Atomistic Machine Learning for Modelling Disordered Tetrahedral Networks

Zoé Faure Beaulieu

St Catherine's College

A thesis submitted to the University of Oxford

for the degree of Doctor of Philosophy

Supervisors: Prof. Volker Deringer, Dr. Fausto Martelli

Inorganic Chemistry Laboratory

University of Oxford

Declaration

The work presented in this thesis was carried out between October 2022 and July 2025 in the Inorganic Chemistry Laboratory, University of Oxford under the joint supervision of Prof. Volker L. Deringer and Dr. Fausto Martelli. This dissertation is the result of my own original work, and where it draws on the work of others, this is acknowledged at appropriate points in the text. This dissertation has not been submitted in whole or in part for a degree at this or any other institution.

Abstract

Disordered tetrahedral networks are pervasive across a broad range of chemically diverse materials, including elemental solids, inorganic glasses, metal–organic frameworks, and liquids. While many such systems lack long-range order, some can exhibit extended correlations that resemble those found in crystalline solids. This structural complexity and variability pose significant challenges for computational modelling, particularly due to the presence of local heterogeneity and the breakdown of traditional symmetry-based approaches. In this thesis, I explore the application of machine learning (ML) methods to the atomistic modelling of disordered tetrahedral networks, with a particular focus on how learned representations can capture, compare, and generalise information across systems that share local tetrahedral geometry but differ widely in chemistry.

I begin by examining water and its amorphous phases. Using bond-orientational order parameters as inputs to neural network models, I classify local environments in amorphous ices and explore the structural relationships between low-, medium-, and high-density amorphous phases. I then train a graph neural network (GNN) potential for liquid water that accurately reproduces key experimental signatures, including the splitting of the principal peak in the structure factor. Crucially, I go beyond simply reproducing this feature and provide new understanding on the microscopic origin of the splitting. By analysing the topology of hydrogen-bonded rings, I show that the peak splitting arises from distinct topological rearrangements within the network.

I go on to explore structural and energetic analogies between zeolitic imidazolate frameworks (ZIFs) and zeolites through a coarse-graining approach. Using Gaussian Process Regression (GPR), I construct ML models at multiple levels of resolution. I find that AB_2 coarse-grained representations, which preserve tetrahedral connectivity, retain much of the predictive power of fully atomistic models, while A-only representations lose critical structural information. These results quantitatively support the idea that topology can be used to model the energetic landscape of tetrahedral frameworks.

Building on this, I examine the transferability of ML potentials across chemically distinct disordered tetrahedral materials. Using transfer learning, I adapt GNNs trained on one material, such as carbon, silicon, silica, or water, to another. I identify key factors controlling successful cross-domain transfer and demonstrate that the tetrahedral motif can embed sufficient chemical and structural information to support generalisation under certain conditions. Finally, I show that atomistic foundation models can be distilled on water data to enable large-scale, accurate simulations at reduced computational cost, pointing toward a future of universal, efficient ML interatomic potentials for disordered materials.

Together, these studies advance our understanding of how tetrahedral coordination governs structure in disordered materials and demonstrate the potential of ML-based approaches to uncover transferable, chemically informed insights across a spectrum of complex systems.

Acknowledgements

Firstly, I would like to express my sincere gratitude to my supervisors, Volker Deringer and Fausto Martelli, for their unwavering support, encouragement, and guidance throughout the course of my DPhil. Volker's enthusiasm for science has been inspiring, and I am grateful for the opportunity to have been a part of such an exciting and dynamic research group. Fausto has been an outstanding mentor, always offering generous support and taking the time to answer my many questions. His willingness to help has been invaluable to my development.

I have had the privilege of working alongside a fantastic team during my DPhil. I would like to thank all of my colleagues in the Deringer group for making my time here so enjoyable, for the many stimulating discussions we've shared, and for making the office a place I genuinely looked forward to going to each day. In particular, I am grateful to Joe, Zak, Daniel, Tom, and John, who have been there since my early days as a Part II student, always willing to lend an ear and patiently answer my endless questions. To Louise, your infectious enthusiasm for science has been a true inspiration, but even more than that, your kindness, support, and friendship have meant the world to me.

Thank you to all my friends in Oxford for helping make this city my home over the past seven years. A special thanks goes to my teammates at the Oxford University Swimming Club, who have been a constant source of motivation and support. To Max, Jake, Toby, and Jonathan, thank you for always being up for a board game night and a good chat. To Elysia, thank you for always being there to listen, to laugh, and to support me. To Rob, thank you for constantly pushing me to be better (if only to earn that elusive Kudos). To Louisa, for always showing me the way and being a steady source of support. And finally, to Ines, for being there long before all this began, and for many more years to come.

I am endlessly grateful to my family for their unconditional love and support. To Mum, thank you for always backing me in everything I do, without question or hesitation. Your strength and determination have been my greatest inspiration. To Nina, 13,000 kilometres haven't stopped you from being my best friend, biggest cheerleader, and fiercest advocate. And to Dad, your interest in AI has often exceeded my own, I'm always grateful to have you in my corner.

And finally, to John, without whom, none of this would have been possible. Thank you for your endless patience, love, and support through all the ups and downs of the last three years. You are my rock, and I am so grateful to have you by my side. I can't wait to see what the future holds for us.

Publications

The following publications and pre-prints contributed directly to this thesis:

- Z. Faure Beaulieu, V. L. Deringer, F. Martelli;
  "High-dimensional order parameters and neural network classifiers applied to amorphous ices";
  J. Chem. Phys. **2024**, 160, 081101.

- Z. Faure Beaulieu, T. C. Nicholas, J. L. A. Gardner, A. L. Goodwin, V. L. Deringer;
  "Coarse-grained versus fully atomistic machine learning for zeolitic imidazolate frameworks";
  Chem. Commun. **2023**, 59, 11405-11408.

- J. L. A. Gardner, D. F. Toit, C. Ben Mahmoud, Z. Faure Beaulieu, V. Juraskova, L. Pasca, L. A. M. Rosset, F. Duarte, F. Martelli, V. L. Deringer;
  "Distillation of atomistic foundation models across architectures and chemical domains";
  ArXiv **2025**, arXiv:2506.10956.

I was involved in the following publication that developed the use of synthetic local energies as a learning target for NNs discussed in Chapter 5:

- J. L. A. Gardner, Z. Faure Beaulieu, V. L. Deringer;
  "Synthetic data enable experiments in atomistic machine learning";
  Digital Discovery **2023**, 2, 651-662.

Abbreviations

ACE atomic cluster expansion

ACSF atom-centered symmetry functions

AIMD *ab initio* molecular dynamics

AL active learning

a-C amorphous carbon

a-Si amorphous silicon

BAS balanced accuracy score

BCC body-centre cubic

BOO bond–orientational order

cg coarse-graining

CRN continuous random network

DFT density functional theory

dia-C diamond carbon

dia-Si diamond-type silicon

DTN disordered tetrahedral networks

EMD Earth Mover's Distance

FCC face-centred cubic

FSDP first sharp diffraction peak

GAP Gaussian Approximation Potential

GNN graph neural network

GP Gaussian Process

GPR Gaussian Process Regression

gra-C graphitic carbon

HBN hydrogen-bond network

HDA high-density amorphous

HDL high-density liquid

HDNNP high-dimensional neural network potential

HCP hexagonal close-packed

ISF intermediate scattering function

IZA-SC Structure Commission of the International Zeolite Association

KDE kernel density estimate

LAMMPS Large-scale Atomic/Molecular Massively Parallel Simulator

LDA low-density amorphous

LDL low-density liquid

LLCP liquid–liquid critical point

LJ Lennard-Jones

MAE mean absolute error

MD molecular dynamics

MDA medium-density amorphous

ML machine learning

MLIP machine-learned interatomic potential

MOF metal-organic framework

MPNN message passing neural network

MRO medium-range order

MSD mean square displacement

NequIP neural equivariant interatomic potential

NMR nuclear magnetic resonance

NN neural network

NQE nuclear quantum effect

OQMD Open Quantum Materials Database

PBE Perdew-Burke-Ernzerhof

PCA principal component analysis

PES potential energy surface

PFI permutation feature importance

PIMD path-integral molecular dynamics

QSAR quantitative structure-activity relationship

RDF radial distribution function

ReLU rectified linear unit

RMSE root-mean-square error

RMSD root-mean-square deviation

SDA shear-driven amorphous

SCAN strongly constrained and appropriately normed

SMILES simplified molecular-input line-entry system

SOAP smooth overlap of atomic positions

SOTA state-of-the-art

SRO short-range order

SW Stillinger–Weber

ta-C tetrahedral amorphous carbon

TMD temperature of maximum density

vdW van der Waals

XPS X-ray photoelectron spectroscopy

ZIF zeolitic imidazolate framework

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The development of new materials has long been a cornerstone of technological and societal advancement. From the Stone Age to the Silicon Age, the materials we engineer have shaped the tools we use and the challenges we can overcome [1]. Today, as humanity confronts pressing global challenges – climate change, sustainable energy, global health – the demand for materials with novel, highly tailored properties is more urgent than ever [2, 3]. It is estimated that 70% of all technological innovations and 20% of the industrial economy rely on advanced materials [4], making them foundational to a resilient and sustainable future. This pivotal role was recently reaffirmed in the *Advanced Materials 2030 Manifesto* [5], which highlights the strategic importance of materials innovation for environmental sustainability, economic security, and industrial competitiveness.

In response to these demands, the natural sciences are undergoing a profound transformation, driven by the convergence of high-performance computing, large-scale data infrastructures, and artificial intelligence. This shift has given rise to what is now widely referred to as the "fourth paradigm" of scientific discovery [6, 7], where data-driven methodologies play a central role in generating new knowledge. A key enabler of this paradigm is the rapid growth of massive, open-access datasets, which allow researchers to train, validate, and benchmark computational models at an unprecedented scale. In materials science, platforms such as the Materials Project [8] and the Open Quantum Materials Database (OQMD) [9] provide extensive repositories of computed and experimental materials properties, freely accessible to the global research community. These resources not only democratise access to high-quality data but also accelerate the development of predictive models and automated discovery pipelines.

At the heart of this transformation towards data-driven discovery lies machine

learning (ML), a rapidly expanding field that enables computers to identify patterns, make predictions, and generate new hypotheses without explicit programming. Long imagined as a distant goal of artificial intelligence research, ML is now a practical and increasingly indispensable tool in scientific research [10]. In materials science, it has opened powerful new avenues for exploring atomic-scale phenomena, enabling the modelling, prediction, and design of materials with remarkable speed and precision [11]. Crucially, ML methods can seamlessly integrate and exploit the wealth of data now available through open-access repositories, accelerating the feedback loop between data generation, model refinement, and scientific insight. Combined with today's computational capabilities, this approach allows researchers to perform *in silico* experiments that are not only faster and more cost-effective than traditional techniques, but also capable of revealing patterns and mechanisms beyond the reach of conventional laboratory methods [10]. As a result, research cycles that once spanned decades can now be compressed to a matter of months, fundamentally reshaping the pace and scale of materials innovation [12]. This paradigm shift has already transformed our ability to interpret complex data and simulate matter at the atomic scale yet, it merely scratches the surface of what ML can contribute to materials science. Powerful AI approaches, such as large language modelling [13] and reinforcement learning [14], have achieved widespread success in other domains including robotics [15] and game-play [16] but remain largely untapped in materials discovery. This vast potential signals a promising future, one that not only accelerates innovation but also invites a reimagining of the scientific process itself – one in which computation is not merely a tool for validation, but a true partner in discovery.

This thesis sits at the intersection of these developments, applying machine learning to one of the field's enduring frontiers: understanding the structure of amorphous materials. These systems, which lack long-range periodic order at the atomic scale, do not produce sharp Bragg diffraction peaks and remain challenging to characterise

using conventional crystallographic techniques. While some amorphous materials exhibit extended structural correlations, the overall absence of translational symmetry and the presence of structural heterogeneity make them difficult to model with traditional methods. By leveraging data-driven methods and modern ML architectures, this work aims to contribute to the wider ML for materials science movement [17, 18] that reimagines how we model, understand, and ultimately design the materials of the future.

1.1 Atomistic Modelling

Atomistic modelling is a computational approach that represents materials as assemblies of atoms whose interactions are described by mathematical models known as interatomic potentials or force fields. These potentials are typically derived from quantum mechanical calculations, experimental data, or a combination of both, and enable prediction of material properties and behaviour at the atomic scale. Such modelling routinely enables the study of structural, thermodynamic, kinetic, and transport properties of matter arising from atomic interactions, offering insights into material behaviours that are often difficult or impossible to access experimentally. By providing spatial resolution down to individual atomic positions and temporal resolution on the order of femtoseconds, along with full control over system composition and conditions, atomistic simulations serve as a unique tool for exploring phenomena across disciplines, from materials science and molecular biology to geochemistry and energy technology [19].

The origins of atomistic modelling trace back to the mid-20th century, when Alder and Wainwright first demonstrated molecular dynamics (MD) simulations of hard spheres [20], and Rahman performed the first realistic simulation of a molecular liquid, argon, in 1964 [21]. These pioneering studies laid the foundation for computational materials science, initiating the transformation of atomistic modelling from a theoretical curiosity into a core methodology. Over time, simulation has come to be recognised as the "third pillar" of scientific discovery [19], complementing the-

ory and experiment. Since then, the field has grown in scope and sophistication, benefitting enormously from the concurrent rise of high-performance computing. Today's supercomputers enable simulations of systems with millions of atoms over microsecond timescales, making atomistic modelling an indispensable tool for materials discovery, design, and characterisation [12, 22]. The central role of modelling in modern science was further recognised by the 2013 Nobel Prize awarded to Karplus, Levitt, and Warshel for pioneering multiscale and atomistic frameworks combining quantum and classical mechanics. [23–27].

One of the most important choices a practitioner needs to make when modelling materials at the atomic scale is the functional form of the interatomic potential they are going to use. Since its inception, the field has evolved through several key paradigms, with methods from each building upon those of its predecessors to improve the balance between computational efficiency and predictive accuracy.

Empirical interatomic potentials express the potential energy, $U(r)$, of a system as an analytical function of atomic positions. Atomic forces, the negative gradient of this energy with respect to atomic position, are therefore also given by an analytical expression. Empirical potentials are parametrised to reproduce experimental data or quantum-mechanical results, allowing for efficient simulations of large atomic systems. To approximate the true many-body energy, empirical potentials decompose interactions into additive terms, typically two-body, three-body, or electrostatic contributions, that reflect the dominant physical forces at play [28]. For example, covalent bonding may be represented by a harmonic spring between atoms, with the equilibrium bond length and spring constant encoding chemically specific details such as bond order or hybridization.

Commonly used empirical models include the Lennard-Jones (LJ) [29, 30], Stillinger–Weber (SW) [31], and Tersoff [32] potentials. The Lennard-Jones potential captures van der Waals interactions with a simple two-body form, while the SW potential models covalent bonding in materials like silicon through a combination of two-

and three-body terms. The Tersoff potential extends this by incorporating bond-order dependence, improving transferability across different bonding environments. For molecular systems such as water, empirical models like SPC [33, 34], TIP3P, TIP4P [35], and their variants are widely used. These models typically treat water as a rigid or semi-flexible molecule with point charges placed on atomic or virtual sites to reproduce key properties such as density, dipole moment, and hydrogen bonding behaviour. These force fields are computationally inexpensive and can be used to simulate systems with tens or hundreds of thousands of atoms over nanosecond to microsecond timescales [36].

However, the key limitation of empirical potentials is their fixed functional form: because the model is constrained to a predefined mathematical structure, it often fails to capture important classes of interactions, for instance many-bodied interactions, hydrogen bonding, π-stacking and more. As a result, empirical force fields are typically reliable only for systems and conditions similar to those used during parameter fitting – for example, known bond lengths, angles, melting points, or elastic constants in crystalline or well-characterised phases [19]. This lack of transferability becomes especially problematic in the modelling of disordered or amorphous systems, where atoms experience a wide variety of bonding environments that can differ significantly from each other. Because empirical potentials rely on a limited set of fitted parameters and functional forms, they struggle to accurately describe these diverse configurations. Despite these limitations, the computational efficiency of empirical potentials makes them a mainstay in large-scale simulations across chemistry, biology, and materials science [19].

At the other end of the computational spectrum from empirical potentials are quantum mechanical methods that aim to solve the many-body Schrödinger equation numerically. These approaches provide a more accurate alternative by explicitly accounting for the electronic structure of materials [37]. Unlike classical force fields, which depend on parameterised functional forms, quantum methods com-

pute system energies from first principles, enabling predictive modelling of chemical bonding, charge transfer, and other quantum phenomena. In principle, solving the Schrödinger equation yields exact ground-state properties, but in practice, the exponential scaling of the many-electron wavefunction with system size renders numerically exact solutions infeasible for all but the smallest systems [38]. This challenge becomes especially acute when modelling complex, disordered, or amorphous materials that require large atomic configurations to describe accurately.

To address this limitation, Density Functional Theory (DFT) was developed as a more tractable quantum-mechanical framework [39]. Introduced in the 1960s [40, 41], DFT reformulates the many-body problem in terms of the electron density, a function of only three spatial variables, rather than the full many-electron wavefunction. This reformulation replaces the complex interacting electron system with an equivalent system of non-interacting electrons that yields the same ground-state density, significantly reducing computational cost while preserving key quantum effects [39]. As a result, DFT has become one of the most widely used tools for simulating condensed matter systems [37, 42, 43], supporting applications that range from calculating electronic and structural properties to enabling high-throughput screening and *in silico* materials design [44–46].

Nevertheless, this quantum level of accuracy still comes with substantial computational expense. Solving the self-consistent field equations of DFT typically scales as $O(N^3)$ with the number of electrons N [38], which limits practical simulations to systems containing only a few hundred atoms. This makes routine application of DFT infeasible for the large, disordered structures commonly encountered in studies of amorphous materials, complex interfaces, or extended defects.

Moreover, DFT relies on approximations to account for electron–electron interactions through exchange-correlation functionals [37]. Widely used choices such as the Local Density Approximation and the Generalized Gradient Approximation [47, 48] offer good efficiency and broad applicability, but suffer from known shortcomings.

For example, they often fail to capture long-range dispersion forces and perform poorly in strongly correlated systems. More advanced functionals can improve performance, but selecting the most appropriate one is not always straightforward and often involves trade-offs between accuracy and computational feasibility [39]. These challenges motivate the development of more scalable methods that retain DFT-level accuracy, particularly when simulating large, complex materials.

ML methods have emerged as a powerful alternative to traditional approaches for parameterising interatomic interactions, effectively bridging the gap between the computational efficiency of empirical potentials and the accuracy of quantum mechanical methods [11, 49]. In essence, these models aim to learn structure–property relationships by identifying patterns in data, specifically, the connections between chemical descriptors and target properties, thereby approximating the underlying principles of quantum mechanics [50]. Crucially, ML approaches bypass the need to explicitly solve the governing physical equations during simulation, enabling significant computational speedups while maintaining high levels of predictive accuracy [51]. By training flexible models on high-quality quantum datasets, ML-based potentials can implicitly capture complex many-body interactions that would be difficult or impossible to encode using traditional analytical forms. As a result, ML has seen rapid and widespread adoption across the chemical and materials sciences [49].

One of the earliest and most impactful applications of ML in atomistic modelling has been the construction of machine-learned interatomic potentials (MLIPs), trained on quantum-mechanical labels, to enable efficient MD simulations. The pioneering work of Behler and Parrinello [52] introduced high-dimensional neural network potentials that laid the groundwork for a rapidly growing area of research [51, 53]. Rather than relying on fixed, physically motivated forms, these models learn directly from first-principles data, most commonly generated using DFT, to represent the potential energy surface (PES) with high fidelity.

These advances have enabled simulations of systems and phenomena that were

previously inaccessible using either empirical potentials or quantum methods alone. For example, Morrow *et al.* [54] performed a million-atom simulation to investigate the structure of amorphous silicon, a system that is notoriously difficult to model due to its structural heterogeneity, and which demands both large system sizes and accurate interatomic interactions [55, 56]. Similarly, Zhou *et al.* [57] carried out a device-scale simulation of a phase-change material, demonstrating that ML potentials can directly capture technologically relevant processes in memory devices.

Beyond the development of MLIPs, a broad spectrum of machine learning methods has been leveraged to accelerate materials discovery, characterisation, and design. These approaches often bypass atomistic simulations entirely by predicting key material properties – such as band gaps [58], elastic constants [59], and electronic densities of states [60] – directly from structural or compositional descriptors. By enabling rapid screening of vast chemical spaces, ML models facilitate the identification of novel compounds and the optimisation of materials for specific functionalities [61, 62]. Unsupervised learning techniques, including clustering and dimensionality reduction, have proven valuable for uncovering hidden structure–property relationships and organising large materials databases [63, 64]. More recently, generative models such as MatterGen [65] have been employed for inverse design, enabling the proposal of candidate materials with tailored properties prior to synthesis. These data-driven approaches complement atomistic modelling by offering fast, scalable alternatives, particularly advantageous when quantum-level accuracy is impractical or unnecessary.

1.2 Disordered Tetrahedral Networks

Disordered materials encompass a broad range of materials that lack conventional long range translational and orientational order (i.e. there are no Bragg peaks in their scattering intensity) [66]. These include amorphous solids, glasses, and some liquids. In contrast to crystalline materials, whose structures are defined by periodic atomic arrangements and described using unit cells and space group symme-

tries, amorphous materials lack translational symmetry and cannot be completely represented by any finite periodic model. In this thesis, the term "amorphous" will be used throughout to describe systems without translational symmetry or periodic atomic arrangements. While glasses are typically understood as amorphous solids that have undergone kinetic arrest into a metastable state (e.g., via rapid cooling), not all amorphous materials are glasses. This distinction is important and will be revisited where relevant, but the term amorphous will serve as the general descriptor unless otherwise specified.

A notable subclass of amorphous materials is disordered tetrahedral networks (DTNs), where each atom or molecular unit tends to have four nearest neighbours arranged in a roughly tetrahedral geometry. This seemingly simple structural motif gives rise to a hugely diverse range of systems, from elemental solids to molecular liquids. For example, amorphous silicon and tetrahedral amorphous carbon adopt a diamond-like coordination, with each atom covalently bonded to four neighbours. Amorphous silica (SiO_2) forms a continuous random network (CRN) of corner-sharing SiO_4 tetrahedra [67], while zeolitic imidazolate frameworks (ZIFs) feature metal ions tetrahedrally coordinated by imidazolate linkers [68, 69]. Amorphous, or glassy, water also falls within this class: although molecular in nature, its structure is defined by a disordered hydrogen-bonded network with locally tetrahedral coordination [70, 71]. Even in the liquid state, water retains transient tetrahedral character, as each H_2O molecule is, on average, hydrogen-bonded to four neighbours, forming a fluctuating local structure [72, 73].

Although disordered networks lack conventional long-range order, they are far from random at shorter lengthscales. Nearest-neighbour bond lengths in amorphous and crystalline phases are often similar, but greater disorder typically appears in bond angles and at longer distances [74]. These disordered networks exhibit short-range order (SRO), usually involving the first one or two co-ordination shells, and intermediate- or medium-range order (MRO), spanning 5–20 Å (∼2–20 neigh-

bours) [66]. The extent and scale of this order depend strongly on the material's chemical interactions.

In DTNs, SRO is defined by local tetrahedral coordination: units like SiO_4 maintain consistent bond lengths and angles (liquid water, an exception, will be discussed later). Beyond this, MRO is dictated by the connectivity and packing of tetrahedra over nanometre scales. In amorphous silica, for example, SiO_4 units form a network with rings, predominantly six-membered, which give rise to distinct features in scattering data [75], such as the first sharp diffraction peak (FSDP) at low Q (a measure of spatial frequency or inverse length scale in scattering experiments). The FSDP broadly encodes medium-range correlations and is a hallmark of ordering in many DTNs [76]. Recent work on amorphous ices has also shown that it is not only the geometry but also the topology of the network, such as the presence of non-trivial motifs like entangled loops and links, that can influence properties in amorphous systems [77]. Thus, even without crystal periodicity, DTNs show hierarchical structure: stable local geometry extends into medium-range motifs (e.g., rings or hydrogen-bonded clusters), whose topological characteristics ultimately influence bulk behaviour.

It is worth noting that some DTNs, such as amorphous silicon and amorphous ices, exhibit suppressed density fluctuations at large length scales despite lacking Bragg peaks and conventional long-range order [78, 79]. This phenomenon, known as disordered hyperuniformity [80, 81], represents a hidden form of structural organisation that is independent of translational symmetry, and is typically identified by the vanishing of the structure factor $S(Q)$ as $Q \to 0$. Such suppression reflects a collective arrangement of local motifs extending across the system and has been associated with enhanced electronic, optical, and mechanical properties [78, 82–85]. However, the analysis of disordered hyperuniformity in materials remains in its infancy, and highlights the difficulty of probing long-range structure and behaviour in disordered systems, even in prototypical systems like amorphous silicon.

Understanding the atomic arrangements of amorphous materials is critical for explaining their physical and chemical properties. However, their lack of long-range periodicity makes conventional crystallographic techniques ineffective. In crystalline solids, the structure can be determined by solving the unit cell, which is then repeated periodically in space [86]. In contrast, amorphous materials produce broad, diffuse features in diffraction patterns rather than sharp Bragg peaks, reflecting only average pairwise correlations. These diffuse signals do not offer a direct mapping to atomic positions, making it impossible to "solve" an amorphous structure using X-ray or neutron crystallography. Moreover, the atomic-scale structure, and thus the resulting properties, of an amorphous material can vary significantly depending on its preparation protocol [87–89]. These processing-dependent differences present an added layer of complexity, further emphasising the challenge of constructing realistic atomistic models that align with experimental data in the study of disordered systems.

In the absence of definitive structural data, atomistic simulations have become invaluable tools for investigating DTNs. Simulations allow researchers to explore disordered configurations at scales and resolutions beyond experimental reach. Typically, amorphous materials are modelled via melt-quench protocols, where a high-temperature liquid is rapidly cooled into a glassy state. Achieving realistic IRO in such models, however, demands both large supercells, to approximate the non-periodic nature of the material, and long simulation times to capture slow structural relaxation during cooling [66, 74].

MLIPs have recently emerged as powerful tools to overcome these limitations. By combining the accuracy of first-principles calculations with the computational efficiency of empirical models, MLIPs facilitate the generation of realistic amorphous structures, including during simulated cooling processes where conventional *ab initio* methods are computationally prohibitive and empirical models are not sufficiently accurate. MLIPs have been successfully developed for a variety of dis-

ordered systems, including carbon [90], silicon [91], silicate glasses [55, 92], battery materials [93], and phase-change alloys [57, 94].

While the primary focus of this thesis is on understanding disordered tetrahedral networks, crystalline phases are also examined, particularly where they offer useful structural context or contrast. Although crystalline and amorphous materials are often treated as fundamentally distinct, they are, in many cases, very similar on the short-range scale [95, 96]. Crystalline systems provide well-defined local coordination environments and bonding motifs that serve as valuable reference points when analysing disordered structures. Conversely, modelling amorphous phases can offer insights into the behaviour of polycrystalline materials under realistic, non-ideal conditions, where complex, high-energy grain boundaries disrupt perfect periodicity [97]. Recent studies on stress correlations further blur the boundary between crystalline and amorphous materials [98]. Despite their differing local order, both regimes display universal long-range stress correlation behaviour, underscoring commonalities in mechanical response that transcend microscopic disorder. Drawing connections between these two regimes enhances our understanding of the broader landscape of structure and disorder in tetrahedral materials.

In the following sections, I introduce the specific DTNs studied in this thesis. The discussion is divided into three parts: (i) water, (ii) ZIFs and silica, and (iii) carbon and silicon. In some cases, two systems are grouped together because they are directly compared within the thesis, with structural and chemical similarities used to inform understanding of one or both. This comparative approach is especially important in Chapter 6, where I explore the transfer of knowledge across different chemical domains.

1.2.1 Water

Water is among the most familiar substances in daily life, yet its physical behaviour remains one of the most puzzling and extensively studied in science. Despite decades of study, water's unique characteristics continue to challenge our understanding and

remain a focal point of scientific inquiry. A central source of this complexity lies in the nature of the hydrogen bond, which gives rise to a rich and intricate phase diagram – the most complex known for any pure substance – encompassing over 20 crystalline ice phases, several amorphous and glassy forms, and at least two distinct liquid states [99–103].

Water also exhibits a range of thermodynamic anomalies that defy classical expectations. Most notably, its density reaches a maximum near 4 °C at ambient pressure and decreases both upon heating and cooling from this point. At lower temperatures and pressures, water can even display a density minimum [104]. In the deeply supercooled regime – a region referred to as "no-man's land" due to rapid crystallisation precluding direct experimental access [102, 105] – several thermodynamic response functions, including isothermal compressibility, heat capacity, and thermal expansivity, show sharp, non-monotonic behaviour. These anomalies are increasingly interpreted through the lens of a hypothesised liquid-liquid critical point (LLCP), beyond which a line of maxima in response functions, known as a Widom line, emerges [106]. This Widom line demarcates the smooth but pronounced crossovers between high-density and low-density liquid structures, offering a unifying framework for understanding the anomalous behaviour of water across a broad region of its phase diagram. Accurately reproducing this phenomena remains a stringent benchmark for computational models and a vivid illustration of the delicate many-body physics at play in this hydrogen-bonded system [107–109].

Beyond reproducing known behaviour, simulations have demonstrated remarkable predictive power in regimes that remain experimentally challenging. One notable example is the prediction of superionic ice (a phase in which hydrogen atoms become diffusive within an ordered oxygen lattice) under the extreme pressure and temperature conditions relevant to planetary interiors [110]. The plastic phase of ice VII, characterised by rotational disorder of water molecules, was likewise first proposed through computational studies, long before its eventual experimental con-

firmation [111–113]. Simulations have also forecasted the existence of a ferroelectric glass phase in supercooled water, offering insight into possible arrested states of orientational order [114].

Perhaps most significantly, simulations have opened a window into the so-called "no-man's land". Within this inaccessible regime, theoretical investigations have predicted the presence of a liquid–liquid critical point (LLCP), marking a boundary between high-density and low-density liquid phases. While the experimental exploration of this region remains constrained by the timescales of ice nucleation, recent advances are beginning to shrink the limits of no-man's land, providing growing support for the LLCP hypothesis [115].

In 1969, Barker and Watts [116] kickstarted the field of water simulation by using classical statistical mechanics in conjunction with a simple rigid non-polarisable model to describe water–water interactions. This work was followed by Rahman and Stillinger [117], who employed the Ben-Naim–Stillinger potential [118] and later introduced the ST2 model [119]. These early studies used classical statistical mechanics in combination with simple, rigid, non-polarisable models to represent water-water interactions.

Building on this foundation, the development of fixed-charge models with interaction sites located on or near atomic centres gained momentum. Among the most influential were the SPC [33, 34] and TIPnP families [35, 120–123]. Particularly impactful was TIP4P [35], which introduced a massless off-centre site to better reproduce the experimental quadrupole moment. Its derivatives, including TIP4P/2005 [120], TIP4P/Ice [124], TIP5P [125], and various flexible variants [126, 127], extended the applicability of these models across a range of thermodynamic conditions and phases. These models were typically parameterised to reproduce key properties of water at ambient conditions, such as density, radial distribution functions, and enthalpy of vaporization, and proved remarkably effective for simulating bulk liquid water at relatively low computational cost.

One of the most stringent benchmarks for any water model is its ability to reproduce homogeneous ice nucleation, a process that occurs extremely slowly in real water. Only two molecular simulation studies have successfully captured this rare event: the work of Yagasaki *et al.* [128], using the TIP4P/2005 [120] model, and that of Palmer and Martelli [129], employing the ST2 model [119]. These simulations not only reproduced homogeneous nucleation kinetics but also provided strong support for the existence of a LLCP, based on unbiased sampling of the supercooled liquid. Moreover, Palmer and Martelli offered detailed insight into the kinetics of critical nucleus formation, a crucial aspect of understanding metastable liquid behaviour.

In parallel with atomistic models, coarse-grained approaches have provided valuable insight into water's behaviour while reducing computational overhead. A particularly notable example is the mW model [130], a monoatomic water model developed as a reparameterisation of the Stillinger–Weber potential [31]. Originally formulated to describe tetrahedral semiconductors such as silicon, carbon, and germanium, the Stillinger–Weber potential encodes tetrahedral bonding via a three-body term, which was retained in mW to mimic hydrogen bonding implicitly. While lacking explicit electrostatics, the mW model successfully reproduces many structural and thermodynamic anomalies of water, and offers a compelling demonstration of chemical transferability — a core theme in this thesis — by bridging water and group-IV tetrahedral solids within a unified potential framework.

Ab initio molecular dynamics (AIMD) has also emerged as a powerful alternative to MD driven by classical potentials. The first AIMD simulations of liquid water, performed in the 1990s using DFT, marked a major advance in the field [131, 132]. Since then, AIMD has been widely applied to investigate the structure, dynamics, and electronic properties of water [133–135], as well as more complex aqueous systems [136]. In principle, AIMD offers predictive accuracy without the need for empirical parameterisation. However, in practice, its reliability is constrained by the approximations built into current exchange–correlation functionals [137, 138].

A persistent challenge has been reproducing the equilibrium density of liquid water using standard DFT. Achieving agreement with experiment typically requires the inclusion of dispersion interactions [109, 138, 139]. Moreover, both the structural and dynamical properties predicted by AIMD are highly sensitive to the choice of functional [138, 140, 141]. For example, the widely used PBE functional [142] tends to over-structure the liquid and underestimate diffusivity due to its overestimation of hydrogen bond strength [134, 143]. More advanced functionals, such as the strongly constrained and appropriately normed (SCAN) functional [144], and approaches that include van der Waals interactions [145, 146], have mitigated some of these deficiencies [141, 147], yielding more accurate descriptions of liquid water.

A related and long-standing issue is the so-called density inversion problem, wherein standard DFT incorrectly predicts hexagonal ice to be denser than liquid water, a reversal of the experimental trend [148–151]. This discrepancy persists even when using hybrid functionals [148], and is only alleviated by more recent developments such as SCAN, due to its ability to describe vdW interactions on intermediate length scales [133].

Another critical difficulty is the treatment of nuclear quantum effects (NQEs), which can strongly influence the structure, dynamics, and thermodynamics of water, especially at low temperatures or in hydrogen-bond-dominated environments [152, 153]. Standard AIMD treats nuclei as classical particles, neglecting zero-point energy and quantum delocalisation. Capturing NQEs typically requires path-integral molecular dynamics (PIMD) [154–156], a computationally intensive extension that further compounds the cost and complexity of simulations.

In response to this accuracy-efficiency trade-off, MLIPs have rapidly gained prominence. Early models such as T4NN [157] used neural networks to capture polarization corrections to classical force fields. Gaussian Approximation Potentials (GAPs) [158] incorporated high-level quantum mechanical data to improve the accuracy of cluster energetics and liquid structure [159]. A major breakthrough came

with the development of high-dimensional neural network potentials (HDNNPs) by Morawietz *et al.* [109], trained directly on DFT data with and without dispersion corrections. These models enabled the accurate prediction of properties such as the melting temperature (within 10 K of experimental value) and converged values for the dielectric constant not previously obtained via AIMD. By drastically lowering the cost of large-scale simulations, HDNNPs also enabled systematic benchmarking of exchange–correlation functionals across the water phase diagram. These studies revealed that standard GGA functionals without dispersion dramatically underestimated liquid densities (by up to 40%), failed to reproduce the density maximum, and consistently overpredicted melting temperatures. In contrast, functionals with van der Waals corrections achieved much closer agreement with experimental results, underscoring the essential role of dispersion interactions in accurately modelling water's anomalous thermodynamic behaviour [109].

Since then, neural network potentials have been widely adopted to study water across a broad range of thermodynamic conditions [160], providing detailed insights into phenomena such as hydrogen-bond dynamics [161], isotope effects [162], and interfacial kinetics [163]. A central tool in this effort has been the DeePMD framework [164], introduced in 2018, which enables simulations at scales far beyond the reach of AIMD while retaining *ab initio*-level accuracy. DeePMD has been used to simulate large portions of the water phase diagram – including liquid water, numerous ice polymorphs, and both low- and high-density amorphous phases – with near-quantitative agreement with experiment [165, 166]. However, it has notable limitations: for instance, it failed to recover the plastic phase of ice VII, instead predicting a non-physical structure in its place. Moreover, the original DeePMD architecture lacks rotational equivariance, a key symmetry in molecular systems, which may limit its ability to generalise across diverse environments. This limitation has only recently begun to be addressed in newer, symmetry-aware variants.

GNN architectures offer a compelling solution to address the lack of rotational

equivariance in models. Equivariant GNNs such as NequIP are able to capture the directional nature of hydrogen bonding with high fidelity, and require significantly less training data than earlier approaches like DeepMD [167]. Recent frameworks like Allegro [168] and MACE-MP-0 [169] further demonstrate how equivariant models trained on general or limited datasets can nevertheless generalise well to aqueous systems. For instance, Allegro, trained solely on liquid water, reproduces vapour–liquid equilibrium densities and the relative stability of several ice phases [170]. MACE-MP-0, a model trained only on Materials Project data [8] which consists primarily of inorganic crystals and is skewed heavily towards oxides, shows good agreement with DFT in simulations of liquid water, multiple ice polymorphs, proton transfer, and aqueous interfaces [169]. GNN-based models thus represent a promising direction for accurate, transferable, and data-efficient water potentials. However, they are not yet widely adopted in practice, and trade-offs remain: while they can match or surpass DeePMD in accuracy, they are often slower than empirical models and still lack comprehensive benchmarks across the breadth of water's anomalous properties.

Motivated by this recent progress, the present work explores how neural network-based models can be leveraged to probe the complexity of disordered water phases. Chapter 3 employs neural network-based methods to investigate the local structure of amorphous ices and their interrelations, while Chapter 4 builds on recent advances in GNN-based modelling to address open questions surrounding the structure of liquid water.

1.2.2 Silica & Zeolitic Imidazolate Frameworks (ZIFs)

Silica is a prototypical tetrahedral network material composed of SiO_4 tetrahedra linked via shared oxygen atoms to form extended three-dimensional frameworks. It is one of the most abundant materials on Earth and exhibits remarkable versatility, with applications ranging from glassmaking to catalysis [171]. The SiO_4 tetrahedron is the fundamental building block of silica, and its arrangement – particularly the Si–O–Si bond angle – plays a crucial role in determining the material's structure and

properties. In crystalline forms such as quartz or cristobalite, these corner-sharing tetrahedra adopt regular, ordered networks, giving rise to characteristic ring motifs and channel-like voids [171, 172]. In the amorphous form, long-range periodicity is lost, but local tetrahedral coordination remains largely intact [173]. This structure is well described by the CRN model proposed by Zachariasen [67].

The structural and chemical regularity of silica has also made it a natural focus for atomistic modelling. A broad range of techniques, from empirical force fields [174] to *ab initio* molecular dynamics [175] and MLIPs [55], have been developed to capture its structural and thermodynamic behaviour. In the amorphous phase, these simulations reveal that silica maintains strong short-range order and exhibits medium-range correlations that manifest in features such as well-defined ring statistic distributions and the FSDP observed in scattering experiments [176, 177]. These modelling efforts have laid the groundwork for much of our understanding of tetrahedral network formation and disorder.

Despite its many technological applications [171], silica also exhibits important limitations. The rigidity of the Si–O–Si framework restricts the set of accessible topologies, with only a small fraction of the theoretically possible four-connected networks [178–180] ever realised experimentally. Furthermore, silica's purely inorganic nature limits opportunities for chemical functionalisation and reduces compatibility with organic or hybrid materials. These constraints have motivated the search for more chemically versatile analogues.

In response to these limitations, ZIFs have emerged as promising alternatives [68, 181]. ZIFs are a subclass of metal–organic frameworks (MOFs) that form structurally analogous but chemically distinct tetrahedral networks. They consist of tetrahedrally coordinated metal cations (typically Zn^{2+} or Co^{2+}) linked by imidazolate (Im^{-}) ligands. While preserving the corner-sharing tetrahedral geometry characteristic of SiO_2 networks, ZIFs offer significantly greater chemical flexibility [68]. Their modular design allows functional groups to be introduced via the

organic linkers, enabling fine-tuning of adsorption, reactivity, and mechanical response, and expanding the range of structural and functional behaviour far beyond what is possible in silica [182–184].

A key insight offered by Park *et al.* [68] was that the M–Im–M bridging angle (145°) closely matched the Si–O–Si angle found in many silicate frameworks. This suggested that, under suitable conditions, metal–imidazolate frameworks could replicate the open, topologically rich networks observed in silica-based materials. This structural analogy has guided the rational design of ZIFs, with many frameworks mimicking known silicate nets such as sodalite, quartz, and analcime [133, 182].

The appeal of ZIFs lies not only in their functional properties but also in the scientific challenges they present. Unlike pure silica, ZIFs span a vast and chemically diverse design space [184]. The range of available organic linkers and the complexity of linker–linker interactions enable structures and properties that go well beyond those of purely inorganic networks.

Most theoretical and synthetic approaches to designing ZIFs rely on heuristic mappings to silicate topologies, based on the shared tetrahedral AB_2 framework. While this analogy provides a useful geometric guide for structure enumeration and synthesis [185, 186], it is limited in scope. Specifically, it does not account for differences in bonding chemistry, electronic structure, or thermodynamic stability that arise from the distinct chemical environments in ZIFs compared to silicates. The question of why certain ZIF topologies are experimentally accessible, while many others are not, remains open. Despite their geometric analogy to silica, ZIFs and silicates often favour different topologies [68, 187], and the reasons for this divergence remain poorly understood [188].

This thesis explores whether the silica–ZIF analogy extends beyond structural similarities to energetics. Chapter 5 investigates whether coarse-grained machine learning models, trained on various atomistic resolutions, can accurately reproduce the energetic landscape of ZIFs. While this chapters focus primarily on crystalline

ZIFs and their relationship to well-characterised systems like silica, the longer-term goal is to extend this understanding to their disordered counterparts. Amorphous ZIFs remain poorly understood, both structurally and energetically [189]. By developing tools and frameworks that clarify the energetic and topological behaviour of crystalline ZIFs, particularly in comparison to silica, we hopefully take an important step toward building transferable models and conceptual tools that could one day enable the systematic study of amorphous ZIFs.

1.2.3 Carbon and Silicon

Carbon and silicon are two of the most consequential elements in the natural and technological world. Carbon, the basis of all known life [190], forms the structural and chemical foundation of organic molecules, biopolymers, and cellular systems. Beyond biology, carbon's structural versatility enables a wide range of technological applications, from battery electrodes [191] to advanced optical technologies [192, 193]. Silicon, in contrast, underpins the modern digital era: its semiconducting properties and structural versatility make it indispensable to solar cells, integrated circuits, and photonic devices [194, 195]. Despite their vastly different domains of influence, these elements share a common feature at the atomic scale: the ability to form extended tetrahedral networks. In their elemental crystalline phases, both adopt the diamond structure, where each atom is covalently bonded to four neighbours in a tetrahedral geometry. However, when this long-range periodicity is disrupted, as in the amorphous phase, the structural similarities between carbon and silicon begin to diverge in fundamental ways [196].

Amorphous silicon (a-Si) has traditionally been understood through the lens of the CRN model [197], in which atoms retain fourfold coordination while losing long-range order. This model, originally developed by Zachariasen [67], preserves short range order with local tetrahedral geometry while introducing topological disorder at medium and long lengthscales, and remains the dominant conceptual framework for interpreting a-Si structure. Nevertheless, despite decades of experimental

and theoretical study, key questions remain unresolved [198, 199]. The extent of medium-range order, the role and distribution of coordination defects, and the possible presence of nanoscale heterogeneity or paracrystalline regions continue to be subjects of debate. Recent advances in atomistic modelling, particularly the development of MLIPs, have enabled simulations of a-Si at previously inaccessible nanometre scales, encompassing hundreds of thousands of atoms [54, 89]. While features such as bond angle distributions, ring statistics, and the FSDP had been studied in smaller systems, these large-scale models now provide statistically meaningful and experimentally consistent descriptions of such structural metrics. These models suggest a rich structural landscape where disorder coexists with subtle, locally ordered motifs – such as small paracrystalline clusters – which leave detectable signatures in structural and scattering data [89].

Tetrahedral amorphous carbon (ta-C) represents the closest analogue to a-Si in its high-density, sp^3-rich form, where most carbon atoms are fourfold coordinated in a disordered, diamond-like network. Yet even in the amorphous phase, carbon exhibits far greater structural diversity than silicon, owing to its ability to easily adopt sp, sp^2, and sp^3 hybridisations [200]. This bonding flexibility gives rise to a wide configurational space, ranging from ta-C to graphitic, sp^2-dominated networks, and extending to low-dimensional structures such as nanotubes and fullerenes. While this diversity underpins carbon's functional versatility, it also poses significant challenges for structural modelling. In the high-sp^3 regime, ta-C can be approximately described using CRN principles [201, 202], similar to those employed for a-Si. However, even dense ta-C films rarely exceed 90% sp^3 content under realistic deposition conditions [203, 204], indicating that ideal CRN-like "amorphous diamond" structures are best viewed as limiting or idealised cases. The central modelling challenge for amorphous carbon, therefore, lies not in capturing one structure type, but in representing its full hybridisation landscape within a single transferable framework capable of describing its full bonding diversity [205]. Recent MLIPs, such as C-GAP-

20 [206], have made meaningful progress in this direction. GAP-20 reproduces the formation energies of key carbon phases – including diamond, graphite, fullerenes, and nanotubes – with errors of only a few meV per atom. It also achieves comparable accuracy for a range of crystalline and amorphous surfaces, as well as for point defects, where its predictions are markedly more accurate than those from empirical potentials. In addition, it captures the behaviour of high-temperature liquid carbon across wide ranges of temperature and density. This accuracy, however, comes with substantial computational cost relative to empirical methods (1 to 2 orders of magnitude), highlighting a persistent trade-off between fidelity and efficiency that continues to drive the search for scalable, transferable models.

Both amorphous carbon and silicon have been studied extensively within their respective domains, each presenting unique modelling challenges as discussed above. However, direct comparisons and transferable insights between the two systems have received comparatively little attention. Early efforts using variants of the Stillinger–Weber potential sought to provide a unified framework for modelling tetrahedral elements such as Si and C by tuning interaction cutoffs and angular parameters to reflect local bonding environments [207, 208]. While these models successfully reproduced key properties of sp^3-bonded phases like diamond, they lacked the flexibility to capture the full hybridisation landscape of carbon, particularly in non-tetrahedral environments. As a result, the broader question of whether structural or energetic knowledge from one tetrahedral element can inform predictions about another remains largely unaddressed. From an ML perspective, this gap presents a valuable opportunity: the shared tetrahedral coordination and relatively simple local structures of amorphous carbon and silicon make them ideal testbeds for alchemical transfer learning – an approach in which models trained on one set of chemical elements are leveraged to improve performance on another.

A recent proof-of-concept by Gardner *et al.* [209] demonstrated the feasibility of this approach; they showed that a message-passing neural network trained on

both elements, can interpolate between the energy landscapes of carbon and silicon. However, the broader applicability of this approach, particularly across different amorphous and crystalline phases, was not explored.

Chapter 6 builds on this foundation by providing a more detailed analysis of model transferability between carbon and silicon, before extending these strategies to more complex AB_2 tetrahedral systems, such as silica, and water.

1.3 Research aims and thesis structure

The central aim of this thesis is to advance the atomistic understanding of disordered tetrahedral networks using machine learning. At the heart of this work is the recurring motif of the tetrahedral building block and a desire to understand its structural diversity across chemical systems, as well as its influence on material properties. By examining systems that share this local geometry, from molecular liquids to solids, this thesis investigates how the tetrahedral motif adapts across chemical contexts and whether these variations can be captured and generalised through machine learning models.

The research is developed in two stages. First, I apply machine learning techniques to study specific systems: amorphous ices, liquid water and ZIFs. These case studies illuminate the challenges of modelling disordered materials and demonstrate how data-driven approaches can reveal structural insights beyond the reach of empirical or quantum methods. In the second phase, I expand the scope to explore how chemically distinct tetrahedral networks relate to one another. The key question becomes whether structural and energetic patterns learnt in one system can be transferred to another, testing the extent to which tetrahedral topology encodes chemically transferable information.

Chapter 2 lays the methodological foundation for the thesis, and in particular provides details on the development of MLIPs used in this work. I introduce the three core components of MLIP construction – reference data, structural descriptors, and fitting algorithms – before explaining how their interplay governs model accuracy

and transferability.

In Chapter 3, I present a machine learning framework for classifying local atomic environments in amorphous ices, focusing on low-density (LDA), high-density (HDA), and the recently discovered medium-density amorphous (MDA) phase. Using bond orientational order parameters and a neural network classifier, I show that while HDA is clearly a distinct phase, LDA and MDA exhibit significant structural overlap. Applying this model to compression trajectories suggests that MDA-like environments emerge transiently during the LDA–HDA transition, indicating that MDA may be a metastable intermediate rather than a distinct phase. This study illustrates both the power and the limitations of local descriptors in resolving subtle structural differences in disordered materials.

Chapter 4 applies a GNN potential to model the structure of liquid water. Although GNN-based potentials have been used previously, this work presents one of the first in-depth applications aimed at extracting new structural insights in liquid water. The resulting potential enables accurate simulations across a temperature range of 260 K to 350 K, successfully reproducing key experimental observables including the radial distribution function, density isobar, and diffusion coefficient. I use this model to investigate the temperature-dependent splitting of the principal peak in the structure factor, a hallmark of structural heterogeneity in liquid water. Analysis of hydrogen-bonded ring statistics reveals that this splitting arises from distinct topological rearrangements within the hydrogen-bond network. By linking medium-range topological order to measurable scattering features, this work offers new insight into the microscopic origins of water's structural anomalies.

Chapter 5 examines the analogy between ZIFs and zeolites through the lens of structural coarse-graining. Using Gaussian Process Regression (GPR), I build models at three levels of resolution: fully atomistic, an AB_2 coarse-grained representation preserving tetrahedral connectivity, and a minimal A-only model omitting it. I find that the AB_2 model retains much of the predictive accuracy of the atomistic one

while dramatically simplifying the representation. By contrast, the A-only model fails to capture local energetics, underscoring the essential role of tetrahedral connectivity. These findings provide quantitative support for the view that ZIFs and zeolites can be described within a unified topological framework and supports the idea that topology-driven representations can form the basis of transferable models.

Chapter 6 explores the transferability of machine learning models across different chemical domains that share tetrahedral structures. Using transfer learning, I test how GNNs trained on one material – such as carbon, silicon, silica, or water – can be adapted to another. This cross-domain study probes the limits of alchemical transfer learning and evaluates whether the tetrahedral motif carries enough embedded chemical information to enable generalisation. I identify key factors governing successful transfer and examine how GNNs represent structure and chemistry across systems. Finally, I demonstrate how a foundation GNN model can be fine-tuned on water data to enable large-scale simulations with modest computational cost, pointing toward more general and efficient machine-learned potentials.

Chapter 7 concludes the thesis by reflecting on the main findings and outlining directions for future research, including how the tools and approaches developed here may inform broader efforts in materials modelling.

Each research chapter introduces the relevant scientific context, followed by the methods, main results, and their interpretation, offering a systematic exploration of how machine learning can be used to capture, analyse, and generalise structural patterns in complex materials.

Chapter 2

Methods

2.1 Introduction

This chapter provides a technical introduction to the core tools and techniques that underpin this thesis as a whole. It outlines the foundational methods and concepts that will be referenced throughout the work, establishing a cohesive framework for the research. While this chapter focuses on the general methodological approach, specific techniques pertinent to individual chapters will be introduced in those chapters as needed.

2.2 Building machine-learned interatomic potentials

An MLIP aims to approximate the true PES of a material by learning a function that maps atomic configurations to their associated energies and forces. This function is not derived analytically from first principles, but instead is trained on a finite set of reference data, typically obtained from DFT or higher-level quantum chemical methods. The trained model then serves as a surrogate for the full electronic structure calculation, enabling rapid evaluation of energies and forces at a fraction of the computational cost.

The process of constructing an ML potential involves three essential components, illustrated schematically in Fig. 2.1, and discussed in detail in the following sections.

2.2.1 Construction of a reference dataset

The foundation of any MLIP is its reference dataset, which encodes all the physical information the model will learn to reproduce. The quality, diversity, and relevance of this dataset are critical to the potential's downstream accuracy, transferability,

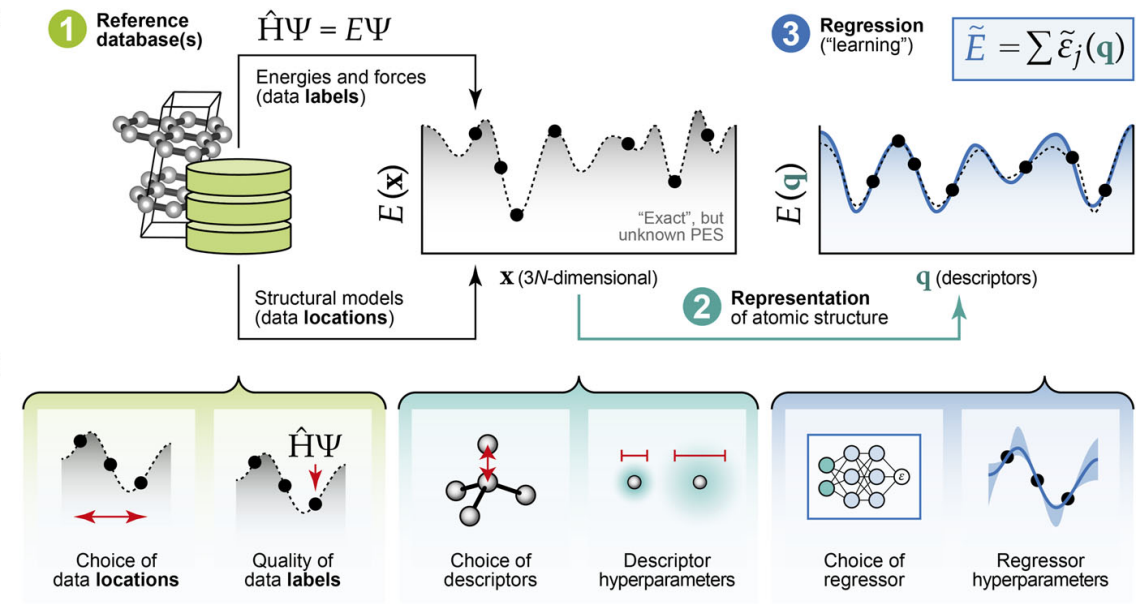

Figure 2.1: Building machine learning interatomic potentials. The process involves three main steps: (1) construction of a reference dataset, (2) representation of chemical environments, and (3) regression of the MLIP [51]. The reference dataset provides the ground truth for the MLIP, while the representation of chemical environments allows the MLIP to learn the underlying potential energy surface. The regression step uses the reference dataset and the representation of chemical environments to learn the MLIP. The figure is adapted from Ref. [210].

and robustness. Consequently, much effort has been devoted to designing datasets that are both compact and comprehensive.

A typical dataset consists of atomic configurations paired with quantum mechanical labels, usually total energies, atomic forces, and stress tensors, which serve as ground truth during training. These configurations should provide a representative sampling of the PES relevant to the intended application, allowing the model to make accurate and physically meaningful predictions across the targeted domain. If key regions of this domain are poorly sampled, the trained potential may be forced to extrapolate, often resulting in large errors or unphysical behaviour [211].

Constructing such a dataset involves two main tasks: (1) selecting the configurations to include (data locations), and (2) computing their corresponding quantum-mechanical labels. Together, these steps determine the coverage and fidelity of the resulting potential.

The first choice to make is the selection of data locations, which governs the configurational diversity of the dataset. To ensure reliability in downstream simulations, the data locations must be representative of the system of interest, capturing both typical and edge-case configurations that may arise under relevant thermodynamic or mechanical conditions. These can be chosen manually using domain expertise, for example, by selecting representative configurations from known phases, surfaces, or defects, or through automated techniques such as active learning [212–215] , random structure search [216–218], or metadynamics-driven sampling [219]. In practice, hybrid strategies are often most effective: domain expertise ensures inclusion of key structural motifs, while algorithmic exploration helps identify non-intuitive but relevant configurations [57, 220]. Despite best efforts, any dataset will inevitably be incomplete relative to the true configuration space of a complex material. As a result, the trained model may need to extrapolate in regions it has not seen during training, often leading to significant errors or unphysical behaviour. Rigorous validation and uncertainty quantification are therefore essential to assess the reliability and limitations of any ML potential.

The second key aspect is the quality of the reference labels. The ML potential cannot exceed the accuracy of the data on which it is trained; it merely reproduces the level of theory used to generate the dataset. For instance, if a DFT functional systematically underestimates dispersion interactions, the resulting potential will inherit this bias. It is therefore essential to ensure that the reference calculations are both consistent and numerically precise.

Together, the accuracy of the labels and the coverage of the configurations define the upper bounds of the ML potential's performance and transferability. A well-constructed reference dataset enables the model to interpolate reliably within its training domain, but also to exhibit stable behaviour in simulations that venture toward its boundaries.

In this thesis, no new reference data sets are generated. Instead, all datasets

are sourced from existing literature. Details on the specific datasets used, including their origin, composition, and suitability are provided in the respective chapters where they are used.

2.2.2 Representing Chemical Environments

The second step in developing an MLIP involves encoding the local atomic environments from the dataset into a numerical representation that can serve as input to the machine learning model.

An (uncharged, isolated) atomic structure, S, is most simply defined by the position, $\mathbf{r}_i$, and atomic number, Z_i, of each of its N component atoms:

$$S = \{(\mathbf{r}_i, Z_i) \quad \forall\, i \in [1, N]\}$$

However, while complete, this raw representation poses several challenges for computational analyses: it lacks critical invariances, such as rotational, translational, and permutation symmetry, and tends to be neither smooth nor fixed in length. Consequently, this makes direct use of atomistic configurations difficult, particularly for machine learning models which require standardised, consistent, and invariant inputs.

To address these challenges, we employ descriptors, also known as fingerprints [221]. These are transformations of the original atomistic configurations into structured, numerical representations more suited to computational methods. Effective descriptors should satisfy specific criteria [222, 223], including invariance to symmetries, uniqueness to prevent ambiguous representations, smoothness to facilitate stable gradient-based learning methods, faithful representation of underlying physics and chemistry, and strong correlations with system properties of interest. Developing descriptors which meet these conditions significantly enhances our ability to leverage computational tools, particularly machine learning algorithms, to predict system properties and behaviours as each configuration of interest is mapped to a

unique point in a high-dimensional feature space.

