

Characterising and understanding the microscopic structure of amorphous materials

Thomas Christian Nicholas

Worcester College



A thesis submitted to the University of Oxford
for the degree of Doctor of Philosophy

Supervisors: Prof. Andrew Goodwin & Prof. Volker Deringer

Inorganic Chemistry Laboratory

University of Oxford

Declaration

The work presented in this thesis was carried out between October 2020 and July 2024 in the Inorganic Chemistry Laboratory, University of Oxford under the joint supervision of Prof. Andrew L. Goodwin and Prof. Volker L. Deringer. 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.

Thomas Nicholas

Abstract

Amorphous materials are ubiquitous and play an important role in various applications. Fully exploiting these materials requires a thorough understanding of their structure, which is challenging due to the lack of long-range order that complicates the direct interpretation of experimental data.

This thesis explores the use of hybrid Reverse Monte Carlo (HRMC) refinement, a method in which atomic coordinates in a model system are refined to match experimental total scattering data while also minimising the system energy, for obtaining realistic amorphous models. The key to the success of this method is the use of accurate force fields to describe the energetics.

I use the HRMC approach to refine a model of amorphous calcium carbonate (ACC). I find that the principal component of structural order in ACC is the distribution of calcium ions. To understand this structure better, I extract an effective $\text{Ca} \cdots \text{Ca}$ interaction potential. This potential is well described by a Lennard-Jones–Gauss functional form, and the relevant empirical parameters found for ACC are known from theory to promote structural complexity. In this way, I show that the complex structure of ACC arises from geometric frustration.

Next, I explore the relationship between local geometric preferences and network topology in AB_2 metal-organic frameworks (MOFs). I develop an automated coarse-graining routine to reduce complex structures to their fundamental building units. By combining these coarse-grained representations with atom-density-based descriptors and dimensionality reduction algorithms, I demonstrate how to visualise and quantify structural diversity in MOFs. This investigation reveals that building blocks with similar local geometric preferences can favour distinctly different topologies.

With this result in mind, the final two chapters focus on modelling zeolitic imidazolate frameworks (ZIFs), a subset of MOFs known for their ability to amorphise, melt, and vitrify. In Chapter 4, I detail the development of a fast and accurate machine-learned (ML) potential for ZIFs. I show that HRMC refinement can be used to iteratively improve the ML model’s ability to describe the amorphous ZIF structure.

In Chapter 5, I use the ML potential alongside molecular dynamics and HRMC to generate an amorphous ZIF model, offering insights into the structure and topology of the amorphous MOF. I contextualise the findings by introducing a topological metric based on established crystalline descriptors that is able to discriminate between amorphous networks with different topological features. This enables a discussion of the (dis)similarities between amorphous Si, amorphous SiO_2 , and amorphous ZIF, suggesting that the landscape of AB_2 continuous random networks may be more nuanced than previously understood.

Acknowledgements

First, I would like to thank my supervisors, Andrew Goodwin and Volker Deringer at the University of Oxford, for their unwavering support, guidance, and scientific enthusiasm over the past few years. Their mentorship provided me with the opportunity to work in an exciting field of research and helped me discover my own scientific interests and passions, for which I am sincerely grateful.

I have had the pleasure of working alongside two fantastic teams during my DPhil in the Goodwin and Deringer groups. To all group members, past and present, thank you for your thought-provoking discussions and friendship. I have greatly benefited from the breadth and diversity of scientific interests within the groups. In particular, I would like to thank Ari, Ella, Emily, Johny, Nikolaj, and Simon for always showing me the way ahead. Joe, John, and Yuxing, thank you for your constant encouragement throughout the four years. Elodie, your endless kindness has been invaluable. Quentin, your energy and humour have been uplifting. Daniel, Louise, and Zak, I appreciate the pub trips, hot chocolates, and chats. John and Zoé, thank you for the sports motivation. Chiheb, thank you for putting up with my constant questions. Finally, I want to thank all the Part IIs for your enthusiasm.

I have been privileged to work with many collaborators who have generously contributed their advice and support: Alexander Shevchenko, Adam Stones, Davide Proserpio, Dirk Aarts, Eugeny Alexandrov, F. Marc Michel, Linus Erhard, Martin Dove, Richard Reeder, and Vladislav Blatov.

A special thank you goes to Andrew Seel. Without your constant encouragement and willingness to teach and support me, I wouldn't be where I am today.

The mistakes in this thesis are my own, and occur because I didn't follow the advice and example of all the above people carefully enough.

Thank you to all my Oxford friends for their endless support and for making this city truly a home for the last eight years. A special thanks goes to Liam for your constant friendship; Frances for always being there to talk to; and Jack and Matt for the music, friendship, and dinners.

Thank you to my family: Mum, Dad, James, Mark, and Chip, for your constant love and support, and for your willingness to hear me talk about my science at all times.

Finally, thank you to Sophie for sticking through this with me. If we can write up together, there is surely nothing we cannot do. I cannot wait to start the next chapter with you.

Publications

The following list of my publications are directly relevant to this thesis.

- 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;
“Geometrically frustrated counteractions drive structural complexity in amorphous calcium carbonate”
Nat. Chem. **16**, 36–41 (2024)
- T. C. Nicholas, E. V. Alexandrov, V. A. Blatov, A. P. Shevchenko, D. M. Proserpio, A. L. Goodwin, V. L. Deringer;
“Visualization and Quantification of Geometric Diversity in Metal–Organic Frameworks”
Chem. Mater. **33**, 8289–8300 (2021)

I also include additional publications, not discussed in this thesis, but based on work developed herein:

- P. M. Maffettone, W. J. U. Fletcher, T. C. Nicholas, V. L. Deringer, J. R. Allison, L. J. Smith and A. L. Goodwin;
“When can we trust structural models derived from pair distribution function measurements?”
Faraday Discuss. Accepted manuscript doi: 10.1039/D4FD00106K (2024)
- Z. F. Beaulieu, T. C. Nicholas, J. L. A. Gardner, A. L. Goodwin, V. L. Deringer;
“Coarse-grained *versus* atomistic machine learning for zeolitic imidazolate frameworks”
Chem. Commun. **59**, 11405–11408 (2023)
- E. M. Meekel, T. C. Nicholas, B. Slater, A. L. Goodwin;
“Torsional flexibility in zinc-benzenedicarboxylate metal–organic frameworks”
CrystEngComm **26**, 673–680 (2024)

Abbreviations

ACC	amorphous calcium carbonate
<i>a</i> -Si	amorphous Si
<i>a</i> -SiO ₂	amorphous silica
<i>a</i> -ZIF	amorphous zeolitic imidazolate framework
ACE	atomic cluster expansion
AIMD	<i>ab initio</i> molecular dynamics
ASE	atomic simulation environment
CBU	composite building unit
CHIC	coarse-graining hybrid and inorganic crystals
CIF	crystallographic information file
CP	cross-polarisation
CRN	continuous random network
CSD	Cambridge Structural Database
DFT	density functional theory
EPSR	empirical potential structure refinement
EXAFS	X-ray absorption fine structure
FSDP	first sharp diffraction peak
FWHM	full-width half-maximum
GAP	Gaussian Approximation Potential
GGA	generalised gradient approximation
<i>h</i> -ZIF	hypothetical zeolitic imidazolate framework
HETCOR	heteronuclear correlation
HRMC	hybrid reverse Monte Carlo
Im	imidazolate
IZA-SC	Structure Commission of the International Zeolite Association
LAMMPS	Large-scale Atomic/Molecular Massively Parallel Simulator

LDA	local density approximation
LJG	Lennard-Jones–Gauss
MAE	mean absolute error
MAS	magic-angle spinning
MC	Monte Carlo
MD	molecular dynamics
MDS	multi-dimensional scaling
meta-GGA	meta-generalised gradient approximation
ML	machine learning
MOF	metal–organic framework
NMR	nuclear magnetic resonance
OPP	oscillating pair potential
PAW	projector-augmented wave
PCF	pair correlation function
PDF	pair distribution function
PES	potential energy surface
RDF	radial distribution function
RMC	reverse Monte Carlo
RMSE	root-mean-square error
SBU	secondary building unit
SCAN	strongly constrained and appropriately normed semilocal density
SEM	scanning electron microscopy
SOAP	smooth overlap of atomic positions
SSB	broad spinning sideband
t-SNE	t-distributed stochastic neighbour embedding
UMAP	uniform manifold approximation and projection
VASP	Vienna Ab initio Simulation Package
XPS	X-ray Photoelectron Spectroscopy
ZIF	zeolitic imidazolate framework

Contents

1	Introduction	1
1.1	Canonical amorphous materials	3
1.2	Building structural models of amorphous materials	21
1.3	Aims and structure of this thesis	52
2	The structure of ACC	54
2.1	Acknowledgements	54
2.2	Introduction	54
2.3	Methods	57
2.4	Results	62
2.5	Conclusions	73
3	Geometric Diversity in MOFs	75
3.1	Acknowledgements	75
3.2	Introduction	75
3.3	Methods	78
3.4	Results	84
3.5	Conclusions	95
4	An ML potential for ZIFs	96
4.1	Acknowledgements	96
4.2	Introduction	96
4.3	Methods	99

4.4	Results	114
4.5	Conclusions	131
5	The structure of <i>a</i>-ZIF-4	132
5.1	Acknowledgements	132
5.2	Introduction	132
5.3	Methods	133
5.4	Results	135
5.5	Conclusions	142
6	Conclusion and outlook	143
6.1	HRMC simulations	143
6.2	ACC	144
6.3	MOFs	145
6.4	Final remarks and future directions	147
	Bibliography	148

List of Figures

1.1	Zachariasen CRN model	4
1.2	X-ray diffraction data for <i>a</i> -Si	5
1.3	X-ray diffraction patterns for <i>a</i> -SiO ₂ and cristobalite	7
1.4	SiO ₂ SBUs and CBUs	8
1.5	Tiling representation of the SOD and FAU frameworks	9
1.6	Commonly used topological descriptors	11
1.7	Building units in MOFs and ZIFs	13
1.8	Diverse structural behaviour of ZIF-4	15
1.9	Structural diversity in calcium carbonate	17
1.10	Experimental characterisation of ACC	18
1.11	Overview of empirical potentials	26
1.12	Bond and bond angle terms in empirical potentials.	28
1.13	The Perdew ladder of density functional approximations.	31
1.14	Liquid and glassy ZIF structural models from AIMD.	33
1.15	Three key components in ML potential fitting	34
1.16	Modelling disordered ZIFs with ML potentials	39
1.17	Hand-built models of CRNs	41
1.18	WWW bond-switch move	43
1.19	The decorate and relax method	44
1.20	RMC models of ACC and <i>a</i> -ZIF	48
1.21	EPSR models of ACC	51

2.1	Evolution of system properties during HRMC and MD simulations . .	61
2.2	HRMC refinement for ACC	63
2.3	Characteristics of water distribution in the HRMC ACC model	64
2.4	Key bond-length and bond-angle distributions in ACC	65
2.5	Local structure analysis of ACC	66
2.6	Ca-pair correlation functions in ACC	67
2.7	Effective Ca $\cdots$ Ca interactions in ACC	68
2.8	Comparing HRMC and MC-driven LJG configurations	71
2.9	ACC in the context of the complex LJG potential phase-space	72
3.1	Overview of the cg-SOAP-based methodology	78
3.2	Geometric diversity within isorecticular groups of AB ₂ MOFs	85
3.3	Geometric diversity within MOFs of dia topology	87
3.4	Visualising geometric diversity using different embedding schemes . .	89
3.5	Using cg-SOAP analysis to identify relationships in RefCode families	91
3.6	Structure property distributions for different degrees of interpenetration	93
3.7	Distributions of A-site heterogeneity for different topologies	94
4.1	Rigid body effects in PDF calculations	113
4.2	Structure map of the ZIF ML potential training dataset	115
4.3	A data-driven protocol for modelling amorphous structures	121
4.4	Numerical errors during ML potential iterative training	125
4.5	Hyperparameter optimisation of ACE ML potential models	128
4.6	Learning curves for the final ACE ML potential	129
5.1	Optimal data and energy weightings in HRMC	134
5.2	Temperature profile of melt-quench protocol for <i>a</i> -ZIF-4 models . . .	136
5.3	Experimental validation of <i>a</i> -ZIF-4 models	137
5.4	The structure of <i>a</i> -ZIF	138
5.5	Comparing CRNs of <i>a</i> -Si, <i>a</i> -SiO ₂ , and <i>a</i> -ZIF-4	139

List of Tables

1.1	Details of different ZIF ML potentials in the literature.	37
2.1	Radial cut-offs for determining local coordination environments. . . .	62
3.1	Characteristics of commonly occurring AB_2 MOF topologies	86
4.1	The final GAP ML potential fitting parameters.	103
4.2	GAP regularisation parameters.	105
4.3	Summary of the composition of the ML potential reference dataset. .	123
4.4	Optimised ACE hyperparameters.	127
4.5	Comparing different ZIF ML potentials.	130
5.1	Topology metric for crystalline and amorphous materials.	141

Chapter 1

Introduction

Amorphous (or ‘non-crystalline’) materials are metastable, kinetically-stabilised forms of matter characterised by the absence of long-range order in the positions of their constituent atoms [1, 2]. This broad definition encompasses a wide variety of chemistries and spans different length scales.

