bonding network. Although appealing due its simplicity, this classification has been challenged by measurements performed with various spectroscopic techniques.[14–18]

One of the most striking examples of the uncertainties in the current understanding of specific ion effects is perhaps represented by the ongoing debate around the distribution of ions at the air/water interface. Surface tension values larger than those for pure water were measured for salt solutions by Heydweiller more than one hundred years ago.[19] Importantly, while the surface tension was found to be independent of the nature of the cations, it varied significantly depending on the type of anion present in solution. Moreover, the effects of the anions on the surface tension were found to follow an inverse Hofmeister series. First Wagner,[20] and later Onsager and Samaras[21] proposed that image charges at the air/water interface are responsible for the local depletion of ions in the interfacial region which, in turn, leads to the observed variation of the surface tension in salt solutions. However, subsequent experiments found that the electrostatic potential difference across the air/water interface measured for solutions of halide salts (with the exception of fluoride salts) is more negative than that for pure water, thus suggesting that larger halide ions have higher propensity for the interface than cations, in contradiction with predictions derived from Wagner, and Onsager and Samaras theories.[22] Since then, various experimental approaches have been used to characterize the physical mechanisms that determine the distribution of different ions at the water surface, sometimes with conflicting results.[23–25]

Pioneering molecular dynamics (MD) simulations carried out for halide–water clusters using polarizable force fields (FFs) predicted that all halide ions, except fluoride,

are preferentially located at the surface of the clusters.[28–32] In contrast, MD simulations carried out for $Cl^-(H_2O)_n$ clusters with nonpolarizable FFs found that the chloride ion is always located in the interior of the clusters.[26,27] The different results obtained from MD simulations with polarizable and nonpolarizable FFs were interpreted as an indication of the importance of many-body effects in the hydration of halide ions.

Higher propensity for the water surface relative to the bulk was found for larger halide ions from MD simulations of concentrated salt solutions as well as from calculations of single-ion potentials of mean force (PMFs) carried out using polarizable FFs.[33,34] In particular, the surface propensity was found to increase from Cl^- to I^-, with F^- being repelled from the interface. Similar results were later obtained with various polarizable models.[35–37] More recent *ab initio* MD simulations based on density functional theory (DFT) found a much lower propensity of the iodide ion, for the air/water interface, compared to predictions obtained with polarizble FFs.[263] However, these MD-DFT simulations were carried out with the dispersion-corrected BLYP functional which has been shown to suffer from some intrinsic limitations when applied to the modeling of liquid water.[129,264,265] A lower surface propensity than that calculated for larger halide ions using MD simulations with polarizable FFs has also been predicted by an extended dielectric continuum (DC) theory which takes into account both the dimension and the polarizability of the ions.[38,39] An alternative model emphasizing the impact that ions may have on surface fluctuations has also been proposed to explain the experimental observations of selective ion adsorption at the air/water interface.[266]

Despite much recent effort in characterizing the molecular driving forces that contribute to modulating both structural and thermodynamic properties of ions in solution, it has become increasingly apparent that the development of a unified, molecular theory of ion hydration requires a quantitative description of the interplay between ion–water and water–water interactions, which has so far remained elusive. In this context, the last decade has witnessed the emergence of many-body potential energy functions (PEFs)

which, rigorously derived from the many-body expansion (MBE) of the energy, hold great promise for predictive MD simulations of aqueous systems, from small clusters in the gas phase to bulk solutions and interfaces.[40–42] In particular, building upon the MB-pol PEF for water,[97,98,106] we developed two families of many-body PEFs, the TTM-nrg[267,268] and MB-nrg[107,108] PEFs, which have been shown to accurately reproduce both structural and thermodynamics properties of $X^-(H_2O)_n$ (X = F, Cl, Br, I) and $M^+(H_2O)_n$ (M = Li, Na, K, Rb, Cs) clusters, including quantum-mechanical effects in H-bonding rearrangements and isomeric equilibria.[269–273]

In this study, we present a systematic analysis of many-body effects in the hydration properties of Cl^- through detailed comparisons of different models of chloride–water interactions, from simple point-charge FFs to polarizable FFs, and explicit many-body PEFs. For this purpose, we introduce an extended MB-nrg PEF that, building upon the results of Ref. 107, includes an explicit 3-body (3B) term as well as a refined 2-body term (2B) derived from an expanded training set of dimer configurations generated using a recently developed active learning scheme for many-body PEFs.[120]

3.2 Methods

The MBE expresses the energy, E_N, of a system containing N (atomic or molecular) monomers as the sum of individual n-body energies, ϵ^{nB}, where $n \leq N$,[219]

$$E_N(r_1,..,r_N) = \sum_{i=1}^{N} \epsilon^{\mathrm{1B}}(r_i) + \sum_{i<j}^{N} \epsilon^{\mathrm{2B}}(r_i,r_j) + \sum_{i<j<k}^{N} \epsilon^{\mathrm{3B}}(r_i,r_j,r_k) + ... + \epsilon^{\mathrm{NB}}(r_1,..,r_N) \quad (3.1)$$

Here, r_i collectively represents the coordinates of all atoms in monomer i, ϵ^{1B} represents the distortion energy of an isolated (molecular) monomer, and the n-body energies ϵ^{nB} are

defined as

$$\epsilon^{\mathrm{nB}} = E_n(r_1, ..., r_n) - \sum_{i=1}^{N} \epsilon^{\mathrm{1B}}(r_i) - \sum_{i<j}^{N} \epsilon^{\mathrm{2B}}(r_i, r_j) - \ldots - \sum_{i<j<k<\ldots}^{N} \epsilon^{(\mathrm{n\text{-}1})\mathrm{B}}(r_i, r_j, r_k, ...) \quad (3.2)$$

Since the MBE converges quickly for systems with localized electron densities and large band gaps, Eq. 3.1 can be used as a effective theoretical/computational framework for developing many-body PEFs where each individual term of the MBE is independently fitted to the corresponding electronic structure data.[110,274]

As discussed in detail in the original studies, both TTM-nrg[267,268] and MB-nrg[107,108] PEFs are derived from Eq. 3.1, and use the MB-pol PEF for representing water–water interactions.[97,98,106] MB-pol has been shown by us and others to correctly reproduce the properties of water,[110,158] from small clusters in the gas phase[159–161,220–227,229] to liquid water,[162,230–232,275–277] the air/water interface,[163–165,233,234], and ice.[166–169] In the TTM-nrg PEFs all ion–water many-body contributions, i.e., ϵ^{2B} to ϵ^{NB}, are described by an implicit NB term represented by classical polarization.[267,268] The MB-nrg PEFs instead approximate Eq. 3.1 with the sum of explicit low-order terms, generally up to the 3B term, along with the same implicit many-body term used by the TTM-nrg PEFs to represent all higher-body interactions,[107,108]

$$E_N = \sum_{i=1}^{N} \epsilon_i^{\mathrm{1B}} + \sum_{i>j}^{N} \epsilon_{i,j}^{\mathrm{2B}} + \sum_{i>j>k}^{N} \epsilon_{i,j,k}^{\mathrm{3B}} + V_{\mathrm{pol}} \quad (3.3)$$

Each term of Eq. 3.3 is fitted to reproduce the corresponding nB reference energies that, as discussed below, are calculated at the explicitly correlated coupled cluster level of theory including single, double, and perturbative triple excitations, i.e., CCSD(T)-F12b.[136,137] Since the theoretical/computational framework behind the MB-pol[97,98] and MB-nrg[107,108] PEFs is described in the original references, we will only discuss here specific details related to the development of the present chloride–water MB-nrg PEF.

3.2.1 2-body energies

Following Refs. 107 and 108, ϵ^{2B} in Eq. 3.3 is represented by three terms:

$$\epsilon^{2B} = V_{\rm sr}^{2B} + V_{\rm elec} + V_{\rm disp} \tag{3.4}$$

Here, $V_{\rm sr}^{2B}$ describes quantum-mechanical short-range 2B interactions (e.g., Pauli repulsion, and charge transfer and penetration) that arise from the overlap of the electron densities of the chloride ion and a water molecule, which cannot be represented in terms of classical expressions.[41,42] In the MB-nrg PEF, $V_{\rm sr}^{2B}$ is represented by a permutationally invariant polynomial, $V_{\rm PIP}^{2B}$, that is dampened to zero at long range by a switching function, $s_2(R_{\rm Cl^-O})$, of the distance $R_{\rm Cl^-O}$ between the chloride ion (Cl^-) and the oxygen atom (O) of the water molecule within a Cl^-–H_2O dimer,

$$V_{\rm sr}^{2B} = s_2\left(R_{\rm Cl^-O}\right) \cdot V_{\rm PIP}^{2B} \tag{3.5}$$

where

$$s_2(R_{\rm Cl^-O}) = \begin{cases} 1, & \text{if } t_2(R_{\rm Cl^-O}) < 0 \\ cos^2\left[t_2(R_{\rm Cl^-O})\pi/2\right], & \text{if } 0 \le t_2(R_{\rm Cl^-O}) < 1 \\ 0, & \text{if } 1 \le t_2(R_{\rm Cl^-O}) \end{cases} \tag{3.6}$$

and

$$t_2(R_{\rm Cl^-O}) = \frac{R_{\rm Cl^-O} - R_i^{2B}}{R_{\rm out}^{2B} - R_{\rm in}^{2B}} \tag{3.7}$$

Here, $R_{\rm in}^{2B} = 5.8$ Å and $R_o^{2B} = 7.8$ Å are the predefined inner and outer cutoff distances of the switching function. These cutoff distances, which differ slightly from those used in Ref. 107, were found to guarantee a smooth and continuous representation of ϵ^{2B}. In particular, both $R_{\rm in}^{2B}$ and the range of the switching function were increased compared to the original values[107] since the gain in stability provided by a more slowly varying

switching function was found to overcome the higher computational cost associated with a larger number of 2B interactions to compute. As in Ref. 107, $V_{\mathrm{PIP}}^{\mathrm{2B}}$ is a function of all pairwise distances among the physical atoms (H, O, and Cl^-) and the lone-pair sites of the MB-pol water molecule (L_1 and L_2)[97] within a Cl^-–H_2O dimer. $V_{\mathrm{PIP}}^{\mathrm{2B}}$ contains 496 symmetrized monomials (ξ_i): 3 first-degree monomials, 15 second-degree monomials, 49 third-degree monomials, 130 fourth-degree monomials, and 299 fifth degree monomials. By construction, $V_{\mathrm{PIP}}^{\mathrm{2B}}$ thus contains 496 linear fitting parameters (c_i) and 9 nonlinear fitting parameters.[107]

As in Ref. 107, V_{elec} in Eq. 3.4 represents permanent electrostatics between the negative (-1e) charge of the chloride ion and the MB-pol geometry-dependent point charges of the water molecule which reproduce the *ab initio* dipole moment of an isolated water molecule.[185]

The last term in Eq. 3.4, V_{disp}, describes the 2B dispersion energy:

$$V_{\mathrm{disp}} = -f(\delta_{\mathrm{Cl^-O}})\frac{C_{6,\mathrm{Cl^-O}}}{R^6_{\mathrm{Cl^-O}}} - f(\delta_{\mathrm{Cl^-H_1}})\frac{C_{6,\mathrm{Cl^-H_1}}}{R^6_{\mathrm{Cl^-H_1}}} - f(\delta_{\mathrm{Cl^-H_2}})\frac{C_{6,\mathrm{Cl^-H_2}}}{R^6_{\mathrm{Cl^-H_2}}} \tag{3.8}$$

where $R_{\mathrm{Cl^-O}}$, $R_{\mathrm{Cl^-H_1}}$, and $R_{\mathrm{Cl^-H_2}}$ are the distances between the Cl^- ion and the O, and the two H atoms of the water molecule within the dimer, and $f(\delta)$ and C_6 are the corresponding Tang-Toennies damping functions[278] and dispersion coefficients.

3.2.2 3-body energies

Building upon the same theoretical framework used in the development of MB-pol and the Cs^+–H_2O MB-nrg PEF, ϵ^{3B} in Eq. 3.3 is represented by a 3B short-range term that effectively takes into account 3B energy contributions of quantum-mechanical origin arising from the overlap of the electronic densities of the chloride ion and two water (*a*

and b) molecules at a time as well as short-range 3B dispersion energy contributions,

$$\epsilon^{3B} = [s_3(R_{\mathrm{Cl^-O}_a})s_3(R_{\mathrm{Cl^-O}_b}) + s_3(R_{\mathrm{Cl^-O}_a})s_3(R_{\mathrm{O}_a\mathrm{O}_b}) + s_3(R_{\mathrm{Cl^-O}_b})s_3(R_{\mathrm{O}_a\mathrm{O}_b})] \cdot V^{3B}_{\mathrm{PIP}} \quad (3.9)$$

Here, s_3 is a 3-body switching function given by

$$s_3(R_{kl}) = \begin{cases} 1, & \text{if } t_3(R_{kl}) < 0 \\ cos^2\big[t_3(R_{kl})\pi/2\big], & \text{if } 0 \leq t_3(R_{kl}) < 1 \\ 0, & \text{if } 1 \leq t_3(R_{kl}) \end{cases} \quad (3.10)$$

where

$$t_3(R_{kl}) = \frac{R_{kl} - R_i^{3B}}{R_{\mathrm{out}}^{3B} - R_{in}^{3B}} \quad (3.11)$$

In Eqs. 3.9 and 3.10, R_{kl} is the distance between any (k, l) pair of Cl^- and O atoms within a $Cl^-(H_2O)_2$ trimer, R_{in}^{3B} and R_{out}^{3B} are the inner and outer cutoff distances. As an optimal compromise between accuracy and computational cost, 3B effects are only included within the first hydration shell of the chloride ion which is achieved by setting R_{in}^{3B} and R_{out}^{3B} to 2.5 Åand 4.5 Å, respectively. Although the following analyses demonstrate that the 2B and 3B cutoff ranges adopted in the present study allow for the accurate representation of the hydration structure of a chloride ion both in gas-phase clusters and in solution, it should be noted that the MB-nrg framework gives the user complete freedom in the choice of the inner and outer cutoffs.

V^{3B}_{PIP} is a function of all 41 pairwise distances between the physical atoms (H, O, and Cl^-) and the lone-pair sites of the two water molecules (L_1 and L_2) within a $Cl^-(H_2O)_2$ trimer. V^{3B}_{PIP} contains 1575 symmetrized monomials, ξ_i: 39 second-degree monomials, 613 third-degree monomials, and 923 fourth-degree monomials. Therefore, V^{3B}_{PIP} contains 1575 linear fitting parameters and 13 nonlinear fitting parameters.

Specific details about V^{2B}_{PIP} and V^{3B}_{PIP}, along with the definition of all monomials are

given in the Supplementary Material.

3.2.3 Reference energies

The reference dimer configurations used in the parameterization of the 2B energy term, ϵ^{2B}, were selected using the active learning approach described in Ref. 120, starting from a pool of 150 000 configurations generated by sampling a spherical grid between 2–8 Å from the chloride ion as well as the normal modes of the Cl^-–H_2O dimer. 9059 and 854 dimer configurations were used in the training and test sets, respectively. The reference 2B energies were calculated at the CCSD(T)-F12b level of theory[136,137] in the complete basis set (CBS) limit that was achieved via a two-point extrapolation.[138,139] The CCSD(T)-F12b calculations were performed with the augmented correlation-consistent polarized valence triple-/quadruple-ζ (aug-cc-pV[T/Q]Z) basis sets.[238,239,279,280]

A Cl^-–H_2O MB-nrg PEF, without the explicit 3B term, ϵ^{3B}, was initially developed and used in MD simulations with a single Cl^- ion in water which were carried out at ambient conditions in the isobaric-isothermal (NPT) ensemble to generate the 3B pool. 13140 and 1240 $Cl^-(H_2O)_2$ trimer configurations were extracted from the MD trajectories and included in the training and test sets, respectively. The 3B energies were calculated at the CCSD(T)-F12b level of theory[136,137] using the aug-cc-pVTZ basis set.[238,239,279,280]

All CCSD(T)-F12b electronic structure calculations were carried out using MOLPRO (version 2020.1).[281]

3.2.4 Fitting procedure

We followed the same fitting procedure used in the development of MB-pol[97,98] and other MB-nrg PEFs.[107–109,143,144] Specifically, the linear parameters were optimized through singular value decomposition, while the non-linear parameters were optimized using the simplex algorithm. The following regularized weighted sum of squared deviations

was calculated and minimized:

$$\chi^2 = \sum_{n \in \mathcal{S}} w_n [\epsilon_{\text{model}}(n) - \epsilon_{\text{ref}}(n)]^2 + \Gamma^2 \sum_{l=1}^{L} c_l^2 \tag{3.12}$$

Here, $\mathcal{S}$ represents the training set and L is the number of linear parameters. The weights w_n are introduced in Eq. 3.12 to emphasize configurations with lower binding energies according to[282]

$$w(E_i) = \left[\frac{\Delta E}{E_i - E_{\text{min}} + \Delta E} \right]^2 \tag{3.13}$$

Here, E_{min} is the lowest binding energy in the training set and ΔE is a parameter that defines the range of favorably weighted energies. $\Delta E = 35$ kcal/mol and $\Delta E = 47.5$ kcal/mol were used for the 2B and 3B energies, respectively. The regularization parameter Γ was set to 0.0005 to reduce the variation of the linear parameters while preserving the overall accuracy.

3.2.5 MD simulations and analysis

All MD simulations were carried out in the NPT ensemble for a box containing a single chloride ion and 277 water molecules at 298.15 K and 1.0 atm, corresponding to a ∼0.2 M solution. The velocity-Verlet algorithm[283] was used to propagate the equations of motion in the MD simulations with the TTM-nrg and MB-nrg PEFs. A timestep δt of 0.2 fs was used, which guarantees a correct sampling of the molecular degrees of freedom as well as the TTM-nrg and MB-nrg induced dipole moments that were propagated according to the always stable predictor–corrector algorithm.[284] Nosé-Hoover chains with 4 thermostats attached to each degree of freedom were used to control the temperature while the pressure was controlled using the algorithm described in Ref. 285. The path-integral molecular dynamics (PIMD) simulations with the MB-nrg PEF were carried out using the normal-mode representation, with each atom being described by a ring-polymer with 32 beads.[286] All MD simulations with the TTM-nrg and MB-nrg PEFs were carried out with

an in-house version of DL_POLY 2.0[245].

For comparison, MD simulations were also carried out using the Cl^-–H_2O empirical parameterization compatible with the TIP4P/Ew water model[61] which was introduced in Ref. 66 These simulations were carried out with AMBER[287] using a global Langevin thermostat and a Monte Carlo barostat to control the temperature and the pressure, respectively.

The FEFF software was used to calculate the EXAFS signals.[288–290] Following Ref. 235, all FEFF calculations were performed using clusters containing the chloride ion and its 33 closest water molecules which were extracted from the corresponding MD and PIMD trajectories at intervals of 0.5 ps.

3.3 Results

Fig. 3.1 shows the correlation plots between the reference CCSD(T)-F12b/CBS 2B energies and the corresponding MB-nrg values for both training (panel a) and test (panel b) sets. Root-mean-square errors (RMSEs) of 0.2387 kcal/mol and 0.2027 kcal/mol for the training and test sets, respectively, demonstrate that the present MB-nrg PEF is able to describe the Cl^-–H_2O 2B energies with coupled cluster accuracy over a wide range of values, without overfitting.

Fig. 3.1c shows one-dimensional potential energy radial scans for different (θ, ϕ) orientations of Cl^- relative to H_2O (see Fig. 3.1d for the definition of the coordinate system). Independently of the relative orientation, the MB-nrg PEF quantitatively reproduces the CCSD(T)/CBS values at all Cl^-–H_2O separations, which provides further evidence for the overall high accuracy of the present MB-nrg PEF at the 2B level.

After assessing the accuracy of the 2B term of the MB-nrg PEF, Fig. 3.2 shows the correlation plots between the CCSD(T)-F12b 3B reference energies and the corresponding MB-nrg values for both training (panel a) and test (panel b) sets. Also in this case, the

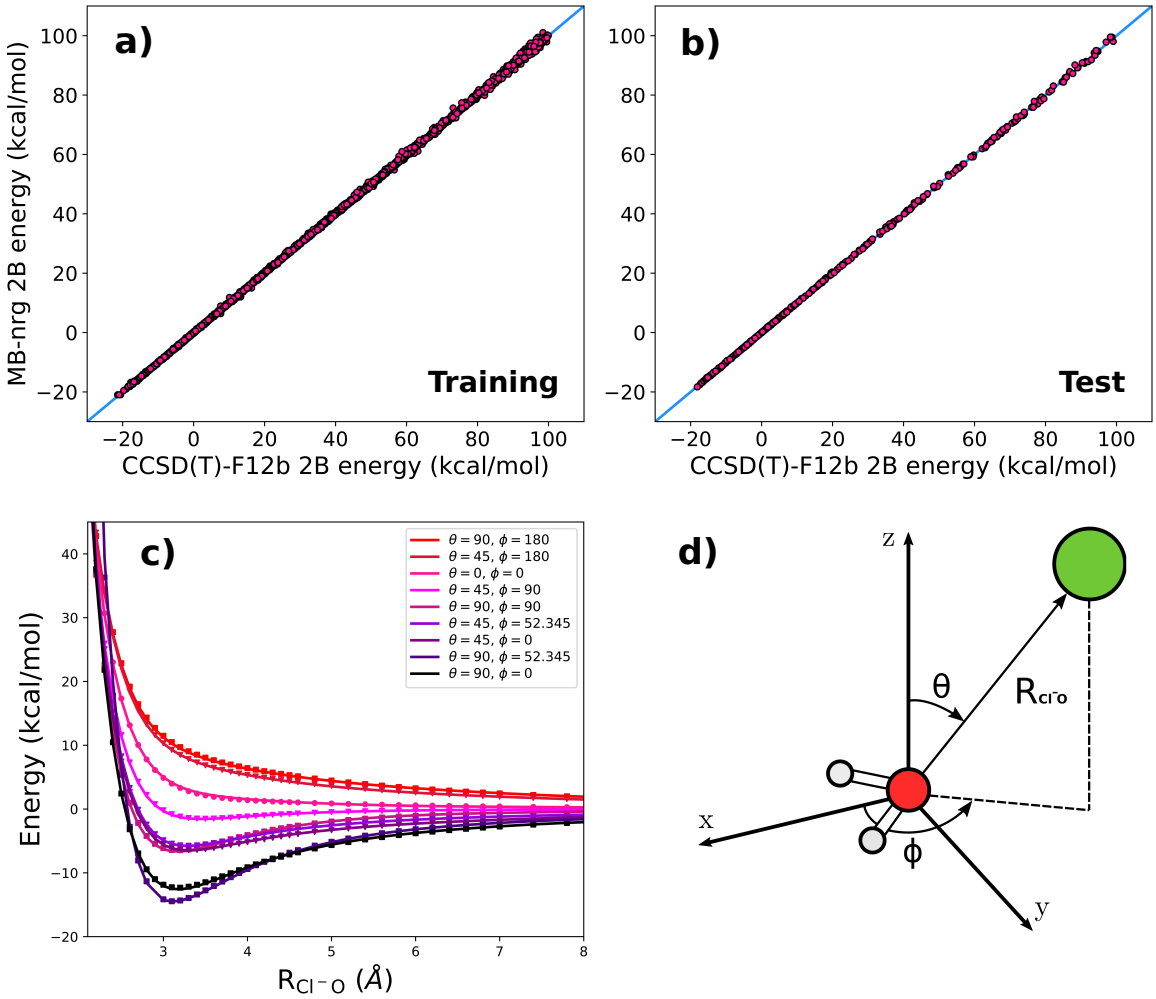

Figure 3.1. Top panels: 2B energy correlation plots between the CCSD(T)-F12b/CBS reference values (x-axis) and corresponding MB-nrg values (y-axis) for the training (a) and test (b) sets. Bottom panels: Cl^-–H_2O potential energy scans (c) along the Cl^-–O distance (R_{Cl^-O}) for different orientations θ and ϕ (d). The symbols corresponds to the CCSD(T)-F12b/CBS reference energies, while the corresponding MB-nrg values are shown as solid lines.

present Cl^-–H_2O MB-nrg PEF is able to quantitatively reproduce the CCSD(T)-F12b values over the a wide range of 3B energies, with RMSEs of 0.0655 kcal/mol and 0.0506 kcal/mol for the training and test sets, respectively. As for the 2B energies, this analysis indicates that the high accuracy achieved by the MB-nrg PEF does not result from overfitting. MB-nrg 2B and 3B energy error plots are reported in the Supplementary Material.

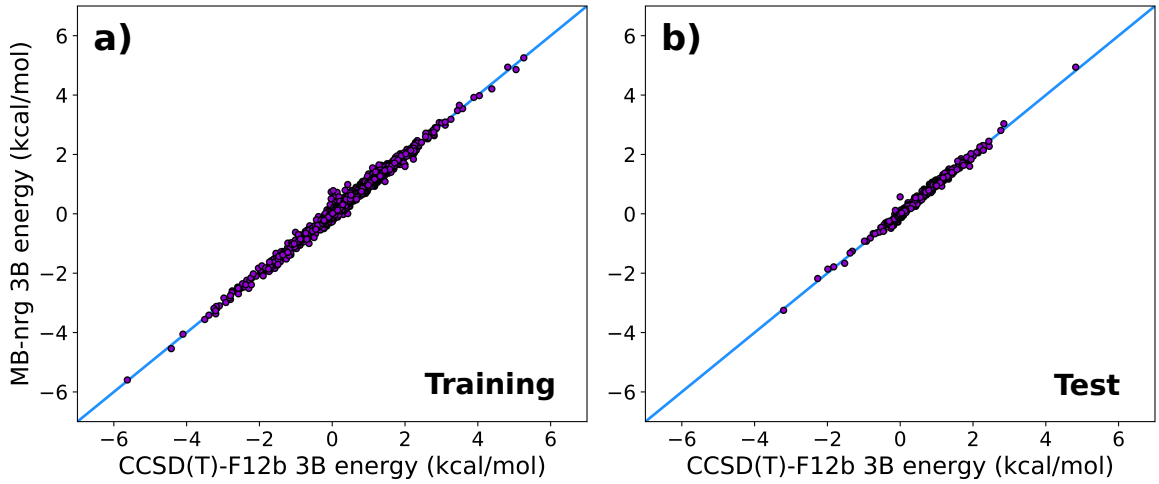

Figure 3.2. 3B energy correlation plots between the CCSD(T)-F12b reference values (x-axis) and corresponding MB-nrg values (y-axis) for the training (a) and test (b) sets.

Fig. 3.3 shows a comparison between the interaction energies calculated for the low-energy isomers of the $Cl^-(H_2O)_n$ clusters ($n = 1 - 4$) using the empirical TIP4P/Ew-based model of Ref. 66, the TTM-nrg PEF of Ref. 267, and the present MB-nrg PEF. To investigate the role played by short-range many-body effects, in this and following analyses, we consider two versions of the MB-nrg PEF: the (2B+NB)-MB-nrg PEF that includes the explicit 2B term of Eq. 3.4 in addition to the classical NB polarization term of Eq. 3.3, and the (2B+3B+NB)-MB-nrg PEF that includes both explicit 2B (Eq. 3.4) and 3B (Eq. 3.9) terms in addition to the classical NB polarization term of Eq. 3.3. The results obtained with these four different representations of the Cl^-–H_2O interactions are compared with the the corresponding CCSD(T)-F12b reference values.[40,270] As expected, being an empirical pairwise additive model developed for bulk simulations, the TIP4P/Ew-based model is unable to correctly reproduce the energetics of $Cl^-(H_2O)_n$ clusters, independently of the size and H-bonding arrangements. By including NB effects through a classical polarization term, the TTM-nrg PEF clearly provides a more accurate representations of all clusters, which somewhat deteriorates for structures with more cooperative H-bonding arrangements, as occurs in the isomer 4f of $Cl^-(H_2O)_4$. The agreement with the CCSD(T)-F12b results effectively becomes quantitative with the inclusion of the explicit 2B term

in the (2B+NB)-MB-nrg PEF, with the exception of the isomer 4f of $Cl^-(H_2O)_4$, for which only the (2B+3B+NB)-MB-nrg PEF is able to accurately reproduce the reference interaction energy. While emphasizing the role of many-body effects in chloride–water interactions, this analysis also highlights the limitations of a representation of many-body effects which is entirely based on classical polarization, as in the TTM-nrg PEF and other common polarizable models, and further demonstrates the importance of correctly describing short-range low-order (i.e., 2B and 3B) interactions.[40] In this context, it should be noted that the analyses reported in Ref. 40 demonstrate that all nB interactions with $n > 3$ in $Cl^-(H_2O)_n$ clusters are correctly described by the classical polarization term, V_{pol} employed by the TTM-nrg and MB-nrg PEFs.