Global descriptors represent one class of such transformations, summarising the entire system in a single, often compact, numerical vector or string. Global descriptors have become prominent in particular applications: for example, Quantitative Structure-Activity Relationship (QSAR) modelling in drug discovery frequently utilises simplified molecular-input line-entry system (SMILES) strings [224–226] due to their concise representation of molecular topology. Electrostatic interaction analyses commonly employ Coulomb matrices to quantify pairwise atomic interactions effectively [227]. Additionally, polymer science and protein folding studies often use descriptors such as the radius of gyration, capturing global structural features essential for understanding macromolecular behaviours [228]. Despite their utility in high-throughput screening and rapid analyses, global descriptors inherently sacrifice local structural information, as they typically represent averaged or aggregated system properties. This loss is especially problematic in systems lacking long-range order, such as amorphous or disordered materials, where properties are critically influenced by local structural motifs.

To capture detailed local interactions, particularly in disordered or heterogeneous systems, local descriptors are used. These focus on individual atoms and their surrounding environments, typically defined by a cutoff radius, enabling models to learn structure–property relationships that depend on local atomic arrangements. Simple descriptors may include coordination numbers or distance-based features (e.g., bag-of-bonds [229]), while more expressive approaches rely on many-body expansions, such as the smooth overlap of atomic positions (SOAP) descriptor [230] or atom-centred symmetry functions (ACSF) [231]. Local descriptors have become central to atomistic machine learning and have been applied across diverse domains, from modelling amorphous silicon [54] to characterising grain boundaries in alloys [232]. By encoding the geometric and chemical structure of an atom's local environment, these descriptors allow models to infer the underlying interatomic interactions, both

short- and many-body, which ultimately govern mechanical, electronic, and thermodynamic behaviour.

The optimal choice of descriptor remains task-specific, requiring careful consideration of both the chemical context and the targeted predictive property. Consequently, selecting the best descriptor is not straightforward and remains an active area of research [233]. For a comprehensive review of the diverse range of structural descriptors used in atomistic machine learning, see Ref. [222]. In the following sections, I introduce the three descriptors used in this thesis: the Steinhardt order parameters, the SOAP descriptor, and the atomic cluster expansion (ACE) descriptor.

Steinhardt Order Parameters

Bond–orientational order (BOO) parameters, commonly known as *Steinhardt parameters*, were introduced in the early 1980s by Steinhardt, Nelson, and Ronchetti [234] to quantitatively characterise the local orientational order in particle systems. These parameters address the shortcomings of traditional measures, e.g., $g(r)$, in capturing the local structure of systems like liquids and glasses, where long-range order is absent yet local structural motifs persist. Since their introduction, Steinhardt parameters have become a widely used and powerful tool for identifying various crystalline phases, including face-centred cubic (fcc), hexagonal close-packed (hcp), and body-centred cubic (bcc) structures [235–238].

Steinhardt parameters employ spherical harmonics to encode the symmetries of the local structures into a complex vector $\mathbf{q}(i)$ indexed by $l \in \mathbb{N}$ and $m \in [-l, l]$, with components given by:

$$q_{lm}(i) = \frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} Y_l^m(\mathbf{r}_{ij}) \tag{2.1}$$

where $\mathcal{N}_i$ is the set of neighbours of particle i (typically defined by some user-specified distance cutoff), and $Y_l^m(\mathbf{r}_{ij})$ are the spherical harmonics evaluated at the

relative position vector $\mathbf{r}_{ij}$ pointing from particle i to j.

Thermal fluctuations can blur the distributions of these order parameters, complicating the task of distinguishing local crystalline structures. To overcome this, Lechner and Dellago [236] proposed an averaging scheme in which the bond order parameters of neighbouring particles are combined. The averaged bond order parameter is given by

$$\bar{q}_{lm}(i) = \frac{1}{|\mathcal{N}_i| + 1} \sum_{j \in \mathcal{N}_i \cup \{i\}} q_{lm}(j) \tag{2.2}$$

While $q_{lm}(i)$ holds the information about of the structure of the first shell around particle i, $\bar{q}_{lm}(i)$ also takes into account the structure of the second shell. This averaging substantially reduces the overlap between the order parameter distributions of different phases, often by several orders of magnitude [236], enhancing the accuracy of phase identification at the expense of a modest reduction in spatial resolution.

Evaluations of both $q_{lm}(i)$ and $\bar{q}_{lm}(i)$ depend on the choice of reference frame. It is therefore standard practice to average over the magnetic quantum number m to obtain rotationally invariant measures of local symmetry:

$$q_l(i) = \sqrt{\frac{4\pi}{2l+1} \sum_{m=-l}^{l} |q_{lm}(i)|^2} \qquad \text{and} \qquad \bar{q}_l(i) = \sqrt{\frac{4\pi}{2l+1} \sum_{m=-l}^{l} |\bar{q}_{lm}(i)|^2} \tag{2.3}$$

The spherical harmonics of a given degree l form an orthonormal basis for the $(2l+1)$-dimensional representation of the rotation group SO(3). This basis is naturally aligned with the symmetries observed in crystalline structures. Consequently, by choosing different l values, one can tailor the sensitivity of these order parameters to various crystal symmetries. For example, the parameter q_6 has been used as a prominent indicator when searching for glass transitions [239, 240] and crystalline clusters [241, 242] in glasses and super-cooled liquids.

In addition to the q_l parameters, which quantify the overall magnitude of l-fold symmetry, one can define a cubic order parameter, $w_l(i)$ [234]. This parameter

captures the relative phases of the spherical harmonic components by combining three spherical harmonic coefficients via the Wigner $3j$ symbol [243]. The cubic order parameter is given by

$$w_l(i) = \frac{\sum_{m_1+m_2+m_3=0} \begin{pmatrix} l & l & l \\ m_1 & m_2 & m_3 \end{pmatrix} q_{lm_1}(i) q_{lm_2}(i) q_{lm_3}(i)}{\left(\sum_{m=-l}^{l} |q_{lm}(i)|^2 \right)^{3/2}} \tag{2.4}$$

This parameter is sensitive not only to the magnitude of order but also to detailed angular correlations, enabling the discrimination between environments that exhibit similar $\bar{q}_l$ values but possess distinct local symmetries.

Similarly to $\bar{q}_l$, an averaged version of the cubic order parameter, $\bar{w}_l(i)$ can be defined using the averaged bond order parameters, $\bar{q}_{lm}(i)$:

$$\bar{w}_l(i) = \frac{\sum_{m_1+m_2+m_3=0} \begin{pmatrix} l & l & l \\ m_1 & m_2 & m_3 \end{pmatrix} \bar{q}_{lm_1}(i) \bar{q}_{lm_2}(i) \bar{q}_{lm_3}(i)}{\left(\sum_{m=-l}^{l} |\bar{q}_{lm}(i)|^2 \right)^{3/2}}, \tag{2.5}$$

This formulation further refines the characterization of local symmetry by incorporating both magnitude and phase information over an extended neighbourhood.

A combination of these parameters can be used to describe a local environment depending on the application. In practical use, a Steinhardt descriptor is often constructed as a feature vector composed of one or more scalar bond-order metrics, such as $\bar{q}_l(i)$ and $\bar{w}_l(i)$, for various values of l. The choice of l values can be tuned to enhance sensitivity to specific local symmetries or structural motifs. For instance, a descriptor vector might include $\bar{q}_4$, $\bar{q}_6$, and $\bar{w}_6$, or a broader set depending on the desired discriminatory power. This vector, encapsulating local orientational and angular correlation information, serves as a compact numerical representation of a particle's local environment. As such, it is well-suited for use as input to machine learning models, where it enables the automated classification, clustering, or regression analysis of structural phases across diverse particle systems [244].

SOAP

The *Smooth Overlap of Atomic Positions* (SOAP) descriptor was originally developed by Bartók *et al.* [230] in the context of ML potentials, but has since been used for classifying and understanding complex structures across a wide range of chemical systems [245].

Without any loss of information or generality, the description of the local environment of an atom can be converted from a set of discrete neighbours (locations and atomic identities) to a continuous density function, $\rho_i(\mathbf{r})$, created by centering a Gaussian function of width σ on each neighbour j up to a smoothed cutoff radius:

$$\rho_i(\mathbf{r}) = \sum_{j=1}^{N_{env}} \exp\left(-\frac{|\mathbf{r}-\mathbf{r}_{ij}|^2}{2\sigma^2}\right) f_{\text{cut}}(r_{ij}) \tag{2.6}$$

(where I have dropped the explicit dependence on the atomic number Z_i for simplicity: see Ref. [230] for more details).

According to Hilbert space theory, this atomic neighbour density can be expanded using an appropriate radial basis set, R_n, and spherical harmonics, Y_l^m:

$$\begin{aligned} \rho_i(\mathbf{r}) &= \sum_{nlm}^{n_{\max},l_{\max}} c_{nlm}^{(i)} R_n(r) Y_l^m(\hat{\mathbf{r}}) \\ c_{nlm}^{(i)} &= \int \rho_i(\mathbf{r}) R_n(r)^* Y_l^m(\hat{\mathbf{r}})^* \mathrm{d}\mathbf{r}. \end{aligned} \tag{2.7}$$

Truncating at finite $n_{\max}$ and $l_{\max}$ gives a a fixed-length descriptor, converting the continuous spatial distribution into a discrete set of coefficients, c_{nlm}, that characterise local atomic arrangements.

To achieve rotational invariance, the expansion coefficients, c_{nlm} are combined into a power spectrum, $\mathbf{p}^{(i)}$:

$$p_{n_1 n_2 l}^{(i)} = \frac{1}{\sqrt{2l+1}} \sum_m (c_{n_1 lm}^i)^* c_{n_2 lm}^i. \tag{2.8}$$

This power spectrum captures all the essential structural details of the local environment around atom i.

A particularly convenient feature of this representation is that the dot product of the power spectra of two environments provides an approximation to the integral of the product of their densities over all possible three-dimensional rotations. This dot product (typically raised to a power of 2 or 4 and denoted as $k(i,j)$) is known as the SOAP kernel, and serves as a similarity measure between two local atomic environments taking values between 0 (completely dissimilar) and 1 (identical local environments).

$$k(i,j) = \mathbf{p}_i \cdot \mathbf{p}_j \propto \int \left| \int \rho_i(\mathbf{r}) \rho_j \left(\hat{\mathbf{R}}\mathbf{r} \right) dr \right|^2 d\hat{R}. \tag{2.9}$$

In the limit of a complete basis set (i.e. $n_{\max} \to \infty$ and $l_{\max} \to \infty$), this approximation becomes exact.

While the preceding description assumes a single element for simplicity, SOAP naturally extends to multi-element systems. A straightforward extension assigns separate density channels for each element-element pair, although this significantly increases the descriptor size. Advanced strategies, such as linear combinations of densities, can effectively reduce dimensionality without significant loss of information [246].

There are two main hyperparameters that need to be chosen when using SOAP descriptors. These are often system-specific and need to be chosen carefully to ensure the descriptor is sensitive to the relevant features of the system.

The Gaussian width σ influences how smoothly neighbour atoms contribute to the local density. Smaller values of σ result in sharp, well-defined atomic peaks, making the descriptor sensitive to minor positional changes, whereas larger values lead to smoother, broader peaks that provide robustness against small perturbations in chemical environment but at the expense of sensitivity to local structure.

The cutoff radius r_{cut} determines the extent of the local environment considered.

A larger cutoff includes more neighbours, capturing longer-range interactions at the cost of higher computational complexity and potentially diminishing sensitivity to local atomic detail.

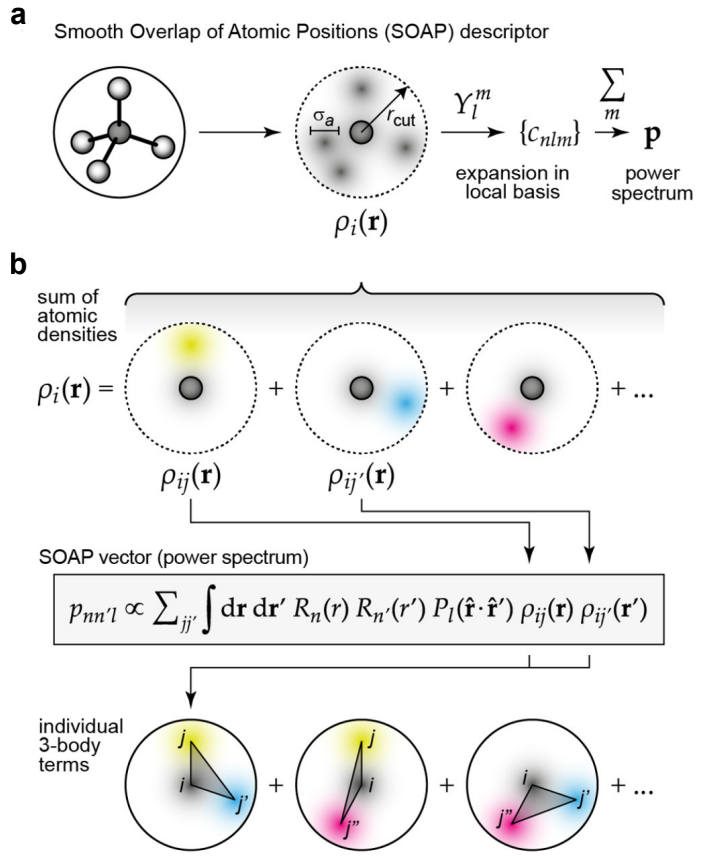

Figure 2.2: Visualisation of the SOAP descriptor for atomistic structures. (a) Schematic representation of the SOAP descriptor. The neighbour density $\rho(i)$ around atom i is constructed to be invariant to permutations of atoms. Expanding $\rho(i)$ in a basis of spherical harmonics (Y_l^m) and radial functions (commonly spherical Bessel functions) and summing over the magnetic quantum number m yields a rotationally invariant power spectrum p. This is conceptually similar to combining p orbitals into a spherically symmetric electronic configuration. (b) Illustration showing that the SOAP power spectrum encodes three-body correlations, capturing angular relationships between triplets of atoms. Adapted from Ref. [247].

Finally, the maximum angular degree, $l_{\max}$, and the maximum radial degree, $n_{\max}$, limit the complexity of the descriptor. Increasing these values allows for more detailed descriptions but at the expense of increased computational cost. Optimisation usually involves balancing accuracy with computational feasibility.

Atomic Cluster Expansion

Atomic Cluster Expansion (ACE) is a unified formalism for representing the local atomic environment in a complete, rotationally, translationally, and permutationally invariant manner [248].

At the core of the ACE approach is a body-ordered expansion, where the total energy contribution of an atom is written as a sum over one-body, two-body, three-body (and higher) terms. Each atomic energy, E_i, is expressed as a sum over increasingly complex interactions involving the neighbours of atom i, formally:

$$E_i = V_0(Z_i) + \sum_{j_1} V_1(\mathbf{r}_{ij_1}, Z_{j_1}; Z_i) + \sum_{j_1<j_2} V_2(\mathbf{r}_{ij_1}, \mathbf{r}_{ij_2}, Z_{j_1}, Z_{j_2}; Z_i) + \cdots \tag{2.10}$$

where $\mathbf{r}_{ij}$ is the vector from atom i to a neighbour j, and Z_j denotes the atomic number. Each term V_n encodes interactions between n neighbouring atoms. In practice, computing these terms directly becomes infeasible for $n > 3$ due to combinatorial scaling.

To address this, ACE introduces a density projection trick, which simplifies the many-body expansion by expressing the local environment around atom i as a neighbour density:

$$\rho_i(\mathbf{r}) = \sum_{j\neq i} \delta(\mathbf{r} - \mathbf{r}_{ij}), \tag{2.11}$$

This density is projected onto a set of basis functions $\phi_\nu(\mathbf{r}_{ij})$, typically involving radial and angular components. These basis functions are then combined through a tensor product to construct higher-body descriptors:

$$\Phi_\nu(\mathbf{r}_{ij_1}, \ldots, \mathbf{r}_{ij_\nu}) = \prod_{t=1}^{\nu} \phi_{\nu_t}(\mathbf{r}_{ij_t}), \tag{2.12}$$

The final atomic energy can then be written as a linear combination of these descriptors:

$$E_i = \sum_{\nu} \sum_{\Phi_\nu} \theta_\nu \, \Phi_\nu(\mathbf{r}_{ij_1}, \ldots, \mathbf{r}_{ij_\nu}), \tag{2.13}$$

where θ_ν are coefficients learned during model fitting. This formulation allows ACE to retain the physical interpretability of body-order expansions while scaling linearly with the number of neighbours, rather than combinatorially.

The ACE descriptors [248] are not explicitly used or calculated at any point in this thesis. However, they form the basis of the MACE graph neural network model which is introduced in the following section and used extensively in the work presented in this thesis.

2.2.3 Regression Techniques

The final step in the development of a machine learning model is to use the training data to learn a mapping from the input descriptors to the target property. This is known as regression.

Regression models are trained to approximate an unknown functional relationship between a set of structured inputs, such as local or global chemical environment descriptors (as introduced in the previous section), and a target quantity of interest, such as energy and forces. Formally, the goal of regression is to learn a mapping

$$f : \mathbb{R}^i \to \mathbb{R}^o$$

from a high-dimensional input space, $\mathbb{R}^i$, to a scalar or vector-valued target, $\mathbb{R}^o$. The function f is trained or inferred from a dataset of N observations $(\mathbf{x}_i, y_i)_{i=1}^N$, where $\mathbf{x}_i \in \mathbb{R}^i$ is the input descriptor and $y_i \in \mathbb{R}^o$ is the corresponding target value. A concrete example of this is the prediction of the potential energy of a system, $E \in \mathbb{R}$, from a set of local environment, i.e. SOAP, descriptors, $\mathbf{x}$. ML regression models differ in how they parametrise f, how uncertainty is quantified (if at all),

and how prior knowledge is incorporated.

Many ML models, particularly those with a large number of parameters or flexible functional forms, can achieve near-perfect fits to the training data. However, such performance is often misleading: a model that merely memorises training examples (known as overfitting) but does not smoothly interpolate between training points will fail to capture the underlying relationships needed to make reliable predictions on new, unseen inputs. This ability to generalise, to make accurate predictions beyond the training set, is a fundamental criterion for evaluating the usefulness of a regression model, especially in scientific applications where the ultimate goal is often extrapolation or interpolation to novel configurations or thermodynamic conditions.

To systematically assess generalisation during model development, the available dataset is typically partitioned into three disjoint subsets: the training set, the validation set, and the test set. The training set is used to optimise the model's parameters by minimising a loss function over the observed data. The validation set is used to monitor model performance during training and to tune hyperparameters in order to prevent overfitting. Since the validation data is not directly used to update model weights, it serves as a proxy for the model's performance on unseen data. Once model selection and training are complete, final performance is evaluated on the test set. The test set is held out entirely during training and validation, ensuring that it provides an unbiased estimate of how the model will perform in practical deployment. It is essential that the test set reflects the target application domain as closely as possible; discrepancies between training and test distributions (known as distribution shift) can severely undermine performance in real-world settings.

Irrespective of the specific regression technique employed, most models possess a set of hyperparameters. These are settings that govern the model's architecture, learning dynamics, or prior assumptions. Unlike model parameters, which are learned directly from data through optimisation (e.g., weights in a neural network or the posterior function in Gaussian Process Regression), hyperparameters are

not learned during training but must instead be selected beforehand. Their values strongly influence how well the model captures underlying patterns and, crucially, how well it generalises to previously unseen data.

The optimisation of hyperparameters is a critical step in model development. Poorly chosen hyperparameters can lead to overfitting, underfitting, slow convergence, or instability during training. To identify appropriate values, a model is typically trained on a training set for many different hyperparameter configurations and evaluated on a separate validation set. Strategies for hyperparameter optimisation range from simple grid or random search to more sophisticated approaches such as Bayesian optimisation or evolutionary algorithms, depending on the complexity of the model and computational budget. A number of software packages now exist to automate efficient hyperparamter optimisation for atomistic ML [249, 250].

In this thesis, I focus on three model classes: Gaussian Process Regression (GPR), feed-forward Neural Networks (NNs), and Graph Neural Networks (GNNs). The following subsections describe each in turn. I will not describe full theoretical details of each ML architecture, but rather provide a brief overview of the key concepts and principles relevant to the work presented in this thesis.

Gaussian Process Regression

GPR is a non-parametric, probabilistic model that defines a distribution over functions, enabling both predictions and associated uncertainty estimates. It is particularly well-suited for applications where datasets are often small and noisy, and where quantifying uncertainty is crucial for downstream decision-making and model interpretability. Unlike parametric models, which learn a fixed set of parameters to approximate a function, GPR infers a posterior distribution over functions directly. This posterior combines prior assumptions about the function's behaviour (encoded in a kernel) with evidence from observed training data.

A Gaussian process is formally defined as a collection of random variables, any finite subset of which follows a joint Gaussian distribution. It is fully specified by a

mean function, $m(\mathbf{x})$, and a covariance function, or kernel, $k(\mathbf{x}, \mathbf{x}')$:

$$f(x) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')) \tag{2.14}$$

The mean function is often set to zero without loss of generality, under the assumption that any global trends can be captured by the covariance structure or through a simple linear model pre-fit to the data. The kernel plays a central role by defining the model's assumptions about function smoothness, complexity, periodicity, and the characteristic length scale of variation.

In essence, the kernel quantifies the similarity between two inputs, $\mathbf{x}$ and $\mathbf{x}'$, and determines how information is shared between them. A core assumption in GPR is that nearby inputs (according to the kernel-defined similarity) are likely to have correlated outputs. As such, training points that are similar to a test point have a stronger influence on the prediction at that point. This structure allows GPR to make informed predictions without learning an explicit mapping from inputs to outputs. For a comprehensive introduction to GPR, see Ref. [251].

In this thesis, I employ a GPR model using the SOAP kernel [230], which is specifically designed for atomistic systems. The SOAP kernel provides a similarity measure between local atomic environments that is invariant to rotation, translation, and permutation of identical atoms – symmetries that are essential for physical consistency. For two environments i and j, the SOAP kernel is defined as:

$$k(i, j) = (\mathbf{p}_i \cdot \mathbf{p}_j)^{\zeta}, \tag{2.15}$$

where $\mathbf{p}_i$ and $\mathbf{p}_j$ are the SOAP power spectra (Eq. 2.8) of environments i and j, and ζ is a positive integer controlling the sharpness of the similarity measure. The kernel yields values in the interval $[0, 1]$, where $k(i, j) = 1$ implies identical environments and lower values indicate increasing dissimilarity.

Using the SOAP kernel equips the GPR model with a physically meaningful measure of structural similarity that aligns closely with the types of variations that influence material properties. Because the SOAP descriptors are constructed directly from structural data, no manual feature engineering is required, and the resulting model naturally benefits from strong correlations between structural similarity and target properties.

GPR combined with the SOAP kernel has been successfully applied across a broad range of atomistic machine learning tasks. One of its most prominent applications is the accurate prediction of potential energy surfaces, enabling quantum-level precision in the simulation of molecular and condensed-phase systems at significantly reduced computational cost [57, 158, 252]. Beyond energy prediction, SOAP-GPR models have been used to infer a variety of materials properties, including adsorption energies on surfaces [253], nuclear magnetic resonance (NMR) chemical shifts [254], electronic densities of states [60], and even X-ray photoelectron spectroscopy (XPS) spectra [255].

Neural Networks

While GPR models have long been valued for their interpretability, uncertainty quantification, and strong performance in low-data regimes, they are increasingly limited by scalability and expressiveness when applied to large, high-dimensional datasets. In particular, the computational cost of full GPR training scales cubically with the number of training samples [247], and the fixed structure of kernel-based similarity imposes constraints on the complexity of the functions that can be learned. Practical implementations such as the Gaussian Approximation Potential framework address this issue through sparsification (replacing the full training set with a smaller set of representative environments) so that the computational cost scales with the number of these representative points rather than the entire dataset. Nonetheless, even with such improvements, GPR-based models (including those using expressive kernels like SOAP) are increasingly being outpaced by NNs and other flexible models

in many high-throughput or large-scale materials informatics tasks, due to their superior scalability and expressive capacity [256, 257].

NNs offer a highly flexible and scalable alternative. Their parametric nature allows them to learn arbitrarily complex, non-linear mappings from input features to target properties, given sufficient data. Advances in NN architectures along with optimised implementations and hardware acceleration (e.g., GPUs, TPUs), enable training on datasets several orders of magnitude larger than those tractable for GPR. Unlike kernel methods, NNs can automatically extract multi-scale and hierarchical representations of input data, without relying on hand-crafted similarity metrics [258, 259].

At their core, NN models are composed of several "layers", each of which acts as a non-linear mapping function, $f : \mathbb{R}^{\text{in}} \to \mathbb{R}^{\text{out}}$. The first layer maps inputs to hidden representations, $\mathbf{h}^{(1)}$, with subsequent layers generating further hidden representations. The action of the l'th layer is given by:

$$\mathbf{h}^{(l)} = \phi\left(\mathbf{W}^{(l)}\mathbf{h}^{(l-1)} + \mathbf{b}^{(l)}\right), \tag{2.16}$$

i.e., as an affine transformation parametrised by weight and bias matrices, $\mathbf{W}^{(l)}$ and $\mathbf{b}^{(l)}$, followed by a non-linear activation function, ϕ (e.g. ReLU, sigmoid, tanh).

The final output is task dependent. In the case of scalar regression, a simple linear readout can be used:

$$\hat{y} = \mathbf{W}^{(L+1)}\mathbf{h}^{(L)} \tag{2.17}$$

To train such a model, one must first define a loss function, typically the mean squared error for regression tasks:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N}\sum_{i=1}^{N}\left(f(\mathbf{x}_i;\boldsymbol{\theta}) - y_i\right)^2, \tag{2.18}$$

where $f(\mathbf{x}_i;\boldsymbol{\theta})$ denotes the neural network prediction for input $\mathbf{x}_i$ given model parameters $\boldsymbol{\theta}$, and y_i is the corresponding true target value.

Gradients with respect to this loss function can then be analytically derived for each model parameter, allowing for the use of (typically first-order) optimisers such as stochastic gradient descent or Adam to iteratively update model parameters, minimising the expected loss over the training dataset [260].

NNs excel in high-dimensional settings, are able to learn complex hierarchical features, and are provably universal function approximators given sufficient parameters [261]. However, they are often treated as black-box models, and lack the interpretability and uncertainty estimates of probabilistic methods such as GPR [262]. Moreover, their performance is often highly sensitive to architecture choice and hyperparameter tuning.

Graph Neural Networks

GNNs generalise NNs to operate on graph-structured data, making them particularly well-suited for molecular and atomistic systems. In such systems, entities like atoms and their relationships (whether through covalent bonds or spatial proximity) are naturally represented as nodes and edges in a graph. Formally, a molecular graph can be defined as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of nodes (atoms), and $\mathcal{E}$ is the set of edges (bonds or neighbour interactions). Each node $v \in \mathcal{V}$ is associated with a feature vector $\mathbf{h}_v$, and each edge $(u, v) \in \mathcal{E}$ may carry edge features $\mathbf{e}_{uv}$ describing properties such as bond type or interatomic distance.

GNNs offer a powerful alternative to conventional machine learning approaches that rely on hand-crafted molecular descriptors as input to models such as GPR or NNs described above. Rather than depending on pre-defined representations, GNNs directly take as input the full molecular graph. Through training, they learn task-relevant representations tailored to specific predictive goals [263]. Their ability to operate directly on chemically meaningful structures enables them to capture the intricacies of atomic-scale interactions and global structural phenomena.

Given their versatility and effectiveness, a wide variety of GNN architectures have been proposed [167, 256, 264–268]. Most GNNs in chemistry and materials science

can be unified under the Message Passing Neural Network (MPNN) framework [266], which I briefly outline below.

Unlike regular data structures such as grids or sequences, graphs are inherently irregular: they lack fixed dimensionality, ordering, or size. MPNNs address this by learning node-level representations (embeddings) through a process known as *message passing*, where each node updates its representation by aggregating information from its local neighbourhood.

A widely used formulation for a single GNN layer updates a node's embedding $\mathbf{h}_v^{(l)}$ at layer l based on its previous state $\mathbf{h}_v^{(l-1)}$ and the embeddings of its neighbouring nodes $\mathcal{N}(v)$:

$$\mathbf{h}_v^{(l)} = \sigma\left(\mathbf{W}^{(l)} \sum_{u \in \mathcal{N}(v) \cup \{v\}} \mathbf{h}_u^{(l-1)}\right), \tag{2.19}$$

where $\sigma(\cdot)$ is a non-linear activation function, $\mathbf{W}^{(l)}$ is a trainable weight matrix at layer l, and $\deg(v)$ denotes the degree of node v, i.e., the number of neighbouring atoms within a specified cutoff.

This iterative process allows the model to encode both local atomic environments and long-range dependencies, with node embeddings gradually capturing broader chemical context as they get updated layer by layer. Consequently, GNNs learn hierarchical, many-body representations that are particularly effective for modelling complex molecular and material structures.

Owing to their inherent respect for symmetries in chemical graphs – such as permutation invariance of atoms – and their capacity to model intricate interatomic dependencies, GNNs have emerged as the state-of-the-art for numerous applications. These include materials property prediction [269–271], drug discovery [272], phase classification [115], and the acceleration of *ab initio* molecular dynamics [167, 273].

Recent advances have further enriched GNNs by incorporating equivariant representations of geometric information, such as atomic positions, distances, and angular correlations, allowing for more physically informed and data-efficient learning. In this thesis, I employ the MACE architecture [256], and compare it with two other

leading models: Neural Equivariant Interatomic Potential (NequIP) [167] and Polarizable atom interaction Neural Network (PaiNN) [267]. While all three are based on message passing, they differ in how they encode and process geometric data, as well as in their respective trade-offs between accuracy, computational efficiency, and architectural complexity. The following paragraphs provide a concise introduction to each of these models.

PaiNN is an equivariant message passing neural network that models atomic interactions by explicitly separating scalar and vector features. Each atom is represented by a collection of (i) scalar and (ii) vectorial features, which together evolve through equivariant message updates and interaction blocks. Messages are computed using learned radial functions and are applied via equivariant convolutions, ensuring that the model respects the symmetries of 3D space. The vector components transform equivariantly under rotations, and their updates are governed by operations that couple scalar and vector features while preserving equivariance. The architecture avoids spherical harmonics and uses relatively lightweight vector operations, making it computationally efficient (around 10× faster than MACE [274]).

NequIP uses features that transform according to irreducible representations of the 3D rotation group, SO(3), via spherical harmonics. Each atomic feature is a set of tensors with specific rotation behaviours (e.g., scalars, vectors, higher-order tensors). Messages are passed and updated using tensor products that maintain equivariance at each layer. The core operation involves:

$$\mathbf{h}_i^{(l+1)} = \sum_{j \in \mathcal{N}(i)} \sum_{l_1, l_2} \mathbf{C}^{l}_{l_1, l_2} \left(\mathbf{h}_i^{(l)} \otimes \mathbf{h}_j^{(l)} \right), \tag{2.20}$$

where $\mathbf{C}^{l}_{l_1, l_2}$ are Clebsch–Gordan coefficients enforcing equivariant coupling between tensorial features.

MACE extends the message passing framework by introducing a hierarchical, many-body message construction mechanism that scales beyond pairwise interac-

tions. Each message is built from a sum of learnable functions, known as ACE descriptors (Section 2.2.2), acting on increasingly complex combinations of neighbour features, capturing $(v+1)$-body correlations via symmetric tensor products. To maintain computational efficiency, MACE leverages a tensor product factorisation, avoiding the exponential scaling typical in high-order expansions. Edge embeddings use a learnable radial basis and spherical harmonics, while node features are aggregated via Clebsch–Gordan-weighted sums that preserve rotational equivariance. This results in a rich, equivariant representation that fully encodes geometric information up to the desired correlation order. MACE has been shown to match or exceed NequIP's accuracy (see later Section 4.4.2) with improved training stability and lower computational overhead in some settings [256].

2.3 Density Functional Theory

The reference datasets used to train machine-learned interatomic potentials in this thesis are labelled using electronic structure methods, and in particular density functional theory (DFT). A concise overview of the DFT framework is therefore provided here, both to contextualise its role as the source of ground-truth data and to clarify its inherent approximations and limitations, particularly with regards to the systems considered in this thesis.

2.3.1 Conceptual Framework

The fundamental task of electronic structure theory is to solve the many-electron Schrödinger equation for a system of nuclei (at locations $\vec{R_i^n}$) and electrons (at locations $\vec{r_j^e}$):

$$\begin{aligned}\hat{H}\Psi &= E\Psi \\ &= \hat{T}_n\Psi + \hat{T}_e\Psi + \hat{V}_{nn}\Psi + \hat{V}_{ne}\Psi + \hat{V}_{ee}\Psi\end{aligned}$$

where $\Psi = \Psi(\{\vec{R}_i^n\}_{i=1}^N, \{\vec{r_j^e}\}_{j=1}^M)$ for a system of N nuclei and M electrons.

In the context of this thesis, we are interested in finding the ground-state energy for a given system. In principle, we could obtain this by solving the above equation, either exactly or numerically, but this is intractable for all but the simplest of systems due to the non-linear electron–electron interactions, and the exponential scaling of the wavefunction with particle count respectively [38].

DFT circumvents this difficulty by reformulating the problem in terms of the electron density, $\rho(\vec{r})$, a three-dimensional scalar field. The Hohenberg–Kohn theorems (1964) [40] establish that:

(i) the ground-state energy of an interacting electron system is a unique functional of $\rho(\vec{r})$, and

(ii) the exact ground-state density minimises this functional.

These results shift the focus from the many-electron wavefunction to the more tractable density.

In practice, the Kohn–Sham formalism (1965) [41] is used: the interacting electron system is mapped onto a fictitious non-interacting system that reproduces the same ground-state density. The total energy is written as

$$E[\rho] = T_s[\rho] + E_{\text{ext}}[\rho] + E_H[\rho] + E_{xc}[\rho], \tag{2.21}$$

where T_s is the kinetic energy of the non-interacting electrons, E_{ext} is the energy due to external potentials (e.g. nuclei), E_H is the classical Hartree energy of Coulomb repulsion, and E_{xc} is the exchange–correlation (XC) energy functional, which accounts for all many-body effects missing from the other terms.

Calculating the total energy for a given density (and exchange–correlation functional) requires a straightforward application of the above equation. DFT is (relatively) expensive because:

(i) diagonalising the Hamiltonian matrix extracted from the Kohn–Sham equations scales cubically with the number of electrons, $\mathcal{O}(N^3)$ [38];

(ii) finding the density that minimises the total energy is a non-trivial problem that requires iteratively updating estimates for the density until convergence is reached.

2.3.2 Approximate Functionals

The exact form of $E_{xc}[\rho]$ is unknown, and all practical implementations of DFT rely on approximations.

Common families of XC functionals include:

- **Local density approximation (LDA)**: assumes E_{xc} at a point depends only on the local density, using the uniform electron gas as reference [41, 275].
- **Generalised gradient approximation (GGA)**: extends LDA by including dependence on the density gradient, e.g. PBE [142] and BLYP functionals [276, 277].
- **Meta-GGA and hybrid functionals**: introduce dependence on kinetic energy density or incorporate exact Hartree–Fock exchange, e.g. SCAN [144]

Each class balances accuracy, transferability, and computational cost. The following section summarises the key limitations of standard DFT approximations relevant to this work.

2.3.3 Known Limitations

For the systems considered in this thesis, important limitations of standard DFT approximations include:

- **Water:** GGAs tend to overbind hydrogen bonds, leading to overstructured radial distribution functions and underestimated diffusion coefficients [134, 143]. Inclusion of dispersion corrections (e.g., DFT-D3 [145, 146]) or use of hybrid/meta-GGA functionals improves agreement with experiment [141, 147].

- **Silica (SiO_2):** the balance between covalent and ionic character is sensitive to the functional. Standard GGAs capture qualitative trends but may misrepresent defect energetics and high-pressure phases [278].

- **Band gaps and excited states:** ground-state DFT systematically underestimates band gaps due to self-interaction error and lack of derivative discontinuity [279, 280]. While not the primary concern for MLIP training, this limits transferability to electronic properties.

In summary, DFT provides an efficient and widely adopted route to high-quality atomistic reference data. However, its approximations imprint characteristic biases into the datasets, and therefore into MLIPs trained upon them. Recognising these limitations is essential when interpreting ML-driven simulations.

2.4 Molecular Dynamics Simulations

Molecular dynamics (MD) is a powerful computational technique used to simulate the time evolution of many-body systems at the atomic scale. MD models atoms as classical particles whose trajectories evolve according to Newton's equations of motion,

$$\mathbf{F}_i = m_i \mathbf{a}_i = -\nabla_i U(\mathbf{r}_1, \ldots, \mathbf{r}_N), \tag{2.22}$$

where $U(\mathbf{r}_1, \ldots, \mathbf{r}_N)$ is the PES that defines the total potential energy as a function of the atomic coordinates. Starting from an initial configuration of positions and velocities, the equations of motion are integrated over discrete time steps (typically on the order of femtoseconds), generating correlated trajectories that describe the dynamical evolution of the system.

The accuracy of MD critically depends on the fidelity of the PES. Empirical force fields are efficient and interpretable, but often limited in accuracy and transferability [19]. Quantum mechanical approaches such as DFT provide higher accuracy by explicitly treating electronic structure [37], but scale poorly with system size [38].

ML models, particularly GNN-based potentials, aim to bridge this gap by learning high-fidelity PES representations from quantum reference data [49].

2.4.1 Statistical Foundations and Thermodynamic Ensembles

The theoretical foundation of MD lies in statistical mechanics, which provides the connection between microscopic particle motion and macroscopic thermodynamic observables. Instead of describing a system by a single configuration, statistical mechanics considers an ensemble: a large collection of microstates consistent with specified macroscopic constraints. Each ensemble corresponds to particular physical conditions and defines a probability distribution over accessible phase space.

The three most relevant ensembles for MD are:

- **Microcanonical Ensemble (NVE):** Number of particles (N), volume (V), and total energy (E) are constant. The system is isolated, and all microstates with total energy E are equally probable:

$$\rho_{\mathrm{NVE}}(p, q) \propto \delta(E - H(p, q)),$$

  where $H(p, q)$ is the Hamiltonian.

- **Canonical Ensemble (NVT):** N and V are fixed, but the system exchanges energy with a thermal reservoir at temperature T. The probability of microstate i with energy E_i is given by the Boltzmann distribution:

$$P_i = \frac{e^{-E_i/k_B T}}{Z}, \quad Z = \sum_i e^{-E_i/k_B T},$$

  where k_B is Boltzmann's constant and Z is the partition function.

- **Isothermal–Isobaric Ensemble (NPT):** N, P, and T are constant. The

system can exchange both energy and volume with its surroundings, appropriate for condensed-phase or experimental conditions.

In all ensembles, the average of an observable A is given by

$$\langle A \rangle = \sum_i P_i A_i, \tag{2.23}$$

which corresponds to macroscopic quantities such as energy, pressure, or enthalpy in the thermodynamic limit.

2.4.2 Ensemble Control: Thermostats and Barostats

A basic MD simulation inherently samples the microcanonical (NVE) ensemble, as total energy is conserved. However, most physical experiments occur at constant temperature or pressure. To reproduce these conditions, MD employs thermostats and barostats, which modify the dynamics to sample canonical (NVT) and isothermal–isobaric (NPT) ensembles, respectively.

Thermostats such as Nosé–Hoover [281] and Berendsen [282] regulate system temperature by rescaling or dynamically adjusting particle velocities. Stochastic approaches like the Langevin thermostat [283] introduce random forces to mimic collisions with a heat bath, ensuring proper Boltzmann sampling. Barostats such as Parrinello–Rahman [284] or Martyna–Tuckerman–Klein [285, 286] adjust the simulation cell's volume or shape to maintain target pressure, introducing extended degrees of freedom in the equations of motion.

2.4.3 Ergodicity and Ensemble Sampling

The equivalence between MD trajectories and statistical ensembles relies on the ergodic hypothesis [287]: over sufficiently long times, the trajectory explores all microstates consistent with the ensemble constraints. When ergodicity holds, time

averages over a trajectory equal ensemble averages,

$$\langle A \rangle_{\text{ensemble}} = \lim_{\tau \to \infty} \frac{1}{\tau} \int_0^\tau A(t)\, dt. \tag{2.24}$$

For complex systems with slow conformational dynamics or high energy barriers, ergodicity may be limited. In such cases, enhanced sampling methods such as replica exchange, umbrella sampling, or metadynamics are employed to improve statistical convergence.

2.4.4 Implementation

All MD simulations presented in this chapter are performed using the `Large-scale Atomic/Molecular Massively Parallel Simulator` (LAMMPS) software package [288, 289], with appropriate thermostats and barostats to enforce desired ensemble conditions. The resulting trajectories enable direct computation of structural, thermodynamic, and dynamic properties, providing a computational bridge between statistical mechanics and experimentally accessible thermodynamics.

Chapter 3

Classification of Amorphous Ices

3.1 Acknowledgements

The work presented in this chapter has been published in The Journal of Chemical Physics [244]. Portions of the text and several figures have been reused; where appropriate, figures are labelled as "Adapted". I am grateful to John Gardner for his assistance with implementing the Steinhardt parameter calculations.

3.2 Introduction

Water is one of the most familiar substances in everyday life, yet its properties continue to challenge scientific understanding due to a wealth of anomalous behaviours. This complexity is vividly illustrated in its phase diagram, which is arguably the most intricate among all pure substances [290]. Since Bridgman's early discovery of ice polymorphism in 1912 [291], experimental investigations have revealed over 20 distinct crystalline phases of ice [290], including four that have been identified within the past five years alone [292–295].

In addition to its crystalline diversity, water also exhibits polyamorphism (i.e. the existence of multiple amorphous forms) under deeply supercooled conditions. This phenomenon is considered indicative, though not conclusive, of the presence of a hypothesised liquid–liquid critical point (LLCP) [296]. The two main categories of amorphous ices, low-density amorphous (LDA) and high-density amorphous (HDA), comprise several structural variants that differ in their formation pathways, densities, and local configurations [297, 298]. LDA is thought to be the most prevalent form of ice in the universe, commonly forming as water vapour condenses onto interstellar dust grains [299]. It can also be synthesised at ambient conditions via rapid quenching of liquid water and its family comprises LDA-I and LDA-II, obtained via

heating HDA and vHDA respectively [300], as well as a more-ordered low-density phase obtained upon heating ice VIII [301]. HDA, by contrast, is typically produced through compression-induced transitions from LDA or hexagonal ice Ih, involving a substantial increase in density [302]. Further subtypes within the HDA family include expanded HDA (eHDA) and vHDA [303, 304]. In simulations, however, the structural differences between these subtypes are subtle, and thus they are grouped under the broader labels of LDA and HDA.

The structural distinctions between LDA and HDA are both significant and nuanced. LDA exhibits high tetrahedral coordination, with clearly separated first and second hydration shells. HDA, on the other hand, displays greater local disorder, with molecules occupying the interstitial space between these shells in motifs reminiscent of ice IV [305–307]. The hydrogen-bond network topologies of LDA and HDA are also markedly different [308], even though both forms exhibit a similar suppression of long-range density fluctuations [79, 308].

Polyamorphism in water has attracted significant attention, in part due to its implications for the existence of distinct liquid states. LDA and HDA are generally regarded as the glassy counterparts of the low-density liquid (LDL) and high-density liquid (HDL) phases, respectively [309]. Understanding the nature of the LDA–HDA separation is fundamentally important: if it results from a first-order phase transition, the corresponding transition line could extend into the so-called "no-man's land" and terminate at a liquid–liquid critical point (LLCP). While an LLCP has been established in several computational models of water [105, 310], definitive experimental verification remains elusive. Nevertheless, a growing body of experimental evidence lends strong support to this hypothesis [115, 311]. Both experimental [312] and simulation studies [77, 308, 313–315] strongly indicate that the LDA–HDA transformation is indeed first-order. Additionally, signatures of metastable critical behaviour have been observed in the long-range structural properties of both LDA and HDA [79, 308, 316].

A recent development in this field is the identification of a third amorphous ice phase with a density intermediate between that of LDA and HDA. This new phase, named medium-density amorphous ice (MDA), is created by ball milling hexagonal ice at low temperatures [317]. Simulations suggest this process mimics the effects of random mechanical shear on the ice lattice. Several interpretations have been proposed, including the idea that MDA could be the true glassy state of liquid water. If so, this would challenge the prevailing LLCP hypothesis. Within the framework of the two-state model, MDA would need to exhibit a glass transition temperature above the LLCP and might represent a metastable phase poised to separate into LDA and HDA as the system approaches this critical point [317]. However, recent work [318] by Eltareb *et al.* suggests that MDA is not a unique phase but rather part of a continuum of intermediate amorphous ices that structurally resemble a transitional state between LDA and HDA.

In this chapter, I analyse the local structures of LDA, HDA, and MDA using Steinhardt bond-orientational order (BOO) parameters [234]. BOO parameters serve as invariants under the SO(3) group, and are intimately connected to the symmetry properties of crystalline environments, thus providing a systematic framework for assessing local structural order. This approach enables a detailed and quantitative comparison of the structural characteristics of these amorphous phases, contributing to a deeper understanding of the relationship between local structure and phase behaviour in amorphous ice.

3.3 Methods

3.3.1 Database

The database used in this chapter was compiled from two sources. The first is the work of Martelli *et al.* [319], which includes an extensive collection of LDA, HDA ices, and liquid water structures. The second source is the study by Rosu-Finsen *et al.* [317], which provides a series of MDA ice structures. Brief descriptions

of the simulation protocols used to generate these structures are provided below. Where MD simulations were employed, water molecules were modelled using the TIP4P/2005 interaction potential [120].

LDA and HDA

The majority of the database was sourced from the work of Martelli *et al.* [319], which comprises numerous LDA and HDA configurations, each consisting of 8192 water molecules. LDA structures were obtained by quenching equilibrated liquid water from $T = 300\,\mathrm{K}$ to $T = 80\,\mathrm{K}$ at a cooling rate of $1\,\mathrm{K/ns}$. HDA structures were generated by isothermal compression of LDA and ice Ih at four temperatures: 80 K, 100 K, 120 K, and 140 K. To sample configurations across a range of pressures, isothermal decompression of HDA was simulated from 2.0 GPa down to 10^{-4} GPa. Additionally, liquid water configurations were extracted from MD simulations performed at 300 K and 10^{-4} GPa. For a detailed account of the simulation methodology, the reader is referred to Ref. [319].

MDA

Complementing the LDA and HDA data, a set of MDA configurations was taken from Rosu-Finsen *et al.* [317]. MDA was generated through repeated shearing of randomly selected layers within a simulation box of ice Ih containing 2880 water molecules. After each shear step, local geometry optimisation was performed. This shearing process continued until key structural metrics converged, indicating full amorphisation. In the 2880-molecule system, full amorphisation was typically achieved after 100 shear steps (see Fig. 3.1).

Following the amorphisation process, the resulting structures underwent *NPT* MD simulations at 125 K and 0 atm using a 2 fs time step. MDA configurations were sampled from 2 ns production runs. Further details can be found in Ref. [317].

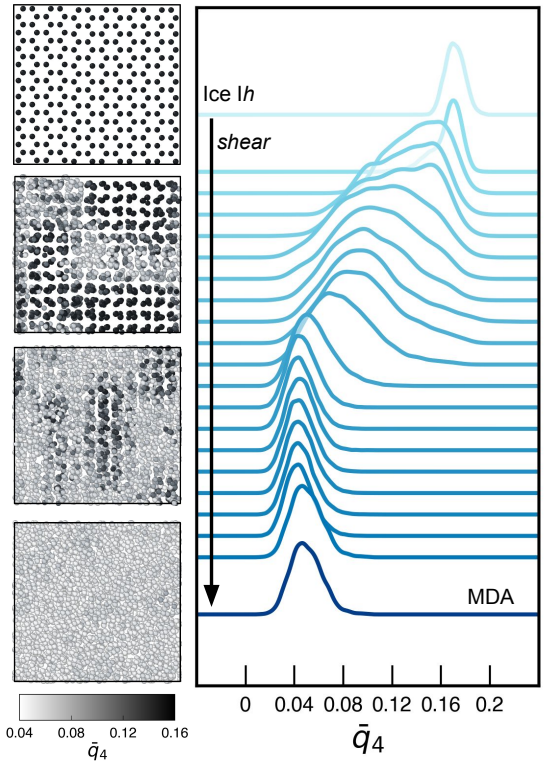

Figure 3.1: Steinhardt $\bar{q}_4$ values for ice I*h* and their evolution during repeated shearing to produce MDA. Left: Structural snapshots taken at different stages of the shearing process, with atoms coloured by their $\bar{q}_4$ value. Right: Kernel density estimates (KDE) of $\bar{q}_4$ values in a 2880-molecule system at selected points along the shearing trajectory (data from Ref. [317]). Adapted from Ref. [244].

Dataset Compilation and Sampling

The final compiled dataset, comprising LDA, HDA, and MDA structures, was split into training, validation, and test partitions containing 1285, 20, and 1245 structures, respectively. From each of these partitions, local atomic environments of oxygen atoms were randomly sampled, generating training, validation and test sets of environments containing an equal proportion of HDA, LDA, and MDA environments. In total, this yielded 24,000 training environments (8,000 from each state) and 7,500 (2,500 from each state) environments each for validation and testing. This stratified sampling strategy was designed to mitigate data imbalance and ensure uniform representation of each structural class.

An additional dataset was prepared comprising only HDA, LDA, and liquid

water environments. Following a similar sampling protocol, this dataset consisted of 32,000 training, 10,000 validation, and 10,000 test environments.

3.3.2 Classification

Neural Network Architecture

NNs are a class of parametric models capable of approximating complex, non-linear functions, making them well-suited for classification tasks. In this context, a neural network is trained to assign input data to one of K predefined categories by learning a mapping from input features to output class probabilities. Training is performed on a labeled dataset $\{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, where $\mathbf{x}_i \in \mathbb{R}^d$ is a feature vector describing the local environment of atom i, and $y_i \in \{1, \ldots, K\}$ is the corresponding class label.

The architecture adopted here is a feed-forward neural network, discussed in Section 2.2.3, with the rectified linear unit (ReLU) activation function [320]:

$$\phi(z) = \max(0, z) \tag{3.1}$$

The final output layer produces a vector of unnormalised scores (logits) $\mathbf{z} \in \mathbb{R}^K$, one for each class. These are transformed via the softmax function to produce a probability distribution over the classes:

$$\hat{p}_k = \frac{\exp(z_k)}{\sum_{j=1}^{K} \exp(z_j)}, \quad k = 1, \ldots, K, \tag{3.2}$$

where $\hat{p}_k$ represents the model's predicted probability that the input belongs to class k.

Model parameters are learned by minimising the categorical cross-entropy loss. For a single training example with true class label y and predicted class probabilities $\hat{\mathbf{p}}$, the loss is defined as:

$$\mathcal{L}(\hat{\mathbf{p}}, y) = -\log(\hat{p}_y), \tag{3.3}$$

or, equivalently, for one-hot encoded labels $\mathbf{y} \in \{0, 1\}^K$:

$$\mathcal{L}(\hat{\mathbf{p}}, \mathbf{y}) = -\sum_{k=1}^{K} y_k \log(\hat{p}_k). \tag{3.4}$$

This loss function penalises incorrect predictions by comparing the predicted distribution to the true class label. The total loss over the dataset is minimised using stochastic gradient descent with backpropagation. In this work, the Adam optimiser [260] was employed for parameter updates, as implemented in `PyTorch` [321].

In the current study, the classification task involves assigning each atomic environment to one of three amorphous ice structures: HDA, LDA, or MDA ice. The feature vectors $\{\mathbf{x}_i\} \in \mathbb{R}^{30}$ consist of the following Steinhardt BOO parameters: $q_l(i)$ and $\bar{q}_l(i)$ with $l \in [3, 12]$, $w_l(i)$ and $\bar{w}_l(i)$ with l even and $l \in [4, 12]$ (see Section 2.2.2). The labels $\{y_i\}$ are encoded as one-hot vectors indicating the corresponding structural class.

A schematic of the model architecture mapping BOO parameters to class probabilities is provided in Fig. 3.2.

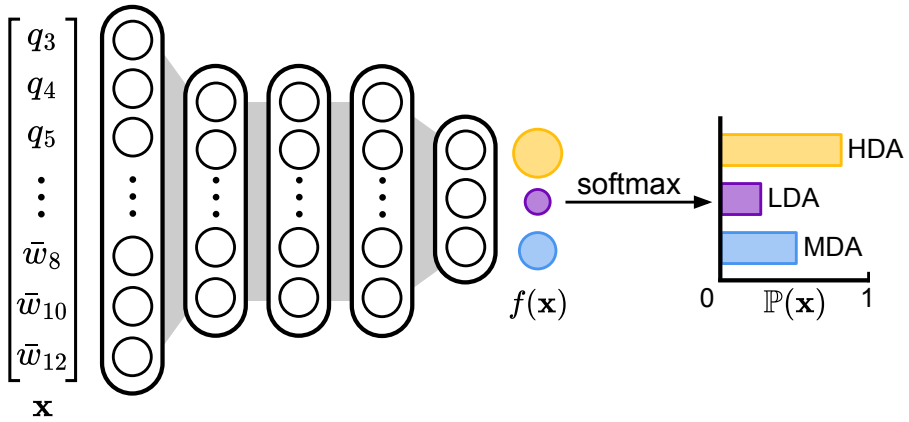

Figure 3.2: Schematic representation of the neural network employed in this chapter. The input layer ($\mathbf{x}$) consists of 30 nodes, each representing one of the structural BOO parameters. The output layer has three nodes for the HDA, LDA, and MDA phases respectively. A softmax activation function is used on the output layer to convert the raw outputs, $f(\mathbf{x})$, to predicted probabilities, $\mathbb{P}(\mathbf{x})$. Adapted from Ref [322].

Performance Metrics

To evaluate the classification performance of the trained neural network, I focus on metrics that account for class imbalances and provide a nuanced view of model behaviour across different structure types.

Recall A fundamental metric in multi-class classification is *recall*, which measures the ability of the model to correctly identify all instances of a given class. For a given class k, recall is defined as:

$$\text{Recall}_k = \frac{\text{TP}_k}{\text{TP}_k + \text{FN}_k}, \tag{3.5}$$

where TP_k is the number of true positives (i.e., correctly predicted instances of class k), and FN_k is the number of false negatives (i.e., instances of class k that were incorrectly classified as another class). High recall indicates that most instances of a class are being correctly identified by the model.

Recall is particularly valuable in this context because it provides class-specific insight into model performance, allowing us to identify which structural environments (e.g., LDA, HDA, MDA) are more prone to misclassification.

Balanced Accuracy While recall offers a per-class evaluation, a single scalar metric is often desirable for comparing overall model performance. To this end, I use the *balanced accuracy score* (BAS), which is the average recall across all classes:

$$\text{BAS} = \frac{1}{K} \sum_{k=1}^{K} \text{Recall}_k = \frac{1}{K} \sum_{k=1}^{K} \frac{\text{TP}_k}{\text{TP}_k + \text{FN}_k}, \tag{3.6}$$

where K is the number of classes.

Balanced accuracy is especially important when class distributions are uneven, as it ensures each class contributes equally to the final score. This mitigates the risk of overestimating model performance due to dominance by a majority class. A balanced accuracy of 1.0 indicates perfect classification, while a score of 0.5 corresponds to random guessing in a three-class problem with equal weights.

Confusion Matrix To complement these quantitative metrics, I also examine *confusion matrices*, which offer a detailed visual summary of classification outcomes. Each entry in a confusion matrix represents the proportion of samples from a true class (rows) that were predicted as a given class (columns). The diagonal entries correspond to correct predictions (true positives), while off-diagonal elements represent misclassifications.

Confusion matrices allow for easy identification of systematic errors, such as one class being consistently confused with another, and help interpret how well the model separates structurally similar classes like LDA and MDA. This visualisation is especially useful when assessing multi-class problems, where summary statistics alone may obscure underlying trends.

3.4 Results

3.4.1 Hyperparameter Optimisation

The first step of the analysis was to optimise the NN to ensure accurate predictions. To do so, the categorical cross-entropy loss (Eq. 3.4) was minimised over the training set, before quantifying model accuracy on the validation set. Four key hyperparameters were optimised: the number of hidden layers, the number of neurons per layer, the learning rate, and the weight decay. These control, respectively, the model's capacity to learn complex patterns, the expressiveness of each layer, the rate at which the model's parameters are updated, and the extent of regularisation applied to prevent overfitting. These were sampled from the ranges given in Table 3.1, with final values being selected via Bayesian optimisation as implemented in `Optuna` [250].

Interestingly, the optimisation revealed that the model is very insensitive to hyperparameter selection. Among 110 optimisation iterations, 97% of models achieved test set accuracy within 3% of the best model's performance. This shows that the model and task are very robust to hyperparameter selection and large-scale fine-tuning is not required in this case.

Table 3.1: Optimised NN hyperparameters for classification of amorphous ices.

Hyperparameter	Range	Optimised Value
# hidden layers	[1, 5]	3
# neurons per layer	[8, 256]	82
Weight decay	$[10^{-8}, 0.1]$	1.36×10^{-4}
Learning rate	$[10^{-5}, 0.1]$	6.36×10^{-3}

The final model comprises 3 hidden layers and 82 neurons per layer (the optimal depth and width), trained with a learning rate of 6.36×10^{-3} and a weight decay of 1.36×10^{-4}.

3.4.2 MDA Classification

With the model optimised, I proceeded to the primary classification task: identifying the structural phase (HDA, LDA, or MDA) from which a given atomic environment had been sampled. As a control, a separate model was also trained to classify environments among HDA, LDA, and liquid water. In both cases, the models were trained on input–label pairs $(\mathbf{x}_i, y_i)$, where $\{\mathbf{x}_i\}$ are the 30-dimensional vectors of BOO parameters describing the local environment of atom i (Section 2.2.2), and $\{y_i\}$ are one-hot encoded vectors representing the corresponding class labels. Additional details regarding the dataset construction and training procedure are provided in Section 3.3.1 and Section 3.3.2.

Figure 3.3 presents the performance of this NN trained to classify local atomic environments in amorphous ices based on Steinhardt bond order parameters. In both cases, the confusion matrices (panel a) and prediction confidence distributions (panel b) are used to assess the classification accuracy and confidence of the model when distinguishing between different structural phases: HDA, LDA, and either MDA ice or high-temperature liquid water.