On the atomic scale, systems like doped amorphous Si (*a*-Si) are key electronic materials for thin-film transistors and solar photovoltaics [3]. Another example is amorphous SiO₂ (*a*-SiO₂), a material used for window glass since the Bronze Age (*ca.* 3000 BC) [4]. A glass may be defined as an amorphous solid exhibiting a glass transition: a phenomenon where a solid amorphous phase shows an abrupt change in derivative thermodynamic properties (e.g., heat capacity), from crystal-like to liquid-like values with a change in temperature [2].

On the molecular scale, the family includes systems with charged components interacting *via* electrostatic interactions, such as calcium carbonate and calcium phosphate, important precursors in biomineralisation processes [5], or those forming covalent extended networks with metals linked by organic molecules, such as metal–organic frameworks (MOFs) [6, 7]. On larger length scales, there are polymer and colloidal systems whose assembly is governed by the competition between long-range van der Waals attraction and electrostatic repulsion.

To develop a detailed understanding of the physical and chemical properties of a material, it is essential to know the structural arrangement of its atoms. For crystalline materials, this analysis is now routine. By determining the structure within the unit cell, which typically contains relatively few atoms, the entire crystal structure can be generated by repeating the unit cell periodically in space [8]. The absence of structural periodicity in amorphous materials, however, means their structure cannot be described in terms of a finite-sized unit cell. Solving for the

structure is therefore a significant challenge.

It is important to note that although an amorphous material is non-periodic, it is not entirely random in a statistical sense. Most amorphous materials contain some degree of local order. For example, the structure of a -SiO₂ in the glassy state was anticipated as being composed of SiO₄ tetrahedra, the fundamental building unit, connected together with some degree of randomness, long before a ‘complete’ understanding of the structure was obtained [9]. In general, some degree of local order will often be found, although the extent and length scale over which it occurs will strongly depend on the chemical interactions present in the system.

It is therefore helpful to distinguish between two different length scales: short-range order (on the order of a few Ångströms) and intermediate-range order (on length scales of 5–20 Å) [2, 10]. The short-range order in amorphous materials is often probed using diffraction experiments, such as those involving X-rays, neutrons, or electrons, all of which have wavelengths comparable to interatomic spacings [2, 11]. These experiments can be used to identify the fundamental building units. While this is a useful starting point, it is often insufficient by itself to provide a comprehensive understanding of a material. The challenge then becomes determining how these building units assemble on longer length scales. This problem is most often addressed through atomistic modelling, where the goal is to develop a model of the structure that agrees with the experimental data available for the material [4]. These two aspects of amorphous material characterisation are central to the thesis presented herein.

This introductory chapter is structured as follows. In Section 1.1, I discuss four amorphous materials relevant to this thesis: a -Si, a -SiO₂, an amorphous zeolitic imidazolate framework (a -ZIF), and amorphous calcium carbonate (ACC). Through these examples, I will identify the key properties and structural characteristics that can—or cannot—be directly inferred from scattering and PDF analysis, as well as other experimental techniques where appropriate.

Excluding ACC, these materials are typically described as ‘network materials’: they feature well-defined bonds, marked by a clear separation between the nearest and next-nearest neighbours. These bonds have covalent (directional) characteristics, connecting atoms or molecules to form extended networks [12]. The structure of network materials are commonly analysed with topological approaches. I will therefore introduce examples of topological metrics throughout this chapter.

In Section 1.2, I introduce computational techniques that are used to build structure models for amorphous materials. This discussion will begin by introducing ‘direct methods’: protocols that use a description of the potential energy of a given system to drive the simulation. There are two components to be discussed here: first, how the interactions between atoms are modelled; second, how these models for interactions are used, alongside experimental data, to generate models of amorphous materials. I will then introduce ‘inverse methods’: protocols that start with experimental data and work backwards to infer the structure of the material. I will conclude this chapter by summarising the aims of this thesis.

1.1 Canonical amorphous materials

1.1.1 *a*-Si

a-Si is a canonical example of a covalently bonded disordered material. Its structure is generally approximated as a continuous random network (CRN) model of tetrahedrally coordinated Si centres [13]. CRNs were introduced in 1932 by Zachariasen to describe the atomic arrangement in vitreous oxide glasses, as depicted in Fig. 1.1 [9]. The CRN model provided an alternative to the ‘crystallite hypothesis’, which describes the structures of glasses as consisting of finite-sized (approximately 15 Å) crystalline domains [14]. Instead, Zachariasen proposed that the network in glass is characterised by an infinitely large unit cell containing an infinite number of atoms. The interatomic forces in the glass are of a similar type and magnitude to those in the crystal, resulting in similar local atomic environments described in terms of

the packing of coordination polyhedra. The principal difference lies in the relative arrangement of these polyhedra: in crystalline forms, the relative orientation of polyhedra repeats throughout the entire structure, whereas in vitreous solids, the relative orientations of polyhedra can vary widely.

High-energy X-ray diffraction experiments provide evidence supporting the CRN description of *a*-Si [15, 16]. Recording the coherent scattering up to momentum transfers $Q = 4\pi \sin(\theta/\lambda) = 55 \text{ \AA}^{-1}$ [Fig. 1.2(a)], where θ is the scattering angle between the incident and scattered X-ray beam and λ is the wavelength of the incident X-rays, offers sufficient spatial resolution to perform a quantitative analysis on the PDF. The reduced radial distribution function $G(r)$ is obtained by taking the Fourier transform of the structure factor, $S(Q)$. The radial distribution function (RDF), as plotted in Fig. 1.2(b), is given by $J(r) = 4\pi r^2 \rho_0 + rG(r)$, where ρ_0 is the average atomic density of the sample.

Focusing initially on the first-neighbour peak in the RDF, two key observations can be made. First, the average Si-Si bond length in the annealed sample of *a*-Si is identical to that in crystalline Si (*c*-Si). Second, integration of the first-neighbour peak gives a coordination number approaching the ‘ideal’ $N = 4$ coordination number found in *c*-Si. On average, therefore, the network is fourfold-connected and

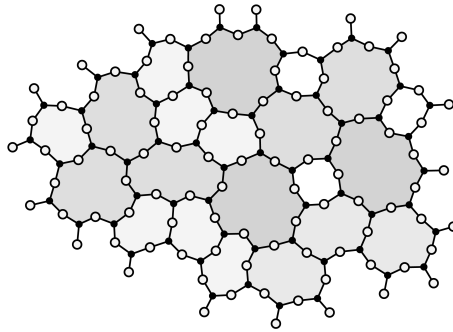


Figure 1.1: Zachariasen’s CRN model for a two-dimensional oxide glass. Silicon and oxygen atoms are depicted in black and white, respectively. The rings are shaded according to ring size: lighter shades are smaller rings, darker shades are larger rings. The image is a digitised version of the original sketch shown in Ref. 9.

resembles the local geometric motif in *c*-Si. Nonetheless, the lack of Bragg peaks in $S(Q)$ [Fig. 1.2(a)] implies the absence of periodicity and symmetry in the network connectivity of these motifs. Indeed, the broad signals observed for magic-angle spinning (MAS) solid-state ^{19}Si nuclear magnetic resonance (NMR) experiments performed on an *a*-Si sample are consistent with a broader range of coordination environments within the amorphous form compared to *c*-Si [17].

The ‘perfect’ CRN model—defined as maintaining the ideal coordination number at each node (here, $N = 4$)—has proven capable of reproducing many of the aforementioned experimental observations, including being in agreement with electronic structure calculations relevant to its characteristic electronic and optical properties. Nevertheless, the roles of structural defects cannot be overlooked. Diffraction experiments clearly identify the presence of under-coordinated sites [15], which helps to explain the slightly lower density of *a*-Si compared to *c*-Si.

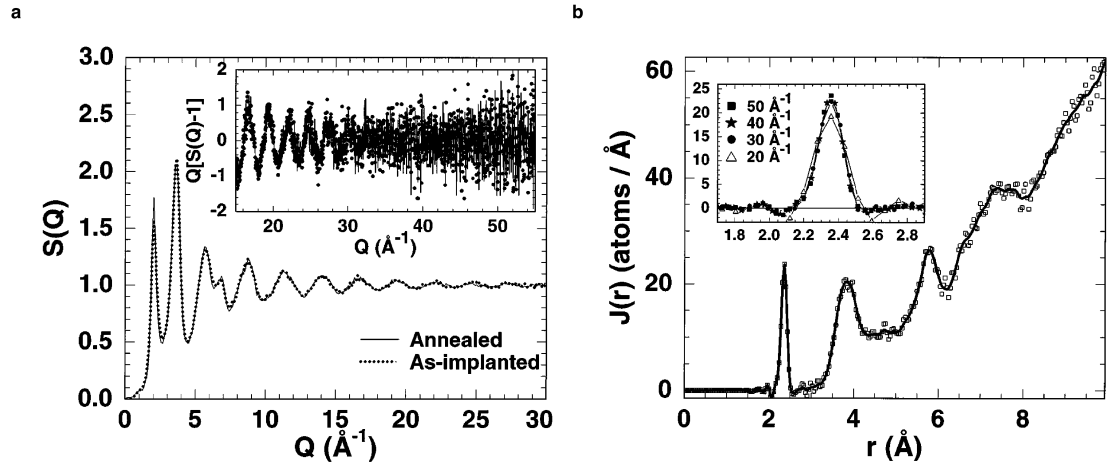


Figure 1.2: X-ray diffraction data for *a*-Si. **a**, The structure factor, $S(Q)$, of the as-implanted (dots) and annealed (solid line) *a*-Si samples. The inset shows the corresponding interference functions, $F(Q)$, of the two samples in the high- Q region. **b**, The radial distribution function, $J(r)$, of annealed *a*-Si calculated before (squares) and after (solid line) filtering the high- Q noise from $S(Q)$. The inset shows how the size and shape of the first-neighbour peak varies with different Q_{max} limits when computing $J(r)$. The figures are reproduced from Ref. 16 with permission from the American Physical Society.

1.1.2 α -SiO₂

α -SiO₂, or silica, is a prototypical glass-forming material, similarly studied for the role of network topology on its structure and properties. Silica is one of the most abundant materials on Earth, playing a crucial role in various natural and industrial processes [18]. This versatile material is of interest for its various applications, including in window glass, oscillators [19], and catalysts [20].

The versatility of silica is largely attributed to its rich polymorphism, that is, its ability to adopt multiple structural forms. Considering first the crystalline polymorphs: the thermodynamically stable phase at ambient conditions, α -quartz is a piezoelectric material [21]; the α -cristobalite polymorph has a negative Poisson ratio, an unusual property for minerals* where the material contracts laterally when compressed and *vice versa* [22]; and a variety of low-density, porous zeolite polymorphs are widely used as catalysts in the petrochemical industry [23]. α -SiO₂ finds application in optical fibres [24], and recent work on amorphising zeolites has led to the formation of glassy materials with specific cavity sizes, defining the properties and functionalities [25–27].

The primary building unit found in the majority of silica forms, and particularly those relevant to this thesis, is the tetrahedral SiO₄ motif. This structural unit can be directly inferred from diffraction experiments, as I will come to describe below, regardless of the sample’s crystallinity. This is a key strength of the PDF method. The PDF method does not require assumptions about the material’s state, making it applicable to liquids, amorphous solids, and powdered crystalline materials alike. To illustrate this point, I will describe the original X-ray diffraction experiment and analysis on α -SiO₂.

Figure 1.3(a) shows the X-ray diffraction patterns for α -SiO₂ (*top*) and cristobalite (*bottom*) captured using monochromatic Cu K- α radiation using a vacuum

*Such materials are commonly referred to as ‘auxetics’ and usually feature uniquely oriented, hinged structures.

camera in 1936 [28]. After obtaining and correcting the scattering intensity from these microphotometer records, the PDF was derived, as shown in Fig. 1.3(b). Analysis of the peak positions and integrated areas, along with comparing to the PDFs of known crystalline SiO_2 polymorphs [30], collectively suggested that α - SiO_2 consists of vertex-sharing SiO_4 tetrahedra [29].* It should be noted that the consideration of the peak widths in the structure factors and the absence of small-angle scattering also provided compelling evidence in support of the CRN model of α - SiO_2 , compared to the crystallite hypothesis [29].