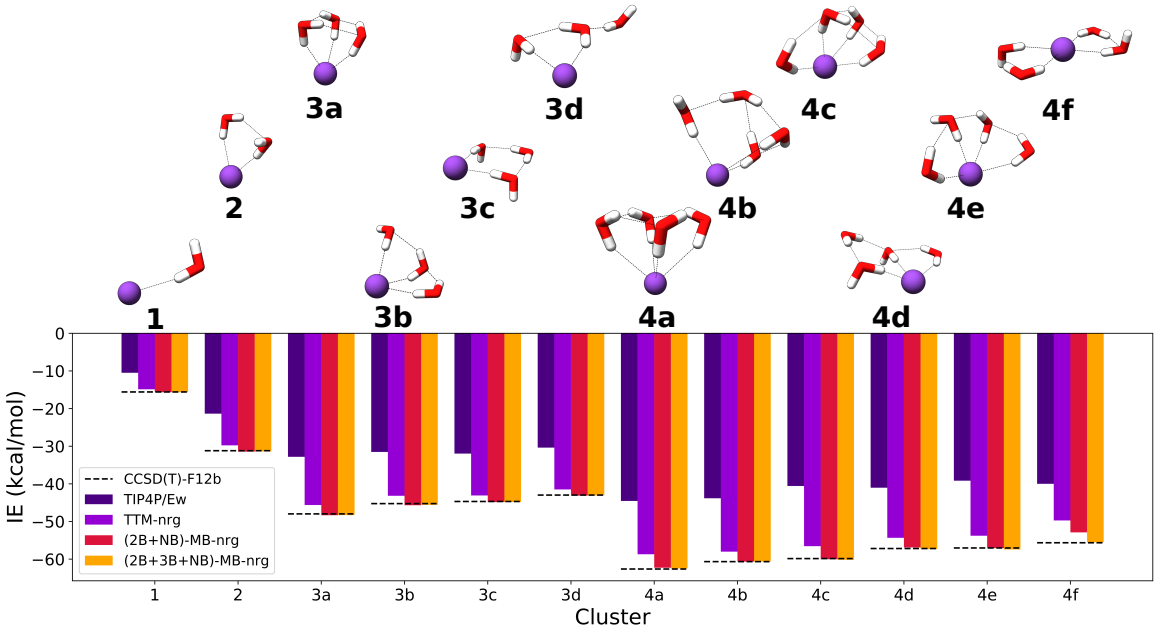

Figure 3.3. Comparison between the interaction energies calculated for the low-energy isomers of $Cl^-(H_2O)_n$ clusters (with $n = 1 - 4$) using the empirical TIP4P/Ew-based model of Ref. 66, the TTM-nrg PEF of Ref. 267, and the present (2B+NB)-MB-nrg and (2B+3B+NB)-MB-nrg PEFs. For each cluster, the CCSD(T)-F12b reference values[40,270] are shown as horizontal dashed lines.

Although the analyses presented in Figs. 3.1-3.3 provide a quantitative assessment of the ability of the present MB-nrg PEF to reproduce, at the fundamental level, many-body

interactions in $Cl^-(H_2O)_n$ clusters for which calculations at the coupled cluster level of theory are feasible, the MB-nrg PEFs also enable computer simulations of condensed-phase systems for which CCSD(T)-level calculations are currently out of reach. In absence of high-quality *ab initio* reference data for bulk simulations, the following analysis uses available EXAFS data to assess the reliability of the MB-nrg PEF in predicting the hydration structure of Cl^-. As for the analysis in Fig. 3.3, comparisons are made with results obtained from MD simulations carried out with the TIP4P/Ew-based model, the TTM-nrg PEF, and the two (2B+NB)-MB-nrg and (2B+3B+NB)-MB-nrg PEFs. In addition, results from PIMD simulations carried out with the (2B+3B+NB)-MB-nrg PEF are used to quantify the role played by nuclear quantum effects in determining the local hydration structure of Cl^-.

Fig. 3.4a shows the Cl^-–O radial distribution functions (RDFs) calculated from the five sets of NPT simulations that were carried out in this study. Within 8 Å from the Cl^- ion, the TIP4P/Ew-based model predicts highly structured hydration shells located at 3.15 Å, 4.80 Å, and 7.15 Å. The inclusion of many-body effects through classical polarization in the TTM-nrg PEF leads to significant disruption of the first hydration shell compared to the TIP4P/Ew RDF, which is accompanied by a shift of the second hydration shell to larger distances (5.13 Å). This effect becomes more pronounced with the inclusion of explicit 2B and 3B contributions as shown by the RDFs calculated with the (2B+NB)-MB-nrg and (2B+3B+NB)-MB-nrg PEFs, respectively. In particular, the (2B+3B+NB)-MB-nrg PEF predicts more diffuse first and second hydration shells which indicates a significantly less structured spatial arrangement of the water molecules around the chloride ion compared to that predicted by the TIP4P/Ew-based model. Similar trends are found in the Cl^-–H RDFs shown in Fig. 3.4b, with the TIP4P/Ew-based model and (2B+3B+NB)-MB-nrg PEF predicting the most and least structured first Cl^-–H shells, respectively. Importantly, while the TIP4P/Ew-based model predicts a well-defined and more compact second Cl^-–H shell along with a still distinguishable third Cl^-–H shell, all

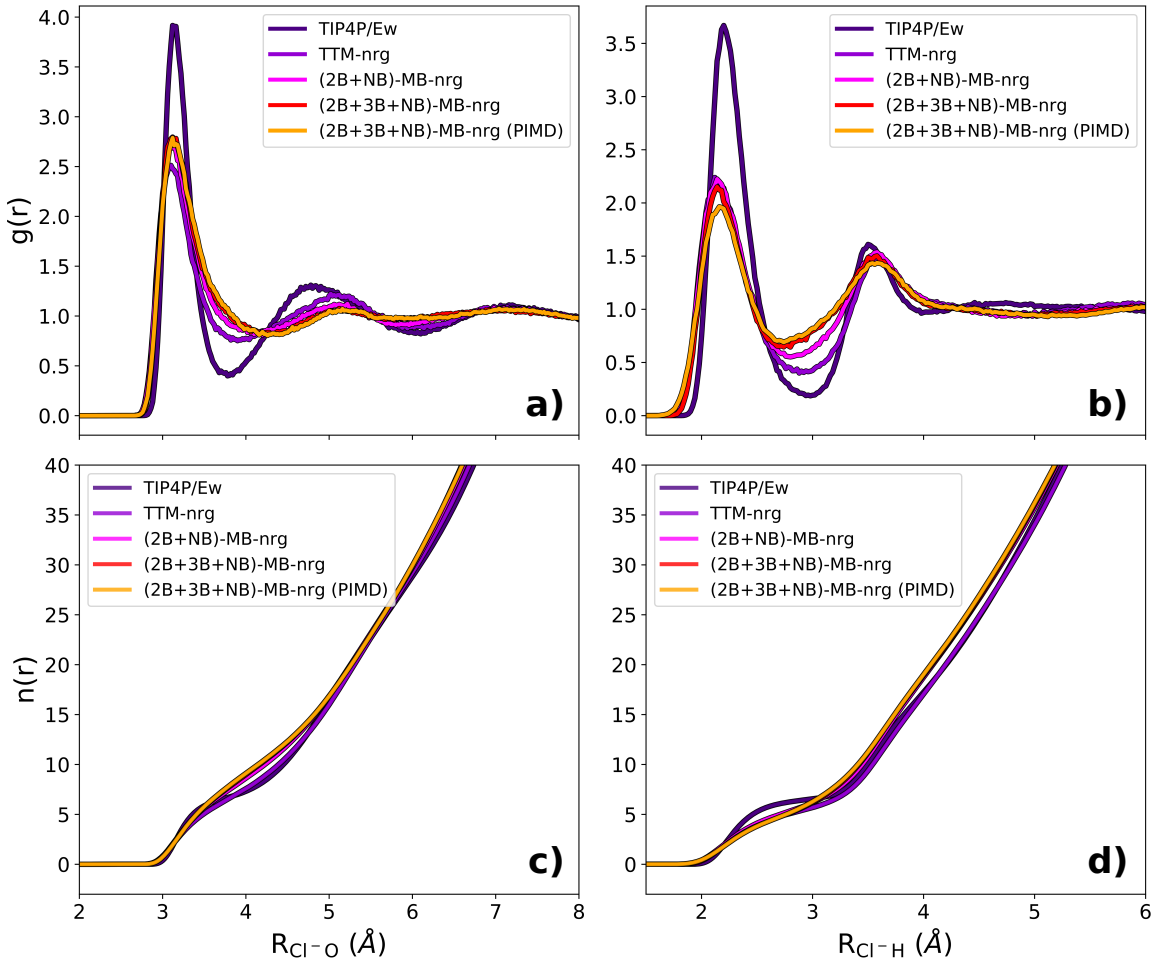

Figure 3.4. Top panels: Cl-O (a) and Cl-H (b) radial distribution functions (RDFs) calculated from MD simulations carried out with the TIP4P/Ew-based model, and TTM-nrg, (2B+NB)-MB-nrg, and (2B+3B+NB)-MB-nrg PEFs as well as from PIMD simulations with the (2B+3B+NB)-MB-nrg PEF. Bottom panels: Corresponding Cl-O (c) and Cl-H (d) cumulative distribution functions (CDFs).

many-body PEFs predict more diffuse distributions of the water hydrogen atoms of the water molecules that more closely hydrate the Cl^- ion. It should be noted that MD and PIMD simulations carried out with the (2B+3B+NB)-MB-nrg PEF effectively provide the same progression of hydration shells, with minor differences only visible in the first peak of the Cl^--H RDF, which suggests that nuclear quantum effects play a negligible role in determining the hydration structure of Cl^-.

Figs. 3.4c-d show the corresponding Cl^-–O and Cl^-–H cumulative distribution

functions (CDFs). The structural features highlighted in the analysis of the Cl^-–O RDFs clearly have a direct effect on the coordination number. In particular, due to the underlying more ordered hydration structure, the TIP4P/Ew-based model predicts that 7 water molecules constitute the first shell. A similar evolution of the Cl^-–O CDF is obtained with the TTM-nrg PEF model although, due to a more diffuse arrangement of the water molecules, it is difficult to precisely determine the coordination number within the first hydration shell of Cl^-, with $6 < n_{1st} < 8$. As it could be inferred from the analyses of the RDFs, a broader first hydration shell, containing ~ 11 water molecules, is predicted by both MB-nrg PEFs.

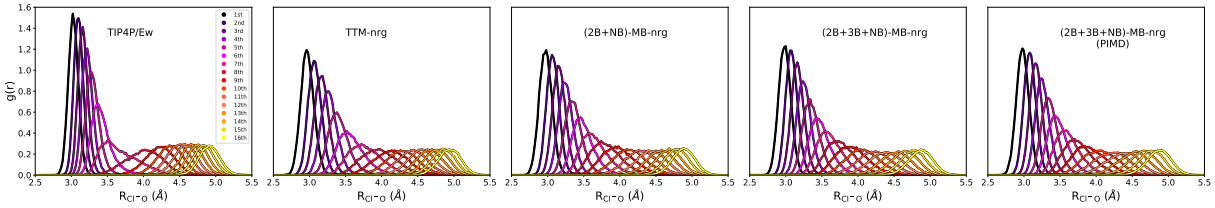

Figure 3.5. Incremental radial distribution functions (iRDFs) calculated from MD simulations with the TIP4P/Ew-based model, and TTM-nrg, (2B+NB)-MB-nrg, and (2B+3B+NB)-MB-nrg PEFs as well as from PIMD simulations with the (2B+3B+NB)-MB-nrg PEF.

To further explore the structural properties of the hydrated chloride ion, we calculated the incremental radial distribution functions (iRDFs), which describe individual contributions to the total Cl^-–O RDF associated with each water molecule i as a function of its distance ($R_{Cl^-O_i}$) Cl^-, and the radial-angular distribution functions (RADFs) in the first hydration shell region. The iRDFs shown in Fig. 3.5 indicate that the TIP4P/Ew-based model predicts a clear separation between the first and second hydration shells which is located between the seventh and eighth water molecule. As indicated by the CDFs in Fig. 3.4c, the separation between first and second hydration shells becomes increasingly less distinguishable and shifts to larger distances as many-body contributions to the Cl^-–H_2O interactions are progressively included, from the TTM-nrg to the (2B+3B+NB)-MB-nrg PEFs. It should also be noted that the TIP4P/Ew-based model predicts significantly

narrower iRDFs for water molecules located in the first hydration shell of Cl^- ($i = 1 - 6$) as well as much broader distributions for the 7th and 8th water molecules located in the transition region between the first and second hydration shells compared to those obtained with the many-body PEFs. This analysis thus provides further evidence for the TTM-nrg and MB-nrg PEFs predicting less tightly bound water molecules in the first hydration shell of Cl^- than the TIP4P/Ew-based model.

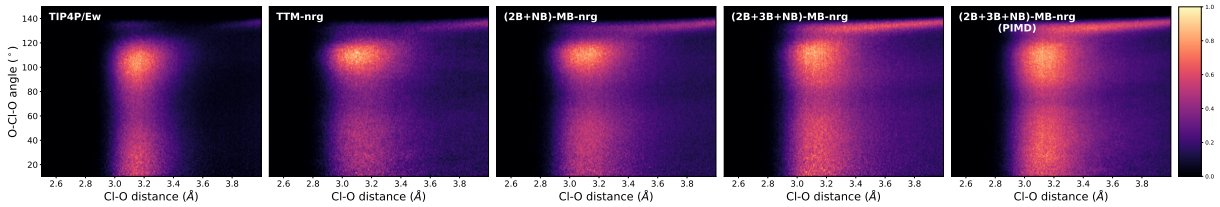

Figure 3.6. Radial-angular distribution functions (RADFs) of the first hydration shell calculated from MD simulations with the TIP4P/Ew-based model, and TTM-nrg, (2B+NB)-MB-nrg, and (2B+3B+NB)MB-nrg PEFs as well as from PIMD simulations with the (2B+3B+NB)-MB-nrg PEF. The Cl-O distance (in Å) is shown on the x-axis, and the O-Cl-O angle (in degrees) is shown in the y-axis.

The analysis of the RADFs shown Fig. 3.6 provides direct insights into the average distribution of water molecules within the first hydration shell of Cl^-. While predicting a tighter first hydration shell along the Cl^--O distance, as already inferred from the analyses of both RDFs and iRDFs, the TIP4P/Ew-based model is also characterized by a slightly broader distribution along the angular coordinate than both TTM-nrg and MB-nrg PEFs, which results in a relatively higher intensity between 55^o and 75^o. Particularly evident is the lack of the feature at 130° that becomes more pronounced as many-body contributions are progressively included in the description of the Cl^-–H_2O interactions. Due to the angular and radial diffuseness of the first hydration shell, it is difficult to extract contributions from individual molecules to the RADFs, and the average distribution of oxygen atoms around the chloride ion cannot be easily inferred. However, the different features exhibited by the different models show that short-range low-order interactions directly influence the geometry of the hydration complex. It should be noted that the RADF calculated from

PIMD simulations with the (2B+3B+NB)-MB-nrg PEF is effectively indistinguishable from that obtained from the corresponding MD simulations, which provides further support for nuclear quantum effects playing a negligible role in determining the hydration structure of Cl^-.

To determine which representation of the Cl^-–H_2O interactions provides the most realistic description of the hydration structure of Cl^- in solution, Fig. 3.7 shows comparisons of the K-edge EXAFS spectra calculated with the TIP4P/Ew-based model, and TTM-nrg and MB-nrg PEFs with the corresponding experimental data from Ref. 291. This analysis shows that the TIP4P/Ew-based model is unable to correctly reproduce the amplitude of the EXAFS spectra and slightly underestimates the period of the oscillations, with the calculated peaks appearing at relatively smaller k values. On the other hand, the agreement with the experimental data systematically improves as many-body effects are progressively included, with the MD and PIMD simulations carried out with the (2B+3B+NB)-MB-nrg PEFs providing nearly quantitative agreement with the experimental EXAFS spectrum. The comparisons shown in Fig. 3.5 also indicate that 3B interactions have minimal impact on the simulated EXAFS spectra, while the inclusion of nuclear quantum effects in the PIMD simulations improves the agreement with the experimental data at larger k but worsens it for k values between 2 and 4 $Å^{-1}$.

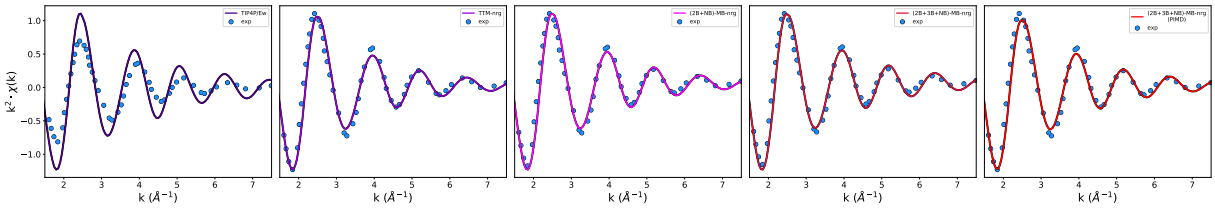

Figure 3.7. K-edge EXAFS spectra, $k^2\chi(k)$, calculated from MD simulations with the TIP4P/Ew-based model, and TTM-nrg, (2B+NB)-MB-nrg, and (2B+3B+NB)-MB-nrg PEFs as well as from PIMD simulations with the (2B+3B+NB)-MB-nrg PEF. The experimental data from Ref. 291 are shown as blue circles.

3.4 Conclusions

In this study, we have introduced a new many-body MB-nrg PEF describing chloride–water interactions which includes explicit 2B and 3B terms derived from CCSD(T)-F12b data along with an implicit NB term based on classical polarization. Although the new MB-nrg PEF is trained only on $Cl^-(H_2O)$ dimer and $Cl^-(H_2O)_2$ trimer configurations, it is able to quantitatively reproduce the interaction energies of small $Cl^-(H_2O)_n$ clusters, with $n = 1 - 4$. Importantly, when used in MD and PIMD simulations of a single Cl^- ion in water, we have demonstrated that the new MB-nrg PEF enables the calculation of EXAFS spectra that are in close agreement with the available experimental data, correctly reproducing both the amplitude and phase of the EXAFS oscillations.

Comparisons with the results obtained with a popular empirical force field[66] based on the TIP4P/Ew model of water[61] show that pairwise additive representations of chloride–water and water–water interactions are inadequate for representing chloride hydration structure in both gas-phase clusters and solution, underestimating the strength of the interactions in the first case while predicting an overly tight first hydration shell in the second case. On the other hand, comparisons with the results obtained with the TTM-nrg PEF[267] emphasize the importance of many-body effects in determining the hydration structure of Cl^- but, at the same time, highlight the limitations associated with a representation of these effects entirely based on classical many-body polarization.

We believe that the analyses presented here, along with results reported in previous studies,[107,108,235,269–273,292] provide further evidence that, as the MB-pol PEF has enabled an accurate description of the properties of water across different phases,[110] the MB-nrg PEFs for ion–water interactions can enable more realistic simulations of ionic aqueous systems from gas-phase clusters to bulk solutions and interfaces.

3.5 Acknowledgements

Chapter 3, in full, is a reprint of the material as it appears in "Data-driven many-body models enable a quantitative description of chloride hydration from clusters to bulk" A. Caruso, and F. Paesani. In: *J. Chem. Phys.* 155.6 (2021), p. 064502. The dissertation author is the primary investigator and author of this paper.

This research was supported by the National Science Foundation through Grant No. CHE-1453204. The simulations used resources of the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by the National Science Foundation through Grant No. ACI-1053575; 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; and the Triton Shared Computing Cluster (TSCC) at the San Diego Supercomputer Center.

Chapter 4

Accurate Modeling of Bromide and Iodide Hydration with Data-Driven Many-Body Potentials

4.1 Introduction

Halide ions are among the most studied electrolytes, both in experiments and in simulations, due to their role in various natural and industrial processes.[293–296] They are frequently encountered in biological systems[297] and found to play important roles in electrochemistry,[12] and environmental chemistry.[8–10] Moreover, due to their spherical symmetry and short-lived interactions with water molecules, halide ions are frequently used as benchmarks in fundamental studies of hydration phenomena of negative ions.[298]

Developing a molecular-level understanding of the properties of hydrated ions from small gas-phase clusters to aqueous solutions poses several challenges to both experiment and simulation.[17,23,298–308] Specific ion effects on the local structure of the hydrogen-bond (H-bond) network of water have been the focus of extensive investigations.[14,16,28,29,31,32,272,298,309–311] It is established that the presence of ionic species in aqueous environments gives rise to structural rearrangements of the water H-bond network. However, the extent of these rearrangements remains matter of debate. Until recently, the common view has been to classify ions as "structure makers" or "structure breakers" de-

pending on whether they strengthen or disrupt the H-bond network of the surrounding water molecules. According to this classification, "structure makers" correspond to ions with high charge density while "structure breakers" are ions with small and diffuse charge density. This classification, inspired by Hofmeister's series on protein stability,[262] has often appeared to be too simplistic and has been repeatedly challenged by experimental measurements.[14–18]

Several theoretical and computational studies of the hydration properties of bromide and iodide ions based on either force fields (FFs) or density functional theory (DFT) have been reported in the literature, starting from pioneering simulations showing enhanced propensity of these ions for the water surface.[28–32]. A polarizable FF derived from first principles[312] was used to characterize the hydration structure and calculate the extended X-ray absorption fine structure (EXAFS) spectrum of bromide in solution which was found to be in qualitative agreement with the corresponding experimental data.[313] Molecular dynamics (MD) simulations carried out with the AMOEBA polarizable force field found that the interactions among water molecules in the first solvation shell around a bromide ion are similar to those in pure bulk water.[314] A systematic analysis of the hydration properties in ion–water clusters, including $Br^-(H_2O)_n$ clusters, carried out at the DFT level found that the differences in dipole moments between molecules residing inside and outside of the first hydration shell of the ion become smaller as the cluster size increases, which was interpreted as evidence in support of the use of nonpolarizable FFs in MD simulations.[315] Car-Parrinello MD simulations of bromide in water carried out with the BLYP functional were used to calculate the EXAFS spectrum that was found to be in better agreement with the experimental data than the analogous spectrum calculated using the TIP3P and OPLS models.[316] The BLYP functional was also used in DFT and quantum mechanics/molecular mechanics (QM/MM) simulations of iodide in water.[317] Somewhat surprisingly, it was found that the EXAFS spectrum calculated from the QM/MM simulations was in better agreement with the experimental data than the analogous spectrum calculated from the

DFT simulations. DFT simulations carried out with the BLYP-D functional were used to calculate the EXAFS spectra of iodide in water which were found to be in qualitative agreement with the experimental data, displaying some differences in both phases and amplitudes.[318] More recently, polarizable FFs based on the BK3 water model were used to determine the hydration structure of both halide and alkali-metal ions.[319] A subsequent study found that these BK3-based FFs overpredict water structuring around the ions in solution and concluded that the ion–water interactions are not adequately represented by these FFs.[320] MD simulations carried out with the ONIOM-XS approach found that bromide–water interactions in solution are weak and give rise to a loosely bound first shell.[321]

In previous studies,[107,108,143,144] we introduced the many-body energy (MB-nrg) theoretical/computational framework for data-driven many-body potential energy functions (PEFs) that are rigorously derived from the many-body expansion (MBE) of the energy calculated using coupled cluster theory, including single, double, and perturbative triple excitations, i.e., CCSD(T), which is currently considered as the "gold standard" for molecular interactions.[117] When used to model halide–water interactions, the MB-nrg PEFs were shown to provide high accuracy, quantitatively reproducing the energetics of small $X^-(H_2O)_N$ clusters, with X = F, Cl, Br, and I,[40,270] as well as the vibrational spectra and tunneling splitting of halide monohydrate[292] and dihydrate complexes.[270,272] More recently, we demonstrated that the original chloride–water MB-nrg PEF[107] could be improved by refining the training set of the 2-body energies via active learning[120] and including an explicit 3-body energy term.[113] The new chloride–water MB-nrg PEF was shown to achieve CCSD(T) accuracy in representing the interaction energies of $Cl^-(H_2O)_n$ clusters and, when used in MD simulations, it predicted the correct hydration structure of chloride in water as demonstrated by the quantitative agreement between the experimental and calculated extended EXAFS spectra.[113]

In this study, we continue our analysis of many-body effects in ion hydration by

introducing second-generation bromide–water and iodide–water MB-nrg PEFs. The article is structured as follows: in the "Methods" section, we present the new MB-nrg PEFs that are developed from expanded training sets for 2-body and 3-body energies generated using the active learning scheme of Ref. 120. We assess the overall accuracy of the new MB-nrg PEFs in the "Results" section by analyzing their ability to reproduce the energetics of small $Br^-(H_2O)_n$ and $I^-(H_2O)_n$ clusters as well as the hydration structure of each ion in solution. In the "Conclusion", we summarize our work and provide an outlook for potential future applications of the MB-nrg PEFs.

4.2 Methods

4.2.1 MB-nrg PEFs

The MB-nrg PEFs for bromide and iodide in water were developed following Ref. 113. Within the MBE, the energy of a system is obtained as the sum of individual n-body energy terms, ϵ^{nB}, according to

$$E_N(1,..,N) = \sum_{i=1}^{N} \epsilon^{\mathrm{1B}}(i) + \sum_{i<j}^{N} \epsilon^{\mathrm{2B}}(i,j) + \sum_{i<j<k}^{N} \epsilon^{\mathrm{3B}}(i,j,k) + ... + \epsilon^{\mathrm{NB}}(1,..,N), \qquad (4.1)$$

Since the MB-nrg theoretical/computational framework has already been described in the literature,[107,108,143,144,322] we will only describe here the salient features and provide details specific to the bromide–water and iodide–water PEFs. Briefly, the MB-nrg PEFs use a combination of short-range permutationally invariant polynomials (PIPs)[186] trained on electronic structure data and physics-based functions to represent the 1B, 2B, and 3B terms of the MBE in Eq. 4.1,[274,322] while all other nB terms with $n > 3$ are represented by an implicit many-body term derived from the Thole model of classical polarization.[323]

Within the MB-nrg theoretical/computational framework, Eq. 4.1 is thus expressed

as

$$E_N = \sum_{i=1}^{N} \epsilon^{1\mathrm{B}}(i) + \sum_{i>j}^{N} \epsilon^{2\mathrm{B}}(i,j) + \sum_{i>j>k}^{N} \epsilon^{3\mathrm{B}}(i,j,k) + V_{\mathrm{pol}}^{>3\mathrm{B}}, \tag{4.2}$$

where $\epsilon^{1\mathrm{B}}(i)$ is the distortion energy of the ith monomer in the system, and all other nB terms are defined recursively as

$$\begin{aligned} \epsilon^{\mathrm{nB}}(1,\ldots,n) &= E_n(1,\ldots,n) - \sum_i \epsilon^{1\mathrm{B}}(i) - \sum_{i<j} \epsilon^{2\mathrm{B}}(i,j) - \ldots \\ &- \sum_{i<j<k} \epsilon^{3\mathrm{B}}(i,j,k) - \ldots - \epsilon^{(\mathrm{n}-1)\mathrm{B}}(1,\ldots,n-1) \end{aligned} \tag{4.3}$$

Here, ϵ^{nB} is the n-body energy, and $\mathrm{E}_n(1,\ldots,n)$ is the energy of a subsystem containing n monomers. Since the halide ions are monoatomic species, the 1-body term of the halide–water MB-nrg PEFs contains contributions only from the intramolecular distortions of the water molecules, which are described by the MB-pol PEF[97,98,106] through the model developed by Partridge and Schwenke.[185]

In Eq. 4.2, the 2-body energy, $\epsilon^{2\mathrm{B}}$, takes the form

$$\epsilon^{2\mathrm{B}} = V_{\mathrm{sr}}^{2\mathrm{B}} + V_{\mathrm{elec}}^{2\mathrm{B}} + V_{\mathrm{disp}}^{2\mathrm{B}} + V_{\mathrm{pol}}^{2\mathrm{B}} \tag{4.4}$$

where $V_{\mathrm{sr}}^{2\mathrm{B}}$ is a short-range term expressed by a PIP that is fitted to reproduce CCSD(T) 2-body energies and switched off when the distance between the halide ion (X) and the oxygen atom (O) of the water molecule within the dimer is larger than a predefined cutoff,

$$V_{\mathrm{sr}}^{2\mathrm{B}} = s_2\left(R_{\mathrm{X^-O}}\right) \cdot V_{\mathrm{PIP}}^{2\mathrm{B}} \tag{4.5}$$

The switching function, $s_2\left(R_{\text{X}^-\text{O}}\right)$, is given by[113]

$$s_2(\text{R}_{\text{X}^-\text{O}}) = \begin{cases} 1, & \text{if } t_2(R_{\text{X}^-\text{O}}) < 0 \\ cos^2\left[t_2(R_{\text{X}^-\text{O}})\pi/2\right], & \text{if } 0 \leq t_2(R_{\text{X}^-\text{O}}) < 1 \\ 0, & \text{if } 1 \leq t_2(R_{\text{X}^-\text{O}}) \end{cases} \tag{4.6}$$

with

$$t_2(R_{\text{X}^-\text{O}}) = \frac{R_{\text{X}^-\text{O}} - R_{\text{in}}^{\text{2B}}}{R_{\text{out}}^{\text{2B}} - R_{\text{in}}^{\text{2B}}}. \tag{4.7}$$

$R_{\text{in}}^{\text{2B}}$ and $R_{\text{out}}^{\text{2B}}$ are the inner and outer cutoffs of $s_2\left(R_{\text{X}^-\text{O}}\right)$ which are chosen in order to ensure a smooth and continuous variation of ϵ^{2B} in the switching region. ($R_{\text{in}}^{\text{2B}}$, $R_{\text{out}}^{\text{2B}}$) = (5.9 Å, 7.9 Å) and (6.2 Å, 8.2 Å) for the bromide–water and iodide–water MB-nrg PEFs, respectively.

V_{elec} in Eq. 4.4 represents electrostatic interactions between the negative (-1 e) charge of the halide ions and the geometry-dependent point charges of the water molecule which are obtained by fitting the *ab initio* dipole moment of an isolated water molecule calculated in Ref. 185.

