

Accelerating coupled cluster molecular dynamics of condensed phase aqueous systems using cluster-based Δ -learning protocols

SZABÓ ANDRÁS

Supervisors:

JÁNOS DARU

BENCE BALÁZS MÉSZÁROS

Department of Organic Chemistry

Eötvös Loránd University



EÖTVÖS LORÁND
UNIVERSITY | BUDAPEST

Contents

1	Introduction	1
1.1	Coupled Cluster Molecular Dynamics	1
1.2	Machine Learning Interatomic Potentials	5
1.2.1	Behler and Parrinello’s symmetry functions	5
1.2.2	Atomic Cluster Expansion - ACE	6
1.3	Density Functional Theory	10
1.3.1	Hohenberg-Kohn Theorem	10
1.3.2	Kohn-Sham Equations	10
1.3.3	Minimization and Kohn-Sham Potential	11
1.3.4	Self-Consistent Kohn-Sham Equations	11
1.3.5	Total Energy Expression	11
1.3.6	Summary	11
1.4	The SCAN Functional	11
1.4.1	Exchange-Correlation Energy Functional	12
1.4.2	Exchange and Correlation Components	12
1.4.3	Exchange Component	13
1.4.4	Correlation Component	13
1.4.5	Summary	13
1.5	Coupled Cluster Methods	13
1.6	Many-Body Expansion (MBE)	15
1.6.1	MB-pol	15
1.7	Path-Integral Molecular Dynamics	18
2	Methodology	20
2.1	Molecular Dynamics Simulations	20
2.2	Training Procedure	20
2.3	Hardwer Specifications	20
2.4	Definition of Spatial correlation functions	21
2.4.1	Radial Distribution Function Between Two Atoms	21

2.4.2	Angular Distribution Function	21
2.4.3	Tetrahedral Order Parameter	22
2.5	Definition of Error Metrics	22
2.5.1	Mean Absolute Error of Forces: $\text{MAE}(F)$	23
2.5.2	Root Mean Square Error of Forces: $\text{RMSE}(F)$	23
2.5.3	Maximum Absolute Error of Forces: $\text{MAX}(F)$	23
2.5.4	Pearson Correlation Coefficient of Forces: $r(F)$	23
2.5.5	Mean Absolute Error of Energies: $\text{MAE}(E)$	23
2.5.6	Root Mean Square Error of Energies: $\text{RMSE}(F)$	23
2.5.7	Maximum Absolute Error of Energies: $\text{MAX}(E)$	23
2.5.8	Pearson Correlation Coefficient of Energies: $r(E)$	24
3	Results and Discussion	25
3.1	Overview	25
3.2	Preliminary Studies	25
3.2.1	Effect of cluster-based learning	25
3.2.2	Effect of training without forces	26
3.3	Optimization of the Δ -Learning Protocol	28
3.3.1	Cutoff Radius	28
3.3.2	Cluster Size and Number of Reference Calculations	29
3.3.3	Comparison with MACE	31
3.4	Performance of the Machine Learning Potentials	32
3.4.1	Radial Distribution Functions	32
3.4.2	Angular Distribution Functions	35
3.4.3	Tetrahedral Order Parameter	36
3.5	Summary and Outlook	38
	Acknowledgements	39
	References	40

Chapter 1

Introduction

1.1 Coupled Cluster Molecular Dynamics

Understanding chemical reactions at the molecular scale is a fundamental goal across all branches of chemistry, providing significant advantages to both industry and academia. This understanding accelerates drug discovery, optimization of catalysts, advancement of energy storage solutions, and promotes the development of environmentally sustainable chemical processes. Yet, most relevant reactions occur in the condensed phase, posing substantial challenges for accurate in-silicon modeling. Although static calculations using density functional theory (DFT) or wavefunction theory (WFT) can achieve high accuracy for gas-phase electronic energies, incorporating solvent effects through continuum solvation models introduces considerable errors. This approach fails to account for specific solvent interactions and lacks dynamic information while complicating the accurate treatment of vibrational degrees of freedom and conformational flexibility.

Alternatively, another technique, *ab initio* molecular dynamics, can handle complex systems — such as solutes and surfaces — while providing valuable dynamical insights. However, the substantial computational demands of these simulations often necessitate compromises in sampling time and the theoretical level. Common simplifications include employing a lower level of theory, such as (meta-)generalized gradient approximation or (meta-)GGA, and neglecting nuclear quantum effects (NQE).

A recent technique introduced by Daru et al. has significantly advanced *ab initio* molecular dynamics (AIMD) by both lowering computational costs and enhancing accuracy.¹ This method, known as coupled cluster molecular dynamics (CCMD), integrates modern local coupled cluster (local CC) calculations with machine learning while effectively incorporating nuclear quantum effects (NQE) through converged path integral molecular dynamics.² Unlike traditional approaches, this technique does not rely on reference electronic structure calculations for periodic systems at the highest level. Instead, it uses finite clusters derived from

periodic boundary condition (PBC) simulations (see Figure 1.1). This enables the local CC potential energy surface (PES) to be learned based on these clusters and then applied within a PBC simulation using a machine learning potential.

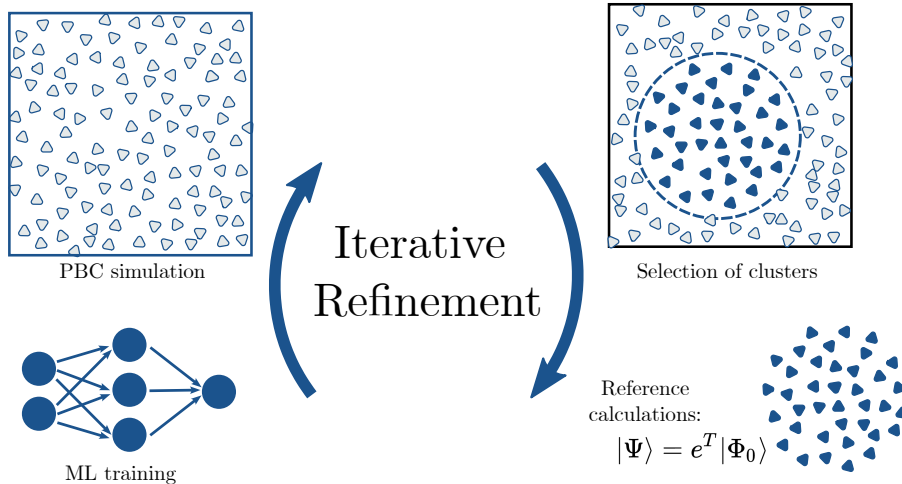


Figure 1.1: Schematic representation of the original CCMD workflow, show the cycle of sampling, machine learning, and simulation until the results are satisfying

This innovative technique was initially tested on liquid water, a challenging system for AIMD simulations. Inadequate treatment of NQEs and high-level electron correlation in simulations often result in water structures and dynamics that resemble amorphous ice. In contrast, CCMD simulations achieved excellent agreement with experimental data, accurately reproducing both structural (radial distribution functions) and dynamical properties (such as diffusion constants and orientational relaxation times) (see Figure 1.2). Recently, this highly precise approach was also applied to heavy water, allowing researchers to explore isotope effects with accurate NQE treatment.³

With its exceptional agreement with experimental properties, the CCMD approach holds substantial promise for accurately modeling reactions in the condensed phase. However, its broader application is constrained by the substantial computational demands of the reference calculations, which in the original study required over 3.5 million core hours. Thus, optimizing the method and achieving considerable speed improvements are critical for extending its usability to a wider range of systems.

To achieve this, we focused on two primary areas for optimization. The first is to adopt DFT as a more economical baseline method instead of the originally used DLPNO-MP2,⁵ which served as the initial level in a two-step Δ -learning protocol. In the original work, the

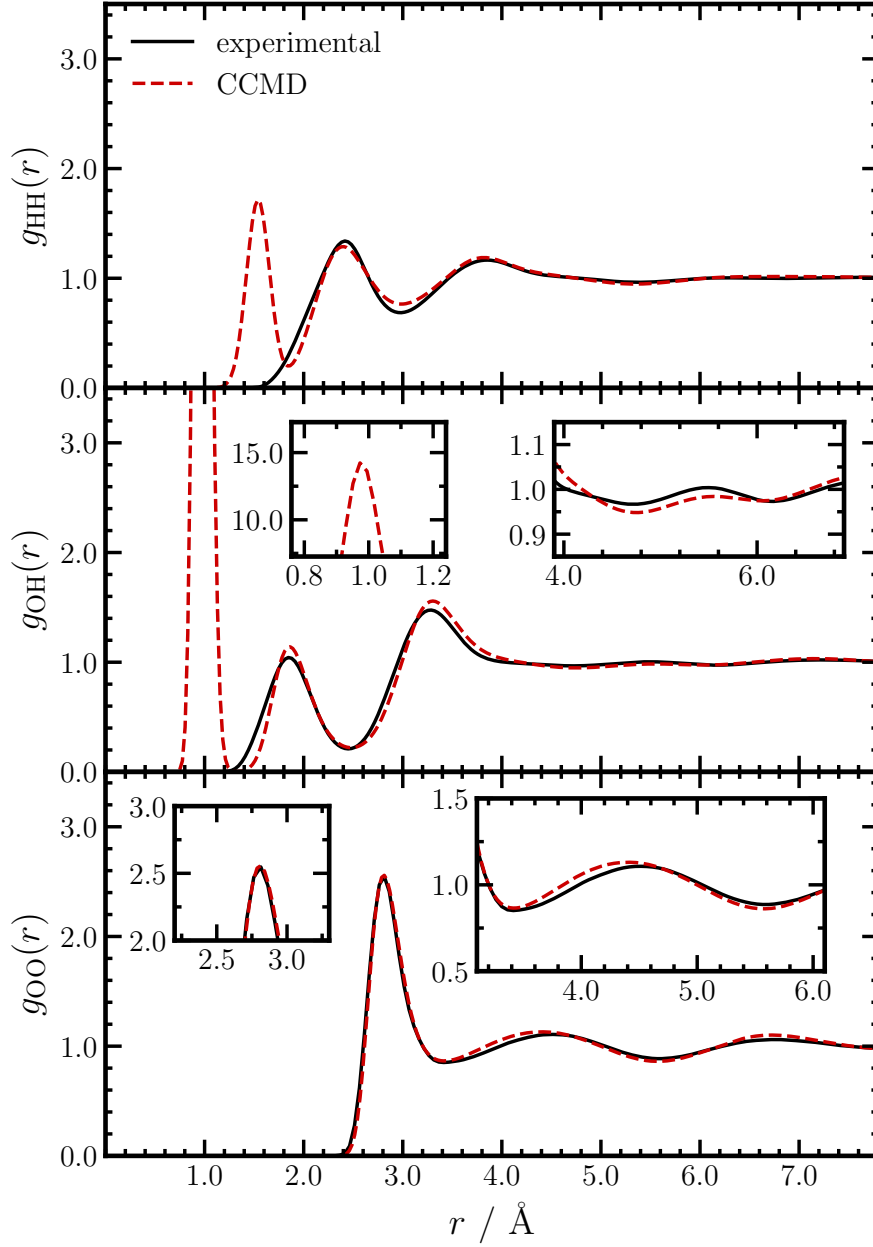


Figure 1.2: Radial distribution functions of water including nuclear quantum effects from path integral CCMD simulations¹ at 298 K (red dashed lines) compared to experimental data⁴ from neutron diffraction ($g_{\text{HH}}(r)$ and $g_{\text{OH}}(r)$) and X-ray diffraction ($g_{\text{OO}}(r)$). The RDFs have been computed using a radial bin size of $0.01 \text{ \AA}$ as depicted and are not smoothened.

difference between DLPNO-MP2 and an analytical two-body potential was initially fitted. We aim to substitute these reference calculations with an approximate periodic electronic structure calculation. Although there are implementations for periodic MP2 or RPA, we are considering more economical SCF methods, such as HF, hybrid functionals, or (meta-)GGAs. Although our updated protocol does not depend on a specific DFT functional, we selected the meta-GGA SCAN functional for our initial studies. SCAN satisfies all known constraints for a meta-GGA functional. It also accurately describes complex bonding environments and van der Waals interactions without relying on bonded information or empirical parameters. Still, SCAN is economical enough to run AIMD simulations for tens or hundreds of picoseconds (ps).

The second optimization step centers on transferring to the coupled cluster PES. In the original study, DLPNO-CCSD(T)⁶ calculations were performed on approximately 3,000 clusters, each composed of 64 water molecules. To improve efficiency, we are currently testing the LNO-CCSD(T)⁷⁻⁹ method which is the TDK topic of Maté Kígyósi working in our laboratory. Finally, the focus of this study is to investigate potential savings by reducing *both the size of the water clusters and the number of required calculations* by using more advanced machine learning potentials.