In the first case (Fig. 3.3a), where the NN is trained on HDA, LDA, and MDA, the model performs exceptionally well in identifying HDA-like environments, achiev-

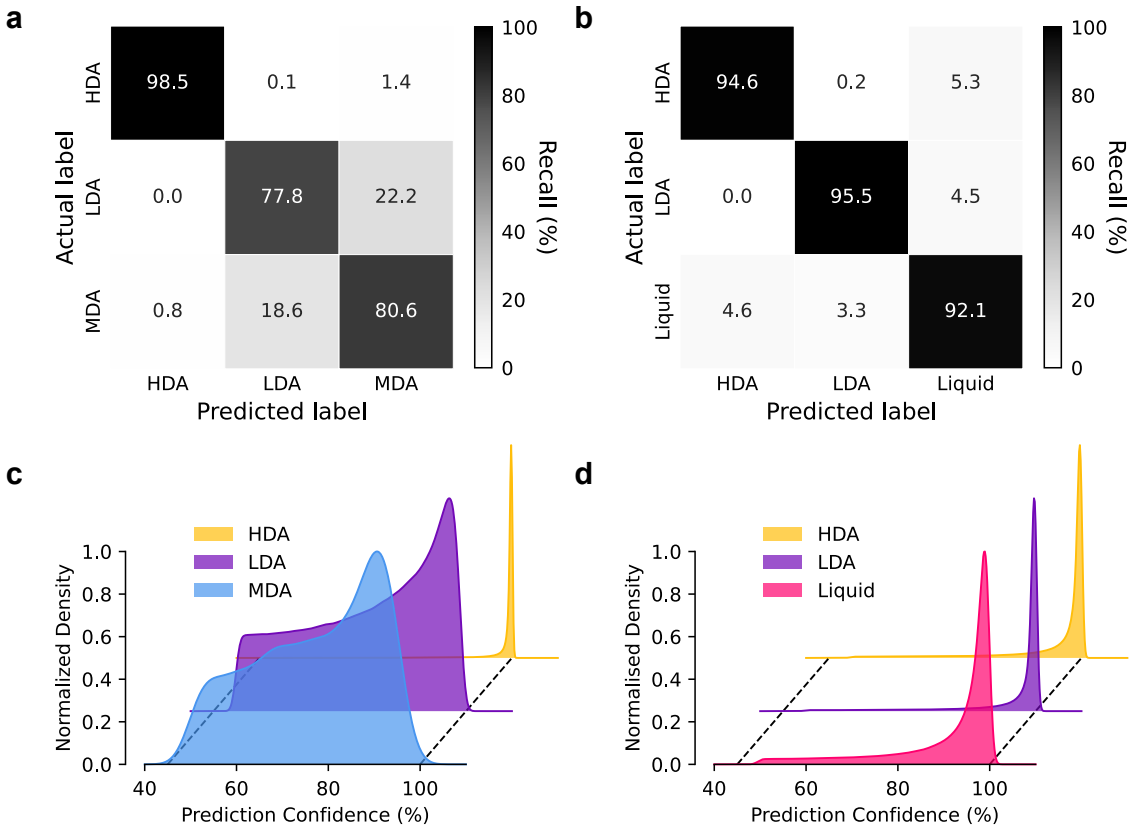

Figure 3.3: Classification performance and prediction confidence for local water environments. (a) Confusion matrix for a model trained to distinguish between HDA, LDA, and MDA environments. Values represent recall percentages per class. (b) Confusion matrix for a separate model trained to classify HDA, LDA, and liquid water environments. (c) Distribution of prediction confidence scores for HDA, LDA, and MDA environments, showing overlaps particularly between LDA and MDA. (d) Prediction confidence distribution for HDA, LDA, and liquid water environments, with more distinct class separation compared to (c). Adapted from Ref. [244].

ing a recall of 98.5%. However, the classification between LDA and MDA proves significantly more challenging, with notable misclassification rates of 22.2% (LDA predicted as MDA) and 18.6% (MDA predicted as LDA). This indicates a high degree of structural similarity between LDA and MDA that the model struggles to resolve using local bond order descriptors. In contrast, Fig. 3.3b, which replaces MDA with high-temperature liquid water, shows greatly improved performance across all three classes. Misclassification rates fall below 5.3% for all pairwise comparisons, and the model achieves over 92% recall for each class, suggesting that liquid water is more easily distinguishable from LDA and HDA based on local structure.

This distinction is further reinforced by the prediction confidence distributions

(panels c and d). For the HDA class, the model exhibits consistently high confidence in both classification tasks, reflected by a narrow peak near 100%. However, in the MDA-inclusive model, the prediction confidence for LDA and MDA is broadly distributed, indicating the model's uncertainty in distinguishing between these two phases. In contrast, when the model is trained with LDA and liquid water (Fig. 3.3d), the confidence distributions for all three phases are sharply peaked, particularly for LDA, which shows a significant gain in classification certainty. This demonstrates that the neural network can effectively learn and identify LDA environments, but its performance degrades when MDA is introduced, implying that LDA and MDA share significant structural overlap in local structure.

This structural ambiguity can be directly observed by examining the distributions of three Steinhardt BOO parameters - $\bar{q}_4$, $\bar{q}_6$, and $\bar{q}_8$ - across the three phases (Fig. 3.4). In the absence of MDA, the distributions of HDA and LDA are largely distinct, with minimal overlap. This clear separation simplifies the classification task for the neural network, as local environments can be effectively distinguished based on their BOO values. However, the inclusion of MDA introduces significant overlap, particularly between the LDA and MDA distributions across all three parameters. This suggests that MDA shares substantial local structural similarity with LDA, blurring the boundaries between the two in the feature space. Some separation is observed in the $\bar{q}_4$ distribution, reflecting partial differences in local tetrahedral ordering. This distinction may help the neural network retain some capacity to differentiate between the two phases, but overlap in $\bar{q}_6$ and $\bar{q}_8$ remains pronounced. As a result, the model finds it challenging to confidently distinguish between LDA and MDA, leading to increased misclassification rates. These findings suggest that, at least from the perspective of local structural descriptors, MDA does not exhibit clearly distinct features that would support its interpretation as a separate form of glassy water. Rather, its strong structural resemblance to LDA implies that MDA may represent a variation or distorted form of LDA, rather than a fundamentally

new amorphous phase.

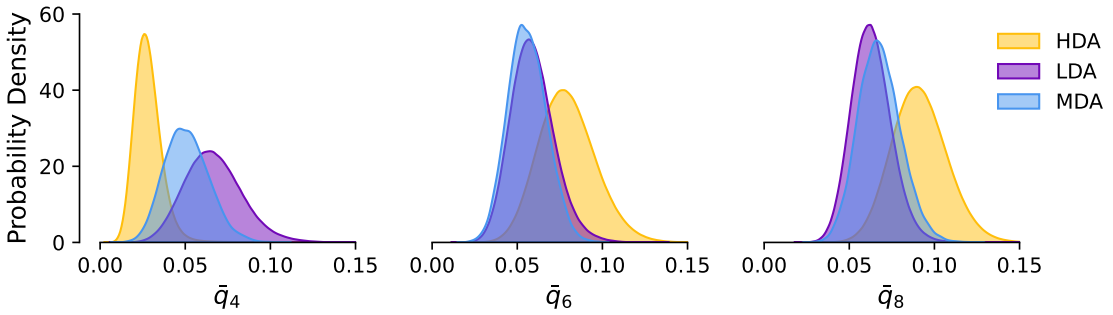

Figure 3.4: Local structure analysis of amorphous ices. KDE plots of BOO parameters $\bar{q}_4$, $\bar{q}_6$, and $\bar{q}_8$ for the test set comprising HDA, LDA and MDA structures. Adapted from Ref. [244].

3.4.3 Benchmarking Studies

I benchmarked the performance of the NN model against six well-established classification algorithms using implementations provided by `scikit-learn` [323].

As shown in Table 3.2, all six baseline models perform comparably to the NN in identifying HDA environments, with recall rates exceeding 95%. However, classification performance drops for the more challenging LDA and MDA categories. In these cases, recall scores across all models remain below 81%.

Table 3.2: Recall and BAS for various classification methods.

	Recall (%)			
	HDA	**LDA**	**MDA**	**BAS (%)**
Neural Network (NN)	**98.5**	**77.8**	**80.6**	**85.6**
k-Nearest Neighbors (k-NN)	98.0	72.4	70.6	80.3
Logistic Regression	98.0	76.7	80.4	85.0
Random Forest	97.7	77.4	79.7	84.9
Gaussian Naive Bayes	97.4	74.6	**80.6**	84.2
Decision Tree	95.0	67.0	65.1	75.7
Support Vector Machines	97.9	76.8	80.6	84.9

Among the baseline methods, Logistic Regression, Random Forest, Gaussian Naive Bayes, and Support Vector Machines demonstrate performance close to that of the NN, particularly in the MDA and LDA categories. In contrast, k-NN and Decision Tree classifiers show significantly lower recall in these categories, suggesting limited robustness in more nuanced classification scenarios.

Notably, the NN model achieves the highest BAS at 85.6%, indicating superior overall performance in correctly classifying instances across all three environment types. This suggests that the NN model offers a more balanced and consistent approach to multiclass classification in this context.

3.4.4 Sensitivity Analysis

To further investigate the defining structural features of the different amorphous ice phases, I conducted a sensitivity analysis of the NN input features using permutation feature importance (PFI), as implemented in the `scikit-learn` library [323]. PFI gauges feature importance by measuring the reduction in a model's accuracy when a single feature value is randomly shuffled [324]. By disrupting the relationship between the feature and the target, this process reveals the extent to which the model relies on that specific feature. Since each input feature corresponds to a specific BOO parameter, this analysis provides insight into which local symmetries are most critical for distinguishing between structural phases.

To perform this analysis, three separate neural network models were trained, each configured to perform binary classification by predicting the presence or absence of one structure type (HDA, LDA, or MDA) against all others. Each model used the same optimised hyperparameters described previously. After training, PFI values were computed for all 30 input features, which span the four sets of symmetry descriptors: $q_{[3-12]}$, $\bar{q}_{[3-12]}$, $w_{[4-12]}$, and $\bar{w}_{[4-12]}$.

As shown in Fig. 3.5, several features emerged as consistently informative across all three phases. Notably, $\bar{q}_4$ and $\bar{q}_{12}$ exhibited strong importance in all models, suggesting that these parameters capture key aspects of local structure relevant for

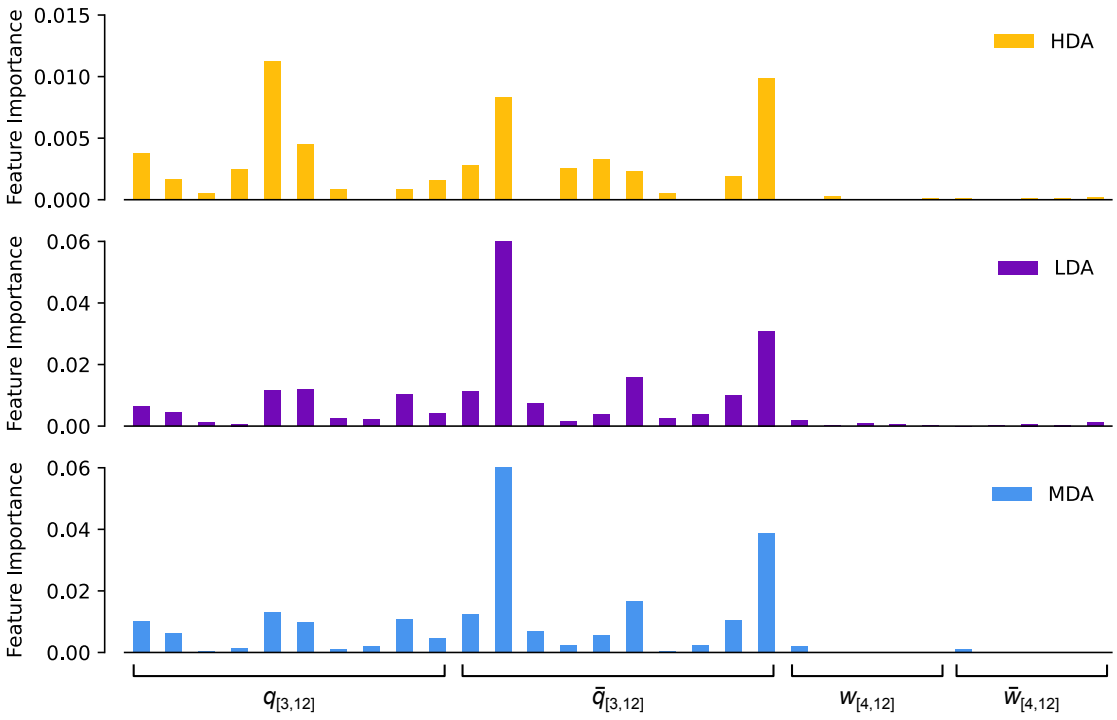

Figure 3.5: PFI scores of the 30 Steinhardt descriptors for each of the three binary classification tasks. Each panel shows the relative importance of features for predicting HDA (top), LDA (middle), and MDA (bottom). Adapted from Ref. [244]

distinguishing amorphous ice phases. In the context of BOOs, $\bar{q}_4$ is sensitive to local tetrahedral and cubic symmetries, while $\bar{q}_{12}$ probes more complex icosahedral and higher-order angular correlations. Their relevance here indicates that distinctions among amorphous ice phases may be largely governed by variations in tetrahedral ordering and longer-range angular coherence. Beyond these shared features, phase-specific patterns of importance became apparent.

For the HDA classifier, certain lower-order descriptors – particularly q_7 and $\bar{q}_7$ – played a significantly more prominent role than in the LDA or MDA models. This suggests that HDA may exhibit local structural motifs associated with characteristic q_7 values, enabling the model to reliably identify HDA environments in contrast to LDA and MDA. The $\bar{q}_7$ parameter is sensitive to sevenfold angular symmetry, which is not compatible with common crystalline lattices and is often associated with disordered or frustrated local arrangements. Its importance in the HDA model implies that such motifs may be more prevalent in high-density amorphous ice,

contributing to its distinct local structural signature.

In contrast, the LDA and MDA models focused more narrowly on a few dominant features, especially $\bar{q}_4$ and $\bar{q}_{12}$, with a slight preference for $\bar{q}_8$ over other descriptors. Importantly, the set of influential features used to distinguish LDA and MDA is nearly identical, which would not pose an issue if the distributions of these descriptors differed. However, as shown in Fig. 3.4, the BOO parameter distributions for LDA and MDA are highly similar, with substantial overlap across all three parameters. This suggests that, at the level of local structure, LDA and MDA are nearly indistinguishable based on the angular symmetries captured by these descriptors. The model's difficulty in differentiating between the two, reflected in both the confusion matrices and prediction confidence distributions (Fig. 3.3), thus stems not from limitations in model capacity, but from an inherent lack of structural distinctiveness. Taken together, these observations provide further evidence that MDA does not represent a structurally independent form of glassy water, but rather a disordered state closely resembling LDA in its local atomic arrangements.

Interestingly, the cubic order parameters w_l and $\bar{w}_l$ consistently showed low importance across all models, indicating they contribute little to the classification of local environments in amorphous ices. These descriptors could likely be omitted in future work with minimal impact on classification performance.

Overall, this analysis highlights that model performance is driven by a small subset of descriptors, with substantial overlap in feature importance between LDA and MDA complicating accurate phase discrimination.

3.4.5 Compression Trajectories

To further examine the structural characteristics of MDA and its relationship to other amorphous ice phases, the trained NN classifier was applied to a set of configurations obtained from LDA compression trajectories reported in Ref. 319. These trajectories were generated by isothermal compression of LDA at three temperatures ($T = 100\,\mathrm{K}$, $T = 120\,\mathrm{K}$, and $T = 140\,\mathrm{K}$), from 10^{-4} to $2.0\,\mathrm{GPa}$, at a compression

rate of 0.01 GPa/ns. Each oxygen atomic environment in each structure along the trajectory was classified into one of the three target amorphous phases – LDA, HDA, or MDA – using the neural network. Figure 3.6a presents an illustrative series of snapshots from the 100 K trajectory, with atoms colour-coded according to their predicted class.

The population fractions of the three structural classes, computed as a function of pressure at $T = 100$ K, are shown in Fig. 3.6b. As expected, LDA-like environments dominate at low pressures, progressively declining as pressure increases. Around 0.75–0.85 GPa, a sharp phase transition is observed: the fraction of LDA-like atoms drops rapidly to zero, coinciding with a steep rise in HDA-like environments. This behaviour is consistent with prior reports of a first-order-like LDA-to-HDA transformation [77, 79, 305, 325].

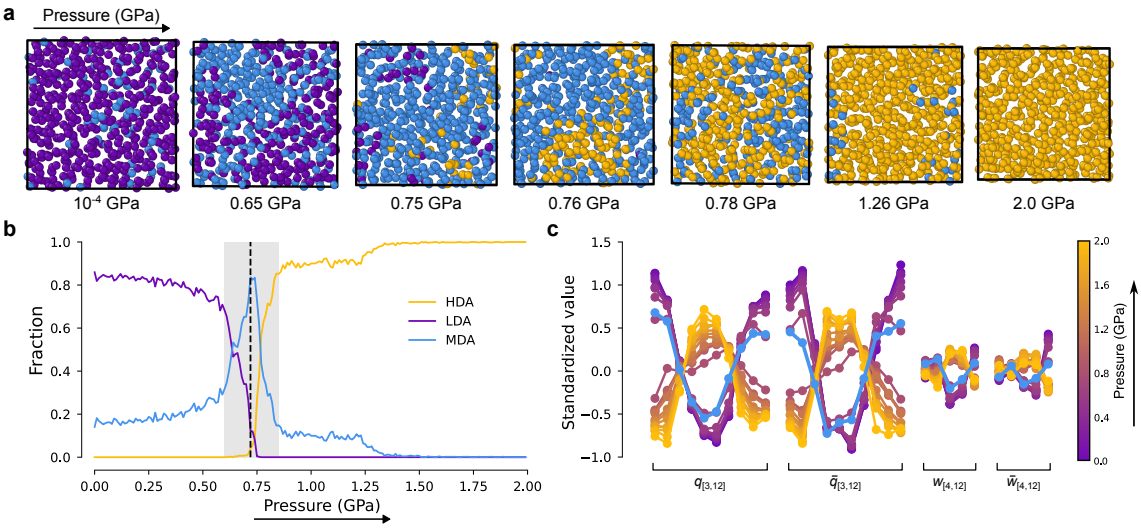

Figure 3.6: Compression trajectory of LDA. (a) Illustrative set of structure snapshots along the trajectory, where each atom is color-coded according to the class of amorphous ice as which the environment was classified. (b) Fraction of local environments as classified by the NN as a function of pressure. The shaded region corresponds to the LDA-to-HDA transition. The dashed line marks the point of maximum structural similarity between compressed LDA and MDA (c) Evolution of all BOO parameters considered herein as a function of pressure for $T = 100$ K. Adapted from Ref. [244].

In contrast, MDA-like environments exhibit a distinct trend. At low pressures, there is a gradual increase in the fraction of MDA-like environments. Strikingly, during the LDA-to-HDA transition region (grey shaded area in Fig. 3.6b), a clear

peak in the MDA-like fraction emerges, followed by a rapid decline as the HDA phase becomes dominant. This suggests that MDA-like local environments are transiently populated during the transition, hinting at a possible structural or energetic intermediate phase.

To provide a more detailed picture of the structural evolution during compression, Fig. 3.6c shows the pressure-dependent behaviour of all 30 BOO parameters, standardised across configurations from the 100 K trajectory. Two distinct clusters are clearly visible: one corresponding to low-pressure, LDA-like environments (purple) and the other to high-pressure, HDA-like environments (yellow). The BOO parameters reveal pronounced differences in local structure between these two phases, consistent with established understanding of their respective short-range order.

These well-separated structural signatures help explain the neural network's high classification accuracy when distinguishing between LDA and HDA environments (in the absence of MDA). In BOO parameter space, the two phases occupy nearly orthogonal regions, making their separation by the NN relatively straightforward. In contrast, the BOO parameters associated with MDA-like environments (blue) show substantial overlap with those of LDA. This pattern not only accounts for the observed classification ambiguity (Fig. 3.3) but also supports the interpretation that MDA represents a structural intermediate between LDA and HDA, rather than a distinct glassy phase. According to these results, MDA does not occupy a unique region in feature space but instead bridges the local structural characteristics of the low- and high-density forms. This continuity reinforces the conclusion that MDA is not a separate amorphous state, but a transitional configuration with mixed structural features.

The structural evolution analysis is consistent with the earlier feature importance results (Fig. 3.5). In particular, $\bar{q}_4$ and $\bar{q}_{12}$, emerged as the most important features for discriminating between LDA and HDA, consistent with their marked differences in values between the two phases as shown in Fig. 3.6c. On the other hand, the

cubic parameters, w and $\bar{w}$ display minimal variation and significant noise across the pressure range, aligning with their low predictive value observed in the sensitivity analysis.

To identify the thermodynamic conditions under which the compressed LDA structures most closely resemble MDA, we compared the BOO parameters of the evolving configurations with those of MDA. The dashed line in Fig. 3.6b marks the point of maximum structural similarity between compressed LDA and MDA, which occurs in the middle of the LDA-to-HDA transformation region (shaded area). This finding supports the interpretation of MDA as a metastable or transient intermediate state that emerges during the compression of LDA. Identical trends were observed for the $T = 120\,\mathrm{K}$ and $T = 140\,\mathrm{K}$ compression trajectories (see Fig. A1).

3.5 Conclusion and Outlook

This chapter introduced an ML framework for classifying local atomic environments in amorphous ices using BOO parameters and a NN classifier. By analysing extensive structural data for LDA, HDA, and MDA ice phases, I achieved robust classification performance, particularly in distinguishing HDA from the other forms. Crucially, the substantial overlap in BOO parameter distributions for LDA and MDA revealed that these two phases are structurally very similar at the local level. These findings point toward an intrinsic structural similarity between LDA and MDA, and provide strong evidence that MDA is not a distinct form of glassy water, but rather a transitional or intermediate form closely resembling LDA.

This interpretation has since been reinforced by the work of two separate groups. De Almeida Ribeiro *et al.* [326] demonstrated through molecular simulations that MDA is best understood as a shear-driven amorphous (SDA) phase – a nonequilibrium steady state governed not only by temperature and pressure, but also by shear rate. Their work shows that MDA can be reproducibly generated from ice $\mathrm{I_h}$, LDA, or HDA via shearing, and that the properties of the resulting amorphous material (density, enthalpy, and structural features) span a continuous spectrum depending

on the applied shear rate. These shear-induced states can reach densities and structures inaccessible by cooling or compression alone, revealing shear as a new axis in water's phase diagram.

Similarly, recent work by Eltareb *et al.* [318] employed MD simulations to show that a continuum of amorphous ices can be generated through isobaric cooling of liquid water at varying pressures. These intermediate amorphous (IA) states smoothly interpolate between LDA and HDA in both density and local structure, and the IA generated at ∼125 MPa was found to be structurally and thermodynamically indistinguishable from experimentally observed MDA. From a potential energy landscape perspective, this IA occupies an intermediate region between the well-separated LDA and HDA basins, indicating that MDA does not correspond to a unique, well-defined glassy state.

The structural ambiguity identified in this chapter, particularly the difficulty in separating LDA and MDA using local BOO metrics, provided early evidence that MDA does not constitute a distinct amorphous phase. The subsequent studies by de Almeida Ribeiro *et al.* [326] and Eltareb *et al.* [318] offer strong support for this interpretation: both independently concluding that MDA is not a separate glassy state of water, but rather a member of a broader family of intermediate, nonequilibrium states that can emerge via shear, cooling, or compression. These later findings reinforce my central conclusion and highlight the broader limitations of using static local descriptors to classify inherently continuous amorphous structures.

Looking forward, these developments imply that future structural classification efforts might go beyond traditional order parameters to incorporate non-local, dynamic, or shear-sensitive features. Beyond water, such approaches could be broadly applicable: for example, complex structural transitions under pressure are well documented in amorphous silicon [327, 328], and bond orientational order parameters initially developed for water [329] have proven useful for characterising local tetrahedrality in phase-change memory materials [330]. Hence, the methodology developed

here could form the foundation for a general and systematic framework for studying amorphous materials across diverse systems.

Chapter 4

Topological origin of peak splitting in the structure factor of liquid water

4.1 Acknowledgements

I am grateful to John Gardner for his assistance with coding, particularly in the implementation of ring analysis calculations. I thank Louise Rosset for her valuable insights into molecular dynamics implementations and the interpretation of structure factor data. Finally, I thank Dr Fausto Martelli for providing the code for the ring analysis of the hydrogen bond network.

4.2 Introduction

X-ray and neutron scattering are among the most powerful experimental techniques for investigating the structural characteristics of liquids, providing direct access to the structure factor, $S(q)$. In simple liquids described by Lennard-Jones interactions, the structure factor typically exhibits a primary peak at a wavenumber associated with the mean distance between particles. In tetrahedrally coordinated liquids such as water [331], silicon [332], and silica [333–335] however, this pattern does not hold. Instead the dominant peak in $S(q)$ often appears at a lower wavenumber. This lower-wavenumber feature is known as the first sharp diffraction peak (FSDP), a structural hallmark whose origin has long been debated [336–339].

Shi and Tanaka have provided important insights into the structural basis of the FSDP in network-forming liquids and glasses. Their work indicates that the FSDP is composed of two overlapping contributions: one stemming from disordered

local configurations lacking tetrahedral symmetry, and another arising from tetrahedrally coordinated domains [331, 332]. This dual contribution supports a two-state description of water, in which the liquid consists of a dynamic mixture of locally ordered and less ordered structures [340–342].

Experimental observations on supercooled water reveal that the main peak in $S(q)$ splits into two distinct maxima upon cooling [311, 343, 344]. This behaviour is consistent with the two-state model: the lower-q peak corresponds to more open, tetrahedral environments, while the higher-q maximum is associated with more compact, disordered regions [311, 331, 332, 343, 344]. The growing separation of these peaks upon cooling reflects increasing spatial correlations and structural heterogeneity – phenomena that are closely tied to water's thermodynamic anomalies and the hypothesised LLCP [309].

Despite these insights, the microscopic mechanisms driving the peak splitting remain unclear. In particular, the role of medium-range structural features has not been investigated. Recent work on silicate glasses [345] has shown that ring-size distributions significantly influence the FSDP, highlighting a strong connection between medium-range topological order and scattering signatures. This raises an interesting question for liquid water: which hydrogen-bonded topological motifs are responsible for the observed splitting in $S(q)$ during cooling?

In this chapter, I employ advanced atomistic simulations powered by a graph-network-based interatomic potential to investigate the influence of hydrogen-bonded rings of varying sizes on the structure factor, $S(q)$. By explicitly linking specific topological motifs to the splitting observed in the FSDP, the results demonstrate that the topology of the hydrogen-bond network (HBN) plays a central role in determining water's structural features. This work establishes a direct connection between topological and structural characteristics, is consistent with experimental findings, and offers deeper insight into the complex nature of liquid water.

4.3 Methods

4.3.1 Data set

The dataset used in this chapter is taken directly from the work of Ibrahim *et al.* [346], who developed an active learning workflow to train an accurate ACE [248, 347] interatomic potential for water [348]. The dataset includes 2,575 DFT-labelled structures comprising approximately 174,000 atoms, spanning a broad range of water configurations – including proton-disordered ice structures generated via `GenIce` [349, 350], molecular dimers, and snapshots sampled from NPT molecular dynamics (summarised in Table 4.1). Densities range from 0.16 to 1.82 g/cm^3; for comparison, ice I$_h$ has a density of 0.92 g/cm^3, and HDA reaches 1.17 g/cm^3. All structures were labelled using static DFT calculations with the PBE+D3 functional [142, 145, 146], chosen for its well-documented accuracy in modelling water [351]. Full details of the dataset construction are available in Ref. [346].

Table 4.1: Summary of the dataset generation steps. Table adapted from Ref. [346].

AL step	Structure type	Number of configurations	Number of atoms
0	initial ices	581	49686
1	highly deformed ices	2335	157174
2	O-O, O-H, H-H dimers	2432	155068
3	liquid water	2575	173686

4.3.2 Structural Analysis

Structure Factor

Some of the most broadly used techniques for characterising internal structure, especially in disordered systems, are scattering experiments, typically using X-rays, neutrons, or electrons. These techniques provide access to the structure factor, $S(\mathbf{q})$,

which quantifies how density fluctuations in a material are spatially correlated in reciprocal (Fourier) space.

The static structure factor is defined as

$$S(\mathbf{q}) = \frac{1}{N} \left\langle \sum_{j=1}^{N} \sum_{k=1}^{N} e^{-i\mathbf{q}\cdot(\mathbf{r}_j - \mathbf{r}_k)} \right\rangle, \tag{4.1}$$

where N is the number of particles, $\mathbf{q}$ is the scattering vector, $\mathbf{r}_j$, $\mathbf{r}_k$ are particle positions, and $\langle \cdot \rangle$ indicates averaging over time or ensembles (since instantaneous measurements are not typically experimentally accessible). This expression accounts for correlations between all particle pairs and, in isotropic systems, depends only on the magnitude $q = |\mathbf{q}|$.

In crystalline materials, $S(q)$ displays sharp Bragg peaks, reflecting long-range order and periodic lattice symmetry. In contrast, amorphous solids and liquids exhibit broad, diffuse features in $S(q)$, indicative of short-range order and a lack of periodicity. This makes it considerably more challenging to extract detailed structural information from the structure factor in disordered systems. Nevertheless, meaningful insights can still be drawn, particularly from the FSDP. Although its exact origin has long been debated, the FSDP is now loosely interpreted as a structural signature of medium-range correlations beyond the nearest-neighbour scale.

As discussed in the introduction, Shi and Tanaka offered a compelling microscopic interpretation of the FSDP in tetrahedral liquids [331, 332]. Their work revealed that this feature arises from the superposition of two distinct contributions: one from density fluctuations associated with locally favoured tetrahedral structures, and another from disordered configurations lacking tetrahedral symmetry. The geometric regularity of the former produces characteristic density modulations that give rise to well-defined features in reciprocal space, distinguishable from the broader scattering patterns of disordered regions.

Intermediate Scattering Function

The intermediate scattering function (ISF), $F(\mathbf{q}, t)$, extends the insight offered by the structure factor into the temporal domain, capturing how spatial correlations evolve over time. While $S(\mathbf{q})$ characterises the *instantaneous* structure, $F(\mathbf{q}, t)$ reveals the *relaxation dynamics* of those structures, making it especially useful for diagnosing equilibration and dynamical behaviour in molecular simulations.

It is defined as

$$F(\mathbf{q}, t) = \frac{1}{N} \left\langle \sum_{j=1}^{N} e^{-i\mathbf{q}\cdot[\mathbf{r}_j(t) - \mathbf{r}_j(0)]} \right\rangle, \tag{4.2}$$

where $\mathbf{r}_j(t)$ is the position of the jth particle at time t, and $\mathbf{q}$ is the wavevector. Choosing $|\mathbf{q}| = \frac{2\pi}{L}$, where L is the box length, probes fluctuations on the scale of the entire simulation box.

At short times, $F(\mathbf{q}, t)$ reflects vibrational motion and transient caging effects, while its decay at longer times indicates structural relaxation and diffusive behaviour. In an equilibrated system, the function is expected to decay to zero, indicating a loss of correlation with the initial configuration. Thus, $F(\mathbf{q}, t)$ complements the structure factor by providing dynamic information about how the system transitions between configurations, essential for understanding phenomena such as glassy dynamics, diffusion, and structural arrest.

Radial Distribution Function

The radial distribution function (RDF), $g(r)$, provides a real-space counterpart to the structure factor, quantifying the probability of finding a particle at distance r from a reference particle, normalised by the expectation for an ideal gas at the same density. It is defined as

$$g(r) = \frac{1}{4\pi r^2 \rho N} \left\langle \sum_{i=1}^{N} \sum_{j \neq i}^{N} \delta(r - |\mathbf{r}_i - \mathbf{r}_j|) \right\rangle, \tag{4.3}$$

where ρ is the bulk number density, and the delta function selects particle pairs separated by distance r. In practice, $g(r)$ is obtained via binning and ensemble averaging.

The shape and features of $g(r)$ provide insight into the degree of structural order in a system. In crystalline solids, $g(r)$ displays a series of sharp, regularly spaced peaks that persist over long distances, reflecting the repeating periodic structure of the lattice. The positions and intensities of these peaks correspond to specific interatomic distances characteristic of the crystal's symmetry and unit cell. In contrast, disordered systems lack long-range order, and while the first few peaks in $g(r)$ still indicate local packing (e.g., nearest neighbours), the function quickly decays to unity, reflecting a loss of coherence beyond a few coordination shells. Similarly, liquids show a well-defined first peak in $g(r)$ – indicative of short-range ordering – followed by smooth decay.

Notably, $g(r)$ and $S(q)$ are mathematically connected via a Fourier transform:

$$S(\mathbf{q}) = 1 + \rho \int_V \mathrm{d}\mathbf{r} \;\; e^{-i\mathbf{q}\mathbf{r}} g(\mathbf{r}) \tag{4.4}$$

which for an isotropic system (e.g. a liquid) resolves to:

$$S(q) = 1 + \frac{4\pi\rho}{q} \int_V \mathrm{dr} \;\; r \sin(qr) \cdot [g(r) - 1] \tag{4.5}$$

This relationship allows direct comparison between simulation (often expressed through $g(r)$) and experiment (typically via $S(q)$). As a result of this relationship, information from either function can, in principle, be used to reconstruct the other. This connection is particularly useful in comparing simulation results with experimental scattering data, allowing validation and refinement of atomistic models.

Ring Statistics

In the previous sections, I discussed how the RDF and the structure factor are used to probe the arrangement of atoms in systems. These functions encode in-

formation about atomic correlations and are fundamentally related through Fourier transforms. While they effectively capture local and extended pairwise correlations, characterising medium-range structural features, such as network connectivity and atomic ring configurations, often requires complementary approaches.

Medium-range order plays a vital role in determining the structure–property relationships in amorphous materials, where the absence of conventional long-range order complicates traditional crystallographic analyses. Zachariasen first showed that under certain thermodynamic conditions, the mechanical properties of glasses are comparable to those of crystals [67], highlighting the importance of atomic connectivity in the determination of such properties. One way to characterise this connectivity is through the analysis of the system's topological network in the form of rings. First introduced by King in 1967 [352], ring statistics have since become a well-established tool for quantifying medium-range order [353–356]. They have been extensively applied to investigate the network topology of amorphous systems [357–359], as well as to analyse continuous random networks [360–362].

Ring analysis involves identifying and counting closed loops or "rings" formed through interatomic connections. These rings reveal how structural units are linked across multiple coordination shells, capturing recurring motifs that extend beyond local order. A common approach to computing ring statistics involves counting all rings up to a given size. However, as King noted [352], this method quickly leads to a proliferation of rings per atom due to redundant or compound rings formed by overlapping smaller rings. These larger, overlapping rings can obscure meaningful structural features of the network.

The key challenge, therefore, lies in defining a criteria that captures all the primitive rings while filtering out redundant ones. Given the abundance of definitions of ring counting schemes in the literature [352, 354, 357, 363–366] it is no surprise that highly variable, and sometimes inconsistent, ring statistics are reported even for simple crystalline structures, let alone more complex network structures like amorphous

silicates [363, 366–368] or water [369–371]. Notably, recent work by Formanek and Martelli [372] has shown that different counting schemes reveal distinct structural information and that using multiple definitions can offer complementary insights into the network topology of water.

In this chapter, I focus on the hydrogen-bond network (HBN) of water, a particularly intricate and directional network due to the nature of hydrogen bonding. Before defining a ring-counting scheme for water, it is essential to address the ambiguity in defining a hydrogen bond itself. Although there is no single, quantitative measure for the number of hydrogen bonds a given oxygen atom is involved in, the qualitative agreement among various proposed definitions – based on geometry, topology, and electronic structure – has been considered satisfactory across a broad range of thermodynamic conditions [373]. For this work, I therefore adopt the widely used geometric definition proposed by Luzar and Chandler [374], where a hydrogen bond is defined by both a distance criterion, $R_{\mathrm{OO}} < 3.5\,\text{Å}$, and an angular constraint, $\beta < 30°$, where $\beta \equiv \angle \mathrm{O_A} \cdots \mathrm{O_D} - \mathrm{H_D}$.

With this definition, I construct ring statistics in the HBN of water as follows: starting from a tagged water molecule, I recursively traverse its hydrogen-bond connections, when the traversal returns to the original molecule, a ring is formed. This process continues until either a ring is completed or the maximum ring size (set to 12 water molecules in this case) is exceeded (Figure 4.1).

Orientational Tetrahedral Order Parameter

The orientational tetrahedral order parameters is one of the most commonly used metrics for quantifying tetrahedral order in water [70, 375–378]. It was originally introduced by Chau and Hardwick [379] and subsequently rescaled by Errington and Debenedetti [378].

It is mathematically defined as:

$$q = 1 - \frac{3}{8} \sum_{j=1}^{3} \sum_{k=j+1}^{4} \left(\cos \psi_{jk} + \frac{1}{3} \right)^2 , \tag{4.6}$$

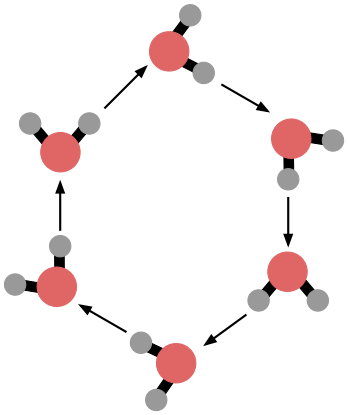

Figure 4.1: Schematic illustration of ring construction within the HBN of water. Oxygen atoms are shown in red, hydrogen atoms in grey. Directional hydrogen bonds (black arrows) are followed from hydrogen to oxygen atoms, tracing paths through the HBN until either a closed loop is formed – returning to the starting molecule – or the path exceeds a ring size of 12. In this example, a hexagonal ring consisting of six water molecules is identified.

where ψ_{jk} is the angle formed by the lines joining the central oxygen atom, i, and those of the nearest neighbours j and k (≤ 4). The sum is over distinct pairs of the four closest neighbours of molecule i (there are 6 possible O-O-O angles per molecule). The average value:

$$\langle q \rangle = \frac{1}{N} \sum_{i=1}^{N} q_i, \tag{4.7}$$

quantifies the orientational order of the system based on the molecules in the first coordination shell. In this thesis, the notation q will automatically represent $\langle q \rangle$, where averaging is done over all atoms and all configurations.

In an ideal gas, where molecular orientations are random and uncorrelated, the six angular terms are independent, resulting in $q = 0$. In contrast, for a perfect tetrahedral arrangement, $\cos(\psi_{jk}) = -1/3$ for all angles, yielding $q = 1$. Thus, q quantifies the degree of local tetrahedral order, serving as a direct measure of deviation from ideal tetrahedrality in the first coordination shell of a central oxygen atom.

4.3.3 Transport Properties

Self-Diffusion Coefficient

Transport properties of a system can be investigated by analysing the mean square displacement (MSD) of particles as the system evolves dynamically. The MSD quantifies the average squared distance that particles travel over time and is mathematically defined as

$$\mathrm{MSD}(t) = \left\langle \Delta\mathbf{r}^2(t) \right\rangle = \left\langle \frac{1}{N} \sum_i \left| \mathbf{r}_i(t_0 + t) - \mathbf{r}_i(t_0) \right|^2 \right\rangle, \tag{4.8}$$

where $\mathbf{r}_i(t)$ is the position of the ith atom at time $t \in \mathbb{R}_{\geq 0}$, N is the number of atoms in the system and $\langle \cdot \rangle$ denotes averaging over many different starting times, t_0. The MSD provides insight into the dynamical regime of the system, distinguishing between ballistic, diffusive, and anomalous transport behaviours.

In systems where diffusion arises from random thermal motion, the MSD grows linearly with time, and the diffusion coefficient D, which characterises the rate of particle spreading, can be obtained using the Einstein relation:

$$\mathrm{MSD} = 6Dt \tag{4.9}$$

This leads to an expression for computing D as the slope of an MSD versus time plot in the linear regime.

$$\mathrm{D} = \frac{1}{6t} \left\langle \frac{1}{N} \sum_i \left| \mathbf{r}_i(t_0 + t) - \mathbf{r}_i(t_0) \right|^2 \right\rangle. \tag{4.10}$$

It is essential to extract D from the region where the MSD exhibits linear time dependence, corresponding to normal (Fickian) diffusion [380]. At very short times, particles often move ballistically, travelling in straight lines due to inertia before undergoing significant collisions, resulting in a superlinear MSD growth that does not reflect true random diffusive behaviour. At long timescales, deviations from

linearity may also occur due to subdiffusive behaviour, where constraints such as confinement or viscoelastic effects hinder motion. By focusing on the linear regime, one ensures that the computed diffusion coefficient accurately reflects steady-state diffusion, free from transient or anomalous influences.

Rotational Lifetime

The rotational lifetime of a molecule is a dynamical quantity that characterises the timescale over which the molecule retains memory of its initial orientation. Physically, it represents the time required for the angular correlation between a molecule's initial and subsequent orientations to decay significantly, typically due to rotational diffusion. This property is particularly important for understanding the dynamic behaviour of molecules like water, where rotational mobility influences hydrogen bonding, dielectric relaxation, and spectroscopic responses. In water, rotational dynamics are intimately connected to the continual reorganisation of the hydrogen bond network, making the rotational lifetime a sensitive probe of molecular interactions.

To determine the rotational lifetime, I compute the second-order rotational correlation function:

$$C(t) = \langle \mathbf{u}(t_0) \cdot \mathbf{u}(t + t_0) \rangle \,, \tag{4.11}$$

where $\mathbf{u}(t)$ is a unit vector fixed in the molecular frame: I chose the arithmetic mean of the two O→H bond vectors. This correlation function captures how the orientation of a molecule evolves over time, and is averaged over many molecules and initial time origins to ensure statistical reliability.

In isotropic systems, this correlation function often exhibits an exponential decay:

$$C(t) \approx e^{-t/\tau}, \tag{4.12}$$

where τ is the rotational correlation time, also referred to as the rotational lifetime. The characteristic time τ thus quantifies the rate at which rotational memory is lost.

In practice, extracting τ assumes that $C(t)$ follows a single-exponential decay and typically involves performing a linear fit over an intermediate time window where this behaviour is approximately valid:

$$\tau = -\frac{1}{\log C(t)} t \tag{4.13}$$

This approach is suitable here since simulations are confined to relatively high temperatures, thus rotational dynamics remain relatively homogeneous. At lower temperatures, where relaxation becomes more heterogeneous, the decay of $C(t)$ may deviate from exponential behaviour, and more sophisticated fitting models, such as stretched exponentials, are required to capture the distribution of relaxation times.

4.4 Results

4.4.1 Hyperparameter Optimisation

The initial step in developing a machine-learned interatomic potential involves the identification and optimisation of key hyperparameters. Hyperparameters exert a substantial influence on the performance and generalisability of the model; thus, selecting an appropriate set is critical to ensuring that the resulting potential is both accurate and computationally efficient.

The optimisation strategy employed in this work was informed by a combination of insights from the literature and empirical tuning through trial and error. The overarching objective was to achieve a balance between predictive accuracy and computational efficiency, particularly with respect to inference time.

I began by tuning the hyperparameters of the MACE model. MACE includes a range of hyperparameters that influence its accuracy and expressive power. Most of these come with well-tested default values that have been shown to perform robustly across a range of systems by the original developers of the MACE framework [381]. Therefore, I focused on three key hyperparameters that are commonly recommended

for adjustment [256] based on the specific task or system: the number of channels, the maximum order of spherical expansion (ℓ_{max}), and the cutoff radius. These parameters have the most pronounced effect on both the model's performance and its inference cost.

Figure 4.2 summarises the performance and computational cost of MACE models with different settings of these hyperparameters. Computational cost is quantified as the inference time and reported in microseconds per atom per step.

The first panel (Figure 4.2a) explores the impact of varying the number of channels (16, 32, and 64), which determines the internal dimensionality, or "width", of the model and serves as a proxy for its representational capacity. As shown, increasing the number of channels reduces both energy (left) and force (middle) root-mean-square error (RMSE), indicating improved prediction accuracy as the model's ability to learn complex interatomic interactions increases. However, this improvement comes at the expense of a sharp increase in inference time (right), highlighting the trade-off between model expressivity and computational efficiency.

Panel b investigates the influence of the maximum angular momentum quantum number ℓ_{max}, which governs the angular resolution in the spherical harmonics expansion of atomic environments. A value of $\ell_{\text{max}} = 0$ corresponds to using only rotationally invariant features. Moving from $\ell_{\text{max}} = 2$ to $\ell_{\text{max}} = 3$ provides modest gains in force accuracy and slight improvements in energy predictions, suggesting enhanced modelling of angular interactions. These gains, however, are accompanied by a noticeable increase in inference time, underscoring the computational cost of higher angular complexity.

Finally, panel c evaluates the effect of increasing the cutoff distance (3.5–6.5 Å), which defines the spatial range over which neighbour atoms contribute to message passing. Expanding the cutoff leads to substantial improvements especially in energy accuracy up to 5.5 Å, beyond which gains diminish. Inference time, however, increases monotonically with cutoff, reflecting the growing number of interactions

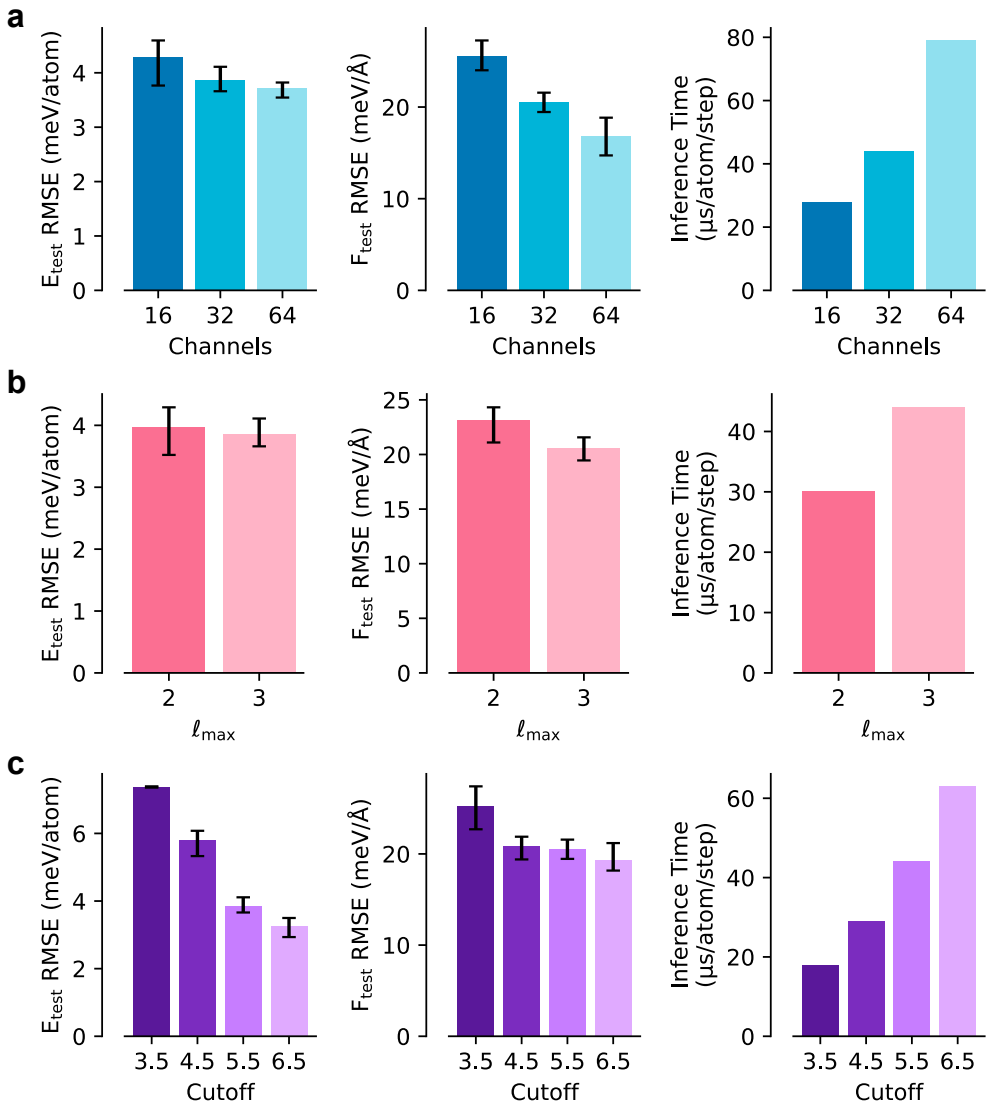

Figure 4.2: Optimisation of MACE hyperparameters. Performance and computational cost of MACE models evaluated with varying hyperparameters. Computational cost is measured in terms of inference time (μs/atom/step). (a) Impact of the number of channels (16, 32, 64) on energy RMSE, force RMSE, and computational time. (b) Effect of increasing the maximum angular momentum quantum number (ℓ_{max}) from 2 to 3. (c) Influence of different cutoff distances (3.5, 4.5, 5.5, 6.5 Å) on model accuracy and efficiency. The error bars represent the minimum and maximum values of the RMSE for each hyperparameter setting and are centred on the mean of 5 independent training runs.

considered. A cutoff of 5.5 Å appears to strike an optimal balance between accuracy and efficiency, coinciding with the end of the second coordination shell in the RDF of water. Most of the relevant interactions are already captured within the 5.5 Å cutoff, so increasing it further primarily adds weakly interacting neighbours, which

dilutes the meaningful signal. This reduces the signal-to-noise ratio and can exceed the model's capacity to extract useful patterns, limiting further improvements in accuracy.

In all MACE hyperparameter studies, model selection was based on optimal numerical performance, with accuracy prioritised even when it incurred increased inference cost. This decision was justified by the availability of sufficient computational resources, enabling the training and deployment of more computationally intensive models without practical constraints. This means in this case I used a model with 64 channels, $\ell_{\max} = 3$ and a cutoff distance of 5.5 Å.

Following architectural optimisation, training-related hyperparameters were investigated. While these do not affect inference time directly, they play a critical role in the model's training dynamics, convergence behaviour, and overall stability. Specifically, batch size and learning rate were examined due to their well-established influence on model generalisation [382, 383].

Figure 4.3a shows the effect of batch size (5, 16, 32) on energy RMSE (left), force RMSE (centre), and the number of training steps to convergence (right). All batch sizes yield similar energy RMSE, with only slight variation. However, force RMSE increases with batch size, indicating that smaller batches better capture the force information. Convergence speed, on the other hand, improves substantially with larger batches: a batch size of 32 requires far fewer training steps than batch sizes 5 or 16. The MACE documentation recommends a batch size of 5 which is consistent with the results here. Larger batch sizes may guide the optimiser toward suboptimal minima due to overly smooth gradient updates.

Figure 4.3b evaluates the impact of learning rate (0.005 vs. 0.01). The higher learning rate results in marginally better energy RMSE, slightly worse force RMSE, and faster convergence. These differences are relatively minor, suggesting that the model is robust to learning rate variation within this range.

Overall, batch size has a more pronounced effect on model performance than

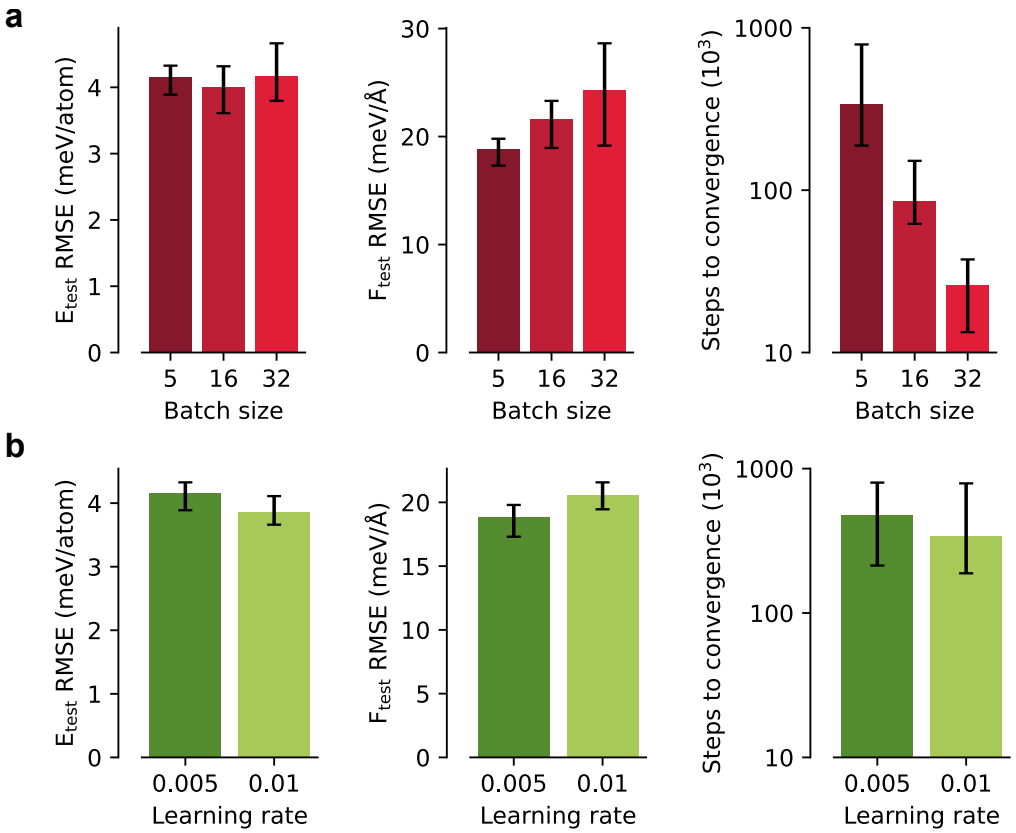

Figure 4.3: Optimisation of training hyperparameters. Analysis of the effect of training hyperparameters on MACE model performance and training efficiency. (a) Influence of batch size (5, 16, 32) on energy RMSE, force RMSE, and steps to convergence. Larger batch sizes reduce training time but may compromise accuracy. (b) Comparison of two learning rates (0.005 and 0.01) showing minimal impact on final errors or total training steps, indicating robustness to learning rate within this range. The error bars represent the minimum and maximum values of the RMSE for each hyperparameter setting and are centred on the mean of 5 independent training runs.

learning rate. While large batch sizes accelerate convergence, they compromise force accuracy. Learning rate can be adjusted for efficiency without significantly affecting final performance. Based on these observations, a batch size of 5 and a learning rate of 0.005 were selected for subsequent training, balancing accuracy and convergence speed.

4.4.2 MLIP Validation

Once the optimal hyperparameters were identified, the model was subjected to a comprehensive validation procedure, incorporating both numerical evaluation and

physically informed testing. Validating MLIPs is a complex and nuanced task that requires more than just high numerical accuracy. Recent work has emphasised the importance of combining rigorous quantitative metrics with physically motivated validation to ensure that models are not only statistically accurate but also physically reliable and interpretable [210, 384].

Standard numerical metrics – such as the RMSE on energies and forces – offer a clear and objective measure of predictive accuracy, particularly when benchmarking against established state-of-the-art (SOTA) models. However, the utility of these metrics is inherently tied to the nature and scope of the test dataset, which may not fully capture the diversity of configurations encountered in practical applications. Consequently, strong numerical performance does not necessarily imply physical robustness or generalisability.

To address this, numerical assessment must be complemented with physically guided validation, which probes the model's ability to reproduce known physical behaviour. These tests draw on domain-specific knowledge to assess whether the potential can recover key structural, thermodynamic, or dynamical properties of the target system. Such validation helps bridge the gap between raw numerical performance and real-world applicability, ensuring that the model's predictions remain physically meaningful across a broad range of scenarios.

In the following sections, I first present the model's numerical performance, then demonstrate its ability to qualitatively reproduce key physical properties of liquid water.

Numerical Errors

Five models were trained using different random seeds on a dataset of 2317 water configurations randomly drawn from the ACE dataset (see Section 4.3.1). Model hyperparameters were selected based on the optimised values identified in the previous section. Each model was evaluated on a held-out test set of 128 configurations, which were not used during training.

Table 4.2 summarises the average energy and force mean absolute errors (MAEs) for the evaluated models. These results are compared against reference values reported in the original ACE publication, which also provided the training data for the present MACE model. MAEs are used instead of RMSEs to maintain consistency with the ACE benchmark. Additionally, the table reports the inference time for each model, defined as the average time required to compute energies and forces for a single atom per simulation step on an NVIDIA RTX A6000 GPU. For broader context, the table also includes performance data for two widely used GNN architectures in atomistic modelling as implemented in `GraphPES` [385]: NequIP [167] and PaiNN [267].

Table 4.2: Comparison of numerical performance for different ML potentials of water. Reported metrics include MAE in predicted energies and forces, as well as inference time in microseconds per atom per simulation step.

ML Potential	**Energy MAE (meV/atom)**	**Force MAE (meV/Å)**	**Inference Time (μs/atom/step)**
MACE	0.3	7.7	78.9
NequIP	0.4	11.5	123.8
PaiNN	1.7	49.7	5.3
ACE [346]	2.5	16.7	0.5

From these results, it is evident that MACE achieves the lowest energy and force errors, outperforming all other models considered especially in force accuracy. Despite a higher inference cost when compared to the original ACE model, the improved accuracy highlights the strength of MACE in capturing the underlying potential energy surface of water. In contrast, the PaiNN model is significantly faster than other GNN-based architectures but suffers from substantially higher prediction errors, indicating a trade-off between computational efficiency and accuracy.

Radial Distribution Function

Having established the numerical accuracy of the MACE model, the next step was to assess its ability to reproduce known structural properties of liquid water. Structural validation provides critical insight into whether the model captures the underlying physical interactions that govern molecular arrangement. To this end, I performed MD simulations using the best-performing model from the five independent training runs. The simulation employed a Nosé–Hoover thermostat [281] with a timestep of 1 fs. A randomised box of 1024 water molecules was used, and the system was equilibrated at 300 K and 1 bar for 600 ps in the NPT ensemble, followed by a 1 ns production run in the NVE ensemble.

Equilibration was monitored using the oxygen-oxygen intermediate scattering function, $\mathrm{F}(\mathbf{q}, t)$, as described in Section 4.3.2. Equilibration was deemed to have occured when $\mathrm{F}(\mathbf{q}, t)$ reached 0 (Fig. 4.4).