$V_{\text{disp}}^{\text{2B}}$ in Eq. 4.4 represents the 2-body dispersion energy and is expressed as

$$V_{\text{disp}}^{\text{2B}} = -f(\delta_{\text{X}^-\text{O}})\frac{C_{6,\text{X}^-\text{O}}}{R_{\text{X}^-\text{O}}^6} - f(\delta_{\text{X}^-\text{H}_1})\frac{C_{6,\text{X}^-\text{H}_1}}{R_{\text{X}^-\text{H}_1}^6} - f(\delta_{\text{X}^-\text{H}_2})\frac{C_{6,\text{X}^-\text{H}_2}}{R_{\text{X}^-\text{H}_2}^6} \tag{4.8}$$

where $R_{\text{X}^-\text{O}}$, $R_{\text{X}^-\text{H}_1}$, and $R_{\text{X}^-\text{H}_2}$ are the distances between the ion (X^-), and the oxygen (O) and the two hydrogen (H) atoms of the water molecule within a X^-–H_2O dimer, and $f(\delta)$ and C_6 are the corresponding Tang-Toennies damping functions[278] and dispersion coefficients determined in Ref. 107.

The 3-body energy in Eq. 4.2, ϵ^{3B}, is given by

$$\epsilon^{\text{3B}} = V_{\text{sr}}^{\text{3B}} + V_{\text{pol}}^{\text{3B}} \tag{4.9}$$

As in ϵ^{2B} of Eq. 4.4, $V_{\rm sr}^{3B}$ is a short-range term expressed by a PIP that is fitted to reproduce CCSD(T) 3-body energies and switched off when two or more of the X-O and O-O distances are larger than a predefined cutoff value according to:

$$\begin{aligned} V_{\rm sr}^{3B} = [&s_3(R_{\rm X\text{-}O_a})s_3(R_{\rm X\text{-}O_b}) \\ &+ s_3(R_{\rm X\text{-}O_a})s_3(R_{\rm O_aO_b}) \\ &+ s_3(R_{\rm X\text{-}O_b})s_3(R_{\rm O_aO_b}) \\ &- 2\cdot s_3(R_{\rm X\text{-}O_a})s_3(R_{\rm X\text{-}O_b})s_3(R_{\rm O_aO_b})]\cdot V_{\rm PIP}^{3B} \end{aligned} \tag{4.10}$$

A combination of switching functions $s_3(R_{kl})$ acting on each (X,O) and (O,O) pair is used to guarantee that ϵ^{3B} transitions smoothly between $(V_{\rm PIP}^{3B} + V_{\rm pol}^{3B})$ at short range and $V_{\rm pol}^{3B}$ at long range, with $s_3(R_{kl})$ given by

$$s_3(R_{kl}) = \begin{cases} 1, & \text{if } t_3(R_{kl}) < 0 \\ cos^2\big[t_3(R_{kl})\pi/2\big], & \text{if } 0 \le t_3(R_{kl}) < 1 \\ 0, & \text{if } 1 \le t_3(R_{kl}) \end{cases} \tag{4.11}$$

Here, the variable $t_3(R_{kl})$ depends on the inner $(R_{\rm in}^{3B})$ and outer $(R_{\rm out}^{3B})$ cutoff values according to:

$$t_3(R_{kl}) = \frac{R_{kl} - R_i^{3B}}{R_{\rm out}^{3B} - R_{in}^{3B}} \tag{4.12}$$

As discussed in Ref. 113, the MB-nrg framework provides the user with complete freedom for the choice of the inner and outer cutoffs. To account for the relatively large size and polarizability of the two halide ions, ($R_{\rm in}^{3B}$ $R_{\rm out}^{3B}$) were set equal to (3.9 Å, 5.9 Å) and (5.5 Å, 6.0 Å) for bromide and iodide, respectively.

Finally, $V_{\rm pol}^{2B}$ in Eq. 4.4, $V_{\rm pol}^{3B}$ in Eq. 4.9, and $V_{\rm pol}^{>3B}$ in Eq. 4.2 are implicitly included in a classical N-body polarization term, $V_{\rm pol}^{\rm NB}$, derived from the Thole model.[323] The effective atomic polarizabilities, $\alpha^{\rm eff}$, for the bromide and iodide ions in water were determined

from exchange-dipole moment (XDM)[324,325] calculations carried out with Gaussian 16[326] and postg[327,328] for $Br^-(H_2O)_n$ and $I^-(H_2O)_n$ clusters. Specifically, clusters of increasingly larger radius were extracted from MD simulations for systems containing a single ion and 277 water molecules carried out in the isothermal-isobaric (NPT) ensemble at 298 K and 1 atm using the TTM-nrg PEFs.[267] For each radius, 20 clusters were randomly selected and the corresponding effective atomic polarizabilities were determined as[329]

$$\alpha^{\text{eff}} = \alpha^{\text{free}} \left(\frac{V^{\text{eff}}}{V^{\text{free}}} \right)^{4/3} \tag{4.13}$$

Here V^{eff} and V^{free} are the effective and free volumes of the halide ions, respectively, obtained from the XDM calculations. Fig. 4.1 shows the average values of α^{eff} for the bromide and ions as a function of the cluster size. The bulk values of α^{eff} were determined from the asymptotic limit of the two curves, which results in the values of 3.7819 Å^3 and 5.9563 Å^3 for Br^- and I^-, respectively.

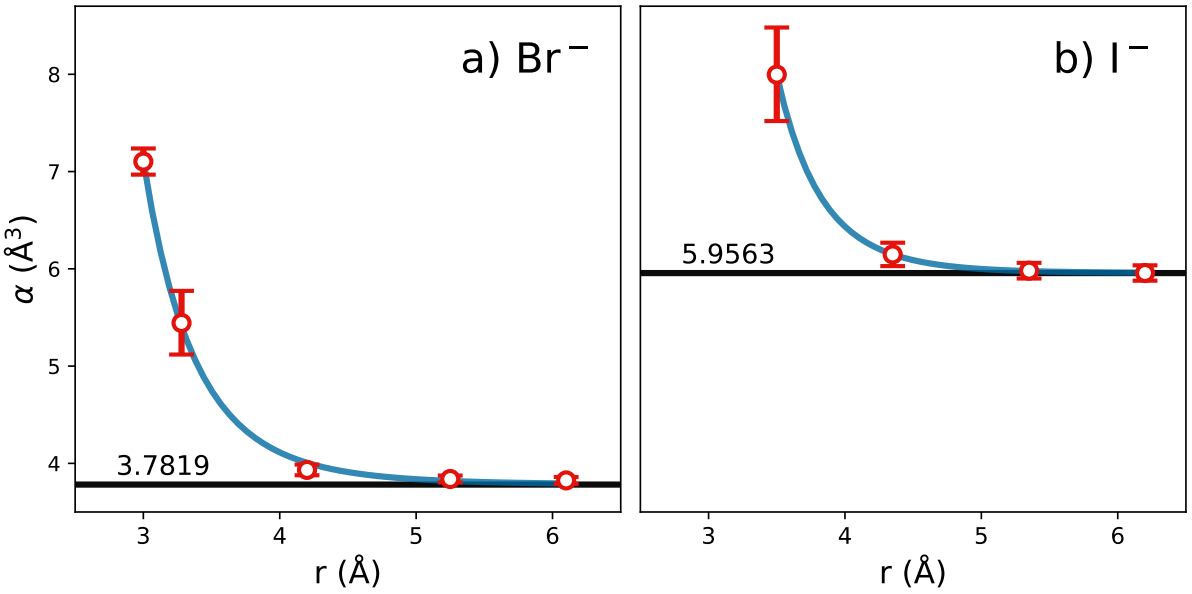

Figure 4.1. Variation of bromide (panel a) and iodide (panel b) polarizabilities calculated with XDM as a function of the radius (r) of the corresponding $Br^-(H_2O)_n$ and $I^-(H_2O)_n$ clusters. The error bars are determined as 95% confidence interval.

4.2.2 Pemutationally invariant polynomials

Both $V_{\mathrm{PIP}}^{\mathrm{2B}}$ and $V_{\mathrm{PIP}}^{\mathrm{3B}}$ are functions of the pairwise distances between the ion (X), the hydrogen and oxygen atoms (H and O), and the lone-pair sites of the MB-pol water molecules (L_1 and L_2)[97]. $V_{\mathrm{PIP}}^{\mathrm{2B}}$ contains 496 symmetrized monomials (ξ_i): 3 first-degree monomials, 15 second-degree monomials, 49 third-degree monomials, 130 fourth-degree monomials, and 299 fifth degree monomials. $V_{\mathrm{PIP}}^{\mathrm{2B}}$ thus contains 496 linear fitting parameters (c_i) and 9 nonlinear fitting parameters.[107,113] $V_{\mathrm{PIP}}^{\mathrm{3B}}$ contains 1575 symmetrized monomials, ξ_i: 39 second-degree monomials, 613 third-degree monomials, and 923 fourth-degree monomials. Therefore, $V_{\mathrm{PIP}}^{\mathrm{3B}}$ contains 1575 linear fitting parameters and 13 nonlinear fitting parameters.

4.2.3 Fitting procedure

As in MB-pol[97,98] and other MB-nrg PEFs,[107–109,113,143,144] the linear parameters of the 2- and 3-body PIPs were fitted through singular value decomposition while the simplex algorithm was used for the non-linear parameters.

The regularized weighted sum of squared deviations, χ^2, was minimized over the training set $\mathcal{S}$, while the L linear parameters were regularized with $\Gamma = 0.0005$:

$$\chi^2 = \sum_{n\in\mathcal{S}} w_n[\epsilon_{\mathrm{model}}(n) - \epsilon_{\mathrm{ref}}(n)]^2 + \Gamma^2 \sum_{l=1}^{L} c_l^2 \tag{4.14}$$

Here, w_n are weights to emphasize low binding energy configurations according to

$$w(E_i) = \left[\frac{\Delta E}{E_i - E_{\mathrm{min}} + \Delta E}\right]^2. \tag{4.15}$$

where E_{min} is the lowest binding energy in $\mathcal{S}$ and ΔE is the range of favorable configurations which was set to 30 kcal/mol and 42.5 kcal/mol for the 2-body and 3-body energies, respectively. The fitting process was performed using a development version of the MB-Fit

software.[322,330]

4.2.4 Reference energies

The training sets for the 2-body energies were generated using the method described in Refs. 120 and 113. A pool of $\sim$ 150000 configurations was generated from different sources: a spherical grid with the water molecule in a $\sim 2-8$ Å shell from the ion, normal modes of the ion–water dimer, and MD simulations carried out in periodic boundary conditions using the TTM-nrg PEFs[267] for a box containing a single ion and 277 water molecules (see Section "Molecular dynamics simulations". The reference 2-body energies were obtained at the CCSD(T)-F12b level of theory[136,137] in the complete basis set (CBS) limit that was achieved via a two-point extrapolation[138,139] between 2-body energies calculated with the augmented correlation-consistent polarized valence triple- (aug-cc-pVTZ) and quadruple-ζ (aug-cc-pVQZ) basis sets.[238,239,279,280] The final 2-body training sets consist of 17057 bromide–water and 15810 iodide–water dimers, while the corresponding test sets consist of 1795 and 1668 dimer configurations, respectively.

3-body training sets consisting of 33830 $Br^-(H_2O)_2$ and 33985 $I^-(H_2O)_3$ trimers were also generated from MD simulations carried with the TTM-nrg PEFs.[267] The corresponding tests consist of 3576 and 3621 trimer configurations, respectively. The 3-body energies were calculated at the CCSD(T)-F12b level of theory[136,137] using the aug-cc-pVTZ basis set.[238,239,279,280] All CCSD(T)-F12b calculations were carried out using MOLPRO (version 2020.1).[281]

4.2.5 Molecular dynamics simulations

All the MD simulations were carried out in the NPT ensemble at 298.15 K and 1.0 atm for an orthorhombic box containing a single ion and 277 water molecules, which corresponds to a concentration of $\sim$0.2 M concentration. The velocity-Verlet algorithm was used to propagate the equations of motion with a time step of 0.5 fs according to

Ref. 191. The temperature and pressure were maintained using a global Nosé–Hoover chain of 3 thermostats with a relaxation time of 0.05 ps, and a global Nosé–Hoover barostat with a relaxation time of 0.5 ps which was thermosttated by a chain of three thermostats. The NPT simulations consisted of 0.1 ns of equilibration followed by 1 ns of production. Short-range interactions were evaluated with a real-space cutoff of 9 Å, while long-range interactions (including electrostatic, dispersion, and polarization contributions) were calculated in reciprocal space using a particle–particle particle–mesh solver.[331] All MD simulations were carried out using the Large-scale Atomic/Molecular Massively Parallel Simulator (LAMMPS)[192] package interfaced with the MBX software for many-body PEFs.[189]

4.2.6 Extended X-ray absorption spectroscopy

Two different EXAFS analysis methods were used to compare the experimentally measured structures with the structures obtained from MD simulations. The first method is exact and consists in generating an ensemble average of EXAFS spectra calculated for a set of snapshots extracted from the simulated trajectory (MD-EXAFS). The MD-EXAFS method has previously been described in Ref. 332. From the equilibrated portion of the MD trajectory, 2000 equally spaced frames are selected and the Cartesian coordinates of the halide ion, and oxygen and hydrogen atoms of the water molecules are retrieved. Each set is used as input to the EXAFS scattering code, FEFF9,[288–290] in order to generate the ensemble average $\chi(k)$ spectra. As in Refs. 235 and 113, all FEFF calculations were performed using clusters containing the halide ion and its 33 closest water molecules, extracted from the corresponding NPT trajectories.

The second method involves fitting a small set of theoretical standards to measured or calculated spectra. The EXAFS analysis software Artemis was used for this approach.[333] The method adopts a Gaussian model and thus approximates the position of each atom with a normal distribution centered on its average value. The limitations of this approach

have been discussed with respect to its application to disordered systems.[334,335] Fitting to the theoretical standards requires some a priori insight into the chemical makeup of the system in order to judiciously select a set of the most important nearby neighbor atoms for which 3 or 4 scattering paths are created. Then using a least-squares fitting procedure, refinements to coordination numbers, distances, and disorder are used to provide a best-fit to the measured or simulated spectrum.

4.3 Results

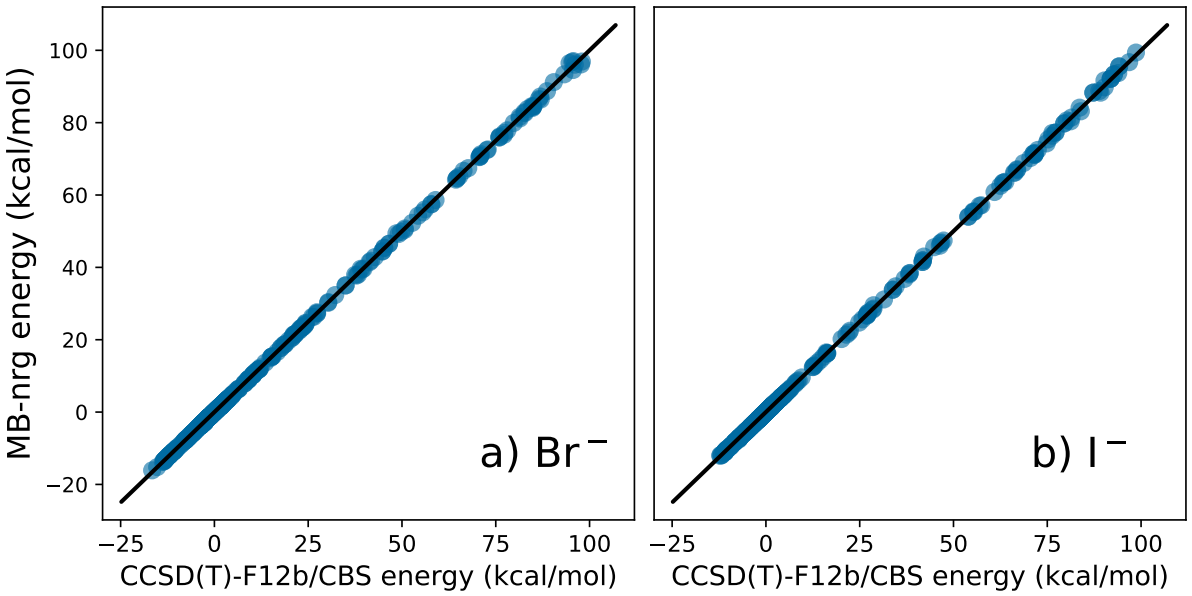

Figure 4.2. 2-body energy correlation plots between the CCSD(T)-F12b/CBS reference values (x axis) and corresponding MB-nrg values (y axis) for the bromide–water (panel a) and iodide–water (panel b) test sets.

To assess the accuracy of the 2-body terms of the MB-nrg PEFs, Fig. 4.2 shows correlation plots between the CCSD(T)-F12b/CBS reference 2-body energies and the corresponding MB-nrg values calculated for the bromide–water and iodide–water dimers in the test sets. The associated root mean square errors (RMSEs) for both training and test sets are reported in Table 4.1. Both MB-nrg PEFs achieve CCSD(T)/CBS accuracy over the entire energy range from -25 kcal/mol to 105 kcal/mol. The ability of the MB-nrg PEFs

Table 4.1. Root mean square errors (RMSEs) associated with bromide–water and iodide–water 2-body and 3-body energies calculated with the MB-nrg PEFs relative to the corresponding CCSD(T)-F12b/CBS reference values of the training and test sets.

MB-nrg PEF	2-body RMSE (kcal/mol)		3-body RMSE (kcal/mol)	
	Training	Test	Training	Test
bromide–water	0.1947	0.1871	0.0404	0.0470
iodide–water	0.2010	0.2287	0.0513	0.0677

to provide a high-fidelity representation of the CCSD(T)/CBS dimer multidimensional energy landscape is further demonstrated in Fig. 4.3 that shows comparisons between CCSD(T)/CBS and MB-nrg one-dimensional potential energy radial scans calculated for various orientations (θ, ϕ) of each ion relative to the water molecule within a ion–water dimer.

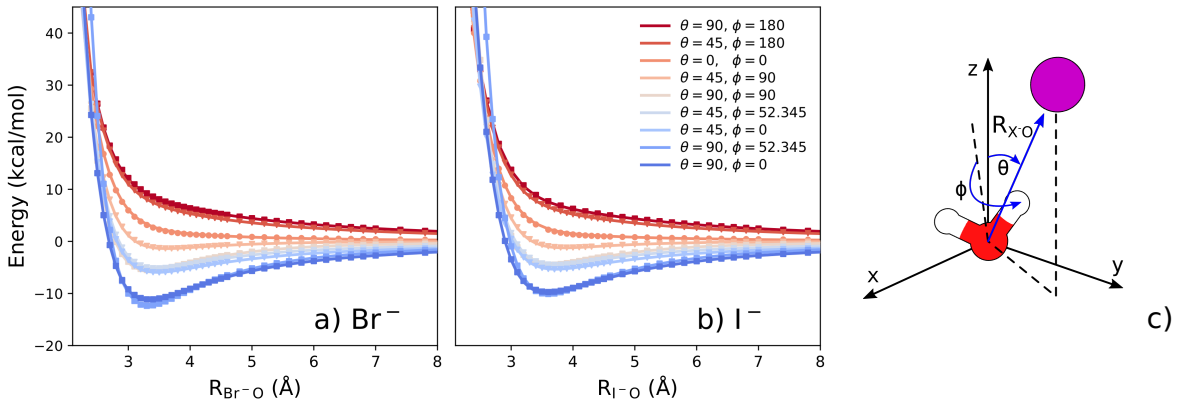

Figure 4.3. Interaction energy scans along the X^-–O distance (R_{X^-O}) for selected orientations (θ, ϕ) of the halide ion relative to the water molecule in a $X^-(H_2O)$ dimer, with X = Br (panel a) and I (panel b). R_{X^-O}, θ, and ϕ are defined in panel c). The symbols correspond to the CCSD(T)-F12b/CBS reference interaction energies, while the corresponding MB-nrg values are shown as solid lines.

Fig. 4.4 shows correlation plots between the CCSD(T)-F12b/CBS reference 3-body energies and the corresponding MB-nrg values calculated for the bromide–water and iodide–water trimers in the test sets, while the RMSEs calculated for both training and test sets are reported in Table 4.1. The comparisons shown in Figs. 4.2 and 4.4 demonstrate that the both bromide–water and iodide–water MB-nrg PEFs are able to quantitatively reproduce

the corresponding CCSD(T)/CBS 2-body and 3-body energies, without overfitting.

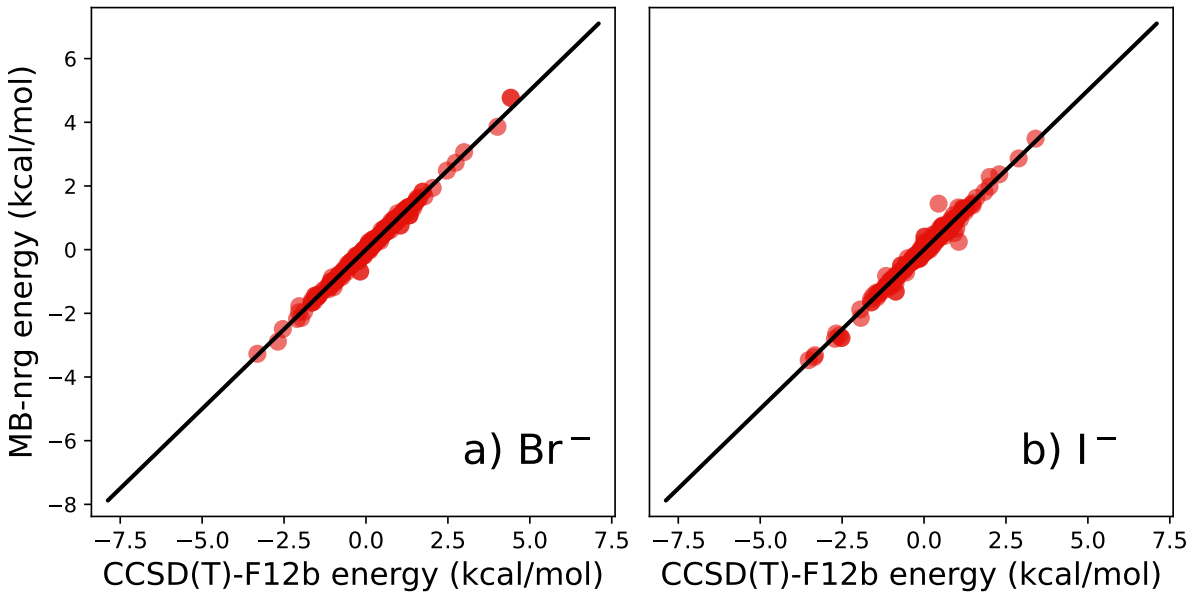

Figure 4.4. 3-body energy correlation plots between the CCSD(T)-F12b/CBS reference values (x axis) and corresponding MB-nrg values (y axis) for the bromide–water (panel a) and iodide–water (panel b) test sets.

While the accuracy exhibited by the 2-body and 3-body terms of the bromide–water and iodide–water PEFs is certainly remarkable, it is also somewhat expected since these terms are explicitly fitted to reproduce the corresponding CCSD(T)-F12b/CBS reference energies. In this context, one of the most arduous challenges for data-driven PEFs is achieving full transferability across phases and/or thermodynamic state points different from those represented in the training sets. To address this challenge, we first analyze the ability of the MB-nrg PEFs to correctly reproduce the CCSD(T)-F12b/CBS interaction energies of small $X^-(H_2O)_n$ clusters shown in Fig. 4.5. It is important to emphasize that the calculations carried out for systems containing a single ion and more than two water molecules correspond to actual predictions since, by construction, the MB-nrg PEFs were only trained up to the 3-body energies in the trimers and all higher many-body terms are represented by classical polarization. Fig. 4.6 shows comparisons between the CCSD(T)-F12b/CBS reference interaction energies[40,270] and the corresponding values calculated

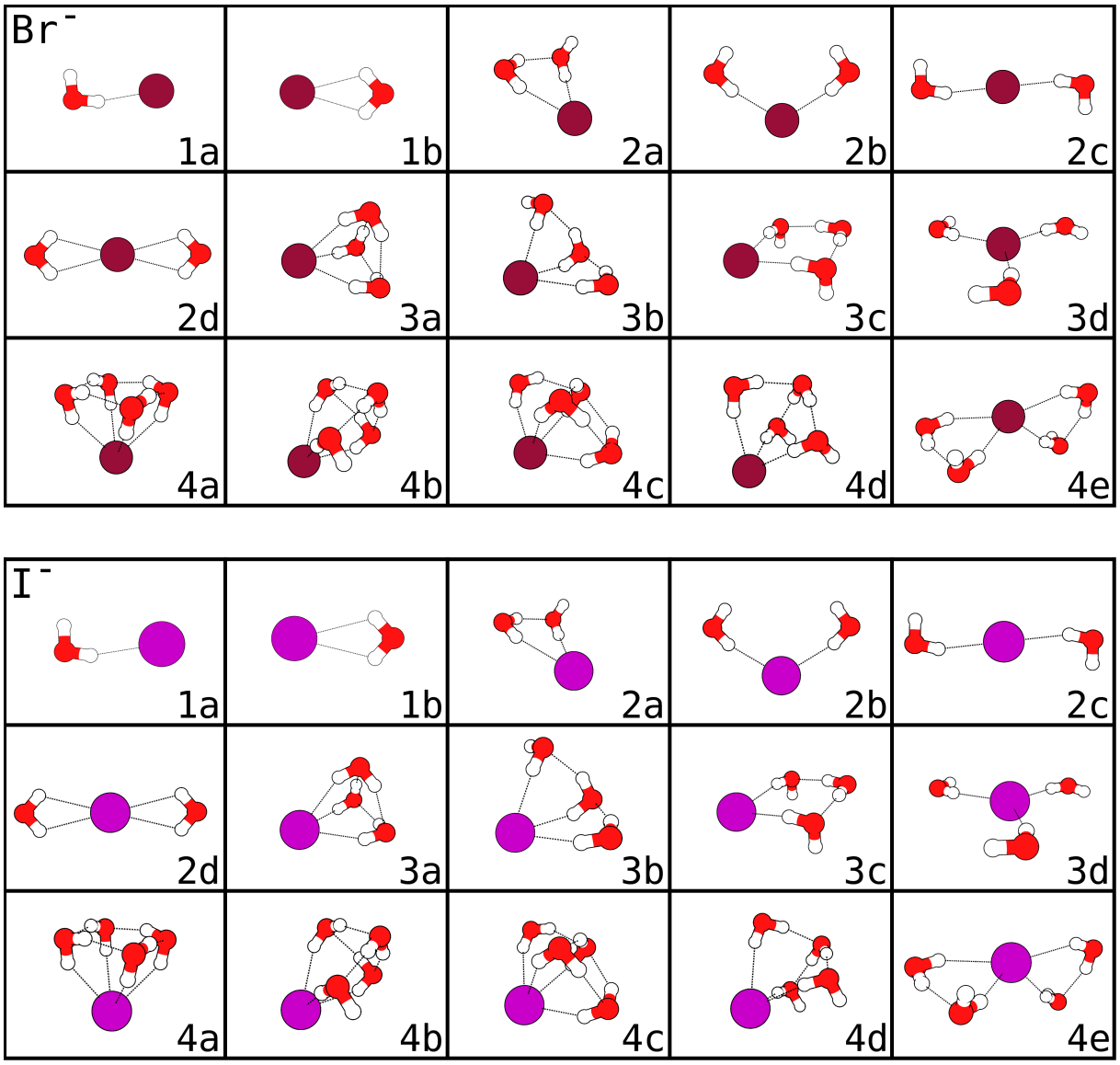

Figure 4.5. Low-lying energy isomers of $Br^-(H_2O)_n$ (top panels) and $I^-(H_2O)_n$ (bottom panels) clusters ($n = 1 - 4$).

with the MB-nrg PEFs for 15 different $X^-(H_2O)_n$ clusters (with $n = 1 - 4$), with X = Br and I. Besides the full MB-nrg PEFs, hereafter referred to as (2B+3B+NB)-MB-nrg PEFs, which include explicit PIP-based representations of both 2-body and 3-body energies (Eqs. 6-14), Fig. 4.6 also shows results obtained with MB-nrg PEFs that include PIP terms only for the 2-body energies and are hereafter referred to as (2B+NB)-MB-nrg PEFs. The comparisons in Fig. 4.6, therefore, allow for assessing not only the overall

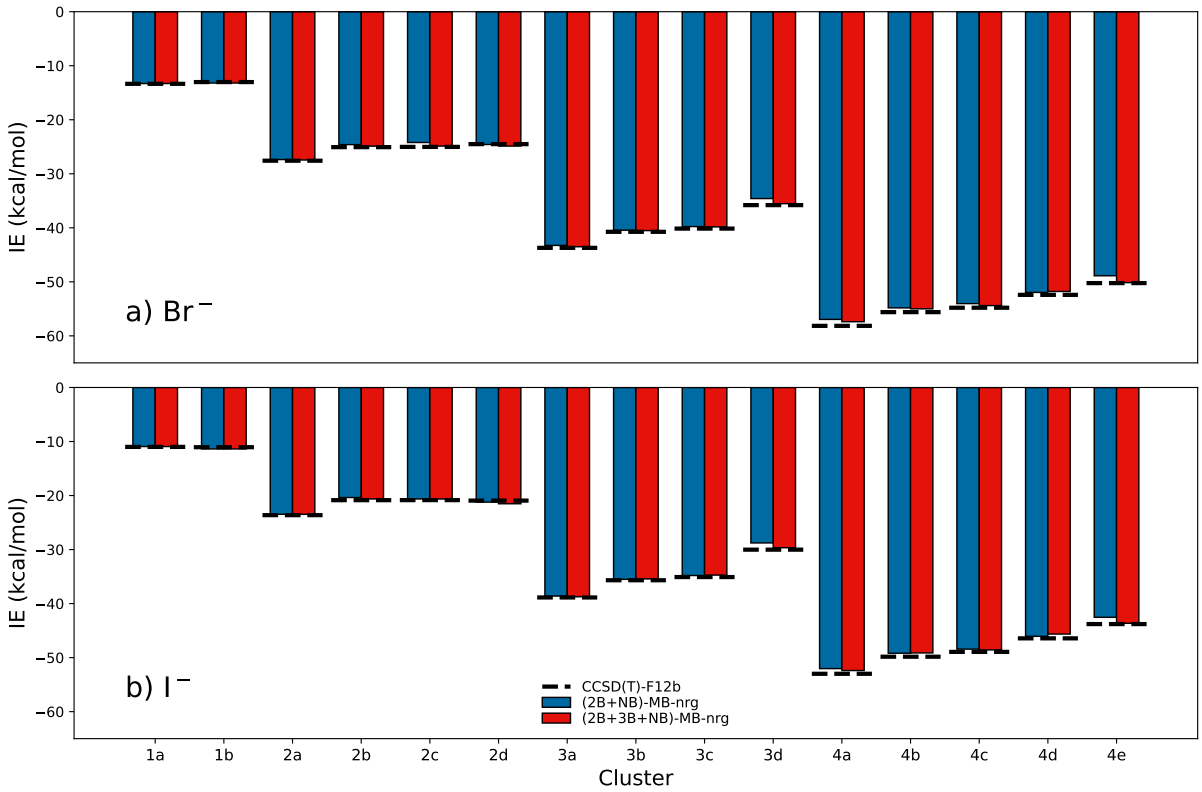

Figure 4.6. Comparison between the interaction energies calculated for the low-energy isomers of $Br^-(H_2O)_n$ (panel a) and $I^-(H_2O)_n$ (panel b) clusters ($n = 1 - 4$) using the (2B+NB)-MB-nrg and (2B+3B+NB)-MB-nrg PEFs. For each cluster, the CCSD(T)-F12b reference values[40,270] are shown as horizontal dashed lines.

accuracy of the full MB-nrg PEFs but also the relative importance of 2-body, 3-body, and higher n-body contributions to the interactions energies of larger bromide–water and iodide–water systems. The (2B+3B+NB)-MB-nrg PEFs quantitatively reproduce the CCSD(T)-F12b/CBS reference energies of both $Br^-(H_2O)_n$ and $I^-(H_2O)_n$, independently of the cluster size and structure. Importantly, the performance of the (2B+3B+NB)-MB-nrg and (2B+NB)-MB-nrg PEFs is very similar, with only small differences found for the isomers with relatively higher interaction energy (e.g., isomers 3d and 4e), which indicates that n-body energy contributions with $n > 2$ are primarily due to classical electrostatic interactions.