The structural diversity in silica, particularly in low-density zeolite polymorphs and α - SiO_2 , can be understood through the various ways in which SiO_4 tetrahedra assemble to form extended networks. Theoretically, there are over one million possible fourfold-connected nets that silica could adopt [34–36], but only a much smaller number (approximately 60 pure-siliceous crystalline frameworks) have been experimentally realised. It is therefore beneficial to identify which topological motifs

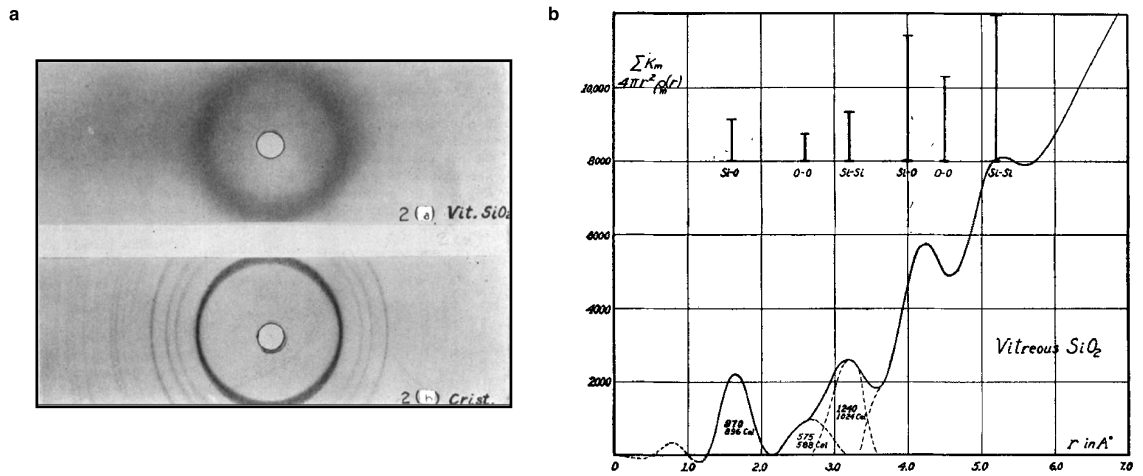


Figure 1.3: X-ray diffraction patterns for α - SiO_2 and cristobalite. **a**, X-ray diffraction patterns for α - SiO_2 (top) and cristobalite (bottom) [28, 29]. **b**, The corresponding PDF computed from the diffraction pattern in panel **a**. The figures are reproduced from Ref. 29 with permission from the Journal of the American Ceramic Society.

*More recent diffraction experiments make use of isotopic substitution to determine the Faber-Ziman partial structure factors [31, 32]. The Fourier transform of these structure factors yields the partial differential distribution functions $D_{\alpha,\beta}(r)$ which support the conclusions of the earlier experiments [33].

occur more frequently. This task is well-suited for analysis within the framework of mathematical tiling theory [35, 37, 38].

Tiling refers to the periodic subdivision of space into bounded, connected regions without holes, known as tiles [35]. These tiles are formed by assembling the primary building unit, SiO_4 , into chain- or layer-like components called secondary building units (SBUs). Figure 1.4(a) illustrates all the SBUs required to construct known zeolite frameworks. SBUs are theoretical topological building units, derived under the assumption that the entire framework consists solely of one type of SBU. Composite building units (CBUs) are a related concept that relaxes this requirement, allowing for the construction of entire frameworks from a combination of different units. The CBU representation facilitates the visualisation of how common building units assemble to form framework structures [Fig. 1.4(b)] [40]. Both representations

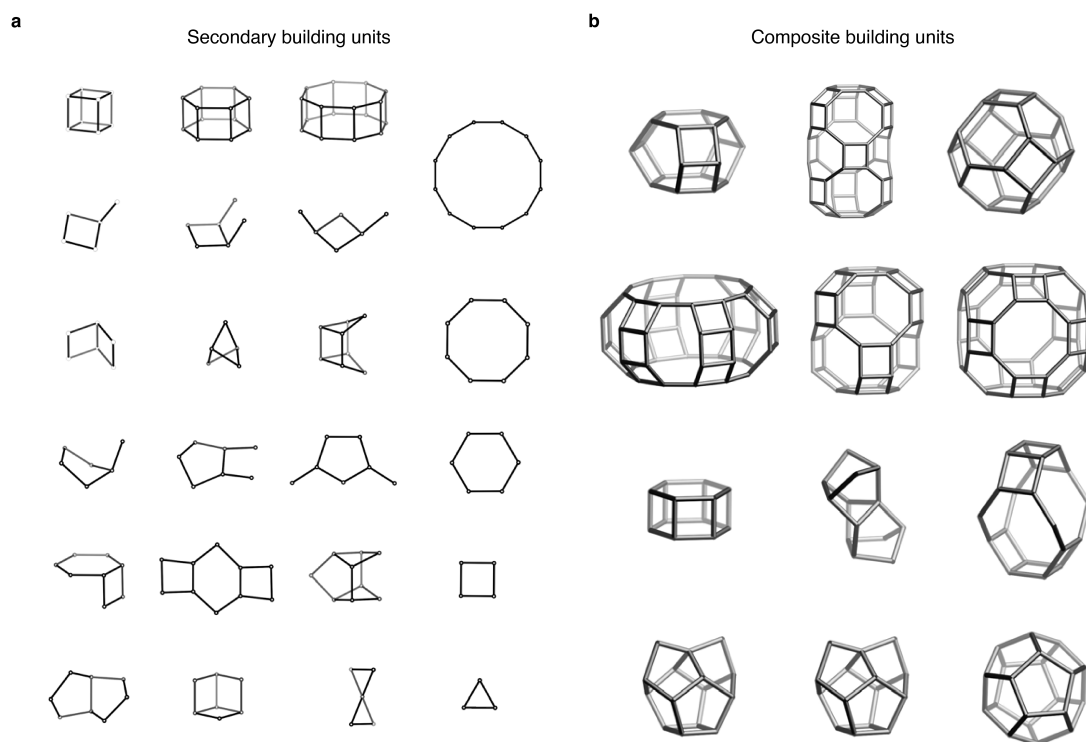


Figure 1.4: Building units in SiO_2 . **a**, Secondary building units (SBUs) are chain- or layer-like assemblies of polyhedra which are derived assuming the entire framework is made up of one type of SBU only [39]. **b**, Composite building units (CBUs) are commonly occurring structural motifs that provide useful relationships between framework types, and provide a more visually obvious link to the tiling representation of frameworks.

highlight that certain ring sizes are more prevalent and that specific geometric relationships between adjoining rings of particular sizes are favoured. For example, 4-, 5-, 6-, and 8-membered rings are observed with high frequency in zeolite frameworks [41, 42]. Such considerations are often taken into account when assessing the synthetic feasibility of a new framework, or for validating amorphous models when directly probing the topology experimentally is not possible.

The CBUs then assemble to tile three-dimensional space by sharing faces. In Fig. 1.5 I have illustrated how CBUs assemble for the sodalite (**SOD**) and faujasite (**FAU**) frameworks. This tiling representation is particularly valuable as it highlights a fundamental characteristic of these structures: namely, their porosity. For a faithful tiling in Euclidean space, tiles must also fill the void spaces within the network. Tiles that accommodate the voids within the **SOD** and **FAU** nets are

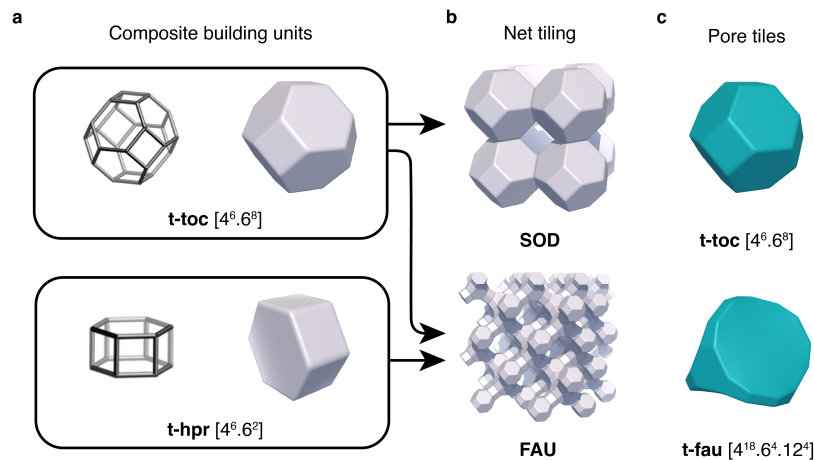


Figure 1.5: Tiling representation of the SOD and FAU frameworks. **a**, Illustration of the composite building units found in the **SOD** and **FAU** frameworks. Each tile is identified by a face symbol of the form $[M^m.N^n\dots]$, which details the configuration of m faces forming M -membered rings, n faces forming N -rings, and so forth [38]. **b**, Illustration of how these tiles assemble to form three-dimensional zeolite networks. The **SOD** net is constructed by linking **t-toc** tiles *via* their square faces, whereas the **FAU** net combines **t-toc** and **t-hpr** tiles. **c**, Zeolite pore sizes depend on the type and arrangement of tiles. The **SOD** network features pores identical to the **t-toc** tiles, accommodating spheres up to 6.3 Å in diameter [43]. Conversely, the **FAU** network forms a more expansive three-dimensional pore system with **t-fau** tiles, suitable for spheres up to 11.2 Å in diameter [43]. The SBU and CBU framework illustrations were taken from Ref. 39. The tiling illustrations were generated using the Gavrog software [44].

shown in Fig. 1.5(c). These voids can vary considerably in size and shape, with diameters ranging in magnitude from a few Ångströms to several nanometers. This diversity in topological structure significantly influences the physical and mechanical properties of the materials. Consequently, zeolites find widespread application in various industrial settings, including catalysis [45–47], ion-exchange [48, 49], and molecular sieving [50]. By considering their ideal topology, along with factors such as the Si/Al molar ratio, one can select zeolites with optimal properties for specific applications.

Quantitative relationships between topology and structural properties have been made possible by encoding the framework structure in terms of local topology descriptors [38, 51, 52]. Three commonly used and complementary descriptors are the coordination sequence, point symbol, and vertex symbol [51]. I proceed to explain the meanings of each of these descriptors in turn.

The coordination sequence of a graph measures how many vertices the graph has at each distance from a fixed vertex [53]. In Fig. 1.6(a) I show four examples of four-coordinate networks where the coordination sequences differ at the second layer. The maximum number of second-nearest neighbours a four-connected network can have is 12 (i.e., each of the nearest neighbours is also fourfold-connected), and with specific connections this number may decrease. Coordination sequences have been successfully used to determine the topological density of crystals [55], correlating well with properties such as lattice energy [56] and catalytic activity of porous materials [57, 58].

The point symbol describes the cycles of a graph around a given node. A cycle is defined as a sequence of vertices in a graph where each vertex v_i is connected to the next vertex v_{i+1} by an edge, and no vertex appears more than once in the sequence, except for the first and last vertices, which are the same. For an n -coordinated vertex in a three-dimensional net, there are $n(n-1)/2$ angles. The point symbol takes the form $A^a.B^b.C^c....$, indicating that there are a angles with shortest cycles

of size A , b angles with shortest cycles of size B , and so on, where $A < B < C < \dots$ and $a + b + c + \dots = n(n - 1)/2$. To illustrate this, in Fig. 1.6(b) I have shown a ‘topology cluster’ for the topologically unique node (yellow) in the **FAU** zeolite net [54]. It is a four-coordinated vertex, and therefore has six angles. Three of those angles have four-membered cycles as the shortest cycle. The remaining three angles have six-membered cycles as the shortest cycles. The **FAU** net therefore has the point symbol $\{4^3.6^3\}$ (because all nodes are equivalent).

The vertex symbol is similar to the point symbol, but with two main differences. First, it considers the *rings* about a node rather than *cycles*. A *ring* (sometimes called a fundamental circuit) is a cycle that is not the sum of *two* smaller cycles. The vertex symbol describes the shortest rings at each angle around each vertex. It is formatted as $[A_a.B_b.C_c]$, where A is the ring size and a is the count of these rings around the first angle, with B and C representing subsequent angles similarly.