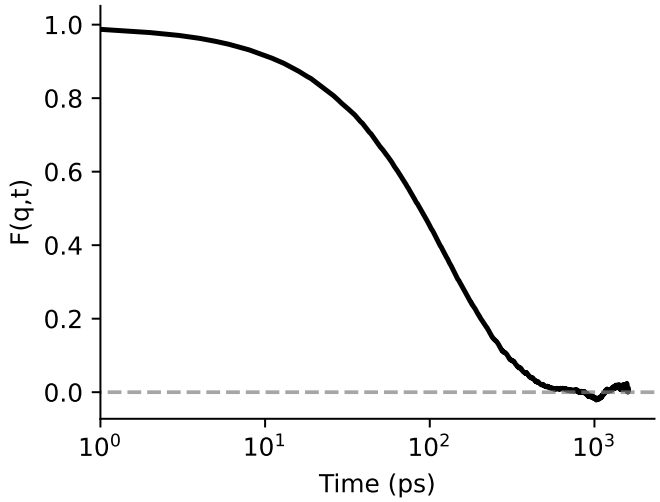

Figure 4.4: Intermediate scattering function $\mathrm{F}(\mathbf{q}, t)$ for oxygen atoms in liquid water at 300 K and 1 bar. The wavevector $q = 0.20\,\text{Å}^{-1}$ corresponds to density fluctuations on the scale of the full simulation box.

Figure 4.5 shows the RDFs computed from the MD trajectory, alongside those from the ACE model [346] and experimental reference data [343, 386]. The MACE

model shows good agreement with the experimental RDFs and captures the key structural features of liquid water, although some deviations remain.

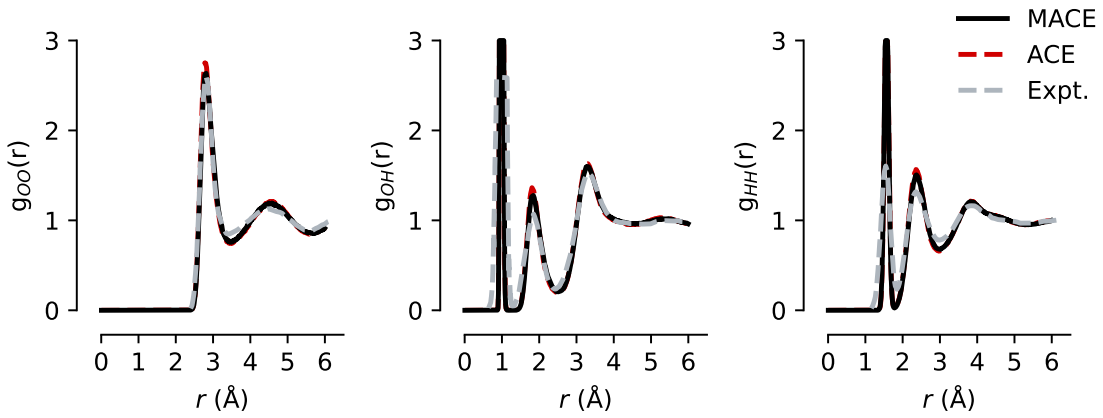

Figure 4.5: RDFs for O–O, O–H, and H–H atomic pairs in liquid water at 300 K and 1 bar. Results are compared against the ACE model [346] and experimental measurements (O–O from Skinner *et al.* [343]; O–H and H–H from Soper *et al.* [386]).

The first peak in $g_{\mathrm{OO}}(r)$ and the second peak in $g_{\mathrm{OH}}(r)$ are associated with correlations between hydrogen-bonded water molecules. The MACE model captures the positions of these features well, though it slightly overestimates the height of the second peak in $g_{\mathrm{OH}}(r)$. This suggests a mild overemphasis on the strength or prevalence of hydrogen bonding, resulting in modest overstructuring of the hydrogen-bond network.

Between the first and second peaks of $g_{\mathrm{OO}}(r)$ lies an interstitial region predominantly occupied by water molecules that do not participate in hydrogen bonding. These molecules are governed primarily by van der Waals (vdW) interactions. As a consequence of the MACE model's slight overprediction of hydrogen bonding, the population density within this interstitial region is reduced compared to experimental observations.

This tendency towards over-organisation may arise from the predominance of ice-like configurations in the training dataset (see Section 4.3.1), which could introduce a structural bias toward more ordered environments. In addition, the model may place disproportionate emphasis on these configurations during training, making it

more difficult to capture the full range of disordered liquid structures. Together, these factors may limit the model's ability to fully generalise across thermodynamic conditions. This observation is further supported by the close agreement between the RDFs predicted by MACE and those from the ACE model [346], which were trained on the same dataset.

Significant deviations are also observed in the heights of the first peaks in the O–H and H–H RDFs. These discrepancies are expected, as the MD simulations presented here do not account for nuclear quantum effects, phenomena known to significantly influence short-range structural correlations in water [152, 387, 388].

The microscopic structure of water emerges from a delicate balance of covalent bonding, hydrogen bonding, and vdW interactions. The ability of the MACE model to closely replicate the experimental RDFs demonstrates its effectiveness in capturing this complex interplay of forces.

Hydrogen Bond Network Analysis at 300 K

Having confirmed that the MACE model shows good agreement with experimental RDFs, I now examined its ability to capture the more detailed, underlying structure of the liquid, specifically, the nature of the HBN. The HBN plays a central role in determining the structural and dynamical properties of water [77, 370, 389, 390], and any realistic water model must be capable of reproducing its characteristic features.

Unlike ice, where hydrogen bonds form a rigid and ordered lattice, the HBN in liquid water is disordered and highly dynamic. Hydrogen bonds in the liquid continuously break and reform, generating a fluctuating tetrahedral network that governs both structure and dynamics [370, 389]. While local coordination remains predominantly tetrahedral, molecules in the liquid typically form around 10% fewer hydrogen bonds than in ice, resulting in transiently broken bonds and coordination defects [101, 391, 392]. These deviations have a direct influence on water's fluidity and are strongly linked to its unique dynamical behaviour [393]. As such, capturing the key features of the HBN is essential for any model aspiring to accurately simulate

liquid water.

To characterise the HBN at 300 K, I adopt two complementary approaches, both applied to the *NVE* trajectory described in the previous section. First, I analyse ring statistics (Section 4.3.2) by computing the normalised probability distribution, $P(n)$, of finding closed hydrogen-bonded rings consisting of $n \in [3, 12]$ water molecules (Figure 4.6a). These statistics are averaged over the last 500 ps of the production run.

Second, I quantify coordination defects, departures from the ideal tetrahedral arrangement, by classifying each water molecule according to the number of hydrogen bonds it donates and accepts (Figure 4.6c). The ideal configuration, A_2D_2, corresponds to a molecule that donates two and accepts two hydrogen bonds. Deviations such as A_1D_1, A_1D_2, A_2D_1, and A_3D_2 are also identified and quantified. While other configurations exist, their occurrence is negligible [394]. The relative abundance of these motifs offers a direct measure of the degree of HBN disruption.

Figure 4.6a shows that, as expected, 5–7-membered rings dominate the distribution, with 6-membered rings being the most frequent [395–397]. This is consistent with the local tetrahedral coordination in liquid water. The relatively high proportion of 6-membered rings (∼20%) also aligns with previous findings for *ab initio* models and exceeds values typically observed in simulations using empirical potentials, which tend to predict higher populations of larger (e.g. 8-membered) rings at ambient conditions [371, 397, 398]. The prominence of the A_2D_2 configuration (∼55%) further reflects a strong tendency towards ideal tetrahedral coordination. These results are consistent with the RDF analysis and confirm a mild overstructuring of the HBN by the MACE model compared to experiment.

Nonetheless, the model clearly improves upon AIMD simulations based on the PBE and PBE0 functionals, which are known to overestimate hydrogen bond strength due to limitations in their treatment of long-range interactions [394]. As shown in Figure 4.6c, the proportion of ideal A_2D_2 motifs decreases from PBE+vdW to

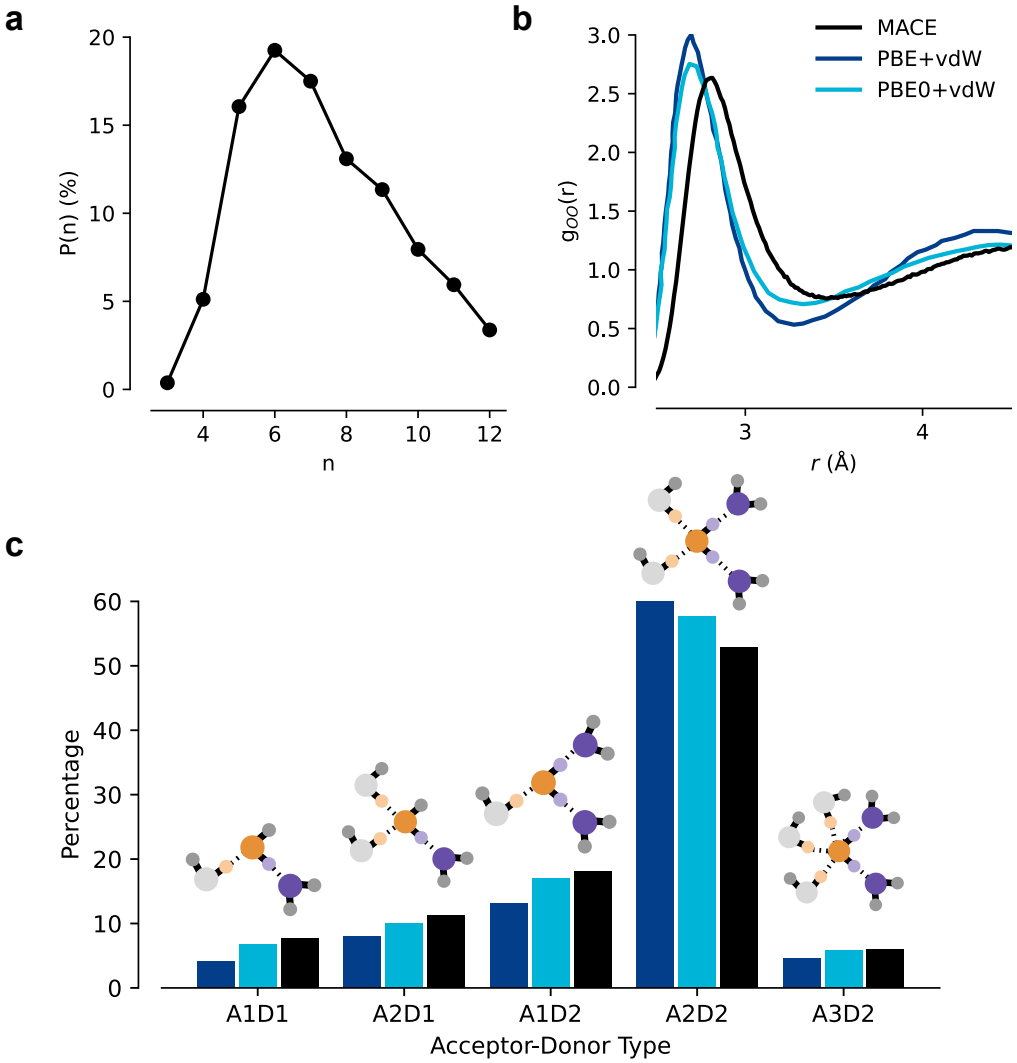

Figure 4.6: HBN analysis of liquid water. (a) Probability distribution $P(n)$ of closed hydrogen-bonded rings containing n water molecules, with $n \in [3, 12]$. (b) Oxygen–oxygen radial distribution function (RDF) focused on the first minimum. Results from the MACE model (black) are compared with AIMD simulations using PBE+vdW (dark blue) and PBE0+vdW (light blue) functionals, reproduced from Ref. [394]. (c) Distribution of intact hydrogen bonds per water molecule, categorised by acceptor (A) and donor (D) roles. Each bar labeled A_xD_y represents the percentage of water molecules forming x acceptor and y donor hydrogen bonds. Schematic diagrams of these configurations are shown above each bar.

PBE0+vdW and drops further with the MACE model. This is accompanied by a compensatory increase in partially coordinated species, including A_1D_2, A_2D_1, and A_1D_1. In contrast, the population of overcoordinated A_3D_2 motifs remains largely unchanged across all methods, as previously noted by DiStasio *et al.* [394].

This trend signals a systematic reduction in local tetrahedral order, corroborated by the average number of intact hydrogen bonds per water molecule: 3.74 for PBE+vdW, 3.67 for PBE0+vdW, and 3.54 for MACE. The increased disorder is also reflected in the oxygen–oxygen RDF, particularly in the reduced depth of the first minimum (Figure 4.6b), which indicates a higher density of molecules in the interstitial region, i.e., outside the primary hydrogen-bonding shell.

Overall, these results demonstrate that the MACE model alleviates some of the known deficiencies of the PBE-class functionals by introducing a greater degree of structural disorder into the hydrogen-bond network. While a slight overstructuring relative to experiment remains, the model represents a clear improvement over simulations based on PBE-class functionals in capturing the fluctuating, disordered nature of liquid water.

Density Isobar

Another key benchmark for validating water models is their ability to reproduce the temperature of maximum density (TMD), a well-known anomaly of liquid water. While the TMD is an important test of model fidelity, its precise value is highly sensitive to the underlying interaction potential. Even *ab initio* methods show considerable variability in their TMD predictions [109, 140]. For instance, Montero de Hijes *et al.* [140] demonstrated that TMD estimates can differ by as much as 100 K depending on the density functional employed. Notably, models trained on the same functional but with different dispersion damping schemes (e.g., zero vs. Becke–Johnson damping) can also yield significantly different TMDs. The primary focus of this section is to assess whether the model qualitatively reproduces the presence of a density maximum and the correct general shape of the isobar.

I computed a 1 bar isobar across a temperature range of 260–350 K, in 10 K increments, using simulations of 256 water molecules with a 1 fs timestep. This system size was chosen based on previous studies indicating that even 128 molecules are sufficient to obtain well-converged density isobars [109]. Simulations were conducted

for both the MACE model and the ACE model from Ref. [346], which share the same training dataset.

All simulations were carried out in the *NPT* ensemble using the equations of motion developed by Shinoda *et al.* [286], with a Nosé–Hoover thermostat and barostat [281]. For simulations at 280 K and above, the initial configurations were randomly generated. For lower temperatures (270 K and 260 K), the simulations were initialised from the final configuration of the preceding higher-temperature run to accelerate equilibration. Production runs were ten times longer than the equilibration period, with total simulation times ranging from 1 ns to 25 ns depending on temperature. Final densities were averaged over the last 300 ps of the production trajectory.

As in the RDF simulations, equilibration was monitored via the ISF, which was required to decay to zero. Figure 4.7 shows the ISF as a function of time for all NPT simulation runs, highlighting the temperature dependence of structural relaxation. At long times, $F(\mathbf{q}, t)$ becomes increasingly noisy due to the reduced number of particle pairs that retain significant correlation. As the system relaxes and the ISF approaches zero, the signal-to-noise ratio decreases because statistical averaging becomes less effective when the underlying correlations have largely vanished.

Figure 4.8 shows the predicted isobar along with comparisons to ML models trained on RPBE+D3 [109] and revPBE+D3 [140] functionals and to experimental data [399]. These comparisons illustrate the significant variability in predicted densities and TMDs that arises from the choice of functional, even among closely related methods. The MACE model, trained on PBE+D3-labelled data [145, 146], reproduces the qualitative features of the isobar and clearly exhibits a density maximum. This behaviour is consistent with physical expectations and confirms the model's ability to capture the key thermodynamic signature of water.

Notably, the MACE model performs substantially better than the ACE model in capturing the isobaric behaviour of water, particularly at lower temperatures. This

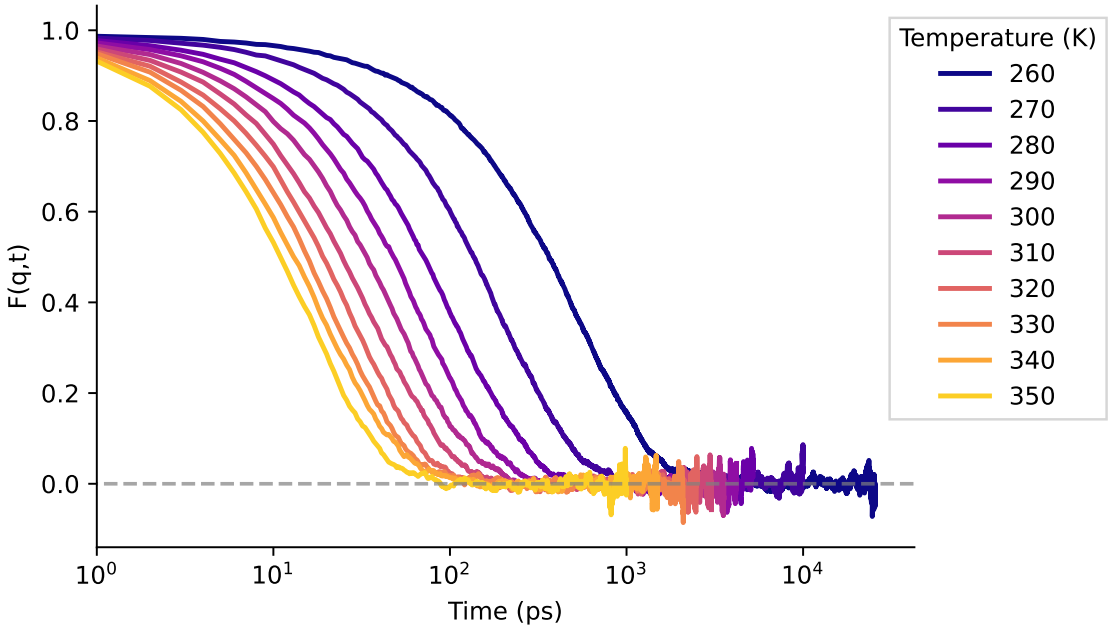

Figure 4.7: Intermediate scattering function $\mathrm{F}(\mathbf{q}, t)$ for oxygen atoms at various temperatures in *NPT* simulations at 1 bar. Each curve represents a distinct temperature from 260 K to 350 K in 10 K increments. The wavevector $q = 0.32\,\text{Å}^{-1}$ corresponds to density fluctuations on the scale of the full simulation box. Higher temperatures lead to faster relaxation dynamics, as reflected in the more rapid decay of the ISF.

improved accuracy arises despite both models being trained on the same dataset. Unlike ACE, which lacks message passing and is constrained by a shorter effective cutoff, MACE's message-passing architecture enables long-range correlations to be modelled more effectively. At high temperatures, where thermal motion disrupts long-range structure, the differences between models are less pronounced. However, at lower temperatures, where the development of extended hydrogen-bond networks becomes more important, MACE maintains physical realism more effectively, closely following the expected curvature and location of the TMD.

As summarised in Table 4.3, the MACE model slightly overestimates both the TMD and the corresponding density, with deviations of approximately 5%. These values fall well within the range reported in the literature. For example, Montero de Hijes *et al.* [140] found that even the best-performing functionals deviate by around 4% in density from experiment. The overestimation of the TMD is consistent with

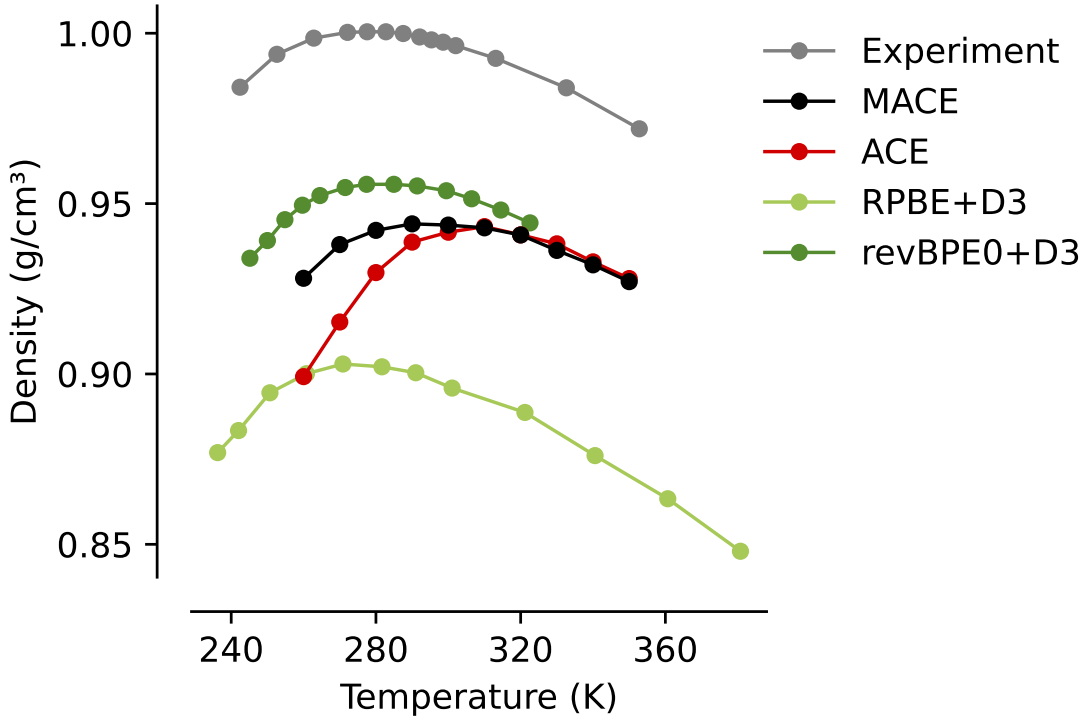

Figure 4.8: Density isobars at 1 bar across the temperature range 240–370 K. Results from the MACE model are shown alongside those from the ACE model, and machine-learned models trained on RPBE+D3 [109] and revPBE0+D3 [140] reference data. Experimental measurements [399] are included for comparison.

prior studies, which have noted similar trends across both AIMD and machine-learned models trained with various functionals [109, 133, 140, 165, 388].

Table 4.3: TMD and corresponding density at the TMD for the MACE and ACE models, as well as for models trained on RPBE+D3 [109] and revPBE0+D3 [140] functionals. Experimental values [399] are also included for reference. The TMDs for MACE and ACE were determined by Gaussian Process fitting to the density isobar.

Method	**TMD (K)**	**Density (g/cm^3)**
MACE	291	0.945
ACE	307	0.944
RPBE+D3 [109]	274	0.901
revPBE0+D3 [140]	280	0.958
Experimental [399]	277	1.00

Transport Properties

Figure 4.9a shows the temperature dependence of the self-diffusion coefficient of liquid water obtained from MACE simulations, compared with experimental values from Refs. 400–402. The diffusion coefficients were computed from the MSD using the Einstein relation (Section 4.3.3), and averaged over ten independent 100 ps *NVE* simulations at each temperature. The model captures the correct qualitative trend across the temperature range and shows good quantitative agreement with experiment. The MACE model tends to slightly underestimate the rate of diffusion, consistent with the minor overstructuring observed in the HBN. Overall, the close match to experimental data provides strong evidence that the model accurately captures the dynamical behaviour of water under ambient conditions.

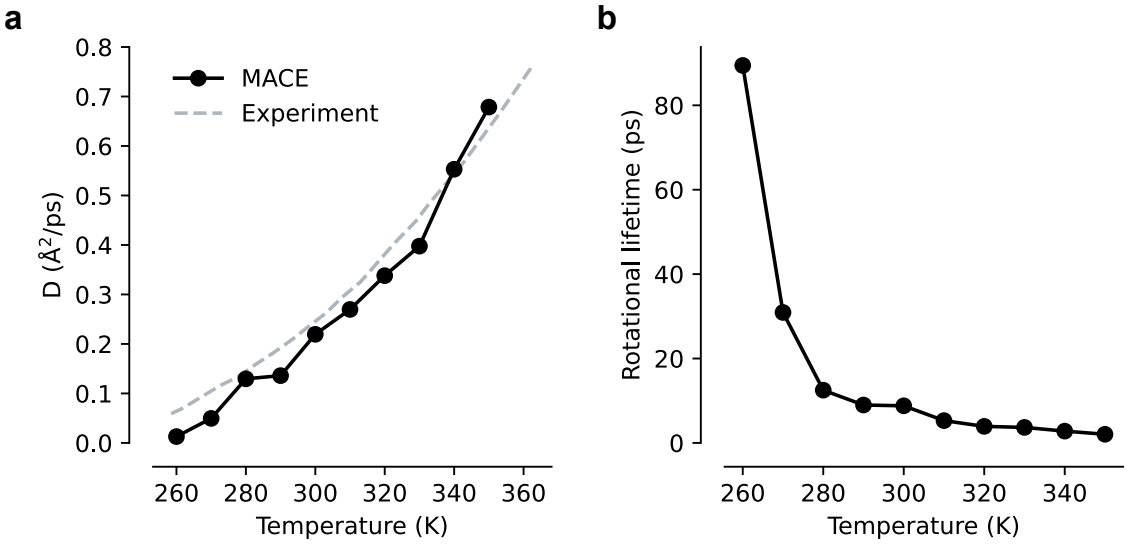

Figure 4.9: Temperature dependence of transport properties of liquid water predicted by the MACE model. (a) Self-diffusion coefficient compared with experimental data from Refs. [400–402]. (b) Rotational lifetime as a function of temperature. Experimental values are not shown in (b) due to variability in reported definitions and measurement techniques.

To assess rotational dynamics, I computed the temperature-dependent rotational lifetime of liquid water using the method described in Section 4.3.3. Figure 4.9b shows that the rotational lifetime decreases rapidly with increasing temperature, consistent with expectations for thermally activated orientational dynamics. Al-

though direct comparison to experiment is complicated by the diversity of experimental techniques and definitions used to extract rotational lifetimes, the observed trend is physically reasonable and reinforces the model's ability to capture key aspects of water's dynamical behaviour.

Orientational Tetrahedral Order Parameter

To conclude the structural validation of the MACE model, I evaluated the local tetrahedral ordering of water across the 260–350 K range using the orientational tetrahedral order parameter [378], q, defined in Section 4.3.2. This parameter ranges from 0 for completely disordered (gas-like) configurations to 1 for perfect tetrahedral geometry, and provides a sensitive measure of the degree of local structural order in the HBN, complementing other observables such as the RDF and ring statistics.

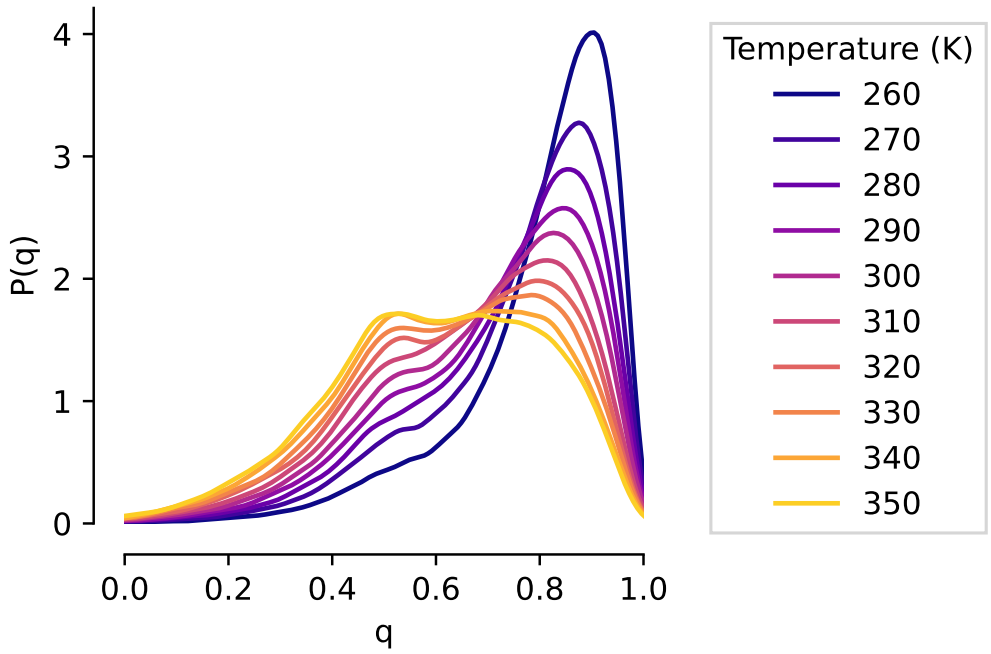

Figure 4.10: Kernel density estimates, $P(q)$, of the tetrahedral order parameter (q), ranging from 0 (ideal gas-like disorder) to 1 (perfect tetrahedral geometry).

As shown in Figure 4.10, the shape and position of the $P(q)$ distributions evolve systematically with temperature. At lower temperatures (260–280 K), the distributions are sharply peaked near high q values, reflecting a high degree of local

tetrahedral order characteristic of well-structured hydrogen-bond networks. As the temperature increases, the peaks broaden and shift toward lower q values, indicating a progressive loss of tetrahedral coordination due to enhanced thermal motion and structural disorder. This trend is consistent with experimental observations and prior simulation studies [378], highlighting the MACE model's ability to capture the temperature-dependent structural behaviour of liquid water.

4.4.3 Structure Factor of Liquid Water

With the structural validation of the MACE model complete and confidence established in its ability to reproduce key structural and thermodynamic properties of liquid water, I proceeded to replicate the structure factor of liquid water, with the aim of reproducing the characteristic peak splitting and the associated linear increase in separation observed at lower temperatures [311, 343, 344].

Figure 4.11a presents the calculated structure factor $S(q)$ for liquid water at two representative temperatures, obtained using the `Debye Calculator` package [403]. I compare my model predictions directly with experimental measurements from Ref. 343 and to simulations done with the ACE model from Ref. 346 trained on the same dataset as the MACE model. Structure factors for the whole temperature range are shown in Fig. B1 of the Appendix.

The MACE model reproduces the experimental structure factor with good fidelity, with some discrepancies observed in the low-q region indicative of a slight overstructuring of the HBN. This overstructuring is consistent with the overstructuring observed in the RDF and ring statistics analyses shown previously and likely originates from the ice-rich training set used. At high temperatures, the MACE and ACE models yield virtually indistinguishable predictions. However, at low temperatures, the MACE model demonstrates improved agreement with experiment relative to ACE, a trend that is also reflected in the density isobar predictions shown in Fig. 4.8.

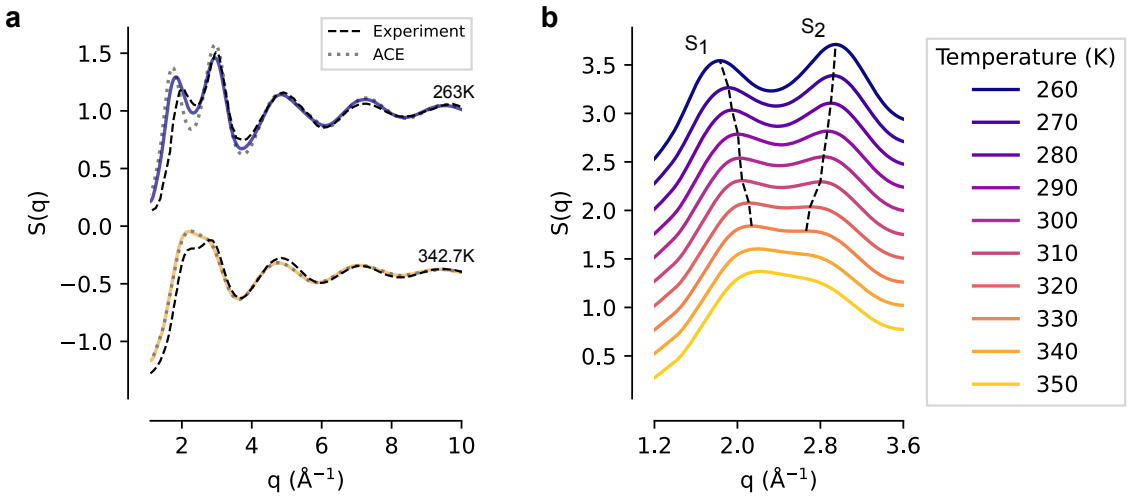

Figure 4.11: Peak splitting in the structure factor, $S(q)$, of liquid water at 1 bar (a) Comparison between simulated from MACE (solid lines), ACE (dotted lines) and experimental (dashed lines) structure factors at two representative temperatures. Experimental data taken from Ref. [343] with corresponding experimental temperatures indicated to the right of each curve. (b) Enlarged view of the low-q region (1.2–3.6 Å^{-1}), illustrating the temperature-dependent divergence of peaks S_1 and S_2. Their positions are extracted via GP fits. At temperatures above 330 K, the fit no longer resolves two distinct features.

Despite being trained on a limited set of liquid environments, the model accurately captures the structure factor of liquid water in the temperature range $T \in [260, 350]$ K. Crucially, the model reproduces the experimentally observed splitting of the principal diffraction peak into two distinct maxima upon cooling, indicating a decoupling of local structural signatures associated with two distinct local environments [311, 331, 343, 344].

A closer inspection of the low-q region in Fig. 4.11b reveals that the two peaks become increasingly separated below 330 K. Using a Gaussian Process (GP) fit to determine their positions, I observed a clear divergence in peak locations with decreasing temperature. This behaviour is quantitatively captured in Fig 4.12, which shows the temperature dependence of the distance between the two peaks, labelled S_1 and S_2. The model successfully replicates the experimentally observed linear increase in separation [311, 343, 404].

At temperatures above 330 K, thermal fluctuations dominate and obscure the structural distinctions that give rise to the peak splitting. As a result, the GP fit

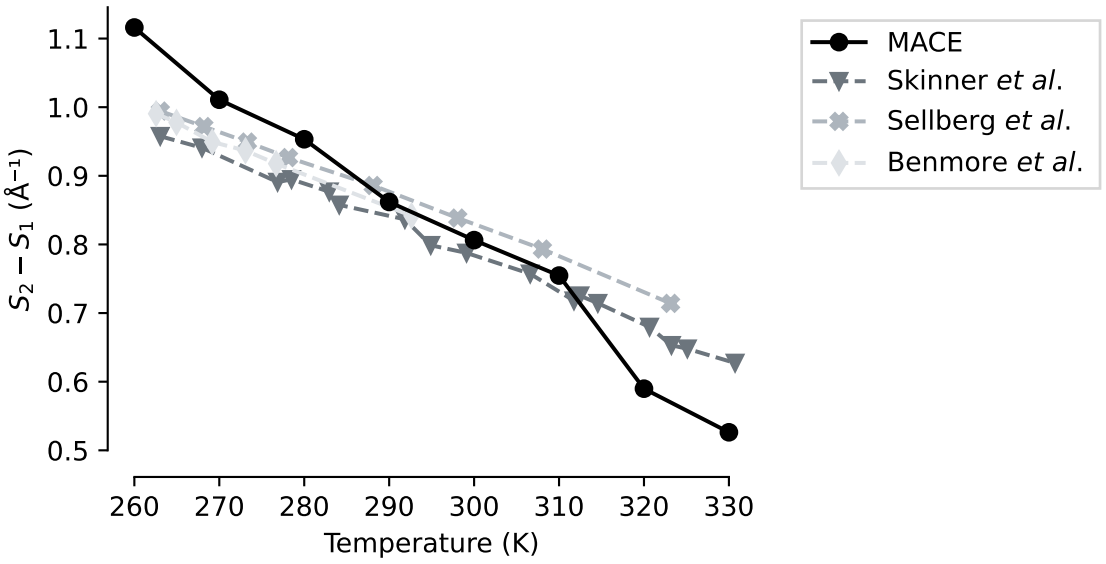

Figure 4.12: Temperature dependence of the peak separation $S_2 - S_1$ comparing the current MACE model (black circles) with experimental data from Skinner *et al.* [343] (triangles), Sellberg *et al.* [311] (squares), and Benmore *et al.* [404] (diamonds)

identifies only a single broad maximum, which reflects an average O$\cdots$O separation rather than distinct local environments.

Topological analysis of the hydrogen-bond network

Having established that the model accurately reproduces the structural properties of liquid water across across the temperature range $T \in [260, 350]$ K, I subsequently conducted a detailed structural analysis. This analysis lays the groundwork for interpreting the underlying network topology of the liquid. By identifying the dominant topological motifs at each temperature, I establish a hitherto unclear cause-and-effect relationship between network topology and structural properties.

Figure 4.13a presents how the distribution of hydrogen-bonded ring sizes in liquid water evolves with temperature. As the system cools, the distribution becomes more sharply peaked, with a pronounced preference for rings containing 5 to 7 molecules, consistent with their relative stability. In contrast, elevated temperatures result in broader distributions and reduced peak intensities, indicating that thermal fluctuations increasingly disrupt the HBN and facilitate the formation of larger, less

stable rings.

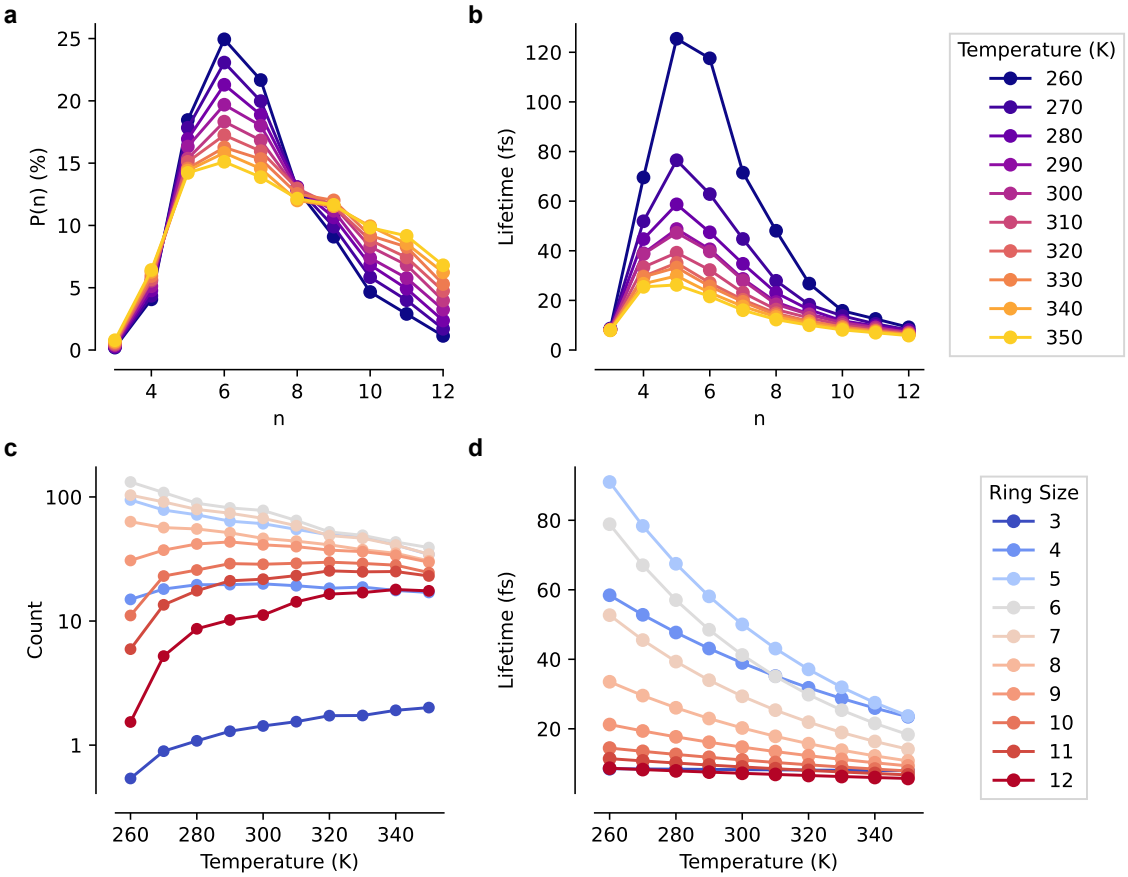

Figure 4.13: Hydrogen-bonded ring statistics in liquid water from 260 K to 350 K. (a) Probability, $P(n)$, of observing hydrogen-bonded rings composed of n water molecules where $n \in [3, 12]$. (b) Average lifetimes of rings of size n as a function of temperature. (c) Total count of rings of each size as a function of temperature. (d) Lifetimes of individual ring sizes as a function of temperature.

This broadening of the ring-size distribution is mirrored by a homogenisation in ring lifetimes (Fig. 4.13b). At 260 K, lifetimes vary significantly with ring sizes ranging from $\sim$ 10 fs for smaller or larger rings up to $\sim$ 120 fs for hexagonal rings, indicating some topological selectivity and local stability. By 330 K, however, these lifetimes converge to a narrow range of 10–30 fs, suggesting no particular ring topology remains particularly long-lived. At high temperatures, rings form and break rapidly, reflecting a highly transient network in which topological preferences are greatly reduced. Nonetheless, six-membered rings consistently remain the most probable across all temperatures, reflecting the underlying hexagonal organisation

characteristic of both cubic and hexagonal ice – water's crystalline ground states at ambient pressure. Interestingly, the population of 8-membered rings remains largely unchanged across the temperature range studied.

To quantify absolute changes in network connectivity, I also computed the total number of rings of each size (Fig. 4.13c). While the normalised distribution, $P(n)$, captures relative topological preferences, total ring counts reflect the overall connectivity of the HBN. As temperature decreases, the total number of rings rises sharply, indicating the progressive establishment of the HBN. This overall increase is primarily driven by the growing number of 5–8-membered rings, which dominate the distribution and increase substantially in absolute abundance. Concurrently, the population of 3-membered and large ($n \geq 9$) rings declines. The reduction in 3-membered rings reflects the annealing of highly strained, defect-like motifs that often arise from transient or disordered bonding configurations. As the system cools, thermal fluctuations diminish and the network undergoes a reorganisation toward more energetically favourable motifs. Similarly, the reduction of larger ring sizes signals suppression of structural disorder that is otherwise sustained by thermal noise at higher temperatures and a transition toward a more compact and tetrahedral network, characteristic of a well-developed HBN.

Figure 4.13b displays the corresponding lifetimes of these ring motifs. The HBN remains highly dynamic throughout, a necessary condition for the simulated structure factor to remain consistent with experimental observations [405]. Among the different motifs, five-membered rings exhibit the greatest stability, followed by six- and four-membered rings, across all temperatures. Although 6- and 5-membered rings display similar average lifetimes, the longer persistence of the latter may be attributed to statistical considerations: forming a pentagonal ring requires fewer constituent molecules than a hexagonal one, thereby reducing the number of potential disruption points. With fewer molecules involved, the likelihood of a random molecular displacement breaking the ring is diminished, lending greater stability to

smaller rings purely through probabilistic effects.

Notably, the lifetimes of both pentagonal and hexagonal rings increase rapidly upon cooling, becoming markedly more persistent at lower temperatures. This trend highlights their growing structural significance in supercooled regimes. In particular, pentagonal and hexagonal motifs emerge as the dominant long-lived units. This behaviour is significant: while hexagonal rings enhance local ordering and are associated with slower molecular dynamics, pentagonal rings introduce geometric frustration that inhibits crystallisation. At around 260 K, the near-equality of their lifetimes suggests a subtle competition between order-promoting and disorder-promoting effects that may play a central role in the anomalous dynamic and thermodynamic properties of supercooled water.

Tetragonal rings, despite being energetically less favourable, display lifetimes that rival or exceed those of hexagonal rings at temperatures down to 270 K. As further shown in Fig. 4.13d, the temperature dependence of tetragonal ring lifetimes is much weaker than that observed for five- and six-membered rings, indicating that their stability is less affected by thermal fluctuations. This unusual kinetic behaviour suggests the presence of additional stabilising mechanisms, such as local geometric constraints or strain-induced rigidity, that suppress rearrangements or dissolution of these motifs, counteracting their energetic instability and allowing them to persist longer than anticipated.

To further explore the influence of local topology, I analysed the immediate environment surrounding rings of selected sizes. Figure 4.14 presents the expected number of neighbouring rings of different sizes for central rings comprising 4, 5, or 6 water molecules, across the full temperature range examined. Additional data for other central ring sizes are included in Fig. B2 of the Appendix.

Rings of most sizes (from pentagon to hendecagon) display broadly similar neighbourhood distributions (see Fig. B2). These distributions consistently exhibit a preference for hexagonal neighbours, particularly at lower temperatures. This re-

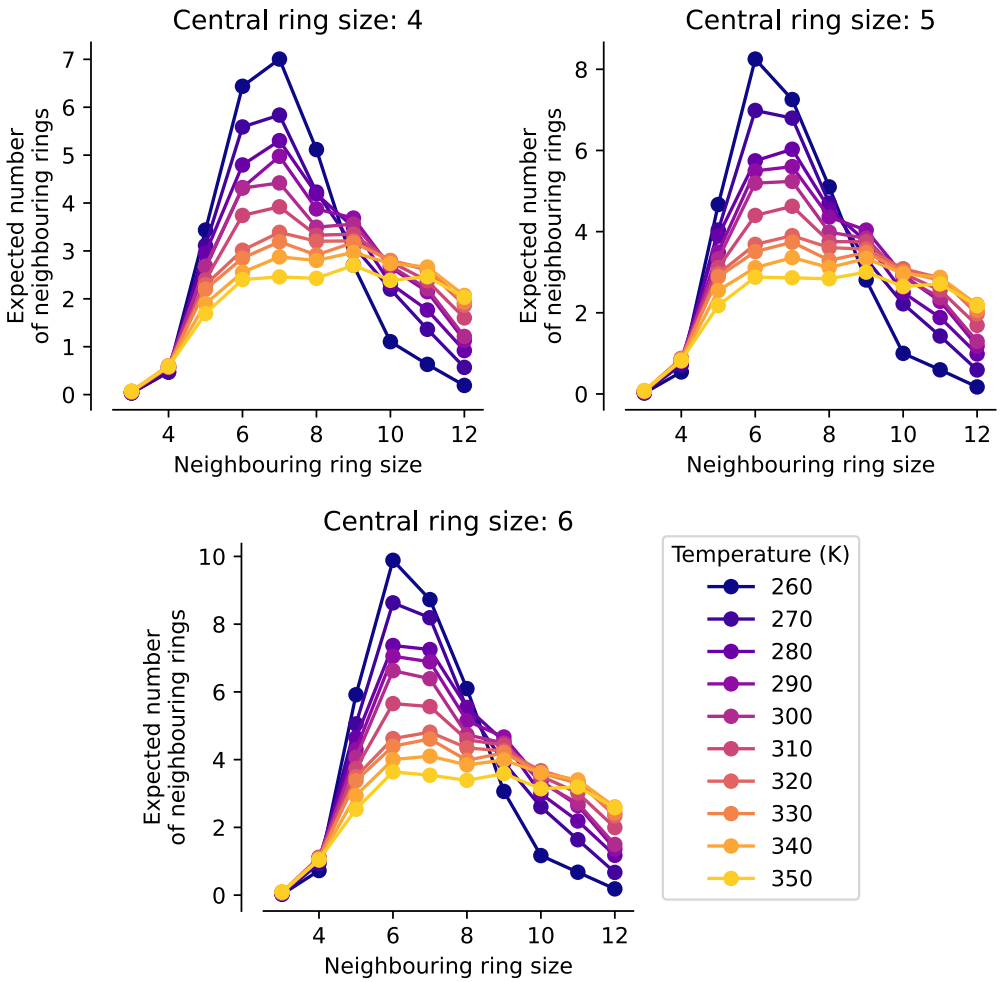

Figure 4.14: Expected number of neighbouring hydrogen-bonded rings given a central ring of size 4, 5, or 6, across the 260–350 K temperature range.

curring pattern reflects water's underlying inclination towards tetrahedral coordination, which is most naturally accommodated through hexagonal ring geometries in the HBN. The resemblance in neighbouring-ring profiles across a wide range of central ring sizes highlights the structural influence of this tetrahedral framework, suggesting a common local motif persists throughout the liquid regardless of the specific ring involved.

An interesting deviation to the general trend is observed for tetragonal rings, which exhibit a preference for heptagonal neighbours over hexagonal ones – an inversion of the trend observed for all other ring sizes, which, below $\sim 290\,\mathrm{K}$, favour hexagonal neighbours. Figures 4.15a and b illustrate this contrast, showing repre-

sentative configurations where a tetragonal ring connects to a heptagonal ring, and a pentagonal ring connects to a hexagonal one, respectively. This unusual embedding may explain the unexpectedly long lifetimes of tetragonal rings (Fig. 4.13b): their integration within stabilising hexagonal and heptagonal surroundings could inhibit reconfiguration despite their inherent energetic unfavourability. Moreover, the cooperative embedding between tetragonal and heptagonal rings could further stabilise these motifs, prolonging the persistence of tetragonal rings in the structure.

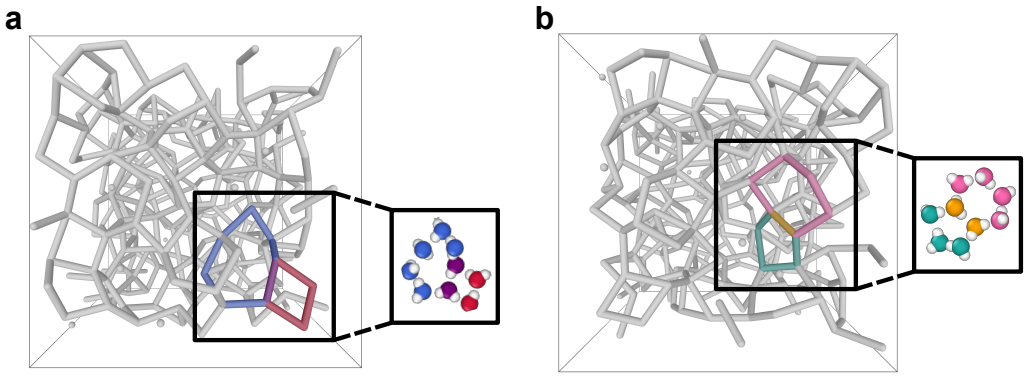

Figure 4.15: Illustrative snapshots of adjacent (a) 4- and 7-membered rings and (b) 5- and 6-membered rings embedded within the broader HBN (light grey) at 300 K. Hydrogens are omitted for clarity. Insets show the individual water molecules which make up the rings.

Structure Factor Decomposition

To investigate how persistent topological motifs influence experimentally accessible structural observables, I examined their contributions to the static structure factor. In particular, I considered whether the characteristic peak splitting observed in the total structure factor, $S(q)$, at low temperatures could be attributed to specific ring sizes, and whether the persistence of these motifs correlated with signatures of structural heterogeneity in the liquid. To this end, I adopted the methodology introduced by Zhou *et al.* [345], which enables the decomposition of the structure factor according to ring size. For each ring size n, I identified the atoms forming n-membered rings, excluded all other atoms from the configuration, and computed the corresponding partial structure factor $S_n(q)$ for the resulting substructure.

It is important to emphasise that this ring-wise decomposition results in varying number of atoms per structure. As a consequence, the absolute intensities of the partial structure factors $S_n(q)$ lack direct physical significance and cannot be quantitatively compared to the total $S(q)$. The meaningful quantity in this context is the position of the peaks in each $S_n(q)$, which reflects the dominant intermolecular length scales associated with each ring size. The analysis therefore focuses on these peak positions, and in particular, on the emergence and persistence of peak splitting, as indicators of the structural roles played by specific ring topologies.

Figure 4.16 shows the resulting decompositions at the two temperature extremes, 260 K and 330 K, i.e., where the full structure factor still exhibits clearly resolved bimodal peaks. Decompositions at intermediate temperatures are shown in Fig. B3 of the Appendix.

At 260 K, the structure factors associated with pentagonal, hexagonal, and heptagonal rings most closely resemble the full $S(q)$, indicating these motifs dominate the local structural organisation at low temperature. Their substantial contributions are consistent with their high relative populations (Fig. 4.13a), long lifetimes (Fig. 4.13b), and elevated absolute counts (Fig. 4.13c). This convergence of both stability and frequency make it such that these topologies are central to the underlying tetrahedral network at lower temperatures.

Larger rings, such as 8- and 9-membered motifs, also exhibit bimodality in $S_n(q)$, suggesting that they too participate in the structural heterogeneity underlying the observed peak splitting. In contrast, 3- and 4-membered rings contribute little to the overall peak structure. 10-, 11-, and 12-membered motifs exhibit a prominent shoulder near the high-q peak, S_2, suggesting their association with denser, interstitial-rich structures.

At 330 K, the decomposition of the structure factor reveals that the contributions from individual ring sizes become increasingly indistinct, with most $S_n(q)$ curves closely resembling the total $S(q)$. This loss of differentiation coincides with a

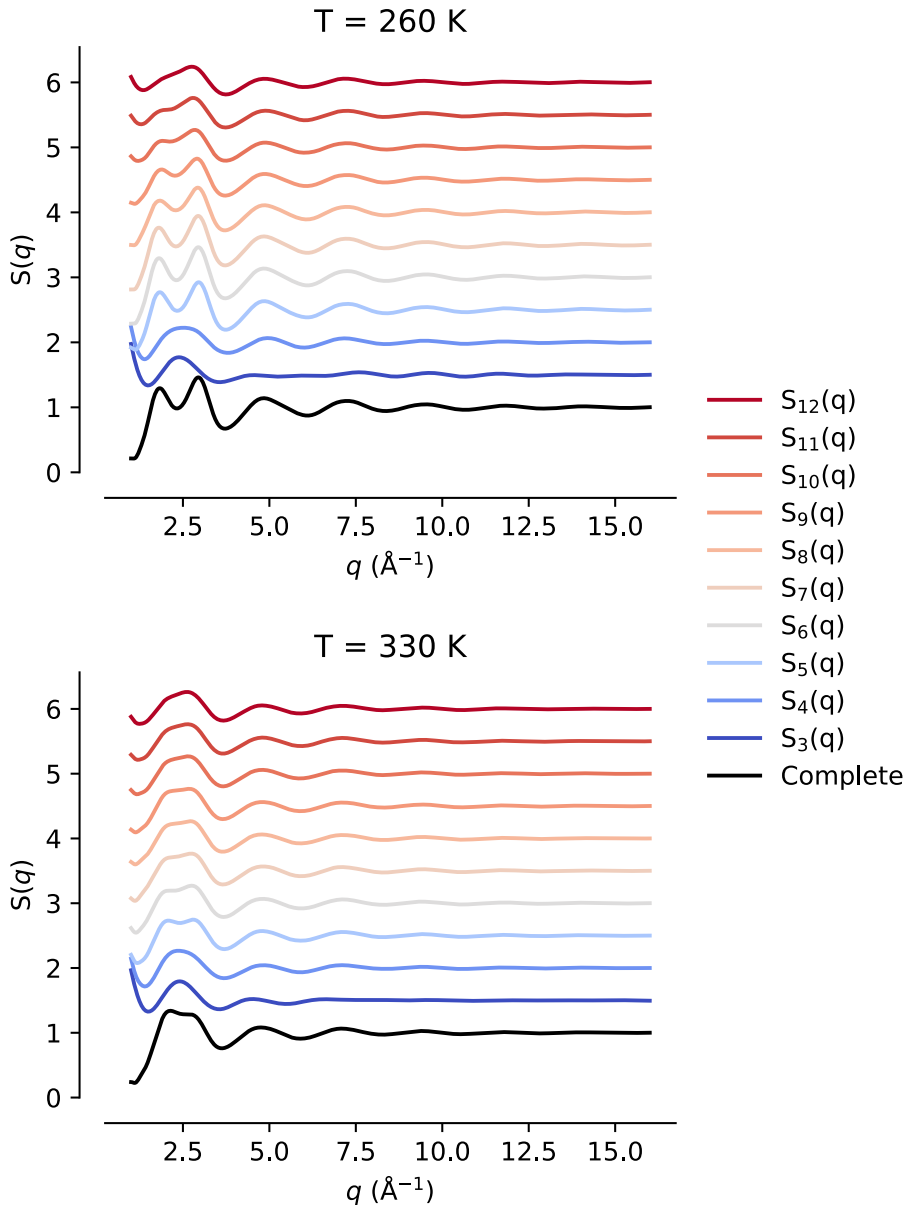

Figure 4.16: Structure factor $S(q)$ of liquid water at 1 bar and at two temperatures (260 K and 330 K) decomposed into contributions from individual ring sizes, $S_n(q)$ where $n \in [3, 12]$. 330 K is the last temperature at which the splitting of the first peak into two distinct maxima was resolved. The black curves indicate the total original structure factor, while coloured lines represent individual ring contributions.

broadening and flattening of the ring-size distribution (Fig. 4.13a), indicating a diminished preference for specific topological motifs and a shift towards increased structural disorder. A wider range of ring sizes becomes comparably probable, reflecting the absence of dominant local geometries. This structural homogenisa-

tion is further supported by the convergence of ring lifetimes across different sizes (Fig. 4.13b), indicating a more transient network in which rings form and break rapidly, and no particular topology persists long enough to influence the liquid's organisation significantly.

These microscopic structural changes align with macroscopic dynamical trends. Between 260 and 330 K, the self-diffusion coefficient increases by a factor of four (Fig. 4.9a), while the rotational correlation time decreases sharply from over 80 ps to less than 10 ps (Fig. 4.9b). Since ring structures can be disrupted by both translational and rotational motion, the combination of accelerated dynamics in both modes amplifies the breakdown of persistent ring topologies. Consequently, the decomposition of $S(q)$ becomes increasingly featureless, echoing the breakdown of well-defined local motifs and the emergence of topological disorder.

Finally, Fig. 4.17 presents the separation between the two resolved peaks, S_1 and S_2, in each ring-specific structure factor, $S_n(q)$. This analysis explicitly identifies which ring topologies contribute most significantly to the divergence in peak positions, thereby revealing the ring sizes that most clearly reflect the coexistence of low- and high-density local environments in liquid water.

Only rings containing 5 to 10 water molecules display peak separations large enough to be recognised as two distinct maxima. Among these, pentagonal rings stand out as exhibiting the most persistent splitting; their partial structure factor retains both peaks up to at least 350 K, while the total structure factor loses this resolution above 330 K. This suggests that the local ordering within pentagonal rings remains distinctly decoupled over a broad temperature range, even when such bimodal behaviour is obscured in the total structure factor.