In summary, our new protocol combines two machine learning (ML) potentials: a baseline model trained on periodic reference data using SCAN and a cluster-based Δ -learning model that incorporates electron correlation and transfers to the CC PES (see Figure 1.3). Additionally, given that these calculations are very expensive, we are using the many body expansion models MB-pol and MB-SCAN as surrogate models for CC and SCAN, respectively. This not only allows the saving of computational time, but provides a CC-level periodic reference through MB-pol PIMD simulations.

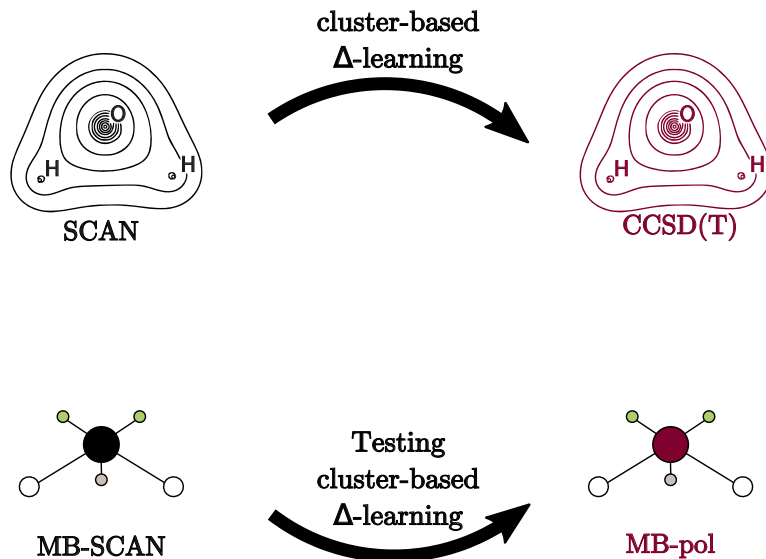


Figure 1.3: Our proposed methodology, Δ -learning on a difference between MB-SCAN and CCSD(T) water clusters potential surfaces, to achieve a correction method to the SCAN function, which can produce similar results as an MB-pol calculation in periodic water evaluation.

1.2 Machine Learning Interatomic Potentials

Machine learning models have been evolving at an unprecedented pace in recent years, driven by advancements in algorithms, computing power, and data availability. Machine learning interatomic potentials use descriptors (usually referred to as fingerprints) to capture the local atomic environment, allowing the prediction of atomic properties with high accuracy. Their applications include several fields, e. g. biomolecular MD simulations,^{10,11} and quantum-chemical calculations with deep tensor neural networks.¹² We see studies to reduce the gap between ab initio and classic force field methods.¹³ There is also an effort to develop universal machine-learning force fields across organic molecules.¹⁴

1.2.1 Behler and Parrinello’s symmetry functions

The first atomistic machine learning potentials was pioneered by Behler and Parrinello.¹⁵ Their symmetry functions form the core of their neural network interatomic potential, enabling accurate molecular simulations. These functions are designed to be invariant under symmetry operations such as translations and rotations, which is crucial for ensuring the neural network can learn the potential energy surface in a general, transferable way.

The two-body symmetry functions are:

$$G_1 = \sum_j f_c(r_{ij}) \quad (1.1)$$

$$G_2 = \sum_j \exp(-\eta(r_{ij} - r_s)^2) f_c(r_{ij}) \quad (1.2)$$

where $f_c(r_{ij})$ is a smooth cutoff function and η controls the width of the Gaussian function.

For three-body interactions, the symmetry functions are:

$$G_3 = 2(1 - \zeta) \sum_{jk} (1 + \lambda \cos \theta_{ijk})^\zeta \exp(-\eta(r_{ji}^2 + r_{kj}^2 + r_{ki}^2)) f_c(r_{ji}) f_c(r_{kj}) f_c(r_{ki}) \quad (1.3)$$

$$G_4 = 2(1 - \zeta) \sum_{jk} (1 + \lambda \cos \theta_{ijk})^\zeta \exp(-\eta(r_{ji}^2 + r_{ki}^2)) f_c(r_{ji}) f_c(r_{ki}) \quad (1.4)$$

These functions describe the local atomic environments, including both radial and angular dependencies. The angular terms, $(1 + \lambda \cos \theta)^\zeta$, can be expanded using Legendre or Chebyshev polynomials, providing a flexible description of angular interactions.

The key importance of symmetry is that it ensures the descriptors are invariant to rotations, translations, and permutations of identical atoms. This symmetry property allows the neural network to generalize across different systems and configurations, ensuring accurate predictions of atomic interactions regardless of the atomic ordering or orientation. Behler and Parinello's symmetry functions are limited to third order terms (angles), therefore two different structures can share the same symmetry function set.

1.2.2 Atomic Cluster Expansion - ACE

In order to overcome the above-mentioned limitation, Drautz introduced the atomic cluster expansion (ACE), a complete and interpretable descriptor that scales efficiently with the number of neighbors, offering an alternative to previous machine learning potentials.¹⁶ To provide a theoretical background, we start with the definition of a single-bond basis function between atoms i and j , defined as a product of radial function $\mathbf{R}_{nl}$ and the spherical harmonics $\mathbf{Y}_{lm}$:¹⁷

$$\phi_{\mu_i \mu_j n l m}(\mathbf{r}_{ij}) = \mathbf{R}_{nl}^{\mu_i \mu_j}(r_{ji}) \mathbf{Y}_{lm}(\hat{\mathbf{r}}_{ji}) \quad (1.5)$$

Here, μ_i and μ_j denote the index of chemical species of the two atoms, and nlm specify the type of basis function (like quantum numbers). r_{ji} is the distance between atoms, $\mathbf{r}_{ji}$ is the vector from atom i to atom j , and $\hat{r}_{ji} = \mathbf{r}_{ji}/r_{ji}$ is the bond direction. The radial part is a linear combination of $k_{\max}^{\mu_i\mu_j}$ number of radial basis functions:

$$\mathbf{R}_{nl}(r) = \sum_{k=1}^{k_{\max}^{\mu_i\mu_j}} c_{nlk}^{\mu_i\mu_j} g_k(r), \quad (1.6)$$

where $c_{nlk}^{\mu_i\mu_j}$ are the radial expansion coefficients, and $g_k(r)$ denote the radial basis functions, which are smoothly taken to zero outside the cutoff distance r_{cut} .

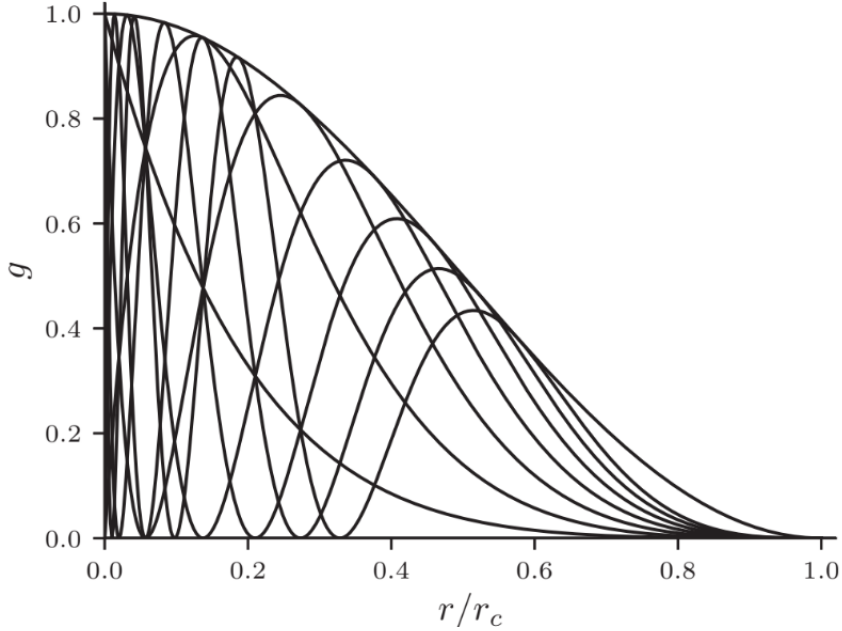


Figure 1.4: Chebyshev polynomial radial functions from g_1 to g_8 .¹⁶

Different forms of the radial basis functions can be used, Chebyshev polynomials and the Bessel functions are often used. With the single-bond basis function defined, we derive an atomic base as the sum of single-bond basis functions over the neighbours j of the central atom i :

$$A_{i\mu nlm} = \sum_{j \neq i} \phi_{\mu_i\mu_j nlm} \quad (1.7)$$

Now we derive products of the atomic basis functions as permutation-invariant many-body basis functions of order ν :

$$\mathbf{A}_{i\mu nlm}^\nu = \prod_t^\nu A_{i_t\mu_t n_t l_t m_t} \quad (1.8)$$

From this, we construct a rotationally invariant or equivariant basis function $\mathbf{B}_{i\mu nL}^\nu$, by integrating $\mathbf{A}_{i\mu nlm}^\nu$ over all rotation in three-dimensional space. In practice, this is done by multiplication by generalized Glebsch-Gordan coefficients. Now, an atomic property $\varphi_i^{(p)}$ is expanded as a linear combination of rotationally invariant basis functions:

$$\varphi_i^{(p)} = \sum_{\nu\mu nlm} \tilde{c}_{i\mu nlm}^{(p)} \mathbf{A}_{i\mu nlm}^\nu = \sum_{\mu nL} c_{\mu nL}^{(p)} \mathbf{B}_{i\mu nL}^\nu \quad (1.9)$$

The atomic energy can be expressed either directly by this property as done in linear ACE

$$E_i = \varphi_i^{(1)}, \quad (1.10)$$

For faster convergence, a Finnis-Sinclair-type expansion can be applied by partitioning the energy into contributions from two different properties

$$E_i = \varphi_i^{(1)} + \sqrt{\varphi_i^{(2)}} \quad (1.11)$$

Finally, the total energy is the sum of atomic energies:

$$E = \sum_i E_i \quad (1.12)$$

While for training the rotationally invariant or equivariant basis set is used, the evaluation can be made significantly faster by using the permutationally invariant basis functions $\mathbf{A}_{i\mu nlm}$. In this case, the expansion coefficients $\tilde{c}_{i\mu nlm}$ ensure the rotational invariance.

$$\varphi_i^{(p)} = \sum_{\nu\mu nlm} \tilde{c}_{i\mu nlm}^{(p)} \mathbf{A}_{i\mu nlm}^\nu = \sum_{\nu\mu nlm} \tilde{c}_\nu^{(p)} \prod_{t=1}^\nu A_{i_t\mu_t n_t l_t m_t} \quad (1.13)$$

Considering N interacting neighbors with order ν , the original many-atom expansion incurs an $O(N^\nu)$ scaling, leading to prohibitive computational costs. In ACE however, the cost of calculating an atomic property $\varphi_i^{(p)}$ scales linearly with both the number of neighbors and the body order ($O(N\nu)$). This is illustrated in the next equation, where on the left side we have the classical atom-order expansion with the $\sum_{j_1, \dots, j_\nu}$ summing over all possible ν -atom clusters involving the central atom i in a brute force manner. This is rewritten in the middle part of

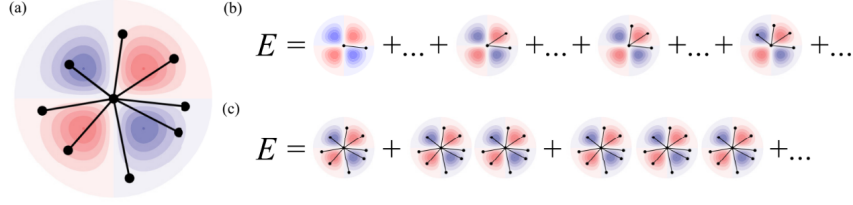


Figure 1.5: Illustration of the atomic cluster expansion. (a) Basis function evaluated for single-bond basis functions to neighbors of the central atom. (b) Atomic cluster expansion as a sum of pair, three-body, four-body, etc., contributions illustrating Eq. 1.14, left side. (c) Atomic cluster expansion as polynomial of the atomic base illustrating Eq. 1.14, right side. Figure adapted from the original ACE publication.¹⁶

the equation to the product of a single sum of the single-bond basis functions, therefore the product of atomic basis functions, as in the right side.

$$\sum_{\nu\mu nlm} \tilde{c}_{\nu}^{(p)} \sum_{j_1, \dots, j_{\nu}} \prod_{t=1}^{\nu} \phi_{\mu_i \mu_j n l m} = \sum_{\nu\mu nlm} \tilde{c}_{\nu}^{(p)} \prod_{t=1}^{\nu} \sum_{j \neq i} \phi_{\mu_i \mu_j n l m} = \sum_{\nu\mu nlm} \tilde{c}_{\nu}^{(p)} \prod_{t=1}^{\nu} A_{i_t \mu_t n_t l_t m_t} \quad (1.14)$$

This is a critical point, as the improvement of computational efficiency allows the calculation of higher-order terms easily (e.g. up to order 7).¹⁸

The accuracy and applicability of the method were first demonstrated on copper.¹⁶ But there are many applications for different materials like magnesium,¹⁹ or quantum accurate large scale simulation of carbon.²⁰ The method was also applied to describe water and ice,²¹ hybrid materials like Pt–Rh catalysts.²² There are also examples of using them for organic molecules’ dihedral torsional simulation.,²³ moreover it was also applied to reactive systems, including a Diels-Alder.²⁴ It can also be applicable to describe catalytic interaction between hydrogen and platina.²⁵ The ACE model was further developed to achieve even higher accuracy, including the Multi-ACE (MACE) with message passing,^{14,26,27} ACE with cartesian base²⁸ or graph-ACE.²⁹ The above-cited works have demonstrated that atomic cluster expansion can achieve accuracy comparable to neural network-based machine-learning approaches while retaining physical and chemical interpretability.