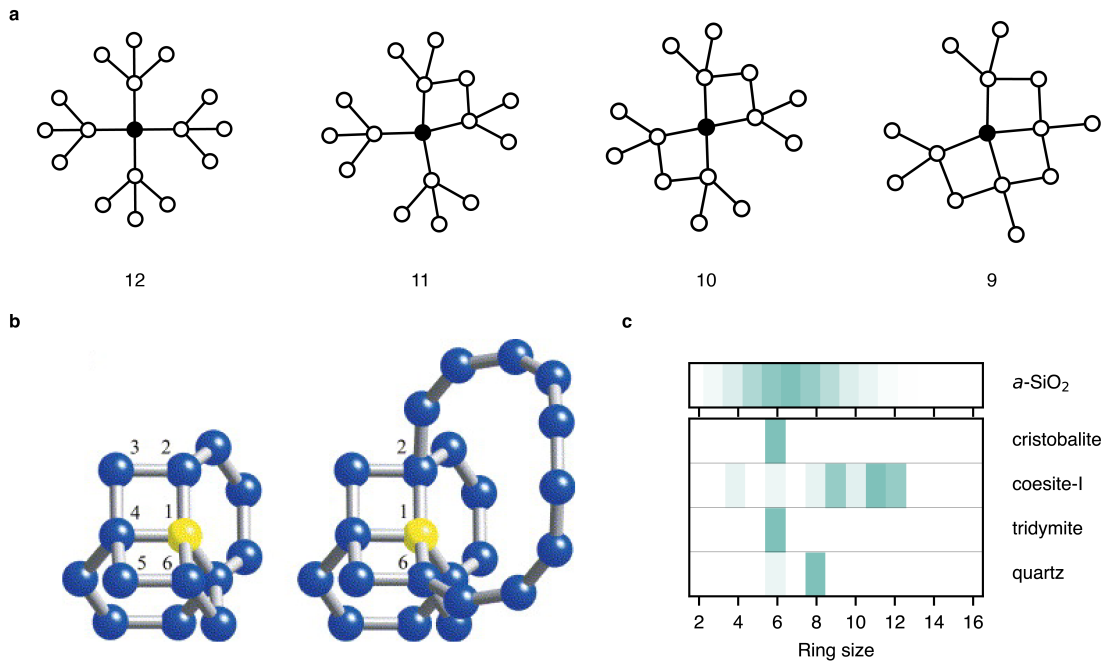


Figure 1.6: Example local topological descriptors for graphs. **a**, Coordination sequences with the numerical values detailing the number of tetrahedra in the second coordination layer from central (black) node. **b**, Topology cluster for the unique node in the **FAU** net [54]. **c**, Ring size distribution for crystalline and amorphous silica. The figure in panel **b** is reproduced with permission from Ref. 54.

Returning to the **FAU** example, we have six angles and therefore expect six elements in the vertex symbol. Here, similarly to the circuits, we have three angles that have four-membered rings as the shortest rings. Two of the remaining angles have six-membered rings as the shortest rings. The final angle, formed between the indices labelled 2–1–6, requires closer attention. Although this has a six-membered cycle, this cycle is built from two edge-sharing four-membered rings (i.e., a short-cut along the bond 4–1 exists). Searching further, the smallest ring is a 12-membered ring. The vertex symbol is therefore [4.4.4.6.6.12].

Together, these symbols provide a unique identifier for the topology of a given structure. Each unique net is labelled with a three-letter code, which facilitates efficient communication about the extended structure [59]. Because the topology is based on the underlying graph representation of a structure, for deformations sufficiently small that bonds are not broken or formed, it remains invariant to bending, stretching, squeezing, and chemical composition. This makes it a useful structural descriptor for comparing materials with different chemical building units.

The practical utility of the topological symbols introduced in this section is largely derived from the crystallinity of the materials studied. For example, the point symbol for the **FAU** net corresponds to the point symbol of a single node, as there is only one unique node. If there were several unique nodes, the label would become cumbersome. This complexity is inherent in amorphous networks. Therefore, statistical averages, such as ring size distributions, are more commonly employed for such materials. For instance, in Fig. 1.6(c) I show the ring size distribution for a model of *a*-SiO₂ [60] alongside the ring size distributions of crystalline silica polymorphs. The amorphous distribution is broader than that of any individual crystalline polymorph but is centred around similar values. This indicates that the intermediate-range order in the amorphous form shares features with the crystalline polymorphs.

1.1.3 *a*-ZIF

A single topology can be translated into a variety of conceptually-related materials by ‘decorating’ the network with different node and linker chemistries. This conceptual mapping is referred to as the ‘reticular chemistry approach’ [61, 62]. MOFs are a family of materials that exemplify this approach [63]. MOFs are polymeric hybrid materials comprised of metal ions or metal clusters and organic linkers that assemble to form periodic networks [Fig. 1.7(a)].

For example, zeolitic imidazolate frameworks (ZIFs) are a subfamily of MOFs, with tetrahedral metal nodes (e.g., Zn^{2+} or Co^{2+}) linked by imidazolate-derived linkers (e.g., $\text{Im}=\text{C}_3\text{H}_3\text{N}_2^-$) [Fig. 1.7(b)], forming extended networks similar to those found for zeolites [64–68]. The mapping from zeolites to ZIFs can be understood in terms of their shared bridging geometry, as shown in Fig. 1.7(c) [66]. Indeed, like zeolites, ZIFs have been synthesised with a multitude of tetrahedral topologies, typically with larger void space and greater mechanical flexibility than zeolites [67–69]. For example, ZIF-8 [$\text{Zn}(\text{mIm})_2$; mIm =2-methylimidazolate] is the prototypical ZIF with the sodalite (**SOD**) topology [66]; the same topology is adopted by the naturally occurring aluminosilicate mineral, zeolite chlorosodalite [70]. ZIF-8 has a bulk modulus (high-pressure crystallography: 6.5 GPa [71]; Brillouin scattering: 7.7 GPa [72]) that is ≈ 7 –9 times greater than the naturally occurring zeolite (ultra-

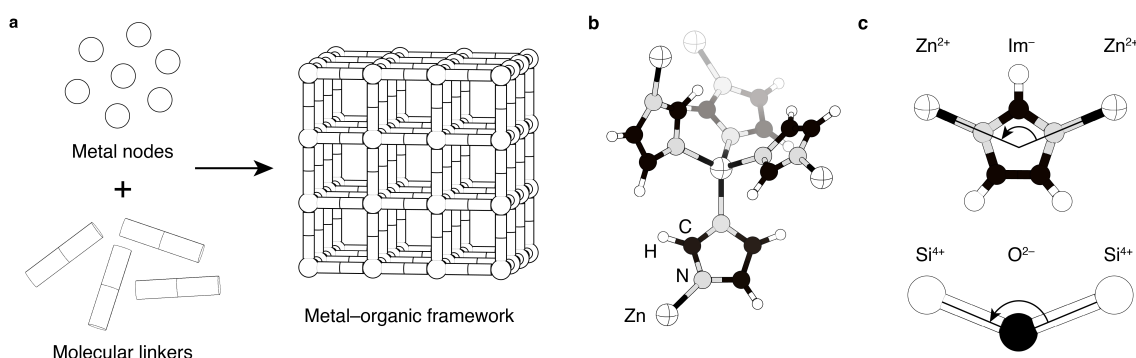


Figure 1.7: Building units in MOFs and ZIFs. **a**, Schematic representation of the assembly of a MOF from the constituent nodes and linkers. **b**, The $\text{Zn}(\text{Im})_4$ tetrahedral motif in ZIFs. **c**, ZIFs and SiO_2 share similar node–linker–node bridging geometries.

sonic measurements: 55.3 GPa [70]), thereby representing a much lower resistance of the structure against volumetric strains under a hydrostatic pressure [73].

Despite the conceptual and geometric similarities described above, zeolites and ZIFs are found empirically to favour different tetrahedral network geometries [66, 74]. The reason why these material families adopt different topologies remains an open question [69]. It is important to note that, for both zeolites and ZIFs, the various polymorphs are typically kinetic phases that are strongly dependent on synthetic conditions, including the role of solvent in the pores, structure-directing agents, and temperature [74]. For example, variations in synthesis conditions alone yields 18 polymorphs for the $\text{Zn}(\text{Im})_2$ composition, each with distinct structures, densities, and topologies [75]. This is suggestive of a complex and nuanced energetic landscape where small external triggers (e.g., temperature, solvent choice) can promote different structural phases. This readiness to form and trap kinetic polymorphs might, in part, explain the susceptibility for ZIFs to undergo amorphisation [71, 76].

ZIF-4 is the most energetically stable *microporous* $\text{Zn}(\text{Im})_2$ polymorph [77]. ZIF-4 demonstrates diverse structural behaviour under non-ambient conditions, including various amorphous forms, as shown in its pressure–temperature phase diagram [Fig. 1.8(a)] [75]. When subjected to hydrostatic pressure at ambient temperature, ZIF-4 undergoes two transitions to different closed-pore phases before amorphising [Fig. 1.8(a), *top left*, blue shading] [75, 78]. Heating ZIF-4 to 300 °C causes an irreversible crystal-to-amorphous transition, resulting in an *a*-ZIF form that is stable at ambient conditions [Fig. 1.8(a), *middle*, blue shading] [76]. Further heating of the *a*-ZIF to 400 °C induces a transition to the dense ZIF-**zni** crystalline phase.

Both the ZIF-4 and ZIF-**zni** polymorphs can be melted by heating to ≈ 600 °C in an argon tube furnace [Fig. 1.8(a), *bottom right*, blue shading] [79, 80]. Quenching this liquid state below the decomposition temperature leads to the formation of melt-quenched ZIF glasses. Notably, these glassy states have a higher density compared to the previously mentioned heat-induced *a*-ZIF form.

The structure of *a*-ZIF—be it the temperature-induced forms or the glassy structures—can be understood in terms of CRNs analogous to that of *a*-SiO₂ [76, 80]. That the amorphous form maintains, on average, a Zn(Im)₄ tetrahedral coordination can be inferred from the experimental PDF. I show in Fig. 1.8(b) and (c) the neutron total scattering function, $F_N(Q)$, and the corresponding neutron differential correlation function $D(r)$, respectively, obtained for the temperature-induced *a*-ZIF form [Fig. 1.8(a), *middle*, blue shading]. The low-*r* region remains essentially unchanged compared to the crystalline phases [76]. This suggests the local tetrahedral motif is also present in *a*-ZIF. Furthermore, comparing $D(r)$ for *a*-ZIF and *a*-SiO₂, where the latter data has been scaled in *r* by a factor of two to reflect the approximate difference in Zn···Zn and Si···Si distances [Fig. 1.8(c)], suggests that the two materials share key features in their intermediate-range structure [76].

NMR has been used to provide complementary insight into the structure of ZIF glasses to that gained from diffraction data [81–84]. High field (35.2 T) ⁶⁷Zn MAS

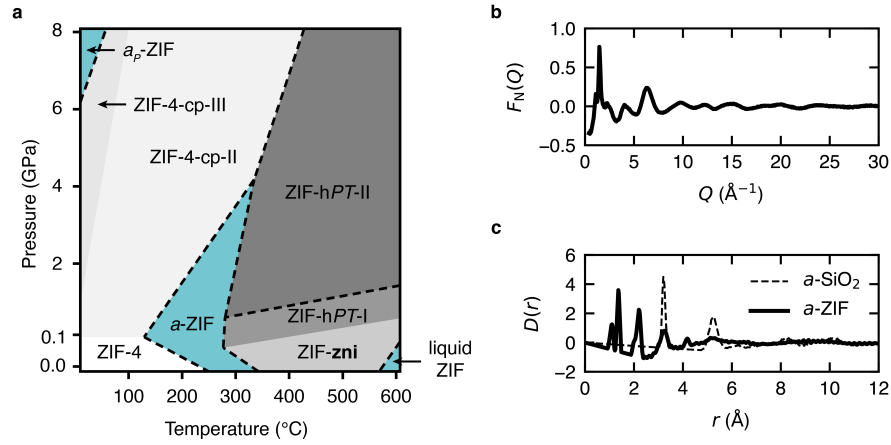


Figure 1.8: Diverse structural behaviour of ZIF-4. **a**, The pressure–temperature phase diagram of ZIF-4, as reported in Ref. 75. Different shades indicate different phases. *a*-ZIF forms are shown in blue. Dashed lines indicate irreversible, reconstructive transitions. The pressure range from 0 to 0.1 GPa has been magnified for better visibility and is therefore not to scale. Panels **b** and **c** show the neutron total scattering structure factor, $F_N(Q)$, and the corresponding neutron differential correlation function $D(r)$, respectively, for the temperature-induced *a*-ZIF form [76]. The $D(r)$ function for *a*-SiO₂ has been included (dashed line) for comparison, scaled in *r* by a factor of two to reflect the approximate difference in Zn···Zn and Si···Si distances.