This finding aligns with ongoing discussions in the literature concerning the dual role of pentagonal rings within water's HBN [370]. Pentagonal rings are understood to introduce geometric frustration, thereby impeding the formation of long-range crystalline order. At the same time, they contribute to the structural diversity and

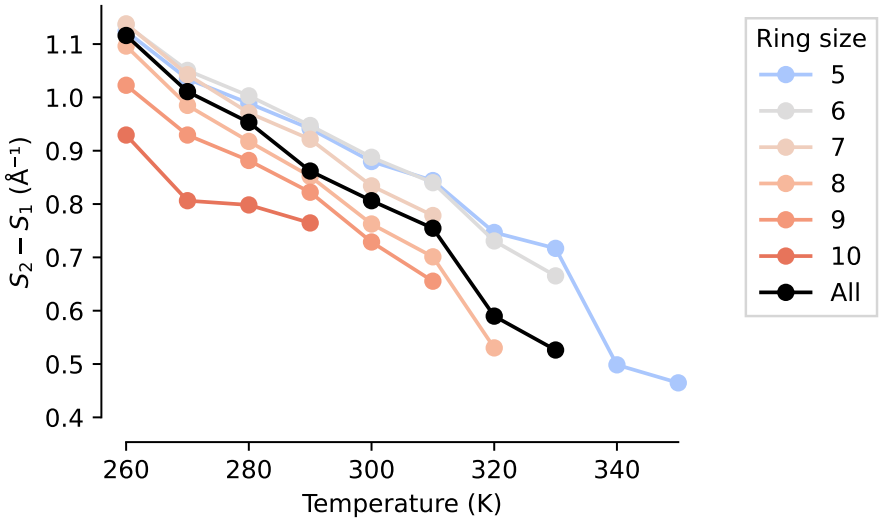

Figure 4.17: Distance between the two resolved maxima, S_1 and S_2, in the structure factor of water as a function of temperature for different ring sizes. The black line represents the total structure factor as shown in Fig. 4.12

thermodynamic stability of the liquid phase.

The persistence of peak splitting in the structure factor associated with pentagonal rings, even at elevated temperatures, indicates that these structures retain a distinct form of local ordering despite increasing network disorder. This behaviour highlights the central role of pentagonal motifs in shaping water's anomalous properties, particularly in the supercooled regime where the interplay between order-promoting and disorder-promoting structural features becomes increasingly pronounced.

As the ring size increases, the distinction between the characteristic length scales associated with locally ordered and disordered environments becomes less evident. This results in the gradual merging of the two peaks into a single resolved feature, even at relatively low temperatures.

4.5 Conclusion and Outlook

The anomalous properties of liquid water have been the subject of extensive investigation, with the splitting of the principal peak in the static structure factor, $S(q)$, recognised as a key experimental indicator of its underlying structural com-

plexity. In this work, I build upon earlier interpretations [311, 331, 332, 343, 344], and demonstrate that the observed peak splitting in $S(q)$ originates from specific medium-range topological features within the network. Notably, pentagonal rings emerge as persistent contributors to this phenomenon, underscoring their critical role in maintaining structural heterogeneity through geometric frustration. This supports the notion of a delicate interplay between locally ordered (hexagonal) and disordered (pentagonal) configurations in supercooled water [370]. Furthermore, the surprising stability of tetragonal rings – especially their tendency to occur adjacent to heptagonal ones – suggests a more complex and cooperative network architecture, wherein even topologically strained motifs can be locally stabilised by their surroundings.

These findings are consistent with recent studies on other disordered materials. For instance, Nicholas *et al.* reported that variations in ring-size distributions systematically account for differences in intermediate-range order across a range of amorphous materials, including zeolitic imidazolate frameworks, silicon, and silica. While their approach involved a formal topological descriptor to quantify structural diversity, the ring-resolved decomposition of $S(q)$ presented here offers a complementary perspective by linking persistent network motifs directly to experimentally accessible scattering features. Similarly, studies by Zhou *et al.* [345] and Tavanti *et al.* [406] have shown that changes in ring statistics correlate with diffraction features in silicate and chalcogenide glasses. By directly associating network topology with structural heterogeneity, the present analysis contributes to this broader effort to interpret disordered systems and provides a transferable framework for probing the structure of complex liquids and glasses.

This methodology may be extended in future studies to biological and soft-matter systems, where water plays a vital role in mediating structural and dynamical processes such as protein folding [407, 408], enzymatic function [409], and membrane organisation [398]. Moreover, examining how external conditions, such as applied

pressure, influence topological motifs and their connection to bulk properties could also be a promising direction for further research.

Chapter 5

Coarse-graining as a bridge between ZIFs and zeolites

5.1 Acknowledgements

The work presented in this chapter has been published in *Chemical Communications* [322]. Portions of the text and several figures have been reused; where appropriate, figures are labelled as "Adapted".

This research builds upon a preliminary study conducted during my MChem Part II project, in which the concept of coarse-graining ZIFs was first introduced. The MChem thesis focused on analysing an existing ZIF database and exploring various coarse-graining strategies of the imidazolate ring. The present work represents a substantial advancement beyond that initial study. It is based on an entirely new dataset, and all machine learning models, optimisation procedures, hyperparameter searches, and analyses have been developed independently as part of my PhD research. No material from the MChem thesis is reused in this chapter; all work presented here is original and was carried out during the PhD.

I would like to thank Thomas Nicholas for providing the dataset used in this chapter and for valuable discussions. I am also grateful to John Gardner for his insightful input and assistance with the code.

5.2 Introduction

ZIFs [68, 410–412], a subclass of MOFs, have attracted considerable attention for their potential in applications ranging from gas storage and separation [187, 413] to catalysis [414, 415] and next-generation batteries [416, 417] (see Section 1.2.2).

ZIFs uniquely combine characteristics of both zeolites and MOFs: they offer the

high surface areas and tunability typical of MOFs, while retaining the exceptional thermal and chemical stability associated with zeolites [68]. This duality arises from a combination of factors: the use of extended organic linkers enables large pore volumes, while strong Zn–N covalent bonds contribute to their robustness. Underpinning these properties is a tetrahedral network topology analogous to the SiO_4 units found in silicates, which organises the local coordination environment and governs the extended framework geometry, whether in crystalline, glassy [69, 418–420], or even liquid forms [421].

Structurally, ZIFs are constructed from Zn^{2+} cations and imidazolate anions, forming AB_2 tetrahedral networks that mirror the vertex-sharing $[SiO_4]_{Td}$ connectivity in zeolites (Fig. 5.1) [68]. This conceptual mapping to zeolites has guided both the rational design and synthetic targeting of ZIFs by enabling the replication of known zeolite topologies using metal–imidazolate building units [181, 182]. It has also informed structural classification and provided a framework for understanding their porosity and stability. However, the extent to which this analogy holds quantitatively remains an open question – it is yet unclear whether the energy landscape of ZIFs can be quantified without a fully atomistic description, and whether established stability trends in zeolites [422] map onto hybrid ZIF phases, especially given that the two material classes access different crystal topologies.

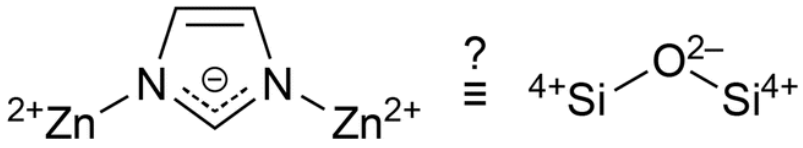

Figure 5.1: Schematic comparison between the chemically simplest ZIF material $Zn(Im)_2$ (left) and the Si–O–Si bridge in silicates (right). The structural analogy forms the basis for describing ZIFs as a hybrid inorganic-organic analogue of the SiO_2 framework in zeolites. Figure adapted from Ref. [322].

One powerful way to interrogate this conceptual mapping is through coarse-graining (cg), a modelling technique in which complex atomic structures are reduced to simplified representative units or "beads" (see Section 5.3.1). This technique,

common in biomolecular simulations [423, 424], facilitates computational efficiency by reducing structural resolution. Coarse-graining has also been applied to disordered networks, such as amorphous calcium carbonate [425], to rationalise interactions at larger scales. In the context of MOFs, cg models have been developed to describe mechanical properties, pore dynamics, and thermodynamic stability under different conditions [426, 427].

In the context of ZIFs, coarse-graining naturally highlights their underlying tetrahedral architecture. By representing metal nodes and linkers as A and B beads in AB_2 networks, it becomes possible to ask whether the tetrahedral scaffold alone suffices to capture the local energetic landscape of these materials. Previous studies suggest that cg-ML models can uncover structural relationships between ZIFs and inorganic AB_2 networks [63, 428], but a quantitative benchmark is still lacking.

In this chapter, I investigate whether tetrahedral AB_2 coarse-grained representations of ZIFs can be used to construct predictive machine-learning models of local energetics. Using GPR, I assess the performance of different levels of coarse-graining – from fully atomistic, to AB_2, to A-only models that omit AB_2 tetrahedral information entirely. These models are trained on local energy targets derived from a classical force field and evaluated across a systematically constructed, structurally diverse dataset. The results demonstrate that maintaining tetrahedral connectivity is critical: it enables accurate energy predictions even in highly reduced representations, whereas discarding it leads to a sharp decline in performance. These results provide quantitative validation of the long-standing analogy between ZIFs and zeolites, demonstrating that their shared tetrahedral topology carries predictive value. More broadly, this work contributes to the development of coarse-grained force fields [429–434] and advances our understanding of hybrid framework materials.

5.3 Methods

5.3.1 Coarse-graining

Coarse-graining is a modelling strategy that simplifies a material's atomistic structure by replacing groups of atoms with single representative particles, or "beads". The process follows a mapping scheme, which defines how atoms are grouped into beads [435]. The number of atoms per bead determines the degree of coarse-graining, with coarser mappings reducing resolution but improving computational efficiency. While mapping schemes can be flexible, they should preserve the key physical and chemical characteristics of the system [436]; choices are often tailored to the specific properties or behaviours under investigation [437, 438].

In MOFs, a natural approach is to represent each rigid or semi-rigid building unit as a bead. For instance, Dürholt *et al.* [426] constructed a maximally coarse-grained model of HKUST-1 by representing each copper paddlewheel (a rigid inorganic cluster) as a single bead. In ZIF-8, where Zn^{2+} nodes are linked by methylimidazolate ligands in a sodalite-type structure, several mapping options exist. Alvares *et al.* [427] explored models with varying resolution, ranging from merging a Zn node and its coordinating imidazolate into one bead, to finer models where nodes and linkers were treated separately.

A primary advantage of cg models is the reduction in the number of particles and their interactions, leading to significant computational savings, often scaling with the square of the particle reduction [436]. This efficiency enables faster MD simulations, while still reproducing key properties of atomistic models. Moreover, cg models provide smoother potential energy surfaces and simplified interaction potentials, allowing for larger integration time steps (typically 5–20 fs) [439–441].

In this chapter, coarse-graining was performed by replacing each zinc atom with an "A"-bead and each imidazolate molecule with a "B"-bead (Fig. 5.2). The bead positions are assigned using the geometric centroids of the constituent atoms (with-

out mass-weighting), a straightforward and widely used method in both crystal net analysis [442] and coarse-grained MD [436]. This choice offers a practical compromise between computational simplicity and physical accuracy: it preserves the topology and connectivity of the original framework, enables efficient modelling, and faithfully reproduces the Zn–Im–Zn bond angle, which maps closely to the Si–O–Si angle in analogous silicate networks.

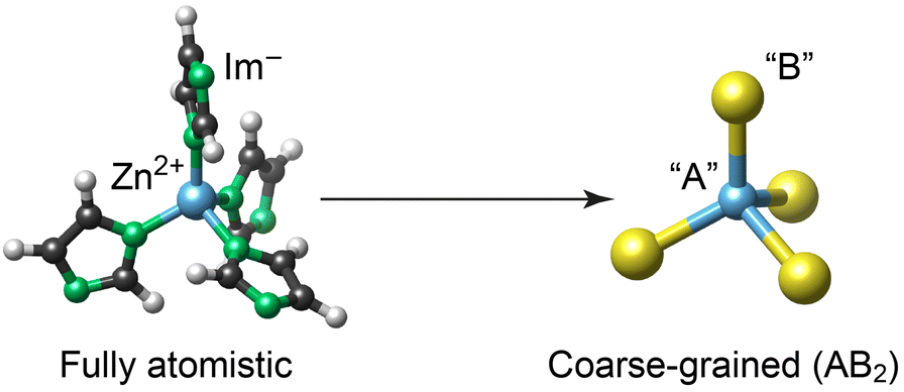

Figure 5.2: Illustration of the coarse-graining approach applied to a tetrahedral ZIF unit cell. Left: fully atomistic representation of a Zn^{2+} cation coordinated by imidazolate (Im^{-}) linkers. Right: coarse-grained (AB_2-type) model where the Zn^{2+} node is represented as a central "A" bead and the imidazolate linkers as peripheral "B" beads. Schematic generated using VESTA [443]. Figure adapted from Ref. [322].

All structural coarse-graining described in this chapter was carried out using the `CHIC` package [444].

5.3.2 Data Generation

The data set for this chapter was developed by Thomas Nicholas. A summary of the generation process is outlined below; for a more detailed description the reader is referred to https://github.com/tcnicholas/hZIF-data.

The initial dataset was assembled from two sources: a curated set of coarse-grained AB_2 MOFs with unique topologies, compositions, and space groups, and idealised zeolite frameworks approved by the Structure Commission of the International Zeolite Association (IZA-SC) [445]. These structures provided the template networks for further decoration.

Each template was then "decorated" by substituting A and B sites with zinc atoms and imidazolate molecules, respectively. The imidazolate was positioned such that its ring centre aligned with the original B site, the ring plane was parallel to the B–A bond vectors, and the molecule's C_2 axis bisected the bond angle. After decoration, an energy minimisation was performed using the MOF-FF for ZIFs empirical force field [446]. This model evaluates the total system energy based on contributions from bond stretching, angle bending, torsional rotations, van der Waals interactions, and electrostatics. Although the force field was initially parametrised for ZIF-8, its formulation includes explicit bonding terms, which help guide structures toward chemically sensible geometries during relaxation, even for hypothetical frameworks, by avoiding atomic overlaps and unphysical bond distortions. As such, MOF-FF serves as a robust reference approach for capturing the underlying potential energy surface across a range of network topologies.

The final atomistic structures were then obtained using three different procedures in order to sample a broader region of the potential energy surface. All began with coarse-graining the energy-minimised structures using the `CHIC` package, followed by re-decoration with idealised imidazolate molecules.

In the first procedure, no further modifications were made. In the second, the coarse-grained structures were re-decorated and random perturbations were applied to the atomic positions and cell parameters. These distortions (commonly referred to as "rattling") slightly displace atoms from their local-minimum configurations, a strategy commonly employed to improve the generalisability of machine-learned potentials by exposing them to a broader region of configuration space [274, 447, 448]. In the third procedure, rattling was applied to the cg configuration, followed by re-decoration. For procedures involving rattling, three perturbation magnitudes were used (small, medium, and large). Five independent batches were generated by systematically varying the random seed. Finally, the structures were labelled using MOF-FF for ZIFs empirical force field [446] and coarse-grained using `CHIC` to the

varying levels of abstraction.

Atomic energies were mapped to the coarse-grained structures by summing constituent atoms' energies for each site. In this case, the A site energy equals the zinc atom energy, and the B site energy equals the sum of the imidazolate molecules atoms' energies. See the next section for more detail.

To validate the use of the MOF-FF potential in this work, Figure 5.3 shows that the majority of bond lengths and bond angles in the dataset fall well within the parameter ranges over which the MOF-FF model was originally tested [446]. While a small number of configurations extend slightly beyond this validated range, these outliers are infrequent and are not expected to significantly affect the overall accuracy of the energy predictions.

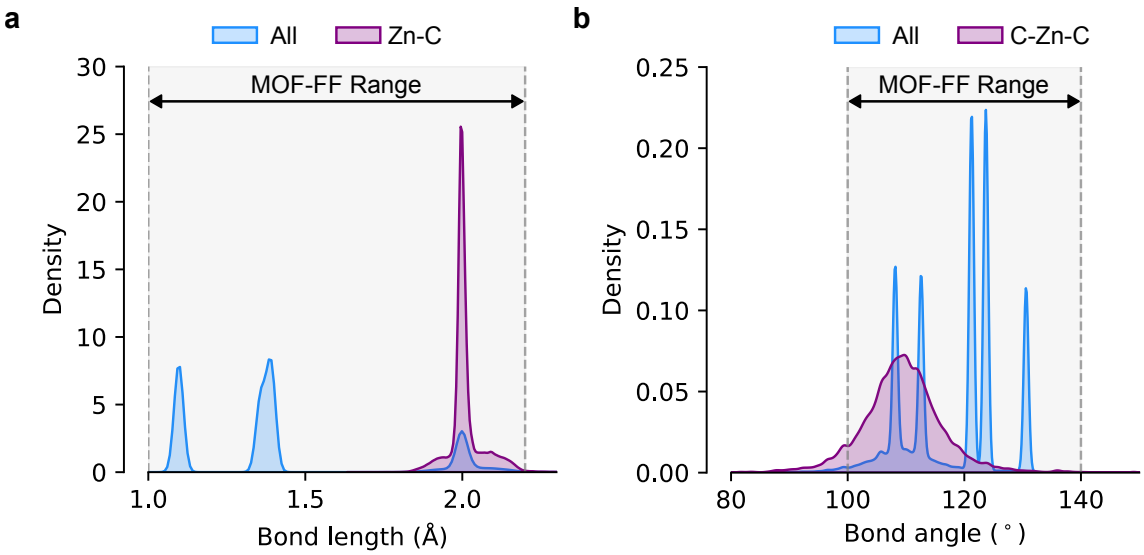

Figure 5.3: Distribution of structural properties in the database used in this chapter compared to the range tested for the MOF-FF model (light grey shading) in Ref. [446]. (a) Distribution of bond lengths. (b) Distribution of bond angles. Blue curves represent the distribution for all bonds or angles, while purple curves show only those involving Zn^{2+} (Zn–C bonds and C–Zn–C angles, respectively). Figure adapted from Ref. [322].

5.3.3 Learning Local Energies

To investigate the local chemical environments in ZIFs, ML models were trained to predict the local energies of Zn^{2+} cations. This analysis leveraged the MOF-

FF force field, which provides not only total (per-cell) energies but also atom-wise energy contributions.

Although these local energies are not quantum-mechanical observables and lack the physical rigour of DFT total energies, they serve as useful empirical proxies. Specifically, they arise from force-field-based decompositions of the total system energy into atom-centred components:

$$E = \sum_{i=1}^{N} e_{\text{local}}^{(i)}, \tag{5.1}$$

where $e_{\text{local}}^{(i)}$ denotes the local energy attributed to atom i in a structure with N atoms. While this partitioning is not guaranteed to be unique, it supports the development of linear-scaling ML models and aligns with the notion that atomic-level energies can encode chemically meaningful information [449]. Recent studies have further explored the transferability and reliability of such localised predictions, particularly regarding how local errors affect global properties [450].

In this chapter, the local energy of each Zn^{2+} cation in a ZIF is defined as the sum of its individual atomic energy and half the energy contributions from each of its coordinated imidazolate (Im^-) linkers. The energy of an Im^- molecule is calculated by summing the local energies of its constituent atoms (C, N, and H), and then equally dividing that total between the two Zn^{2+} centres it bridges. The resulting expression is:

$$e_{\text{local}}^{(i)} = e_{\text{Zn}}^{(i)} + \frac{1}{2} \sum_{j=1}^{4} e_{\text{Im}}^{(j)}, \tag{5.2}$$

where $e_{\text{Zn}}^{(i)}$ is the energy of the i-th Zn^{2+} atom, and $e_{\text{Im}}^{(j)}$ is the energy of the j-th imidazolate linker attached to it. This decomposition ensures that summing all $e_{\text{local}}^{(i)}$ across a unit cell recovers the total energy, and mirrors the locality assumptions used in the development of ML interatomic potentials. Importantly, by incorporating both the Zn^{2+} atomic energy and contributions from its coordinated imidazolate linkers, this definition is more sensitive to variations in the local chemical environ-

ment than using Zn^{2+} energies alone. This increased sensitivity is especially valuable for capturing the effects of structural coarse-graining.

5.4 Results

5.4.1 Proof of concept

I began by investigating how structural resolution influences the performance of GPR models (see Section 2.2.3), with a particular interest in whether the energetics of ZIFs can be predicted using only their underlying tetrahedral connectivity.

To this end, I trained GPR models using three structural representations: a fully atomistic model; a cg AB_2 model (see Section 5.3.1); and a more extreme cg A-only model that omits linker atoms entirely, thereby eliminating explicit AB_2 tetrahedral connectivity. This A-only model retains only the Zn positions, providing no chemical context of the local neighbourhood and making it a stringent test of how much structural detail can be removed before predictive accuracy collapses.

Figure 5.4 presents parity plots comparing the predictive accuracy of these three models. For each representation, the SOAP descriptors were calculated using a cutoff radius of 6.75, 7.00, and 9.75 Å and a Gaussian smearing width of 0.30, 0.25, and 0.90 Å for the fully atomistic, AB_2, and A-only models respectively. The models were trained on 32,000 atomic environments and evaluated using 5-fold cross-validation.

As expected, the fully atomistic model achieves the highest predictive accuracy, with an R^2 value of 0.93. This indicates that the atomistic representation successfully encodes the key structural features influencing local energy contributions. The AB_2 model maintains a solid performance, with an R^2 of 0.85, representing a modest drop in accuracy of less than 9%, despite the reduced structural detail.

In contrast, the A-only model shows a dramatic decline in performance, with an R^2 value of just 0.49, a 47% reduction relative to the fully atomistic baseline. This sharp decline highlights the importance of preserving the tetrahedral connectivity and local chemical context. While coarse-graining can still yield accurate predictions

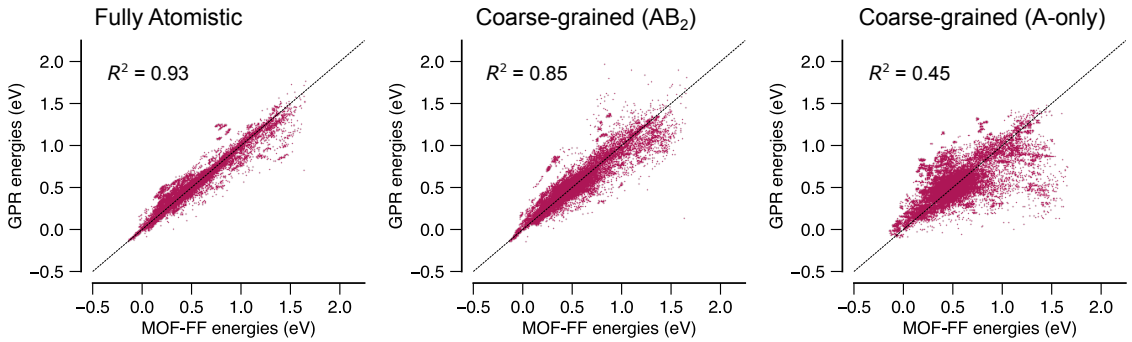

Figure 5.4: Scatter plots of local-environment energies as defined in Eq. 5.2 (the "ground-truth" learning target) on the horizontal axis, and the GPR ML predictions on the vertical axis. The values were obtained by 5-fold cross-validation. From left to right, the models are based on: a fully atomistic description; a cg description where the linker molecules are described by single "B" beads (Fig. 5.2); and a more aggressively coarse-grained model where only A-site species are represented. All models were trained on 32,000 atomic environments and evaluated using 5-fold cross-validation.

when the essential topological features are maintained, overly aggressive simplification – such as omitting the underlying framework – results in significant information loss that severely degrades model performance.

Figure 5.5 further supports these findings by showing learning curves for each model. All three reach convergence with respect to training data size implying that residual errors likely stem from the inherent locality assumptions of the SOAP featurisation and the effects of regularisation, rather than from insufficient data. The atomistic model achieves an RMSE of $\sim$0.047 eV, corresponding to the commonly accepted threshold for chemical accuracy ($\sim$1 kcal/mol) [451], while the AB_2 model converges at $\sim$0.071 eV. This represents a moderate increase in prediction error, roughly a factor of 1.5 higher than the atomistic model, but still within acceptable limits for many applications. The A-only model, however, plateaus at $\sim$0.128 eV, nearly 3x higher than the atomistic model. Its early saturation suggests that this representation fundamentally lacks the structural context required for high-fidelity predictions.

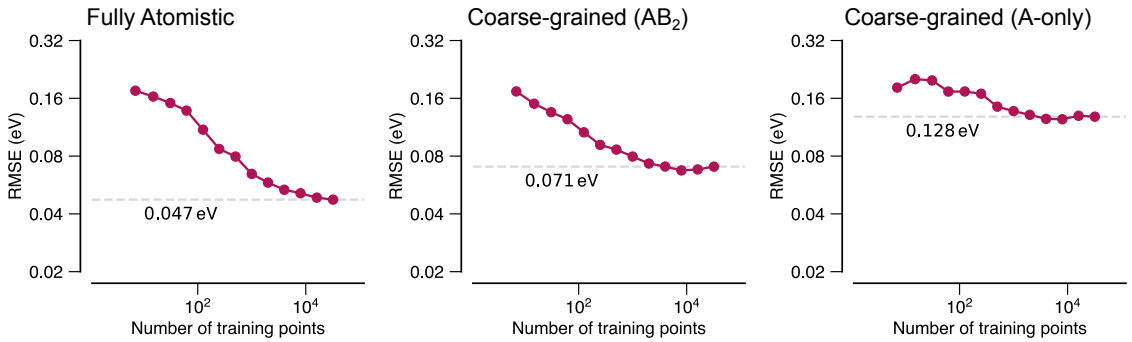

Figure 5.5: Learning curves showing the evolution of the RMSE as a function of training set size for the fully atomistic, AB_2 and A-only models. The RMSE for the largest number of training points is indicated by a dashed grey line in each panel.

5.4.2 Varying Model Hyperparameters

Following the initial validation of the GPR models, the dependence of model performance on structural descriptor hyperparameters was systematically investigated. In particular, two central parameters governing the behaviour of the SOAP descriptor [230] were examined: the cut-off radius (r_{cut}) and the smoothness of the atomic neighbour density (σ_{at}). The cut-off radius determines the spatial locality of the information used to describe atomic environments, while the smoothness parameter controls the resolution of the atomic density (for more details see Section 2.2.2).

The regularisation strength was also tuned and found to have only a weak influence on model performance; a fixed value of 0.2 eV was used consistently across all models for simplicity, as the focus was on the more impactful effects of r_{cut} and σ_{at}, which are directly tied to the degree of coarse-graining.

Although the results in Fig. 5.4 were obtained using individually optimised hyperparameters for each model type, Fig. 5.6 presents the outcome of a broader grid search across these hyperparameters to more comprehensively map the relationship between model error and SOAP descriptor settings. Specifically, the search spanned values up to 15 Å for r_{cut} and up to 2 Å for σ_{at}, thereby covering a substantially wider parameter space than typically employed in conventional SOAP-based machine learning potentials (e.g., $r_{\mathrm{cut}} \sim 5$ Å, $\sigma_{\mathrm{at}} \sim 0.5$ Å) [247].

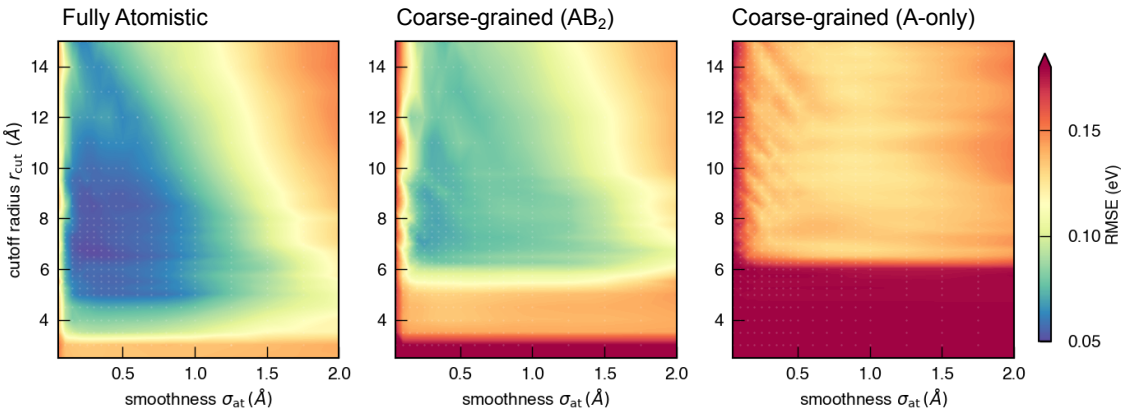

Figure 5.6: Survey of the hyperparameter space for fully atomistic GPR models versus cg-GPR models. The two decisive choices are the cut-off radius (vertical axis) and the smoothness of the atomic neighbour density (horizontal axis). The results of a grid search are given by colour coding, with individual grid points highlighted by small white markers. For these scans, a more economical setting of $N = 10,000$ training points, compared to $N = 32,000$ otherwise.

The heatmaps in Fig. 5.6 are consistent with the performance trends seen in Fig. 5.4 and Fig. 5.5: the fully atomistic model achieves the lowest RMSEs, followed by the AB_2 model, with the A-only model performing substantially worse across the entire hyperparameter space. Even under optimal settings, the A-only model retains a high error, reaffirming that omitting linker information imposes an intrinsic limitation that no choice of r_{cut} or σ_{at} can fully overcome.

Interestingly, the loss landscapes become progressively flatter as the model becomes more coarse-grained, indicating reduced sensitivity to hyperparameter tuning. This trend reflects the growing effective interatomic distances in cg models: for atomistic descriptions, a 1 Å increase in r_{cut} can result in many new neighbours being considered, significantly altering the descriptor and thus the prediction. For cg models with sparser environments the same increase has a smaller effect, sometimes adding only one or two neighbours, or none at all, as in the A-only case. This is especially evident in the A-only grid, which shows no sharply defined optimal region.

A notable feature in both the AB_2 and A-only grid search spaces is the marked reduction in RMSE around 6 Å. This corresponds to the minimum Zn–Zn distance typically observed in ZIF structures. For any cutoff smaller than this, the model

fails to capture even the first nearest neighbour. This is particularly problematic in the A-only model, which entirely excludes B sites and the model may only "see" the central atom.

The final optimised hyperparameters are given in Table 5.1. It is important to note, however, that the RMSE landscapes remain relatively flat as discussed above in the vicinity of the minima and therefore that the notion of a single "optimal" point should be interpreted as indicative rather than definitive.

Table 5.1: Optimised cut-off radius and smoothness parameter for the three GPR models investigated.

	Fully atomistic	**AB_2**	**A-only**
r_{cut} (Å)	6.75	7.00	9.75
σ_{at} (Å)	0.30	0.25	0.90

To further assess the robustness and generalisability of the optimised hyperparameters, I performed a transferability test by applying the hyperparameters from one structural representation to the others. The results are summarised in Table 5.2.

Table 5.2: RMSEs for different GPR models for local-environment energies in ZIFs. The columns show results for the three different representations. The rows correspond to hyperparameters, $\mathcal{H}$, optimised for the respective representation.

	RMSE for $\varepsilon^{(i)}_{\mathrm{local}}$ (eV)		
	Fully atomistic	**AB_2**	**A-only**
$\mathcal{H}$ for atomistic	**0.047**	0.074	0.138
$\mathcal{H}$ for AB_2	0.047	**0.071**	0.135
$\mathcal{H}$ for A-only	0.063	0.078	**0.128**

The fully atomistic model proved the most sensitive to this transfer: using coarse-grained hyperparameters led to a noticeable increase in RMSE (from 0.047 eV to

0.063 eV). By contrast, the AB_2 and A-only models were more tolerant to mismatched settings, showing only minor degradation in performance.

These results are consistent with the loss landscape trends observed in Fig. 5.6, reinforcing that hyperparameter optimisation is more critical for high-resolution models. Coarse-grained models, with their lower structural complexity and fewer degrees of freedom, exhibit broader optima and are thus more robust to variations in descriptor settings.

Finally, to evaluate the role of angular resolution in model performance, I conducted an ablation study by systematically reducing the angular complexity of the SOAP descriptor. This was done by varying $l_{\max}$ – the f maximum degree of the spherical harmonics expansion – from 8 down to 0, where $l_{\max} = 0$ effectively removes all angular information. Figure 5.7 shows learning curves for each structural model across this range.

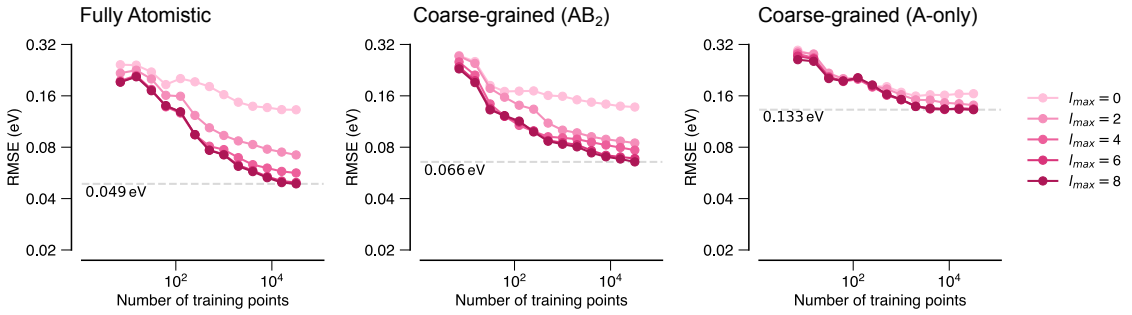

Figure 5.7: Learning curves showing the model root mean square error as a function of training set size for increasing $l_{\max}$ for the fully atomistic, AB_2 and A-only models. The RMSE for the largest number of training points is indicated by a dashed grey line in each panel. $l_{\max} = 8$ was used in all the optimised models.

The results underscore the importance of angular features. In the atomistic model, reducing $l_{\max}$ from 8 to 0 leads to a ≈250% increase in RMSE, reflecting the loss of directional bonding information. The AB_2 model shows a similar but smaller effect (140%), while the A-only model is effectively insensitive to angular resolution: the RMSE changes by less than 20%. This suggests that in representations composed solely of a single atomic species (e.g. Zn) with no explicit linker geometry, angular

detail offers little benefit.

These findings reinforce a broader theme emerging from the hyperparameter grid search and transferability study: the sensitivity of a GPR model to descriptor complexity scales with the structural resolution of the input representation. Fully atomistic models capture detailed local environments and benefit significantly from high angular resolution and careful parameter tuning. In contrast, coarse-grained models, by design, lack the structural complexity needed to benefit from high-order descriptors. As a result, they are less dependent on fine-grained features, making simpler, more computationally efficient descriptors not only sufficient but potentially preferable.

Overall, the combined results highlight a key trade-off: as structural resolution decreases, so does the need for rich feature representations. This has important implications for designing machine learning potentials that balance accuracy with computational efficiency depending on the level of coarse-graining.

5.4.3 Varying the data set composition

The second part of the study assessed the impact of the training data on model performance. The dataset used for training and evaluating the GPR models was constructed by decorating coarse-grained AB_2 networks via a back-mapping procedure, wherein atomistic detail was reintroduced based on predefined templates. The generation process is described in detail in Section 5.3.2, and includes a deliberate "rattling" of structures.

To evaluate the impact of structural distortions, configurations were categorised into three groups based on the magnitude of the applied perturbation: small, medium, and large. From this classification, two datasets were constructed. The "main" dataset consists exclusively of medium-distorted structures. This is the dataset that was used for all the results presented in the previous section. Only one rattling scheme was chosen for the proof of concept to ensure exceedingly similar structures were not present that could bias the results and make the model

appear more accurate than it truly is, due to exposure to near-duplicate examples during training. In contrast, the extended dataset includes structures from all three distortion levels, providing a broader representation of local environments.

Figure 5.8 presents an evaluation of model robustness across this distortion spectrum. In panel a, models were trained on the "main" dataset (i.e medium-rattled structures only) and tested on the individual distortion subsets. While the quantitative performance varies slightly across the different distortion levels (as seen by the R^2 values), the qualitative trends remain the same. The fully atomistic model consistently outperforms the coarse-grained variants, followed by the AB_2 model, with the A-only model noticeably trailing behind in predictive accuracy.

As the degree of distortion in the test set increases, predictive performance declines for all models. This degradation is especially pronounced for coarse-grained representations which lack the structural detail necessary to accurately model highly distorted local environments. Notably, the A-only model yields a negative R^2 value when tested on the large-rattled dataset, indicating it fails to capture any meaningful relationship and performs worse than a simple mean predictor.

Panel b presents a complementary evaluation where models are trained on an extended dataset comprising all distortion levels (small, medium, and large) and tested via 5-fold cross-validation. This broader training set leads to improved generalisability across model types, particularly when compared to models trained exclusively on medium distortions and evaluated on either medium or large distortions, as shown in panel a. The results demonstrate that incorporating a wider range of structural variations, especially both small and large distortions, enhances model robustness relative to training on a narrower subset. This underscores the importance of dataset diversity in developing reliable and generalisable models, while highlighting the limitations of training on restricted structural regimes when aiming for broader applicability.

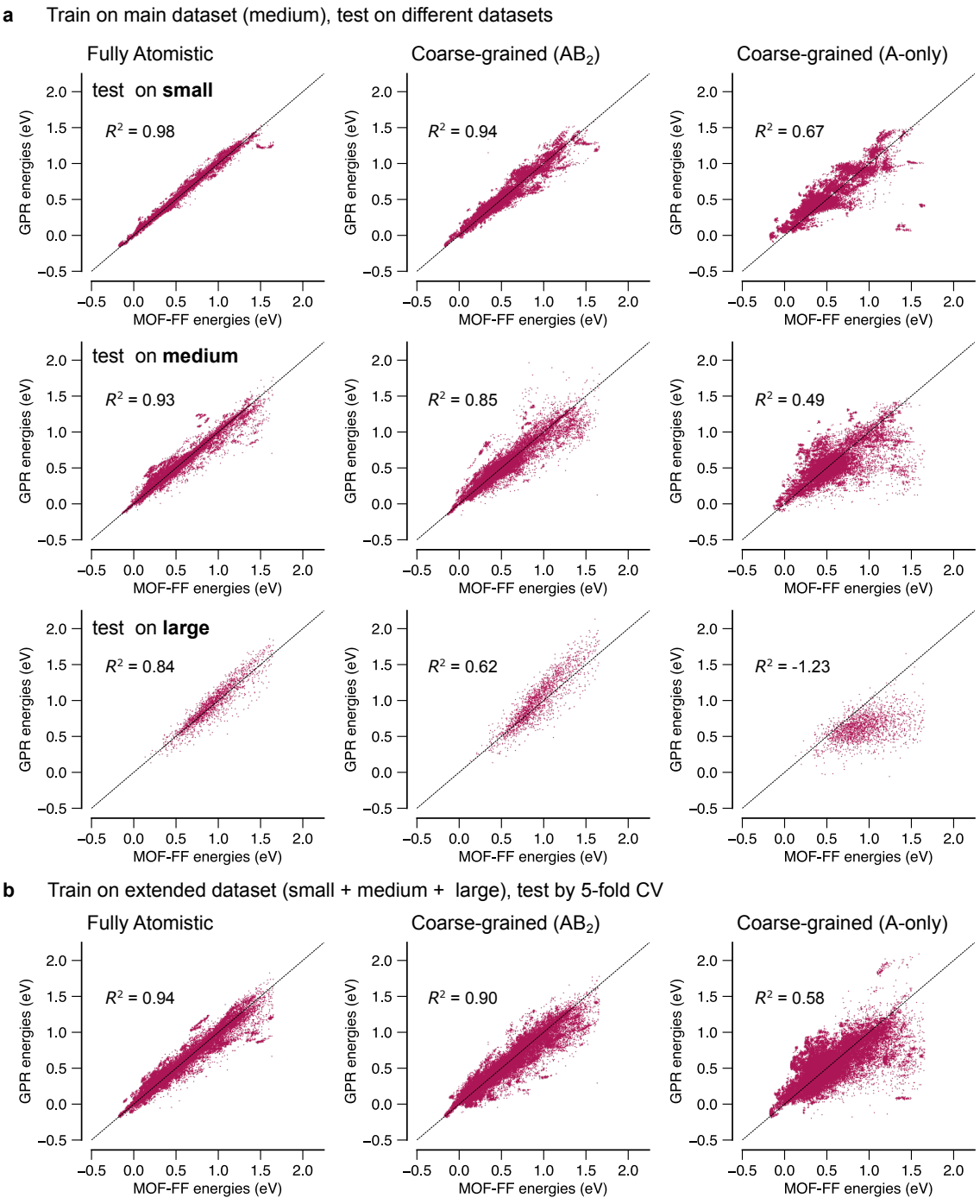

Figure 5.8: Scatter plots of local-environment energies as defined in the main text and the associated GPR ML predictions. From left to right, GPR models are characterised based on: a fully atomistic description; a cg description where the linker molecules are described by single "B" beads; and one where only A-site species are represented. (a) Tests for the GPR model described in the main text, fitted using 32,000 data points from the main ("medium") dataset. Tests for this model are shown from top to bottom: on structures with smaller distortions than in the training; on the same dataset (using 5-fold cross-validation), as shown in Fig. 5.4; and on structures with larger distortions than in the training. (b) As before, but now for training and 5-fold cross-validation for the extended dataset containing all relevant configurations.

Notably, across all settings, the A-only coarse-grained model consistently underperforms relative to the fully atomistic and AB_2 coarse-grained models. While expanding the training dataset improves overall performance, it does not overcome the intrinsic representational limitations of the A-only model.

To test how optimal hyperparameters generalise to more diverse training data, the grid search was repeated on the extended dataset, which includes configurations with small, medium, and large distortions. Figure 5.9 visualises the RMSE across the same SOAP parameter space as in Fig. 5.6, but for this extended dataset. The effect of varying cut-off radius (r_{cut}) and atomic density smoothness (σ_{at}) was again evaluated across the three structural representations.

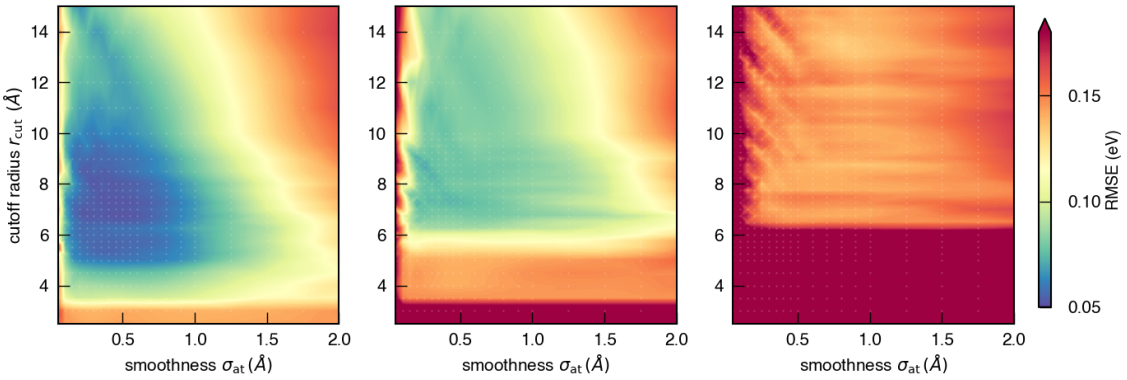

Figure 5.9: Survey of the hyperparameter space for fully atomistic GPR models versus cg-GPR models using the extended dataset. The two decisive choices are the cut-off radius (vertical axis) and the smoothness of the atomic neighbour density (horizontal axis). The results of a grid search are given by colour coding, with individual grid points highlighted by small white markers.

Table 5.3 summarises the optimal r_{cut} and σ_{at} values obtained from the grid search for the extended dataset. For ease of comparison, the corresponding optimal values from the main dataset (previously shown in Table 5.1) are also included. Notably, the fully atomistic model maintains consistent optimal cut-off values regardless of dataset scope. In contrast, the AB_2 and A-only coarse-grained models exhibit a clear shift towards larger optimal r_{cut} values when trained on the extended dataset. This shift reflects the inclusion of highly distorted ("rattled") environments in the extended dataset, which are more challenging to model accurately. To

compensate, the coarse-grained models require a larger spatial context – i.e., more neighbouring atoms – to adequately capture local energy contributions, resulting in an increased cut-off radius.

Table 5.3: Optimised SOAP hyperparameters obtained via grid searches using N = 10,000 training datapoints. Note that the above surfaces are very shallow, especially in terms of the σ_{at} dependence, and therefore those "optimised" values are only given for completeness

	Main dataset		**Extended dataset**	
	r_{cut} (Å)	σ_{at} (Å)	r_{cut} (Å)	σ_{at} (Å)
Fully atomistic	6.75	0.30	6.75	0.50
AB_2	7.00	0.25	9.00	0.20
A-only	9.75	0.90	13.50	0.80

5.4.4 Varying the learning target

The final part of the analysis examines the effect of varying the learning target. In Section 5.3.3, I introduced a synthetic learning target, denoted as $\varepsilon^{(i)}_{\mathrm{local}}$. This local energy combines the energy of the Zn^{2+} ion with half the energy of its four bonded imidazolate molecules. The rationale behind this formulation is that it offers a more accurate representation of the Zn^{2+} ion's chemical environment than using the isolated Zn^{2+} energy alone.

In this section, I explore how switching the learning target to the Zn^{2+} energy alone influences model performance, and how this simpler target behaves under structural coarse-graining.

Table 5.4 presents a comparison of the RMSEs from GPR models trained on both energy definitions. As anticipated, the fully atomistic models achieve the lowest RMSEs in both tasks, followed by the AB_2 model and then a dramatic increase to the A-only model. Models trained on the Zn-only energy consistently achieve lower RMSEs, reflecting the reduced complexity of this learning target. Notably, the A-only model trained on Zn-only energy performs comparably to the fully atomistic model trained on the local energy target. This highlights how much easier it is

to learn the simpler Zn-only energy, which omits information from the now-absent linkers. These results suggest that the Zn-only energy retains more predictive accuracy in highly simplified models, likely because it excludes the nuanced contribution of linkers that are no longer explicitly represented.

Table 5.4: Cross-validation RMSE values (in eV) for models trained on the main dataset using different coarse-graining levels.

	Fully atomistic	**AB_2**	**A-only**
Local environment, $\varepsilon^{(i)}_{\text{local}}$	0.047	0.071	0.128
Zn^{2+} energies only	0.011	0.016	0.048

The most pronounced performance gap between the local-environment energy target and the Zn-only energy target is observed in the atomistic and AB_2 models, where the RMSE for predicting $\varepsilon^{(i)}_{\text{local}}$ over 4 times higher than for the Zn-only energies. This is likely due to the fact that the A-only model is so severely coarse-grained that neither energy target can be predicted with high accuracy and so the nuance between the two learning targets is overridden by the general breakdown in predictive power.

The performance drop from the fully atomistic model to the A-only coarse-grained model is particularly stark for the Zn-only energy target, where the RMSE increases by 340%, compared to a 170% increase for the local environment energy target. This discrepancy likely arises from the differing degrees of implicit information available in each target. In the case of the local environment energy, the target inherently includes contributions from both the metal node and the linker atoms. Consequently, even when the model is trained on A-only structures (which lack explicit linker representation), it may still benefit from residual information about the linkers encoded in the energy target itself. This indirect signal allows the model to retain some predictive capability. In contrast, for the Zn-only energy models, the energy target isolates the contribution of the metal node. Although the Zn energy is still geometrically influenced by nearby linker positions this information cannot

be learned or inferred by the A-only model, since the linker atoms are absent from both the input structures and the energy decomposition. As a result, the model lacks the necessary structural and energetic context to make accurate predictions, leading to a much larger degradation in performance.

For completeness, I performed the same analysis on the extended dataset. The results are shown in Table 5.5. The same trends are observed as for the main dataset.

Table 5.5: Cross-validation RMSE values (in eV) for models trained on the extended dataset using different coarse-graining levels.

	Fully atomistic	**AB_2**	**A-only**
Local environment, $\varepsilon^{(i)}_{\text{local}}$	0.049	0.066	0.133
Zn^{2+} energies only	0.013	0.018	0.053

5.5 Conclusion and Outlook

This chapter set out to determine how much structural detail is truly necessary to machine-learn the energetics of ZIFs. Central to this investigation is the role of tetrahedral connectivity, a defining feature of ZIFs and their zeolitic analogues. I trained GPR models on three levels of structural resolution – fully atomistic, coarse-grained AB_2, and a minimalist A-only mapping – and showed that accurate energy predictions are possible even with significantly reduced structural detail, as long as tetrahedrality is preserved. Specifically, local-environment energies in ZIFs can be learned from the AB_2 coarse-grained representation with less than a factor-of-two loss in accuracy compared to atomistic models, despite a drastic reduction in the number of structural degrees of freedom, from 51 coordinates per $Zn(Im)_2$ unit to just 9 per AB_2 equivalent.

The preservation of tetrahedral connectivity between nodes and linkers emerges as the irreducible minimum for reliable coarse-grained modelling. When linkers are removed entirely as in the A-only models, predictive power collapses, underlining

the critical structural role that connectivity plays in encoding energetic information. The success of the AB_2 description provides quantitative support for the long-standing analogy between ZIFs and zeolites: both materials can be understood as tetrahedral AB_2 networks, and it is this network, not the detailed chemistry of the linker, that carries the bulk of the energetic information.

Looking ahead, a natural extension is to move beyond unsubstituted imidazolate linkers to more complex and anisotropic ones (e.g., methyl-, ethyl-, or benzimidazolate). Preliminary tests show that treating methylimidazolate (mIm) as a single "B" bead is insufficient for accurate coarse-grained models of $Zn(mIm)_2$. Future work could explore more nuanced coarse-graining strategies for such ligands, as shown, e.g., by Semino *et al.* [452], who used an atom-to-bead ratio of ≈ 2.6 for a carboxylate MOF, and Alvares *et al.* [427] who studied different cg models for ZIFs.

While this chapter focused on learning synthetic local energies, similar machine learning approaches can be applied to other atomistic observables, such as NMR chemical shifts [453], which would aid in the interpretation of experimental studies particularly for glassy ZIFs. Extending the present energy-learning framework to include forces would allow the construction of fully coarse-grained ML potentials capable of accessing much larger time and length scales than current atomistic methods. Recent work by Mohamed *et al.* [454] has shown that such cg-ML force fields can reproduce both static and dynamic responses of ZIF-8 under pressure, including structural amorphisation. Taken together, these developments suggest that integrating energy and force learning within a unified cg framework could offer a powerful and transferable toolset for simulating structural, thermal, and mechanical behaviour across a wide range of ZIF chemistries.

While this chapter focuses on crystalline ZIFs, an improved understanding of their energetic landscape and structural representations could prove a useful stepping stone to the modelling of amorphous ZIFs. These disordered analogues remain far less characterised, both experimentally and computationally, and few reliable

structural models exist. Recent advances, such as the active-learning-based structural refinement of *a*-ZIF by Nicholas *et al.* [189], highlight the importance of capturing intermediate-range topology, which often departs significantly from crystalline analogues. By investigating the extent to which coarse-grained models of crystalline ZIFs can predict local energetics, this work contributes foundational tools for interpreting and eventually simulating the disordered frameworks of amorphous ZIFs.

More broadly, this chapter highlights a general strategy for identifying transferable structural representations across material classes: by isolating and preserving the topological features that govern local energetics, one can construct simplified yet predictive models that bridge chemistry, structure, and function. This approach may prove valuable not only in understanding known materials but also in the computational discovery of new ones.

Chapter 6

Learning across chemical domains with MLIPs

6.1 Acknowledgements

Section 6.4.2 of this chapter has been published as a preprint on arXiv [274]. Portions of the text and several figures have been reused; where appropriate, figures are labelled as "Adapted." I am not the first author of this work and will only discuss the results that I directly contributed to. In the Methods section, I describe certain core methodologies that were developed by others but are essential for understanding the presented work. These methods are clearly attributed to the original authors where appropriate. I would like to thank John Gardner and Daniel Thomas du Toit for their development of the model distillation pipeline, and Daniel in particular for running the MD simulations using the distilled ACE model.

Section 6.4.1 of this chapter is new and thus far unpublished work I have carried out separately from the above preprint.

6.2 Introduction

As I hope to have shown in the previous chapters, MLIPs have become indispensable tools for atomistic simulation, offering near *ab initio* accuracy at a fraction of the computational cost [51, 53]. Yet, despite their promise, MLIPs face important limitations. Chief among these is their reliance on large, diverse datasets that span relevant chemical environments and thermodynamic conditions for their downstream use. For chemically complex or disordered systems, assembling such datasets is a significant bottleneck, both time-consuming and resource-intensive [211]. This challenge has spurred growing interest in training strategies that improve data efficiency

without sacrificing predictive power [209, 274, 455].

Two principal approaches to this challenge have emerged in the literature. The first is transfer learning, which reuses knowledge acquired from one system to accelerate learning in another [456]. For MLIPs, this typically involves leveraging a model trained on a well-characterised system to assist in learning a related, data-scarce system. When successful, transfer learning can significantly reduce the volume of training data required, improve predictive accuracy, and enable faster model deployment [457–459]. For instance, Smith *et al.* [458] demonstrated that a model trained on DFT-level data could be effectively fine-tuned using a relatively small number of high-accuracy coupled-cluster [CCSD(T)] labels, thereby elevating the overall fidelity of the resulting MLIP. A particularly interesting direction within transfer learning is *alchemical* transfer learning, wherein knowledge is transferred across elements by exploiting shared bonding or structural features [209, 460]. This approach aims to harness underlying chemical similarities, such as shared coordination environments or structural motifs, to enable cross-element generalisation.

Recent studies have demonstrated the potential of this approach. Gardner *et al.* [209] introduced a synthetic pre-training strategy in which a model trained on a large synthetic dataset for carbon was fine-tuned on DFT-labelled silicon structures. Their results showed that such pre-training can substantially improve accuracy in low-data regimes and enhance numerical stability, particularly when structural rescaling is used to align the atomic environments of chemically distinct systems. Röcken and Zavadlav [460] extended this strategy to silicon and germanium, showing that transfer learning significantly improved force and energy predictions and increased the robustness of molecular dynamics simulations across temperatures.

Despite these promising results, key questions remain. What degree of structural or energetic similarity is necessary for successful transfer? Is a shared geometric motif, such as tetrahedral coordination, sufficient to enable cross-element generalisation? Previous studies have largely focused on chemically similar elements within

the same periodic group. In this chapter, I explore whether tetrahedral topology can provide a meaningful scaffold for transferability across broader chemical domains. Through pre-training and fine-tuning experiments on both elemental systems (carbon, silicon) and more complex AB_2-type materials (silica, water), I systematically assess the effectiveness and limitations of alchemical transfer learning in tetrahedral networks. Ultimately, by analysing transferable representations in both simple and complex tetrahedral systems, this chapter offers insights into the structural and energetic features that govern cross-domain learning. These insights lay the groundwork for developing more generalisable, efficient, and robust MLIPs tailored to disordered tetrahedral materials.

The second, more recent, approach to enhancing data efficiency in MLIPs involves foundation models: large, general-purpose potentials trained on broad, heterogeneous datasets spanning most of the periodic table [169, 461–463]. Inspired by developments in natural language processing and computer vision [464–466], these models leverage large-scale pretraining to enable strong generalisation in downstream tasks. Their development in the atomistic domain has been accelerated by the rise of open-access materials datasets [8, 9, 467] and the adoption of expressive graph-based neural networks capable of modelling complex interatomic interactions [50, 167, 256].

Foundation models aim to be universally applicable: instead of training a separate MLIP for each new chemical system, a single pretrained model can be fine-tuned with relatively little additional data, offering both improved data efficiency and reduced development time [468]. As a result, a major challenge now is making foundation models lightweight and broadly usable, such that they become accessible tools for the wider atomistic simulation community.

A promising solution to this challenge is model distillation, the process of compressing the knowledge encoded in a large foundation model into a smaller, task-specific model that retains high accuracy while being significantly more efficient to

train and deploy [469, 470]. In this chapter, I demonstrate such an approach using liquid water as a test case – an archetypal example of a chemically and structurally complex system. By fine-tuning a foundation model on water and distilling it into a compact MLIP, I show that the chemical insights captured during large-scale pre-training can be transferred into lightweight, high-performance models suitable for practical simulations. Beyond its immediate utility, this work also reveals the types of chemical knowledge that foundation models can learn to apply across domains, and how that knowledge can be adapted for more specialised applications.

Together, the transfer learning and foundation model strategies explored in this chapter present complementary pathways for sharing chemical knowledge across domains, and help illuminate the structural and energetic features that underpin successful generalisation in atomistic machine learning. More broadly, they point toward a future where MLIPs are not only more data-efficient but also more widely accessible and applicable to a diverse array of complex materials.

6.3 Methods

6.3.1 Data sets

This chapter uses a variety of data sets to investigate the transferability of PES models across different chemistries. They are all taken from existing literature. The following sections briefly introduce the data sets, their origin and the type of information they contain.

Carbon

The carbon data are taken from the study by Rowe *et al.* [206]. The dataset comprises over 6,100 unique carbon structures, representing a wide range of bonding environments, including sp, sp^2, and sp^3 hybridisations.

Crystalline structures include various allotropes such as diamond, graphite, and lonsdaleite, as well as high-pressure phases like bc8 and sc16. Amorphous configurations span low- and high-density amorphous carbon, with structures generated from

melt-quench simulations. The dataset also includes molecular and cluster structures (including fullerenes and nanotubes) and selected defective and surface configurations. Each structure was labelled with DFT using the PBE exchange-correlation functional [142].

Silicon

The silicon data is taken from the study by Bartók *et al.* [91]. The dataset encompasses a diverse collection of atomic configurations, totalling more than 170,000 environments sampled from crystalline, liquid, and amorphous phases, as well as from various defect structures.

Crystalline structures span multiple phases such as diamond, β-Sn, bc8, and st12, with both equilibrium and strained geometries. Amorphous and liquid states are sampled from MD simulations across a range of temperatures and pressures. The dataset also includes surfaces ((111), (110), and (100) orientations), point defects (vacancies and interstitials), dislocation cores, and crack tips. Each configuration is labelled using DFT calculations employing the PW91 exchange-correlation functional [471].

Water

Two water data sets are used in this chapter. The first is the dataset by Ibrahim *et al.* [346] which is described in Section 4.3.1 and contains predominantly ice structures.

The second is sourced from the study by Cheng *et al.* [388]. The dataset comprises 1,593 configurations of liquid water, each containing 64 molecules, with energies and forces evaluated at the revPBE0 [142, 472] level of theory with a Grimme D3 dispersion correction [145, 146]. To represent proton disorder in ice phases, 16 proton-ordered configurations were generated for both ice Ih and ice Ic, each also consisting of 64 water molecules. In contrast to the first dataset, this dataset contains predominantly liquid water structures.