1.3 Density Functional Theory

Kohn-Sham theory provides a practical approach to solving the many-electron Schrödinger equation in Density Functional Theory (DFT). It replaces the complex many-body system with an auxiliary system of non-interacting particles that has the same ground-state electron density as the real system.

1.3.1 Hohenberg-Kohn Theorem

The Hohenberg-Kohn theorem states that the ground-state energy E of an electron system is a unique functional of the electron density $\rho(\mathbf{r})$:

$$E[\rho] = T[\rho] + V_{\text{ext}}[\rho] + U[\rho] \quad (1.15)$$

where:

- $T[\rho]$: kinetic energy functional,
- $V_{\text{ext}}[\rho] = \int \rho(\mathbf{r}) v_{\text{ext}}(\mathbf{r}) d\mathbf{r}$: external potential energy,
- $U[\rho] = \frac{1}{2} \int \int \frac{\rho(\mathbf{r})\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} d\mathbf{r} d\mathbf{r}'$: electron-electron interaction energy.

1.3.2 Kohn-Sham Equations

To simplify, Kohn and Sham proposed using non-interacting electrons that replicate the exact density of the interacting system. The energy functional is then rewritten as:

$$E[\rho] = T_s[\rho] + V_{\text{ext}}[\rho] + U[\rho] + E_{\text{xc}}[\rho] \quad (1.16)$$

where:

- $T_s[\rho]$: kinetic energy of the non-interacting system,
- $E_{\text{xc}}[\rho]$: exchange-correlation energy, accounting for the difference between $T[\rho]$ and $T_s[\rho]$, as well as for electron correlation effects.

1.3.3 Minimization and Kohn-Sham Potential

Minimizing $E[\rho]$ with respect to ρ leads to the *Kohn-Sham potential* $v_{\text{KS}}(\mathbf{r})$:

$$v_{\text{KS}}(\mathbf{r}) = v_{\text{ext}}(\mathbf{r}) + v_{\text{H}}(\mathbf{r}) + v_{\text{xc}}(\mathbf{r}) \quad (1.17)$$

where:

- $v_{\text{H}}(\mathbf{r}) = \int \frac{\rho(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d\mathbf{r}'$: Hartree potential,
- $v_{\text{xc}}(\mathbf{r}) = \frac{\delta E_{\text{xc}}[\rho]}{\delta \rho(\mathbf{r})}$: exchange-correlation potential.

1.3.4 Self-Consistent Kohn-Sham Equations

The electrons in the non-interacting system follow single-particle Kohn-Sham equations:

$$\left(-\frac{\hbar^2}{2m} \nabla^2 + v_{\text{KS}}(\mathbf{r}) \right) \phi_i(\mathbf{r}) = \epsilon_i \phi_i(\mathbf{r}) \quad (1.18)$$

where $\phi_i(\mathbf{r})$ are the Kohn-Sham orbitals, and the electron density is:

$$\rho(\mathbf{r}) = \sum_{i=1}^N |\phi_i(\mathbf{r})|^2 \quad (1.19)$$

1.3.5 Total Energy Expression

The total energy can be expressed as:

$$E[\rho] = \sum_{i=1}^N \epsilon_i - \frac{1}{2} \int \int \frac{\rho(\mathbf{r})\rho(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d\mathbf{r} d\mathbf{r}' + E_{\text{xc}}[\rho] - \int \rho(\mathbf{r}) v_{\text{xc}}(\mathbf{r}) d\mathbf{r} \quad (1.20)$$

1.3.6 Summary

Kohn-Sham theory reduces the many-body problem to a self-consistent field problem of non-interacting particles under an effective potential, where the accuracy depends on the choice of the exchange-correlation functional $E_{\text{xc}}[\rho]$.

1.4 The SCAN Functional

The SCAN functional is a meta-Generalized Gradient Approximation (meta-GGA) exchange-correlation functional used in Density Functional Theory (DFT). It improves upon standard

GGA functionals by incorporating the kinetic energy density $\tau(\mathbf{r})$, in addition to the electron density $\rho(\mathbf{r})$ and its gradient $\nabla\rho(\mathbf{r})$. This additional dependence on $\tau(\mathbf{r})$, which reflects the local variation of the electron density, enhances the accuracy of the functional, particularly for systems with rapidly varying electron densities. The SCAN functional is specifically designed to satisfy all known exact constraints for meta-GGA functionals, including the uniform electron gas limit and correct asymptotic behavior for exchange and correlation energies. Overall, SCAN represents a significant advancement in approximating the exchange-correlation functional in DFT, bridging the gap between standard GGAs and exact functionals.

1.4.1 Exchange-Correlation Energy Functional

The SCAN functional's exchange-correlation energy $E_{\text{xc}}^{\text{SCAN}}[\rho]$ is expressed as:

$$E_{\text{xc}}^{\text{SCAN}}[\rho] = \int \rho(\mathbf{r}) \varepsilon_{\text{xc}}^{\text{SCAN}}(\rho(\mathbf{r}), \nabla\rho(\mathbf{r}), \tau(\mathbf{r})) d\mathbf{r} \quad (1.21)$$

where:

- $\varepsilon_{\text{xc}}^{\text{SCAN}}$: SCAN exchange-correlation energy density per particle,
- $\rho(\mathbf{r})$: electron density,
- $\nabla\rho(\mathbf{r})$: gradient of the electron density,
- $\tau(\mathbf{r}) = \frac{1}{2} \sum_{i=1}^N |\nabla\phi_i(\mathbf{r})|^2$: kinetic energy density.

1.4.2 Exchange and Correlation Components

The SCAN functional separates the exchange and correlation components:

$$E_{\text{xc}}^{\text{SCAN}}[\rho] = E_{\text{x}}^{\text{SCAN}}[\rho] + E_{\text{c}}^{\text{SCAN}}[\rho] \quad (1.22)$$

Each component is constructed to satisfy exact constraints and to interpolate between different physical limits, such as:

- *Uniform electron gas* limit,
- *Non-bonded interactions* (van der Waals interactions),
- *Hydrogenic atoms* limit.

1.4.3 Exchange Component

The exchange energy $E_x^{\text{SCAN}}[\rho]$ depends on the local spin density and gradient of the density, along with the reduced density gradient $s(\mathbf{r})$:

$$s(\mathbf{r}) = \frac{|\nabla\rho(\mathbf{r})|}{2k_F(\mathbf{r})\rho(\mathbf{r})} \quad (1.23)$$

where $k_F(\mathbf{r}) = (3\pi^2\rho(\mathbf{r}))^{1/3}$ is the local Fermi wavevector. SCAN tunes the dependence on $s(\mathbf{r})$ to recover the correct limits and improve accuracy for real materials.

1.4.4 Correlation Component

The correlation energy $E_c^{\text{SCAN}}[\rho]$ in SCAN incorporates the kinetic energy density $\tau(\mathbf{r})$ to distinguish between regions of electron localization (bonding, non-bonding, or metallic). SCAN's correlation functional uses:

$$\alpha(\mathbf{r}) = \frac{\tau(\mathbf{r}) - \tau^{\text{W}}(\mathbf{r})}{\tau^{\text{UEG}}(\mathbf{r})} \quad (1.24)$$

where:

- $\tau^{\text{W}}(\mathbf{r}) = \frac{|\nabla\rho(\mathbf{r})|^2}{8\rho(\mathbf{r})}$: von Weizsäcker kinetic energy density,
- $\tau^{\text{UEG}}(\mathbf{r}) = \frac{3}{10}(3\pi^2)^{2/3}\rho(\mathbf{r})^{5/3}$: kinetic energy density of the uniform electron gas.

The parameter $\alpha(\mathbf{r})$ allows SCAN to interpolate between different electron localization regimes, enhancing its accuracy in complex bonding environments.

1.4.5 Summary

The SCAN functional is a meta-GGA functional designed to satisfy known constraints for exchange and correlation energies, with dependence on the electron density $\rho(\mathbf{r})$, density gradient $\nabla\rho(\mathbf{r})$, and kinetic energy density $\tau(\mathbf{r})$. This makes it an accurate and versatile functional for a wide range of materials.

1.5 Coupled Cluster Methods

In the Coupled-Cluster (CC) method, electron correlation is approximated by incorporating contributions from excited determinants. This is achieved through the definition of a cluster

operator for an n -electron system:

$$\hat{T} = \hat{T}_1 + \hat{T}_2 + \hat{T}_3 + \cdots + \hat{T}_n, \quad (1.25)$$

where each term, $\hat{T}_i$, generates all possible i th excitations from the reference Hartree-Fock (HF) determinant Φ_0 as follows:

$$\begin{aligned} \hat{T}_1 \Phi_0 &= \sum_{i,a} t_i^a \Phi_i^a, \\ \hat{T}_2 \Phi_0 &= \sum_{i<j} \sum_{a<b} t_{ij}^{ab} \Phi_{ij}^{ab}, \\ &\vdots \end{aligned} \quad (1.26)$$

Amplitude coefficients marked as t , quantifying the contribution of each excitation. The CC wavefunction, Ψ_{CC} , is expressed through the exponential ansatz:

$$\Psi_{CC} = e^{\hat{T}} \Phi_0. \quad (1.27)$$

The energy corresponding to the coupled-cluster wavefunction, E_{CC} , is determined by:

$$E_{CC} = \langle \Phi_0 | \hat{H} e^{\hat{T}} | \Phi_0 \rangle. \quad (1.28)$$

Coupled cluster (CC) methods, with CCSD(T) as a prominent example, represent a cornerstone of high-accuracy quantum chemistry, often referred to as the "gold standard" for achieving prediction uncertainties of approximately 1 kcal/mol for small molecular systems. Despite its precision, CCSD(T) is computationally demanding, with scaling that becomes prohibitive as molecular size increases, typically restricting its direct application to systems containing no more than 20–30 atoms. To address these challenges, significant research has focused on enhancing the scalability of CCSD(T) by leveraging the inherently short-range nature of electron correlation. Among the strategies developed are local correlation techniques (DLPNO,⁶ LNO⁷), including those based on natural orbitals (NOs) and pair natural orbitals (PNOs). These approaches reduce the computational expense of CCSD(T), thereby enabling its application to larger systems and expanding its practical utility in quantum chemistry.

1.6 Many-Body Expansion (MBE)

The energy of a molecular complex can be expressed rigorously as a sum of N -body contributions:

$$E_N(1, \dots, N) = \sum_{i=1}^N V^{1B}(i) + \sum_{i<j}^N V^{2B}(i, j) + \sum_{i<j<k}^N V^{3B}(i, j, k) + \dots + V^{NB}(1, \dots, N). \quad (1.29)$$

In this expression, $V^{1B}(i) = E(i) - E_{\text{eq}}(i)$ represents the one-body (1B) energy, which corresponds to the energy required to deform a monomer from its equilibrium geometry. The higher-order terms, V^{nB} , represent n -body interactions and are defined recursively as:

$$\begin{aligned} V^{nB}(1, \dots, n) = & E_n(1, \dots, n) - \sum_i V^{1B}(i) - \sum_{i<j} V^{2B}(i, j) - \dots \\ & - \sum_{i<j<\dots<n-1} V^{(n-1)B}(i, j, \dots, (n-1)). \end{aligned} \quad (1.30)$$

This approach allows the interaction energy of a system to be decomposed into contributions from clusters of different sizes, such as monomers, dimers, trimers, and so on. While the method provides an exact N -body solution when all terms are included, it is often practical to truncate the expansion. For systems like water, the series converges rapidly, and typically only the first three or four terms are needed. Truncation is essential in the many-body expansion (MBE) approach to achieve a balance between computational cost and accuracy.