The last challenge that remains to be addressed in order to assess the transferability

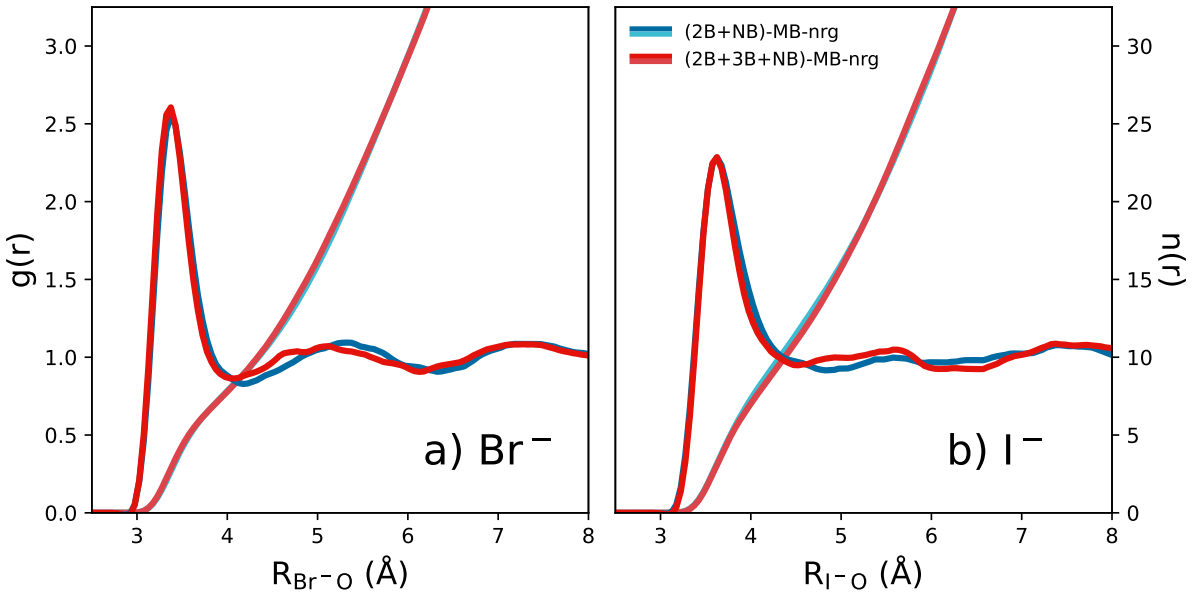

Figure 4.7. Bromide-oxygen (panel a) and iodide-oxygen (panel b) radial distribution functions, $g(r)$, and corresponding coordination numbers, $n(r)$, calculated from NPT simulations carried out at 298 K and 1 atm with the (2B+NB)-MB-nrg (blue) and (2B+3B+NB)-MB-nrg (red) PEFs.

of the MB-nrg PEFs is to determine if the high accuracy displayed in predicting the interaction energies of small clusters in the gas phase translates into realistic descriptions of the hydration structure of the bromide and iodide ions in solution at finite temperature.

In this regard, Fig. 4.7 shows the radial distribution functions (RDFs) and corresponding coordination numbers calculated from NPT simulations carried out with both (2B+NB)-MB-nrg and (2B+3B+NB)-MB-nrg PEFs at 298 K and 1 atm for a box containing a single ion and 277 water molecules in periodic boundary conditions. The RDFs describing spatial 2-body correlations between the Br^- ion and the oxygen (O) atoms of the water molecules exhibit a well-define hydration sttructure, with a prominent, first-shell peak at ∼3.4 Å. Both (2B+NB)-MB-nrg and (2B+3B+NB)-MB-nrg PEFs effectively predict the same position and shape for this first peak. Some minor differences exist between the two MB-nrg PEFs in the region of the second, broader peak corresponding to the second hydration shell that extends between ∼4.2 Å and ∼6.2 Å. Specifically, the

(2B+3B+NB)-MB-nrg PEF predicts a small shift of the second hydration shell towards shorter distances, signalling relatively stronger bromide–water interactions. The variation of the water coordination number as a function of the distance from the bromide ion shows an inflection point at ∼4.2 Å, indicating that the first hydration shell contains ∼ 7-8 water molecules.

The iodide–oxygen RDFs predicted by both (2B+NB)-MB-nrg and (2B+3B+NB)-MB-nrg PEFs shows a first peak similar to that observed in the corresponding bromide–oxygen RDFs, but at slightly larger distances due to the larger size of the iodide ion. However, the evolution of the subsequent hydration shells around iodide is appreciably different. In particular, both (2B+NB)-MB-nrg and (2B+3B+NB)-MB-nrg PEFs predict a shallow hydration structure beyond 4.5 Å, with the presence of a second and third hydration shell only barely visible in the (2B+3B+NB)-MB-nrg PEF. In this regard, it should be noted that the presence of the "kink" at ∼5.5 Å in the iodide-oxygen RDF calculated with the (2B+3B+NB)-MB-nrg PEF might also be due to a less-than-perfect transition from the data-driven component (i.e., PIPs + polarization) to the purely classical component (i.e., polarization) of the 3-body term in Eq. 4.9. The role of the switching functions in the 2-body (Eq. 4.5) and 3-body (Eq. 4.10) terms of the MB-nrg potentials will be the subject of a forthcoming study.

As expected from the shape of the iodide-oxygen RDFs calculated with the (2B+NB)-MB-nrg and (2B+3B+NB)-MB-nrg PEFs, the water coordination number calculated with both PEFs effectively shows a monotonic increase as a function of the distance from the iodide ion, with only a weak inflection point at ∼4.5 Å which precludes a precise determination of the first-shell coordination number. The differences between the bromide–oxygen and iodide–oxygen RDFs primarily arise from the competition between 2-body halide–water and water–water interactions that are further modulated by 3-body iodide–water–water interactions. As shown in Refs. 272, the strength of 2-body halide–water interactions decreases with the size of the halide ions and becomes comparable to the

strength of 2-body water–water interactions in the case of the iodide ion. The competition between iodide–water and water–water interactions thus results in a shallower iodide–oxygen RDF beyond the first hydration shell. When compared to the theoretical models of Ref. 318, the MB-nrg model predicts a nearly identical first peak. However, DFT-based MD and the Dang and Chang (D/C) model predict less interstitial water between the first and the second solvation shell, as reported in the Supporting Information.

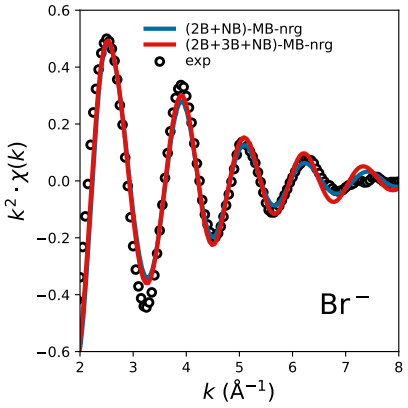

Figure 4.8. K-edge EXAFS spectrum, $k^2\chi(k)$, of bromide in water calculated from NPT simulations carried out at 298 K and 1 atm with the (2B+NB)-MB-nrg and (2B+3B+NB)-MB-nrg PEFs. The experimental EXAFS spectrum from Ref. 336 is shown as black circles.

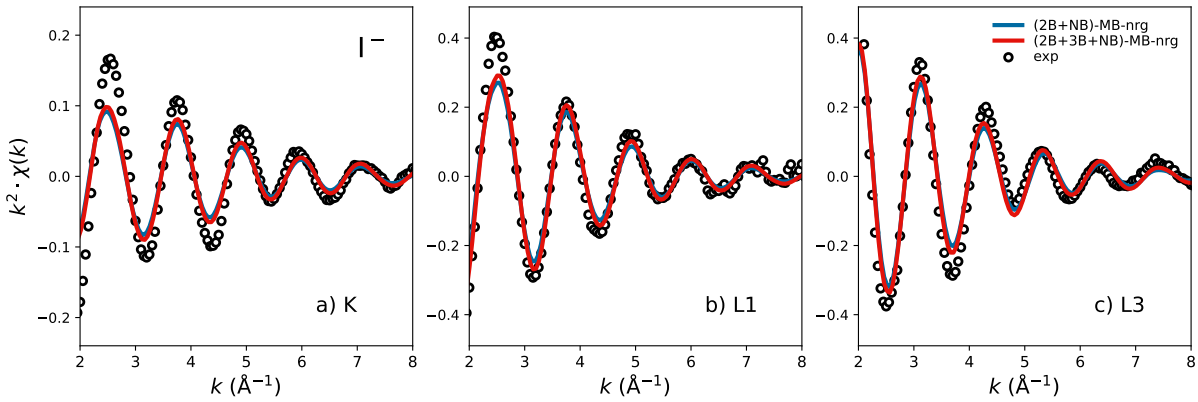

Figure 4.9. K-edge (panel a), L1-edge (panel b), and L3-edge (panel c) EXAFS spectra, $k^2\chi(k)$, of iodide in water calculated from NPT simulations carried out at 298 K and 1 atm with the (2B+NB)-MB-nrg and (2B+3B+NB)-MB-nrg PEFs. The corresponding experimental EXAFS spectra from Ref. 318 are shown as black circles.

While the analysis of the RDFs discussed above allows for gaining insights into the

hydration structure of bromide and iodide ions in solution, it does not provide any evidence for the accuracy and realism of the RDFs predicted by the MB-nrg PEFs. To address this last challenge, in Figs. 4.8 and 4.9, we show comparisons between the experimental and simulated EXAFS spectra, calculated with the MD-EXAFS methodology introduced in the Methods section, for bromide and iodide in water, respectively. The K-edge spectra calculated with the (2B+NB)-MB-nrg and (2B+3B+NB)-MB-nrg PEFs for bromide in water are effectively indistinguishable from each other and in quantitative agreement with the experimental spectrum, with only small deviations for k in the 3-3.5 Å^{-1} range. Similar agreement between the experimental and MB-nrg results was reported for the K-edge spectrum of chloride in water.[113]

In Fig. 4.9, we report results relative to three different absorption edges for iodide (K-, L1-, and L3-edges). With regard to the EXAFS single scattering paths, there is a simple 90° phase shift in the $\chi(k)$ oscillations from the K- or L1- edge spectra (1s and 2s initial states, respectively) with respect to those observed for the L3-edge (2p initial state). However, the symmetry selection rules dictate that the K-, L1-, and L3-edge spectra provide independent, non-redundant, measurements of the local symmetry with respect to multiple scattering paths.[337]

All three EXAFS spectra calculated with the iodide–water MB-nrg PEFs are also in good agreement with the experimental data, with the agreement improving as k increases. However, while the phase of each spectrum is quantitatively reproduced by the MB-nrg PEFs over the entire range of k values, some discrepancies in the amplitudes of the oscillations exist at small values of k, especially in the case of the K-edge spectrum. Interestingly, the differences seen in the second shells of the RDFs predicted by the (2B+NB)-MB-nrg and (2B+3B+NB)-MB-nrg PEFs appear to only have minimal effects on the corresponding EXAFS spectra. This provides further evidence for the local nature of the EXAFS measurements that are primarily sensitive to the first hydration shell. The amplitude discrepancies for the K- and L1-edge spectra (1s and 2s initial states,

respectively), especially in the region around 2.5 $Å^{-1}$ and 4.2 $Å^{-1}$, are due to multi-electron excitations that are part of the atomic background function present in the experimental spectra.[318,338]. Importantly, the multi-electron excitations for the L3-edge (2p initial state) have different distributions and intensities of transitions. The complete set of K-, L1- and L3-edge atomic-background multi-electron excitation features are also observed in the residuals for the model fit to the experimental spectra that is shown in the $\chi(k)$ spectra of Fig. S6a in the Supporting Information. In the light of these moderate distortions of the experimental EXAFS spectra, the MB-nrg provides excellent overall agreement.

For comparison, EXAFS spectra calculated only including water molecules within the the first hydration shell of the halide ions are shown in Figs. S2 and S3 in the Supporting Information. Although very similar to those calculated by including 33 water molecules, i.e., including water molecules residing beyond the first hydration shell, the EXAFS spectra calculated with only water molecules within the first solvation shell show small differences in the amplitudes as well as at large k values for both halide ions. This suggests that, when possible, including a larger number of water molecules beyond the first hydration shell in the FEFF calculations may lead to better converged EXAFS spectra. For comparisons, the EXAFS spectra calculated for I^- using DFT-based simulations as well as simulations with the D/C polarizable model reported in Ref. 318 are shown in Fig. S4.

It is important to emphasize that all the structural details of the RDFs are encoded into the various regions of the calculated EXAFS spectra. For instance, the EXAFS spectra contain information about the rather broad and asymmetric first peak in the calculated RDFs for which both halide ions show a relatively large amount of disorder. Furthermore, the EXAFS spectra are sensitive to interstitial water molecules that reside in the region from 4 to 4.5 Å. At larger distances, however, the lack of any or the presence of very weak structures makes detection more difficult. The signals $\chi(k)$ also contain angular correlation components that are a feature of photoelectron multiple scattering paths.[318]

Table 4.2. Top: Comparison of the structural parameters from K-edge fits to experimental EXAFS gathered from Ref. 336 and MD-EXAFS from this work for the aqueous Br^- first-shell structure. Bottom: Comparison of the structural parameters from simultaneous K-, L1-, and L3-edge fits to experimental EXAFS and MD-EXAFS gathered from Ref. 318 with MD-EXAFS fits from this work for the aqueous I^- first-shell structure. N is the coordination number, R refers to the measured X-H and X-O distances, σ^2 is the Debye-Waller factor, $\phi_{\text{X-H-O}}$ is the X-H-O angle, and the subscript XHO refers to three-leg I-H-O paths. The goodness of the fit, $\mathcal{R}$, is calculated as the sum of the errors squared scaled by the magnitude of the data. See main text for details.

System	Structure						
	Scatterer	N*	R (Å)	$\sigma^2 \times 10^3$ (Å^2)	$\sigma^2_{\text{BrHO}} \times 10^3$ (Å^2)	$\phi_{\text{Br-H-O}}$	$\mathcal{R}$
0.5m RbBr	O_{H_2O}	6	3.274(024)	13.0(4.2)	12.4(2.0)	158°	0.020
	H_{H_2O}	6	2.357(049)	21.6(9.6)			
MB-nrg MD	O_{H_2O}	6	3.392(017)	13.2(2.4)	18.4(4.8)	154°	0.016
	H_{H_2O}	6	2.501(037)	27.1(11.5)			
System	Structure						
	Scatterer	N	R (Å)	$\sigma^2 \times 10^3$ (Å^2)	$\sigma^2_{\text{IHO}} \times 10^3$ (Å^2)	$\phi_{\text{I-H-O}}$	$\mathcal{R}$
0.4m NaI	O_{H_2O}	6.3(0.9)	3.498(025)	17.1(4.7)	20.9(6.8)	148°	0.046
	H_{H_2O}	6.3(0.9)	2.649(028)	35.6(4.5)			
DFT MD	O_{H_2O}	6.0(0.4)	3.584(007)	14.4(1.1)	15.9(1.6)	156°	0.008
	H_{H_2O}	6.0(0.4)	2.686(026)	36.6(5.1)			
D/C MD	O_{H_2O}	5.0(0.4)	3.526(007)	12.7(1.0)	21.4(2.7)	163°	0.013
	H_{H_2O}	5.0(0.4)	2.597(017)	22.7(2.7)			
MB-nrg MD	O_{H_2O}	5.6(0.5)	3.569(010)	16.6(1.4)	19.1(1.0)	147°	0.011
	H_{H_2O}	5.6(0.5)	2.723(029)	34.9(5.4)			

* Fixed parameter.

In order to explore these characteristics of the EXAFS spectra, in Table 4.2 we report the structural parameters obtained from fitting the K-edge of bromide and simultaneously fitting the K-, L1-, and L3-edge of iodide to the theoretical standards, using the second methodology discussed in the Methods section. Specifically, listed in Table 4.2 are the coordination number N, the measured I-H and I-O distances R, the Debye-Waller factor σ^2, and the X-H-O angle $\phi_{\text{X-H-O}}$. The resulting fits are shown in Figs. S5 and S6 of the Supporting Information. For both bromide and iodide (X = Br, I), the X-O, X-H, and X-H-O scattering paths were selected. The bromide K-edge EXAFS data were weighted by k^2, and fitted over the range $1.8 < k < 10.0$ Å^{-1}. Both the real and imaginary parts of $\chi(R)$ are in the region of $1.25 < R < 4.5$ Å. A value $S_0^2 = 0.91$ of the core hole factor was

used. Given the small number of independent points collected, the coordination number was kept fixed at 6 for bromide; for the same reason, the fitted parameters generally show a larger error when compared to iodide. In the case of iodide, the K-, L1- and L3-edge experimental data were weighted by k^3, and fitted simultaneously over the ranges $1.9 < k < 8.8$, $1.9 < k < 8.6$, and $1.9 < k < 7.8$ Å^{-1}, respectively. In all cases, both the real and imaginary parts of the resulting $\chi(R)$ are in the region of $1.25 < R < 4.5$ Å. A value $S_0^2 = 1.0$ of the core hole factor was used. It is important to note that, since the exact same fitting model was applied to the experimental, DFT-based, D/C model, and MB-nrg model spectra, their differences can be quantitatively compared. As shown in Table 4.2, the I-O and I-H distance from the MB-nrg model fit are within 0.07 Å of the experimental values. The Debye Waller factors for I-O, I-H, and I-H-O from MB-nrg are identical to those of the experimental values within fitting errors; this is especially striking when compared to other reported MD-EXAFS results. Similar agreement is observed for the bromide-water structure. Moreover, the MB-nrg iodide-water model is capable of faithfully reproducing the I-H-O angle. The small differences observed in the coordination number are within the error recorded for the experimental fit.

4.4 Conclusions

In this work, we have introduced second-generation bromide–water and iodide–water MB-nrg PEFs. Within the MB-nrg theoretical/computational framework, the two MB-nrg PEFs are derived from the MBE of the energy and includes explicit data-driven 2-body and 3-body energy terms along with an implicit term describing all n-body energy contributions with $n > 3$. The 2-body and 3-body terms are represented by PIP optimized to reproduce the corresponding CCSD(T)-F12b reference data, while the implicit term is represented by classical many-body polarization.

The two MB-nrg PEFs are able to quantitatively reproduce the interaction energies

of $Br^-(H_2O)_n$ and $I^-(H_2O)_n$, effectively achieving CCSD(T)-F12b accuracy in all cases examined in this study. A systematic analysis of the interaction energies show that 3-body energy contributions are primarily due to classical polarization. However, the inclusion of an explicit, short-range 3-body term is shown to be important for retrieving specific features of the hydration shells around the two ions in solution. The second solvation shell is found to be particularly sensitive to 3-body interactions: while in the case of bromide the shell slightly shifts to shorter distances when the short-range PIP is included in the MB-nrg PEF, neglecting short-range iodide-water-water interactions leads to a shallow and nearly featureless iodide–oxygen radial distribution function beyond the first hydration shell.

The structural features that characterize the hydration structure of bromide and iodide in solution predicted by the MB-nrg PEFs are confirmed by the agreement between the experimental and simulated EXAFS spectra. It should, however, be noted that while the phases of the EXAFS spectra simulated with the MB-nrg PEFs correctly reproduce the experimental values, small variations in the amplitudes, especially in the case of the K-edge of iodide, exist, which may be due to inaccuracies in the MB-nrg PEFs and/or multielectron scattering effects that are not accounted for in the simulated spectra.

We believe that the results presented in this study further demonstrate the ability of the MB-nrg PEFs to correctly predict the physics of hydrated halide ions, providing a quantitative representation of halide–water interactions from the gas to the condensed phase and enabling affordable MD simulations of ionic aqueous solutions with CCSD(T) accuracy.

4.5 Acknowledgements

Chapter 4, in full, is a reprint of the material as it appears in "Accurate Modeling of Bromide and Iodide Hydration with Data-Driven Many-Body Potentials" A. Caruso, X.

Zhu, J. Fulton, F. Paesani. In: *J. Phys. Chem. B* 126.41 (2022), pp. 8266–8278. The dissertation author is the primary investigator and author of this paper.

We thank Greg Schenter for stimulating discussions about the calculation and interpretation of the EXAFS spectra. This research was supported by the National Science Foundation through Grant No. CHE-1453204. The simulations used resources of the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by the National Science Foundation through Grant No. ACI-1053575, and the Triton Shared Computing Cluster (TSCC) at the San Diego Supercomputer Center.

Chapter 5

Correcting Delocalization Errors in DFT-Based Representations of Ion Hydration

The necessity of a unified approach in the study of the complex and diverse interplay of forces between ions and water at various thermodynamic conditions arises from the myriad of natural and industrial chemical phenomena in which hydrated ions participate, including, but not limited to, biomolecule stabilization,[3–5] catalytic and transport processes,[1,2,257–261] and electrochemical processes.[12] In this context, the advent of modern computing architectures[126] allowed molecular dynamics (MD) simulations to be adopted as the principal method of discovery in computational molecular sciences.[339,340] The accuracy of molecular models ultimately determines the fidelity of all physical quantities calculated from an MD simulation. In this context, conventional force fields (FFs), which usually contain a set of empirically derived parameters that govern simple analytical expressions within a pairwise additive approximation, allow low-cost simulations of large systems at the expense of accuracy.[129,132,134,135,341–345] The limitations of pairwise-additive FFs can be overcome by using data-driven potential energy functions (PEFs) trained on *ab initio* reference data.[79,83,146,346] In this context, while reference energies calculated at the coupled cluster level of theory including single, double, and perturbative triple excitations [CCSD(T)], currently regarded as the "gold standard" of electronic structure

calculations,[117,347] have been used to develop data-driven PEFs for water[97,98] and other molecular systems,[107–109,143,144,235,322] the development of advanced PEFs that are trained on more computationally efficient, while still accurate, "first principles" methods remains of practical interest.

In principle, density functional theory (DFT) provides an exact treatment of the electronic ground-state potential energy surface (PES) of a given molecular system by solving the Kohn-Sham equations.[50,51] However, as the exact density functional $E[n]$ that characterizes the ground-state of a given system is unknown, "practical" DFT relies on approximations that satisfy, to a varying extent, the exact constraints of DFT.[348–350] The ultimate quality of the density functional approximation (DFA) is given by the exchange-correlation potential $\tilde{V}_{\mathrm{XC}}[n]$ that effectively determines the accuracy of any physical property derived from the approximate functional $\tilde{E}[n]$, including the density itself.

Over the years, different classes of $\tilde{V}_{\mathrm{XC}}[n]$ have emerged to tackle major challenges in molecular DFT, such as accurately describing (i) non-covalent interactions,[351,352] and (ii) electronic charge densities,[152,349,352,353] both of which are crucial for describing ion-hydration.[121,354–356] In this context, semi-local DFAs are the *de facto* approach for "first principles" modeling, due to their often-reasonable accuracy relative to (more expensive) post Hartree-Fock methods, and the efficiency provided by their $\mathcal{O}(N^3)$ scaling. However, despite the success and widespread use of semi-local DFAs, it is well known that the incomplete cancellation of self-interaction in $\tilde{V}_{XC}[n]$[357,358] can lead to a poor description of the static correlation energy,[359] and excess delocalization of the electron density.[353,360,361]

Early DFT studies of water and hydrated ion clusters revealed that the self-interaction error (SIE) and the delocalization error (DE) have a crucial effect on the interaction energies and bond topologies of aqueous systems.[362–364] Due to the complex interconnection between the electronic structure of ion–water systems and their physical properties in the thermodynamic limit, the unreliability of semi-local DFAs may limit

DFT from providing insight into ion hydration phenomena. Critical to this matter, it was recently shown that many modern DFAs do not faithfully predict the electron density of a wide range of systems, suggesting deviation of the field from the path toward the exact functional.[349] In this spirit, the last decade has seen significant progress in our understanding of the strengths and limitations of DFAs in terms of systematic errors,[152,354,365–367] and developing of physically robust and strongly constrained DFAs.[368,369] For instance, the SCAN functional was derived to satisfy the 17 exact constraints known for meta-GGA functionals.[368] Being non-empirical and capable of describing mid-range interactions reasonably well, SCAN continues to find applications for predicting the properties of molecular and extended systems alike,[370] and the development of machine-learned potentials for accelerated simulations of complex molecular systems such as water.[150,176,371]

While SCAN has been the DFA of choice in several recent studies on the properties of water,[150,153,154,176,179,194,355,371–373] as well as hydrated ions,[121,356,374,375] it is known that it over-delocalizes the electron density of such systems, leading to unphysical over-binding that is manifested in large 2-body errors.[154,355] Recently, it was shown that the 2-body energies predicted by SCAN for water cannot be systematically improved to approach CCSD(T) reference energies through a naive modulation of fractional Hartree-Fock exchange.[154] Soon after, an analysis of the density-driven errors (ΔE_{D}) and functional-driven errors (ΔE_{F}) in SCAN revealed minimal ΔE_{F} relative to CCSD(T), providing support for the physical robustness of the DFA adopted by SCAN, and large ΔE_{D} that could be remedied via density-correction.[121,150,153] Thus, density-corrected SCAN (DC-SCAN) has since been shown to be the first DFT-based model to accurately predict the properties of water across its phase diagram,[150,346] and has displayed overall higher accuracy for water and ionic aqueous clusters than modern DFAs such as ωB97M-V[376] and the recently developed DM21.[377,378]

However, despite the proven accuracy of DC-SCAN in the description of water across its phase diagram, researchers in the field of ion hydration have yet to widely adopt

the use of density-corrected DFT-based models. This study presents a systematic study of ion hydration using DC-SCAN and shows that it offers a significant improvement over SCAN, currently regarded as the reference density functional across semi-local DFAs for the modeling of aqueous solutions.[264,265,379–387] We demonstrate that the use of DC-SCAN-based data-driven many-body PEFs, MB-SCAN(DC), enables the efficient study of ions in solution with near-chemical accuracy, by combining the DC-SCAN coverage of short- and mid-range contributions with the dispersion term V_{disp}, which is inherent to the MB-nrg many-body framework.[107,108,113,115,346]

Within the DC-DFT formalism,[149] the total error in any DFT calculation can be decomposed as

$$\Delta E_{\mathrm{DFA}} = \tilde{E}[\tilde{n}] - E[n] = \Delta E_{\mathrm{F}} + \Delta E_{\mathrm{D}} \tag{5.1}$$

where $\Delta E_F = \tilde{E}[n] - E[n]$ is the functional-driven error (FE) due to the specific choice of a DFA approximation, and $\Delta E_D = \tilde{E}[\tilde{n}] - \tilde{E}[n]$ is the density-driven error (DE) due to the approximate self-consistent density predicted by the DFA. As recent studies suggest, the breadth of DC-DFT's applicability (particularly in the form of HF-DFT, where the Hartree-Fock density n_{HF} is used as a proxy for the exact density) is still under investigation.[153,388–391]

To begin our analysis, we quantify the sensitivity of ion–water interaction to the choice of an approximate density, $\tilde{n}$. Figure 5.1(a) shows the average density sensitivity $\langle\tilde{S}\rangle$, calculated with SCAN for a set of $Na^+(H_2O)_n$ and $Cl^-(H_2O)_n$ clusters according to the following expression:

$$\langle\tilde{S}_{\mathrm{DFA}}\rangle = \frac{1}{N}\sum_{i=1}^{N} |\tilde{E}_i[\tilde{n}_{\mathrm{LDA}}] - \tilde{E}_i[\tilde{n}_{\mathrm{HF}}]|, \tag{5.2}$$

as $n_{\mathrm{w}} \equiv n$ increases from $1 \rightarrow 16$. To account for structural dependence, the averages in Eq. (5.2) were performed over sets of $N = 20$ randomly selected cluster configurations

extracted from MD simulations performed using the MB-nrg PEFs for sodium–water[115] and chloride–water[113]. As discussed in detail in previous studies, the ion–water MB-nrg PEFs[107,108] build upon the functional form introduced with the MB-pol PEF for water[97,98] and provide a highly accurate description of ion hydration from small clusters to solutions at infinite dilution. It is apparent from Figure 5.1(a) that the density sensitivity is significant and a point of possible concern in modeling ion hydration, as $\langle \tilde{S} \rangle$ is greater than 2 kcal/mol for systems larger than the dimer,[367,389] and characterized by a quasi-linear increase with respect to the number of water molecules in the cluster (n_{w}). As it has been suggested that the delocalization error may be sensitive to system size up to an asymptotic limit,[153,389] Figure 5.1(b) plots $\langle \tilde{S}/n_{\mathrm{w}} \rangle$, illustrating a contribution of about 1 kcal/mol from individual water molecules to the total density error.