NMR spectra were collected for ZIF glasses and the ZIF-4 and ZIF-**zni** polymorphs. Whereas both crystalline NMR line shapes necessitated simulation with at least two sites, corresponding to the fact that each of these crystals contain two crystallographically distinct Zn sites at a 1:1 ratio, the ZIF glass spectra were modelled with a single (average) environment, with a broader distribution and an asymmetric line shape with low-frequency tails, indicative of structural disorder in the glassy state [84]. The ZIF glass also shows small lowering of the isotropic chemical shift (≈ 20 parts per million (ppm)), compared to the crystalline ZIF-4 and ZIF-**zni** spectra [84]. A similar lowering in the isotropic chemical shift has been noted for ZIF-14 [$\text{Zn}(\text{eIm})_2$, eIm =2-ethylimidazolate] compared to ZIF-4 and ZIF-8 [82]. The difference was attributed to the longer Zn–N distances (2.00 to 2.02 Å) in ZIF-14, compared to ZIF-4 and ZIF-8 (1.98 to 1.99 Å). This could therefore imply an increase in the average Zn–N distance in the glass structure.

High field (20 T) ^{13}C and ^{15}N cross-polarisation (CP) MAS NMR has also been used to probe the structural differences between crystalline and amorphous ZIFs [83]. These spectra show broadening upon amorphisation and the coalescence of peaks, in accordance with the ^{67}Zn experiments described above, and the interpretation that *a*-ZIF is well represented by an average Zn environment and an average Im environment. For *a*-ZIF-8, two additional small signals were present in the ^{13}C spectrum that were not present in the crystalline spectra [83]. This was proposed to be indicative of de-coordinated linkers, introducing a small proportion ($< 10\%$ *) of defects into the framework.

1.1.4 ACC

Amorphous calcium carbonate (ACC) is the final amorphous material I will introduce in detail. Unlike the previous examples, ACC is unusual in that the metastable hydrated amorphous form precipitates directly out of solution. ACC has the nominal

*Quantitative analysis was not presented in Ref. 83. This approximate value is therefore obtained by visual inspection of the spectra.

composition of $\text{CaCO}_3 \cdot x \text{H}_2\text{O}$ ($x \approx 1$) [85]. This amorphous form can be stabilised for weeks by incorporating dopants such as Mg^{2+} or PO_4^{3-} [86, 87], or alternatively directed to crystallise into a number of different polymorphs by varying the pH or temperature [88]. This complex phase behaviour is exploited by nature in the context of biomineralisation processes, such as fashioning large-scale hierarchical morphologies [5, 89]. By way of example, in Fig.1.9(a) I show scanning electron microscopy (SEM) images of some stable biogenic ACC samples [89].

The top row shows samples taken from *Pyura pachydermatina*. The left image is a body spicule composed entirely of ACC. The right-hand image is the cross-section of a broken tunic spicule, which is composed of an external calcitic layer, separated by a membrane from an internal ACC core [86]. The bottom left image shows a cystolith from the leaves of *Ficus microcarpa*. Cystoliths are irregular-shaped objects, composed entirely of ACC, and are found in the epidermis of leaves from plants of various families [90]. The cystoliths are stable *in vivo*, but crystallise if extracted from the leaf in a humid environment. The bottom right image shows a granule from the storage structure of *Orchestia cavimana*, a terrestrial crustacean. The crustacean stores calcium in its mid-gut during the moult stage, in the form of concretions composed entirely of stable ACC. The higher solubility of ACC makes

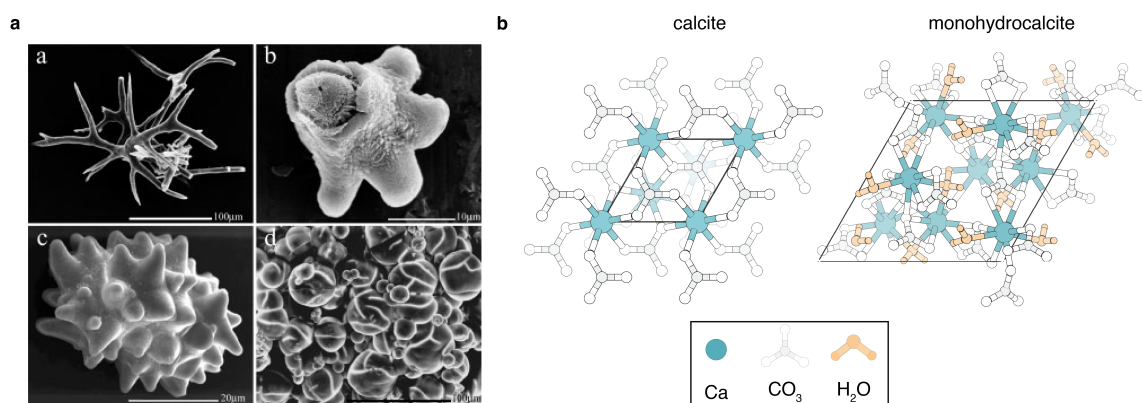


Figure 1.9: Structural diversity in (amorphous) calcium carbonate. **a**, SEM images of various skeletal parts composed of 'stable' ACC [89]. **b**, The crystal structures of calcite and monohydrocalcite. The figure in panel a is reproduced from Ref. 89 with permission from John Wiley and Sons.

it more easily available for mineralisation of new cuticles [91].

There are three anhydrous crystalline polymorphs of CaCO_3 : calcite, aragonite, and vaterite, in order of decreasing thermodynamic stability [92]. There are also two hydrated crystal phases: monohydrocalcite ($\text{CaCO}_3 \cdot \text{H}_2\text{O}$), and ikaite ($\text{CaCO}_3 \cdot 6 \text{H}_2\text{O}$). More recently, a hemi-hydrate phase has also been discovered with a stoichiometry of $\text{CaCO}_3 \cdot \frac{1}{2} \text{H}_2\text{O}$ [93, 94].

In seeking to understand how calcium carbonate supports such complex phase behaviour, including the stabilisation of the amorphous form and the controlled crystallisation to specific polymorphs, synthetic samples of additive-free ACC have been studied by a number of experimental techniques, including X-ray PDF analysis and NMR spectroscopy [95]. I show the Q -weighted X-ray total scattering structure factor, $QF_X(Q)$, in Fig. 1.10(a) [95]. The pattern shows only diffuse maxima and

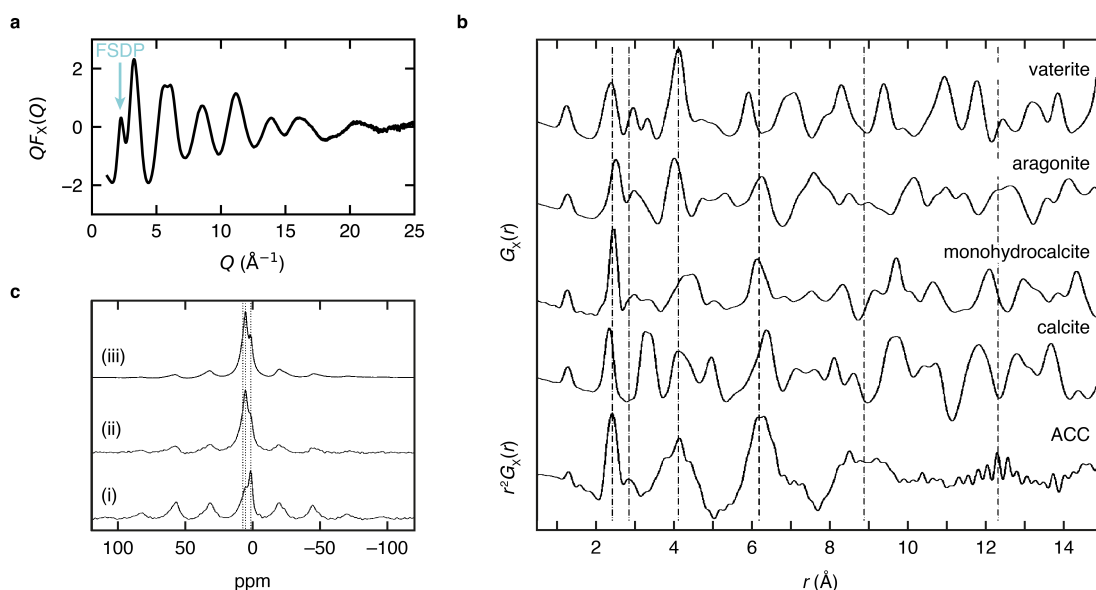


Figure 1.10: Experimental characterisation of ACC. **a**, The Q -weighted X-ray total scattering structure factor for a synthetic sample of ACC [95]. **b**, The effective X-ray PDF function, $G(r)$, for ACC obtained by Fourier transform of the $QF_X(Q)$ shown in panel **a**, compared with the PDFs of four crystalline CaCO_3 phases [95]. **c**, (i) and (ii): The HETCOR summed projections acquired with 1 ms contact time, displayed in absolute intensity mode, where (ii) includes a 50 ms mixing time. (iii) ^1H single pulse MAS NMR spectrum. The figures in panels **b** and **c** are adapted with permission from Ref. 95. Copyright 2008 American Chemical Society.

minima (i.e., contains no Bragg peaks), indicative of an absence of long-range order. There is, however, an observable first sharp diffraction peak (FSDP) at $Q \approx 2.2 \text{ \AA}^{-1}$, emphasised in Fig. 1.10(a) by a blue arrow. The presence of an FSDP usually suggests the presence of intermediate-range order, beyond the nearest-neighbour local environment [10, 96].

Figure 1.10(b) shows the effective ACC PDF function, $G(r)$, obtained by Fourier transform of the $F_X(Q)$ in Fig. 1.10(a), compared to $G(r)$ for the crystalline polymorphs calcite, monohydrocalcite, aragonite, and vaterite [95]. The sharpest peaks in the PDF occur at the low- r values, $r \approx 1.3$ and $2.4 \text{ \AA}$. These values correspond to the intramolecular C–O carbonate bond distance and the Ca–O bond distance, respectively. These distances remain unchanged from the analogous crystalline polymorph distances. The Ca–O bond distance is also supported by Ca K-edge extended X-ray absorption fine structure (EXAFS) analysis of ACC [95, 97–100]. Beyond these first few distinct features in the PDF arising from the local coordination environments, there are two broader and less well-defined features occurring at approximately 4.0 and $6.2 \text{ \AA}$. Correlation at these distances are also observed for the crystalline polymorphs. Beyond $\approx 10 \text{ \AA}$, the features in the ACC PDF are extremely small. Despite the aforementioned similarities, the short- and intermediate-range structure of ACC differs from that of each of these crystalline calcium carbonate phases.

A feature of the ACC structure that is not resolved by X-ray diffraction is the location and interactions of water within the CaCO_3 framework. $^{13}\text{C}\{^1\text{H}\}$ two-dimensional heteronuclear correlation (HETCOR) NMR spectroscopy has been used to probe the features of H atoms associated with carbonate in ACC [95]. There is only one characteristic carbon environment in the ACC ^{13}C NMR spectra ($\delta\text{C}=169 \text{ ppm}$, full-width half-maximum (FWHM)= 3.6 ppm) [95]. Hence, taking the ^1H sum projection (from δC 170.0 to 167.5 ppm) from the $^{13}\text{C}\{^1\text{H}\}$ HETCOR spectrum represents the ^1H spectrum arising from hydrogens in close proximity

to carbonate. Figure 1.10(c,i) shows the ^1H sum projection acquired with 1 ms contact time [95]. This projection contains two narrow peaks at 1.5 and 5.3 ppm, corresponding to hydroxyl groups and mobile molecular water, respectively. The projection also contains a broad spinning sideband (SSB) pattern centred at 6 ppm indicating the presence of rigid molecular H_2O .

The fractions of each of these components can be obtained by adding a mixing time (τ_{mix}) before the contact time in the $^{13}\text{C}\{^1\text{H}\}$ HETCOR experiment. This gives a quantitative, indirectly observed ^1H spectrum from the HETCOR experiment, where the ^1H intensity of each component is proportional to the population of corresponding species. The ^1H sum projection obtained with $\tau_{\text{mix}} = 50$ ms is shown in Fig. 1.10(c,ii). Least-squares fits to this spectrum show that the ratio of protons in hydroxyl groups, mobile molecular water, and rigid structural water is approximately 7:55:38 %. This suggests that rigid H_2O is coordinated to Ca atoms, and the H-bonding plays an important role in the formation and stabilisation of the ACC structure [95]. This is a feature of the monohydrocalcite structure [Fig. 1.9(b)]. Furthermore, the ^1H single pulse MAS spectrum for ACC show in Fig. 1.10(c,iii) is essentially the same pattern as the equilibrated ^1H HETCOR spectrum. This indicates that no significant water is present as a separate fluid phase.