1.6.1 MB-pol

MB-pol combines the coupled-cluster approach for electronic correlations with the many-body expansion (MBE) approach for intermolecular interactions. The following equations and figure collectively describe the modeling approach of the MB-pol framework, a sophisticated method for accurately capturing interactions in water systems.³⁰

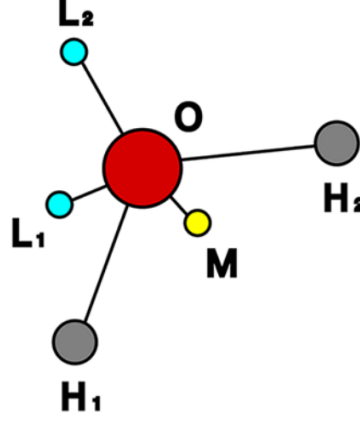


Figure 1.6: Schematic representation of a MB-pol water molecule, showing the location of the fictitious site (M) and lone-pair sites (L1 and L2) relative to the O and H atoms. Note that the L1 and L2 sites occupy opposite positions on a plane perpendicular to the molecular plane containing the O and H atoms (i.e., they point below and above the molecular plane, respectively).³⁰

Figure 1.6 schematically illustrates an MB-pol water molecule. The key features include the placement of fictitious and lone-pair sites that mimic the electrostatic and polarization characteristics of the molecule. Specifically, the lone-pair sites (L1 and L2) are positioned on a plane perpendicular to the molecular plane that contains the oxygen (O) and hydrogen (H) atoms. These sites lie above and below this molecular plane, enabling the model to account for the electronic distribution and geometry of water, crucial for accurately representing intermolecular forces.

The total energy of a system of N water molecules, as expressed in Equation 1.31, is hierarchically decomposed into single-body (1B), two-body (2B), three-body (3B), and four-body (4B) interaction contributions. Additionally, a polarization term, V_{pol} , accounts for the interactions of molecular dipoles due to electric fields generated by neighboring molecules. This partitioning simplifies the complex energy calculation by isolating contributions from smaller clusters of molecules, leveraging physical locality and symmetry properties.

$$\begin{aligned}
 E_N(1, \dots, N) = & \sum_{i=1}^N V^{\text{1B}}(i) + \sum_{i<j}^N V^{\text{2B}}(i, j) \\
 & + \sum_{i<j<k}^N V^{\text{3B}}(i, j, k) + \sum_{i<j<k<l}^N V^{\text{4B}}(i, j, k, l) \\
 & + V_{\text{pol}}(1, \dots, N)
 \end{aligned} \tag{1.31}$$

Equation 1.32 introduces the use of permutationally invariant polynomials (PIP) to represent n -body interaction potentials. These polynomials ensure invariance under permutations of molecular indices. The expansion relies on basis functions (η_l) that depend on geometrical descriptors of the molecular system and coefficients (c_l) optimized during training to accurately model the interactions.

$$V_{\text{PIP}}^{n\text{B}}(M_1, M_2, \dots, M_n) = \sum_{l=1}^L c_l \cdot \eta_l(\xi_1, \xi_2, \dots, \xi_N) \quad (1.32)$$

To further refine the energy components, Equation 1.33 decomposes the two-body potential, $V^{2\text{B}}$, into three terms: $V_{\text{ML}}^{2\text{B}}$, a machine-learned contribution; $V_{\text{perm}}^{2\text{B}}$, which captures permanent electrostatic interactions; and $V_{\text{disp}}^{2\text{B}}$, a dispersion term. This separation facilitates targeted optimization, combining established physical models for electrostatics and dispersion with a machine-learned approach to capture the remaining complexity.

$$V^{2\text{B}} = V_{\text{ML}}^{2\text{B}} + V_{\text{perm}}^{2\text{B}} + V_{\text{disp}}^{2\text{B}} \quad (1.33)$$

Dispersion interactions are described in Equation 1.34. These long-range interactions are modeled using a term proportional to $-C_{6,ij}/R_{ij}^6$, which represents the attractive van der Waals forces between sites i and j in molecules M_1 and M_2 . A damping function $f(\delta_{ij}R_{ij})$ is included to regulate the behavior at short ranges, ensuring physical realism and avoiding divergence.

$$V_{\text{disp}}^{2\text{B}} = \sum_{\substack{i \in M_1 \\ j \in M_2}} -f(\delta_{ij}R_{ij}) \frac{C_{6,ij}}{R_{ij}^6} \quad (1.34)$$

Polarization, a critical many-body effect, is expressed in Equation 1.35 as a sum of contributions from two-body, three-body, four-body, and higher-order interactions. This term is essential for accurately capturing the behavior of water molecules in diverse environments, including liquid water and ice.

$$V_{\text{pol}} = V_{\text{pol}}^{2\text{B}} + V_{\text{pol}}^{3\text{B}} + V_{\text{pol}}^{4\text{B}} + V_{\text{pol}}^{>4\text{B}} \quad (1.35)$$

One of our main focuses with this method was calculations with water. As principals for monomers and dimers³¹ calculation was introduced first when it was created, later added trimers and the usability of it in liquid phase.³² Then extended its for many more properties in solid a liquid phase³³.³¹ Other than water, potentials for alkali³⁴ and halogen³⁵ ions in water

have also been developed. It has also been extended to describe CH_4 ,³⁶ CO_2 ,³⁷ and N_2O_5 ,³⁸ in aqueous solution. The same approach was applied to fit many body potentials on DFT reference data. The MB-SCAN is such a model fitted on the SCAN functional.³⁹ While these are highly accurate potentials, since molecular entities need to be predefined in a simulation, they cannot describe reactive systems.

1.7 Path-Integral Molecular Dynamics

Path Integral Molecular Dynamics (PIMD)² is a computational method that incorporates quantum mechanical effects, such as nuclear delocalization, into molecular simulations by representing particles as a series of interconnected replicas or "beads". This method relies on the isomorphism between the quantum system and a classical ring-polymer system, where the beads are connected by harmonic springs, mimicking the quantum mechanical behavior of nuclei. In quantum statistical mechanics, such as PIMD, we use the *density matrix* ρ to describe the state of a quantum system at thermal equilibrium. For a system with Hamiltonian H at temperature T , the density matrix is:

$$\rho = e^{-\beta H} \quad (1.36)$$

where: $\beta = \frac{1}{k_B T}$ is the inverse temperature, and H is the Hamiltonian, which contains information about both the kinetic and potential energy of the system. The *partition function* Z , which is central to calculating thermodynamic properties, is the trace of the density matrix:

$$Z = \text{Tr}(\rho) = \text{Tr}(e^{-\beta H}) \quad (1.37)$$

This partition function Z essentially sums over all possible quantum states, weighted by their Boltzmann factors $e^{-\beta E}$, and contains all statistical information about the system. To compute the partition function, we typically represent the density matrix in the *position basis* $|x\rangle$. This lets us write the partition function as an integral over position states:

$$Z = \int dx \langle x | e^{-\beta H} | x \rangle \quad (1.38)$$

Using the Trotter decomposition, we can split the Hamilton operator into kinetic and potential energy operators for small steps of imaginary time.

$$e^{-\beta H} = \lim_{N \rightarrow \infty} \left((e^{-\Delta\tau T} e^{-\Delta\tau V})^N \right) \quad (1.39)$$

This becomes an approximation using finite N . Throughout further derivation, we get:

$$Z \propto \prod_{i=1}^N e^{-\frac{m(x_i - x_{i-1})^2}{2\hbar^2 \Delta\tau}} e^{-\Delta\tau V(x_i)} \quad (1.40)$$

So in the end we get, that adjacent beads are connected via a harmonic potential (springs).

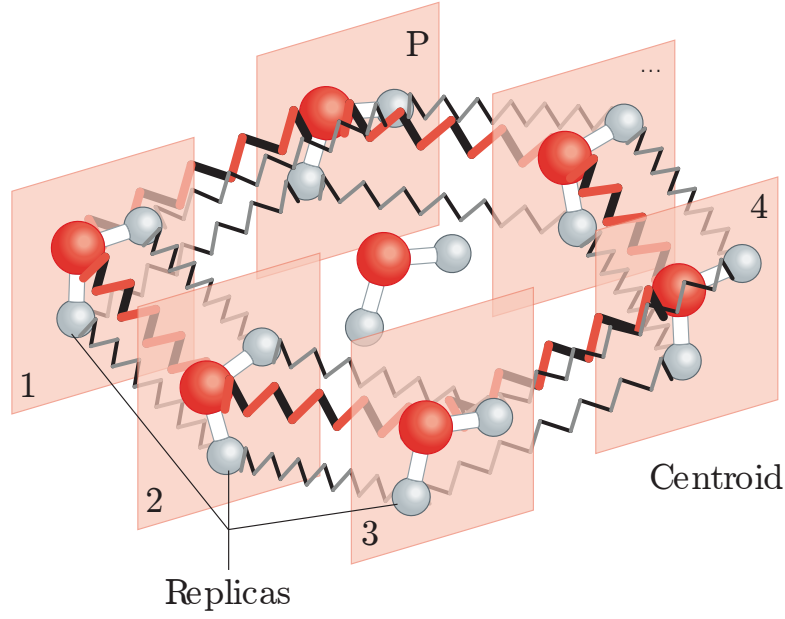


Figure 1.7: Schematic representation for a path integral simulation of water with six beads, the adjacent one being connected by harmonic springs.⁴⁰

Chapter 2

Methodology

2.1 Molecular Dynamics Simulations

Molecular Dynamics Simulations have been performed using the LAMMPS software⁴¹ built with the MBX LAMMPS plugin. We used PBC with an orthogonal cubic box of 19.7295 Å containing 256 water molecules, which corresponds to the experimental density of water at room temperature (0.997 g/cm³). PIMD simulations were performed using 32 beads, with a timestep of 0.25 fs and a local PILE thermostat ($\tau=100$ fs, $T = 298.15$ K).

2.2 Training Procedure

Clusters were randomly selected from 300 ps MB-SCAN PIMD simulation, MB-pol, and MB-SCAN single-point calculations were done using the MBX software package. In the many-body extension, the two-body cutoff was set to 9.0 Å, and the three-body cutoff was 4.5 Å. For the *Delta*-learning procedures the difference between the MB-pol and MB-SCAN energies was used. The fitting of the ACE potentials was performed using the Pacemaker software.

2.3 Hardwer Specifications

Calculations were done on the newly opened Komondor supercomputer in Hungary, a cutting-edge system hosted at the University of Debrecen. PIMD simulations were run on one node with two 64-core AMD EPYC 7763 CPUs. ACE training was performed on an Nvidia A100 GPU.

2.4 Definition of Spatial correlation functions

In the description of different spatial correlation functions $\langle \dots \rangle$ represents an ensemble average of the system.

2.4.1 Radial Distribution Function Between Two Atoms

The radial distribution function (RDF) between two atoms, $g(r)$, quantifies the probability of finding two atoms at a distance r , normalized by the probability expected for an ideal gas at the same density. For two atoms of atom type A and B, the RDF is given by:

$$g_{AB}(r) = \frac{V}{N_A N_B} \left\langle \sum_{i=1}^{N_A} \sum_{j=1}^{N_B} \delta(r - r_{ij}) \right\rangle, \quad (2.1)$$

where V is the volume of the system, N_A , and N_B is the total number of atoms of type A and B, δ is Kronecker-delta, and r_{ij} is the distance between atoms i and j . When these atoms are of the same type, a slight modification is needed:

$$g_{AA}(r) = \frac{V}{N_A(N_A - 1)} \left\langle \sum_{i=1}^{N_A} \sum_{j \neq i}^{N_A} \delta(r - r_{ij}) \right\rangle \quad (2.2)$$

By analyzing $g(r)$ specifically between two atoms, we gain insight into how atom pairs are distributed relative to one another in the system. In practice, $g(r)$ is determined by calculating the distance between all pairs of atoms in the system, binning them into a histogram, and normalizing by the volume of the spherical shell corresponding to the bin.