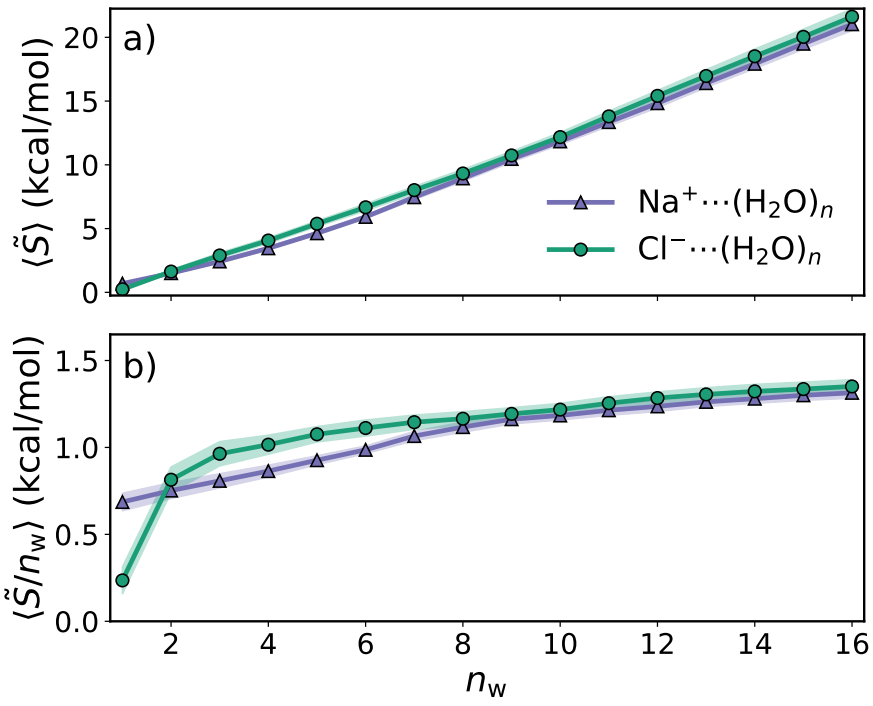

Figure 5.1. Average density sensitivity (a) $\langle \tilde{S} \rangle$ (kcal/mol) associated with the SCAN functional for $Na^+(H_2O)_n$ (purple curve) and $Cl^-(H_2O)_n$ (green curve) clusters where $n_{\mathrm{w}} \equiv n = 1 - 16$, and (b) $\langle \tilde{S}/n_{\mathrm{w}} \rangle$ shows the size dependence of the density sensitivity, illustrating the contributions of individual water molecules to the density error. The error bands are determined as 95% confidence interval.

The significance of the density sensitivity is better appreciated by examining the average error in the interaction energy per water molecule, $\langle \Delta E_{\text{int}}/n_{\text{w}} \rangle$, shown in Figure 5.2 for both SCAN and DC-SCAN relative to the corresponding MB-nrg reference values.[113,115] The per-molecule contribution to the interaction energy error converges beyond the first solvation shell. As expected for semi-local exchange-correlation functionals, SCAN shows excessive attractive ion-water interaction; in this regard, the introduction of the HF density, free of self-interaction, succeeds in remedying the excess delocalization, resulting in a smaller average error. The total interaction energy error of each $Na^+(H_2O)_n$ and $Cl^-(H_2O)_n$ cluster is reported in Fig. S1 of the Supporting Information, highlighting the approximately two-fold reduction in the magnitude of the error.

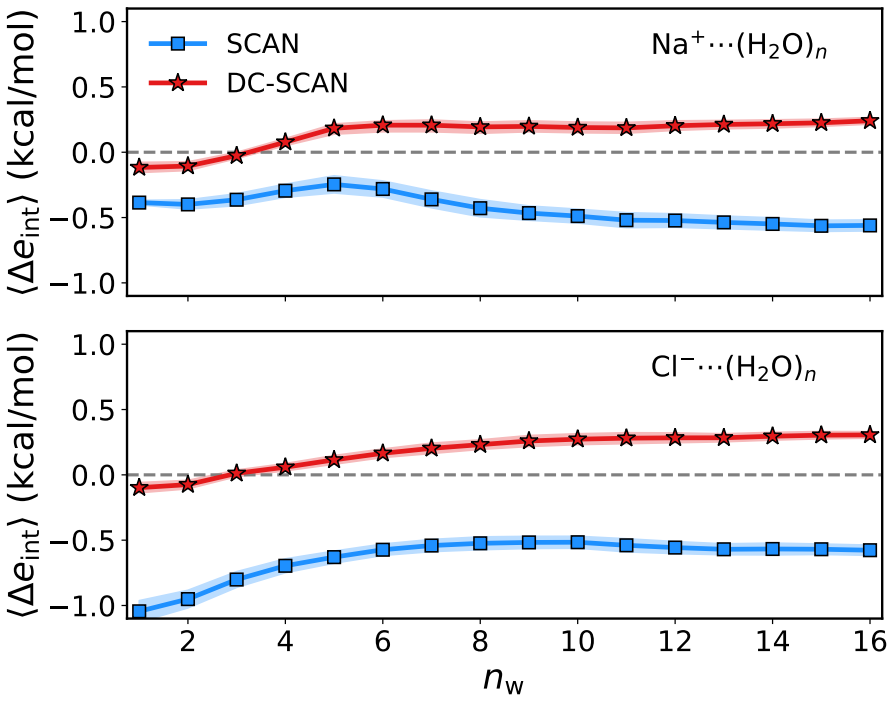

Figure 5.2. Average signed error in the interaction energy per water molecule for clusters with $n_{\text{w}} = 1 - 16$, $\langle \Delta e_{\text{int}} \rangle$, relative to the MB-nrg reference energies (in kcal/mol). The error bands are determined as 95% confidence interval.

Density-driven errors in SCAN are further understood by taking advantage of the many-body expansion (MBE) of the energy,[54] which elegantly allows the total energy E_N

to be expressed as a sum of the individual n-body terms, including 1-body, 2-body, 3-body, and up to the N-body energy:

$$E = \sum_{i=1}^{N} E^{1\mathrm{B}}(i) + \sum_{i>j}^{N} E^{2\mathrm{B}}(i,j) + \sum_{i>j>k}^{N} E^{3\mathrm{B}}(i,j,k) + \ldots + E^{\mathrm{NB}}(1,\ldots,N). \tag{5.3}$$

Figure 5.3 shows the many-body decomposition analysis of the minimum-energy isomers of the $Na^+(H_2O)_4$ and $Cl^-(H_2O)_4$ clusters. In this analysis, we further investigate the different accuracy in energy prediction of SCAN and DC-SCAN with regards to small hydration complexes of alkali-metals and halides.[121] As apparent in panels a) and b) of Fig. 5.3, while DC-SCAN shows errors within ∼1 kcal/mol from MB-nrg reference values, the 2-body errors associated with SCAN are generally larger, reaching up to ∼7 kcal/mol in the case of chloride; the partial correction due to the positive 3-body error only slightly mitigates the effect on the total interaction energy, resulting in the overall better prediction of interaction energies by DC-SCAN relative to the reference values shown in Table 5.1. In gas-phase water clusters, the localized nature of n_{HF} is the foundation to the heightened predictive capabilities of DC-SCAN, as it virtually reduces the excess delocalization of self-consistent SCAN densities.[121] In this context, panels c) and d) of Figure 5.3 show the energy separation of each n-body term into ion–water (iw) and water–water (ww) contributions. A significant fraction of the total error in 2-body and 3-body energies arises from the description of ion–water interactions. The consistently larger error associated with the SCAN and DC-SCAN results for $Cl^-(H_2O)_4$ can be explained by considering that the more diffuse electron density of negatively charged ions results in more pronounced density-driven errors.[121,356,378]

Semi-local exchange-correlation functionals, such as SCAN, are routinely used to perform *ab initio* molecular dynamics (AIMD) simulations of aqueous systems.[372,374,375] Nonetheless, AIMD simulations using DC-SCAN can become computationally prohibitive due to the fact that DC-DFT formally scales as $\mathcal{O}(N^4)$ given n_{HF}, limiting their applicability

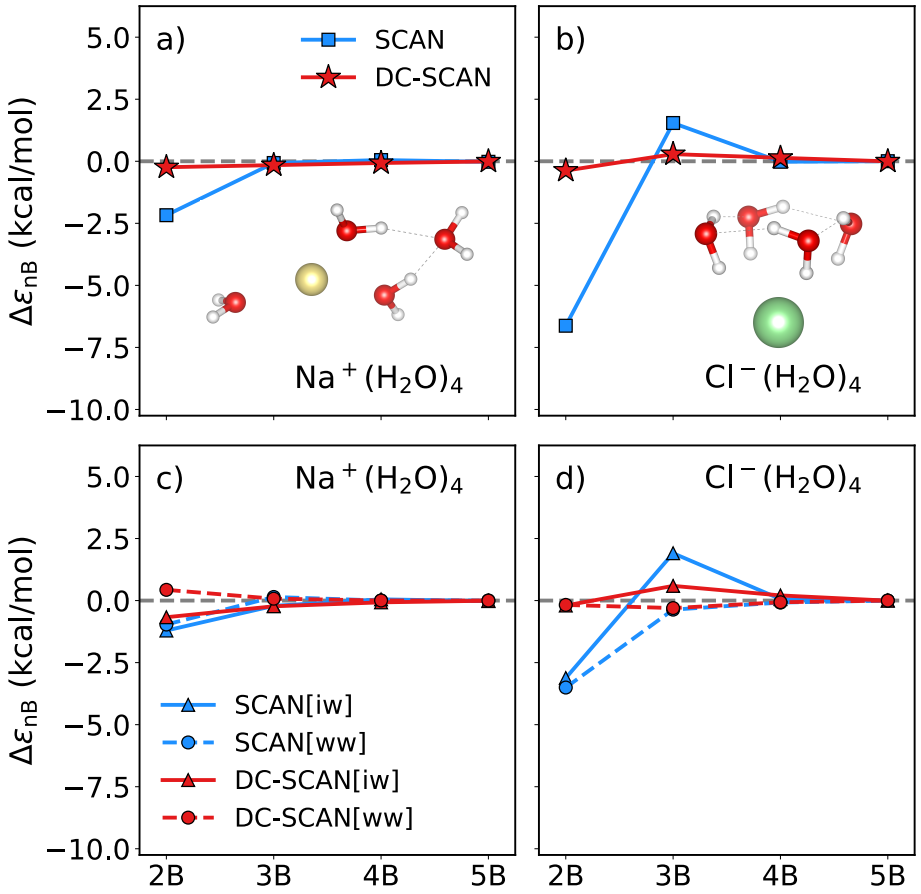

Figure 5.3. Errors (in kcal/mol) associated with individual n-body contributions to the interaction energy of the low-lying energy isomers $Na^+(H_2O)_4$ (panel a), and $Cl^-(H_2O)_4$ (panel b) A detailed breakdown of the contributions of water-water and ion-water interactions to the n-body energies is shown for $Na^+(H_2O)_4$ (panel c), and $Cl^-(H_2O)_4$ (panel d). The errors are shown relative to the MB-nrg reference energies (in kcal/mol).

to small molecular systems. One particularly successful approach involves leveraging the MBE of the energy; this approach has demonstrated its ability to accurately model various aqueous systems. In our theoretical/computational framework, MB-SCAN and MB-SAN(DC) PEFs of generic molecules approximate the MBE defined in Eq. 5.3 as

$$E_N = V^{1B} + V^{2B} + V^{3B} + V_{elec} \tag{5.4}$$

where each of the V^{nB} terms of the MB-DFT and MB-nrg PEFs includes an n-body data-driven term V^{nB}_{PIP} for each n-mer. Specifically, V^{1B} is the monomer distortion energy,

Table 5.1. Interaction energy comparison between MB-nrg, DC-SCAN and SCAN for the $Na^+(H_2O)_4$ and $Cl^-(H_2O)_4$ clusters represented in the n-body analysis shown in Fig. 5.3. The energy values are given in kcal/mol.

	$Na^+(H_2O)_4$			$Cl^-(H_2O)_4$		
	MB-nrg	DC-SCAN	SCAN	MB-nrg	DC-SCAN	SCAN
2B	-83.673	-83.913	-85.846	-65.160	-65.535	-71.786
3B	4.266	4.111	4.197	3.808	4.096	5.350
4B	0.361	0.291	0.411	-0.713	-0.574	-0.733
5B	-0.035	-0.040	-0.056	0.037	0.033	0.041
E_{int}	-79.081	-79.551	-81.294	-62.028	-61.979	-67.128

such that $V^{1B} = 0$ for monatomic ions, and it can be represented either by the Partridge-Schwencke PEF[185] or the MB-SCAN, MB-SCAN(DC) PEFs for the water monomer[153] in MB-MD simulations. In this context, V^{2B} and V^{3B} MB-SCAN and MB-SCAN(DC) PEFs are introduced here for the ion-water interactions. Note that MB-nrg PEFs inherently account for long-range dispersion, such that V^{1B} and V^{2B} are defined as $V^{1B}_{\text{PIP}} + V^{1B}_{\text{disp}}$ and $V^{2B}_{\text{PIP}} + V^{2B}_{\text{disp}}$, respectively,[107,108,113,115] accounting for the contribution to the long-range dispersion that is missing in SCAN and DC-SCAN.[392,393] Finally, in Eq. 5.4, V_{elec} contains terms explicitly treating charge-charge, charge-dipole, dipole-dipole interactions, and the polarization energy.[394] A comprehensive description of the MB-DFT and MB-nrg framework, including a breakdown of the individual terms in MB PEFs can be found in references 346 and 394, respectively. Further details on the computational details and fitting procedure are reported in the Supporting Information.

While previous studies have focused on the effect of density-corrected DFAs on water-water interactions,[153,378] in this work we isolate the ion-water contribution during simulation by using MB-pol as water-water interaction model. Fig. 5.4 shows the RDFs with respect to the sodium-oxygen and chloride-oxygen distances obtained via bulk-phase MD simulations in the NPT ensemble at 298 K and 1 atm using the MB-SCAN and MB-SCAN(DC) models and including n-body corrections of the MB-nrg framework up to the 3-body term in the top panels and up to the 2-body term in the bottom panels. All

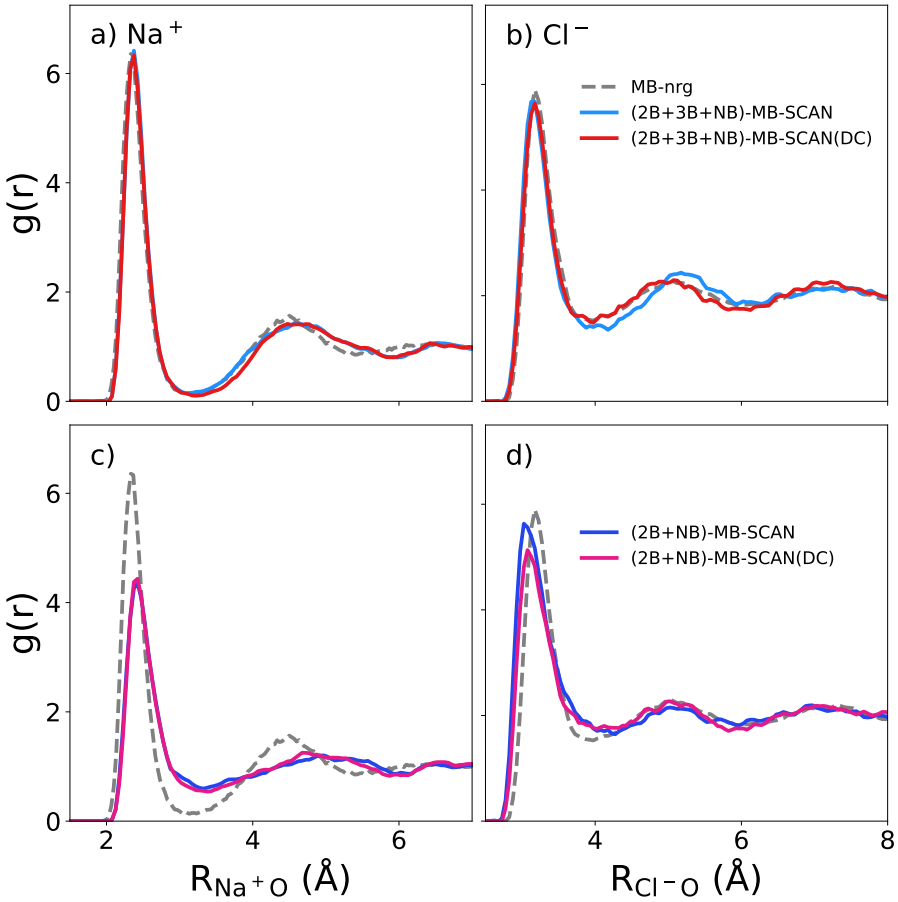

Figure 5.4. Sodium-oxygen (left panels) and chloride-oxygen (right panels) radial distribution functions, $g(r)$, calculated from NPT simulations carried out at 298 K and 1 atm with the MB-SCAN (blue, purple), MB-SCAN(DC) (red, magenta), and MB-nrg (gray) PEFs using both 2-body and 3-body corrections (top panels) and using only 2-body corrections (bottom panels).

the simulation details are reported in the Supporting Information.

Sodium and chloride ions are expected to exhibit significantly different solvation behavior when simulated using SCAN- and DC-SCAN-based MB-nrg PEFs, given the large difference in error shown in the gas-phase analysis of Fig. 5.3 between the energies predicted with the two functionals. Both systems show identical structural features in the first solvation shell using the two models; when compared with the reference interaction energy of the sodium- and chloride-water dimer configurations reported in Refs. 41 and 42 of −24.09 and −15.58 kcal/mol, respectively, it is clear that, when accounting for all

the ion–water pairwise interactions, the total density correction becomes negligible with respect to the dimer interaction energy of the bulk system. In the second solvation shell, however, the number of 3-body contributions becomes significant. In the case of sodium, the small difference in the 3-body effect of the density correction, translates to effectively identical bands. On the other hand, the relevant SCAN 3-body error of the chloride ion in Fig. 5.3 has a repulsive effect on the overall hydration structure and is directly related to a shift of the second solvation shell to larger distances. This is substantiated by the 2-body corrected MB-nrg models, whose RDFs are reported in panels c) and d). While sodium shows superimposing RDFs for both MB-SCAN and MB-SCAN(DC), as expected, given the small 3-body contribution, the same is true in the case of chloride. The small difference in intensity of the first solvation shell can be attributed to a stronger overall 3-body effect, considering the greater dimension of the hydration complex.

In this work, we have explored the contribution of density driven errors to the description of hydration phenomena of monatomic, singly-charged ions, and reported a systematic analysis of SCAN and DC-SCAN functionals applied to the hydration of sodium and chloride ions. The analyses of density sensitivity, interaction and individual many-body energies of hydrated ions further demonstrate that DC-SCAN is a robust tool for the description of gas-phase clusters of non-covalent systems with non-homogenous charge densities, providing an accuracy that goes beyond that of pure DFT. By effectively correcting the delocalization errors in SCAN calculations of ion-water systems using the Hartree-Fock density, DC-SCAN allows DFT to closely approach the accuracy of explicitly-correlated wavefunction-based *ab initio* methods. The almost constant density sensitivity per water molecule illustrates the weight of delocalization error in aqueous solutions of ionic systems, determining a linear size-dependent misrepresentation of the electronic density, and thus the interaction energy. In this spirit, we rationalize the effect that a localized density, free of self-interaction error, plays in the prediction of interaction energies in hydration complexes of increasing sizes. Our findings manifest the proportional correlation

between interaction energy error and delocalization error. Correcting for density-driven errors allows the DC-SCAN method to provide consistent interaction energies with minimal loss of accuracy when compared to reference MB-nrg values, trained on CCSD(T) reference energies, as a function of system size. In addition, many-body analyses of the SCAN and DC-SCAN interaction energies of sodium- and chloride-water clusters show that DC-SCAN predicts 2- and 3-body energies with chemical accuracy when compared to MB-nrg reference values. Thus, DC-SCAN greatly mitigates the many-body energy errors identified in SCAN calculations for both sodium and chloride. It should be emphasized that DC-SCAN achieves remarkable accuracy in 3-body contributions, crucial to the faithful representation of chloride-water interactions, overcoming the significant deviations displayed by SCAN. As the system size approaches the macroscopic limit, the propagation of the density-driven errors is examined by means of MB-MD simulations and RDF analyses. Here, the density delocalization of the SCAN functional introduces spurious repulsions at the 3-body level that affect the topology of the ion-water-water network of the chloride ion, resulting in a larger second solvation shell. This effect appears to be negligible at shorter distances due to the small number of 3-body contributions and the small absolute value of 2-body errors with respect to the total 2-body energy of the system. In turn, DC-SCAN accurately reproduces the expected behavior across all distances, which attests to the robustness of the SCAN non-empirical functional form, upon correction of density-driven errors. These results suggest that the usage of accurate functional forms as in SCAN, together with better system densities, provides a more rigorous representation of the overall ground-state properties of hydrated ions. The significant quality of the physical properties retrieved with DC-SCAN warrants further investigation in its applicability for the representation of generic molecular systems.

Acknowledgements

Chapter 5, in full is currently being prepared for submission for publication of the material. E. Palos, A. Caruso, and F. Paesani. The dissertation author is the co-primary investigator and author of this material.

We thank Saswata Dasgupta, Eleftherios Lambros, Marc Riera, Henry Agnew, Jie Hu, and Steven Swee for useful discussions. This research was supported by the U.S. Department of Energy, Office of Science, Office of Basic Energy Science, through grant no. DE-SC0019490. This research used resources of 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, the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by the National Science Foundation grant number ACI-1548562, and the Triton Shared Computing Cluster (TSCC) at the San Diego Supercomputer Center (SDSC). E.P. acknowledges support from the National Science Foundation (NSF) Graduate Research Fellowship Program (GRFP) under Grant No. DGE-2038238, as well as the Alfred P. Sloan Foundation Ph.D. Fellowship Program under Grant No. G-2020-14067.

Chapter 6

Conclusions

This dissertation dove into the design of new computational frameworks for modelling and understanding the physical and chemical properties of ionic systems in aqueous environments, with a specific focus on the development and application of many-body (MB) interaction potentials for hydrated halides.

We investigated the performance and transferability of a relatively new family of interaction models: deep neural network (DNN) potentials. Taking water as benchmark system, DeePMD-based models were trained liquid water configurations at various thermodynamic conditions, using the highly accurate MB-pol reference. The DNN potentials were found to accurately reproduce structural and thermodynamic properties of liquid water predicted by MB-pol from the boiling point to deeply supercooled temperatures. However, the trained models lack the high level of transferability between phases that MB-pol possesses; they struggle to accurately describe vapor-liquid equilibrium properties and many-body interactions, tending to rely on error compensation among individual many-body energy contributions. While an improved description of vapor-liquid equilibrium (VLE) properties and a better description of individual many-body contributions to the interaction energy was achieved by training on additional VLE configurations and gas-phase clusters, respectively, obtaining a consistently accurate description of water across different phases was not possible. DNN potentials are inherently many-body, but they may

not correctly represent the underlying many-body physics, limiting their transferability over a wide range of thermodynamic conditions.

Among state-of-the-art interaction models, MB potentials are currently a compelling alternative to neural networks; in this regard, the MB-nrg family of interaction potentials provides transferable models by design, explicitly embedding the many-body framework and retrieving the total energy bottom-up, from the many-body expansion (MBE) of the energy. For the construction of such potential energy functions (PEFs), we have developed an active learning (AL) algorithm for generating representative training sets that works in synergy with the present MB-nrg workflow. The AL framework estimates the error on the MB-nrg model and the energy uncertainty through Gaussian process (GP) regression, and assigns a probability to configurations in a large pool generated through scans along relevant collective variables, normal mode sampling, and molecular dynamics (MD) simulations. The learner self-consistently selects configurations from the pool and adds them to the training set, until convergence is reached. The framework was tested on the cesium-water system as a case study; the development of the MB-nrg PEF has enabled efficient identification of the most relevant configurations necessary for accurately representing the target many-body potential energy surfaces (PESs), resulting in significantly smaller training sets than those needed for the development of the original MB-nrg PEF while preserving the accuracy.

Considering the computational expense of reference coupled cluster with single, double and perturbative triple excitations (CCSD(T)) calculations of individual many-body energies, the AL framework was used in the generation of the new MB-nrg PEFs for halide-water interaction. In particular, dimer and trimer training sets of chloride, bromide, and iodide ions with water have been used to build explicit 2-body and 3-body polynomials. Although the MB-nrg PEFs are trained only on gas-phase clusters, they successfully reproduce the interaction energies and solvation structure of hydrated halides from the gas- to the condensed- phase, as the resulting extended x-ray absorption fine structure (EXAFS)

spectra closely align with experimental data and accurately reproduce amplitude and phase of the EXAFS oscillations, outperforming the existing TTM-nrg framework and popular empirical force fields based on the TIP4P/Ew water model. With regards to chloride, pairwise additive representations of ion-water and water-water interactions are insufficient for accurately depicting chloride hydration structure in gas-phase clusters and solution; these models underestimate interaction strengths in the former and predict an overly tight first hydration shell in the latter. While classical many-body polarization significantly improves the description of chloride-water interactions, TTM-nrg still shows limitations when compared to more accurate MB-nrg models; these offer a quantitative representation of the halide ion hydration shell structure and energetics from gas to condensed phase, and enable cost-effective MD simulations of ionic aqueous solutions with chemical accuracy.

For relatively small systems, CCSD(T) reference values allow to generate accurate MB potentials to be used in MD simulations. However, for larger systems and framework that require bulk configurations during the training procedure, as DNNs, we must rely on alternative *ab initio* methods, as density functional theory (DFT). We have shown that for monatomic, singly-charged ions as Na^+ and Cl^-, semi-local density functional approximations (DFAs) as SCAN suffer from large density-driven error that can compromise the accuracy of the calculations. We have shown how, analogously to pure water, extracting SCAN energies from Hartree-Fock densities greatly corrects this behavior. DC-SCAN serves as a robust tool for modeling gas-phase clusters of non-covalent systems with non-homogeneous charge densities, outperforming pure DFT in terms of accuracy and approaching explicitly-correlated wavefunction-based *ab initio* methods. The limitation of non-corrected semi-local DFAs propagates to the macroscopic limit; by constructing MB-SCAN and MB-SCAN(DC) models, we have explored the structural differences in the radial distribution functions (RDFs) of the two ions: SCAN's density delocalization introduces spurious repulsions in chloride's hydration complex, resulting in a larger second solvation shell. On the other hand, DC-SCAN accurately reproduces the expected behavior

across all distances, demonstrating the robustness of the SCAN non-empirical functional form upon correction of density-driven errors. This suggests that combining accurate functional forms like SCAN with improved system densities can provide a more rigorous representation of the overall ground-state properties of hydrated ions.

The findings herein demonstrated the potential and limitations of state-of-the-art interaction models. Advanced computational frameworks as MB-nrg can accurate describe and predict the behavior of ionic systems in aqueous environments; the models that have been presented provide an important foundation for future work, not only for further refinement and optimization, but also for their application to a broader range of phenomena. It is hoped that this work will stimulate the exploration of air-water interfaces, the calculation of the surface propensity of halide ions, and their behaviour at finite concentration.