Silica

The data for this chapter is sourced from the study by Erhard *et al.* [55]. The dataset comprises a diverse set of more than 3,000 silica structures spanning crystalline, amorphous, liquid, and high-pressure phases. Crystalline configurations include relaxed unit cells and distorted variants of polymorphs such as α-quartz, coesite, stishovite, and α-cristobalite. Amorphous structures were generated via melt–quench simulations and include fast-quenched, slowly-quenched, and hybrid-annealed variants (produced by slow quenching with a relatively cheap potential followed by brief annealing using the more expensive potential) to capture realistic glassy behaviour.

The dataset also includes isolated Si–O dimers and structures representative of high-pressure polymorphs such as α-PbO_2-type silica. Both ambient and high-density disordered phases are represented to enable the model to extrapolate under extreme conditions. All configurations were labelled with DFT using the SCAN functional [144].

6.3.2 Alchemical Transfer Learning

Alchemical transfer learning leverages knowledge acquired from one chemical environment to improve performance on a different, previously unseen chemical task. To implement this, I adopt a pre-training followed by fine-tuning approach.

Various advanced methods exist for fine-tuning pre-trained models, including early layer freezing, AutoFreeze [473], the lottery ticket hypothesis [474], and discriminative fine-tuning [475]. In this work, I adopt the simplest possible approach: I fine-tune all weights of the best pre-trained model by continuing standard training on the target dataset. This is the same method used by Gardner *et al.* [209] in their alchemical transfer learning approach. An overview of the process is shown in Figure 6.1.

First, I directly train a model (shown in purple) on a dataset from the target

chemical system D_0 (e.g., silicon), starting from randomly initialised weights. I then evaluate its performance by predicting energies and forces on a hold-out test set.

Next, I pre-train a separate model on a large dataset D_1 from a different chemical system (e.g., carbon), as shown in orange. I then fine-tune this pre-trained model on the target system data D_0 and evaluate its predictions on the same hold-out test set (shown in red).

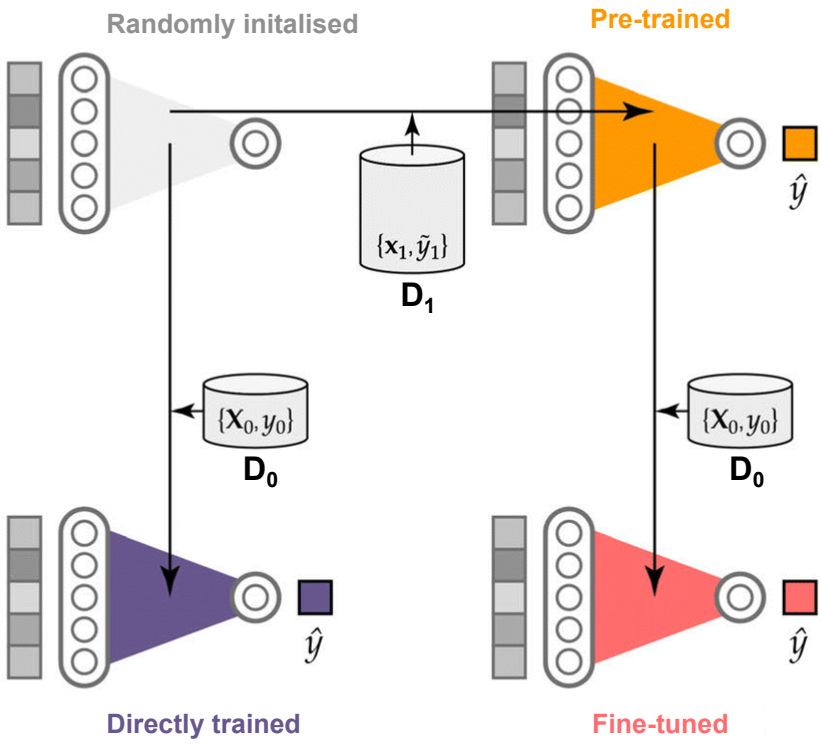

Figure 6.1: Model Training Strategies: Direct Training vs. Pre-training and Fine-tuning. A machine learning model can be trained on a target chemical dataset D_0 (e.g., silicon) either by (left path) direct training (purple), or (right path) by first being pre-trained (orange) on a separate source dataset D_1 (e.g., carbon), and subsequently fine-tuned (red) on D_0. This approach leverages prior knowledge from D_1 to improve performance on D_0. Figure adapted from Ref. [449].

All models are trained using the same hyperparameters and neural network architecture, specifically, the MACE GNN [256], to ensure that any observed differences in performance arise solely from the effects of pre-training and fine-tuning, rather

than from architectural or optimisation differences. The hyperparameters used were $l_{\max} = 2$, 64 channels, batch size of 16, and a learning rate of 0.01. The cutoff distance was system-dependent and chosen based on accepted values from the literature.

It is important to note that pre-processing of the pre-training dataset D_1 is necessary to enable the model to generalise effectively to the target chemical system D_0. To this end, I apply two main steps:

(i) **Transmuting:** In the first step, I replace each atom in D_1 with the corresponding atom type from D_0. This is achieved by mapping the atomic numbers in D_1 to those of D_0 (such that pre-training is performed on the same chemical elements as the target system). This operation leaves the structural parameters, such as the unit cell and interatomic distances, unchanged, but updates the atomic identities so that the model learns the appropriate chemical environment.

(ii) **Rescaling:** The second step involves rescaling the unit cell and interatomic distances to match the characteristic length scales of D_0. Specifically, I scale the structures so that the first maximum in the RDF of the source and target systems align. Gardner *et al.* [209] demonstrated that without this alignment, effective transfer does not occur. This is intuitive: the typical minimum atomic separation in carbon structures is approximately 1.4 Å, whereas in silicon it is around 2.1 Å. For a given cutoff, this results in significantly different numbers of atoms within the receptive field of the GNN model, and thus very different internal representations of local environments. In the carbon-silicon example, each cell (and atomic positions) of each pre-training structure would be scaled by a factor of $2.1/1.4 = 1.5$

After rescaling distances, I also scale the forces by the inverse of the distance scaling factor. Since forces are the derivative of energy with respect to position, $F = -\partial E/\partial r$, the force magnitudes must be adjusted accordingly.

Although this force scaling step was not performed in prior work, I found it to significantly improve transfer effectiveness.

6.3.3 Dimensionality Reduction

In this chapter, dimensionality reduction is used as a tool to explore and compare the structural diversity of different datasets. Each dataset consists of atomic environments represented by high-dimensional SOAP descriptors (see Section 2.2.2). While these descriptors encode detailed information about local atomic structure, their high dimensionality (typically hundreds or thousands of features) makes them difficult to interpret or visualise directly. To facilitate comparison and analysis, a projection onto a low-dimensional space is therefore required.

To this end, I employ dimensionality reduction technique to project the high-dimensional descriptor vectors onto a lower-dimensional manifold while maximally preserving information. Visualisations based on these compressions allow for a compact representation of data, and makes it easier to identify patterns, similarities, and differences between datasets.

In this work I have used principal component analysis (PCA) for dimensionality reductions. This is a linear technique that iteratively finds the directions in the data with the highest variance (i.e. the directions along which the data varies most strongly). These directions are known as *principal components*. By projecting the data onto the first few principal components, we can retain the most important structural variation while reducing the number of dimensions.

More formally, suppose we have N SOAP vectors, each of dimensionality d, represented as $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$, where each $\mathbf{x}_i \in \mathbb{R}^d$. We first standardise the data by subtracting the mean from each feature so that the dataset is centred.

Next, we compute the covariance matrix $\mathbf{C}$ of the dataset:

$$\mathbf{C} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i^\top. \tag{6.1}$$

This matrix describes how the different components of the SOAP vectors vary together.

We then solve the eigenvalue problem:

$$\mathbf{C}\mathbf{u}_k = \lambda_k \mathbf{u}_k, \tag{6.2}$$

where $\mathbf{u}_k$ is the k-th eigenvector (the k-th principal component) and λ_k is its corresponding eigenvalue. The eigenvalues indicate how much variance is captured by each component.

To reduce the dimensionality, we keep only the top K eigenvectors (those with the largest eigenvalues), forming a projection matrix $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_K]$. The original data can then be projected into this K-dimensional space using:

$$\mathbf{z}_i = \mathbf{U}^\top \mathbf{x}_i, \tag{6.3}$$

where $\mathbf{z}_i \in \mathbb{R}^K$ is the lower-dimensional representation of the original SOAP vector $\mathbf{x}_i$.

In the context of this work, I typically choose $K = 2$ so that the resulting vectors $\mathbf{z}_i$ can be visualised in scatter plots. In this reduced space, environments that are structurally similar tend to appear close together, while environments that are very different appear further apart. These plots provide a simple and interpretable way to compare different datasets. For example, if two datasets overlap in PCA space, this suggests they contain similar atomic environments. If they appear as distinct clusters, this indicates structural differences between the datasets.

PCA has several desirable properties for materials and atomistic applications. First, it is deterministic (i.e. there is a single valid decomposition of the data) and computationally efficient, scaling linearly with the number of samples and quadratically with feature dimensionality. Second, the resulting components are orthogonal, ensuring that each captures a unique, uncorrelated mode of variation. Third,

PCA offers a natural ranking of components by explained variance, which allows for straightforward selection of the number of retained dimensions. This ranking also facilitates interpretability by enabling the examination of individual principal components to identify dominant structural motifs or patterns in the descriptor space.

While more advanced nonlinear dimensionality reduction techniques such as t-SNE [476, 477] and UMAP [478] can also be used for this purpose, they tend to be harder to interpret and reproduce. PCA is preferred here for its simplicity and interpretability.

6.3.4 Quantifying Structural Similarity with Earth Mover's Distance

To quantify the similarity between the atomic environments of two chemical systems, I used the *Earth Mover's Distance* (EMD) [479], also known as the 2-Wasserstein distance. This metric measures the minimum "effort" required to transform one distribution into another, where effort is defined in terms of the amount of distribution "mass" moved multiplied by the distance it is moved. In this work, EMD is used to compare the distributions of SOAP descriptors projected into a low-dimensional space via PCA, as described in Section 6.3.3. It is essentially used as a similarity metric, where larger values correspond to greater alignment between structural distributions.

Given two sets of atomic environments, each represented by a collection of high-dimensional SOAP vectors, I first projected both sets into a common 2D PCA space. This projection ensures a consistent basis for comparison and reduces computational complexity.

Let $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ and $\mathcal{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_m\}$ represent two sets of atomic environments, where each environment is a point in PCA space. These two sets correspond to the descriptors from two different chemical systems. To compare them, I treated each set as a probability distribution by assigning equal weight to each point (i.e.

assuming each atomic environment is equally important).

Next, I computed the pairwise distances between all points in $\mathcal{X}$ and $\mathcal{Y}$ using Euclidean distance. This gives a distance matrix M, where each entry M_{ij} is the distance between $\mathbf{x}_i$ and $\mathbf{y}_j$. To make comparisons consistent across different systems, all entries in M are divided by its average value, effectively normalising the distances:

$$\hat{M}_{ij} = \frac{M_{ij}}{\text{mean}(M)}. \tag{6.4}$$

I then computed the EMD [480], which finds the most efficient way to "move" the mass from one distribution to match the other, given the cost defined by $\hat{M}$. This is done by solving an optimisation problem that determines how much mass to move between each pair of points, while preserving the total amount of mass in each set. For intuition, I defined a reversed version of EMD, denoted as revEMD, which is simply $1 - \text{EMD}$; this transforms the score so that 1 corresponds to identical distributions and 0 to maximal dissimilarity.

This method provides a quantitative measure of global structural similarity between datasets to help improve interpretability of the PCA structure space. Unlike simple point-wise comparisons, EMD accounts for the overall geometry of the distributions and is sensitive to shifts and reshaping of the data clouds in PCA space. This makes it particularly well-suited for comparing materials with subtle differences in structural motifs, where overlap in PCA projections may be partial or diffuse.

6.3.5 Model Distillation

In the final part of this chapter, I explore the use of model distillation to compress the knowledge contained in a large foundation model into a smaller, task-specific model. This approach is based on the pre-print by Gardner *et al.* [274], with the methodology developed and implemented by John Gardner and Daniel Thomas du Toit. I provide a brief overview of the process here to support understanding of how the distilled models were created. My contribution involved using these distilled

models to run MD simulations of liquid water, with the results discussed in later sections.

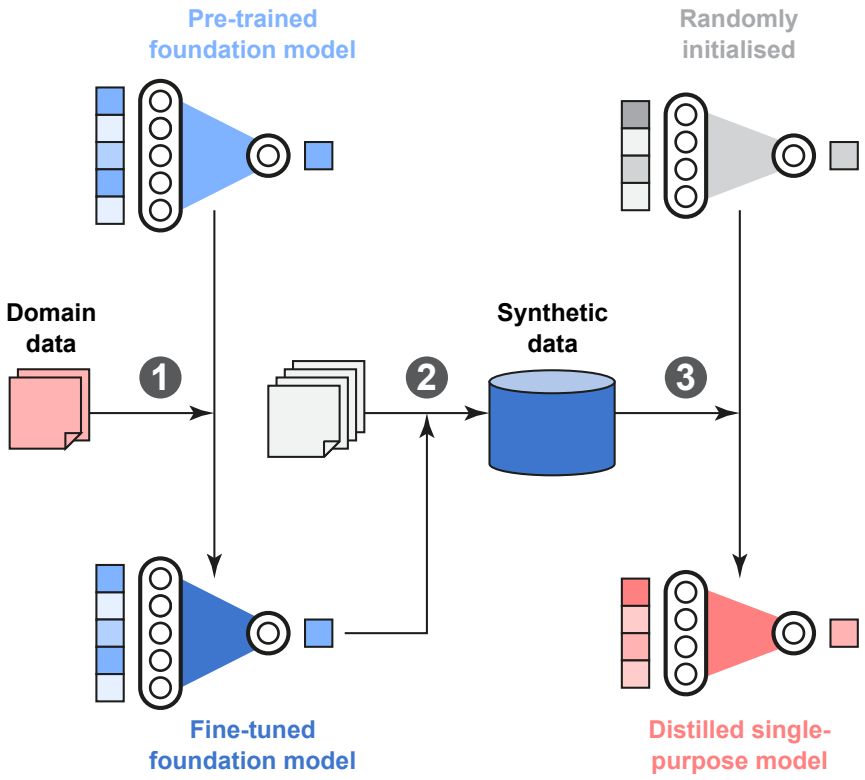

Figure 6.2: Overview of the distillation workflow for atomistic machine learning models. Step 1: A pre-trained foundation model (light blue) is fine-tuned using a limited set of high-fidelity, domain-specific quantum mechanical data (red). Step 2: The resulting domain-adapted model (dark blue) serves as a teacher to generate a large corpus of labelled synthetic data at low cost. Step 3: A compact, task-specific model (red) is trained on this synthetic dataset, resulting in a lightweight distilled model optimised for the target application. Figure adapted from Ref. [274].

As a case study, the MACE-MP-0b3 foundation model [169] was distilled to produce smaller, faster models capable of accurately simulating liquid water at room temperature and pressure, targeting the revPBE0-D3 [145, 472, 481] level of theory. The distillation pipeline, illustrated in Fig. 6.2, proceeds in three stages.

First, the foundation model was fine-tuned on a small, high-quality dataset comprising just 25 quantum-mechanical configurations of bulk liquid water, taken from Ref. [388], with an additional 5 structures used for validation.

Second, the fine-tuned model was used to generate a large synthetic dataset (approximately 10,000 structures, plus 50 dimer configurations for each of H–H, O–O, and O–H pairs) via a sample-efficient protocol involving structural perturbation and relaxation. This approach avoids the high computational cost of generating data via molecular dynamics.

Third, this synthetic dataset was used to train a set of compact student models. To demonstrate the architecture-agnostic nature of the distillation process, we selected both PaiNN [267], a graph neural network, and ACE [248]. The resulting student models achieved force MAEs relative to the DFT close to that of the fine-tuned foundation model, while offering up to 2 orders-of-magnitude speed-up. Importantly, these distilled models were capable of running large and stable MD simulations on a single GPU, unlike the more computationally expensive foundation model, and the poorly performing directly trained models.

Overall, this distillation framework is flexible, efficient, and does not rely on specialised, high-end hardware or a specific model architecture, making it an accessible strategy for developing high-performance models for atomistic simulations.

6.4 Results

6.4.1 Transfer Learning

In the following sections, I present the results of a series of experiments investigating transfer learning between different chemical systems. In all cases, models are trained using the MACE architecture [256], with hyperparameters held constant for a given pair of pre-training and fine-tuning tasks to ensure comparability. Structure maps are generated by applying PCA to SOAP descriptors of atomic environments.

The requirements for positive transfer

Previous work on alchemical transfer for MLIPs has demonstrated that pre-training on one chemical system can sometimes improve performance on another. Notably, Gardner *et al.* [209] explored such transfer by pre-training on a carbon dataset,

before fine-tuning on a silicon dataset. While their results showed some positive transfer, the approach involved pre-training on large, undifferentiated sets of synthetic carbon structures without separating them by structural type (e.g., diamond, graphite, amorphous). Although they performed some coarse separation by density, they did not investigate how specific chemical or structural domains contributed to transfer success or failure. As a result, their analysis did not yet possess the resolution needed to understand which regions of chemical space were responsible for driving or hindering transfer performance.

To build on this work, I proposed that two key criteria must be satisfied (Figure 6.3) to enable positive alchemical transfer: (1) geometric similarity between structures in the source and target chemical systems, and (2) alignment of the PES, such that high-energy structures in one system correspond to high-energy structures in the other and vice versa.

The rationale behind these hypotheses is that for a model to leverage prior knowledge during fine-tuning, it must see structurally similar configurations and associate them with comparable energetic behaviour. To test this hypothesis, I conducted a series of three experiments corresponding to the cases illustrated in Figure 6.3. Each experiment involved pre-training on a different area of carbon chemistry, followed by fine-tuning on diamond-type silicon (dia-Si) structures. I then compared the performance of the fine-tuned models to a baseline trained directly on dia-Si data.

Case 1: Geometric dissimilarity In the first set of experiments, I investigated transfer between graphitic carbon (gra-C) and dia-Si, two systems that differ significantly in their geometric configurations. This is clearly illustrated in the structure space map (Figure 6.4a), where the two datasets occupy distinct and non-overlapping regions. This separation is quantitatively confirmed by the revEMD score (see Section 6.3.4), which yields a value of 0.0433. Since a revEMD of 1 indicates identical structural distributions and 0 corresponds to complete dissimilarity, this result con-

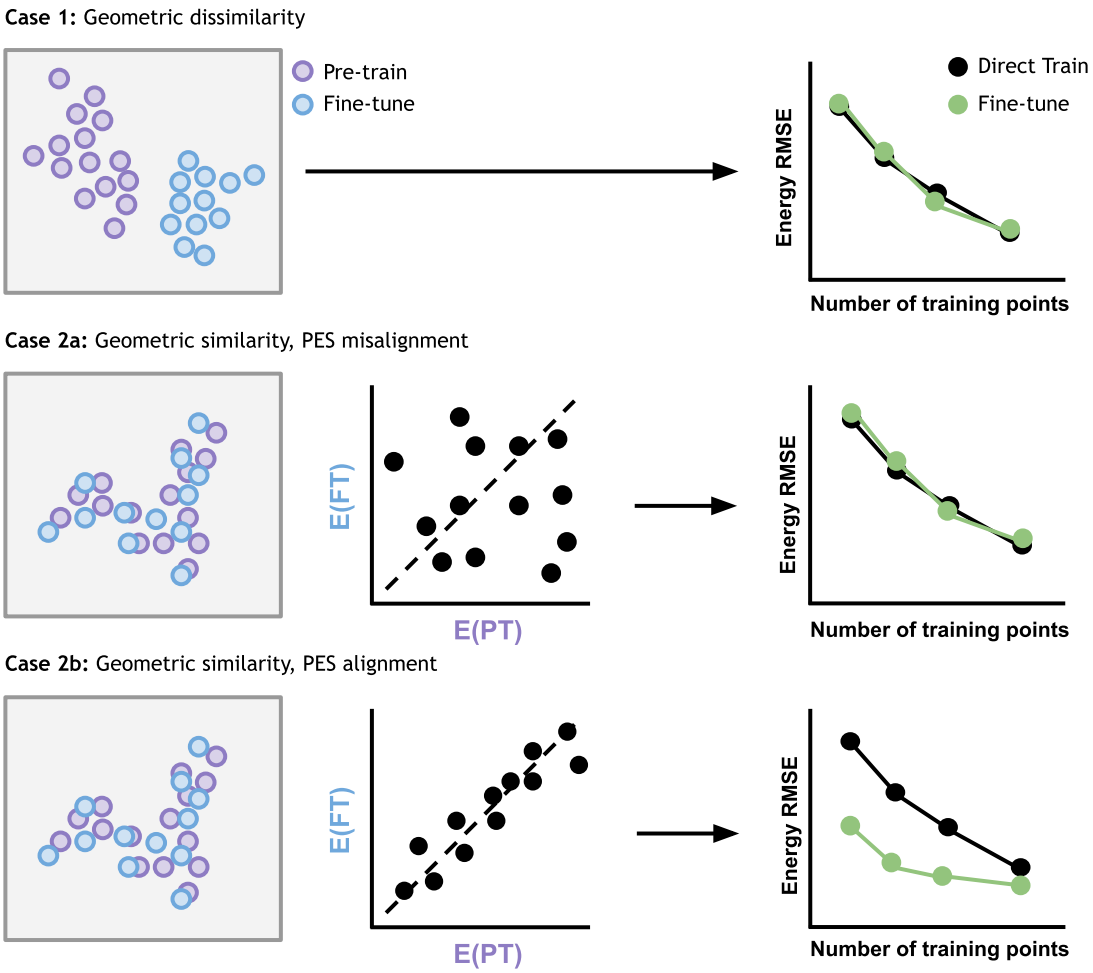

Figure 6.3: Hypothesised conditions for effective alchemical transfer. Case 1 illustrates the scenario where geometric dissimilarity between pre-training (purple) and fine-tuning (blue) datasets are expected to limit transfer effectiveness. In Case 2a, despite geometric similarity, misalignment of the PES may prevent the model from benefitting from pre-training. Case 2b represents the hypothesised ideal condition, where both geometric similarity and PES alignment enable successful knowledge transfer, leading to improved performance during fine-tuning. Schematic energy and force RMSEs are shown as a function of the number of fine-tuning points, comparing direct training (black) and fine-tuning from a pre-trained model (green).

firms that the two systems are nearly maximally different in terms of local atomic environments.

The learning curves (Figure 6.4b) demonstrate that pre-training on increasing amounts of gra-C data fails to yield any improvement in either energy or force accuracy after fine-tuning. In fact, models pre-trained on up to 800 gra-C structures consistently underperform compared to models trained directly on the target silicon

data. These results support the hypothesis that geometric similarity is a necessary precondition for successful alchemical transfer: without it, the model is unable to generalise from the source system and is instead hampered by incompatible structural priors.

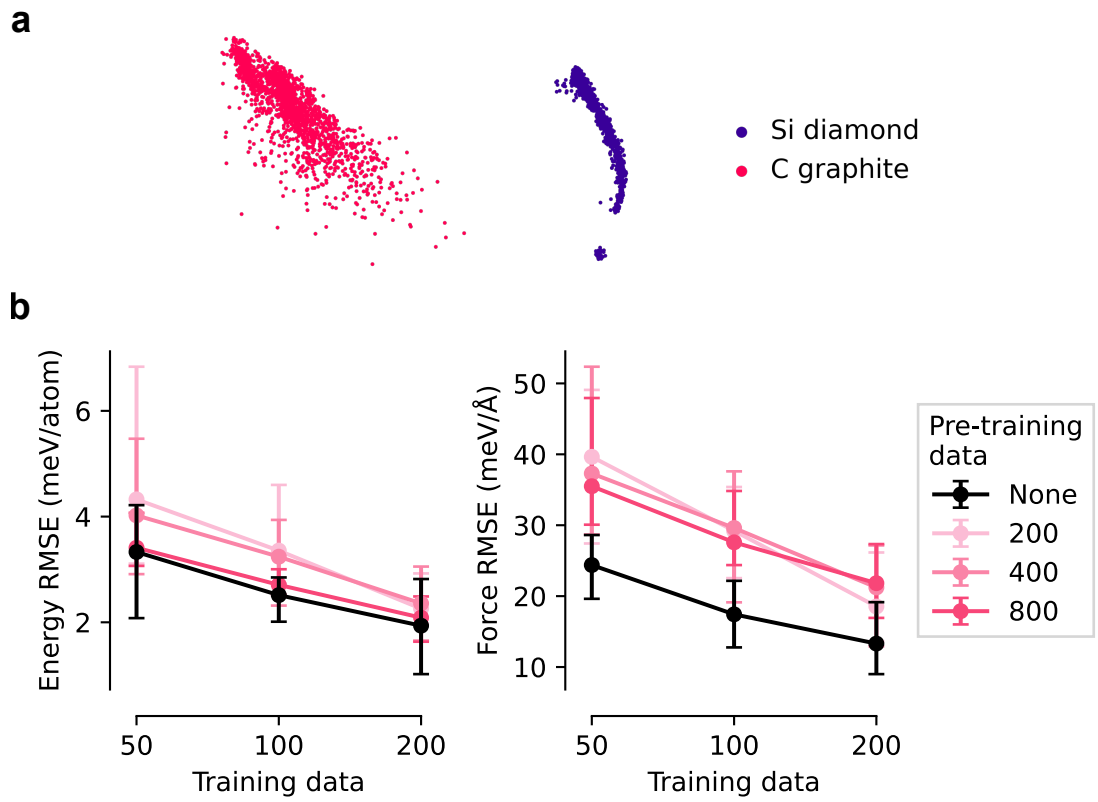

Figure 6.4: (a) PCA structure map showing the distribution of atomic environments in the pre-training dataset (scaled gra-C, red) and the fine-tuning target system (dia-Si, purple). The clear separation between the two clusters highlights their geometric dissimilarity. (b) Energy and force RMSE as a function of the number of fine-tuning data points for models pre-trained on varying amounts of graphitic carbon data.

Case 2a: Geometric similarity but PES misalignment In the second set of experiments, I examined a case where geometric similarity between the pre-training and fine-tuning datasets was high – both were derived from diamond structures – but the PES were deliberately misaligned. This was achieved by labelling the carbon diamond (dia-C) structures used for pre-training with an oxygen-derived potential, thereby assigning them energies and forces that do not reflect the true PES of carbon or silicon. While such a setup is unphysical, it serves to isolate and highlight the

importance of PES alignment in successful transfer. The structure map (Figure 6.5a) shows that the pre-training and fine-tuning structures are geometrically very similar, which is confirmed by a revEMD score of 0.9163. This value indicates a high degree of structural overlap, with only minimal differences in the distributions of local atomic environments.

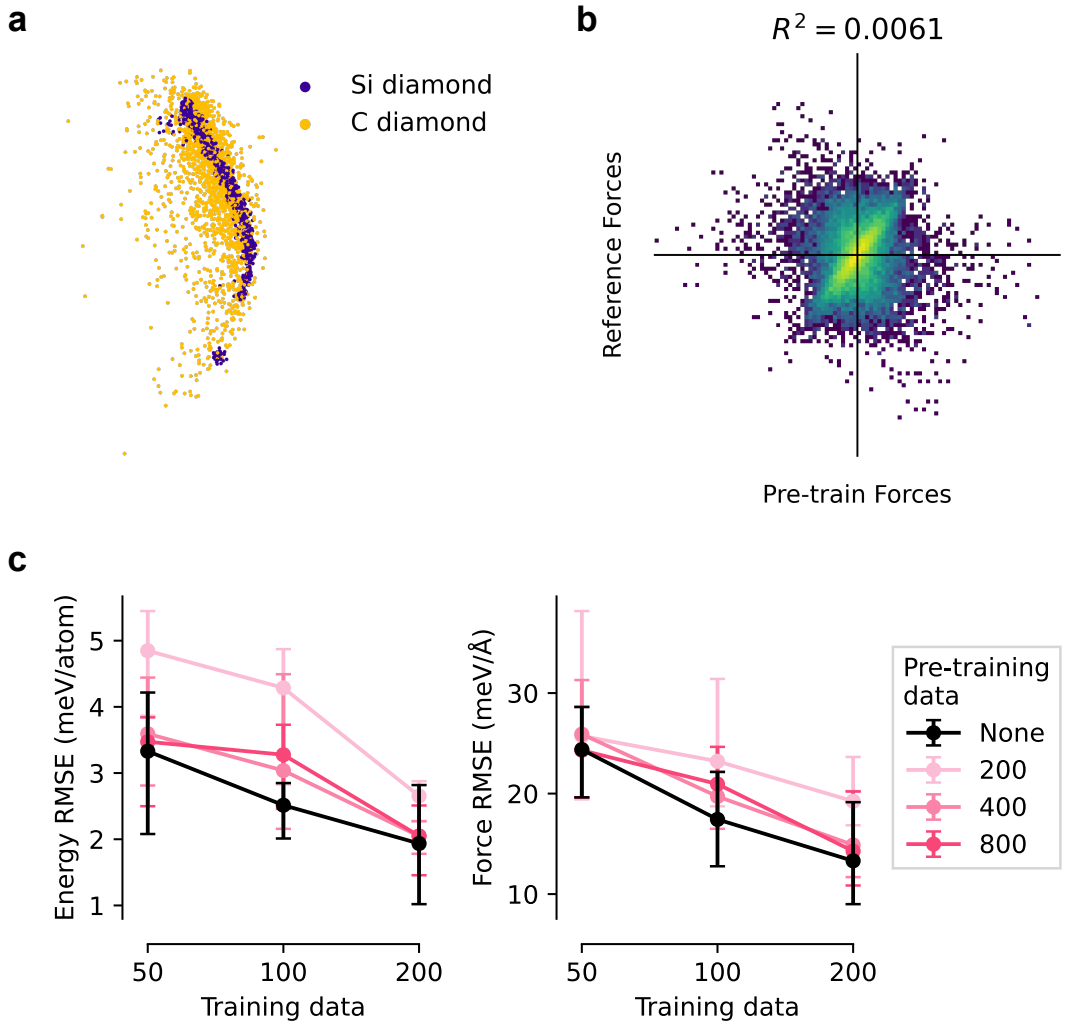

Figure 6.5: Effect of PES misalignment on alchemical transfer. (a) PCA structure map showing high geometric similarity between the pre-training dataset (scaled C-dia, yellow) and the fine-tuning target system (Si-dia, purple). (b) Predicted vs. reference forces from the fine-tuned model, showing essentially no correlation ($R^2 = 0.0061$), indicating strong PES misalignment. (c) Energy and force RMSE as a function of the number of fine-tuning data points for models pre-trained on varying amounts of diamond carbon data.

However, the force predictions from the pre-trained model show no correlation with the true reference forces after fine-tuning (Figure 6.5b, $R^2 = 0.0061$), con-

firming a misalignment of the learned PES. Consequently, the learning curves (Figure 6.5c) show that pre-training on PES-misaligned data provides no benefit and in some cases slightly degrades performance compared to direct training, particularly in the low-data regime. These results support the hypothesis that PES alignment is a necessary complement to geometric similarity, and without it alchemical transfer fails to improve model performance, even when structural overlap is high.

Case 2b: Geometric similarity and PES alignment In the final experiment, I tested a case where both hypothesised requirements for successful alchemical transfer, geometric similarity and PES alignment, are satisfied. Here, the model was pre-trained on dia-C structures labelled with their correct carbon PES, and fine-tuned on dia-Si data. The PCA structure map (Figure 6.6a) is the same as above and shows very high structural similarity (revEMD of 0.9163), while the force correlation plot (Figure 6.6b) shows moderate alignment between the pre-trained predictions and fine-tuning labels ($R^2 = 0.4561$).

Under these conditions, the learning curves (Figure 6.6c) show consistent improvement in both energy and force accuracy when pre-training is used, particularly when a sufficient number of pre-training structures (≥ 400) are included. While the degree of improvement may seem modest, it is in fact quite notable given the low intrinsic complexity of the dia-Si dataset: direct training already performs well, leaving limited room for gains through transfer. These results demonstrate that when both geometric and energetic alignment are present, pre-training can yield clear and reliable performance benefits. In later sections, we will see that in more complex systems with higher learning barriers, the same transfer strategy leads to significantly larger improvements.

The results from the three experiments are summarised in Table 6.1, clearly demonstrating the necessity of both geometric similarity (quantified via revEMD) and PES alignment (quantified via the R^2 value) for effective alchemical transfer. Each entry in the table reflects the best performance achieved across varying

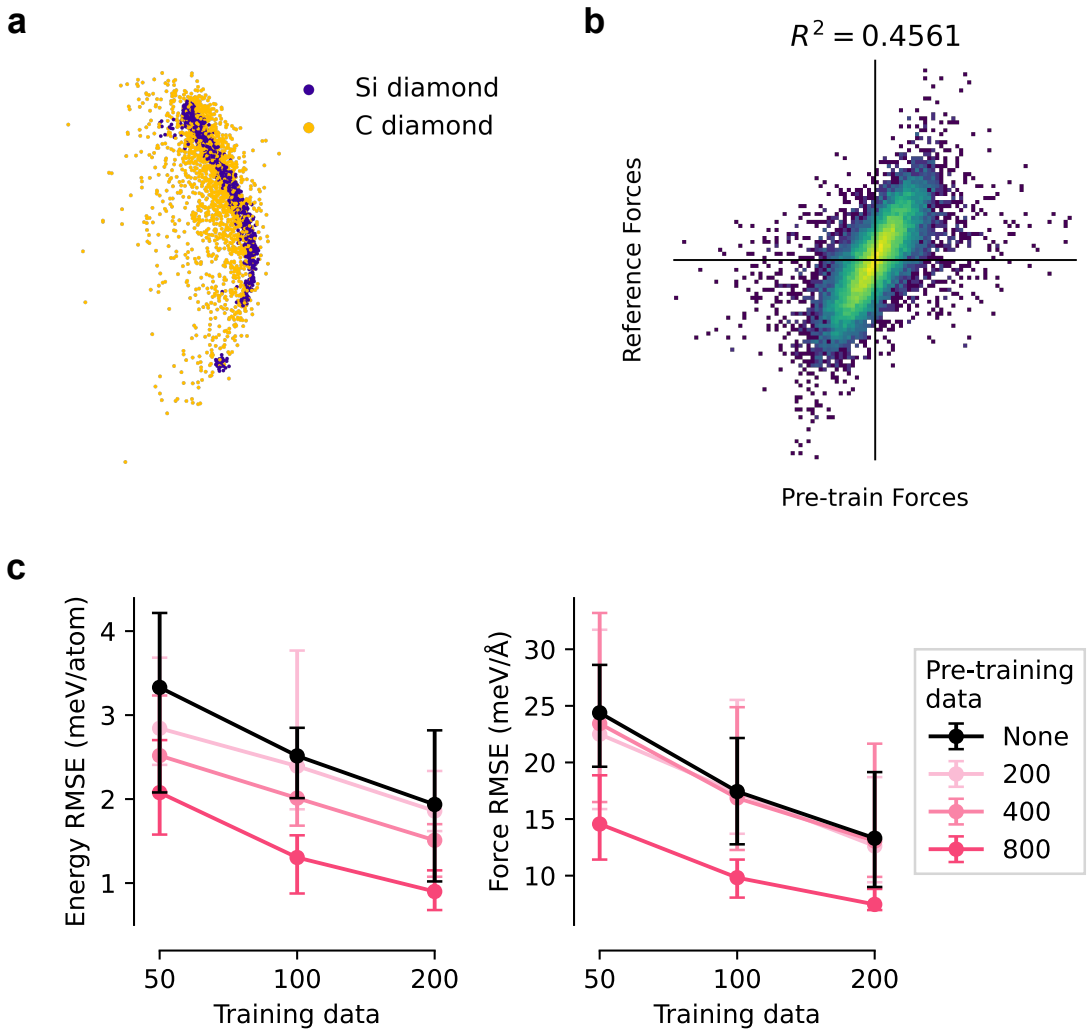

Figure 6.6: Positive transfer when both geometric similarity and PES alignment are satisfied. (a) PCA structure map showing strong overlap between the pre-training dataset (scaled C-dia, yellow) and the fine-tuning system (Si-dia, purple). (b) Correlation between pre-trained model force predictions and reference forces after fine-tuning, with $R^2 = 0.4561$, indicating moderate PES alignment. (c) Energy and force RMSE as a function of the number of fine-tuning data points for models pre-trained on varying amounts of diamond carbon data.

amounts of fine-tuning data. The findings show that geometric dissimilarity alone (Case 1) leads to negative transfer, even with large pre-training sets. In Case 2a, despite strong structural similarity, the absence of PES alignment prevents any performance gains. Only in Case 2b, where both conditions are satisfied, do we observe clear and consistent improvements in energy and force prediction accuracy. This confirms that both structural and energetic compatibility between the source and target systems are required to unlock the benefits of transfer learning in this context.

Table 6.1: Summary of transfer learning performance across three pre-training scenarios when fine-tuning on diamond silicon. Each row corresponds to one of the experimental cases described in Fig. 6.3. Geometric similarity between the pre-training and fine-tuning datasets is quantified by the revEMD score, and alignment of the PES is assessed via the R^2 correlation of predicted versus reference forces after fine-tuning. Energy and force transfer ratios (%) indicate the best-case relative improvement (or degradation) compared to a baseline trained only on dia-Si data. Positive transfer is only observed when both geometric similarity and PES alignment are satisfied (Case 2b).

Pre-train system	**revEMD**	**R^2**	**Energy transfer ratio (%)**	**Force transfer ratio (%)**
C-gra	0.0433	N/A	-2.4	-45.6
C-dia (alt. PES labels)	0.9163	0.0061	-4.2	0.03
C-dia	0.9163	0.4561	+53.5	+43.9

Case Study: Amorphous Carbon and Silicon

Building on the insights from the preceding crystalline systems, I now turn to a more complex and realistic material scenario: a-C and a-Si. These disordered systems present a greater challenge for interatomic potential modelling due to their structural variability and lack of long-range order, making them an ideal testbed for evaluating the practical utility of alchemical pre-training.

In the first experiment, I pre-trained on a-C structures and fine-tuned on a-Si. The learning curves (Figure 6.7c) clearly show strong positive transfer: even pre-training on as few as 200 a-C structures leads to a substantial reduction in both energy and force RMSE during fine-tuning.

Figure 6.7a visualises the structural distributions of the two datasets, revealing a reasonable degree of overlap. Although the revEMD score between them is 0.3025, formally indicating a relatively high level of dissimilarity, this can be somewhat misleading here. The structure map provides a more intuitive understanding of the relationship between the datasets: the a-Si distribution is largely embedded within the broader structural manifold of a-C. The low revEMD score therefore arises from the greater structural diversity of the a-C dataset, which spans a much wider space of

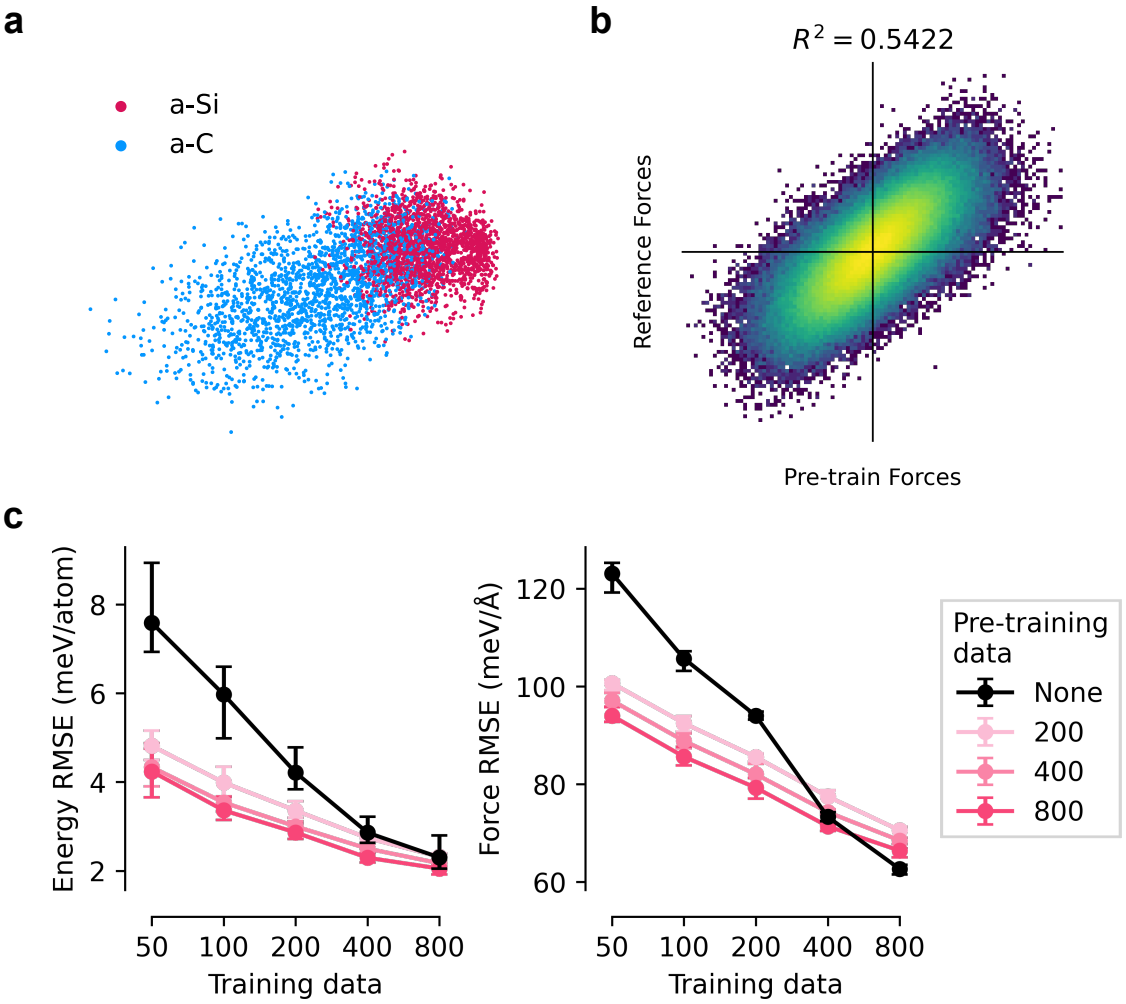

Figure 6.7: Pre-training on a-C and fine-tuning on a-Si. (a) PCA structure map showing the a-Si dataset (red) and a-C data (blue). (b) Correlation between pre-trained model force predictions and reference forces after fine-tuning, with $R^2 = 0.5422$. (c) Energy and force RMSE as a function of the number of fine-tuning data points for models pre-trained on varying amounts of a-C data.

bonding environments and local geometries, but does not account for the asymmetric overlap between the two distributions.

This asymmetry does not impede transfer; rather, it facilitates it. Pre-training on the structurally richer a-C dataset exposes the model to a superset of the environments found in a-Si, equipping it with generalised knowledge that transfers well. This is further supported by the PES alignment results, with the pre-trained model achieving a strong correlation with silicon reference data (Figure 6.7b, $R^2 = 0.5422$).

Notably, the extent of improvement here is far greater than that observed in the

earlier crystalline case (e.g., dia-C to dia-Si), where direct learning was relatively straightforward. In contrast, a-Si is a more challenging target, and the model benefits significantly from the additional prior knowledge provided by pre-training. This is particularly evident in the low-data regime, where the fine-tuned model substantially outperforms the directly trained baseline. As the volume of direct training data increases, the benefit of pre-training naturally diminishes, consistent with the notion that abundant task-specific data can eventually compensate for a lack of prior knowledge.

In the reverse experiment, pre-training on a-Si and fine-tuning on a-C, no meaningful benefit was observed. In fact, this training strategy consistently underperformed relative to direct training (Figure 6.8c). This asymmetry can be explained by the stark difference in structural diversity between the two systems. As discussed above and shown in Figure 6.8a, the a-Si dataset occupies a much narrower and more uniform region of structural space, confined largely to tetrahedral coordination environments. In contrast, the a-C dataset spans a wide range of bonding motifs and topologies. Consequently, pre-training on a-Si offers only a limited glimpse into the broader and more complex a-C configuration space, which restricts the utility of the learned representations for transfer.

This imbalance in structural diversity is clearly reflected in the poorer PES alignment: the pre-trained model exhibits only weak correlation with the carbon reference forces (Figure 6.8b, $R^2 = 0.2028$), in stark contrast to the much stronger alignment achieved when pre-training on a-C and fine-tuning on a-Si (Figure 6.7b, $R^2 = 0.5422$). The asymmetry in transfer performance is thus evident: pre-training on a-C is effective because the model is exposed to a structurally diverse dataset that encompasses the environments present in a-Si. In contrast, pre-training on a-Si is ineffective, as it provides only a narrow and biased sample of the structural configurations needed to model a-C.

Pre-training on such a limited PES appears to over-specialise the model, reducing

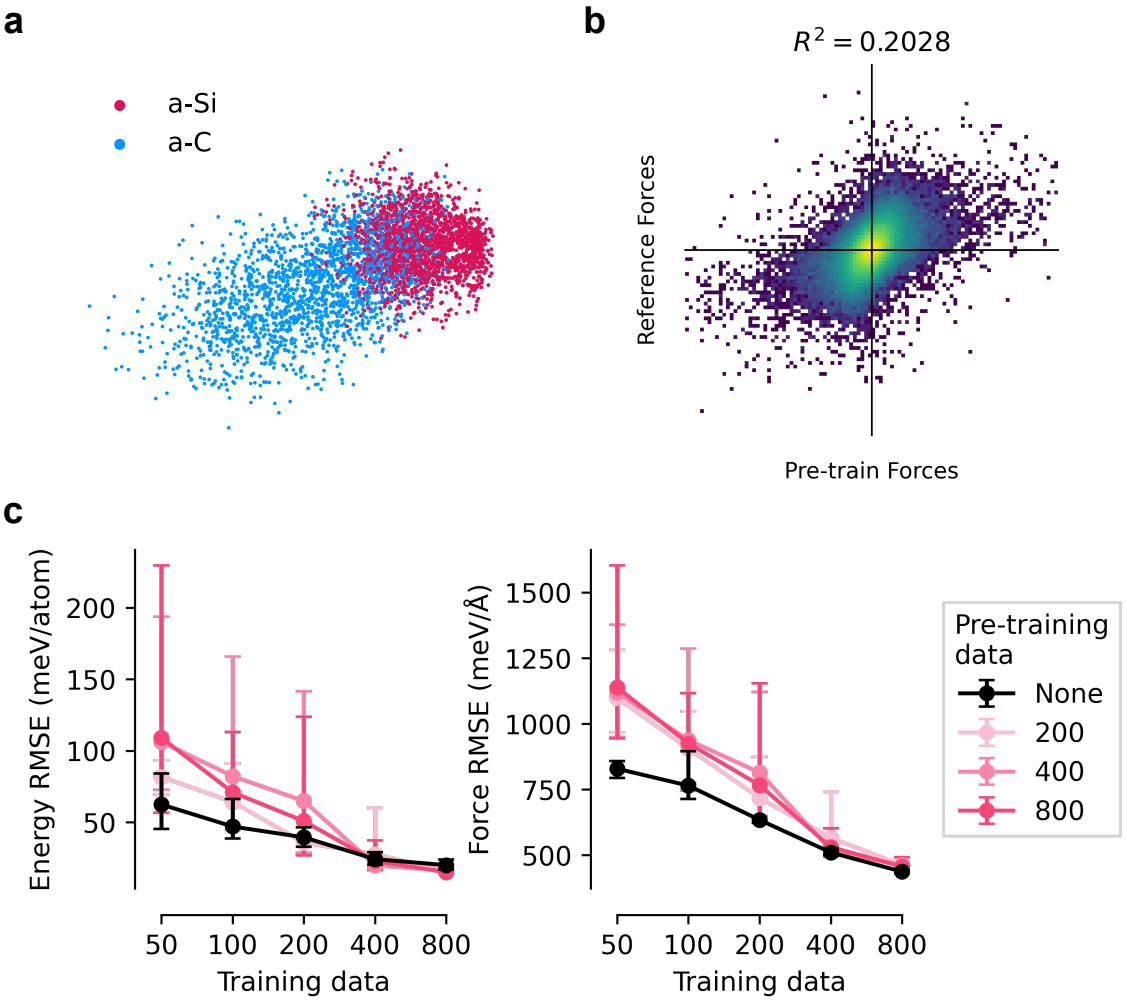

Figure 6.8: Pre-training on a-Si and fine-tuning on a-C. (a) PCA structure map showing the a-C dataset (blue) and a-Si data (red). (b) Correlation between pre-trained model force predictions and reference forces after fine-tuning, with $R^2 = 0.2028$. (c) Energy and force RMSE as a function of the number of fine-tuning data points for models pre-trained on varying amounts of a-Si data.

its flexibility during fine-tuning. The model becomes trapped in a restricted region of weight space and is unable to generalise to the broader chemical and structural variability of carbon, even when provided with substantial fine-tuning data. This result highlights a key caveat: successful transfer in one direction does not imply reciprocal success in the other. For alchemical transfer to be effective, the pre-training domain must be sufficiently representative and chemically rich to encompass the diversity of the target system.

To address the limitations encountered when transferring from a-Si to a-C, I

conducted an additional experiment in which the target a-C dataset was restricted to structures with a density of 2.9 g/cm^3 or higher. This subset corresponds to sp^3-rich carbon structures composed predominantly of tetrahedrally coordinated carbon atoms.

Under these conditions, pre-training on a-Si led to improved transfer performance, particularly in energy prediction (Figure 6.9c). This result aligns with the structure map (Figure 6.9a), which now exhibits more substantial, and crucially, more symmetric overlap between the datasets.

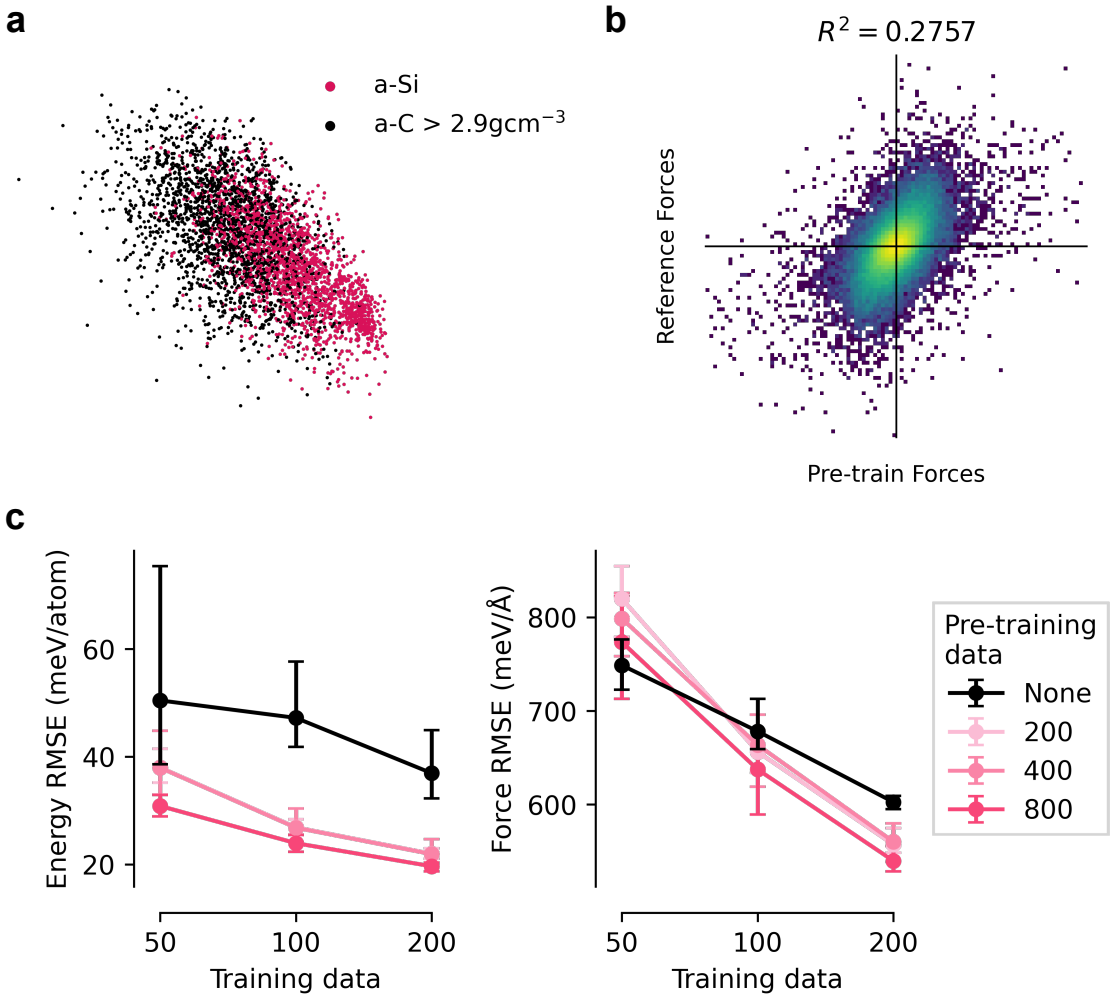

Figure 6.9: Pre-training on a-Si and fine-tuning on a-C with a density of 2.9 g/cm^3 or higher. (a) PCA structure map showing the a-C dataset (black) and a-Si data (red). (b) Correlation between pre-trained model force predictions and reference forces after fine-tuning, with $R^2 = 0.2757$. (c) Energy and force RMSE as a function of the number of fine-tuning data points for models pre-trained on varying amounts of a-Si data.

In this filtered regime, the high-density carbon structures lie well within the structural manifold spanned by the a-Si data. Correspondingly, the revEMD score increases to 0.6309 (compared to 0.3025 for the full a-C dataset), reflecting a greater degree of structural similarity. This refinement also yields a modest improvement in PES alignment (Figure 6.9b, $R^2 = 0.2757$), suggesting more effective knowledge transfer when the structural domains are better matched.

While the structural overlap is improved in the high-density subset, a degree of PES mismatch persists. Even at elevated densities, amorphous carbon retains significantly greater local structural diversity, including sp, sp^2, and sp^3 bonding environments, compared to the predominantly tetrahedral coordination found in amorphous silicon. Nevertheless, this experiment demonstrates that careful curation of the target domain can enhance transferability. When the geometric and energetic spaces of the source and target systems are more closely aligned, pre-training becomes more effective, enabling more accurate downstream predictions.

Extending to AB_2 systems

Building on insights from elemental systems like carbon and silicon, I now addressed a more complex challenge: cross-element transfer learning between chemically distinct AB_2-type systems. In particular, I examine whether representations learned from silica (SiO_2) can be usefully transferred to water (H_2O). While these systems differ significantly in bonding, elemental composition, and thermodynamic behaviour, both exhibit disordered tetrahedral networks. This structural similarity raises a natural question: can a shared network topology compensate for underlying chemical differences and enable successful transfer learning across such disparate systems?

As a first step, I pre-trained a model on the silica dataset from Erhard *et al.* and fine-tuned it on the water dataset from Ibrahim *et al.* A PCA analysis of SOAP descriptors, shown in Fig. 6.10a, reveals limited structural overlap between the two systems. A-type environments (Si in silica, O in water) exhibit partial alignment,

suggesting some shared geometric features. In contrast, B-type environments (O in silica, H in water) remain almost entirely disjoint, reflecting deeper discrepancies in local coordination. In silica, oxygen atoms are tetrahedrally coordinated in a symmetric environment by four silicon atoms, whereas in water, hydrogen atoms experience asymmetric surroundings, one covalent bond to oxygen and one hydrogen bond, resulting in fundamentally different local environments and descriptors

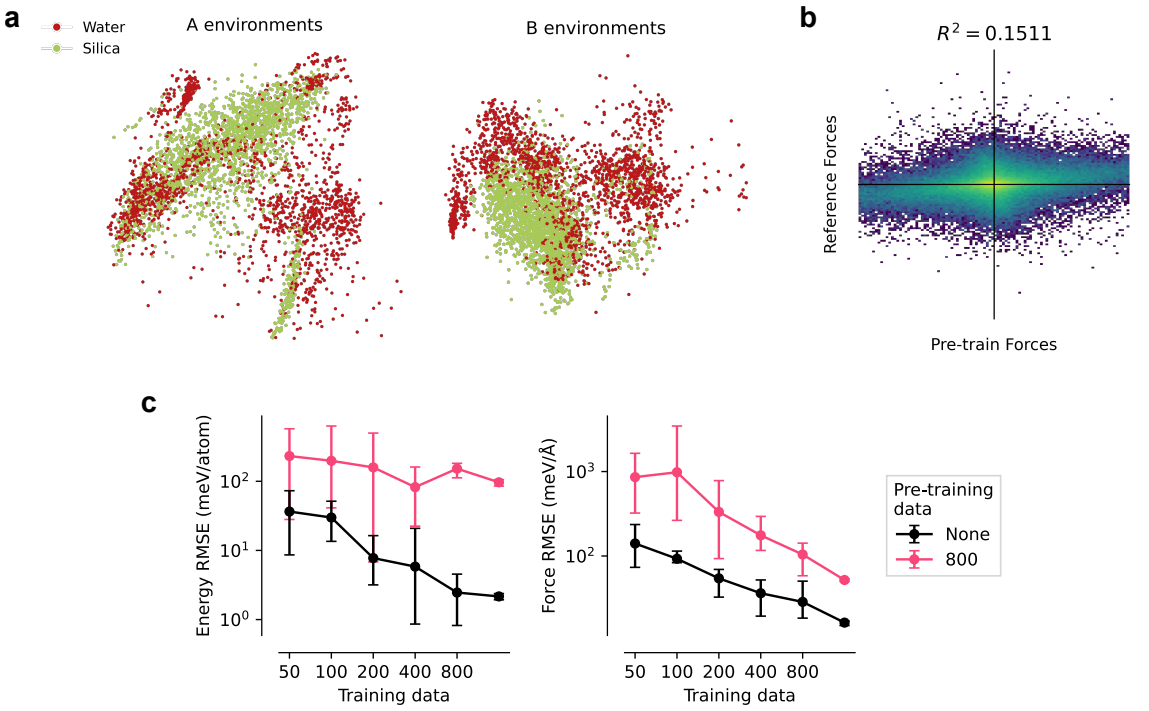

Figure 6.10: Pre-training on silica and fine-tuning on water. (a) PCA maps of SOAP descriptors for A (left) and B (right) atomic environments. (b) Correlation plot of force predictions from the pre-trained model vs. fine-tuning target forces ($R^2 = 0.1511$). (c) Energy and force RMSE as a function of training set size, comparing direct training (black) and fine-tuning after pre-training on silica (pink).