2.4.2 Angular Distribution Function

The angular distribution function (ADF), $P(\theta)$, describes the probability distribution of the angle θ formed by three atoms: a central atom of type B and two neighboring atoms of type A and B within the cutoff radius r_c of the first coordination shell (the first minimum of the corresponding $g(r)$):⁴²

$$P_{ABC}(\theta) = \frac{1}{N_B \sin \theta \int_0^\pi P_{ABC}(\theta) \sin \theta d\theta} \left\langle \sum_{i=1}^{N_B} \sum_{j=1}^{N_A} \sum_{k=1}^{N_C} \delta(\theta - \theta_{ijk}) \delta_{r_{ij}, r_c} \delta_{r_{ik}, r_c} \right\rangle, \quad (2.3)$$

$$\delta_{r_{ij}, r_c} = \begin{cases} 1, & \text{if } r_{ij} \leq r_c, \\ 0, & \text{if } r_{ij} > r_c, \end{cases} \quad (2.4)$$

where θ_{jik} is the angle between atoms i , j , and k , with i being the central atom. The Kronecker deltas δ_{r_{ij}, r_c} and δ_{r_{ik}, r_c} ensure that only those atoms are considered that are in the cutoff distance r_c . The factor $\sin \theta$ in the denominator accounts for the uneven distribution of angles between three randomly distributed points in three-dimensional space. Finally, the expression is normalized, so that $\int_0^\pi P_{ABC}(\theta) \sin \theta d\theta = 1$.

2.4.3 Tetrahedral Order Parameter

The tetrahedral order parameter q_{tet} quantifies the degree of tetrahedral ordering around a particle, commonly used to characterize the local structure in materials like water. It is defined as:^{42–44}

$$q_{\text{tet}} = 1 - \frac{3}{8} \sum_{j=1}^3 \sum_{k=j+1}^4 \left(\cos \theta_{jk} + \frac{1}{3} \right)^2, \quad (2.5)$$

where θ_{jk} is the angle between bonds formed by the central particle and each of its four nearest neighbors. A perfect tetrahedral arrangement yields q_{tet} close to 1, while random orientations yield values closer to 0. The average tetrahedral order parameter for a system is calculated as follows:

$$\langle q_{\text{tet}} \rangle = \frac{1}{N} \left\langle \sum_{i=1}^N q_{\text{tet},i} \right\rangle \quad (2.6)$$

The distribution of the tetrahedral order parameter, $P(q_{\text{tet}})$, represents the probability density of finding a particular value of the tetrahedral order parameter q in the system. It is calculated by binning the values of q across all central atoms i , taking ensemble averages and normalizing by the total number of atoms and the bin width:

$$P(q_{\text{tet}}) = \frac{1}{N \Delta q} \left\langle \sum_{i=1}^N \delta(q - q_i) \right\rangle \quad (2.7)$$

2.5 Definition of Error Metrics

In the error metrics of forces, the summation occurs over the three force components of individual atoms, and also over different MD snapshots.

2.5.1 Mean Absolute Error of Forces: $\text{MAE}(F)$

$$\text{MAE}(F) = \frac{1}{N} \sum_{i=1}^N |F_i - F_i^{\text{ref}}| \quad (2.8)$$

2.5.2 Root Mean Square Error of Forces: $\text{RMSE}(F)$

$$\text{RMSE}(F) = \sqrt{\frac{1}{N} \sum_{i=1}^N |F_i - F_i^{\text{ref}}|^2} \quad (2.9)$$

2.5.3 Maximum Absolute Error of Forces: $\text{MAX}(F)$

$$\text{MAX}(F) = \max(|F_i - F_i^{\text{ref}}|) \quad (2.10)$$

2.5.4 Pearson Correlation Coefficient of Forces: $r(F)$

$$r(F) = \frac{\sum_{i=1}^N (F_i - \bar{F})(F_i^{\text{ref}} - \bar{F}^{\text{ref}})}{\sqrt{\sum_{i=1}^N (F_i - \bar{F})^2} \sqrt{\sum_{i=1}^N (F_i^{\text{ref}} - \bar{F}^{\text{ref}})^2}} \quad (2.11)$$

2.5.5 Mean Absolute Error of Energies: $\text{MAE}(E)$

$$\text{MAE}(E) = \frac{1}{N} \sum_{i=1}^N |E_i - E_i^{\text{ref}}| \quad (2.12)$$

2.5.6 Root Mean Square Error of Energies: $\text{RMSE}(E)$

$$\text{RMSE}(E) = \sqrt{\frac{1}{N} \sum_{i=1}^N |E_i - E_i^{\text{ref}}|^2} \quad (2.13)$$

2.5.7 Maximum Absolute Error of Energies: $\text{MAX}(E)$

$$\text{MAX}(E) = \max(|E_i - E_i^{\text{ref}}|) \quad (2.14)$$

2.5.8 Pearson Correlation Coefficient of Energies: $r(E)$

$$r(E) = \frac{\sum_{i=1}^N (E_i - \overline{E})(E_i^{\text{ref}} - \overline{E^{\text{ref}}})}{\sqrt{\sum_{i=1}^N (E_i - \overline{E})^2} \sqrt{\sum_{i=1}^N (E_i^{\text{ref}} - \overline{E^{\text{ref}}})^2}} \quad (2.15)$$

Chapter 3

Results and Discussion

3.1 Overview

My research aimed to construct a Δ -learning ML potential for describing the difference between the CC and SCAN potential energy surface (PES) in a periodic simulation while learning only on cluster data. While optimizing the protocol, I focused on the predictive ability of potentials regarding reference periodic energies and forces with the ML potentials. Clusters were extracted from the reference PBC simulations using random sampling⁴⁵ and cutting the n closest waters from the central oxygen atom. Moreover, I also limited the maximum cluster size and paid attention to minimizing the number of sampled atoms, to reduce the data set to a minimum, to get the maximum efficiency and therefore get a decent calculation speed. As one of my main goals was to perform accurate and fast simulations, I assessed the best-performing method in PIMD simulations on top of the ACE(SCAN) baseline model.

3.2 Preliminary Studies

3.2.1 Effect of cluster-based learning

By cutting a cluster from a periodic cell, we already involve some errors since we perform single-point calculations in the gas phase for the cluster. To assess this, we analyzed force errors on the central oxygen of the cluster, compared to its periodic reference. In Figure 3.1, we can see that the MAE of CC (MB-pol) forces is very high at small clusters, and it slowly converges around 77 meV/Å at large cluster sizes like 100 waters. Taking the CC-SCAN difference on the other hand, the same value is achieved at a cluster size of 15 waters. So it is very worthwhile to learn only the CC-SCAN difference on clusters and learn a baseline potential periodically. This allows the decrease in cluster size, which is beneficial to us because of the non-linear scaling nature of the used calculations.

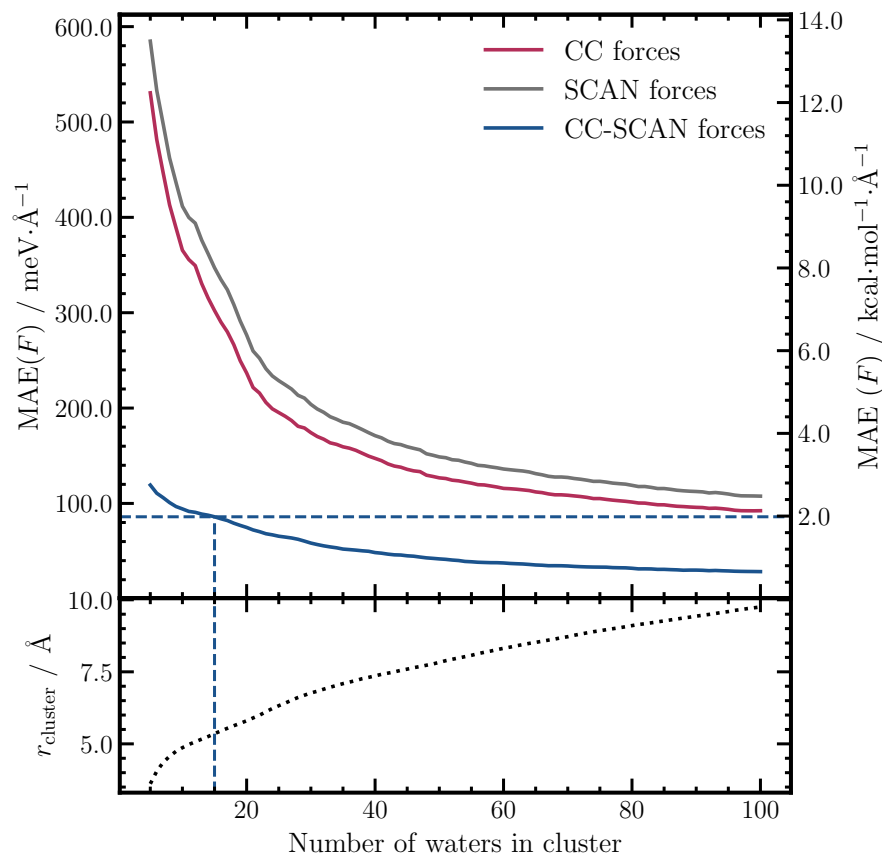


Figure 3.1: Mean absolute error comparison of forces calculated on the central oxygen of clusters wrt. the periodic reference force on the same atom. Calculated as an average of 500 clusters for each cluster size. Abbreviations: SCAN: MB-SCAN, CC: MB-pol

3.2.2 Effect of training without forces

In coupled cluster (CC) methods, calculating forces is prohibitively expensive. Notably, many CC methods do not include a straightforward way to compute forces. Even for those that do, the calculations are extremely resource-intensive, often requiring up to three times the computational cost of an energy calculation for the same system. Without forces, we are limited to fitting cluster energies, which provides significantly less information for each cluster. In a preliminary study, we explored the impact of training without forces on the accuracy. This was possible due to the MB-pol representation of the CC PES in our studies.

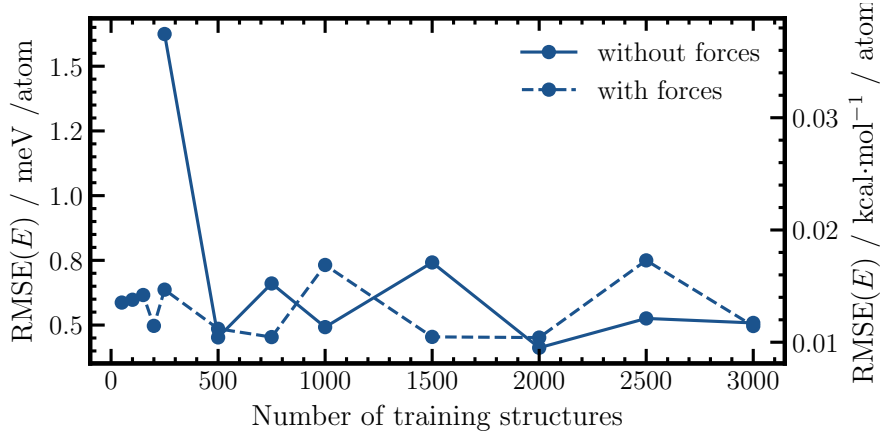


Figure 3.2: Energy RMSE wrt. the number of training structures using clusters containing 64 water molecules, comparing fitting procedures with and without atomic forces.

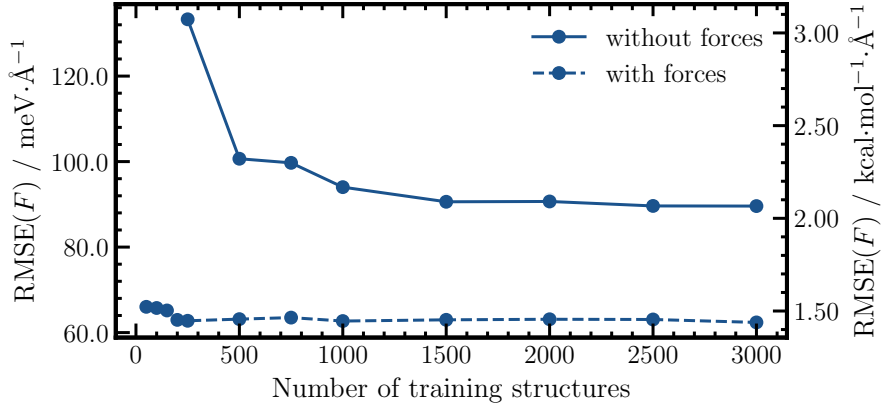


Figure 3.3: Force RMSE wrt. the number of training structures using clusters containing 64 water molecules, comparing fitting procedures with and without atomic forces.