Bibliography

[1] R. A. Sneen, "Substitution at a saturated carbon atom. XVII. Organic ion pairs as intermediates in nucleophilic substitution and elimination reactions," *Acc. Chem. Res.*, vol. 6, no. 2, pp. 46–53, 1973.

[2] M. Pregel, E. Dunn, R. Nagelkerke, G. Thatcher, and E. Buncel, "Alkali-metal ion catalysis and inhibition in nucleophilic displacement reaction of phosphorus-, sulfur-, and carbon-based esters," *Chem. Soc. Rev.*, vol. 24, no. 6, pp. 449–455, 1995.

[3] S. Nahar and H. Tajmir-Riahi, "Do metal ions alter the protein secondary structure of a light-harvesting complex of thylakoid membranes?," *J. Inorg. Biochem.*, vol. 58, no. 3, pp. 223–234, 1995.

[4] S. A. Woodson, "Metal ions and rna folding: A highly charged topic with a dynamic future," *Curr. Opin. Chem. Biol.*, vol. 9, no. 2, pp. 104–109, 2005.

[5] D. E. Draper, "Rna folding: Thermodynamic and molecular descriptions of the roles of ions," *Biophys. J.*, vol. 95, no. 12, pp. 5489–5495, 2008.

[6] G. Siuzdak, Y. Ichikawa, T. J. Caulfield, B. Munoz, C. H. Wong, and K. Nicolaou, "Evidence of calcium (2+)-dependent carbohydrate association through ion spray mass spectrometry," *J. Am. Chem. Soc.*, vol. 115, no. 7, pp. 2877–2881, 1993.

[7] Z.-J. Tan and S.-J. Chen, "Nucleic acid helix stability: effects of salt concentration, cation valence and size, and chain length," *Biophys. J.*, vol. 90, no. 4, pp. 1175–1190, 2006.

[8] R. G. Harrison and H. Tammet, "Ions in the terrestrial atmosphere and other solar system atmospheres," *Space Sci. Rev.*, vol. 137, no. 1-4, pp. 107–118, 2008.

[9] D. J. Tobias, A. C. Stern, M. D. Baer, Y. Levin, and C. J. Mundy, "Simulation and theory of ions at atmospherically relevant aqueous liquid-air interfaces," *Annu. Rev. Phys. Chem.*, vol. 64, pp. 339–359, 2013.

[10] N. S. Shuman, D. E. Hunton, and A. A. Viggiano, "Ambient and modified atmospheric

ion chemistry: From top to bottom," *Chem. Rev.*, vol. 115, no. 10, pp. 4542–4570, 2015.

[11] K. Lehtipalo, L. Rondo, J. Kontkanen, S. Schobesberger, T. Jokinen, N. Sarnela, A. Kürten, S. Ehrhart, A. Franchin, T. Nieminen, F. Riccobono, M. Sipilä, T. Yli-Juuti, J. Duplissy, A. Adamov, L. Ahlm, J. a. Almeida, A. Amorim, F. Bianchi, M. Breitenlechner, J. Dommen, A. J. Downard, E. M. Dunne, R. C. Flagan, R. Guida, J. Hakala, A. Hansel, W. Jud, J. Kangasluoma, V.-M. Kerminen, H. Keskinen, J. Kim, J. Kirkby, A. Kupc, O. Kupiainen-Määttä, A. Laaksonen, M. J. Lawler, M. Leiminger, S. Mathot, T. Olenius, I. K. Ortega, A. Onnela, T. Petäjä, A. Praplan, M. P. Rissanen, T. Ruuskanen, F. D. Santos, S. Schallhart, R. Schnitzhofer, M. Simon, J. N. Smith, G. Tröstl, Jasmin Tsagkogeorgas, A. Tomé, P. Vaattovaara, H. Vehkamäki, A. E. Vrtala, P. E. Wagner, C. Williamson, D. Wimmer, P. M. Winkler, A. Virtanen, N. M. Donahue, K. S. Carslaw, U. Baltensperger, I. Riipinen, J. Curtius, D. R. Worsnop, and M. Kulmala, "The effect of acid–base clustering and ions on the growth of atmospheric nano-particles," *Nat. Commun.*, vol. 7, no. 1, pp. 1–9, 2016.

[12] M. Winter and R. J. Brodd, "What are batteries, fuel cells, and supercapacitors?," *Chem. Rev.*, vol. 104, no. 10, pp. 4245–4270, 2004.

[13] K. D. Collins and M. W. Washabaugh, "The Hofmeister effect and the behaviour of water at interfaces," *Q. Rev. Biophys.*, vol. 18, no. 4, pp. 323–422, 1985.

[14] C. D. Cappa, J. D. Smith, K. R. Wilson, B. M. Messer, M. K. Gilles, R. C. Cohen, and R. J. Saykally, "Effects of alkali metal halide salts on the hydrogen bond network of liquid water," *J. Phys. Chem. B*, vol. 109, no. 15, pp. 7046–7052, 2005.

[15] J. D. Smith, R. J. Saykally, and P. L. Geissler, "The effects of dissolved halide anions on hydrogen bonding in liquid water," *J. Am. Chem. Soc.*, vol. 129, no. 45, pp. 13847–13856, 2007.

[16] D. A. Schmidt, Ö. Birer, S. Funkner, B. P. Born, R. Gnanasekaran, G. W. Schwaab, D. M. Leitner, and M. Havenith, "Rattling in the cage: Ions as probes of sub-picosecond water network dynamics," *J. Am. Chem. Soc.*, vol. 131, no. 51, pp. 18512–18517, 2009.

[17] S. Funkner, G. Niehues, D. A. Schmidt, M. Heyden, G. Schwaab, K. M. Callahan, D. J. Tobias, and M. Havenith, "Watching the low-frequency motions in aqueous salt solutions: The terahertz vibrational signatures of hydrated ions," *J. Am. Chem. Soc.*, vol. 134, no. 2, pp. 1030–1035, 2012.

[18] I. Waluyo, D. Nordlund, U. Bergmann, D. Schlesinger, L. G. Pettersson, and A. Nilsson, "A different view of structure-making and structure-breaking in alkali halide aqueous solutions through X-ray absorption spectroscopy," *J. Chem. Phys.*,

vol. 140, no. 24, p. 244506, 2014.

[19] A. Heydweiller, "Über physikalische eigenschaften von lösungen in ihrem zusammenhang. ii. oberflächenspannung und elektrisches leitvermögen wässeriger salzlösungen," *Ann. Phys.*, vol. 4, no. 33, pp. 145–185, 1910.

[20] C. Wagner, "Die oberflächenspannung verdünnter elektrolytlösungen," *Phys. Z.*, vol. 25, p. 474, 1924.

[21] L. Onsager and N. N. Samaras, "The surface tension of debye-hückel electrolytes," *J. Chem. Phys.*, vol. 2, no. 8, pp. 528–536, 1934.

[22] A. Frumkin, "Phasengrenzkräfte und adsorption an der trennungsfläche luft: Lösung anorganischer elektrolyte," *Z. Phys. Chem.*, vol. 109, pp. 34–48, 1924.

[23] D. Liu, G. Ma, L. M. Levering, and H. C. Allen, "Vibrational spectroscopy of aqueous sodium halide solutions and air-liquid interfaces: Observation of increased interfacial depth," *J. Phys. Chem. B*, vol. 108, no. 7, pp. 2252–2260, 2004.

[24] E. A. Raymond and G. L. Richmond, "Probing the molecular structure and bonding of the surface of aqueous salt solutions," *J. Phys. Chem. B*, vol. 108, no. 16, pp. 5051–5059, 2004.

[25] P. B. Petersen and R. J. Saykally, "Confirmation of enhanced anion concentration at the liquid water surface," *Chem. Phys. Lett.*, vol. 397, no. 1-3, pp. 51–55, 2004.

[26] S. J. Stuart and B. Berne, "Effects of polarizability on the hydration of the chloride ion," *J. Phys. Chem.*, vol. 100, no. 29, pp. 11934–11943, 1996.

[27] S. J. Stuart and B. Berne, "Surface curvature effects in the aqueous ionic solvation of the chloride ion," *J. Phys. Chem. A*, vol. 103, no. 49, pp. 10300–10307, 1999.

[28] L. Perera and M. L. Berkowitz, "Many-body effects in molecular dynamics simulations of $Na^+(H_2O)_n$ and $Cl^-(H_2O)_n$ clusters," *J. Chem. Phys.*, vol. 95, no. 3, pp. 1954–1963, 1991.

[29] L. Perera and M. L. Berkowitz, "Structure and dynamics of $Cl^-(H_2O)_{20}$ clusters: The effect of the polarizability and the charge of the ion," *J. Chem. Phys.*, vol. 96, no. 11, pp. 8288–8294, 1992.

[30] L. Perera and M. L. Berkowitz, "Stabilization energies of Cl^-, Br^-, and I^- ions in water clusters," *J. Chem. Phys.*, vol. 99, no. 5, pp. 4222–4224, 1993.

[31] L. Perera and M. L. Berkowitz, "Structures of $Cl^-(H_2O)_n$ and $F^-(H_2O)_n$ (n= 2, 3,...,

15) clusters. molecular dynamics computer simulations," *J. Chem. Phys.*, vol. 100, no. 4, pp. 3085–3093, 1994.

[32] L. X. Dang and D. E. Smith, "Molecular dynamics simulations of aqueous ionic clusters using polarizable water," *J. Chem. Phys.*, vol. 99, no. 9, pp. 6950–6956, 1993.

[33] P. Jungwirth and D. J. Tobias, "Molecular structure of salt solutions: A new view of the interface with implications for heterogeneous atmospheric chemistry," *J. Phys. Chem. B*, vol. 105, no. 43, pp. 10468–10472, 2001.

[34] P. Jungwirth and D. J. Tobias, "Ions at the air/water interface," *J. Phys. Chem. B*, vol. 106, no. 25, pp. 6361–6373, 2002.

[35] L. X. Dang, "Computational study of ion binding to the liquid interface of water," *J. Phys. Chem. B*, vol. 106, no. 40, pp. 10388–10394, 2002.

[36] L. X. Dang and T.-M. Chang, "Molecular mechanism of ion binding to the liquid/vapor interface of water," *J. Phys. Chem. B*, vol. 106, no. 2, pp. 235–238, 2002.

[37] C. D. Wick, I.-F. W. Kuo, C. J. Mundy, and L. X. Dang, "The effect of polarizability for understanding the molecular structure of aqueous interfaces," *J. Chem. Theory Comput.*, vol. 3, no. 6, pp. 2002–2010, 2007.

[38] Y. Levin, A. P. Dos Santos, and A. Diehl, "Ions at the air-water interface: An end to a hundred-year-old mystery?," *Phys. Rev. Lett.*, vol. 103, no. 25, p. 257802, 2009.

[39] Y. Levin, "Polarizable ions at interfaces," *Phys. Rev. Lett.*, vol. 102, no. 14, p. 147803, 2009.

[40] F. Paesani, P. Bajaj, and M. Riera, "Chemical accuracy in modeling halide ion hydration from many-body representations," *Adv. Phys. X*, vol. 4, no. 1, p. 1631212, 2019.

[41] 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.*, vol. 15, no. 5, pp. 2983–2995, 2019.

[42] 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.*, vol. 16, no. 5, pp. 3055–3072, 2020.

[43] T. E. Gartner III, K. M. Hunter, E. Lambros, A. Caruso, M. Riera, G. R. Medders,

A. Z. Panagiotopoulos, P. G. Debenedetti, and F. Paesani, "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.*, vol. 13, no. 16, pp. 3652–3658, 2022.

[44] S. L. Bore and F. Paesani, "Quantum phase diagram of water." ChemRxiv, https://doi.org/10.26434/chemrxiv-2023-kmmmz, 2023.

[45] J. Pople, R. Krishnan, H. Schlegel, and J. Binkley, "Electron correlation theories and their application to the study of simple reaction potential surfaces," *Int. J. Quantum Chem.*, vol. 14, no. 5, pp. 545–560, 1978.

[46] G. D. Purvis III and R. J. Bartlett, "A full coupled-cluster singles and doubles model: The inclusion of disconnected triples," *J. Chem. Phys.*, vol. 76, no. 4, pp. 1910–1918, 1982.

[47] J. A. Pople, R. Seeger, and R. Krishnan, "Variational configuration interaction methods and comparison with perturbation theory," *Int. J. Quantum Chem.*, vol. 12, no. S11, pp. 149–163, 1977.

[48] C. Møller and M. S. Plesset, "Note on an approximation treatment for many-electron systems," *Phys. Rev.*, vol. 46, no. 7, p. 618, 1934.

[49] M. Head-Gordon, J. A. Pople, and M. J. Frisch, "Mp2 energy evaluation by direct methods," *Chem. Phys. Lett.*, vol. 153, no. 6, pp. 503–506, 1988.

[50] P. Hohenberg and W. Kohn, "Inhomogeneous electron gas," *Phys. Rev.*, vol. 136, no. 3B, p. B864, 1964.

[51] W. Kohn and L. J. Sham, "Self-consistent equations including exchange and correlation effects," *Phys. Rev.*, vol. 140, no. 4A, p. A1133, 1965.

[52] R. G. Parr, "Density functional theory of atoms and molecules," in *Horizons of Quantum Chemistry*, pp. 5–15, Springer, 1980.

[53] N. Mardirossian and M. Head-Gordon, "Thirty years of density functional theory in computational chemistry: an overview and extensive assessment of 200 density functionals," *Molecular Physics*, vol. 115, no. 19, pp. 2315–2372, 2017.

[54] A. Stone, *The Theory of Intermolecular Forces*. Oxford University Press, 2013.

[55] H. J. Berendsen, J. P. Postma, W. F. van Gunsteren, and J. Hermans, "Interaction models for water in relation to protein hydration," in *Intermolecular forces: proceedings of the fourteenth Jerusalem symposium on quantum chemistry and biochemistry*

held in jerusalem, israel, april 13–16, 1981, pp. 331–342, Springer, 1981.

[56] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, "Comparison of simple potential functions for simulating liquid water," *J. Chem. Phys.*, vol. 79, no. 2, pp. 926–935, 1983.

[57] H. Berendsen, J. Grigera, and T. Straatsma, "The missing term in effective pair potentials," *J. Phys. Chem.*, vol. 91, no. 24, pp. 6269–6271, 1987.

[58] L. X. Dang and B. M. Pettitt, "Simple intramolecular model potentials for water," *J. Phys. Chem.*, vol. 91, no. 12, pp. 3349–3354, 1987.

[59] D. M. Ferguson, "Parameterization and evaluation of a flexible water model," *J. Comput. Chem.*, vol. 16, no. 4, pp. 501–511, 1995.

[60] M. W. Mahoney and W. L. Jorgensen, "A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions," *J. Chem. Phys.*, vol. 112, no. 20, pp. 8910–8922, 2000.

[61] H. W. Horn, W. C. Swope, J. W. Pitera, J. D. Madura, T. J. Dick, G. L. Hura, and T. Head-Gordon, "Development of an improved four-site water model for biomolecular simulations: Tip4p-ew," *J. Chem. Phys.*, vol. 120, no. 20, pp. 9665–9678, 2004.

[62] Y. Wu, H. L. Tepper, and G. A. Voth, "Flexible simple point-charge water model with improved liquid-state properties," *J. Chem. Phys.*, vol. 124, no. 2, p. 024503, 2006.

[63] F. Paesani, W. Zhang, D. A. Case, T. E. Cheatham III, and G. A. Voth, "An accurate and simple quantum model for liquid water," *J. Chem. Phys.*, vol. 125, no. 18, p. 184507, 2006.

[64] S. Habershon, T. E. Markland, and D. E. Manolopoulos, "Competing quantum effects in the dynamics of a flexible water model," *J. Chem. Phys.*, vol. 131, no. 2, p. 024501, 2009.

[65] K. Park, W. Lin, and F. Paesani, "A refined ms-evb model for proton transport in aqueous environments," *J. Phys. Chem. B*, vol. 116, no. 1, pp. 343–352, 2012.

[66] I. S. Joung and T. E. Cheatham III, "Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations," *J. Phys. Chem. B*, vol. 112, no. 30, pp. 9020–9041, 2008.

[67] T. P. Lybrand and P. A. Kollman, "Water–water and water–ion potential functions including terms for many body effects," *J. Chem. Phys.*, vol. 83, no. 6, pp. 2923–2933,

1985.

[68] J. Behler and M. Parrinello, "Generalized neural-network representation of high-dimensional potential-energy surfaces," *Phys. Rev. Lett.*, vol. 98, no. 14, p. 146401, 2007.

[69] J. Behler, "Neural network potential-energy surfaces in chemistry: A tool for large-scale simulations," *Phys. Chem. Chem. Phys*, vol. 13, no. 40, pp. 17930–17955, 2011.

[70] J. Behler, "Representing potential energy surfaces by high-dimensional neural network potentials," *J. Phys. Condens. Matter*, vol. 26, no. 18, p. 183001, 2014.

[71] K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O. A. Von Lilienfeld, K.-R. Müller, and A. Tkatchenko, "Machine learning predictions of molecular properties: Accurate many-body potentials and nonlocality in chemical space," *J. Phys. Chem. Lett*, vol. 6, no. 12, pp. 2326–2331, 2015.

[72] J. Behler, "First principles neural network potentials for reactive simulations of large molecular and condensed systems," *Angew. Chem. Int. Ed.*, vol. 56, no. 42, pp. 12828–12840, 2017.

[73] 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.*, vol. 3, no. 5, p. e1603015, 2017.

[74] 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.*, vol. 8, no. 4, pp. 3192–3203, 2017.

[75] J. S. Smith, O. Isayev, and A. E. Roitberg, "ANI-1, a data set of 20 million calculated off-equilibrium conformations for organic molecules," *Sci. Data*, vol. 4, no. 1, pp. 1–8, 2017.

[76] 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.*, vol. 120, no. 14, p. 143001, 2018.

[77] B. Cheng, E. A. Engel, J. Behler, C. Dellago, and M. Ceriotti, "Ab initio thermodynamics of liquid and solid water," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 116, no. 4, pp. 1110–1115, 2019.

[78] 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., vol. 16, no. 1, pp. 88–99, 2019.

[79] M. Haghighatlari and J. Hachmann, "Advances of machine learning in molecular modeling and simulation," *Curr. Opin. Chem. Eng.*, vol. 23, pp. 51–57, 2019.

[80] O. Wohlfahrt, C. Dellago, and M. Sega, "Ab initio structure and thermodynamics of the rpbe-d3 water/vapor interface by neural-network molecular dynamics," *J. Chem. Phys.*, vol. 153, no. 14, p. 144710, 2020.

[81] C. Schran, K. Brezina, and O. Marsalek, "Committee neural network potentials control generalization errors and enable active learning," *J. Chem. Phys.*, vol. 153, no. 10, p. 104105, 2020.

[82] S. Manzhos and T. Carrington Jr., "Neural network potential energy surfaces for small molecules and reactions," *Chem. Rev.*, vol. 121, no. 16, pp. 10187–10217, 2020.

[83] P. Gkeka, G. Stoltz, A. Barati Farimani, Z. Belkacemi, M. Ceriotti, J. D. Chodera, A. R. Dinner, A. L. Ferguson, J.-B. Maillet, H. Minoux, C. Peter, F. Pietrucci, A. Silveira, A. Tkatchenko, Z. Trstanova, R. Wiewiora, and T. Lelièvre, "Machine learning force fields and coarse-grained variables in molecular dynamics: Application to materials and biological systems," *J. Chem. Theory Comput.*, vol. 16, no. 8, pp. 4757–4775, 2020.

[84] Z. L. Glick, D. P. Metcalf, A. Koutsoukas, S. A. Spronk, D. L. Cheney, and C. D. Sherrill, "Ap-net: An atomic-pairwise neural network for smooth and transferable interaction potentials," *J. Chem. Phys.*, vol. 153, no. 4, p. 044112, 2020.

[85] C. Schran, F. L. Thiemann, P. Rowe, E. A. Müller, O. Marsalek, and A. Michaelides, "Machine learning potentials for complex aqueous systems made simple," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 118, no. 38, p. e2110077118, 2021.

[86] T. W. Ko, J. A. Finkler, S. Goedecker, and J. Behler, "A fourth-generation high-dimensional neural network potential with accurate electrostatics including non-local charge transfer," *Nat. Commun.*, vol. 12, no. 1, pp. 1–11, 2021.

[87] J. Behler, "Four generations of high-dimensional neural network potentials," *Chem. Rev.*, vol. 121, no. 16, pp. 10037–10072, 2021.

[88] O. T. Unke, S. Chmiela, H. E. Sauceda, M. Gastegger, I. Poltavsky, K. T. Schütt, A. Tkatchenko, and K.-R. Müller, "Machine learning force fields," *Chem. Rev.*, vol. 121, no. 16, pp. 10142–10186, 2021.

[89] T. Zubatiuk and O. Isayev, "Development of multimodal machine learning potentials: Toward a physics-aware artificial intelligence," *Acc. Chem. Res.*, vol. 54, no. 7,

pp. 1575–1585, 2021.

[90] V. Zaverkin, D. Holzmüller, R. Schuldt, and J. Kästner, "Predicting properties of periodic systems from cluster data: A case study of liquid water," *J. Chem. Phys.*, vol. 156, no. 11, p. 114103, 2022.

[91] L. D. Jacobson, J. M. Stevenson, F. Ramezanghorbani, D. Ghoreishi, K. Leswing, E. D. Harder, and R. Abel, "Transferable neural network potential energy surfaces for closed-shell organic molecules: Extension to ions," *J. Chem. Theory Comput.*, vol. 18, no. 4, pp. 2354–2366, 2022.

[92] S. Batzner, A. Musaelian, L. Sun, M. Geiger, J. P. Mailoa, M. Kornbluth, N. Molinari, T. E. Smidt, and B. Kozinsky, "E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials," *Nat. Commun.*, vol. 13, no. 1, pp. 1–11, 2022.

[93] Y. Zhai, A. Caruso, S. L. Bore, Z. Luo, and F. Paesani, "A "short blanket" dilemma for a state-of-the-art neural network potential for water: Reproducing experimental properties or the physics of the underlying many-body interactions?," *J. Chem. Phys.*, vol. 158, no. 8, p. 084111, 2023.

[94] R. Bukowski, K. Szalewicz, G. C. Groenenboom, and A. Van der Avoird, "Predictions of the properties of water from first principles," *Science*, vol. 315, no. 5816, pp. 1249–1252, 2007.

[95] 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.*, vol. 134, no. 9, p. 094509, 2011.

[96] V. Babin, G. R. Medders, and F. Paesani, "Toward a universal water model: First principles simulations from the dimer to the liquid phase," *J. Phys. Chem. Lett.*, vol. 3, no. 24, pp. 3765–3769, 2012.

[97] 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.*, vol. 9, no. 12, pp. 5395–5403, 2013.

[98] 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.*, vol. 10, no. 4, pp. 1599–1607, 2014.

[99] G. R. Medders, V. Babin, and F. Paesani, "A critical assessment of two-body

and three-body interactions in water," *J. Chem. Theory Comput.*, vol. 9, no. 2, pp. 1103–1114, 2013.

[100] R. Bukowski, K. Szalewicz, G. C. Groenenboom, and A. van der Avoird, "Polarizable interaction potential for water from coupled cluster calculations. I. Analysis of dimer potential energy surface," *J. Chem. Phys.*, vol. 128, no. 9, p. 094313, 2008.

[101] R. Bukowski, K. Szalewicz, G. C. Groenenboom, and A. van der Avoird, "Polarizable interaction potential for water from coupled cluster calculations. II. Applications to dimer spectra, virial coefficients, and simulations of liquid water," *J. Chem. Phys.*, vol. 128, no. 9, p. 094314, 2008.

[102] X. Huang, B. J. Braams, and J. M. Bowman, "Ab initio potential energy and dipole moment surfaces of $(H_2O)_2$," *J. Phys. Chem. A*, vol. 110, no. 2, pp. 445–451, 2006.

[103] Y. Wang, B. C. Shepler, B. J. Braams, and J. M. Bowman, "Full-dimensional, ab initio potential energy and dipole moment surfaces for water," *J. Chem. Phys.*, vol. 131, no. 5, p. 054511, 2009.

[104] Y. Wang and J. M. Bowman, "Towards an ab initio flexible potential for water and post-harmonic quantum vibrational analysis of water clusters," *Chem. Phys. Lett.*, vol. 491, no. 1-3, pp. 1–10, 2010.

[105] Y. Wang and J. M. Bowman, "Ab initio potential and dipole moment surfaces for water. II. Local-monomer calculations of the infrared spectra of water clusters," *J. Chem. Phys.*, vol. 134, no. 15, p. 154510, 2011.

[106] 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.*, vol. 10, no. 8, pp. 2906–2910, 2014.

[107] P. Bajaj, A. W. Gotz, 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.*, vol. 12, no. 6, pp. 2698–2705, 2016.

[108] 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.*, vol. 147, no. 16, p. 161715, 2017.

[109] V. W. D. Cruzeiro, E. Lambros, M. Riera, R. Roy, F. Paesani, and A. W. Gotz, "Highly accurate many-body potentials for simulations of n_2o_5 in water: Benchmarks, development, and validation," *J. Chem. Theory Comput.*, vol. 17, no. 7, pp. 3931–

3945, 2021.

[110] 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.*, vol. 49, no. 9, pp. 1844–1851, 2016.

[111] M. Riera, E. P. Yeh, and F. Paesani, "Data-driven many-body models for molecular fluids: Co2/h2o mixtures as a case study," *J. Chem. Theory Comput.*, vol. 16, no. 4, pp. 2246–2257, 2020.

[112] M. Riera, A. Hirales, R. Ghosh, and F. Paesani, "Data-driven many-body models with chemical accuracy for ch4/h2o mixtures," *J. Phys. Chem. B*, vol. 124, no. 49, pp. 11207–11221, 2020.

[113] A. Caruso and F. Paesani, "Data-driven many-body models enable a auantitative description of chloride hydration from clusters to bulk," *J. Chem. Phys.*, vol. 155, no. 6, p. 064502, 2021.

[114] A. Caruso, X. Zhu, J. L. Fulton, and F. Paesani, "Accurate modeling of bromide and iodide hydration with data-driven many-body potentials," *J. Phys. Chem. B*, vol. 126, no. 41, pp. 8266–8278, 2022.

[115] D. Zhuang, M. Riera, R. Zhou, A. Deary, and F. Paesani, "Hydration structure of na^+ and k^+ ions in solution predicted by data-driven many-body potentials," *J. Phys. Chem. B*, vol. 126, no. 45, pp. 9349–9360, 2022.

[116] E. F. Bull-Vulpe, M. Riera, S. L. Bore, and F. Paesani, "Data-driven many-body potential energy functions for generic molecules: Linear alkanes as a proof-of-concept application," *J. Chem. Theory Comput.*, 2022.

[117] J. Rezac and P. Hobza, "Benchmark calculations of interaction energies in noncovalent complexes and their applications," *Chem. Rev.*, vol. 116, no. 9, pp. 5038–5071, 2016.

[118] J. Han, L. Zhang, and R. Car, "Deep potential: A general representation of a many-body potential energy surface," *arXiv preprint arXiv:1707.01478*, 2017.

[119] L. Zhang, J. Han, H. Wang, W. Saidi, R. Car, and W. E, "End-to-end symmetry preserving inter-atomic potential energy model for finite and extended systems," *Adv. Neural Inf. Process. Syst.*, vol. 31, 2018.

[120] 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.*, vol. 152, no. 14, p. 144103, 2020.

[121] S. Dasgupta, C. Shahi, P. Bhetwal, J. P. Perdew, and F. Paesani, "How good is the density-corrected scan functional for neutral and ionic aqueous systems, and what is so right about the hartree–fock density?," *J. Chem. Theory Comput.*, vol. 18, no. 8, pp. 4745–4761, 2022.

[122] 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.*, vol. 49, no. 11, pp. 5116–5129, 1968.

[123] 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.*, vol. 53, no. 2, pp. 582–594, 1970.

[124] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equation of state calculations by fast computing machines," *J. Chem. Phys.*, vol. 21, no. 6, pp. 1087–1092, 1953.

[125] B. J. Alder and T. E. Wainwright, "Phase transition for a hard sphere system," *J. Chem. Phys.*, vol. 27, no. 5, pp. 1208–1209, 1957.

[126] M. S. Gordon, G. Barca, S. S. Leang, D. Poole, A. P. Rendell, J. L. Galvez Vallejo, and B. Westheimer, "Novel computer architectures and quantum chemistry," *J. Phys. Chem. A*, vol. 124, no. 23, pp. 4557–4582, 2020.

[127] J. W. Ponder, C. Wu, P. Ren, V. S. Pande, J. D. Chodera, M. J. Schnieders, I. Haque, D. L. Mobley, D. S. Lambrecht, R. A. DiStasio Jr, M. Head-Gordon, G. N. I. Clark, M. E. Johnson, and T. Head-Gordon, "Current status of the AMOEBA polarizable force field," *J. Phys. Chem. B*, vol. 114, no. 8, pp. 2549–2564, 2010.

[128] K. Vanommeslaeghe and A. MacKerell Jr, "CHARMM additive and polarizable force fields for biophysics and computer-aided drug design," *Biochim. Biophys. Acta Gen. Subj.*, vol. 1850, no. 5, pp. 861–871, 2015.