These geometric mismatches are echoed in the energy landscape. As shown in Fig. 6.10b, the correlation between force predictions from the pre-trained model and the true target forces after fine-tuning is extremely weak ($R^2 = 0.1511$), highlighting a fundamental incompatibility between the PES of silica and water.

The scale of this dissimilarity is illustrated in Fig. 6.10c, which presents energy and force RMSE across different training set sizes. Models pre-trained on silica (pink) consistently underperform relative to models trained directly on water data

(black). Rather than providing a helpful inductive prior, pre-training introduces biases that actively hinder learning — a clear indication that the learned representations do not transfer effectively, despite the shared tetrahedral topology. For clarity of presentation, only results from pre-training on 800 silica structures are shown; models pre-trained on smaller subsets (200 or 400 structures) performed even worse, severely distorting the figure axes and masking broader trends.

These results highlight a fundamental limitation of alchemical transfer learning: although geometric similarity and PES alignment are necessary conditions for successful transfer, they are rarely met in practice when dealing with chemically complex or compositionally diverse systems. In particular, transfer across AB_2-type compounds such as silica and water proves especially challenging. Consequently, a shared network topology alone is neither a sufficient nor reliable predictor of transferable learned representations.

To further explore the possibility of meaningful alchemical transfer across chemically distinct systems, I revisited a concept introduced in Chapter 5.3.1: coarse-graining. This approach is well established in the context of water modelling, with the mW potential [130] being a prominent example. In the mW model, each water molecule is represented by a single interaction site. Despite this simplification, the model captures many essential properties of real water – such as tetrahedral coordination, anomalous thermodynamics, and glassy behaviour – without explicitly including hydrogen atoms. The design of the mW potential was directly inspired by structural similarities between water and silicon and is, in fact, a modified form of the Stillinger–Weber potential [31], originally developed for silicon.

Motivated by these similarities, I coarse-grained the water dataset from Ibrahim *et al.* by removing hydrogen atoms and re-labelling the remaining oxygen-only structures using the mW potential [130]. The resulting dataset exhibits an energy landscape which should be similar to that of silicon, despite its distinct chemical origin. I used this cg-water data to pre-train a model, followed by fine-tuning on

the same amorphous silicon task introduced earlier.

Figure 6.11 presents the results. Although the force correlation is high ($R^2 = 0.8457$), indicating strong energetic alignment, the PCA structure map shows little geometric overlap between the coarse-grained water and amorphous silicon environments (revEMD score of 0.0816). Correspondingly, the learning curves reveal no statistically significant gains from pre-training.

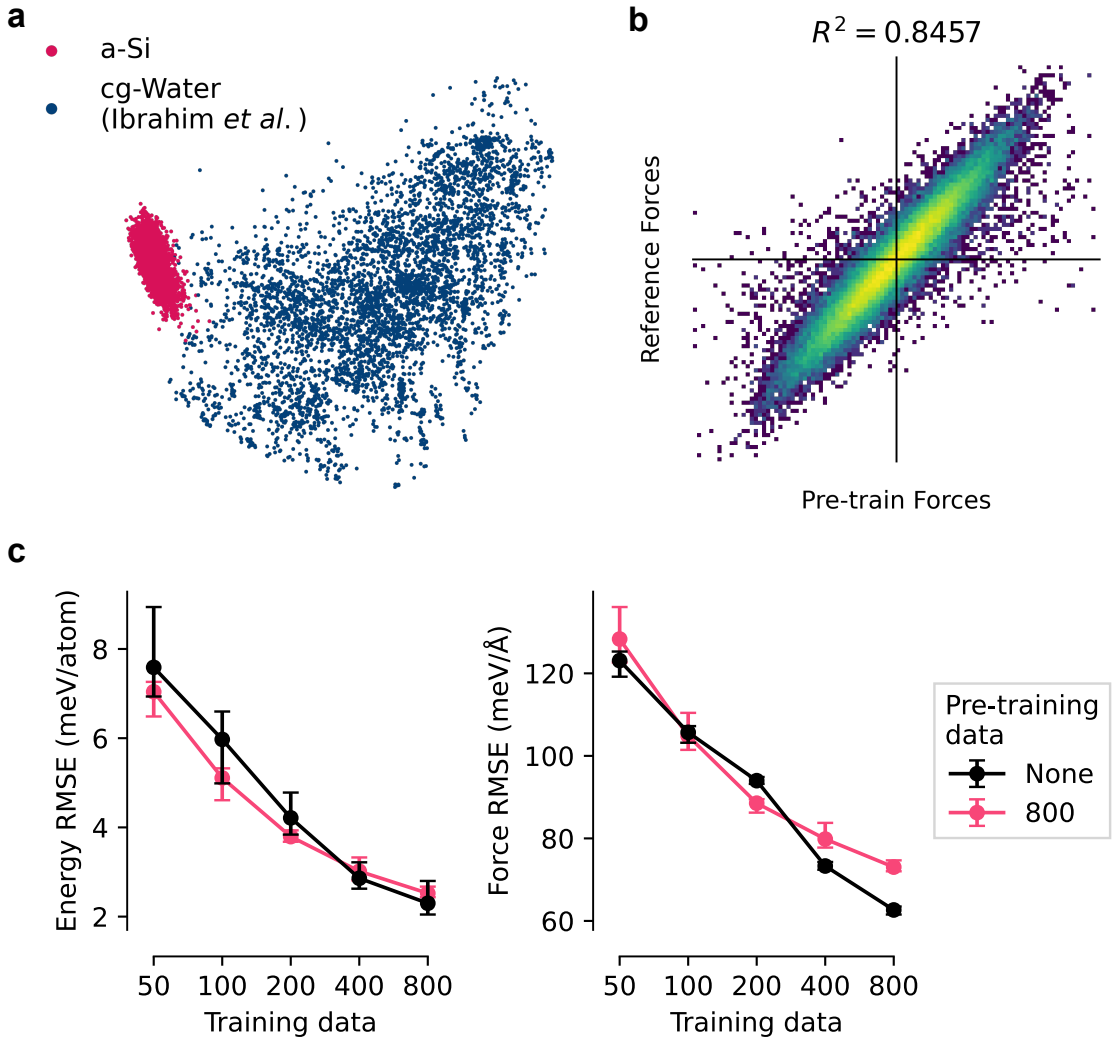

Figure 6.11: Pre-training on cg-water based on the dataset from Ibrahim *et al.* [346] and fine-tuning on a-Si. (a) PCA structure map showing the cg-water dataset (dark blue) and a-Si data (red). (b) Correlation between pre-trained model force predictions and reference forces after fine-tuning, with $R^2 = 0.8457$. (c) Energy and force RMSE as a function of the number of fine-tuning data points for models pre-trained on varying amounts of a-Si data.

This experiment reinforces a central theme of this chapter: geometric similarity is

a prerequisite for effective alchemical transfer. Even when energy landscapes align, a lack of overlap in atomic environments prevents the model from generalizing learned representations. It is the structural similarity between atomic environments, not merely the similarity of potential energy functional forms, that ultimately determines whether alchemical transfer is possible.

These findings suggest that while coarse-graining can be a useful conceptual bridge between chemically distinct systems, it does not eliminate the need for close structural alignment. In this case, the promising link between mW water and silicon does not translate into meaningful transfer performance, unless the structural manifold is also brought into alignment. This insight motivated a final test: coarse-graining a water dataset whose structure more closely resembles that of amorphous silicon.

To this end, I repeated the procedure using the Cheng *et al.* dataset (Section 6.3.1), which primarily consists of liquid water – unlike the ice-rich dataset from Ibrahim *et al.* (Section 4.3.1). As before, hydrogen atoms were removed, and the oxygen-only configurations were re-labelled using the mW potential. [130].

This change had a striking effect. As shown in Figure 6.12 (panel a), the coarse-grained liquid water configurations now occupy a region in PCA space much closer to that of amorphous silicon (revEMD score of 0.5006). Although chemically distinct, the two systems now share disordered, tetrahedral topologies that align more closely in local geometry. The PES alignment remains very strong ($R^2 = 0.8447$), and this time, pre-training leads to clear and consistent improvements in both energy and force predictions during fine-tuning.

This result is significant: it shows that meaningful transfer is possible even between systems that are far more chemically dissimilar than carbon and silicon. However, this experiment also highlights the limitations of the pre-training and fine-tuning approach. It required extensive manual effort, including coarse-graining, surrogate labelling, and careful dataset selection to engineer favourable overlap in

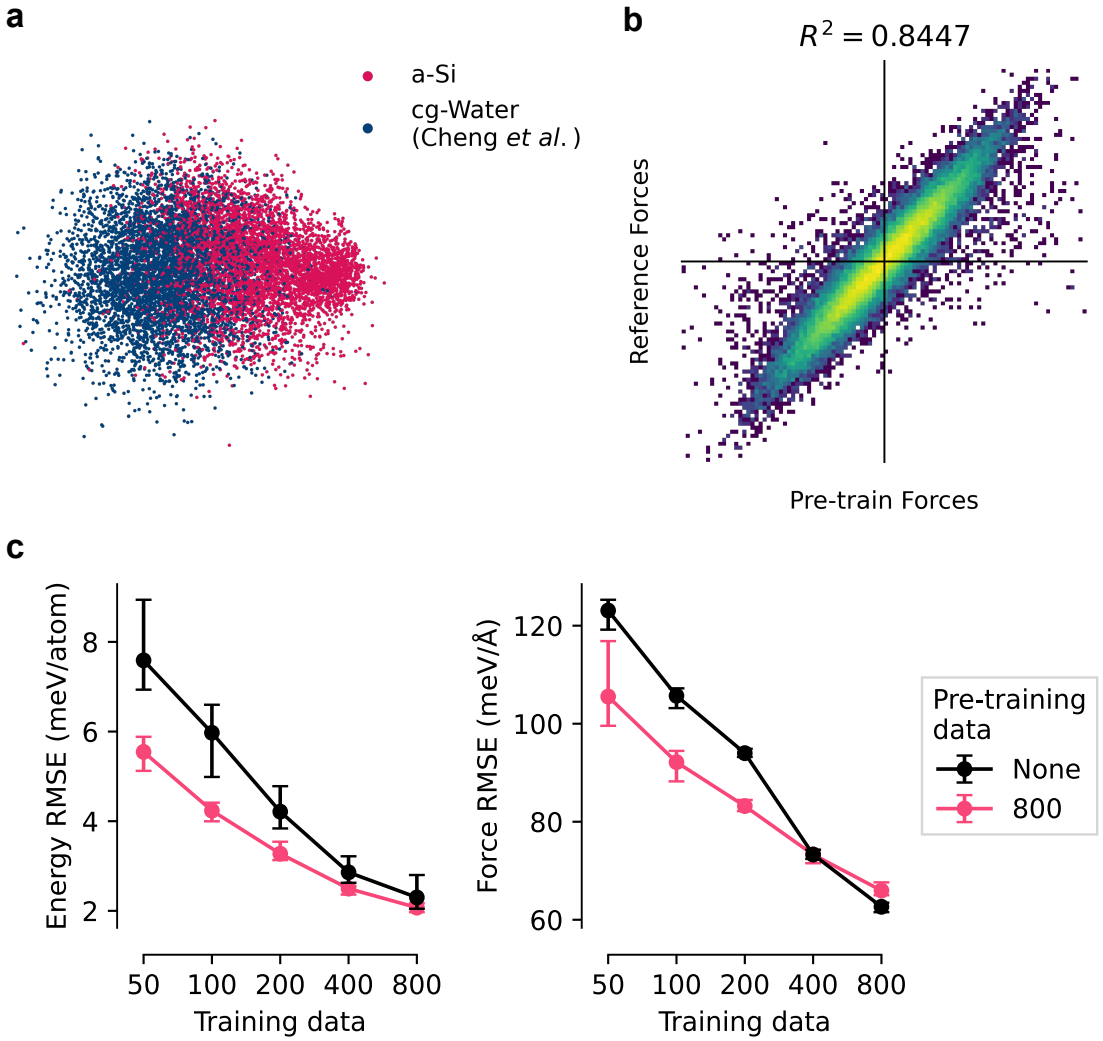

Figure 6.12: Pre-training on cg-water based on the dataset from Cheng *et al.* [388] and fine-tuning on a-Si. (a) PCA structure map showing the cg-water dataset (dark blue) and a-Si data (red). (b) Correlation between pre-trained model force predictions and reference forces after fine-tuning, with $R^2 = 0.8447$. (c) Energy and force RMSE as a function of the number of fine-tuning data points for models pre-trained on varying amounts of a-Si data.

descriptor space.

Across all experiments in this chapter, a clear conclusion emerges: pre-training and fine-tuning for alchemical transfer, while sometimes effective, is rarely practical for real-world applications. Even with deep domain knowledge, achieving meaningful transfer demands careful curation of both structure and PES. Successful cases often depend on prior chemical similarity or artificial alignment strategies, which limit generality. Although pre-training and fine-tuning represented the state-of-the-

art just a few years ago, this work shows that this approach in practice can be challenging and unreliable when applied across truly distinct systems.

In response to these challenges, a new class of models has begun to take centre stage: foundation models. These large, general-purpose interatomic potentials are trained on chemically diverse datasets spanning broad regions of the periodic table. Unlike traditional models tailored to specific materials, foundation models aim to generalise by default, learning chemically flexible representations that can be quickly fine-tuned with limited, high-quality data.

Crucially, foundation models do not require transmutation, coarse-graining, relabelling, or PES alignment tricks. They offer a more principled approach to alchemical transfer, one that scales naturally across domains and dramatically reduces the effort required to train high-accuracy models for new systems. In the next section, I explore this paradigm by fine-tuning and distilling a foundation model for liquid water, showing how the rich chemical knowledge encoded in a large pretrained model can be compressed into a lightweight, task-specific MLIP suitable for efficient molecular dynamics simulations.

6.4.2 Model Distillation

Gardner *et al.* [274] have demonstrated that distilled foundation models can achieve substantial improvements in efficiency and accuracy compared to, respectively, the original foundation model and directly trained counterparts with the same architecture. In particular, the distilled student models were over an order of magnitude faster than MACE-MP-0b3, while surpassing directly trained (on the same dataset) PaiNN and ACE models in accuracy by up to 70%.

To assess the practical utility of these distilled models in MD simulations, I evaluated their ability to reproduce key structural properties of liquid water. Following the distillation procedure outlined by Gardner *et al.* [274] and summarised in Section 6.3.5, the MACE-MP-0b3 model was fine-tuned using the dataset from Cheng *et al.* [388], then distilled into two significantly smaller models with different

architectures: one based on PaiNN [267] and the other on ACE [248].

My contribution focused on validating these distilled models in MD simulations. All models were benchmarked by simulating liquid water under the same conditions as used in Chapter 4 (Section 4.4.2). Each simulation started from a randomised box of 1024 water molecules, equilibrated at 300 K and 1 bar for 600 ps in the NPT ensemble, followed by a 1 ns production run in the NVE ensemble. Model performance was evaluated using structural metrics, including the RDF, ring-size distribution, tetrahedral order parameter, and hydrogen-bonding statistics, as detailed in Section 4.3.2.

All three models – the original MACE-MP-0b3 and its distilled variants – produced RDFs indicative of realistic local structure (Fig. 6.13a). The PaiNN model, with a $\sim$10$\times$ speed-up, closely replicated the teacher's RDF and experimental data. The ACE model, while approximately 100$\times$ faster, yielded a slightly over-structured liquid: the first RDF peak, corresponding to O $\cdots$ O contacts, was more pronounced than in experiment.

Further differences emerged in medium-range order, probed via the distribution of ring sizes (Fig.6.13b) and the tetrahedral order parameter q [378] (Fig. 6.13c). The PaiNN model tended to produce slightly more disordered environments than the teacher, while the ACE model favored more ordered, ice-like local structures, evidenced by a higher prevalence of six-membered rings and a tetrahedral order parameter distribution tending towards unity.

To gain additional insight, I examined the HB network topology using the AkDl classification scheme from Ref. 394 (previously used in Sec. 4.4.2), which characterises each water molecule by its number of HB acceptor (A) and donor (D) interactions. All models captured the characteristic diversity of hydrogen bonding in the liquid phase, in contrast to the uniform A2D2 arrangement of ice Ih (Fig. 6.13d). Consistent with earlier observations, the PaiNN model exhibited a slightly reduced fraction of crystal-like topologies, while the ACE model predicted a somewhat more

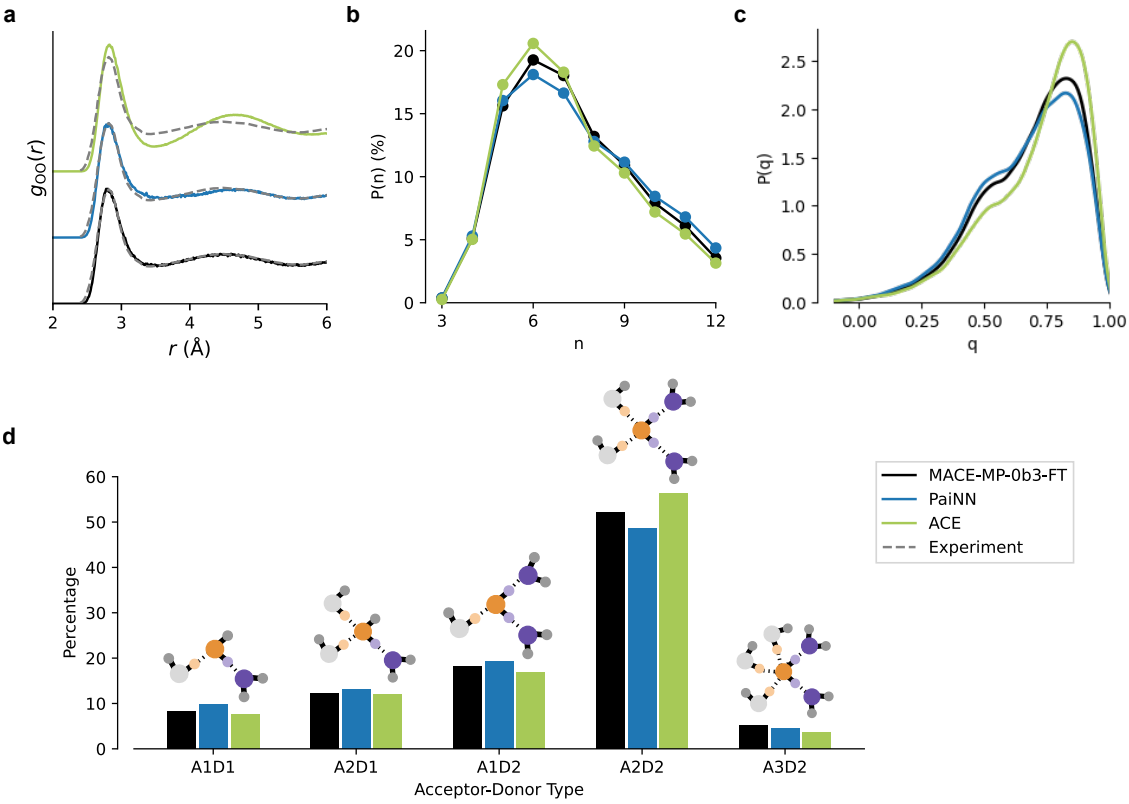

Figure 6.13: (a) Oxygen–oxygen RDF $g(r)$, comparing experimental data (dashed) with simulations using the teacher model (MACE-MP-0b3, black), PaiNN student (blue), and ACE student (green). (b) Probability distribution $P(n)$ of closed hydrogen-bonded rings containing n water molecules, with $n \in [3, 12]$. (c) Tetrahedral order parameter distribution $P(q)$, indicating local structural order. (d) Distribution of intact hydrogen bonds per water molecule, categorised by acceptor (A) and donor (D) roles. Each bar labeled A_xD_y represents the percentage of water molecules forming x acceptor and y donor hydrogen bonds. Schematic diagrams of these configurations are shown above each bar. Figure adapted from Ref. [274].

structured HBN.

Atomistic foundation models represent a conceptual and practical shift. Their broad chemical scope, pre-trained flexibility, and compatibility with distillation unlock new pathways for alchemical transfer that are both scalable and physically meaningful. When distilled into efficient architectures, such as PaiNN and ACE, these models preserve essential physical behaviour across scales and environments – even in the complex, disordered structure of liquid water.

6.5 Conclusion and Outlook

This chapter has explored two complementary strategies aimed at enhancing the generalisability and data efficiency of MLIPs: alchemical transfer learning and foundation model distillation. The central question underpinning this work was whether knowledge acquired in one chemical domain could meaningfully accelerate learning in another, particularly across systems differing in elemental composition but sharing a common tetrahedral motif.

A systematic investigation revealed that successful alchemical transfer hinges on two key factors: geometric similarity between atomic environments and alignment of potential energy surfaces. When both are satisfied, as in the case of amorphous carbon to silicon, significant performance gains were achieved, particularly in low-data settings. However, when either condition was lacking, transfer was ineffective or detrimental. Attempts to extend this approach to chemically dissimilar AB_2 systems like silica and water highlighted its limitations: extensive efforts such as coarse-graining and careful dataset selection were required for marginal performance gains. These results suggest that while alchemical transfer can be valuable in select cases, it remains a fragile and labour-intensive strategy with limited general applicability.

In contrast, foundation models represent a more principled and scalable pathway towards generalisation. Trained on large and chemically diverse datasets, these models learn rich, flexible representations that span a wide range of bonding types, elements, and structural motifs. Unlike traditional pretraining-fine-tuning pipelines, foundation models do not rely on transmutation or re-labelling and are inherently more adaptable to new domains. Furthermore, through model distillation, these models can be efficiently compressed into smaller, task-specific MLIPs suitable for deployment in practical simulations. The distilled models examined here retained high fidelity to their parent foundation model while delivering substantial speed-ups and producing physically realistic liquid water structures.

The rise of foundation models marks a conceptual turning point in the development of MLIPs. Their capacity to generalise across chemistry, coupled with their compatibility with lightweight architectures, positions them as promising default tools for atomistic simulations, particularly in regimes where data are scarce or systems are computationally expensive to generate. Looking ahead, foundation models that support streamlined fine-tuning for targeted applications are likely to replace bespoke pre-training and fine-tuning workflows that currently demand significant manual intervention.

In summary, the strategies discussed in this chapter – alchemical transfer, pretraining-fine-tuning, and foundation model distillation – highlight both the promise and the complexity of achieving robust, generalisable MLIPs.

Chapter 7

Conclusion & Outlook

This thesis has been concerned with two primary aims: first, to advance the atomistic understanding of disordered tetrahedral networks across multiple chemical systems; and second, to evaluate how machine learning can be used not only to model these disordered states accurately, but also to identify transferable structural patterns that transcend specific material classes. By bridging methods development, system-specific investigations, and cross-domain generalisation, the work has delivered a cohesive framework for interrogating and modelling disorder in complex materials. While each chapter has offered its own conclusions and outlined specific directions for future research, this final chapter reflects on the overarching progress made toward the thesis goals, and offers a brief perspective on the future role of machine learning in atomistic modelling.

7.1 Disordered Tetrahedral Networks

The first part of this thesis demonstrated how machine learning methods can be used to extract novel structural insights from individual disordered systems. Across the diverse systems studied — liquid water, amorphous ices, and ZIFs — a recurring theme has been the centrality of the tetrahedral motif. Whether in molecular liquids like water, or extended frameworks such as ZIFs, tetrahedral units form the fundamental building blocks of many structures. Yet, the ways in which these units connect, distort, or reorganise give rise to remarkable physical diversity. A core aim of this work has been to understand how this motif manifests and evolves across chemically and structurally distinct systems.

This investigation begins in Chapter 3 with the question of phase identity in amorphous ices, focusing in particular on the recently discovered MDA phase. To address this, I developed a neural network classifier based on BOO parameters to

distinguish between local atomic environments in amorphous ices. While the model successfully differentiated HDA from other phases, it revealed substantial structural overlap between LDA and MDA. This ambiguity suggests that MDA may not constitute a sharply defined phase, but rather an intermediate or transitional ensemble. This view is supported by recent work from de Almeida Ribeiro *et al.* [326] and Eltareb *et al.* [318], who interpret MDA as a nonequilibrium collection of states rather than a distinct thermodynamic basin. These findings highlight the limitations of static, local descriptors for capturing structural distinctions in systems where disorder is continuous.

Chapter 4 extended the focus to liquid water, developing a GNN based MLIP that reproduced key experimental features across a wide temperature range. Notably, I showed that the characteristic splitting of the structure factor in water is directly attributable to medium-range topological features of the hydrogen bond network, particularly the prevalence of 5–8 membered rings. These findings establish a direct, quantitative link between specific topological motifs and scattering features, demonstrating that ring statistics can serve as transferable descriptors of intermediate-range order. The prominent role of pentagonal rings in maintaining local heterogeneity and inhibiting crystallisation suggests a deeper, topological basis for water's anomalies.

Taken together, Chapters 3 and 4 deliver two complementary contributions. First, the structural classification tools developed for amorphous ices – based on BOO parameters and NNs – provide a scalable and interpretable way to describe and classify local environments in disordered systems. Second, the topological analysis of hydrogen-bonded rings in liquid water offers a framework for connecting medium-range motifs to experimentally accessible observables. This work shows how motif-level descriptions can bridge microscopic configurations and macroscopic signatures. Together, these developments offer both new tools for analysing complex materials and a conceptual foundation for a more unified, data-driven theory

of disordered tetrahedral networks.

I have already provided several possible next steps for each of these projects in the respective outlook sections of Chapters 3 and 4. Perhaps a more general future-facing observation is that these two analysis techniques are highly specialised to the specific tasks at hand, requiring labelled datasets and manual analyses respectively. Future work could focus on developing more automated or generalisable methods for analysis of amorphous structure. A promising approach could be to use un- or self-supervised approaches – e.g. contrastive learning [482–484] or autoencoders [485, 486] – that can discover emergent categories or clusters of local amorphous structure without prior bias. Such approaches remove the need for manual labelling and analyses of structures, unlocking the use of existing, large-scale datasets. By coupling such methods to graph neural networks, which can learn representations of local environment directly from atomic graphs [487], the framework could also progress from hand-engineered Steinhardt descriptors toward fully data-driven embeddings capable of capturing subtle distortions or coordination changes. This expanded toolkit could enable intuitive, automated tracking of structural evolution in a host of important processes – including doping-induced perturbations, nanoconfinement, melting phenomena, and both first- and second-order phase transitions – providing a unified lens through which to interrogate the rich landscape of tetrahedral disorder.

7.2 Transferability of Machine Learning Models

If the first half of this thesis focused on *what* can be learned about disorder in tetrahedral systems, the second half turned to *how* such learning can be made transferable. In particular, the key question became whether models trained on one tetrahedral material could be applied to another, probing the extent to which learned representations generalise across chemically and structurally distinct classes.

Chapter 5 examined this question from a coarse-graining perspective, using GPR models to predict local energies in ZIFs from structural representations of varying

resolution. Remarkably, the coarse-grained AB_2 model, which preserves only tetrahedral connectivity between nodes and linkers, retained much of the predictive accuracy of the fully atomistic model, despite a drastic reduction in the number of degrees of freedom. This result supports the long-standing analogy between ZIFs and zeolites, and more broadly demonstrates that simplified, topology-aware representations can encode the essential energetics of disordered materials.

Chapter 6 extended the idea of transferability to chemically diverse systems by evaluating the potential of alchemical transfer learning. Promising results were observed in structurally similar systems, such as between amorphous carbon and silicon, where both the local environments and underlying energy landscapes were well aligned. However, outside this narrow corridor of compatibility, transfer proved fragile. Water and silica, for example, share a tetrahedral motif but diverge in local chemistry; no amount of surrogate labelling could bridge that gap in a reliable way. These findings highlight that the primary barrier to generalisation is not the learning algorithm itself, but the (in)adequacy of the underlying representation to span multiple chemical domains.

Over the course of my DPhil research, I have witnessed and made use of the rapid development of atomistic foundation models. It is not difficult to envision a (near) future in which such models underpin a vast majority of automated, high-throughput workflows for materials discovery and property prediction, as well as in forming the basis for all manner of novel methodologies (such as the distillation procedure I made use of in Chapter 6). With particular relevance to this thesis, the emergence of foundation models to me marks an inflection point in the modelling of disordered materials: by leveraging large-scale pretraining and flexible architectures, these models promise to drastically accelerate the analysis of amorphous materials by removing the need to train new ML potential models *from scratch* for each new system of interest. Fine-tuning and distillation workflows like those I have used in Chapter 6 are already automated and quick to use; as general-purpose datasets and

model architectures continue to improve, it is even plausible that foundation models will be broadly applicable out-of-the-box, completely removing the need for manual training of new models.

Nevertheless, many challenges still remain in the understanding of amorphous materials, even if foundation models deliver on the promise of near perfect 0-shot performance. Central among these is model interpretability: despite their impressive accuracy, foundation models remain largely "opaque". We currently lack tools to understand what features these models prioritise, how they encode physical priors, or how best to guide their behaviour when extending to new materials classes. Hence, while foundation models allow us to run faithful simulations over varied chemical domains, they are not (yet) able to directly "tell" us which structural features give rise to a given property, to directly guide the design of new materials, or to provide mechanistic insights into important phenomena without detailed analysis of extensive simulation data. For now, at least, expert human analysis is still very much required.

A recurring limitation of most current machine learning interatomic potentials, including those used in this work, is the reliance on a locality assumption: atomic energies are expressed as a sum of contributions depending only on neighbouring environments within a finite cutoff. While this decomposition enables efficient and accurate learning of short-range interactions, it inherently neglects non-local physical effects, and in particular long-range electrostatics, polarisation, which are central to the behaviour of many disordered materials. A promising direction for future research is the explicit integration of such long-range physics into machine learning models, either through hybrid ML–physics formalisms, or via message-passing schemes with extended receptive fields. Combining the flexibility of data-driven representations with explicit physical constraints could yield models that are both transferable and faithful to the underlying physics, providing a more realistic foundation for modelling disorder across chemical and structural domains.

In summary, while the tools for learning from disorder are rapidly maturing, true generalisation across chemistry, structure, and function remains a work in progress. Realising this goal will require not just better models, but better representations, better data, and a deeper dialogue between physics, chemistry, and machine learning. The results in this thesis – both positive and cautionary – point toward a future in which general-purpose, interpretable, and physically grounded ML models are the norm rather than the exception.

Bibliography

[1] M. Miodownik, *The Guardian* (2014). https://www.theguardian.com/science/2014/sep/14/story-of-materials-human-civilisation-mark-miodownik (Accessed: 12/05/2025).

[2] S. Baker, *Nature* **636**, S1 (2024).

[3] S. P. Stier, *et al.*, *Advanced Materials* **36**, 2407791 (2024).

[4] Joint Research Centre (European Commission), *SETIS magazine* (2015). Catalogue number: LD-AD-15001-EN-N (PDF), LD-AD-15001-EN-C (Paper).

[5] AMi2030, Materials 2030 Manifesto: Systemic approach of advanced materials for prosperity — a 2030 perspective (2022). https://beda.org/wp-content/uploads/2023/02/advanced-materials-2030-manifesto.pdf (Accessed: 17/06/2025).

[6] T. Hey, S. Tansley, K. Tolle, J. Gray, *The Fourth Paradigm: Data-Intensive Scientific Discovery* (Microsoft Research, 2009).

[7] Y. Xu, *et al.*, *The Innovation* **2**, 100179 (2021).

[8] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, K. A. Persson, *APL Materials* **1**, 011002 (2013).

[9] S. Kirklin, J. E. Saal, B. Meredig, A. Thompson, J. W. Doak, M. Aykol, S. Rühl, C. Wolverton, *npj Computational Materials* **1**, 15010 (2015).

[10] A. Aldossary, J. A. Campos-Gonzalez-Angulo, S. Pablo-García, S. X. Leong, E. M. Rajaonson, L. Thiede, G. Tom, A. Wang, D. Avagliano, A. Aspuru-Guzik, *Advanced Materials* **36**, 2402369 (2024).

[11] Y. Liu, T. Zhao, W. Ju, S. Shi, *Journal of Materiomics* **3**, 159 (2017).

[12] A. Agrawal, A. Choudhary, *APL Materials* **4**, 053208 (2016).

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention Is All You Need (2023). DOI: 10.48550/arXiv.1706.03762.

[14] R. S. Sutton, F. Bach, *Reinforcement Learning – An Introduction* (Cambridge, Massachusetts, 2014).

[15] S. Gu, E. Holly, T. Lillicrap, S. Levine, Deep Reinforcement Learning for Robotic Manipulation with Asynchronous Off-Policy Updates (2016). DOI: 10.48550/arXiv.1610.00633.

[16] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, D. Hassabis, *Nature* **529**, 484 (2016).

[17] M. H. Mobarak, M. A. Mimona, M. A. Islam, N. Hossain, F. T. Zohura, I. Imtiaz, M. I. H. Rimon, *Applied Surface Science Advances* **18**, 100523 (2023).

[18] A. Jain, *Current Opinion in Solid State and Materials Science* **33**, 101189 (2024).

[19] L. A. Ruiz Pestana, Y. Liao, Z. Li, W. Xia, *Fundamentals of Multiscale Modeling of Structural Materials*, W. Xia, L. A. Ruiz Pestana, eds. (2023), pp. 37–73.

[20] B. J. Alder, T. E. Wainwright, *The Journal of Chemical Physics* **27**, 1208 (1957).

[21] A. Rahman, *Physical Review* **136**, A405 (1964).

[22] G. Hautier, A. Jain, S. P. Ong, *Journal of Materials Science* **47**, 7317 (2012).

[23] A. Warshel, M. Levitt, *Journal of Molecular Biology* **103**, 227 (1976).

[24] A. Warshel, M. Levitt, *Journal of Molecular Biology* **106**, 421 (1976).

[25] Q. Cui, G. Li, J. Ma, M. Karplus, *Journal of Molecular Biology* **340**, 345 (2004).

[26] P. Maragakis, M. Karplus, *Journal of Molecular Biology* **352**, 807 (2005).

[27] H. Hodak, *Journal of Molecular Biology* **426**, 1 (2014).

[28] M. H. Müser, S. , Sergey V., L. and Pastewka, *Advances in Physics: X* **8**, 2093129 (2023).

[29] J. E. Jones, S. Chapman, *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* **106**, 463 (1997).

[30] J. E. Jones, S. Chapman, *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* **106**, 709 (1997).

[31] F. H. Stillinger, T. A. Weber, *Physical Review B* **31**, 5262 (1985).

[32] J. Tersoff, *Physical Review B* **37**, 6991 (1988).

[33] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, J. Hermans, *Intermolecular Forces: Proceedings of the Fourteenth Jerusalem Symposium on Quantum Chemistry and Biochemistry Held in Jerusalem, Israel, April 13–16, 1981*, B. Pullman, ed. (Dordrecht, 1981), pp. 331–342.

[34] H. J. C. Berendsen, J. R. Grigera, T. P. Straatsma, *The Journal of Physical Chemistry* **91**, 6269 (1987).

[35] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, M. L. Klein, *The Journal of Chemical Physics* **79**, 926 (1983).

[36] R. L. C. Vink, G. T. Barkema, M. A. Stijnman, R. H. Bisseling, *Physical Review B* **64**, 245214 (2001).

[37] R. M. Martin, *Electronic Structure: Basic Theory and Practical Methods* (Cambridge, 2004).

[38] M. T. Dove, *Seminarios de la Sociedad Española de Mineralogía* **4**, 7 (2006).

[39] D. S. Sholl, J. A. Steckel, *Density Functional Theory: A Practical Introduction* (2009), first edn.

[40] P. Hohenberg, W. Kohn, *Physical Review* **136**, B864 (1964).

[41] W. Kohn, L. J. Sham, *Physical Review* **140**, A1133 (1965).

[42] M. D. Segall, P. J. D. Lindan, M. J. Probert, C. J. Pickard, P. J. Hasnip, S. J. Clark, M. C. Payne, *Journal of Physics: Condensed Matter* **14**, 2717 (2002).

[43] M. C. Payne, M. P. Teter, D. C. Allan, T. A. Arias, J. D. Joannopoulos, *Reviews of Modern Physics* **64**, 1045 (1992).

[44] A. Zakutayev, X. Zhang, A. Nagaraja, L. Yu, S. Lany, T. O. Mason, D. S. Ginley, A. Zunger, *Journal of the American Chemical Society* **135**, 10048 (2013).

[45] R. Gautier, X. Zhang, L. Hu, L. Yu, Y. Lin, T. O. L. Sunde, D. Chon, K. R. Poeppelmeier, A. Zunger, *Nature Chemistry* **7**, 308 (2015).

[46] Y. Hinuma, T. Hatakeyama, Y. Kumagai, L. A. Burton, H. Sato, Y. Muraba, S. Iimura, H. Hiramatsu, I. Tanaka, H. Hosono, F. Oba, *Nature Communications* **7**, 11962 (2016).

[47] D. C. Langreth, M. J. Mehl, *Physical Review B* **28**, 1809 (1983).

[48] A. D. Becke, *Physical Review A* **38**, 3098 (1988).

[49] F. Noé, A. Tkatchenko, K.-R. Müller, C. Clementi, *Annual Review of Physical Chemistry* **71**, 361 (2020).

[50] K. T. Schütt, S. Chmiela, O. A. Von Lilienfeld, A. Tkatchenko, K. Tsuda, K.-R. Müller, eds., *Machine Learning Meets Quantum Physics*, vol. 968 of *Lecture Notes in Physics* (Cham, 2020).

[51] V. L. Deringer, M. A. Caro, G. Csányi, *Advanced Materials* **31**, 1902765 (2019).

[52] J. Behler, M. Parrinello, *Physical Review Letters* **98**, 146401 (2007).

[53] O. T. Unke, S. Chmiela, H. E. Sauceda, M. Gastegger, I. Poltavsky, K. T. Schütt, A. Tkatchenko, K.-R. Müller, *Chemical Reviews* **121**, 10142 (2021).

[54] J. D. Morrow, C. Ugwumadu, D. A. Drabold, S. R. Elliott, A. L. Goodwin, V. L. Deringer, *Angewandte Chemie International Edition* **63**, e202403842 (2024).

[55] L. C. Erhard, J. Rohrer, K. Albe, V. L. Deringer, *npj Computational Materials* **8**, 1 (2022).

[56] L. C. Erhard, J. Rohrer, K. Albe, V. L. Deringer, *Nature Communications* **15**, 1927 (2024).

[57] Y. Zhou, W. Zhang, E. Ma, V. L. Deringer, *Nature Electronics* **6**, 746 (2023).

[58] Y. Zhuo, A. Mansouri Tehrani, J. Brgoch, *The Journal of Physical Chemistry Letters* **9**, 1668 (2018).

[59] N. Linton, D. S. Aidhy, *APL Machine Learning* **1**, 016109 (2023).

[60] C. Ben Mahmoud, A. Anelli, G. Csányi, M. Ceriotti, *Physical Review B* **102**, 235130 (2020).

[61] A. Talapatra, B. P. Uberuaga, C. R. Stanek, G. Pilania, *Communications Materials* **4**, 46 (2023).

[62] S. G. Baird, T. Q. Diep, T. D. Sparks, *Digital Discovery* **1**, 226 (2022).

[63] T. C. Nicholas, E. V. Alexandrov, V. A. Blatov, A. P. Shevchenko, D. M. Proserpio, A. L. Goodwin, V. L. Deringer, *Chemistry of Materials* **33**, 8289 (2021).

[64] W. Liao, R. Yuan, X. Xue, J. Wang, J. Li, T. Lookman, *npj Computational Materials* **10**, 171 (2024).

[65] C. Zeni, *et al.*, *Nature* **639**, 624 (2025).

[66] S. R. Elliott, *Nature* **354**, 445 (1991).

[67] W. H. Zachariasen, *Journal of the American Chemical Society* **54**, 3841 (1932).

[68] K. S. Park, Z. Ni, A. P. Côté, J. Y. Choi, R. Huang, F. J. Uribe-Romo, H. K. Chae, M. O'Keeffe, O. M. Yaghi, *Proceedings of the National Academy of Sciences* **103**, 10186 (2006).

[69] T. D. Bennett, A. L. Goodwin, M. T. Dove, D. A. Keen, M. G. Tucker, E. R. Barney, A. K. Soper, E. G. Bithell, J.-C. Tan, A. K. Cheetham, *Physical Review Letters* **104**, 115503 (2010).

[70] N. Giovambattista, P. G. Debenedetti, F. Sciortino, H. E. Stanley, *Physical Review E* **71**, 061505 (2005).

[71] P. H. Handle, T. Loerting, F. Sciortino, *Proceedings of the National Academy of Sciences* **114**, 13336 (2017).

[72] T. Head-Gordon, M. E. Johnson, *Proceedings of the National Academy of Sciences* **103**, 7973 (2006).

[73] G. N. Clark, C. , Christopher D., S. , Jared D., S. , Richard J., T. and Head-Gordon, *Molecular Physics* **108**, 1415 (2010).

[74] J. K. Christie, *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **381**, 20220251 (2023).

[75] B. E. Warren, J. Biscce, *Journal of the American Ceramic Society* **21**, 49 (1938).

[76] G. Lucovsky, J. C. Phillips, *physica status solidi (b)* **246**, 1806 (2009).

[77] Y. A. Gutiérrez Fosado, D. Michieletto, F. Martelli, *Physical Review Letters* **133**, 266102 (2024).

[78] R. Xie, G. G. Long, S. J. Weigand, S. C. Moss, T. Carvalho, S. Roorda, M. Hejna, S. Torquato, P. J. Steinhardt, *Proceedings of the National Academy of Sciences* **110**, 13250 (2013).

[79] F. Martelli, S. Torquato, N. Giovambattista, R. Car, *Physical Review Letters* **119**, 136002 (2017).

[80] A. Gabrielli, M. Joyce, F. Sylos Labini, *Physical Review D* **65**, 083523 (2002).

[81] S. Torquato, F. H. Stillinger, *Physical Review E* **68**, 041113 (2003).

[82] S. Gorsky, W. A. Britton, Y. Chen, J. Montaner, A. Lenef, M. Raukas, L. Dal Negro, *APL Photonics* **4**, 110801 (2019).

[83] M. M. Milošević, W. Man, G. Nahal, P. J. Steinhardt, S. Torquato, P. M. Chaikin, T. Amoah, B. Yu, R. A. Mullen, M. Florescu, *Scientific Reports* **9**, 20338 (2019).

[84] Y. Xu, S. Chen, P.-E. Chen, W. Xu, Y. Jiao, *Physical Review E* **96**, 043301 (2017).

[85] N. Muller, J. Haberko, C. Marichy, F. Scheffold, *Advanced Optical Materials* **2**, 115 (2014).

[86] C. Giacovazzo, H. L. Monaco, G. Artioli, D. Viterbo, M. Milanesio, G. Gilli, P. Gilli, G. Zanotti, G. Ferraris, M. Catti, *Fundamentals of Crystallography* (Oxford University Press, Oxford, New York, 2011), third edn.

[87] E. Lerner, I. Procaccia, C. Rainone, M. Singh, *Physical Review E* **98**, 063001 (2018).

[88] H. Tong, S. Sengupta, H. Tanaka, *Nature Communications* **11**, 4863 (2020).

[89] L. A. M. Rosset, D. A. Drabold, V. L. Deringer, *Nature Communications* **16**, 2360 (2025).

[90] V. L. Deringer, G. Csányi, *Physical Review B* **95**, 094203 (2017).

[91] A. P. Bartók, J. Kermode, N. Bernstein, G. Csányi, *Physical Review X* **8**, 041048 (2018).

[92] M. L. Bødker, M. Bauchy, T. Du, J. C. Mauro, M. M. Smedskjaer, *npj Computational Materials* **8**, 192 (2022).

[93] C. G. Staacke, H. H. Heenen, C. Scheurer, G. Csányi, K. Reuter, J. T. Margraf, *ACS Applied Energy Materials* **4**, 12562 (2021).

[94] G. C. Sosso, G. Miceli, S. Caravati, J. Behler, M. Bernasconi, *Physical Review B* **85**, 174103 (2012).

[95] I. Vincze, F. Van der Woude, *Journal of Non-Crystalline Solids* **42**, 499 (1980).

[96] J. Mavračić, F. C. Mocanu, V. L. Deringer, G. Csányi, S. R. Elliott, *The Journal of Physical Chemistry Letters* **9**, 2985 (2018).

[97] P. Keblinski, S. R. Phillpot, D. Wolf, H. Gleiter, *Acta Materialia* **45**, 987 (1997).

[98] R. Maharana, D. Das, P. Chaudhuri, K. Ramola, *Physical Review E* **109**, 044903 (2024).

[99] Pablo G Debenedetti, *Journal of Physics: Condensed Matter* **15**, R1669 (2003).

[100] C. A. Angell, *Annual Review of Physical Chemistry* **34**, 593 (1983).

[101] D. Eisenberg, W. Kauzmann, *The Structure and Properties of Water* (Oxford University Press, Oxford, New York, 2005).

[102] 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, L. G. M. Pettersson, *Chemical Reviews* **116**, 7463 (2016).

[103] J. C. Palmer, P. H. Poole, F. Sciortino, P. G. Debenedetti, *Chemical Reviews* **118**, 9129 (2018).

[104] D. Liu, Y. Zhang, C.-C. Chen, C.-Y. Mou, P. H. Poole, S.-H. Chen, *Proceedings of the National Academy of Sciences* **104**, 9570 (2007).

[105] P. G. Debenedetti, F. Sciortino, G. H. Zerze, *Science* **369**, 289 (2020).

[106] L. Xu, P. Kumar, S. V. Buldyrev, S.-H. Chen, P. H. Poole, F. Sciortino, H. E. Stanley, *Proceedings of the National Academy of Sciences* **102**, 16558 (2005).

[107] R. Shi, H. Tanaka, *Proceedings of the National Academy of Sciences* **117**, 26591 (2020).

[108] A. Nilsson, L. G. M. Pettersson, *Nature Communications* **6**, 8998 (2015).

[109] T. Morawietz, A. Singraber, C. Dellago, J. Behler, *Proceedings of the National Academy of Sciences* **113**, 8368 (2016).

[110] J. Sun, B. K. Clark, S. Torquato, R. Car, *Nature Communications* **6**, 8156 (2015).

[111] Y. Takii, K. Koga, H. Tanaka, *The Journal of Chemical Physics* **128**, 204501 (2008).

[112] J. L. Aragones, M. M. Conde, E. G. Noya, C. Vega, *Physical Chemistry Chemical Physics* **11**, 543 (2009).

[113] A. Henao, J. M. Salazar-Rios, E. Guardia, L. C. Pardo, *The Journal of Chemical Physics* **154**, 104501 (2021).

[114] G. Cassone, F. Martelli, *Nature Communications* **15**, 1856 (2024).

[115] K. H. Kim, *et al.*, *Science* **370**, 978 (2020).

[116] J. A. Barker, R. O. Watts, *Chemical Physics Letters* **3**, 144 (1969).

[117] A. Rahman, F. H. Stillinger, *The Journal of Chemical Physics* **55**, 3336 (1971).

[118] A. Ben-Naim, F. H. Stihinger, M. Hill, *Aspects of the Statistical-Mechanical Theory of Water* pp. 295–330 (1972).

[119] F. H. Stillinger, A. Rahman, *The Journal of Chemical Physics* **60**, 1545 (1974).

[120] J. L. F. Abascal, C. Vega, *The Journal of Chemical Physics* **123**, 234505 (2005).

[121] D. J. Price, C. L. Brooks, III, *The Journal of Chemical Physics* **121**, 10096 (2004).

[122] H. W. Horn, W. C. Swope, J. W. Pitera, J. D. Madura, T. J. Dick, G. L. Hura, T. Head-Gordon, *The Journal of Chemical Physics* **120**, 9665 (2004).

[123] R. Fuentes-Azcatl, J. Alejandre, *The Journal of Physical Chemistry B* **118**, 1263 (2014).

[124] J. L. F. Abascal, E. Sanz, R. García Fernández, C. Vega, *The Journal of Chemical Physics* **122**, 234511 (2005).

[125] M. W. Mahoney, W. L. Jorgensen, *The Journal of Chemical Physics* **112**, 8910 (2000).

[126] Y. Khalak, B. Baumeier, M. Karttunen, *The Journal of Chemical Physics* **149**, 224507 (2018).

[127] M. A. González, J. L. F. Abascal, *The Journal of Chemical Physics* **135**, 224516 (2011).

[128] T. Yagasaki, M. Matsumoto, H. Tanaka, *Physical Review E* **89**, 020301 (2014).

[129] F. Martelli, J. C. Palmer, *The Journal of Chemical Physics* **156**, 114502 (2022).

[130] V. Molinero, E. B. Moore, *The Journal of Physical Chemistry B* **113**, 4008 (2009).

[131] K. Laasonen, M. Sprik, M. Parrinello, R. Car, *The Journal of Chemical Physics* **99**, 9080 (1993).

[132] M. E. Tuckerman, K. Laasonen, M. Sprik, M. Parrinello, *Journal of Physics: Condensed Matter* **6**, A93 (1994).

[133] 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, X. Wu, *Proceedings of the National Academy of Sciences* **114**, 10846 (2017).

[134] S. Yoo, X. C. Zeng, S. S. Xantheas, *The Journal of Chemical Physics* **130**, 221102 (2009).

[135] M. J. McGrath, J. I. Siepmann, I.-F. W. Kuo, C. J. Mundy, J. VandeVondele, J. Hutter, F. Mohamed, M. Krack, *The Journal of Physical Chemistry A* **110**, 640 (2006).

[136] A. A. Hassanali, J. Cuny, V. Verdolino, M. Parrinello, *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **372**, 20120482 (2014).

[137] M. J. Gillan, D. Alfè, A. Michaelides, *The Journal of Chemical Physics* **144**, 130901 (2016).

[138] I.-C. Lin, A. P. Seitsonen, I. Tavernelli, U. Rothlisberger, *Journal of Chemical Theory and Computation* **8**, 3902 (2012).

[139] S. Yoo, S. S. Xantheas, *The Journal of Chemical Physics* **134** (2011).

[140] P. Montero de Hijes, C. Dellago, R. Jinnouchi, G. Kresse, *The Journal of Chemical Physics* **161**, 131102 (2024).

[141] J. Villard, M. P. Bircher, U. Rothlisberger, *Chemical Science* **15**, 4434 (2024).

[142] J. P. Perdew, K. Burke, M. Ernzerhof, *Physical Review Letters* **77**, 3865 (1996).

[143] J. C. Grossman, E. Schwegler, E. W. Draeger, F. Gygi, G. Galli, *The Journal of Chemical Physics* **120**, 300 (2004).

[144] J. Sun, R. C. Remsing, Y. Zhang, Z. Sun, A. Ruzsinszky, H. Peng, Z. Yang, A. Paul, U. Waghmare, X. Wu, M. L. Klein, J. P. Perdew, Scan: An efficient density functional yielding accurate structures and energies of diversely-bonded materials (2015). DOI: 10.48550/arXiv.1511.01089.

[145] S. Grimme, J. Antony, S. Ehrlich, H. Krieg, *The Journal of Chemical Physics* **132**, 154104 (2010).

[146] S. Grimme, S. Ehrlich, L. Goerigk, *Journal of Computational Chemistry* **32**, 1456 (2011).

[147] L. Zheng, M. Chen, Z. Sun, H.-Y. Ko, B. Santra, P. Dhuvad, X. Wu, *The Journal of Chemical Physics* **148**, 164505 (2018).

[148] A. P. Gaiduk, F. Gygi, G. Galli, *The Journal of Physical Chemistry Letters* **6**, 2902 (2015).

[149] J. Schmidt, J. VandeVondele, I.-F. W. Kuo, D. Sebastiani, J. I. Siepmann, J. Hutter, C. J. Mundy, *The Journal of Physical Chemistry B* **113**, 11959 (2009).

[150] J. Wang, G. Román-Pérez, J. M. Soler, E. Artacho, M.-V. Fernández-Serra, *The Journal of Chemical Physics* **134**, 024516 (2011).

[151] G. Miceli, S. de Gironcoli, A. Pasquarello, *The Journal of Chemical Physics* **142**, 034501 (2015).

[152] M. Ceriotti, W. Fang, P. G. Kusalik, R. H. McKenzie, A. Michaelides, M. A. Morales, T. E. Markland, *Chemical Reviews* **116**, 7529 (2016).

[153] R. H. McKenzie, C. Bekker, B. Athokpam, S. G. Ramesh, *The Journal of Chemical Physics* **140**, 174508 (2014).

[154] D. Marx, M. Parrinello, *The Journal of Chemical Physics* **104**, 4077 (1996).

[155] J. A. Morrone, R. Car, *Physical Review Letters* **101**, 017801 (2008).

[156] A. Eltareb, G. E. Lopez, N. Giovambattista, *The Journal of Chemical Physics* **156**, 204502 (2022).

[157] K.-H. Cho, K. T. No, H. A. Scheraga, *Journal of Molecular Structure* **641**, 77 (2002).

[158] A. P. Bartók, M. C. Payne, R. Kondor, G. Csányi, *Physical Review Letters* **104**, 136403 (2010).

[159] A. P. Bartók, M. J. Gillan, F. R. Manby, G. Csányi, *Physical Review B* **88**, 054104 (2013).

[160] A. Omranpour, P. Montero De Hijes, J. Behler, C. Dellago, *The Journal of Chemical Physics* **160**, 170901 (2024).

[161] C. Andreani, G. Romanelli, A. Parmentier, R. Senesi, A. I. Kolesnikov, H.-Y. Ko, M. F. Calegari Andrade, R. Car, *The Journal of Physical Chemistry Letters* **11**, 9461 (2020).

[162] J. Xu, C. Zhang, L. Zhang, M. Chen, B. Santra, X. Wu, *Physical Review B* **102**, 214113 (2020).

[163] P. Montero de Hijes, S. Romano, A. Gorfer, C. Dellago, *The Journal of Chemical Physics* **158**, 204706 (2023).

[164] H. Wang, L. Zhang, J. Han, W. E, *Computer Physics Communications* **228**, 178 (2018).

[165] P. M. Piaggi, A. Z. Panagiotopoulos, P. G. Debenedetti, R. Car, *Journal of Chemical Theory and Computation* **17**, 3065 (2021).

[166] L. Zhang, H. Wang, R. Car, W. E, *Physical Review Letters* **126**, 236001 (2021).

[167] S. Batzner, A. Musaelian, L. Sun, M. Geiger, J. P. Mailoa, M. Kornbluth, N. Molinari, T. E. Smidt, B. Kozinsky, *Nature Communications* **13**, 2453 (2022).

[168] A. Musaelian, S. Batzner, A. Johansson, L. Sun, C. J. Owen, M. Kornbluth, B. Kozinsky, *Nature Communications* **14**, 579 (2023).

[169] I. Batatia, *et al.*, A foundation model for atomistic materials chemistry (2024). DOI: 10.48550/arXiv.2401.00096.

[170] T. Maxson, T. Szilvási, *The Journal of Physical Chemistry Letters* **15**, 3740 (2024).

[171] P. J. Heaney, C. T. Prewitt, G. V. Gibbs, *Silica: Physical Behavior, Geochemistry, and Materials Applications* (2018).

[172] D. S. Wragg, R. E. Morris, A. W. Burton, *Chemistry of Materials* **20**, 1561 (2008).

[173] R. E. Kirk, *Kirk–Othmer Encyclopedia of Chemical Technology* (Wiley-Blackwell, New York, 1998).

[174] P. Vashishta, R. K. Kalia, J. P. Rino, I. Ebbsjö, *Physical Review B* **41**, 12197 (1990).

[175] B. W. H. van Beest, G. J. Kramer, R. A. van Santen, *Physical Review Letters* **64**, 1955 (1990).

[176] J. Du, L. R. Corrales, *Physical Review B* **72**, 092201 (2005).

[177] A. Tadros, M. A. Klenin, G. Lucovsky, *Journal of Non-Crystalline Solids* **64**, 215 (1984).

[178] O. D. Friedrichs, A. W. M. Dress, D. H. Huson, J. Klinowski, A. L. Mackay, *Nature* **400**, 644 (1999).

[179] M. M. Treacy, K. H. Randall, S. Rao, J. A. Perry, D. J. Chadi, *Zeitschrift fur Kristallographie* **212**, 768 (1997).

[180] M. M. J. Treacy, I. Rivin, E. Balkovsky, K. H. Randall, M. D. Foster, *Microporous and Mesoporous Materials* **74**, 121 (2004).

[181] B. Chen, Z. Yang, Y. Zhu, Y. Xia, *Journal of Materials Chemistry A* **2**, 16811 (2014).

[182] A. Phan, C. J. Doonan, F. J. Uribe-Romo, C. B. Knobler, M. O'Keeffe, O. M. Yaghi, *Accounts of Chemical Research* **43**, 58 (2010).

[183] S. Wang, L. Luo, A. Wu, D. Wang, L. Wang, Y. Jiao, C. Tian, *Coordination Chemistry Reviews* **498**, 215464 (2024).

[184] Z. Zheng, Z. Rong, H. L. Nguyen, O. M. Yaghi, *Inorganic Chemistry* **62**, 20861 (2023).

[185] S. Lee, D. Nam, D. C. Yang, W. Choe, *Small* **19**, 2300036 (2023).

[186] S. Lee, H. Jeong, S. Jung, Y. Kim, E. Cho, J. Nam, D. ChangMo Yang, D. Y. Shin, J.-H. Lee, H. Oh, W. Choe, *JACS Au* **5**, 1460 (2025).

[187] R. Banerjee, A. Phan, B. Wang, C. Knobler, H. Furukawa, M. O'Keeffe, O. M. Yaghi, *Science* **319**, 939 (2008).

[188] V. A. Blatov, G. D. Ilyushin, D. M. Proserpio, *Chemistry of Materials* **25**, 412 (2013).

[189] T. C. Nicholas, D. F. T. du Toit, L. A. M. Rosset, D. M. Proserpio, A. L. Goodwin, V. L. Deringer, The structure and topology of an amorphous metal-organic framework (2025). DOI: 10.48550/arXiv.2503.24367.

[190] P. Hawken, *Carbon: The Book of Life* (Viking Press USA, New York, 2025).

[191] W. Zhang, S. Zhu, R. Luque, S. Han, L. Hu, G. Xu, *Chemical Society Reviews* **45**, 715 (2016).

[192] W. Yuan, Y. Zhang, L. Cheng, H. Wu, L. Zheng, D. Zhao, *Journal of Materials Chemistry A* **4**, 8932 (2016).

[193] F. Bonaccorso, Z. Sun, T. Hasan, A. C. Ferrari, *Nature Photonics* **4**, 611 (2010).