In Figure 3.2, we can see that the RMSE of the energy is significantly worse when using less than 500 structures to train the model. Over that value, the two methods are comparable in this case giving an energy RMSE between 0.5 and 0.8 meV/atom. On the other hand, in Figure 3.3, we observe a stark contrast between the two approaches in terms of force RMSE. Here, we also observe significant errors under 500 structures, but the energies-only approach remains inferior even with a large amount of reference data. We can see that the $\text{RMSE}(F)$ converges to a significantly higher value (89 meV/Å) compared to the one with forces (68 meV/Å). Also, convergence is much faster when calculating atomic forces, and 200 clusters provide sufficient information, while without forces one would need at least 1000 of them. In the end, it turned out that without forces, many more data points are needed and the errors are more significant,

so in the future, I paid special attention to the optimization of learning without forces to make it as efficient as possible

3.3 Optimization of the Δ -Learning Protocol

3.3.1 Cutoff Radius

My initial task was to select a cutoff radius for my ML method. This parameter allows to specify the local range of the potential. I have attempted to find the adequate cutoff radius in the 4-6 Å regime (see Figure 3.4).

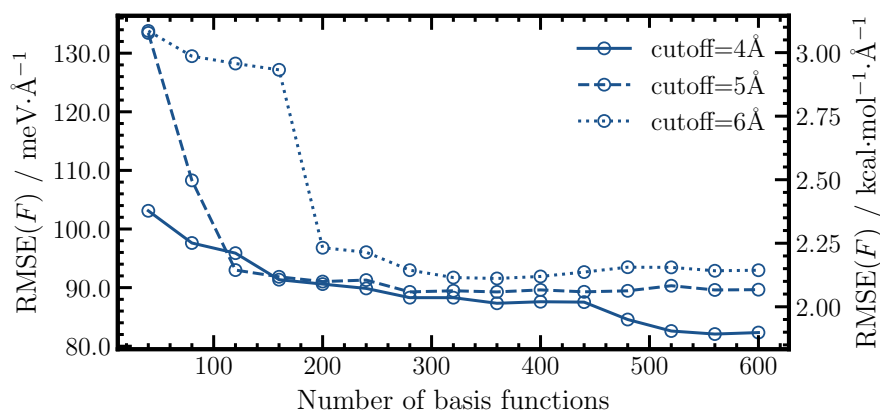


Figure 3.4: Force RMSE comparison between different cutoffs for machine learning, the calculations were run on 64 cluster size and with 2500 training structures.

Unintuitively, large cutoff potentials underperform, because they require more basis functions to adequately describe the large-radius environment. Therefore I chose the most accurate and computationally most efficient 4 Å cutoff, which force RMSE can be as low as (82 meV/Å), in comparison with the 5 and 6 Å (89-93 meV/Å) values. This also means that the CC-SCAN difference in the interaction of water molecules can be described in a relatively short distance cutoff. This can be explained by the fact that the electron correlation decays quickly. From a structural point of view, a 4 Å radius lies between the first and second hydration sphere of water (see Figure 1.2).

Up until this point, I used 2500 clusters with 64 water molecules to analyze different potential settings with an extensive training set and cluster size. Both the reference calculations and the learning procedure are slower for each additional atom and structure in the training data. From here on my work was focused on minimizing these datasets.

3.3.2 Cluster Size and Number of Reference Calculations

In this section, I focused on speeding up the learning process by substantially reducing the size of the training set without significant loss of accuracy. The undermentioned tests were executed to find the smallest cluster size, with the least amount of fitting data, needed. This would be crucial to our research group because it can bring significant improvement in our workflow, as it can exponentially speed up some of our calculations.

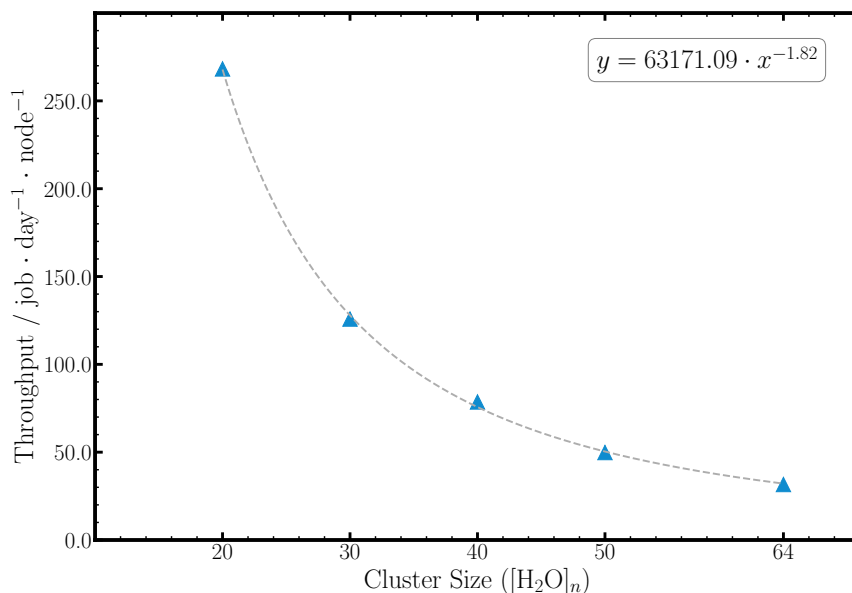


Figure 3.5: The estimated productivity increase measured by our group, when working with reduced cluster sizes. Figure and results from Máté Kígyósi.

As we can see in Figure 3.5, we could do five times the amount of the calculation if I manage to reduce the cluster size from 64 to 20. I also want to emphasize that computational cost scales linearly the number of training structures so we may have better efficiency if we can manage smaller clusters, with a higher number of training data. While we would expect a monotonous decrease in the errors as the number of structures increases, we do not see this at certain points. To remedy this, it would be necessary to examine the deviation of several parallel fits.

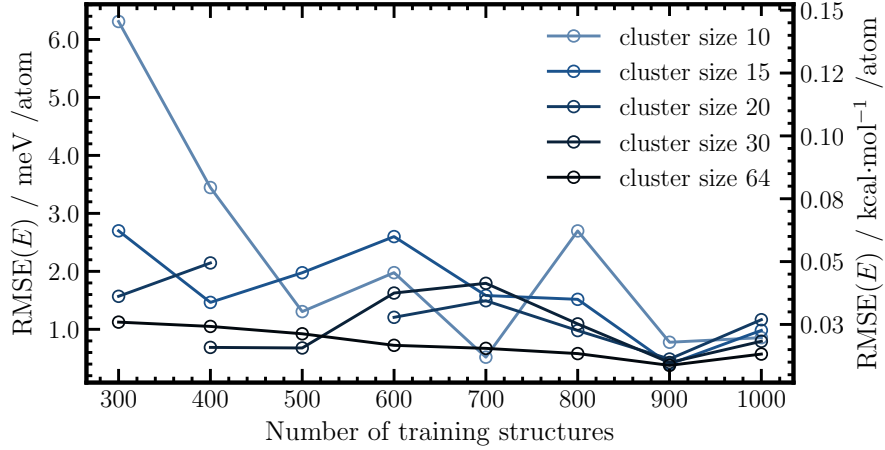


Figure 3.6: Energy RMSE comparison between different sizes of clusters in meV and kcal/mol wrt. the number of training structures. The growth of the cluster size is represented by gradiently darker colors

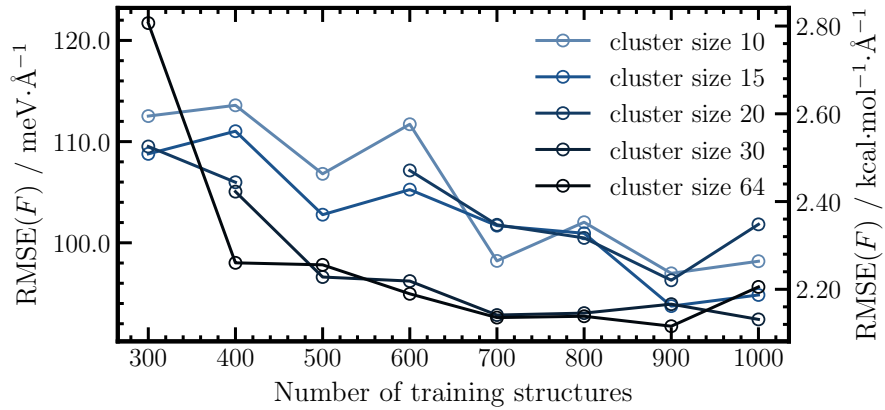


Figure 3.7: Force RMSE comparison between different sizes of clusters in meV and kcal/mol wrt. the number of training structures. The growth of the cluster size is represented by gradiently darker colors

As we see in Figures 3.6 and 3.7, RMSE errors converge in the range of 900-1000 training structures. In Figure 3.6, the cluster 10 fit shows bad performance in the low count training structure range as it has almost three times larger error than with 1000 structures. Figure 3.6 also shows that using a cluster size of 64 waters doesn't get much more accurate in the prediction of the energy when increasing the number of structures.

Moving on to the next figure 3.6, the force errors converge more slowly. Looking at the 64-cluster dataset in Figure 3.7, Force errors improve significantly with the increasing training set. Surprisingly cluster sizes of 30 and 64 are producing fairly similar results. I also want to emphasize that I have very close results across the cluster sizes with 900-1000 training

structures. Specifically, the much smaller 15-cluster potentials show fairly similar performance to the 64-cluster one above 900 training structures. Due to its remarkable cost-accuracy trade-off, I choose the potential trained on 900 15-water clusters for the MD studies ($\text{RMSE}(E) = 0.4 \text{ meV/atom}$, $\text{RMSE}(F) = 93.7 \text{ meV/\AA}$).

3.3.3 Comparison with MACE

In the hope of improving performance, I also tested the Multi Atomic Cluster Expansion (MACE). MACE builds on the ACE formalism but also incorporates message passing, and is known to reduce errors compared to ACE, especially in cluster-based learning procedures.¹⁴ I decided to compare my results from ACE with it with cluster size 15.

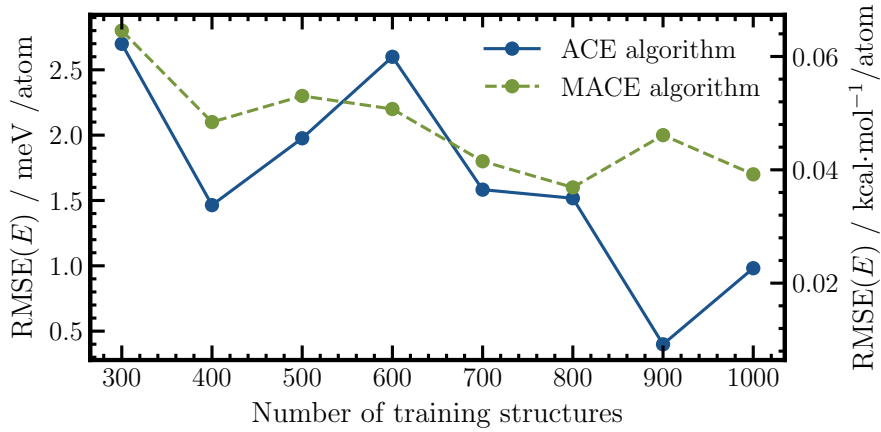


Figure 3.8: Energy RMSE comparison between the two different technologies wrt. to the number of training structures, using a cluster size of 15 waters.

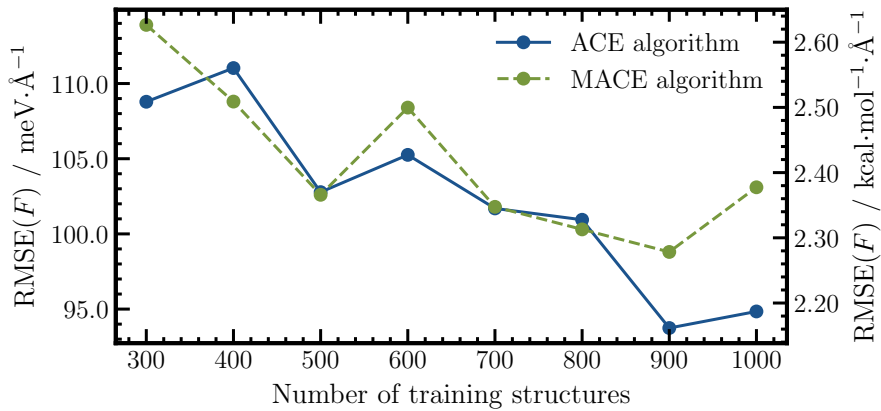


Figure 3.9: Force RMSE comparison between the two different technologies, wrt. to the number of training structures, using a cluster size of 15 waters.