[129] G. A. Cisneros, K. T. Wikfeldt, L. Ojamäe, J. Lu, Y. Xu, H. Torabifard, A. P. Bartók, G. Csányi, V. Molinero, and F. Paesani, "Modeling molecular interactions in water: From pairwise to many-body potential energy functions," *Chem. Rev.*, vol. 116, no. 13, pp. 7501–7528, 2016.

[130] Z. Jing, C. Liu, S. Y. Cheng, R. Qi, B. D. Walker, J.-P. Piquemal, and P. Ren, "Polarizable force fields for biomolecular simulations: Recent advances and applications," *Annu. Rev. Biophys.*, vol. 48, pp. 371–394, 2019.

[131] V. S. Inakollu, D. P. Geerke, C. N. Rowley, and H. Yu, "Polarisable force fields: What do they add in biomolecular simulations?," *Curr. Opin. Struct. Biol.*, vol. 61,

pp. 182–190, 2020.

[132] S. L. Mayo, B. D. Olafson, and W. A. Goddard, "DREIDING: A generic force field for molecular simulations," *J. Phys. Chem.*, vol. 94, no. 26, pp. 8897–8909, 1990.

[133] A. K. Rappé, C. J. Casewit, K. S. 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.*, vol. 114, no. 25, pp. 10024–10035, 1992.

[134] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case, "Development and testing of a general AMBER force field," *J. Comp. Chem.*, vol. 25, no. 9, pp. 1109–1213, 2004.

[135] K. Vanommeslaeghe, E. Hatcher, C. Acharya, S. Kundu, S. Zhong, J. Shim, E. Darian, O. Guvench, P. Lopes, I. Vorobyov, and A. MacKerell Jr, "CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields," *J. Comp. Chem.*, vol. 31, no. 4, pp. 671–690, 2010.

[136] T. Adler, G. Knizia, and H. Werner, "A simple and efficient CCSD(T)-F12 approximation," *J. Chem. Phys.*, vol. 127, no. 22, pp. 221106–221106, 2007.

[137] G. Knizia, T. B. Adler, and H.-J. Werner, "Simplified CCSD(T)-F12 methods: Theory and benchmarks," *J. Chem. Phys.*, vol. 130, no. 5, p. 054104, 2009.

[138] J. G. Hill, K. A. Peterson, G. Knizia, and H.-J. Werner, "Extrapolating MP2 and CCSD explicitly correlated correlation energies to the complete basis set limit with first and second row correlation consistent basis sets," *J. Chem. Phys.*, vol. 131, no. 19, p. 194105, 2009.

[139] U. Góra, R. Podeszwa, W. Cencek, and K. Szalewicz, "Interaction energies of large clusters from many-body expansion," *J. Chem. Phys.*, vol. 135, no. 22, p. 224102, 2011.

[140] D. Manna, M. K. Kesharwani, N. Sylvetsky, and J. M. Martin, "Conventional and explicitly correlated ab initio benchmark study on water clusters: Revision of the BEGDB and WATER27 data sets," *J. Chem. Theory Comput.*, vol. 13, no. 7, pp. 3136–3152, 2017.

[141] J. P. Heindel, K. M. Herman, E. Apra, and S. S. Xantheas, "Guest–host interactions in clathrate hydrates: Benchmark MP2 and CCSD (T)/CBS binding energies of CH_4, CO_2, and H_2S in $(H_2O)_{20}$ cages," *J. Phys. Chem. Lett.*, vol. 12, no. 31, pp. 7574–7582, 2021.

[142] R. Nesbet, "Atomic bethe-goldstone equations," *Adv. Chem. Phys.*, pp. 1–34, 1969.

[143] 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.*, vol. 16, no. 4, pp. 2246–2257, 2020.

[144] M. Riera, A. Hirales, R. Ghosh, and F. Paesani, "Data-driven many-body models with chemical accuracy for CH_4/H_2O mixtures," *J. Phys. Chem. B*, vol. 124, no. 49, pp. 11207–11221, 2020.

[145] S. Yue, M. Riera, R. Ghosh, A. Z. Panagiotopoulos, and F. Paesani, "Transferability of data-driven, many-body models for CO_2 simulations in the vapor and liquid phases," *J. Chem. Phys.*, vol. 156, no. 10, p. 104503, 2022.

[146] F. Noé, A. Tkatchenko, K.-R. Müller, and C. Clementi, "Machine learning for molecular simulation," *Annu. Rev. Phys. Chem.*, vol. 71, pp. 361–390, 2020.

[147] F. Paesani, "Water: Many-body potential from first principles (from the gas to the liquid phase)," in *Handbook of Materials Modeling: Methods: Theory and Modeling* (W. Andreoni and S. Yip, eds.), pp. 635–660, Springer, 2020.

[148] 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.*, vol. 154, no. 3, p. 034111, 2021.

[149] S. Vuckovic, S. Song, J. Kozlowski, E. Sim, and K. Burke, "Density functional analysis: the theory of density-corrected dft," *J. Chem. Theory Comput.*, vol. 15, no. 12, pp. 6636–6646, 2019.

[150] S. Dasgupta, E. Lambros, J. P. Perdew, and F. Paesani, "Elevating density functional theory to chemical accuracy for water simulations through a density-corrected many-body formalism," *Nat. Commun.*, vol. 12, no. 1, p. 6359, 2021.

[151] S. Song, S. Vuckovic, E. Sim, and K. Burke, "Density-corrected DFT explained: Questions and answers," *J. Chem. Theory Comput.*, vol. 18, no. 2, pp. 817–827, 2022.

[152] E. Sim, S. Song, S. Vuckovic, and K. Burke, "Improving results by improving densities: Density-corrected density functional theory," *J. Am. Chem. Soc.*, vol. 144, no. 15, pp. 6625–6639, 2022.

[153] E. Palos, E. Lambros, S. Swee, J. Hu, S. Dasgupta, and F. Paesani, "Assessing the interplay between functional-driven and density-driven errors in dft models of water," *J. Chem. Theory Comput.*, vol. 18, no. 6, pp. 3410–3426, 2022.

[154] E. Lambros, J. Hu, and F. Paesani, "Assessing the accuracy of the scan functional for water through a many-body analysis of the adiabatic connection formula," *J.*

Chem. Theory Comput., vol. 17, no. 6, pp. 3739–3749, 2021.

[155] J. Barker and R. Watts, "Structure of water; a Monte Carlo calculation," *Chem. Phys. Lett.*, vol. 3, no. 3, pp. 144–145, 1969.

[156] A. Rahman and F. H. Stillinger, "Molecular dynamics study of liquid water," *J. Chem. Phys.*, vol. 55, no. 7, pp. 3336–3359, 1971.

[157] P. Gallo, K. Amann-Winkel, C. A. Angell, M. A. Anisimov, F. Caupin, C. Chakravarty, E. Lascaris, T. Loerting, A. Z. Panagiotopoulos, J. Russo, J. A. Sellberg, H. E. Stanley, H. Tanaka, C. Vega, L. Xu, and L. G. M. Pettersson, "Water: A tale of two liquids," *Chem. Rev.*, vol. 116, no. 13, pp. 7463–7500, 2016.

[158] 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.*, vol. 145, no. 19, p. 194504, 2016.

[159] 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*, vol. 351, no. 6279, pp. 1310–1313, 2016.

[160] 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*, vol. 352, no. 6290, pp. 1194–1197, 2016.

[161] 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.*, vol. 139, no. 20, pp. 7082–7088, 2017.

[162] 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.*, vol. 147, no. 24, p. 244504, 2017.

[163] 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.*, vol. 138, no. 11, pp. 3912–3919, 2016.

[164] 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*, vol. 122, no. 15, pp. 4356–4365, 2018.

[165] 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.*, vol. 154, no. 21, p. 211103, 2021.

[166] C. H. Pham, S. K. Reddy, K. Chen, C. Knight, and F. Paesani, "Many-body interactions in ice," *J. Chem. Theory Comput.*, vol. 13, no. 4, pp. 1778–1784, 2017.

[167] 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.*, vol. 8, no. 12, pp. 2579–2583, 2017.

[168] D. R. Moberg, P. J. Sharp, and F. Paesani, "Molecular-level interpretation of vibrational spectra of ordered ice phases," *J. Phys. Chem. B*, vol. 122, no. 46, pp. 10572–10581, 2018.

[169] 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.*, vol. 116, no. 49, pp. 24413–24419, 2019.

[170] R. Speedy and C. Angell, "Isothermal compressibility of supercooled water and evidence for a thermodynamic singularity at -45 °C," *J. Chem. Phys.*, vol. 65, no. 3, pp. 851–858, 1976.

[171] C. Angell, W. Sichina, and M. Oguni, "Heat capacity of water at extremes of supercooling and superheating," *J. Phys. Chem.*, vol. 86, no. 6, pp. 998–1002, 1982.

[172] K. H. Kim, A. Späh, H. Pathak, F. Perakis, D. Mariedahl, K. Amann-Winkel, J. A. Sellberg, J. H. Lee, S. Kim, J. Park, T. Katayama, and A. Nilsson, "Maxima in the thermodynamic response and correlation functions of deeply supercooled water," *Science*, vol. 358, no. 6370, pp. 1589–1593, 2017.

[173] H. Pathak, A. Späh, N. Esmaeildoost, J. A. Sellberg, K. H. Kim, F. Perakis, K. Amann-Winkel, M. Ladd-Parada, J. Koliyadu, T. J. Lane, C. Yang, H. T. Lemke, A. R. Oggenfuss, P. J. M. Johnson, Y. Deng, S. Zerdane, R. Mankowsky, P. Beaud, and A. Nilsson, "Enhancement and maximum in the isobaric specific-heat capacity measurements of deeply supercooled water using ultrafast calorimetry," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 118, no. 6, p. e2018379118, 2021.

[174] K. H. Kim, K. Amann-Winkel, N. Giovambattista, A. Späh, F. Perakis, H. Pathak, M. L. Parada, C. Yang, D. Mariedahl, T. Eklund, T. J. Lane, S. You, S. Jeong, M. Weston, J. H. Lee, I. Eom, M. Kim, J. Park, S. H. Chun, P. H. Poole, and A. Nilsson, "Experimental observation of the liquid-liquid transition in bulk supercooled water under pressure," *Science*, vol. 370, no. 6519, pp. 978–982, 2020.

[175] L. Kringle, W. A. Thornley, B. D. Kay, and G. A. Kimmel, "Reversible structural transformations in supercooled liquid water from 135 to 245 K," *Science*, vol. 369, no. 6510, pp. 1490–1492, 2020.

[176] T. E. Gartner III, L. Zhang, P. M. Piaggi, R. Car, A. Z. Panagiotopoulos, and P. G. Debenedetti, "Signatures of a liquid–liquid transition in an ab initio deep neural network model for water," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 117, no. 42, pp. 26040–26046, 2020.

[177] G. M. Sommers, M. F. C. Andrade, L. Zhang, H. Wang, and R. Car, "Raman spectrum and polarizability of liquid water from deep neural networks," *Phys. Chem. Chem. Phys.*, vol. 22, no. 19, pp. 10592–10602, 2020.

[178] L. Zhang, H. Wang, R. Car, and W. E, "Phase diagram of a deep potential water model," *Phys. Rev. Lett.*, vol. 126, p. 236001, Jun 2021.

[179] C. Zhang, F. Tang, M. Chen, J. Xu, L. Zhang, D. Y. Qiu, J. P. Perdew, M. L. Klein, and X. Wu, "Modeling liquid water by climbing up Jacob's ladder in density functional theory facilitated by using deep neural network potentials," *J. Phys. Chem. B*, vol. 125, no. 41, pp. 11444–11456, 2021.

[180] W. Liang, G. Lu, and J. Yu, "Machine-learning-driven simulations on microstructure and thermophysical properties of $MgCl_2$–KCl eutectic," *ACS Appl. Mater. Interfaces*, vol. 13, no. 3, pp. 4034–4042, 2021.

[181] T. Wen, R. Wang, L. Zhu, L. Zhang, H. Wang, D. J. Srolovitz, and Z. Wu, "Specialising neural network potentials for accurate properties and application to the mechanical response of titanium," *npj Comput. Mater.*, vol. 7, no. 1, pp. 1–11, 2021.

[182] D. Lu, H. Wang, M. Chen, L. Lin, R. Car, E. Weinan, W. Jia, and L. Zhang, "86 PFLOPS deep potential molecular dynamics simulation of 100 million atoms with ab initio accuracy," *Comput. Phys. Commun.*, vol. 259, p. 107624, 2021.

[183] H. Niu, L. Bonati, P. M. Piaggi, and M. Parrinello, "Ab initio phase diagram and nucleation of gallium," *Nat. Commun.*, vol. 11, no. 1, pp. 1–9, 2020.

[184] Z. Guo, D. Lu, Y. Yan, S. Hu, R. Liu, G. Tan, N. Sun, W. Jiang, L. Liu, Y. Chen, *et al.*, "Extending the limit of molecular dynamics with ab initio accuracy to 10 billion atoms," in *Proceedings of the 27th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming*, pp. 205–218, 2022.

[185] H. Partridge and D. W. Schwenke, "The determination of an accurate isotope dependent potential energy surface for water from extensive ab initio calculations and experimental data," *J. Chem. Phys.*, vol. 106, no. 11, pp. 4618–4639, 1997.

[186] B. J. Braams and J. M. Bowman, "Permutationally invariant potential energy surfaces in high dimensionality," *Int. Rev. Phys. Chem.*, vol. 28, no. 4, pp. 577–606, 2009.

[187] H. Wang, L. Zhang, J. Han, and E. Weinan, "DeePMD-kit: A deep learning package for many-body potential energy representation and molecular dynamics," *Comput. Phys. Commun.*, vol. 228, pp. 178–184, 2018.

[188] Y. Zhang, H. Wang, W. Chen, J. Zeng, L. Zhang, H. Wang, and E. Weinan, "DP-GEN: A concurrent learning platform for the generation of reliable deep learning based potential energy models," *Comput. Phys. Commun.*, vol. 253, p. 107206, 2020.

[189] "MBX: A many-body energy and force calculator." https://paesanigroup.ucsd.edu/software/mbx.html.

[190] M. Pinheiro, F. Ge, N. Ferré, P. O. Dral, and M. Barbatti, "Choosing the right molecular machine learning potential," *Chem. Sci.*, vol. 12, no. 43, pp. 14396–14413, 2021.

[191] W. Shinoda, M. Shiga, and M. Mikami, "Rapid estimation of elastic constants by molecular dynamics simulation under constant stress," *Phys. Rev. B*, vol. 69, no. 13, p. 134103, 2004.

[192] A. P. Thompson, H. M. Aktulga, R. Berger, D. S. Bolintineanu, W. M. Brown, P. S. Crozier, P. J. in't Veld, A. Kohlmeyer, S. G. Moore, T. D. Nguyen, *et al.*, "Lammps - a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales," *Comput. Phys. Commun.*, vol. 271, p. 108171, 2022.

[193] J. R. Errington and P. G. Debenedetti, "Relationship between structural order and the anomalies of liquid water," *Nature*, vol. 409, no. 6818, pp. 318–321, 2001.

[194] P. M. Piaggi, A. Z. Panagiotopoulos, P. G. Debenedetti, and R. Car, "Phase equilibrium of water with hexagonal and cubic ice using the SCAN functional," *J. Chem. Theory Comput.*, vol. 17, no. 5, pp. 3065–3077, 2021.

[195] G. S. Kell, "Isothermal compressibility of liquid water at 1 atm.," *J. Chem. Eng. Data*, vol. 15, no. 1, pp. 119–122, 1970.

[196] J. Zinn-Justin, "Precise determination of critical exponents and equation of state by field theory methods," *Phys. Rep.*, vol. 344, no. 4-6, pp. 159–178, 2001.

[197] P. J. Linstrom and W. G. Mallard, "The NIST Chemistry WebBook: A chemical data resource on the internet," *J. Chem. Eng. Data*, vol. 46, no. 5, pp. 1059–1063, 2001.

[198] 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.*, vol. 143, no. 10, p. 104102, 2015.

[199] 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.*, vol. 10, no. 35, pp. 8211–8218, 2019.

[200] E. Lambros and F. Paesani, "How good are polarizable and flexible models for water: Insights from a many-body perspective," *J. Chem. Phys.*, vol. 153, no. 6, p. 060901, 2020.

[201] P. Schienbein and D. Marx, "Liquid–vapor phase diagram of RPBE-D3 water: electronic properties along the coexistence curve and in the supercritical phase," *J. Phys. Chem. B*, vol. 122, no. 13, pp. 3318–3329, 2017.

[202] M. Karplus, "Development of multiscale models for complex chemical systems: From H + H_2 to biomolecules (Nobel lecture)," *Angew. Chem. Int. Ed.*, vol. 53, no. 38, pp. 9992–10005, 2014.

[203] A. Warshel, "Multiscale modeling of biological functions: From enzymes to molecular machines (Nobel lecture)," *Angew. Chem. Int. Ed.*, vol. 53, no. 38, pp. 10020–10031, 2014.

[204] M. Levitt, "Birth and future of multiscale modeling for macromolecular systems (Nobel lecture)," *Angew. Chem. Int. Ed.*, vol. 53, no. 38, pp. 10006–10018, 2014.

[205] 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.*, vol. 108, no. 5, p. 058301, 2012.

[206] G. Montavon, M. Rupp, V. Gobre, A. Vazquez-Mayagoitia, K. Hansen, A. Tkatchenko, K.-R. Müller, and O. A. Von Lilienfeld, "Machine learning of molecular electronic properties in chemical compound space," *New J. Phys.*, vol. 15, no. 9, p. 095003, 2013.

[207] E. Y. Lee, B. M. Fulan, G. C. Wong, and A. L. Ferguson, "Mapping membrane activity in undiscovered peptide sequence space using machine learning," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 113, no. 48, pp. 13588–13593, 2016.

[208] J. P. Janet and H. J. Kulik, "Predicting electronic structure properties of transition metal complexes with neural networks," *Chem. Sci.*, vol. 8, no. 7, pp. 5137–5152, 2017.

[209] Z. Wu, B. Ramsundar, E. N. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing, and V. Pande, "Moleculenet: a benchmark for molecular machine learning," *Chem. Sci.*, vol. 9, no. 2, pp. 513–530, 2018.

[210] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, and A. Walsh, "Machine learning for molecular and materials science," *Nature*, vol. 559, no. 7715, pp. 547–555, 2018.

[211] A. Grisafi, A. Fabrizio, B. Meyer, D. M. Wilkins, C. Corminboeuf, and M. Ceriotti, "Transferable machine-learning model of the electron density," *ACS Cent. Sci.*, vol. 5, no. 1, pp. 57–64, 2018.

[212] L. Tallorin, J. Wang, W. E. Kim, S. Sahu, N. M. Kosa, P. Yang, M. Thompson, M. K. Gilson, P. I. Frazier, M. D. Burkart, *et al.*, "Discovering de novo peptide substrates for enzymes using machine learning," *Nat. Commun.*, vol. 9, no. 1, pp. 1–10, 2018.

[213] M. H. Segler, M. Preuss, and M. P. Waller, "Planning chemical syntheses with deep neural networks and symbolic ai," *Nature*, vol. 555, no. 7698, pp. 604–610, 2018.

[214] F. Noé, S. Olsson, J. Köhler, and H. Wu, "Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning," *Science*, vol. 365, no. 6457, p. eaaw1147, 2019.

[215] J. Wang, S. Olsson, C. Wehmeyer, A. Pérez, N. E. Charron, G. De Fabritiis, F. Noé, and C. Clementi, "Machine learning of coarse-grained molecular dynamics force fields," *ACS Cent. Sci.*, vol. 5, no. 5, pp. 755–767, 2019.

[216] R. Caruana and A. Niculescu-Mizil, "An empirical comparison of supervised learning algorithms," in *Proceedings of the 23rd international conference on Machine learning*, pp. 161–168, 2006.

[217] S. B. Kotsiantis, I. Zaharakis, P. Pintelas, *et al.*, "Supervised machine learning: A review of classification techniques," *Emerging artificial intelligence applications in computer engineering*, vol. 160, no. 1, pp. 3–24, 2007.

[218] H. Almuallim, T. G. Dietterich, *et al.*, "Learning with many irrelevant features.," in *AAAI*, vol. 91, pp. 547–552, Citeseer, 1991.

[219] D. Hankins, J. Moskowitz, and F. Stillinger, "Water molecule interactions," *J. Chem. Phys.*, vol. 53, no. 12, pp. 4544–4554, 1970.

[220] 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.*, vol. 145, no. 6, p. 064308, 2016.

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

[222] C. L. Vaillant and M. T. Cvitaš, "Rotation-tunneling spectrum of the water dimer from instanton theory," *Phys. Chem. Chem. Phys.*, vol. 20, no. 42, pp. 26809–26813, 2018.

[223] C. Vaillant, D. Wales, and S. Althorpe, "Tunneling splittings from path-integral molecular dynamics using a Langevin thermostat," *J. Chem. Phys.*, vol. 148, no. 23, p. 234102, 2018.

[224] 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.*, vol. 148, no. 12, p. 124116, 2018.

[225] 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.*, vol. 148, no. 10, p. 102303, 2018.

[226] 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.*, vol. 148, no. 8, p. 084303, 2018.

[227] N. R. Samala and N. Agmon, "Temperature dependence of intramolecular vibrational bands in small water clusters," *J. Phys. Chem. B*, vol. 123, no. 44, pp. 9428–9442, 2019.

[228] N. R. Samala and N. Agmon, "Thermally induced hydrogen-bond rearrangements in small water clusters and the persistent water tetramer," *ACS omega*, vol. 4, no. 27, pp. 22581–22590, 2019.

[229] M. T. Cvitaš and J. O. Richardson, "Quantum tunnelling pathways of the water pentamer," *Phys. Chem. Chem. Phys.*, vol. 22, no. 3, pp. 1035–1044, 2020.

[230] G. R. Medders and F. Paesani, "Infrared and raman spectroscopy of liquid water through "first-principles" many-body molecular dynamics," *J. Chem. Theory Comput.*, vol. 11, no. 3, pp. 1145–1154, 2015.

[231] 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.*, vol. 121, no. 13, p. 137401, 2018.

[232] K. M. Hunter, F. A. Shakib, and F. Paesani, "Disentangling coupling effects in the infrared spectra of liquid water," *J. Phys. Chem. B*, vol. 122, no. 47, pp. 10754–10761,

2018.

[233] S. Sengupta, D. R. Moberg, F. Paesani, and E. Tyrode, "Neat water–vapor interface: Proton continuum and the nonresonant background," *J. Phys. Chem. Lett.*, vol. 9, no. 23, pp. 6744–6749, 2018.

[234] 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.*, vol. 121, no. 24, p. 246101, 2018.

[235] 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.*, vol. 10, no. 3, pp. 406–412, 2019.

[236] C. J. Burnham, D. J. Anick, P. K. Mankoo, and G. F. Reiter, "The vibrational proton potential in bulk liquid water and ice," *J. Chem. Phys.*, vol. 128, no. 15, p. 154519, 2008.

[237] A. N. Tikhonov, "Solution of incorrectly formulated problems and the regularization method," *Soviet Math.*, vol. 4, pp. 1035–1038, 1963.

[238] T. H. Dunning Jr, "Gaussian basis sets for use in correlated molecular calculations. i. the atoms boron through neon and hydrogen," *J. Chem. Phys.*, vol. 90, no. 2, pp. 1007–1023, 1989.

[239] R. A. Kendall, T. H. Dunning, and R. J. Harrison, "Electron affinities of the first-row atoms revisited. Systematic basis sets and wave functions," *J. Chem. Phys.*, vol. 96, pp. 6796–6806, 1992.

[240] D. E. Woon and T. H. Dunning Jr, "Gaussian basis sets for use in correlated molecular calculations. v. core-valence basis sets for boron through neon," *J. Chem. Phys.*, vol. 103, no. 11, pp. 4572–4585, 1995.

[241] J. G. Hill and K. A. Peterson, "Gaussian basis sets for use in correlated molecular calculations. xi. pseudopotential-based and all-electron relativistic basis sets for alkali metal (k–fr) and alkaline earth (ca–ra) elements," *J. Chem. Phys.*, vol. 147, no. 24, p. 244106, 2017.

[242] 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.*, vol. 19, no. 4, pp. 553–566, 1970.

[243] I. S. Lim, P. Schwerdtfeger, B. Metz, and H. Stoll, "All-electron and relativistic pseudopotential studies for the group 1 element polarizabilities from k to element

119," *J. Chem. Phys.*, vol. 122, no. 10, p. 104103, 2005.

[244] G. J. Martyna, A. Hughes, and M. E. Tuckerman, "Molecular dynamics algorithms for path integrals at constant pressure," *J. Chem. Phys.*, vol. 110, no. 7, pp. 3275–3290, 1999.

[245] W. Smith and T. Forester, "DL_POLY_2. 0: A general-purpose parallel molecular dynamics simulation package," *J. Mol. Graph.*, vol. 14, no. 3, pp. 136–141, 1996.

[246] B. Settles, "Active learning literature survey," Tech. Rep. Computer Sciences Technical Report 1648, University of Wisconsin-Madison Department of Computer Sciences, 2009.

[247] C. E. Rasmussen and C. K. I. Williams, *Gaussian Process for Machine Learning*. MIT Press, 2006.

[248] 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.*, vol. 104, no. 13, p. 136403, 2010.

[249] G. Imbalzano, A. Anelli, D. Giofré, S. Klees, J. Behler, and M. Ceriotti, "Automatic selection of atomic fingerprints and reference configurations for machine-learning potentials," *J. Chem. Phys*, vol. 148, no. 24, p. 241730, 2018.

[250] J. S. Smith, B. Nebgen, N. Lubbers, O. Isayev, and A. E. Roitberg, "Less is more: Sampling chemical space with active learning," *The Journal of chemical physics*, vol. 148, no. 24, p. 241733, 2018.

[251] H. Huo and M. Rupp, "Unified representation for machine learning of molecules and crystals," *arXiv:1704.06439*, 2017.

[252] M. Ceriotti, G. A. Tribello, and M. Parrinello, "Simplifying the representation of complex free-energy landscapes using sketch-map," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 108, no. 32, pp. 13023–13028, 2011.

[253] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2006.

[254] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. The MIT Press, 2016.

[255] M. Ceriotti, G. A. Tribello, and M. Parrinello, "Demonstrating the transferability and the descriptive power of sketch-map," *J. Chem. Theory Comput.*, vol. 9, no. 3, pp. 1521–1532, 2013.

[256] T. T. Nguyen, E. Székely, G. Imbalzano, J. Behler, G. Csányi, M. Ceriotti, A. W. Götz, and F. Paesani, "Comparison of permutationally invariant polynomials, neural networks, and gaussian approximation potentials in representing water interactions through many-body expansions," *J. Chem. Phys.*, vol. 148, no. 24, p. 241725, 2018.

[257] N. Sträter, W. N. Lipscomb, T. Klabunde, and B. Krebs, "Two-metal ion catalysis in enzymatic acyl- and phosphoryl-transfer reactions," *Angew. Chem. Int. Ed. Engl.*, vol. 35, no. 18, pp. 2024–2055, 1996.

[258] R. Hanna and J. A. Doudna, "Metal ions in ribozyme folding and catalysis," *Curr. Opin. Chem. Biol.*, vol. 4, no. 2, pp. 166–170, 2000.

[259] M. Mucha, T. Frigato, L. M. Levering, H. C. Allen, D. J. Tobias, L. X. Dang, and P. Jungwirth, "Unified molecular picture of the surfaces of aqueous acid, base, and salt solutions," *J. Phys. Chem. B*, vol. 109, pp. 7617–7623, 2005.

[260] M. R. Stahley and S. A. Strobel, "Rna splicing: Group I intron crystal structures reveal the basis of splice site selection and metal ion catalysis," *Curr. Opin. Struc. Biol.*, vol. 16, no. 3, pp. 319–326, 2006.

[261] R. K. Sigel and A. M. Pyle, "Alternative roles for metal ions in enzyme catalysis and the implications for ribozyme chemistry," *Chem. Rev.*, vol. 107, no. 1, pp. 97–113, 2007.

[262] F. Hofmeister, "Zur lehre von der wirkung der salze," *Naunyn-Schmiedeberg's Arch. Pharmacol.*, vol. 24, no. 4-5, pp. 247–260, 1888.

[263] M. D. Baer and C. J. Mundy, "Toward an understanding of the specific ion effect using density functional theory," *J. Phys. Chem. Lett.*, vol. 2, no. 9, pp. 1088–1093, 2011.

[264] M. J. Gillan, D. Alfè, and A. Michaelides, "Perspective: How good is DFT for water?," *J. Chem. Phys.*, vol. 144, no. 13, p. 130901, 2016.

[265] M. Galib, T. T. Duignan, Y. Misteli, M. D. Baer, G. K. Schenter, J. Hutter, and C. J. Mundy, "Mass density fluctuations in quantum and classical descriptions of liquid water," *J. Chem. Phys.*, vol. 146, no. 24, p. 244501, 2017.

[266] D. E. Otten, P. R. Shaffer, P. L. Geissler, and R. J. Saykally, "Elucidating the mechanism of selective ion adsorption to the liquid water surface," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 109, no. 3, pp. 701–705, 2012.

[267] 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*, vol. 120, no. 8, pp. 1822–1832, 2016.

[268] 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.*, vol. 18, no. 44, pp. 30334–30343, 2016.

[269] 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*, vol. 122, no. 27, pp. 5811–5821, 2018.