[194] K. Yoshikawa, H. Kawasaki, W. Yoshida, T. Irie, K. Konishi, K. Nakano, T. Uto, D. Adachi, M. Kanematsu, H. Uzu, K. Yamamoto, *Nature Energy* **2**, 1 (2017).

[195] M. Taguchi, K. Kawamoto, S. Tsuge, T. Baba, H. Sakata, M. Morizane, K. Uchihashi, N. Nakamura, S. Kiyama, O. Oota, *Progress in Photovoltaics: Research and Applications* **8**, 503 (2000).

[196] S. J. Clark, J. Crain, G. J. Ackland, *Physical Review B* **55**, 14059 (1997).

[197] F. Wooten, G. A. Fuller, K. Winer, D. Weaire, *Journal of Non-Crystalline Solids* **75**, 45 (1985).

[198] M. M. J. Treacy, K. B. Borisenko, *Science* **335**, 950 (2012).

[199] L. J. Lewis, *Journal of Non-Crystalline Solids* **580**, 121383 (2022).

[200] L. Zhang, X. Wei, Y. Lin, F. Wang, *Carbon* **94**, 202 (2015).

[201] V. Drchal, J. Málek, *Journal of Non-Crystalline Solids* **97–98**, 199 (1987).

[202] J. Ben, A. L. Martinotto, G. L. Rech, J. E. Zorzi, C. A. Perottoni, *Journal of Non-Crystalline Solids* **576**, 121260 (2022).

[203] B. Schultrich, H. J. Scheibe, D. Drescher, H. Ziegele, *Surface and Coatings Technology* **98**, 1097 (1998).

[204] M. A. Caro, G. Csányi, T. Laurila, V. L. Deringer, *Physical Review B* **102**, 174201 (2020).

[205] C. de Tomas, A. Aghajamali, J. L. Jones, D. J. Lim, M. J. López, I. Suarez-Martinez, N. A. Marks, *Carbon* **155**, 624 (2019).

[206] P. Rowe, V. L. Deringer, P. Gasparotto, G. Csányi, A. Michaelides, *The Journal of Chemical Physics* **153**, 034702 (2020).

[207] A. Barnard, S. Russo, G. Leach, *Molecular Simulation* **28**, 761 (2002).

[208] A. S. Barnard, S. P. Russo, *Molecular Physics* **100**, 1517 (2002).

[209] J. L. A. Gardner, K. T. Baker, V. L. Deringer, *Machine Learning: Science and Technology* **5**, 015003 (2024).

[210] J. D. Morrow, J. L. A. Gardner, V. L. Deringer, *The Journal of Chemical Physics* **158**, 121501 (2023).

[211] M. Kulichenko, B. Nebgen, N. Lubbers, J. S. Smith, K. Barros, A. E. A. Allen, A. Habib, E. Shinkle, N. Fedik, Y. W. Li, R. A. Messerly, S. Tretiak, *Chemical Reviews* **124**, 13681 (2024).

[212] E. V. Podryabinkin, A. V. Shapeev, *Computational Materials Science* **140**, 171 (2017).

[213] L. Zhang, D.-Y. Lin, H. Wang, R. Car, W. E, *Physical Review Materials* **3**, 023804 (2019).

[214] E. V. Podryabinkin, E. V. Tikhonov, A. V. Shapeev, A. R. Oganov, *Physical Review B* **99**, 064114 (2019).

[215] J. Vandermause, Y. Xie, J. S. Lim, C. J. Owen, B. Kozinsky, *Nature Communications* **13**, 5183 (2022).

[216] N. Bernstein, G. Csányi, V. L. Deringer, *npj Computational Materials* **5**, 99 (2019).

[217] C. J. Pickard, R. J. Needs, *Journal of Physics: Condensed Matter* **23**, 053201 (2011).

[218] C. J. Pickard, R. J. Needs, *Physical Review Letters* **97**, 045504 (2006).

[219] D. Yoo, J. Jung, W. Jeong, S. Han, *npj Computational Materials* **7**, 131 (2021).

[220] V. L. Deringer, M. A. Caro, G. Csányi, *Nature Communications* **11**, 5461 (2020).

[221] A. Seko, A. Togo, I. Tanaka, *Descriptors for Machine Learning of Materials Data* (Springer, Singapore, 2018).

[222] F. Musil, A. Grisafi, A. P. Bartók, C. Ortner, G. Csányi, M. Ceriotti, *Chemical Reviews* **121**, 9759 (2021).

[223] H. Huo, M. Rupp, *Machine Learning: Science and Technology* **3**, 045017 (2022).

[224] D. Weininger, A. Weininger, J. L. Weininger, *Journal of Chemical Information and Computer Sciences* **29**, 97 (1989).

[225] F. Bazzi-Allahri, F. Shiri, S. Ahmadi, A. P. Toropova, A. A. Toropov, *BMC Chemistry* **18**, 191 (2024).

[226] A. Askarzade, S. Ahmadi, A. Almasirad, *Scientific Reports* **15**, 6573 (2025).

[227] M. Rupp, A. Tkatchenko, K.-R. Müller, O. A. von Lilienfeld, *Physical Review Letters* **108**, 058301 (2012).

[228] M. Yu. Lobanov, N. S. Bogatyreva, O. V. Galzitskaya, *Molecular Biology* **42**, 623 (2008).

[229] K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O. A. von Lilienfeld, K.-R. Müller, A. Tkatchenko, *The Journal of Physical Chemistry Letters* **6**, 2326 (2015).

[230] A. P. Bartók, R. Kondor, G. Csányi, *Physical Review B* **87**, 184115 (2013).

[231] J. Behler, *The Journal of Chemical Physics* **134**, 074106 (2011).

[232] N. Tuchinda, C. A. Schuh, *npj Computational Materials* **10**, 72 (2024).

[233] F. A. Faber, L. Hutchison, B. Huang, J. Gilmer, S. S. Schoenholz, G. E. Dahl, O. Vinyals, S. Kearnes, P. F. Riley, O. A. von Lilienfeld, *Journal of Chemical Theory and Computation* **13**, 5255 (2017).

[234] P. J. Steinhardt, D. R. Nelson, M. Ronchetti, *Physical Review B* **28**, 784 (1983).

[235] W. S. Xu, Z. Y. Sun, L. J. An, *The European Physical Journal E* **31**, 377 (2010).

[236] W. Lechner, C. Dellago, *The Journal of Chemical Physics* **129**, 114707 (2008).

[237] J. S. Van Duijneveldt, D. Frenkel, *The Journal of Chemical Physics* **96**, 4655 (1992).

[238] P. Rein Ten Wolde, M. J. Ruiz-Montero, D. Frenkel, *The Journal of Chemical Physics* **104**, 9932 (1996).

[239] A. Ikeda, K. Miyazaki, *Physical Review Letters* **106**, 015701 (2011).

[240] H. Tanaka, T. Kawasaki, H. Shintani, K. Watanabe, *Nature Materials* **9**, 324 (2010).

[241] P. R. ten Wolde, M. J. Ruiz-Montero, D. Frenkel, *Physical Review Letters* **75**, 2714 (1995).

[242] T. Kawasaki, H. Tanaka, *Journal of Physics: Condensed Matter* **22**, 232102 (2010).

[243] L. D. Landau, E. M. Lifshitz, *Quantum Mechanics: Non-Relativistic Theory*, vol. 3 (Pergamon, 2013), third edn.

[244] Z. Faure Beaulieu, V. L. Deringer, F. Martelli, *The Journal of Chemical Physics* **160**, 081101 (2024).

[245] S. De, A. P. Bartók, G. Csányi, M. Ceriotti, *Physical Chemistry Chemical Physics* **18**, 13754 (2016).

[246] J. P. Darby, J. R. Kermode, G. Csányi, *npj Computational Materials* **8**, 166 (2022).

[247] V. L. Deringer, A. P. Bartók, N. Bernstein, D. M. Wilkins, M. Ceriotti, G. Csányi, *Chemical Reviews* **121**, 10073 (2021).

[248] R. Drautz, *Physical Review B* **99**, 014104 (2019).

[249] D. F. Thomas du Toit, V. L. Deringer, *The Journal of Chemical Physics* **159**, 024803 (2023).

[250] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19 (New York, NY, USA, 2019), pp. 2623–2631.

[251] C. E. Rasmussen, C. K. I. Williams, *Gaussian Processes for Machine Learning* (MIT Press, Cambridge (MA), 2005).

[252] V. L. Deringer, N. Bernstein, A. P. Bartók, M. J. Cliffe, R. N. Kerber, L. E. Marbella, C. P. Grey, S. R. Elliott, G. Csányi, *The Journal of Physical Chemistry Letters* **9**, 2879 (2018).

[253] M. A. Caro, A. Aarva, V. L. Deringer, G. Csányi, T. Laurila, *Chemistry of Materials* **30**, 7446 (2018).

[254] F. M. Paruzzo, A. Hofstetter, F. Musil, S. De, M. Ceriotti, L. Emsley, *Nature Communications* **9**, 4501 (2018).

[255] D. Golze, M. Hirvensalo, P. Hernández-León, A. Aarva, J. Etula, T. Susi, P. Rinke, T. Laurila, M. A. Caro, *Chemistry of Materials* **34**, 6240 (2022).

[256] I. Batatia, D. P. Kovács, G. N. C. Simm, C. Ortner, G. Csányi, Mace: Higher order equivariant message passing neural networks for fast and accurate force fields (2023). DOI: 10.48550/arXiv.2206.07697.

[257] G. Simeon, G. D. Fabritiis, *Thirty-Seventh Conference on Neural Information Processing Systems* (2023).

[258] Y. Bengio, A. Courville, P. Vincent, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **35**, 1798 (2013).

[259] I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning* (MIT Press, 2016). `http://www.deeplearningbook.org`.

[260] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization (2017). DOI: 10.48550/arXiv.1412.6980.

[261] K. Hornik, M. Stinchcombe, H. White, *Neural Networks* **2**, 359 (1989).

[262] J. E. Dobson, *International Journal of Digital Humanities* **5**, 431 (2023).

[263] G. Corso, H. Stark, S. Jegelka, T. Jaakkola, R. Barzilay, *Nature Reviews Methods Primers* **4**, 17 (2024).

[264] M. Gori, G. Monfardini, F. Scarselli, *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.* (31 July-4 Aug. 2005), vol. 2, pp. 729–734 vol. 2.

[265] D. K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik, R. P. Adams, *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, R. Garnett, eds. (Curran Associates, Inc., 2015), pp. 2224–2232.

[266] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, G. E. Dahl, *Proceedings of the 34th International Conference on Machine Learning*, D. Precup, Y. W. Teh, eds. (PMLR, 2017), vol. 70 of *Proceedings of Machine Learning Research*, pp. 1263–1272.

[267] K. T. Schütt, O. T. Unke, M. Gastegger, Equivariant message passing for the prediction of tensorial properties and molecular spectra (2021). DOI: 10.48550/arXiv.2102.03150.

[268] J. Gasteiger, J. Groß, S. Günnemann, Directional message passing for molecular graphs (2022). DOI: 10.48550/arXiv.2003.03123.

[269] J. Schmidt, L. Pettersson, C. Verdozzi, S. Botti, M. A. L. Marques, *Science Advances* **7**, eabi7948 (2021).

[270] R. Wang, Y. Zou, C. Zhang, X. Wang, M. Yang, D. Xu, *Microporous and Mesoporous Materials* **331**, 111666 (2022).

[271] X.-G. Li, B. Blaiszik, M. E. Schwarting, R. Jacobs, A. Scourtas, K. J. Schmidt, P. M. Voyles, D. Morgan, *The Journal of Chemical Physics* **155**, 154702 (2021).

[272] J. Xiong, Z. Xiong, K. Chen, H. Jiang, M. Zheng, *Drug Discovery Today* **26**, 1382 (2021).

[273] C. W. Park, M. Kornbluth, J. Vandermause, C. Wolverton, B. Kozinsky, J. P. Mailoa, *npj Computational Materials* **7**, 73 (2021).

[274] J. L. A. Gardner, D. F. T. du Toit, C. B. Mahmoud, Z. F. Beaulieu, V. Juraskova, L.-B. Paşca, L. A. M. Rosset, F. Duarte, F. Martelli, C. J. Pickard, V. L. Deringer, Distillation of atomistic foundation models across architectures and chemical domains (2025). DOI: 10.48550/arXiv.2506.10956.

[275] D. Bagayoko, *AIP Advances* **4**, 127104 (2014).

[276] A. D. Becke, *Phys. Rev. A* **38**, 3098 (1988).

[277] C. Lee, W. Yang, R. G. Parr, *Phys. Rev. B* **37**, 785 (1988).

[278] M. Fischer, F. O. Evers, F. Formalik, *et al.*, *Theoretical Chemistry Accounts* **135**, 257 (2016).

[279] J. P. Perdew, A. Zunger, *Phys. Rev. B* **23**, 5048 (1981).

[280] A. J. Cohen, P. Mori-Sánchez, W. Yang, *Chemical Reviews* **112**, 289–320 (2012).

[281] S. Nosé, *The Journal of Chemical Physics* **81**, 511 (1984).

[282] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, J. R. Haak, *The Journal of Chemical Physics* **81**, 3684 (1984).

[283] R. L. Davidchack, R. Handel, M. V. Tretyakov, *The Journal of Chemical Physics* **130** (2009).

[284] M. Parrinello, A. Rahman, *Journal of Applied Physics* **52**, 7182 (1981).

[285] G. J. Martyna, D. J. Tobias, M. L. Klein, *The Journal of Chemical Physics* **101**, 4177 (1994).

[286] W. Shinoda, M. Shiga, M. Mikami, *Physical Review B* **69**, 134103 (2004).

[287] L. Boltzmann, *Vorlesungen über Gastheorie* (Leipzig, J. A. Barth, Leipzig, 1896).

[288] S. Plimpton, *Journal of Computational Physics* **117**, 1 (1995).

[289] 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, R. Shan, M. J. Stevens, J. Tranchida, C. Trott, S. J. Plimpton, *Computer Physics Communications* **271**, 108171 (2022).

[290] C. G. Salzmann, *The Journal of Chemical Physics* **150**, 060901 (2019).

[291] P. W. Bridgman, *Proceedings of the American Academy of Arts and Sciences* **47**, 441 (1912).

[292] M. Millot, F. Coppari, J. R. Rygg, A. Correa Barrios, S. Hamel, D. C. Swift, J. H. Eggert, *Nature* **569**, 251 (2019).

[293] R. Yamane, K. Komatsu, J. Gouchi, Y. Uwatoko, S. Machida, T. Hattori, H. Ito, H. Kagi, *Nature Communications* **12**, 1129 (2021).

[294] V. B. Prakapenka, N. Holtgrewe, S. S. Lobanov, A. F. Goncharov, *Nature Physics* **17**, 1233 (2021).

[295] M. Rescigno, A. Toffano, U. Ranieri, L. Andriambariarijaona, R. Gaal, S. Klotz, M. M. Koza, J. Ollivier, F. Martelli, J. Russo, F. Sciortino, J. Teixeira, L. E. Bove, *Nature* **640**, 662 (2025).

[296] H. Tanaka, *The Journal of Chemical Physics.* **153**, 130901 (2020).

[297] K. Amann-Winkel, R. Böhmer, F. Fujara, C. Gainaru, B. Geil, T. Loerting, *Reviews of Modern Physics* **88**, 011002 (2016).

[298] T. Loerting, K. Winkel, M. Seidl, M. Bauer, C. Mitterdorfer, P. H. Handle, C. G. Salzmann, E. Mayer, J. L. Finney, D. T. Bowron, *Physical Chemistry Chemical Physics* **13**, 8783 (2011).

[299] S. Kwok, *Physics and Chemistry of the Interstellar Medium* (University Science Books, Sausalito, Calif, 2007).

[300] K. Winkel, D. T. Bowron, T. Loerting, E. Mayer, J. L. Finney, *The Journal of Chemical Physics* **130**, 204502 (2009).

[301] J. J. Shephard, S. Klotz, M. Vickers, C. G. Salzmann, *The Journal of Chemical Physics* **144**, 204502 (2016).

[302] O. Mishima, L. D. Calvert, E. Whalley, *Nature* **310**, 393 (1984).

[303] R. J. Nelmes, J. S. Loveday, T. Strässle, C. L. Bull, M. Guthrie, G. Hamel, S. Klotz, *Nature Physics* **2**, 414 (2006).

[304] T. Loerting, C. Salzmann, I. Kohl, E. Mayer, A. Hallbrucker, *Physical Chemistry Chemical Physics* **3**, 5355 (2001).

[305] F. Martelli, N. Giovambattista, S. Torquato, R. Car, *Physical Review Materials* **2**, 075601 (2018).

[306] J. J. Shephard, S. Ling, G. C. Sosso, A. Michaelides, B. Slater, C. G. Salzmann, *The Journal of Physical Chemistry Letters* **8**, 1645 (2017).

[307] H. Kobayashi, K. Komatsu, H. Ito, S. Machida, T. Hattori, H. Kagi, *The Journal of Physical Chemistry Letters* **14**, 10664 (2023).

[308] M. Formanek, S. Torquato, R. Car, F. Martelli, *The Journal of Physical Chemistry B* **127**, 3946 (2023).

[309] P. H. Poole, F. Sciortino, U. Essmann, H. E. Stanley, *Nature* **360**, 324 (1992).

[310] J. C. Palmer, F. Martelli, Y. Liu, R. Car, A. Z. Panagiotopoulos, P. G. Debenedetti, *Nature* **510**, 385 (2014).

[311] J. A. Sellberg, *et al.*, *Nature* **510**, 381 (2014).

[312] O. Mishima, L. D. Calvert, E. Whalley, *Nature* **314**, 76 (1985).

[313] N. Giovambattista, H. Eugene Stanley, F. Sciortino, *Physical Review E* **72**, 031510 (2005).

[314] J. Engstler, N. Giovambattista, *The Journal of Chemical Physics* **147**, 074505 (2017).

[315] P. H. Poole, U. Essmann, F. Sciortino, H. E. Stanley, *Physical Review E* **48**, 4605 (1993).

[316] T. E. Gartner, S. Torquato, R. Car, P. G. Debenedetti, *Nature Communications* **12**, 3398 (2021).

[317] A. Rosu-Finsen, M. B. Davies, A. Amon, H. Wu, A. Sella, A. Michaelides, C. G. Salzmann, *Science* **379**, 474 (2023).

[318] A. Eltareb, G. E. Lopez, N. Giovambattista, *Communications Chemistry* **7**, 36 (2024).

[319] F. Martelli, F. Leoni, F. Sciortino, J. Russo, *The Journal of Chemical Physics* **153**, 104503 (2020).

[320] A. F. Agarap, Deep learning using rectified linear units (ReLU) (2019). DOI: 10.48550/arXiv.1803.08375.

[321] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, A. Lerer, *NIPS 2017 Workshop on Autodiff* (2017).

[322] Z. Faure Beaulieu, T. C. Nicholas, J. L. A. Gardner, A. L. Goodwin, V. L. Deringer, *Chemical Communications* **59**, 11405 (2023).

[323] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, *Journal of Machine Learning Research* **12**, 2825 (2011).

[324] L. Breiman, *Machine Learning* **45**, 5 (2001).

[325] N. Giovambattista, F. Sciortino, F. W. Starr, P. H. Poole, *The Journal of Chemical Physics* **145**, 224501 (2016).

[326] I. de Almeida Ribeiro, D. Dhabal, R. Kumar, S. Banik, S. K. R. S. Sankaranarayanan, V. Molinero, *Proceedings of the National Academy of Sciences* **121**, e2414444121 (2024).

[327] S. K. Deb, M. Wilding, M. Somayazulu, P. F. McMillan, *Nature* **414**, 528 (2001).

[328] V. L. Deringer, N. Bernstein, G. Csányi, C. Ben Mahmoud, M. Ceriotti, M. Wilson, D. A. Drabold, S. R. Elliott, *Nature* **589**, 59 (2021).

[329] J. R. Errington, P. G. Debenedetti, *Nature* **409**, 318 (2001).

[330] S. Caravati, M. Bernasconi, T. D. Kühne, M. Krack, M. Parrinello, *Applied Physics Letters* **91**, 171906 (2007).

[331] R. Shi, H. Tanaka, *Journal of the American Chemical Society* **142**, 2868 (2020).

[332] R. Shi, H. Tanaka, *Science advances* **5**, eaav3194 (2019).

[333] A. C. Wright, *Journal of non-crystalline solids* **179**, 84 (1994).

[334] P. Gaskell, D. Wallis, *Physical review letters* **76**, 66 (1996).

[335] Q. Mei, C. Benmore, S. Sen, R. Sharma, J. Yarger, *Physical Review B-Condensed Matter and Materials Physics* **78**, 144204 (2008).

[336] J. C. Phillips, *Journal of Non-Crystalline Solids* **43**, 37 (1981).

[337] S. Elliott, *Physical review letters* **67**, 711 (1991).

[338] S. R. Elliott, *Nature* **354**, 445 (1991).

[339] C. Massobrio, A. Pasquarello, *The Journal of Chemical Physics* **114**, 7976 (2001).

[340] H. Tanaka, *Physical Review Letters* **80**, 5750 (1998).

[341] H. Tanaka, *The Journal of Chemical Physics* **112**, 799 (2000).

[342] H. Tanaka, *The European Physical Journal E* **35**, 1 (2012).

[343] L. B. Skinner, C. J. Benmore, J. C. Neuefeind, J. B. Parise, *The Journal of Chemical Physics* **141**, 214507 (2014).

[344] N. Esmaeildoost, H. Pathak, A. Späh, T. J. Lane, K. H. Kim, C. Yang, K. Amann-Winkel, M. Ladd-Parada, F. Perakis, J. Koliyadu, A. R. Oggenfuss, P. J. M. Johnson, Y. Deng, S. Zerdane, R. Mankowsky, P. Beaud, H. T. Lemke, A. Nilsson, J. A. Sellberg, *The Journal of Chemical Physics* **155**, 214501 (2021).

[345] Q. Zhou, Y. Shi, B. Deng, J. Neuefeind, M. Bauchy, *Science Advances* **7**, eabh1761 (2021).

[346] E. Ibrahim, Y. Lysogorskiy, R. Drautz, *Journal of Chemical Theory and Computation* **20**, 11049 (2024).

[347] G. Dusson, M. Bachmayr, G. Csányi, R. Drautz, S. Etter, C. van der Oord, C. Ortner, *Journal of Computational Physics* **454**, 110946 (2022).

[348] Y. Lysogorskiy, A. Bochkarev, M. Mrovec, R. Drautz, *Physical Review Materials* **7**, 043801 (2023).

[349] M. Matsumoto, T. Yagasaki, H. Tanaka, *Journal of Computational Chemistry* **39**, 61 (2018).

[350] M. Matsumoto, T. Yagasaki, H. Tanaka, *Journal of Chemical Information and Modeling* **61**, 2542 (2021).

[351] L. Ruiz Pestana, N. Mardirossian, M. Head-Gordon, T. Head-Gordon, *Chemical Science* **8**, 3554 (2017).

[352] S. V. King, *Nature* **213**, 1112 (1967).

[353] A. Tadros, M. Klenin, G. Lucovsky, *Proceedings of the international conference on the theory of the structures of non-crystalline solids* **75**, 407 (1985).

[354] D. S. Franzblau, *Physical Review B* **44**, 4925 (1991).

[355] W. D. Luedtke, U. Landman, *Physical Review B* **40**, 1164 (1989).

[356] Z. Zhang, W. Kob, *Physical Review B* **110**, 104203 (2024).

[357] S. Le Roux, P. Jund, *Computational Materials Science* **49**, 70 (2010).

[358] M. Shiga, A. Hirata, Y. Onodera, H. Masai, *Communications Materials* **4**, 91 (2023).

[359] S. Kohara, K. Kato, S. Kimura, H. Tanaka, T. Usuki, K. Suzuya, H. Tanaka, Y. Moritomo, T. Matsunaga, N. Yamada, Y. Tanaka, H. Suematsu, M. Takata, *Applied Physics Letters* **89**, 201910 (2006).

[360] F. Wooten, D. Weaire, *Modeling Tetrahedrally Bonded Random Networks by Computer*, vol. 40 (Academic Press, 1987).

[361] T. S. Hudson, P. Harrowell, *The Journal of Chemical Physics* **126**, 184502 (2007).

[362] G. Opletal, T. C. Petersen, I. K. Snook, D. G. McCulloch, *The Journal of Chemical Physics* **126**, 214705 (2007).

[363] L. Guttman, *Journal of Non-Crystalline Solids* **116**, 145 (1990).

[364] A. Rahman, F. H. Stillinger, *Journal of the American Chemical Society* **95**, 7943 (1973).

[365] K. Goetzke, H.-J. Klein, *Journal of Non-Crystalline Solids* **127**, 215 (1991).

[366] X. Yuan, A. Cormack, *Computational Materials Science* **24**, 343 (2002).

[367] C. S. Marians, L. W. Hobbs, *Journal of Non-Crystalline Solids* **124**, 242 (1990).

[368] C. S. Marians, L. W. Hobbs, *Journal of Non-Crystalline Solids* **106**, 317 (1988).

[369] G. Camisasca, D. Schlesinger, I. Zhovtobriukh, G. Pitsevich, L. G. M. Pettersson, *The Journal of Chemical Physics* **151**, 034508 (2019).

[370] F. Martelli, *The Journal of Chemical Physics* **150**, 094506 (2019).

[371] B. Santra, D. J. , Robert A., M. , Fausto, R. and Car, *Molecular Physics* **113**, 2829 (2015).

[372] M. Formanek, F. Martelli, *AIP Advances* **10**, 055205 (2020).

[373] D. Prada-Gracia, R. Shevchuk, F. Rao, *The Journal of Chemical Physics* **139**, 084501 (2013).

[374] A. Luzar, D. Chandler, *Nature* **379**, 55 (1996).

[375] Z. Yan, S. V. Buldyrev, P. Kumar, N. Giovambattista, P. G. Debenedetti, H. E. Stanley, *Physical Review E* **76**, 051201 (2007).

[376] E. Duboué-Dijon, D. Laage, *The Journal of Physical Chemistry B* **119**, 8406 (2015).

[377] M. Paolantoni, N. F. Lago, M. Albertí, A. Laganà, *The Journal of Physical Chemistry A* **113**, 15100 (2009).

[378] J. R. Errington, P. G. Debenedetti, *Nature* **409**, 318 (2001).

[379] P.-L. Chau, A. J. Hardwick, *Molecular Physics* **93**, 511 (1998).

[380] E. J. Maginn, R. A. Messerly, D. J. Carlson, D. R. Roe, J. R. Elliot, *Living Journal of Computational Molecular Science* **1**, 6324 (2018).

[381] D. P. Kovacs, I. Batatia, E. S. Arany, G. Csanyi, *The Journal of Chemical Physics* **159**, 044118 (2023).

[382] L. N. Smith, A disciplined approach to neural network hyper-parameters: Part 1 – learning rate, batch size, momentum, and weight decay (2018). DOI: 10.48550/arXiv.1803.09820.

[383] N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, P. T. P. Tang, On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima (2017). DOI: 10.48550/arXiv.1609.04836.

[384] X. Fu, B. M. Wood, L. Barroso-Luque, D. S. Levine, M. Gao, M. Dzamba, C. L. Zitnick, Learning Smooth and Expressive Interatomic Potentials for Physical Property Prediction (2025). DOI: 10.48550/arXiv.2502.12147.

[385] J. Gardner, graph-pes: train and use graph-based ML models of potential energy surfaces (2024).

[386] A. Soper, *Chemical Physics* **258**, 121 (2000).

[387] T. E. Markland, M. Ceriotti, *Nature Reviews Chemistry* **2**, 0109 (2018).

[388] B. Cheng, E. A. Engel, J. Behler, C. Dellago, M. Ceriotti, *Proceedings of the National Academy of Sciences* **116**, 1110 (2019).

[389] F. Martelli, *Journal of Molecular Liquids* **329**, 115530 (2021).

[390] M. J. Zimoń, F. Martelli, Predicting Thermodynamics of Liquid Water from Time Series Analysis (2025). DOI: 10.48550/arXiv.2506.21821.

[391] T. Head-Gordon, G. Hura, *Chemical Reviews* **102**, 2651 (2002).

[392] F. H. Stillinger, *Science* **209**, 451 (1980).

[393] F. de los Santos, G. Franzese, *Physical Review E* **85**, 010602 (2012).

[394] R. A. DiStasio, Jr., B. Santra, Z. Li, X. Wu, R. Car, *The Journal of Chemical Physics* **141**, 084502 (2014).

[395] A. C. Belch, S. A. Rice, *The Journal of Chemical Physics* **86**, 5676 (1987).

[396] Y. Gao, H. Fang, K. Ni, *Scientific Reports* **11**, 9542 (2021).

[397] I. Bakó, Á. Bencsura, K. Hermannson, S. Bálint, T. Grósz, V. Chihaia, J. Oláh, *Physical Chemistry Chemical Physics* **15**, 15163 (2013).

[398] F. Martelli, J. Crain, G. Franzese, *ACS Nano* **14**, 8616 (2020).

[399] S. Dasgupta, E. Lambros, J. P. Perdew, F. Paesani, *Nature Communications* **12**, 6359 (2021).

[400] M. Holz, S. R. Heil, A. Sacco, *Physical Chemistry Chemical Physics* **2**, 4740 (2000).

[401] A. J. Easteal, W. E. Price, L. A. Woolf, *Journal of the Chemical Society, Faraday Transactions 1: Physical Chemistry in Condensed Phases* **85**, 1091 (1989).

[402] R. Mills, *The Journal of Physical Chemistry* **77**, 685 (1973).

[403] F. L. Johansen, A. S. Anker, U. Friis-Jensen, E. B. Dam, K. M. O. Jensen, R. Selvan, *Journal of Open Source Software* (2024).

[404] C. Benmore, L. C. Gallington, E. Soignard, *Molecular Physics* **117**, 2470 (2019).

[405] T. Head-Gordon, M. E. Johnson, *Proceedings of the National Academy of Sciences* **103**, 7973 (2006).

[406] F. Tavanti, B. Dianat, A. Catellani, A. Calzolari, *ACS Applied Electronic Materials* **2**, 2961 (2020).

[407] K. A. Dill, *Biochemistry* **29**, 7133 (1990).

[408] Y. Levy, J. N. Onuchic, *Proceedings of the National Academy of Sciences* **101**, 3325 (2004).

[409] H. Hirao, Z. H. Cheong, X. Wang, *The Journal of Physical Chemistry B* **116**, 7787 (2012).

[410] X.-C. Huang, Y.-Y. Lin, J.-P. Zhang, X.-M. Chen, *Angewandte Chemie International Edition* **45**, 1557 (2006).

[411] Y.-Q. Tian, Y.-M. Zhao, Z.-X. Chen, G.-N. Zhang, L.-H. Weng, D.-Y. Zhao, *Chemistry – A European Journal* **13**, 4146 (2007).

[412] H. Hayashi, A. P. Côté, H. Furukawa, M. O'Keeffe, O. M. Yaghi, *Nature Materials* **6**, 501 (2007).

[413] Q. Song, S. K. Nataraj, M. V. Roussenova, J. C. Tan, D. J. Hughes, W. Li, P. Bourgoin, M. A. Alam, A. K. Cheetham, S. A. Al-Muhtaseb, E. Sivaniah, *Energy & Environmental Science* **5**, 8359 (2012).

[414] S. Wang, W. Yao, J. Lin, Z. Ding, X. Wang, *Angewandte Chemie International Edition* **53**, 1034 (2014).

[415] L. T. L. Nguyen, K. K. A. Le, H. X. Truong, N. T. S. Phan, *Catalysis Science & Technology* **2**, 521 (2012).

[416] B. Y. Xia, Y. Yan, N. Li, H. B. Wu, X. W. D. Lou, X. Wang, *Nature Energy* **1**, 15006 (2016).

[417] H.-x. Zhong, J. Wang, Y.-w. Zhang, W.-l. Xu, W. Xing, D. Xu, Y.-f. Zhang, X.-b. Zhang, *Angewandte Chemie International Edition* **53**, 14235 (2014).

[418] T. D. Bennett, J.-C. Tan, Y. Yue, E. Baxter, C. Ducati, N. J. Terrill, H. H.-M. Yeung, Z. Zhou, W. Chen, S. Henke, A. K. Cheetham, G. N. Greaves, *Nature Communications* **6**, 8079 (2015).

[419] T. D. Bennett, Y. Yue, P. Li, A. Qiao, H. Tao, N. G. Greaves, T. Richards, G. I. Lampronti, S. A. T. Redfern, F. Blanc, O. K. Farha, J. T. Hupp, A. K. Cheetham, D. A. Keen, *Journal of the American Chemical Society* **138**, 3484 (2016).

[420] C. Zhou, L. Longley, A. Krajnc, G. J. Smales, A. Qiao, I. Erucar, C. M. Doherty, A. W. Thornton, A. J. Hill, C. W. Ashling, O. T. Qazvini, S. J. Lee, P. A. Chater, N. J. Terrill, A. J. Smith, Y. Yue, G. Mali, D. A. Keen, S. G. Telfer, T. D. Bennett, *Nature Communications* **9**, 5042 (2018).

[421] R. Gaillac, P. Pullumbi, K. A. Beyer, K. W. Chapman, D. A. Keen, T. D. Bennett, F.-X. Coudert, *Nature Materials* **16**, 1149 (2017).

[422] A. Sartbaeva, S. A. Wells, M. M. J. Treacy, M. F. Thorpe, *Nature Materials* **5**, 962 (2006).

[423] S. Riniker, J. R. Allison, W. F. van Gunsteren, *Physical Chemistry Chemical Physics* **14**, 12423 (2012).

[424] W. G. Noid, *The Journal of Chemical Physics* **139**, 090901 (2013).

[425] T. C. Nicholas, A. E. Stones, A. Patel, F. M. Michel, R. J. Reeder, D. G. A. L. Aarts, V. L. Deringer, A. L. Goodwin, *Nature Chemistry* **16**, 36 (2024).

[426] J. P. Dürholt, R. Galvelis, R. Schmid, *Dalton Transactions* **45**, 4370 (2016).

[427] C. M. S. Alvares, G. Maurin, R. Semino, *The Journal of Chemical Physics* **158**, 194107 (2023).

[428] T. C. Nicholas, A. L. Goodwin, V. L. Deringer, *Chemical Science* **11**, 12580 (2020).

[429] S. T. John, G. Csányi, *The Journal of Physical Chemistry B* **121**, 10934 (2017).

[430] T. Lemke, C. Peter, *Journal of Chemical Theory and Computation* **13**, 6213 (2017).

[431] L. Zhang, J. Han, H. Wang, R. Car, W. E, *The Journal of Chemical Physics* **149**, 034101 (2018).

[432] K. K. Bejagam, S. Singh, Y. An, S. A. Deshmukh, *The Journal of Physical Chemistry Letters* **9**, 4667 (2018).

[433] H. Chan, M. J. Cherukara, B. Narayanan, T. D. Loeffler, C. Benmore, S. K. Gray, S. K. R. S. Sankaranarayanan, *Nature Communications* **10**, 379 (2019).

[434] M. Arts, V. Garcia Satorras, C.-W. Huang, D. Zügner, M. Federici, C. Clementi, F. Noé, R. Pinsler, R. van den Berg, *Journal of Chemical Theory and Computation* **19**, 6151 (2023).

[435] H. A. Karimi-Varzaneh, F. Müller-Plathe, *Multiscale Molecular Methods in Applied Chemistry* (Springer, Berlin, Heidelberg, 2012), pp. 295–321.

[436] S. Y. Joshi, S. A. and Deshmukh, *Molecular Simulation* **47**, 786 (2021).

[437] M. Dallavalle, N. F. A. van der Vegt, *Physical Chemistry Chemical Physics* **19**, 23034 (2017).

[438] V. A. Harmandaris, D. Reith, N. F. A. van der Vegt, K. Kremer, *Macromolecular Chemistry and Physics* **208**, 2109 (2007).

[439] H. I. Ingólfsson, C. A. Lopez, J. J. Uusitalo, D. H. de Jong, S. M. Gopal, X. Periole, S. J. Marrink, *WIREs Computational Molecular Science* **4**, 225 (2014).

[440] D. Fritz, K. Koschke, V. A. Harmandaris, N. F. A. van der Vegt, K. Kremer, *Physical Chemistry Chemical Physics* **13**, 10412 (2011).

[441] E. Brini, E. A. Algaer, P. Ganguly, C. Li, F. Rodríguez-Ropero, N. F. A. van der Vegt, *Soft Matter* **9**, 2108 (2013).

[442] O. Delgado Friedrichs, M. O'Keeffe, O. M. Yaghi, *Acta Crystallographica. Section A, Foundations of Crystallography* **59**, 22 (2003).

[443] K. Momma, F. Izumi, *Journal of Applied Crystallography* **44**, 1272 (2011).

[444] T. C. Nicholas, E. V. Alexandrov, V. A. Blatov, A. P. Shevchenko, D. M. Proserpio, A. L. Goodwin, V. L. Deringer, Research data for "visualization and quantification of geometric diversity in metal- organic frameworks" (2021). DOI: 10.5281/zenodo.5271082.

[445] C. Baerlocher, D. Brouwer, B. Marler, L. McCusker, Database of zeolite structures (2020). `https://www.iza-structure.org/databases/` (Accessed: 29/06/2025).

[446] J. P. Dürholt, G. Fraux, F.-X. Coudert, R. Schmid, *Journal of Chemical Theory and Computation* **15**, 2420 (2019).

[447] I. Mosquera-Lois, S. R. Kavanagh, A. Walsh, D. O. Scanlon, *Journal of Open Source Software* **7**, 4817 (2022).

[448] I. Mosquera-Lois, S. R. Kavanagh, A. Walsh, D. O. Scanlon, *npj Computational Materials* **9**, 25 (2023).

[449] J. L. A. Gardner, Z. Faure Beaulieu, V. L. Deringer, *Digital Discovery* **2**, 651 (2023).

[450] S. Chong, F. Grasselli, C. Ben Mahmoud, J. D. Morrow, V. L. Deringer, M. Ceriotti, *Journal of Chemical Theory and Computation* **19**, 8020 (2023).

[451] F. Neese, M. Atanasov, G. Bistoni, D. Maganas, S. Ye, *Journal of the American Chemical Society* **141**, 2814 (2019).

[452] R. Semino, J. P. Dürholt, R. Schmid, G. Maurin, *The Journal of Physical Chemistry C* **121**, 21491 (2017).

[453] C. Ben Mahmoud, L. A. M. Rosset, J. R. Yates, V. L. Deringer, *The Journal of Chemical Physics* **163**, 024118 (2025).

[454] A. M. O. Mohamed, I. G. Economou, H.-K. Jeong, *J. Chem. Phys.* **160**, 204706 (2024).

[455] M. Neumann, J. Gin, B. Rhodes, S. Bennett, Z. Li, H. Choubisa, A. Hussey, J. Godwin, Orb: A Fast, Scalable Neural Network Potential (2024). DOI: 10.48550/arXiv.2410.22570.

[456] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, Q. He, *Proceedings of the IEEE* **109**, 43 (2021).

[457] V. Zaverkin, D. Holzmüller, L. Bonfirraro, J. Kästner, *Physical Chemistry Chemical Physics* **25**, 5383 (2023).

[458] J. S. Smith, B. T. Nebgen, R. Zubatyuk, N. Lubbers, C. Devereux, K. Barros, S. Tretiak, O. Isayev, A. E. Roitberg, *Nature Communications* **10**, 2903 (2019).

[459] M. S. Chen, J. Lee, H.-Z. Ye, T. C. Berkelbach, D. R. Reichman, T. E. Markland, *Journal of Chemical Theory and Computation* **19**, 4510 (2023).

[460] S. Röcken, J. Zavadlav, Enhancing Machine Learning Potentials through Transfer Learning across Chemical Elements (2025). DOI: 10.48550/arXiv.2502.13522.

[461] H. Yang, *et al.*, MatterSim: A Deep Learning Atomistic Model Across Elements, Temperatures and Pressures (2024). DOI: 10.48550/arXiv.2405.04967.

[462] A. E. A. Allen, N. Lubbers, S. Matin, J. Smith, R. Messerly, S. Tretiak, K. Barros, *npj Computational Materials* **10**, 154 (2024).

[463] D. Zhang, *et al.*, *npj Computational Materials* **10**, 293 (2024).

[464] Y. Zhou, M. A. Chia, S. K. Wagner, M. S. Ayhan, D. J. Williamson, R. R. Struyven, T. Liu, M. Xu, M. G. Lozano, P. Woodward-Court, Y. Kihara, A. Altmann, A. Y. Lee, E. J. Topol, A. K. Denniston, D. C. Alexander, P. A. Keane, *Nature* **622**, 156 (2023).

[465] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, R. Girshick, Segment Anything (2023). DOI: 10.48550/arXiv.2304.02643.

[466] T. B. Brown, *et al.*, Language Models are Few-Shot Learners (2020). DOI: 10.48550/arXiv.2005.14165.

[467] A. Merchant, S. Batzner, S. S. Schoenholz, M. Aykol, G. Cheon, E. D. Cubuk, *Nature* **624**, 80 (2023).

[468] R. Bommasani, *et al.*, On the Opportunities and Risks of Foundation Models (2022). DOI: 10.48550/arXiv.2108.07258.

[469] G. Hinton, O. Vinyals, J. Dean, Distilling the Knowledge in a Neural Network (2015). DOI: 10.48550/arXiv.1503.02531.

[470] Z. Allen-Zhu, Y. Li, Towards Understanding Ensemble, Knowledge Distillation and Self-Distillation in Deep Learning (2023). DOI: 10.48550/arXiv.2012.09816.

[471] J. P. Perdew, J. A. Chevary, S. H. Vosko, K. A. Jackson, M. R. Pederson, D. J. Singh, C. Fiolhais, *Physical Review B* **46**, 6671 (1992).

[472] C. Adamo, V. Barone, *The Journal of Chemical Physics* **110**, 6158 (1999).

[473] Y. Liu, S. Agarwal, S. Venkataraman, AutoFreeze: Automatically Freezing Model Blocks to Accelerate Fine-tuning (2021). DOI: 10.48550/arXiv.2102.01386.

[474] J. Howard, S. Ruder, Universal Language Model Fine-tuning for Text Classification (2018). DOI: 10.48550/arXiv.1801.06146.

[475] J. Frankle, M. Carbin, The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks (2019). DOI: 10.48550/arXiv.1803.03635.

[476] G. E. Hinton, S. Roweis, *Advances in Neural Information Processing Systems*, S. Becker, S. Thrun, K. Obermayer, eds. (2002), vol. 15, pp. 857–864.

[477] L. van der Maaten, G. Hinton, *Journal of Machine Learning Research* **9**, 2579 (2008).

[478] L. McInnes, J. Healy, N. Saul, L. Großberger, *Journal of Open Source Software* **3**, 861 (2018).

[479] Y. Rubner, C. Tomasi, L. J. Guibas, *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)* (0007/1998-01-07), pp. 59–66.

[480] N. Bonneel, M. Van De Panne, S. Paris, W. Heidrich, *ACM Transactions on Graphics (TOG)* **30**, 158 (2011).

[481] M. Ernzerhof, G. E. Scuseria, *The Journal of Chemical Physics* **110**, 5029 (1999).

[482] A. van den Oord, Y. Li, O. Vinyals, Representation Learning with Contrastive Predictive Coding (2019). DOI: 10.48550/arXiv.1807.03748.

[483] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A Simple Framework for Contrastive Learning of Visual Representations (2020). DOI: 10.48550/arXiv.2002.05709.

[484] X. Chen, K. He, Exploring Simple Siamese Representation Learning (2020). DOI: 10.48550/arXiv.2011.10566.

[485] G. E. Hinton, R. R. Salakhutdinov, *Science* **313**, 504 (2006).

[486] D. P. Kingma, M. Welling, Auto-Encoding Variational Bayes (2022). DOI: 10.48550/arXiv.1312.6114.

[487] T. N. Kipf, M. Welling, Variational Graph Auto-Encoders (2016). DOI: 10.48550/arXiv.1611.07308.

Appendix

A Classification of amorphous ices

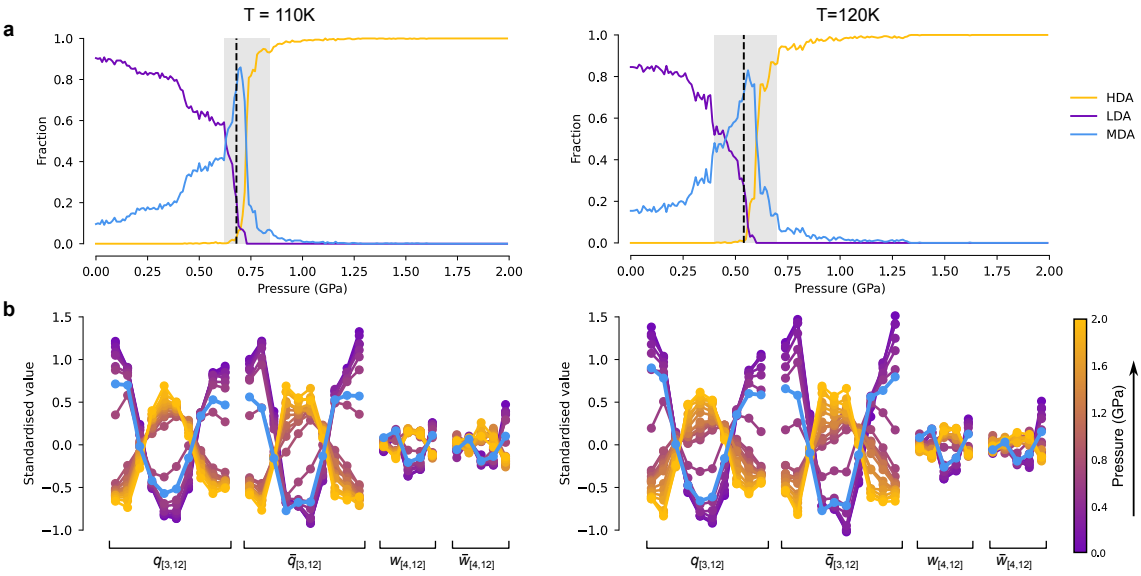

Figure A1: Classification analysis for the isothermal compression of LDA at $T = 120\,\mathrm{K}$ and $T = 140\,\mathrm{K}$. (a) Fraction of local environments as classified by the NN as a function of pressure. The grey band shows the region of the phase transition. The shaded region corresponds to the LDA-to-HDA transition. The dashed line marks the point of maximum structural similarity between compressed LDA and MDA. (b) Evolution of the Steinhardt parameters as a function of pressure.

B Topological origin of peak splitting in the structure factor of liquid water

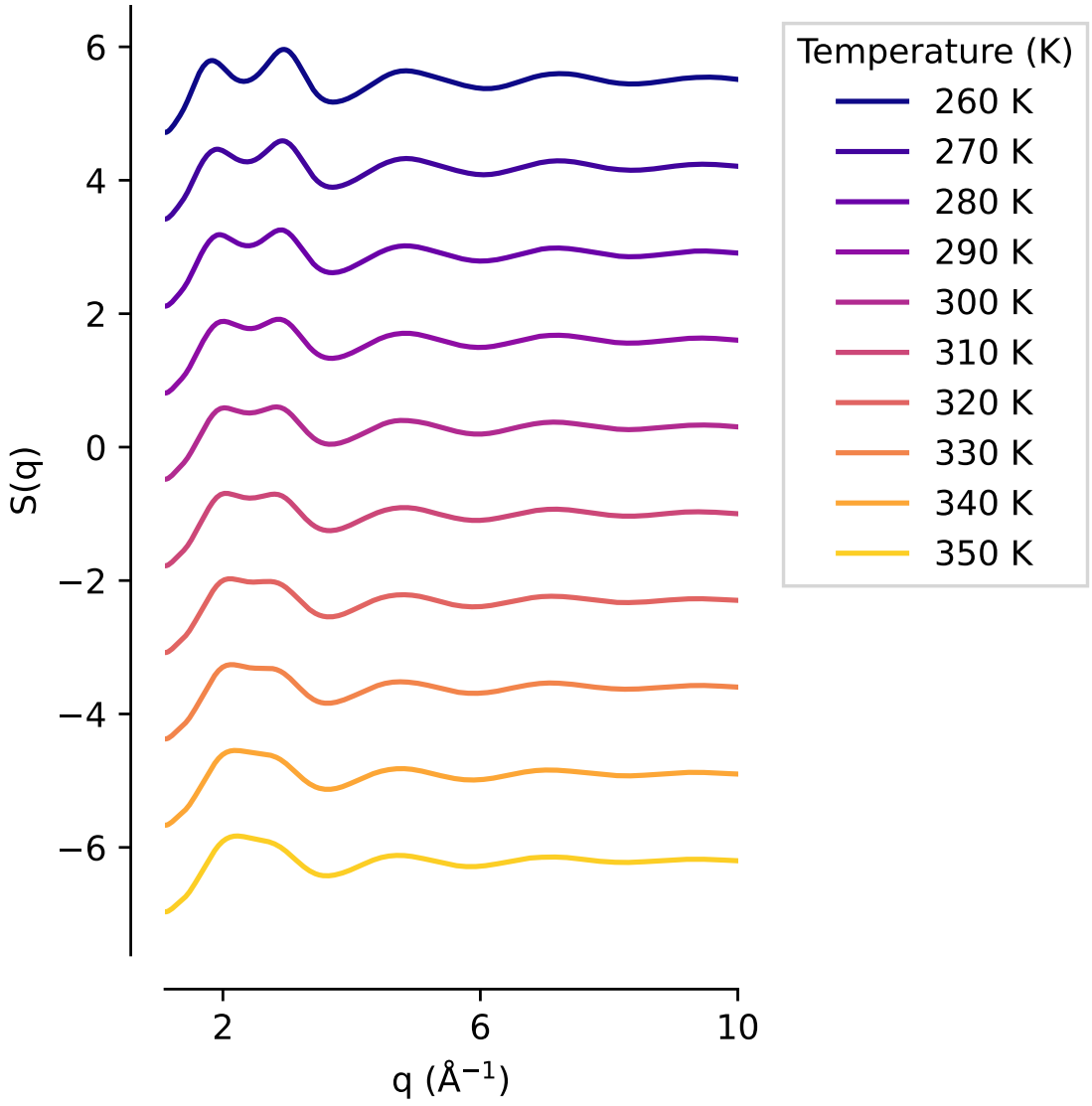

Figure B1: Structure factor $S(q)$ of liquid water at 1 bar for the whole temperature range (260–350 K). The model accurately captures the structure factor of liquid water in the temperature range $T \in [260, 350]$ K and reproduces the experimentally observed splitting of the principal diffraction peak into two distinct maxima upon cooling

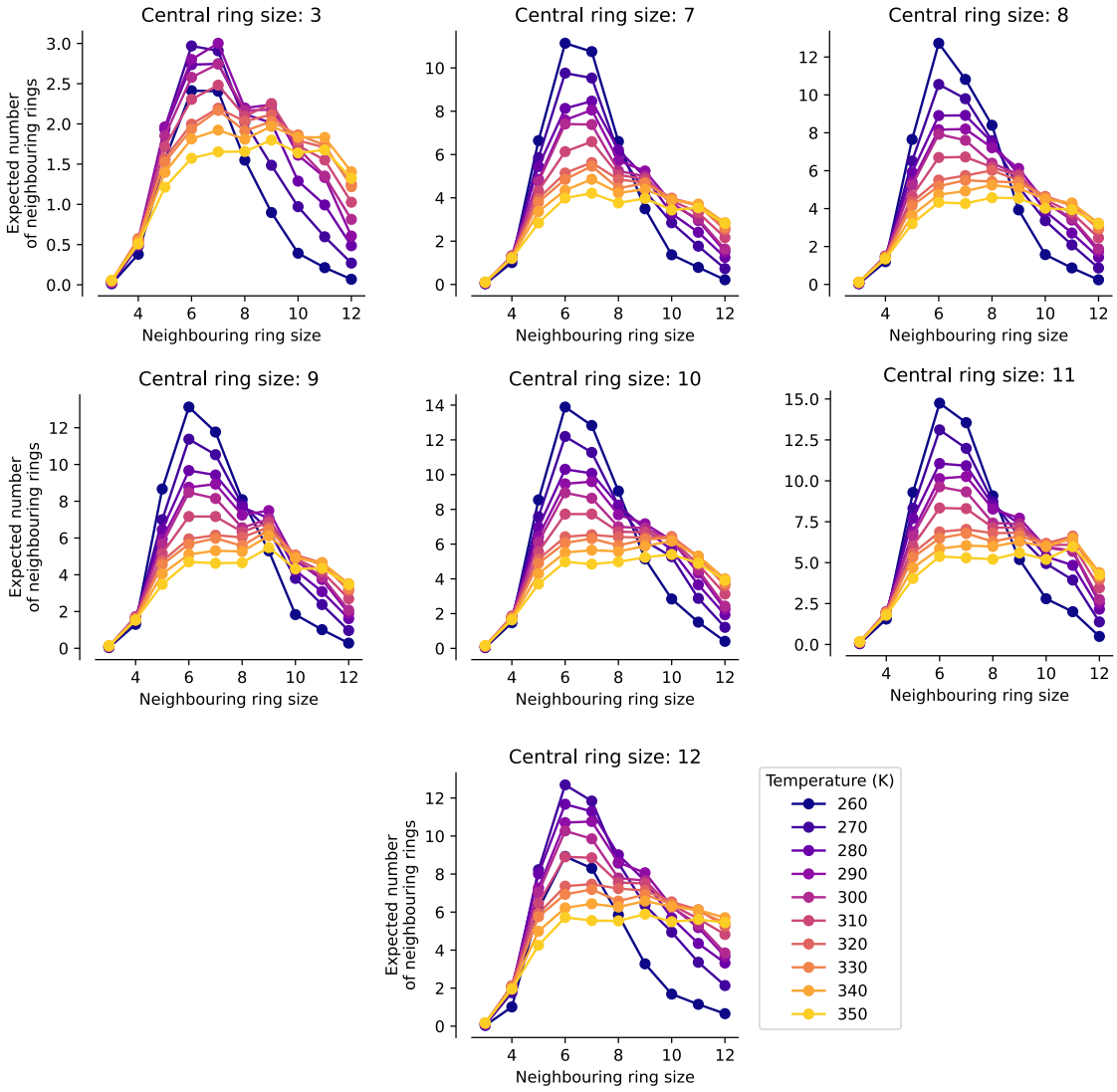

Figure B2: Expected number of neighbouring hydrogen-bonded rings given a central ring size n across the 260–350 K temperature range. A recurring observation across most central ring sizes is the general favouritism for 6-membered neighbouring rings. This preference is consistent with water's inherent tendency towards tetrahedral coordination, which is optimally accommodated by hexagonal arrangements in the hydrogen-bond network.

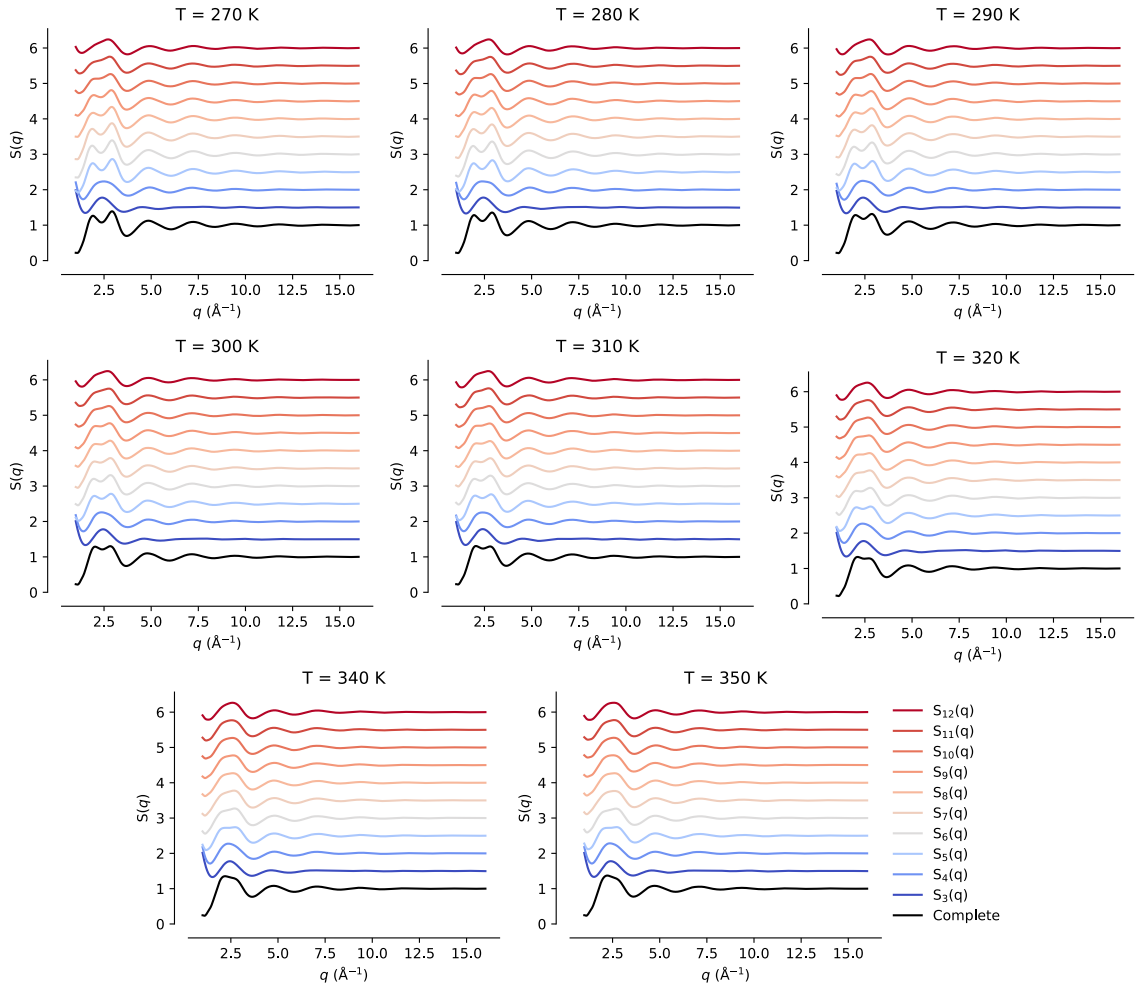

Figure B3: Structure factor $S(q)$ of liquid water at 1 bar decomposed into contributions from individual ring sizes, $S_n(q)$ where $n \in [3, 12]$. The black curves indicate the total original structure factor, while coloured lines represent individual ring contributions.