In Figures 3.8 and 3.9, MACE shows no improvement, even with 900 and 1000 training structures ACE has a slight advantage. In comparison of the force RMSE 3.9 both models show their best performance, but MACE comes short with it's (98.8 meV/Å) to (93.7 meV/Å) accuracy of the ACE. This might be due to the increased complexity of MACE, therefore having a slower learning progress. The fact that we see no improvement can also be attributed to the local nature of CC-SCAN difference, while message passing would help more in long-range interactions. This makes it currently not a primary choice to generate our potential. On top of this, it is computationally more demanding due to its complexity, and can only be efficiently run using GPU.

3.4 Performance of the Machine Learning Potentials

As I mentioned in the preliminary, our goal was to run periodic molecule dynamic simulations. For this, I previously fitted a potential energy surface for water on clusters with ACE, which we will use for the calculations of the energy, forces, and spatial coordinates. We chose to use the cluster size of 15 with 900 training structures as it has the best results as seen from my testing. We set up a PIMD simulation, with a 19 Å wide box. We set our goal to run the calculation for 200 ps. We compared our results with our MB-pol reference simulations used as the ground truth.

3.4.1 Radial Distribution Functions

We chose for our first comparison the radial distribution function (RDF), denoted as $g(r)$, which describes the probability of finding a particle at a distance r from a reference particle, relative to the probability expected in the case of an ideal gas where particles are uniformly distributed. This is important in the description of the behavior of the water molecules because directly relates to intermolecular interactions, such as the hydrogen bonding network or the electromagnetic repulsion of the oxygen-oxygen atoms.

In Figure 3.10 the first thing to observe is that the SCAN and CC models lead to very different structural arrangements. When applying our Δ -ML ACE fit, we get close to perfect agreement with the CC distribution functions, both in the location and the height of the peaks. Based on the OH and OO, it is evident that we capture the structure of the first two hydration spheres well, and remarkably, the long-range ordering is also accurate. I also want to highlight

that we achieved these results with a Δ -ML’s learning cutoff set to 4 Å in a PBC cell with a size of 19.7 Å. This means, that while in the Δ -potential one water molecule sees less, than half of the simulation box, it can correctly describe the whole periodic system. Again, this can be attributed to the idea that the main force and energy contributions in the ACE(CC-SCAN) come from short-range interactions, while long-range interactions can be accurately treated with the ACE(SCAN) baseline model. From the distribution function $g_{\text{OO}}(r)$, it is also clear, that at 4 Å we only cover half of the second hydration sphere.

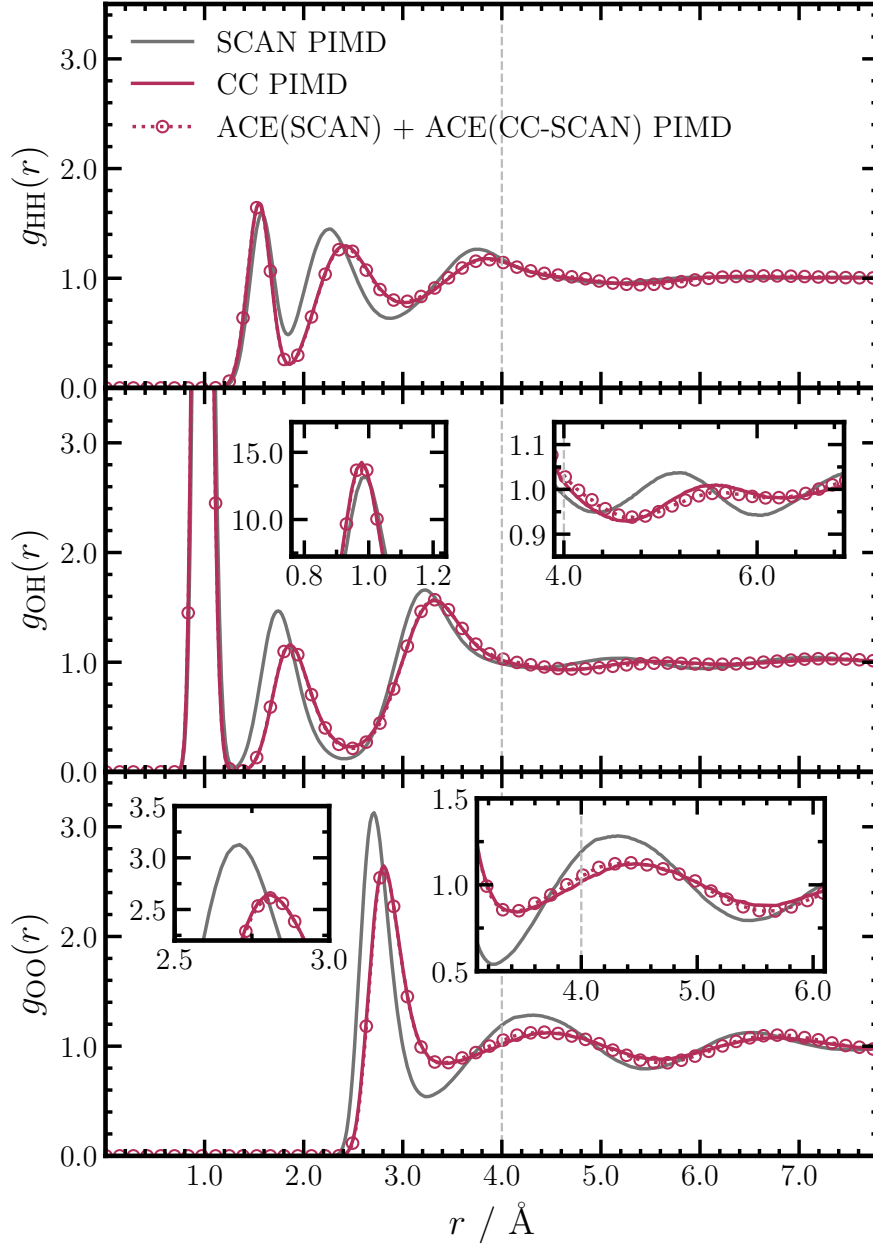


Figure 3.10: Radial distribution functions of liquid water using different methods, calculated from 200 ps PIMD simulation (NVT, 298.15 K). θ is represented in degrees. Abbreviations: SCAN: MB-SCAN, CC: MB-pol, ACE(SCAN): the baseline ACE model fitted on 100 periodic MB-SCAN structures, ACE(CC-SCAN): Δ -learning ACE model fitted on the $E(\text{MB-pol}) - E(\text{MB-SCAN})$ energy differences of 900 clusters each containing 15 water molecules. The vertical dashed line at 4.0 Å indicates the cutoff of the second ACE potential. The RDFs have been computed using a radial bin size of 0.01 Å as depicted and are not smoothened.

3.4.2 Angular Distribution Functions

We decided to have a further assessment of the ML potential, using a more sensitive descriptor, the Angular Distribution Functions (ADF). Unlike the RDF, this depends on the positions of three atoms. We examined three different angular distribution functions: 1. $\text{O} \cdots \text{H}-\text{O}$, representing the angular distribution of hydrogen bonds; 2. HOH , describing the intramolecular angular distribution; 3. OOO , capturing the angular distribution of oxygen triplets. In Figure 3.11 we can see that our model accurately represents the intra-molecular angles (middle panel), which is a key feature for any method, although it is important to highlight that the baseline SCAN model already did well. The advances that were made by the ML-aided calculation are a more accurate description of the hydrogen bond angles and the triplet oxygen angles.

The SCAN functional generally predicts flatter angles for the hydrogen bond, while with the Δ -ML correction, we achieve good agreement with the CC distribution, with only a few degrees of deviation at the maximum. When examining the OOO distribution, we observe a main peak around $100\text{-}110^\circ$, corresponding to the tetrahedral arrangement, and a small shoulder between $50\text{-}60^\circ$, which appears in cases of 5 or higher coordination numbers, likely due to interstitial water molecules in the first hydration shell. The SCAN method results in a narrower distribution, predicting a tetrahedral arrangement with higher probability while underestimating the likelihood of interstitial water formation. The Δ -potential provides a good, but not perfect match to the CC distribution for this sensitive property, slightly underestimating the presence of interstitial water molecules.^{44,46}

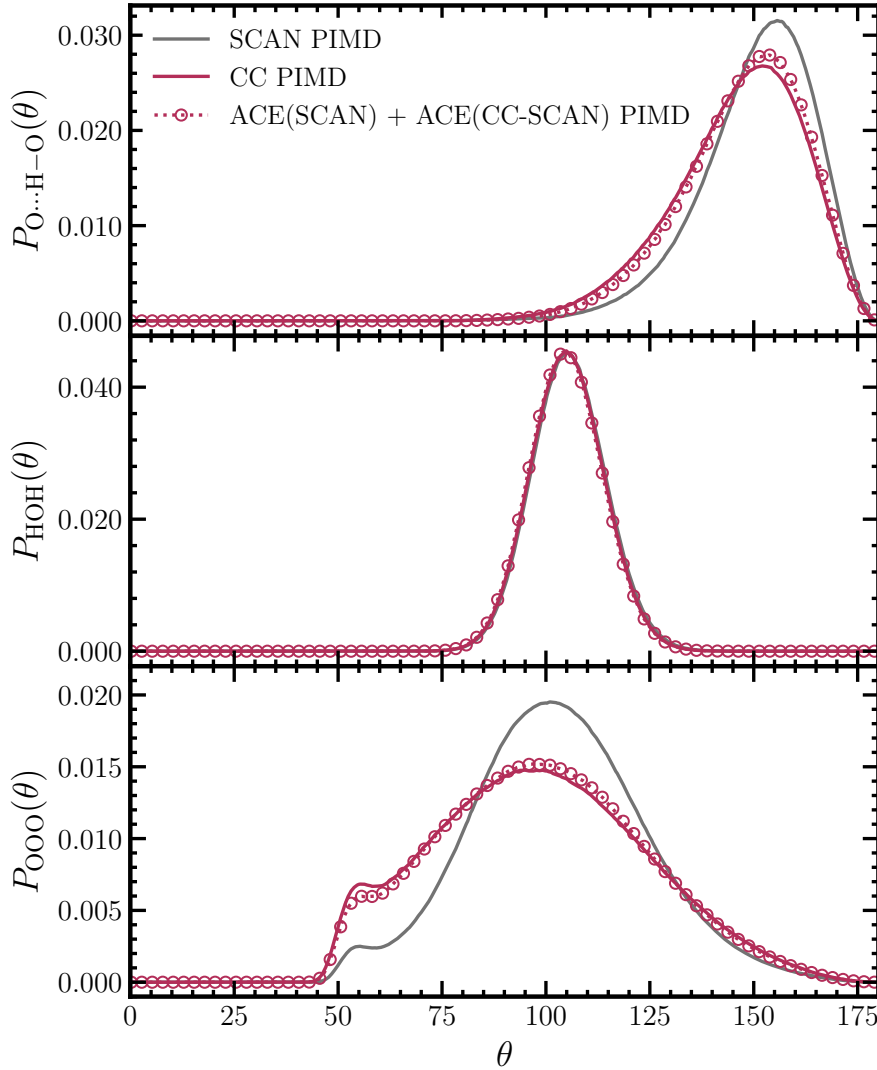


Figure 3.11: Angular distribution functions of liquid water using different methods, calculated from 200 ps PIMD simulation (NVT, 298.15 K). θ is represented in degrees. Abbreviations: SCAN: MB-SCAN, CC: MB-pol, ACE(SCAN): the baseline ACE model fitted on 100 periodic MB-SCAN structures, ACE(CC-SCAN): Δ -learning ACE model fitted on the $E(\text{MB-pol}) - E(\text{MB-SCAN})$ energy differences of 900 clusters each containing 15 water molecules. The ADFs have been computed using an angular bin size of 0.5° as depicted and are not smoothened.

3.4.3 Tetrahedral Order Parameter

Our last test was the tetrahedral order parameter (q_{tet}), which measures the extent to which a molecule and its four closest neighbors adopt a tetrahedral arrangement. It was proposed to analyze the order and structural anomalies in water,⁴³ and it is often applied in the study of other liquids and glasses to measure the local arrangement. This parameter is particularly

useful for characterizing systems where particles tend to form structures that are not long-range ordered. In the case of a perfectly tetrahedral structure, $q_{\text{tet}} = 1$, while in the ideal gas limit $q_{\text{tet}} = 0$.

Last but not least if we take a look at Figure 3.12 we can see, that the original SCAN representation does not catch the bimodal shape of the CC distribution correctly, and it shows a significant deviation, similarly to $P_{\text{OOO}}(\theta)$. Again it shows, that the SCAN PES gives a much more ordered tetrahedral distribution than the CC. But with the help of the Δ -learning potential, we could get close to the reference distribution function of CC.