[270] 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*, vol. 123, no. 13, pp. 2843–2852, 2019.

[271] P. Bajaj, J. O. Richardson, and F. Paesani, "Ion-mediated hydrogen-bond rearrangement through tunnelling in the iodide–dihydrate complex," *Nat. Chem.*, vol. 11, no. 4, pp. 367–374, 2019.

[272] P. Bajaj, D. Zhuang, and F. Paesani, "Specific ion effects on hydrogen-bond rearrangements in the halide–dihydrate complexes," *J. Phys. Chem. Lett.*, vol. 10, no. 11, pp. 2823–2828, 2019.

[273] 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.*, vol. 153, no. 4, p. 044306, 2020.

[274] E. Lambros, S. Dasgupta, E. Palos, S. Swee, J. Hu, and F. Paesani, "General many-body framework for data-driven potentials with arbitrary quantum mechanical accuracy: Water as a case study," *J. Chem. Theory Comput.*, vol. 17, no. 9, pp. 5635–5650, 2021.

[275] 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*, vol. 120, no. 33, pp. 8539–8546, 2016.

[276] A. P. Gaiduk, T. A. Pham, M. Govoni, F. Paesani, and G. Galli, "Electron affinity of liquid water," *Nat. Commun.*, vol. 9, no. 1, pp. 1–6, 2018.

[277] 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.*, vol. 12, no. 16, pp. 3996–4002, 2021.

[278] 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.*, vol. 80, no. 8, pp. 3726–3741, 1984.

[279] B. P. Pritchard, D. Altarawy, B. Didier, T. D. Gibson, and T. L. Windus, "New basis set exchange: An open, up-to-date resource for the molecular sciences community," *J. Chem. Inf. Model.*, vol. 59, no. 11, pp. 4814–4820, 2019.

[280] D. E. Woon and T. H. Dunning, "Gaussian basis sets for use in correlated molecular calculations. III. The atoms aluminum through argon," *J. Chem. Phys.*, vol. 98, pp. 1358–1371, 1993.

[281] H.-J. Werner, P. J. Knowles, G. Knizia, F. R. Manby, and M. Schütz, "Molpro: A general-purpose quantum chemistry program package," *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, vol. 2, no. 2, pp. 242–253, 2012.

[282] A. N. Tihonov, "Solution of incorrectly formulated problems and the regularization method," *Soviet Math.*, vol. 4, pp. 1035–1038, 1963.

[283] M. Tuckerman, B. J. Berne, and G. J. Martyna, "Reversible multiple time scale molecular dynamics," *J. Chem. Phys.*, vol. 97, no. 3, pp. 1990–2001, 1992.

[284] J. Kolafa, "Time-reversible always stable predictor–corrector method for molecular dynamics of polarizable molecules," *J. Comp. Chem.*, vol. 25, no. 3, pp. 335–342, 2004.

[285] G. J. Martyna, M. L. Klein, and M. Tuckerman, "Nosé–Hoover chains: The canonical ensemble via continuous dynamics," *J. Chem. Phys.*, vol. 97, no. 4, pp. 2635–2643, 1992.

[286] B. J. Berne and D. Thirumalai, "On the simulation of quantum systems: Path integral methods," *Annu. Rev. Phys. Chem.*, vol. 37, no. 1, pp. 401–424, 1986.

[287] D. Case, H. Aktulga, K. Belfon, I. Ben-Shalom, S. Brozell, D. Cerutti, T. Cheatham III, V. Cruzeiro, T. Darden, R. Duke, G. Giambasu, M. Gilson, H. Gohlke, A. Goetz, R. Harris, S. Izadi, S. Izmailov, C. Jin, K. Kasavajhala, M. Kaymak, E. King, A. Kovalenko, T. Kurtzman, T. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, M. Machado, V. Man, M. Manathunga, K. Merz, Y. Miao, O. Mikhailovskii, G. Monard, H. Nguyen, K. O'Hearn, A. Onufriev, F. Pan, S. Pantano, R. Qi, A. Rahnamoun, D. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, J. Shen, C. Simmerling, N. Skrynnikov, J. Smith, J. Swails, R. Walker, J. Wang, H. Wei, R. Wolf, X. Wu, Y. Xue, D. York, S. Zhao, and P. Kollman, "Amber 2021." https://ambermd.org, 2021.

[288] J. J. Rehr, J. J. Kas, F. D. Vila, M. P. Prange, and K. Jorissen, "Parameter-free

calculations of X-ray spectra with FEFF9," *Phys. Chem. Chem. Phys.*, vol. 12, no. 21, pp. 5503–5513, 2010.

[289] J. J. Rehr, J. J. Kas, M. P. Prange, A. P. Sorini, Y. Takimoto, and F. Vila, "Ab initio theory and calculations of X-ray spectra," *C. R. Phys.*, vol. 10, no. 6, pp. 548–559, 2009.

[290] J. J. Rehr and R. C. Albers, "Theoretical approaches to X-ray absorption fine structure," *Rev. Mod. Phys.*, vol. 72, no. 3, p. 621, 2000.

[291] L. X. Dang, G. K. Schenter, V.-A. Glezakou, and J. L. Fulton, "Molecular simulation analysis and X-ray absorption measurement of Ca^{2+}, K^{+} and Cl^{-} ions in solution," *J. Phys. Chem. B*, vol. 110, no. 47, pp. 23644–23654, 2006.

[292] P. Bajaj, X.-G. Wang, T. CarringtonJr, 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.*, vol. 148, no. 10, p. 102321, 2017.

[293] Y. Marcus, *Ion Solvation.* Wiley, 1985.

[294] H. T. El-Dessouky and H. M. Ettouney, *Fundamentals of Salt Water Desalination.* Elsevier, 2002.

[295] J. E. Grebel, J. J. Pignatello, and W. A. Mitch, "Effect of halide ions and carbonates on organic contaminant degradation by hydroxyl radical-based advanced oxidation processes in saline waters," *Environ. Sci. Technol.*, vol. 44, no. 17, pp. 6822–6828, 2010.

[296] S. Ghosh and L. Manna, "The many "facets" of halide ions in the chemistry of colloidal inorganic nanocrystals," *Chem. Rev.*, vol. 118, no. 16, pp. 7804–7864, 2018.

[297] D. Duan, C. Winter, S. Cowley, J. R. Hume, and B. Horowitz, "Molecular identification of a volume-regulated chloride channel," *Nature*, vol. 390, no. 6658, pp. 417–421, 1997.

[298] W. H. Robertson and M. A. Johnson, "Molecular aspects of halide ion hydration: The cluster approach," *Annu. Rev. Phys. Chem.*, vol. 54, p. 173, 2003.

[299] H. Ohtaki and T. Radnai, "Structure and dynamics of hydrated ions," *Chem. Rev.*, vol. 93, no. 3, pp. 1157–1204, 1993.

[300] P. Jungwirth and D. J. Tobias, "Specific ion effects at the air/water interface," *Chem. Rev.*, vol. 106, no. 4, pp. 1259–1281, 2006.

[301] P. B. Petersen and R. J. Saykally, "On the nature of ions at the liquid water surface," *Annu. Rev. Phys. Chem.*, vol. 57, pp. 333–364, 2006.

[302] C. J. Fennell, A. Bizjak, V. Vlachy, and K. A. Dill, "Ion pairing in molecular simulations of aqueous alkali halide solutions," *J. Phys. Chem. B*, vol. 113, no. 19, pp. 6782–6791, 2009.

[303] A. B. Wolk, C. M. Leavitt, E. Garand, and M. A. Johnson, "Cryogenic ion chemistry and spectroscopy," *Acc. Chem. Res.*, vol. 47, no. 1, pp. 202–210, 2014.

[304] L. Piatkowski, Z. Zhang, E. H. Backus, H. J. Bakker, and M. Bonn, "Extreme surface propensity of halide ions in water," *Nat. Commun.*, vol. 5, no. 1, pp. 1–7, 2014.

[305] N. F. Van Der Vegt, K. Haldrup, S. Roke, J. Zheng, M. Lund, and H. J. Bakker, "Water-mediated ion pairing: Occurrence and relevance," *Chem. Reviews*, vol. 116, no. 13, pp. 7626–7641, 2016.

[306] Y. Chen, H. I. Okur, N. Gomopoulos, C. Macias-Romero, P. S. Cremer, P. B. Petersen, G. Tocci, D. M. Wilkins, C. Liang, M. Ceriotti, and S. Roke, "Electrolytes induce long-range orientational order and free energy changes in the h-bond network of bulk water," *Sci. Adv.*, vol. 2, no. 4, p. e1501891, 2016.

[307] T. T. Duignan, S. M. Kathmann, G. K. Schenter, and C. J. Mundy, "Toward a first-principles framework for predicting collective properties of electrolytes," *Acc. Chem. Res.*, vol. 54, no. 13, pp. 2833–2843, 2021.

[308] S. E. Weitzner, T. A. Pham, C. A. Orme, S. R. Qiu, and B. C. Wood, "Beyond thermodynamics: Assessing the dynamical softness of hydrated ions from first principles," *J. Phys. Chem. Lett.*, vol. 12, no. 49, pp. 11980–11986, 2021.

[309] J. E. Combariza, N. R. Kestner, and J. Jortner, "Energy-structure relationships for microscopic solvation of anions in water clusters," *J. Chem. Phys.*, vol. 100, no. 4, pp. 2851–2864, 1994.

[310] S. S. Xantheas, "Quantitative description of hydrogen bonding in chloride-water clusters," *J. Phys. Chem.*, vol. 100, no. 23, pp. 9703–9713, 1996.

[311] A. Shalit, S. Ahmed, J. Savolainen, and P. Hamm, "Terahertz echoes reveal the inhomogeneity of aqueous salt solutions," *Nat. Chem.*, vol. 9, no. 3, pp. 273–278, 2017.

[312] R. Ayala, J. M. Martínez, R. R. Pappalardo, and E. Sánchez Marcos, "On the halide hydration study: Development of first-principles halide ion-water interaction potential based on a polarizable model," *J. Chem. Phys.*, vol. 119, no. 18, pp. 9538–

9548, 2003.

[313] P. J. Merkling, R. Ayala, J. M. Martínez, R. R. Pappalardo, and E. Sánchez Marcos, "Interplay of computer simulations and x-ray absorption spectra in the study of the bromide hydration structure," *J. Chem. Phys.*, vol. 119, no. 13, pp. 6647–6654, 2003.

[314] A. Grossfield, "Dependence of ion hydration on the sign of the ion's charge," *J. Chem. Phys.*, vol. 122, no. 2, p. 024506, 2005.

[315] C. Krekeler, B. Hess, and L. Delle Site, "Density functional study of ion hydration for the alkali metal ions (li^+, na^+, k^+) and the halide ions (f^-, br^-, cl^-)," *J. Chem. Phys.*, vol. 125, no. 5, p. 054305, 2006.

[316] P. D'Angelo, V. Migliorati, and L. Guidoni, "Hydration properties of the bromide aqua ion: The interplay of first principle and classical molecular dynamics, and x-ray absorption spectroscopy," *Inorg. Chem.*, vol. 49, no. 9, pp. 4224–4231, 2010.

[317] V. Pham, I. Tavernelli, C. Milne, R. van der Veen, P. D'Angelo, C. Bressler, and M. Chergui, "The solvent shell structure of aqueous iodide: X-ray absorption spectroscopy and classical, hybrid qm/mm and full quantum molecular dynamics simulations," *Chem. Phys.*, vol. 371, no. 1-3, pp. 24–29, 2010.

[318] J. L. Fulton, G. K. Schenter, M. D. Baer, C. J. Mundy, L. X. Dang, and M. Balasubramanian, "Probing the hydration structure of polarizable halides: A multiedge xafs and molecular dynamics study of the iodide anion," *J. Phys. Chem. B*, vol. 114, no. 40, pp. 12926–12937, 2010.

[319] P. T. Kiss and A. Baranyai, "A new polarizable force field for alkali and halide ions," *J. Chem. Phys.*, vol. 141, no. 11, p. 114501, 2014.

[320] S. Yue and A. Z. Panagiotopoulos, "Dynamic properties of aqueous electrolyte solutions from non-polarisable, polarisable, and scaled-charge models," *Mol. Phys.*, vol. 117, no. 23-24, pp. 3538–3549, 2019.

[321] P. Sripa and A. Tongraar, "The "surface"(s) state of the br^- hydration: An oniom-xs md simulation study," *Chem. Phys. Lett.*, vol. 738, p. 136853, 2020.

[322] E. F. Bull-Vulpe, M. Riera, A. W. Götz, and F. Paesani, "Mb-fit: Software infrastructure for data-driven many-body potential energy functions," *J. Chem. Phys.*, vol. 155, no. 12, p. 124801, 2021.

[323] B. T. Thole, "Molecular polarizabilities calculated with a modified dipole interaction," *Chem. Phys.*, vol. 59, no. 3, pp. 341–350, 1981.

[324] A. D. Becke and E. R. Johnson, "Exchange-hole dipole moment and the dispersion interaction revisited," *J. Chem. Phys.*, vol. 127, no. 15, p. 154108, 2007.

[325] A. D. Becke and E. R. Johnson, "A unified density-functional treatment of dynamical, nondynamical, and dispersion correlations," *J. Chem. Phys.*, vol. 127, no. 12, p. 124108, 2007.

[326] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, and D. J. Fox, "Gaussian˜16 Revision C.01," 2016. Gaussian Inc. Wallingford CT.

[327] A. Otero-De-La-Roza and E. R. Johnson, "Non-covalent interactions and thermochemistry using xdm-corrected hybrid and range-separated hybrid density functionals," *J. Chem. Phys.*, vol. 138, no. 20, p. 204109, 2013.

[328] F. O. Kannemann and A. D. Becke, "Van der waals interactions in density-functional theory: Intermolecular complexes," *J. Chem. Theory Comput.*, vol. 6, no. 4, pp. 1081–1088, 2010.

[329] V. V. Gobre, *Efficient modelling of linear electronic polarization in materials using atomic response functions*. Technische Universitaet Berlin (Germany), 2016.

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

[331] R. W. Hockney and J. W. Eastwood, *Computer Simulation Using Particles*. crc Press, 2021.

[332] G. K. Schenter and J. L. Fulton, "Molecular dynamics simulations and xafs (md-xafs)," in *XAFS Techniques for Catalysts, Nanomaterials, and Surfaces*, pp. 251–270, Springer, 2017.

[333] B. Ravel and M. Newville, "Athena, artemis, hephaestus: Data analysis for x-ray absorption spectroscopy using ifeffit," *J. Synchrotron Radiat.*, vol. 12, no. 4,

pp. 537–541, 2005.

[334] V.-A. Glezakou, Y. Chen, J. L. Fulton, G. K. Schenter, and L. X. Dang, "Electronic structure, statistical mechanical simulations, and exafs spectroscopy of aqueous potassium," *Theor. Chem. Acc.*, vol. 115, no. 2, pp. 86–99, 2006.

[335] A. Filipponi, "Exafs for liquids," *J. Phys. Condens. Matter*, vol. 13, no. 7, p. R23, 2001.

[336] V.-T. Pham and J. L. Fulton, "Ion-pairing in aqueous $cacl_2$ and rbbr solutions: Simultaneous structural refinement of xafs and xrd data," *J. Chem. Phys.*, vol. 138, no. 4, p. 044201, 2013.

[337] J. L. Fulton, S. M. Kathmann, G. K. Schenter, and M. Balasubramanian, "Hydrated structure of ag(i) ion from symmetry-dependent, k-and l-edge xafs multiple scattering and molecular dynamics simulations," *J. Phys. Chem. A*, vol. 113, no. 50, pp. 13976–13984, 2009.

[338] P. D'Angelo, A. Zitolo, V. Migliorati, and N. V. Pavel, "Measurement of x-ray multielectron photoexcitations at the i^- k edge," *Phys. Rev. B*, vol. 78, no. 14, p. 144105, 2008.

[339] T. Hansson, C. Oostenbrink, and W. van Gunsteren, "Molecular dynamics simulations," *Curr. Opin. Struc. Biol.*, vol. 12, no. 2, pp. 190–196, 2002.

[340] M. Karplus and J. A. McCammon, "Molecular dynamics simulations of biomolecules," *Nat. Struct. Mol. Biol.*, vol. 9, no. 9, pp. 646–652, 2002.

[341] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman, "A second generation force field for the simulation of proteins, nucleic acids, and organic molecules," *J. Am. Chem. Soc.*, vol. 117, no. 19, pp. 5179–5197, 1995.

[342] W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives, "Development and testing of the opls all-atom force field on conformational energetics and properties of organic liquids," *J. Am. Chem. Soc.*, vol. 118, no. 45, pp. 11225–11236, 1996.

[343] T. A. Halgren and W. Damm, "Polarizable force fields," *Curr. Opin. Struc. Biol.*, vol. 11, no. 2, pp. 236–242, 2001.

[344] S. W. Rick and S. J. Stuart, "Potentials and algorithms for incorporating polarizability in computer simulations," *Rev. Comput. Chem.*, vol. 18, pp. 89–146, 2002.

[345] J. W. Ponder and D. A. Case, "Force fields for protein simulations," *Adv. Protein.*

Chem. Struct. Biol., vol. 66, pp. 27–85, 2003.

[346] E. Palos, S. Dasgupta, E. Lambros, and F. Paesani, "Data-driven many-body potentials from density functional theory for aqueous phase chemistry," *Chem. Phys. Rev.*, vol. 4, no. 1, p. 011301, 2023.

[347] J. F. Stanton, "Why ccsd(t) works: a different perspective," *Chem. Phys. Lett.*, vol. 281, no. 1-3, pp. 130–134, 1997.

[348] J. P. Perdew, R. G. Parr, M. Levy, and J. L. Balduz Jr, "Density-functional theory for fractional particle number: Derivative discontinuities of the energy," *Phys. Rev. Lett.*, vol. 49, no. 23, p. 1691, 1982.

[349] M. G. Medvedev, I. S. Bushmarinov, J. Sun, J. P. Perdew, and K. A. Lyssenko, "Density functional theory is straying from the path toward the exact functional," *Science*, vol. 355, no. 6320, pp. 49–52, 2017.

[350] A. D. Kaplan, M. Levy, and J. P. Perdew, "Predictive power of the exact constraints and appropriate norms in density functional theory," *arXiv preprint arXiv:2207.03855*, 2022.

[351] S. Grimme, "Density functional theory with london dispersion corrections," *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, vol. 1, no. 2, pp. 211–228, 2011.

[352] A. J. Cohen, P. Mori-Sánchez, and W. Yang, "Challenges for density functional theory," *Chem. Rev.*, vol. 112, no. 1, pp. 289–320, 2012.

[353] K. R. Bryenton, A. A. Adeleke, S. G. Dale, and E. R. Johnson, "Delocalization error: The greatest outstanding challenge in density-functional theory," *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, p. e1631, 2022.

[354] M.-C. Kim, E. Sim, and K. Burke, "Ions in solution: Density corrected density functional theory (dc-dft)," *J. Chem. Phys.*, vol. 140, no. 18, p. 18A528, 2014.

[355] K. Sharkas, K. Wagle, B. Santra, S. Akter, R. R. Zope, T. Baruah, K. A. Jackson, J. P. Perdew, and J. E. Peralta, "Self-interaction error overbinds water clusters but cancels in structural energy differences," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 117, no. 21, pp. 11283–11288, 2020.

[356] K. Wagle, B. Santra, P. Bhattarai, C. Shahi, M. R. Pederson, K. A. Jackson, and J. P. Perdew, "Self-interaction correction in water–ion clusters," *J. Chem. Phys.*, vol. 154, no. 9, p. 094302, 2021.

[357] Y. Zhang and W. Yang, "A challenge for density functionals: Self-interaction error

increases for systems with a noninteger number of electrons," *J. Chem. Phys.*, vol. 109, no. 7, pp. 2604–2608, 1998.

[358] A. Ruzsinszky, J. P. Perdew, G. I. Csonka, O. A. Vydrov, and G. E. Scuseria, "Spurious fractional charge on dissociated atoms: Pervasive and resilient self-interaction error of common density functionals," *J. Chem. Phys.*, vol. 125, no. 19, p. 194112, 2006.

[359] A. J. Cohen, P. Mori-Sánchez, and W. Yang, "Fractional spins and static correlation error in density functional theory," *J. Chem. Phys.*, vol. 129, no. 12, p. 121104, 2008.

[360] E. R. Johnson, P. Mori-Sánchez, A. J. Cohen, and W. Yang, "Delocalization errors in density functionals and implications for main-group thermochemistry," *J. Chem. Phys.*, vol. 129, no. 20, p. 204112, 2008.

[361] D. Hait and M. Head-Gordon, "Delocalization errors in density functional theory are essentially quadratic in fractional occupation number," *The journal of physical chemistry letters*, vol. 9, no. 21, pp. 6280–6288, 2018.

[362] K. Laasonen, F. Csajka, and M. Parrinello, "Water dimer properties in the gradient-corrected density functional theory," *Chem. Phys. Lett.*, vol. 194, no. 3, pp. 172–174, 1992.

[363] K. Laasonen, M. Parrinello, R. Car, C. Lee, and D. Vanderbilt, "Structures of small water clusters using gradient-corrected density functional theory," *Chem. Phys. Lett.*, vol. 207, no. 2, pp. 208–213, 1993.

[364] K. Laasonen, M. Sprik, M. Parrinello, and R. Car, ""Ab initio" liquid water," *J. Chem. Phys.*, vol. 99, no. 11, pp. 9080–9089, 1993.

[365] M.-C. Kim, E. Sim, and K. Burke, "Understanding and reducing errors in density functional calculations," *Phys. Rev. Lett.*, vol. 111, no. 7, p. 073003, 2013.

[366] M.-C. Kim, H. Park, S. Son, E. Sim, and K. Burke, "Improved dft potential energy surfaces via improved densities," *J. Phys. Chem. Lett.*, vol. 6, no. 19, pp. 3802–3807, 2015.

[367] E. Sim, S. Song, and K. Burke, "Quantifying density errors in dft," *J. Phys. Chem. Lett.*, vol. 9, no. 22, pp. 6385–6392, 2018.

[368] J. Sun, A. Ruzsinszky, and J. P. Perdew, "Strongly constrained and appropriately normed semilocal density functional," *Phys. Rev. Lett.*, vol. 115, no. 3, p. 036402, 2015.

[369] T. Aschebrock and S. Kümmel, "Ultranonlocality and accurate band gaps from a meta-generalized gradient approximation," *Phys. Rev. Research*, vol. 1, no. 3, p. 033082, 2019.

[370] J. Sun, R. C. Remsing, Y. Zhang, Z. Sun, A. Ruzsinszky, H. Peng, Z. Yang, A. Paul, U. Waghmare, X. Wu, *et al.*, "Accurate first-principles structures and energies of diversely bonded systems from an efficient density functional," *Nat. Commun.*, vol. 8, no. 9, pp. 831–836, 2016.

[371] T. E. Gartner III, P. M. Piaggi, R. Car, A. Z. Panagiotopoulos, and P. G. Debenedetti, "Liquid-liquid transition in water from first principles," *Phys. Rev. Lett.*, vol. 129, no. 25, p. 255702, 2022.

[372] M. Chen, H.-Y. Ko, R. C. Remsing, M. F. Calegari Andrade, B. Santra, Z. Sun, A. Selloni, R. Car, M. L. Klein, J. P. Perdew, *et al.*, "Ab initio theory and modeling of water," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 114, no. 41, pp. 10846–10851, 2017.

[373] P. M. Piaggi, J. Weis, A. Z. Panagiotopoulos, P. G. Debenedetti, and R. Car, "Homogeneous ice nucleation in an ab initio machine-learning model of water," *Proceedings of the National Academy of Sciences*, vol. 119, no. 33, p. e2207294119, 2022.

[374] R. Wang, V. Carnevale, M. L. Klein, and E. Borguet, "First-principles calculation of water pk_a using the newly developed scan functional," *J. Phys. Chem. Lett.*, vol. 11, no. 1, pp. 54–59, 2019.

[375] T. T. Duignan, G. K. Schenter, J. L. Fulton, T. Huthwelker, M. Balasubramanian, M. Galib, M. D. Baer, J. Wilhelm, J. Hutter, M. Del Ben, *et al.*, "Quantifying the hydration structure of sodium and potassium ions: Taking additional steps on jacob's ladder," *Phys. Chem. Chem. Phys.*, vol. 22, no. 19, pp. 10641–10652, 2020.

[376] N. Mardirossian and M. Head-Gordon, "ωB97M-V: A combinatorially optimized range-separated hybrid meta-GGA density functional with VV10 nonlocal correlation," *J. Chem. Phys.*, vol. 144, pp. 214110:1–23, 2016.

[377] J. Kirkpatrick, B. McMorrow, D. H. P. Turban, A. L. Gaunt, J. S. Spencer, A. G. D. G. Matthews, A. Obika, L. Thiry, M. Fortunato, D. Pfau, L. R. Castellanos, S. Petersen, A. W. R. Nelson, P. Kohli, P. Mori-Sánchez, D. Hassabis, and A. J. Cohen, "Pushing the frontiers of density functionals by solving the fractional electron problem," *Science*, vol. 374, no. 6573, pp. 1385–1389, 2021.

[378] E. Palos, E. Lambros, S. Dasgupta, and F. Paesani, "Density functional theory of water with the machine-learned dm21 functional," *J. Chem. Phys.*, vol. 156, no. 16, p. 161103, 2022.

[379] S. B. Rempe and L. R. Pratt, "The hydration number of na+ in liquid water," *Fluid Ph. Equilib.*, vol. 183, pp. 121–132, 2001.

[380] H. B. Schlegel, S. S. Iyengar, X. Li, J. M. Millam, G. A. Voth, G. E. Scuseria, and M. J. Frisch, "Ab initio molecular dynamics: Propagating the density matrix with gaussian orbitals. iii. comparison with born–oppenheimer dynamics," *J. Chem. Phys.*, vol. 117, no. 19, pp. 8694–8704, 2002.

[381] A. Bankura, V. Carnevale, and M. L. Klein, "Hydration structure of salt solutions from ab initio molecular dynamics," *J. Chem. Phys.*, vol. 138, no. 1, p. 014501, 2013.

[382] M. Galib, M. Baer, L. Skinner, C. Mundy, T. Huthwelker, G. Schenter, C. Benmore, N. Govind, and J. L. Fulton, "Revisiting the hydration structure of aqueous na^+," *J. Chem. Phys.*, vol. 146, no. 8, p. 084504, 2017.

[383] R. A. DiStasio, B. Santra, Z. Li, X. Wu, and R. Car, "The individual and collective effects of exact exchange and dispersion interactions on the ab initio structure of liquid water," *J. Chem. Phys.*, vol. 141, no. 8, 2014.

[384] R. C. Remsing, T. T. Duignan, M. D. Baer, G. K. Schenter, C. J. Mundy, and J. D. Weeks, "Water lone pair delocalization in classical and quantum descriptions of the hydration of model ions," *J. Phys. Chem. B*, vol. 122, no. 13, pp. 3519–3527, 2018.

[385] X. Wang, D. Toroz, S. Kim, S. L. Clegg, G.-S. Park, and D. Di Tommaso, "Density functional theory based molecular dynamics study of solution composition effects on the solvation shell of metal ions," *Phys. Chem. Chem. Phys.*, vol. 22, no. 28, pp. 16301–16313, 2020.

[386] T. T. Duignan, C. J. Mundy, G. K. Schenter, and X. S. Zhao, "Method for accurately predicting solvation structure," *J. Chem. Theory Comput.*, vol. 16, no. 8, pp. 5401–5409, 2020.

[387] K. Zhou, C. Qian, and Y. Liu, "Quantifying the structure of water and hydrated monovalent ions by density functional theory-based molecular dynamics," *J. Phys. Chem. B*, vol. 126, no. 49, pp. 10471–10480, 2022.

[388] G. Santra and J. M. Martin, "What types of chemical problems benefit from density-corrected dft? a probe using an extensive and chemically diverse test suite," *J. Chem. Theory Comput.*, vol. 17, no. 3, pp. 1368–1379, 2021.

[389] B. Rana, M. P. Coons, and J. M. Herbert, "Detection and correction of delocalization errors for electron and hole polarons using density-corrected dft," *J. Phys. Chem. Lett.*, vol. 13, no. 23, pp. 5275–5284, 2022.

[390] B. Rana, G. J. Beran, and J. M. Herbert, "Correcting π-delocalisation errors in conformational energies using density-corrected dft, with application to crystal polymorphs," *Molecular Physics*, p. e2138789, 2022.

[391] A. D. Kaplan, C. Shahi, P. Bhetwal, R. K. Sah, and J. P. Perdew, "Understanding density-driven errors for reaction barrier heights," *J. Chem. Theory Comput.*, 2023.

[392] H. Peng, Z.-H. Yang, J. P. Perdew, and J. Sun, "Versatile van der waals density functional based on a meta-generalized gradient approximation," *Phys. Rev. X*, vol. 6, no. 4, p. 041005, 2016.

[393] S. Song, S. Vuckovic, Y. Kim, H. Yu, E. Sim, and K. Burke, "Extending density functional theory with near chemical accuracy beyond pure water," *Nat. Commun.*, vol. 14, no. 1, p. 799, 2023.

[394] M. Riera, C. Knight, E. F. Bull-Vulpe, X. Zhu, D. G. Smith, A. C. Simmonett, and F. Paesani, "Mbx: A many-body energy and force calculator for data-driven many-body simulations." ChemRxiv, https://doi.org/10.26434/chemrxiv-2023-09jh3, 2023.