The variable water content in ACC affects its density, and the orientational order of the carbonate species is a key indicator of how the material organises. For instance, in the anhydrous crystalline polymorphs calcite [Fig. 1.9(b, *left*)] and aragonite, all the carbonate molecules lie in the same plane. In monohydrocalcite [Fig. 1.9(b, *right*)], which contains one water molecule per formula unit, there are two alternating orientations of the carbonate molecules. In vaterite, the structure is still a subject of debate [101], however many possible ordering and stacking patterns for the carbonate molecules have been explored in an attempt to solve the structure [102–104]. In the context of ACC, it is not clear from the aforementioned experiments to what extent the carbonate molecules order in ACC.

1.2 Building structural models of amorphous materials

In the preceding section, I have provided a broad overview of our understanding of the structures of various amorphous materials, with a particular emphasis on insights gained from experimental data alone. In this section, I will explore different methodologies for constructing accurate structural models of amorphous materials.

This section will be structured as follows. In Section 1.2.1, I will introduce the main sampling techniques used in atomistic modelling; namely, molecular dynamics (MD) and Monte Carlo (MC) methods. In Section 1.2.2, I will describe three methods for modelling the interactions between atoms; namely, empirical force fields, quantum mechanical methods, and machine-learned (ML) potentials. In Section 1.2.3, I will describe the ‘direct approach’, which typically uses a description of the interatomic interactions to generate structural models. I will conclude in Section 1.2.4 by describing the ‘inverse approach’, whereby the structural models are generated to achieve consistency with a set of experimental data.

1.2.1 Sampling techniques

Methods for generating amorphous models typically require *simulations*, which are aimed at generating a representative sampling of a system at a finite temperature [105, 106]. There are two primary techniques for achieving such sampling: MC and MD. The choice of sampling technique depends on the property to be studied.

In the context of modelling amorphous materials, both methods can be useful. For example, the configurational energy of a disordered system can be calculated by solving the coupled equations of motion of the atoms and averaging over time (i.e., MD simulations). Alternatively, by first setting up an ensemble of states of the system, where the occurrence of each state is determined by the appropriate probability, the configurational energy can be obtained by averaging uniformly over the ensemble (i.e., MC simulations).

MC simulations

In MC methods, a sequence of states is generated from an initial state by randomly perturbing the coordinates of randomly chosen particles [107]. In the Markov chain variant of MC, a move to a new configuration, ν , from initial configuration μ , is accepted or rejected according to the Metropolis criterion [108]:

$$P(\mu \rightarrow \nu) = \begin{cases} \exp(-\Delta E/k_{\text{B}}T), & \text{if } \Delta E > 0 \\ 1, & \text{otherwise} \end{cases} \quad (1.1)$$

Here, ΔE is the change in energy, T is the simulation temperature, and k_{B} is the Boltzmann constant. The Metropolis criterion ensures that the configurational ensemble obeys a Boltzmann distribution, and the possibility of accepting moves with associated increases in energy allows for the escape from local minima. These simulations typically only require evaluating the system's energy, rather than the forces, which can be advantageous if computing derivatives is computationally expensive.

MC simulations do not require the steps to be physically sensible, enabling them to ‘tunnel’ between states with high energy barriers, which can often lead to better coverage of the configuration space compared to MD [109]. Accordingly, MC methods are often used in cases where configurations only evolve dynamically at an extremely slow rate, for example cation site ordering [110] and topological rearrangements [111, 112]. However, this also means that MC simulations are not suitable for studying the dynamics of a system [109].

It is also possible to constrain certain degrees of freedom, further improving convergence. For example, in bond-switching simulations for *a*-Si [Section 1.2.3], one might impose a constraint to reject moves that generate strained three- and four-membered rings [112]. Recently, a more sophisticated approach was taken whereby two-dimensional networks were constructed by reference to a pre-defined distribution of ring sizes, allowing network topologies to be varied systematically [113].

Molecular dynamics simulations

MD methods generate a series of time-correlated points in phase space by propagating an initial set of coordinates and velocities through finite time steps, following Newton’s second law of motion [105]. A crucial ingredient determining the success of the method is the description of the potential energy surface (PES), the options for which I will describe in Section 1.2.2.

The result of an MD simulation is the evolution of the positions and velocities of an ensemble of atoms through time, typically representing the NVE ensemble (constant number of particles, volume, and energy). To compare with experimental systems, dynamic constraints such as a thermostat or barostat are applied. These constraints create the NVT ensemble (constant number of particles, volume, and temperature) [114, 115] or the NPT ensemble (constant number of particles, pressure, and temperature) [116], ensuring that the average temperature (and pressure for NPT) converges to a given value over time.

MD can be used to compute a variety of properties of a system. First, it is possible to compute equations of state for different temperatures and pressures, and simulate phase-transitions [109]. Second, distribution functions can be computed, such as the distribution of positions or orientations of atoms and molecules, which is particularly useful for analysis of the structure of disordered materials [105]. Third, time-correlation functions can be computed to analyse the dynamics of a system, such as diffusion, thermal conductivity, and phonon density of states [105].

1.2.2 Modelling atomic interactions

Accurately computing the total energy of a system and the interatomic forces is essential to probe the quality of a model for an amorphous material [4]. The interatomic potential arises from the chemical bonding between the atoms in a system. The chemical bond is a spatially non-local entity, involving multi-body forces originating from complicated electronic effects, the details of which depend sensitively on

the local topology. This poses a challenge for defining the interatomic potential for amorphous materials: the presence of disorder implies a wide range of environments, compared to crystalline structures. All of these environments must be well described.

The most reliable approach for modelling the interatomic potential is to determine the electronic structure of a material using quantum-mechanical calculations. However, routine application of these methods is limited to a few hundred atoms, far less than typically required for adequate models of amorphous materials. Empirical approaches are often used instead, where a fixed mathematical form is assumed, and the parameters adjusted to ensure known properties (bond lengths, bond angles, melting points, bulk modulus) in reference systems were reproduced. Empirical potentials therefore tend to be reliable for structures similar to those included in the fitting database, but most likely unreliable for ‘unseen’ structures. This is the concept of *transferability*. It is in this context I will introduce ML potentials. Like empirical potentials, ML potentials rely on a training set of configurations to which they are fitted to. However, ML potentials typically assume very flexible forms with very large numbers of parameters, meaning that they are capable of describing a very large number of diverse atomic environments to an accuracy comparable with quantum mechanical calculations.

In the following, I briefly review some of the approaches to approximating the interatomic potential.

Empirical force fields

Empirical force fields express the electronic energy of a system as a parametric function of the positions of particles, which may be atoms or molecules, by fitting the parameters to experimental or higher-level computational data. They can be routinely used to perform MD on systems containing thousands of atoms and for simulation times on the scale of tens to hundreds of nanoseconds.

One simple model that has worked well for ionic materials, such as SiO_2 , combines a Coulomb term with a covalent, short-range contribution of the (Buckingham) form

$$\Phi_{ij}^{\text{Bh}}(r) = A_{ij} \exp(-B_{ij}r_{ij}) - \frac{C_{ij}}{r_{ij}^6}, \quad (1.2)$$

where r_{ij} is the interatomic spacing, and A_{ij} , B_{ij} , and C_{ij} are parameters to be tuned against some data [117–120]. One such parameterisation, named the ‘CHIK’ potential [120], was obtained by matching the partial pair correlation functions, as obtained from the effective potential, with those obtained from an *ab initio* MD (AIMD) simulation. In the case of silica, the success of such a simple pairwise form is likely attributed to the fact that the basic unit in the crystal, liquid, and glass is the same: the SiO_4 tetrahedron. The ionicity essentially forbids ‘wrong’ homopolar bonds from forming.

Pair potentials such as those in Equation 1.2 are also suitable for modelling interactions between metal ions and molecular units. The empirical potentials developed for (hydrated) metal carbonates are an excellent example [121–124]. The thermodynamic landscape of calcium carbonate is particularly delicately balanced and difficult to describe theoretically. For instance, although aragonite has a lower enthalpy than calcite, the entropic contributions make calcite more stable under ambient conditions [125, 126]. By expressing the calcium–carbonate interactions in a Buckingham form [Equation 1.2], the associated parameters were refined to recover structural and mechanical properties of calcite and aragonite, as well as the energy difference between the two phases. I show the Buckingham pair potential terms used in this force field in Fig. 1.11(a). The force field was then extended to describe the aqueous calcium carbonate system. The carbonate–water interactions were expressed in a Buckingham form, while the calcium–water interactions were expressed in a Lennard-Jones potential form:

$$\Phi_{ij}^{\text{LJ}}(r) = 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^6 \right], \quad (1.3)$$

where ϵ_{ij} is the depth of the pair potential well and σ_{ij} is zero-crossing distance of the pair potential. The water–water interactions were described by the TIP4P-Ew force field [127]. Briefly, this is a four-site rigid water model where an additional, massless site is placed at a fixed distance, d , from the oxygen atom along the bisector of the H–O–H bond angle [Fig. 1.11(b)]. This additional site carries the charge associated with the oxygen atom. The bond lengths and bond angles are set to the values for the experimental gas-phase geometry of the water monomer ($r_{\text{O-H}} = 0.9572$ Å and $\theta_{\text{H-O-H}} = 104.52^\circ$) [128]. The mineral–water interaction parameters were refined using the hydration free energy of the ions and the dissolution enthalpy of the mineral as the objective quantities, because these are the most important properties for studying nucleation and growth of CaCO_3 [121, 122]. The parameterised Lennard–Jones terms of this force field are shown in Fig. 1.11(c). The force field was able to reproduce the free energies of the component ions in solution, as well as the energy difference for the calcite-aragonite phase transition, and was used to model calcite surface growth.

This force field was later re-fitted to include harmonic terms for bond, bond angle, and out-of-plane distortion of the carbonate group [122]. For consistency, the flexible SPC/Fw water model [129] was used in place of the rigid TIP4P-Ew model,

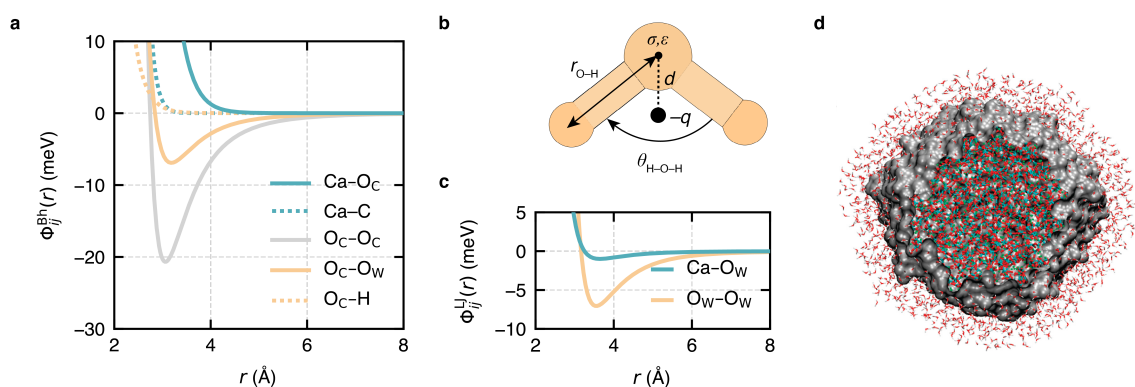


Figure 1.11: Aqueous calcium carbonate empirical potentials. **a**, Buckingham terms [Equation 1.2] in the aqueous CaCO_3 force field [121]. **b**, Depiction of a TIP4P-Ew water molecule [127]. **c**, Lennard-Jones terms [Equation 1.3] in the aqueous CaCO_3 force field [121]. **d**, An ACC nanoparticle simulated using the force field reported in Ref. 122. The figure in panel **d** is reproduced with permission from Ref. 122. Copyright 2010 American Chemical Society.

and the calcium–water and carbonate-water interactions were modified to ensure that the correct free energies of solvation were retained. This revised force field was used in MD simulations to study the stability and growth thermodynamics of ACC [122, 130]. To create the nanoparticles, crystallites of ordered calcite were annealed at 3,000 K to ensure no residual structural order remains. This amorphous particle was then inserted into a simulation cell of water. Water molecules were then removed based on the degree of overlap with the calcium carbonate atoms. By successively excluding water molecules with shorter contact distances, nanoparticles were created with different degrees of hydration. The nanoparticles were then equilibrated.

An example of a simulated solvated ACC nanoparticle is shown in Fig. 1.11(d) [122]. The nanoparticle contains 864 formula units of CaCO_3 and has an approximate diameter of 5.5 nm. The local structure of the nanoparticle is consistent with experimental bond length distributions and coordination numbers [Section 1.1.4] [95]. Importantly, the Ca-pair correlation function, $g_{\text{Ca}}(r)$, computed for the nanocluster shows two main peaks at the expected distances (4 and 6 Å), in agreement with the experimental PDF [Section 1.10(b)] [95]. Interpretation of any correlations beyond these distances are not considered because the presence of the nanoparticle surface begins to contribute to the analysis.