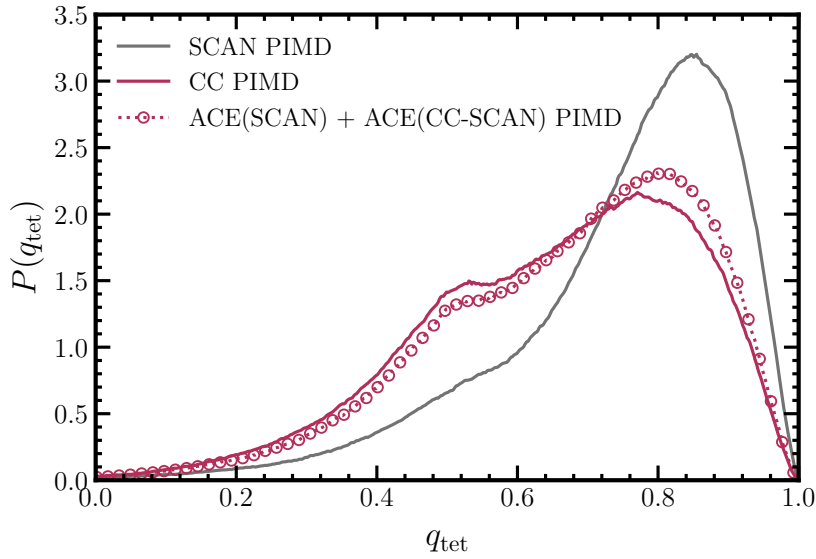


Figure 3.12: Distribution function of tetrahedral order parameter q_{tet} of liquid water using different methods, calculated from 200 ps PIMD simulation (NVT, 298.15 K). Average values of q_{tet} : SCAN: 0.74, CC: 0.65, ACE: 0.67. Abbreviations: SCAN: MB-SCAN, CC: MB-pol, ACE(SCAN): the baseline ACE model fitted on 100 periodic MB-SCAN structures, ACE(CC-SCAN): Δ -learning ACE model fitted on the $E(\text{MB-pol}) - E(\text{MB-SCAN})$ energy differences of 900 clusters each containing 15 water molecules. The distribution functions have been computed using a bin size of 0.003 as depicted and are not smoothened.

3.5 Summary and Outlook

The aim of the research was to further develop and optimize the CCMD-based protocol in the field of cluster-based Δ -learning. Compared to the original CCMD protocol, the number of reference points required for training was reduced to one-third (from 3000 to 900), and the required cluster size was reduced to one-fifth (from 64 to 15). This resulted in a significant speedup: when calculated with quadratic scaling for local CC calculations, we can achieve 75-fold speedup. The final Δ -ML potential can describe the CC-SCAN potential energy surface difference within the typical accuracy of state-of-the-art machine learning potentials ($\text{RMSE}(E) = 0.4 \text{ meV/atom}$, $\text{RMSE}(F) = 93.7 \text{ meV/\AA}$).⁴⁷ We tested the potential in PIMD simulations, and it corrected the significant discrepancies between CC and SCAN in the structural properties of water. It provided near-perfect agreement for pair correlation functions and also yielded good results for different triplet correlation functions. To extend this protocol beyond liquid water, the next step of my research will be to describe ions in water: sodium and chloride.⁴⁸

Acknowledgements

Great thanks to Máté Kígyósi for sharing his results in the benchmarking of local coupled cluster calculations for water clusters. Financial support from the National Research, Development, and Innovation Office (NKFIH, Grant No. FK147031) is gratefully acknowledged. Supported by the DKOP-23 Doctoral Excellence Program of the Ministry for Culture and Innovation from the source of the National Research, Development and Innovation Fund. We acknowledge KIFÜ (Governmental Agency for IT Development, Hungary, <https://ror.org/01s0v4q65>) for awarding us access to the Komondor HPC facility based in Hungary.

References

- [1] Daru, J.; Forbert, H.; Behler, J.; Marx, D. *Physical Review Letters* **2022**, *129*, 226001.
- [2] Ceriotti, M.; Parrinello, M.; Markland, T. E.; Manolopoulos, D. E. *The Journal of Chemical Physics* **2010**, *133*, 124104.
- [3] Stolte, N.; Daru, J.; Forbert, H.; Behler, J.; Marx, D. *arXiv preprint* **2024**, arXiv: 2410.17675.
- [4] Soper, A.; Benmore, C. *Physical Review Letters* **2008**, *101*, 065502.
- [5] Pinski, P.; Neese, F. *The Journal of Chemical Physics* **2019**, *150*, 164102.
- [6] Guo, Y.; Riplinger, C.; Becker, U.; Liakos, D. G.; Minenkov, Y.; Cavallo, L.; Neese, F. *The Journal of Chemical Physics* **2018**, *148*, 011101.
- [7] Nagy, P. R.; Kállay, M. *The Journal of Chemical Physics* **2017**, *146*, 214106.
- [8] Nagy, P. R.; Samu, G.; Kállay, M. *Journal of Chemical Theory and Computation* **2018**, *14*, 4193–4215.
- [9] Nagy, P. R.; Kállay, M. *Journal of Chemical Theory and Computation* **2019**, *15*, 5275–5298.
- [10] Unke, O. T.; Stöhr, M.; Ganscha, S.; Unterthiner, T.; Maennel, H.; Kashubin, S.; Ahlin, D.; Gastegger, M.; Sandonas, L. M.; Berryman, J. T.; Tkatchenko, A.; Müller, K.-R. *Science Advances* **2024**, *10*, eadn4397.
- [11] Wang, T.; He, X.; Li, M.; Li, Y.; Bi, R.; Wang, Y.; Cheng, C.; Shen, X.; Meng, J.; Zhang, H.; Liu, H.; Wang, Z.; Li, S.; Shao, B.; Liu, T.-Y. *Nature* **2024**, 1–9.
- [12] Schütt, K.; Arbabzadah, F.; Chmiela, S.; Müller, K.-R.; Tkatchenko, A. *Nature Communications* **2017**, *8*, 13890.
- [13] Unke, O. T.; Chmiela, S.; Sauceda, H. E.; Gastegger, M.; Poltavsky, I.; Schütt, K. T.; Tkatchenko, A.; Müller, K.-R. *Chemical Reviews* **2021**, *121*, 10142–10186.

-
- [14] Kovács, D. P.; Moore, J. H.; Browning, N. J.; Batatia, I.; Horton, J. T.; Kapil, V.; Witt, W. C.; Magdău, I.-B.; Cole, D. J.; Csányi, G. *arXiv preprint* **2023**, arXiv: 2312.15211.
- [15] Behler, J.; Parrinello, M. *Physical Review Letters* **2007**, *98*, 146401.
- [16] Drautz, R. *Physical Review B* **2019**, *99*, 14104.
- [17] Bochkarev, A.; Lysogorskiy, Y.; Menon, S.; Qamar, M.; Mrovec, M.; Drautz, R. *Physical Review Materials* **2022**, *6*, 013804.
- [18] Lysogorskiy, Y.; Oord, C.; Bochkarev, A.; Menon, S.; Rinaldi, M.; Hammerschmidt, T.; Mrovec, M.; Thompson, A.; Csányi, G.; Ortner, C.; Drautz, R. *NPJ Computational Materials* **2021**, *7*, 97.
- [19] Ibrahim, E.; Lysogorskiy, Y.; Mrovec, M.; Drautz, R. *Physical Review Materials* **2023**, *7*, 113801.
- [20] Qamar, M.; Mrovec, M.; Lysogorskiy, Y.; Bochkarev, A.; Drautz, R. *Journal of Chemical Theory and Computation* **2023**, *19*, 5151–5167.
- [21] Ibrahim, E.; Lysogorskiy, Y.; Drautz, R. *Journal of Chemical Theory and Computation* **2024**, ASAP.
- [22] Liang, Y.; Mrovec, M.; Lysogorskiy, Y.; Vega-Paredes, M.; Scheu, C.; Drautz, R. *Journal of Materials Research* **2023**, *38*, 5125–5135.
- [23] Kovács, D. P.; Oord, C. v. d.; Kucera, J.; Allen, A. E. A.; Cole, D. J.; Ortner, C.; Csányi, G. *Journal of Chemical Theory and Computation* **2021**, *17*, 7696–7711.
- [24] Vargas, S.; Hennefarth, M.; Liu, Z.; Alexandrova, A. *Journal of Chemical Theory and Computation* **2021**, *17*, 6203–6213.
- [25] Vandermause, J.; Xie, Y.; Lim, J. S.; Owen, C. J.; Kozinsky, B. *Nature Communications* **2022**, *13*, 5183.
- [26] Kovács, D. P.; Batatia, I.; Arany, E. S.; Csányi, G. *The Journal of Chemical Physics* **2023**, *159*, 044118.
- [27] Batatia, I.; Batzner, S.; Kovacs, D.; Musaelian, A.; Simm, G.; Drautz, R.; Ortner, C.; Kozinsky, B.; Csányi, G. *Nature* **2022**, *13*, 2453.

-
- [28] Cheng, B. *NPJ Computational Materials* **2024**, *10*, 157.
- [29] Bochkarev, A.; Lysogorskiy, Y.; Drautz, R. *Physical Review X* **2024**, *14*, 021036.
- [30] Palos, E.; Bull-Vulpe, E.; Zhu, X.; Agnew, H.; Gupta, S.; Saha, S.; Paesani, F. *Journal of Chemical Theory and Computation* **2024**, *20*, 9269–9289.
- [31] Babin, V.; Leforestier, C.; Paesani, F. *Journal of Chemical Theory and Computation* **2013**, *9*, 5395–5403.
- [32] Medders, G. R.; Babin, V.; Paesani, F. *Journal of Chemical Theory and Computation* **2014**, *10*, 2906–2910.
- [33] Babin, V.; Medders, G. R.; Paesani, F. *Journal of Chemical Theory and Computation* **2014**, *10*, 1599–1607.
- [34] Riera, M.; Mardirossian, N.; Bajaj, P.; Götz, A. W.; Paesani, F. *The Journal of Chemical Physics* **2017**, *147*, 161715.
- [35] Caruso, A.; Paesani, F. *The Journal of Chemical Physics* **2021**, *155*, 064502.
- [36] Riera, M.; Hiraes, A.; Ghosh, R.; Paesani, F. *The Journal of Physical Chemistry B* **2020**, *124*, 11207–11221.
- [37] Riera, M.; Yeh, E. P.; Paesani, F. *Journal of Chemical Theory and Computation* **2020**, *16*, 2246–2257.
- [38] Cruzeiro, V. W. D.; Lambros, E.; Riera, M.; Roy, R.; Paesani, F.; Gotz, A. W. *Journal of Chemical Theory and Computation* **2021**, *17*, 3931–3945.
- [39] Dasgupta, S.; Lambros, E.; Perdew, J.; Paesani, F. *Nature Communications* **2021**, *12*, 6359.
- [40] Markland, T. E.; Ceriotti, M. *Nature Reviews Chemistry* **2018**, *2*, 0109.
- [41] Thompson, A. P.; Aktulga, H. M.; Berger, R.; Bolintineanu, D. S.; Brown, W. M.; Crozier, P. S.; in 't Veld, P. J.; Kohlmeyer, A.; Moore, S. G.; Nguyen, T. D.; Shan, R.; Stevens, M. J.; Tranchida, J.; Trott, C.; Plimpton, S. J. *Computer Physics Communications* **2022**, *271*, 108171.

- [42] Zheng, L.; Chen, M.; Sun, Z.; Ko, H.-Y.; Santra, B.; Dhuvad, P.; Wu, X. *The Journal of Chemical Physics* **2018**, *148*, 164505.
- [43] Errington, J. R.; Debenedetti, P. G. *Nature* **2001**, *409*, 318–321.
- [44] Dhabal, D.; Singh, M.; Wikfeldt, K. T.; Chakravarty, C. *The Journal of Chemical Physics* **2014**, *141*, 174504.
- [45] Stolte, N.; Daru, J.; Forbert, H.; Marx, D.; Behler, J. *arXiv preprint* **2024**, arXiv:2410.10698.
- [46] Sharp, K. A.; Vanderkooi, J. M. *Accounts of Chemical Research* **2010**, *43*, 231–239.
- [47] Omranpour, A.; Elsner, J.; Lausch, K. N.; Behler, J. *arXiv preprint* **2024**, arXiv:2411.00720.
- [48] Stolte, N.; Forbert, H.; Behler, J.; Marx, D. *The Journal of Physical Chemistry Letters* **2024**, *15*, 12144–12150.