In contrast to the success of these pair potentials for modelling a multi-component system such as ACC, and perhaps counter-intuitively, modelling the interactions in elemental systems such as C and Si is more challenging [4]. This shortcoming is because pair potentials do not account for the directionality of covalent bonding. This is especially true of C for which the sp^3 bonds (as for diamond) and sp^2 bonds (as for graphite) are nearly energetically degenerate [131]. Instead, the approach typically taken is to describe bonds and angles explicitly [Fig. 1.12(a)]. In this way, it is possible to construct accurate potentials for single topologies and small variations about their equilibrium positions. The Keating potential is one such example for Si [132].

A general functional form for the total energy [Equation 1.4] is typically decomposed into bonded terms for interactions that are linked by covalent bonds [Equation 1.5], and non-bonded terms that describe long-range electrostatic and van der Waals forces [Equation 1.6].

$$\Phi_{\text{total}} = \Phi_{\text{bonded}} + \Phi_{\text{nonbonded}} \quad (1.4)$$

$$\Phi_{\text{bonded}} = \Phi_{\text{bond}} + \Phi_{\text{angle}} + \Phi_{\text{dihedral}} \quad (1.5)$$

$$\Phi_{\text{non-bonded}} = \Phi_{\text{Coulomb}} + \Phi_{\text{van der Waals}} \quad (1.6)$$

A variety of force fields exist in this form, initially developed for small organic molecules and later adapted for complex functionalised molecules [135]. Additionally, general force fields have been created to cover all atoms in the periodic table [136], biomolecules [137], and condensed matter [138]. These force fields require parameters for every atom type. For example, an empirical potential named ‘MOF-FF for ZIFs’ [133] describes $\text{Zn}(\text{Im})_2$ systems with 6 atom types, 6 bond types, 9 angle types, and 16 dihedral types, alongside the Coulomb and van der Waals terms

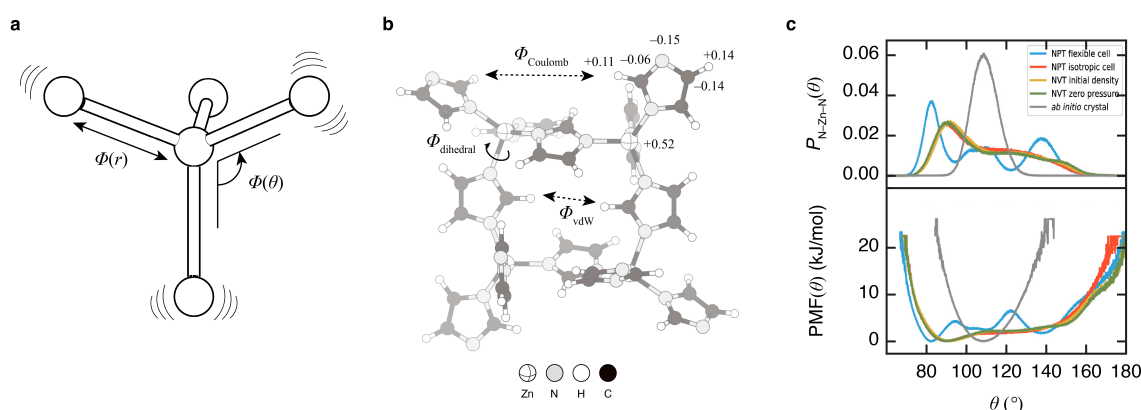


Figure 1.12: Bond length and bond angle terms in empirical potentials. **a**, Schematic of bond length and bond angle terms in a tetrahedral network, such as those used in the Keating potential [132]. **b**, Illustration of the different terms commonly used in more complex empirical potentials, such as MOF-FF for ZIFs [133]. **c**, The bond angle distributions (*top*) and associated potential of mean force (*bottom*) for ZIF glasses produced with the *ReaxFF* force field [134]. The figure in panel **c** is adapted with permission from Ref. 134. Copyright 2022 American Chemical Society.

[Fig. 1.12(b)], with all parameters obtained by fitting to *ab initio* calculations. The force field accurately captures dynamic properties of the flexible ZIF frameworks, including the ‘swing effect’ which allows molecules larger than the geometric window size to diffuse into the framework by cooperative rotation of the Im linkers [139]. Such force fields are more challenging to apply in the context of amorphous modelling however, because they do not permit bonds to be broken and formed.

To simulate bond breaking and forming, reactive force fields are required, such as *ReaxFF* [140]. *ReaxFF* modifies energy terms between bonded atoms to achieve a continuous description of the PES that connects end-points and intermediate structures during a reaction, while ensuring that the bond-stretching energy approaches the dissociating energy as the bond distance tends to infinite separation [141]. All other terms, such as bending and torsional terms [Fig. 1.12(a,b)], must go to zero as the bond-stretching term is gradually replaced by non-bonded electrostatic and dispersion interactions.

A parameterisation of *ReaxFF* for ZIFs has been used to simulate *a*-ZIF structures [134, 142–145]. These simulations generally use a melt-quench protocol [Section 1.2.3] starting from supercells of crystalline ZIFs, in order to generate ZIF glass models. In general, these glass models all share a few key features. First, the long-range order of the crystalline structure is lost. Second, there is an increase in density upon vitrification. Third, the ring size distributions broaden significantly and shift to larger ring sizes compared to the crystalline polymorphs. However, a recent study has indicated that these simulations are highly-sensitive to the employed simulation protocols and parameters, resulting in configurations that differ substantially from those obtained using higher levels of theory (to be discussed in the following section) [80, 134]. Notably, the *ReaxFF* force field specifically under-constrains the directionality of the Zn–N coordination [Fig. 1.12(c)], which likely affects the properties and structural rearrangement in the amorphous form. This is reflected in a longer average Zn–N bond length and an N–Zn–N angle distribution

that bears little resemblance to any crystalline polymorph [Fig. 1.12(c)] [134].

The issues raised here can be considered as general challenges one faces when attempting to develop interatomic potentials, with relatively simple functional forms, that are also suitable for describing a diverse range of atomic environments as encountered in disordered materials.

Quantum mechanical methods

Quantum mechanical methods are a widely used alternative to empirical force fields, because they, in general, provide more accurate calculations without the need for prior parameterisation [109]. This accuracy is achieved by considering the electronic structure of the material to determine the energy (and derived properties) [146].

In quantum simulations of electronic ground-state properties, the wavefunction has to be found so that it satisfies the Schrödinger equation to give the energy of the system [146]. This is a complex task owing to the many-body nature of the interactions between electrons. Furthermore, the large number of atoms needed to adequately describe amorphous systems means directly tackling the many-electron problem is intractable. Instead, a mapping of the many-body problem into an effective one-electron problem is used. This is the approach taken in density function theory (DFT), the mostly widely used *ab initio* method in quantum chemistry applications [147, 148].

DFT determines the ground state energy (and derived properties) of a system by solving for the electron density [149, 150]. To achieve this, a system of non-interacting electrons that produce the same electron density as the real system is considered [150]. This approach invokes approximations, predominantly associated with correcting for the error in approximating the kinetic energy as that of non-interacting electrons, and electron-electron interactions as classical interactions. The accuracy of DFT relies on how these ‘corrections’ are accounted for. There is a ladder [Fig. 1.13] [151] of well-established approaches (i.e., DFT functionals) which increase in complexity (and computational cost) in order to achieve higher accuracy.

The local density approximation (LDA) functionals are the simplest and most computationally efficient option. LDA functionals depend solely on the electron density at each point in space, without considering derivatives of the electron density, which typically leads to an overestimation of bond strengths [146]. The next ‘rung’ of the ladder is the generalised gradient approximation (GGA) functionals [152, 153]. GGA functionals explicitly depend on the electron density and its gradient, thus accounting for non-local exchange and correlation effects. Meta-generalised gradient approximation (meta-GGA) functionals further improve upon GGA functionals by considering the local kinetic energy density, allowing meta-GGAs to more accurately treat various chemical bonds (e.g., covalent, metallic) and intermediate-range interactions compared to LDAs and GGAs [154].

It should be noted that DFT often poorly describes long-range interactions, such as dispersion interactions [154]. To address this, dispersion corrections are typically added [155]. Several dispersion correction methods are available, generally capturing long-range effects using a relatively simple function of interatomic distances with adjustable parameters fitted to interaction and conformational energies computed using high-level wave function methods [155]. Dispersion corrections are often essential, especially when using GGA functionals, as long-range interactions can strongly influence energetic and mechanical properties, particularly for porous structures such as MOFs.

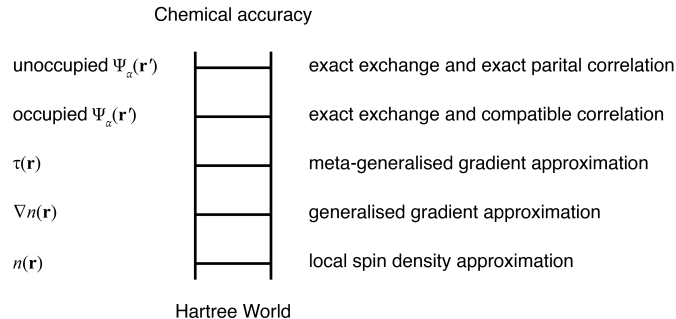


Figure 1.13: The Perdew ladder of density functional approximations. As approximations are lifted and the functionals satisfy more exact conditions (i.e., ascending the ladder), the chemical accuracy of the DFT method generally increases [151].

In practice, it is not always *a priori* clear which functional will be best suited for a given material, and the choice will be often guided by a compromise between accuracy and computational cost [146]. Irrespective of the chosen functional, the variational principle is applied to optimise the electron density self-consistently to minimise the system’s energy, thus enabling the determination of ground state properties and facilitating the study of any material, given only the coordinates and element of each atom, provided a reliable exchange–correlation functional for the system of interest exists [146].

One particular method for modelling amorphous materials with DFT that is commonly used is AIMD [156, 157]. AIMD allows for probing of dynamical properties in situations where sub-nanosecond dynamics are relevant. Briefly, the method works by computing the ground-state electronic structure using DFT, which is then used to perform MD where the ionic motion is described classically, and the ionic and electronic motions are separated according to the Born–Oppenheimer approximation [158]. This can be particularly effective for small-scale modelling of amorphous materials.

AIMD has been used to model the melting of ZIFs, as well as their liquid and glass phases [80, 159–161]. Owing to the high computational cost of AIMD, these melt-quench simulations were performed on primitive cells (272 atoms for ZIF-4). Multiple trajectories (e.g., ten for the ZIF glass models [161]) were therefore required to achieve good statistical sampling. Nevertheless, these models yielded the first reliable description of the melting mechanism in ZIFs, and atomistic insight into the structure of a MOF liquid. The mechanism by which the framework degrades is visualised in Fig. 1.14(a). First, an imidazolate linker de-coordinates and moves away from a fourfold-connected zinc, leaving an under-coordinated zinc site. Second, a neighbouring imidazolate linker with an available nitrogen coordination site binds and restores the fourfold-connectivity of the zinc site [80]. The evolution of the structure during melting is compared to experiment by computing the X-

ray total scattering structure factor, $F_X(Q)$, from models at different temperatures. The experimental and computed $F_X(Q)$ are shown in Fig. 1.14(b) *top* and *bottom*, respectively. A qualitatively similar trend is observed. Upon heating from 304 to 778 K, both the intensity and position of the FSDP, centred at $1.1 \text{ \AA}^{-1}$, remained approximately constant. However, further heating to 856 K resulted in a more pronounced intensity reduction and a shift in the position of the FSDP to $1.3 \text{ \AA}^{-1}$, along with a near-total disappearance of any features at higher Q values. That the computed structure factor is shifted to smaller Q values than the experiment is most likely due to performing the AIMD melt-quench simulation in the NVT ensemble at a density lower than the experimental glass and liquid densities.

In real-space, the most significant change is associated with the broadening of the $g_{\text{Zn}}(r)$ pair correlation function [Fig. 1.14(c)]. Notably, this peak remains centred at $\approx 6 \text{ \AA}$ in the crystal, liquid, and glass forms. Correlations much beyond $6 \text{ \AA}$ cannot be obtained from the AIMD simulations because of the small unit cells considered.

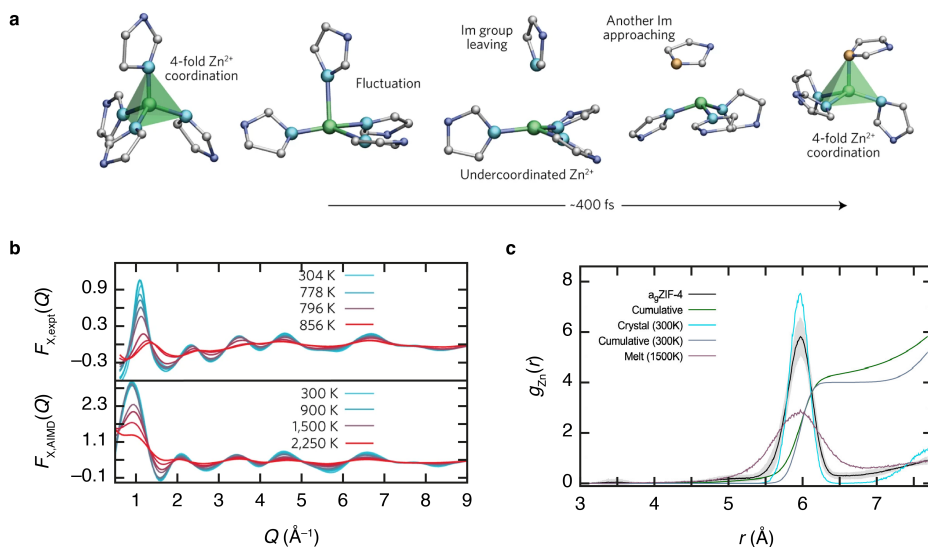


Figure 1.14: Liquid and glassy ZIF structural models from AIMD. **a**, Visualisation of a representative imidazolate exchange event during the melting stage [80]. **b**, Experimental glass (*top*) and computational ZIF-4 (*bottom*) X-ray total scattering structure factors upon heating [80]. **c**, Comparing the $g_{\text{Zn}}(r)$ pair correlation functions for the crystal, melt, and glass forms of ZIF-4 [80]. The figure in panel **a** is reproduced with permission from Ref. 80. The figure in panel **b** is adapted with permission from Ref. 80. The figure in panel **c** is adapted with permission from Ref. 161. Copyright 2020 American Chemical Society.

This is the primary disadvantage of the approach. Often, larger configurations and longer timescales are required than possible with AIMD.

ML potentials

ML potentials are a popular alternative to DFT methods [162–165]. They aim to make the same type of predictions using an approximate or surrogate model, while requiring considerably smaller computational costs [164, 166].

ML potentials accelerate atomistic simulations when trained on quantum-mechanical energies and forces. An overview of the method is shown in Fig. 1.15. The task is formulated as follows. First, like empirical potentials, the system energy is partitioned into a sum of local atomic contributions from each atom i [168, 169]:

$$\Phi_{\text{total}} = \sum_{i=1}^N \epsilon_i. \quad (1.7)$$

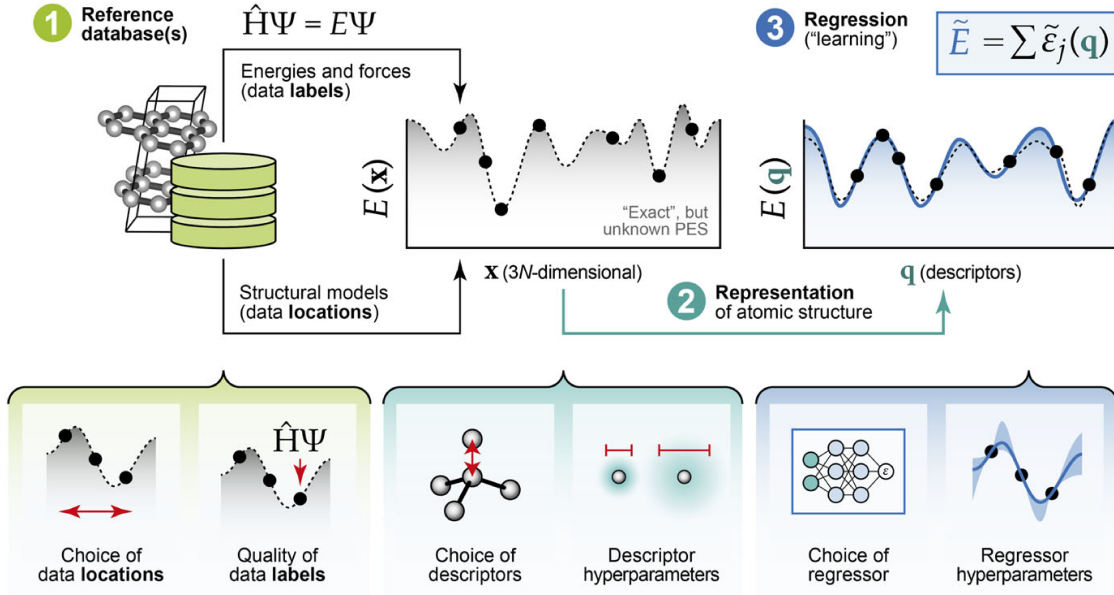


Figure 1.15: Three key components in ML potential fitting. This figure is adapted from Ref. 167. Copyright 2023 Author(s) licensed under a Creative Commons Attribution 4.0 License.

ML potential approximates the form of the PES based on reference data, typically obtained from small-scale DFT computations [166]:

$$\{\mathbf{x}_i, y_i\}_{i=1..N} \longrightarrow \tilde{y}(\mathbf{x}), \quad (1.8)$$

where $\mathbf{x}_i$ is an atomic structure, or ‘data location’ [170], y_i is one of N observations (e.g., energies or forces), and $\tilde{y}$ denotes the ML model fitted with a given method. In this way, given a new data location, $\mathbf{x}$, the ML model can predict the associated observation(s). There are three main components that emerge from Equation 1.8, and which are illustrated in Fig. 1.15 [166, 167].

The first component is the reference dataset of atomic structures to which this fit is made [Fig. 1.15(*left*)] [171]. The goal of fitting an ML potential is to sample the PES for the system of interest, which means it needs to reproduce the system energy, atomic forces, and stress tensor. The training dataset therefore must be ‘complete’ with respect to the configurational domain of application, which I define as the (hypothetical) set of all configurations that will be encountered in the subsequent simulations run with the ML potential.

There are many approaches to building these training datasets. It is common to begin with extracting relevant structures from online repositories, including the QM9 database [172], Inorganic Crystal Structure Database [173], and Cambridge Structural Database (CSD) [174]. These structures alone will not suffice for building a reliable ML potential. A typical approach is to supplement them with structurally-distorted variants, MD trajectories performed under different conditions, and random structure searching. More recently, systematic strategies have been proposed to reduce the need for human input (i.e., ‘domain knowledge’), instead relying on uncertainty quantification by training ML potentials iteratively [175, 176], or to bias the models towards particular regions of phase space [177].

The second component is the representation of local environments of the atoms in the dataset in a mathematical form, i.e., formulating $\mathbf{q}$ [Fig. 1.15(*middle*)] [168,

178, 179]. To this end, there are a wide variety of available descriptors that are suitable. The descriptors should provide a comprehensive description of the atomic arrangements and correlate well with target properties; namely, they should respect the smoothness and symmetries of the PES. Density-based descriptors are one such class that use an atom-centred representation of a system. Examples of this class include the smooth overlap of atomic positions (SOAP) [180], the Behler-Parrinello symmetry functions [168], and the atomic cluster expansion (ACE) [178].

The third component is the fitting of a flexible (highly parameterised) function to the reference data [Fig. 1.15(*right*)]. There are many algorithms for achieving this [181–183]. Irrespective of the specific choice of algorithm, they usually have a set of hyperparameters that need to be optimised in order to ensure the ML potential not only accurately describes structures it was specifically trained on, but also generalises to new structures.

Importantly, when these three components are satisfied, ML potentials ‘unlock’ simulations that were not possible using empirical potentials or DFT. I will give three examples to illustrate this point.

First, ML potentials have provided atomistic insight into complex structural transitions in *a*-Si—a notoriously difficult system to study computationally [see later; Section 1.2.3] [184–186]. Describing both the liquid–amorphous and amorphous–amorphous transitions required an accurate model spanning the full range of diverse atomic environments that was scalable to system sizes of 10 nm length scales [185].

Second, ML potentials have been used to investigate nanoscopic mixtures of *a*-Si and *a*-SiO₂, which are not amenable to a DFT description due to the system sizes (tens of nanometres) required to describe the structural heterogeneity [60, 187]. This heterogeneity presents a challenge—not limited to this particular material—concerning the development of datasets for disordered materials. Ideally, when probing defective sites, finite size effects should be mitigated to ensure a good description of the PES. It can be hard to efficiently sample defective sites in this way

because they are typically rare events. The approach taken for the Si–O system was to detect the onset of defective sites using uncertainty quantification measures [175], and embed them in an annealed ‘amorphous matrix’, which was then labelled and added to the training dataset [187]. In this way, efficiently obtaining high-quality labels of disorder for training ML potentials is possible.

Third, ML potentials have been developed to describe hybrid materials such as MOFs [188, 189]. Not only are these potentials capable of describing the breadth of atomic environments encountered in these flexible materials, but they are also capable of modelling interactions between the framework and guest species such as adsorbed gases, enabling quantitative predictions for diffusion and adsorption properties [190, 191]. ML potentials for disordered ZIFs have also emerged in the last year [Table 1.1] [192, 193]. These ML potentials used deep neural networks [195–197] and were trained on AIMD trajectories of ZIFs under various temperatures and pressures to generate datasets suitable for studying the amorphous forms.

DeepMD uses a deep neural network to learn the interatomic potential from quantum mechanical calculations [198]. DeepMD respects the essential physical symmetries of atomic systems by using carefully engineered input features that capture relative atomic positions within local neighbourhoods using scalar quantities (e.g., distances and angles), and through carefully chosen neural network architectural design. As a result, DeepMD is well-suited for describing a wide range of atomic systems with high efficiency and scalability. However, to achieve accurate and reliable predictions, it requires a substantial amount of training data because the model must learn complex interactions and capture diverse atomic environments using relatively

Table 1.1: Details of different ZIF ML potentials in the literature.

ML architecture	Training data	Database size		Ref.
		Cells	Atoms	
DeePMD	AIMD	>200,000	>54,400,000	193
DeePMD	AIMD + active learning	39,512	8,708,628	194
NequIP/Allegro	AIMD	2,375	646,000	192

simple descriptors. This is evidenced by the very large number of structures required to train the ZIF ML potentials shown in Table 1.1.

The Neural Equivariant Interatomic Potential (**NequIP**) framework takes a different approach by using equivariant atom-centred message passing to encode the atomic structure [195]. In this method, an atomistic graph is constructed by connecting each atom to its neighbours within a finite cut-off sphere. Information is then iteratively propagated along this graph, enabling the model to learn many-body correlations and access non-local information beyond the local cut-off. The use of equivariant convolutions for interactions of geometric tensors creates a more information-rich representation of atomic environments, leading to a significantly more data-efficient method—achieving higher accuracy and better performance in MD simulations compared to models that use invariant convolutions and operate only on scalars. This is evidenced in Table 1.1 by the significantly smaller dataset required for training the **NequIP** ZIF ML potential. However, the iterative propagation increases the effective number of neighbours for each atom, which complicates parallel computation and limits the accessible length scales in MD simulations [197]. To address this, **Allegro** was developed to improve scalability with system size without compromising the accuracy and data efficiency [197]. It achieves this by employing a strictly local equivariant deep neural network architecture that represents many-body potentials through iterated tensor products of learned equivariant representations, eliminating the need for atom-centred message passing.

In the following, I describe the key findings reported for each of the three ZIF ML potentials described in Table 1.1.

The **DeePMD** model reported in Ref. 193 was used to simulate the melting behaviour of a ZIF-4 supercell containing 7,344 atoms in a box with ~ 5 nm side lengths. The MD simulation was performed in the NPT ensemble between temperatures of 300 and 1,700 K. It showed an initial decrease in density during the early heating stages due to greater thermal fluctuations of the network, followed

by a sharp density increase at $\sim 1,700$ K. The transition occurred through the formation of small, heterogeneously distributed high-density domains, rather than a uniform increase in density across the structure. Beyond 1,700 K, non-physical features emerged, such as clusters of H atoms. Notably, the crystalline and glass densities of ZIF-4 predicted by the DeePMD model are significantly lower than those reported experimentally [66, 80].

The second DeePMD model reported in Ref. 194 was employed to perform a melt-quench simulation of ZIF-4 across a range of pressures from 0 to 1 GPa. At zero pressure, similar to the previous DeePMD model, the resulting amorphous structures exhibited significantly lower densities than those observed experimentally. However, applying higher simulation pressures brought the densities into the range of 1.2–1.6 g cm $^{-3}$, aligning more closely with experimental values [76, 80]. The higher pressure simulations also provided a better match between the normalised neutron total scattering function, $S_N(Q)^*$, computed from the models and that measured experimentally for ZIF-4 glass [Fig. 1.16(a)] [80]. While the positions of the peaks were

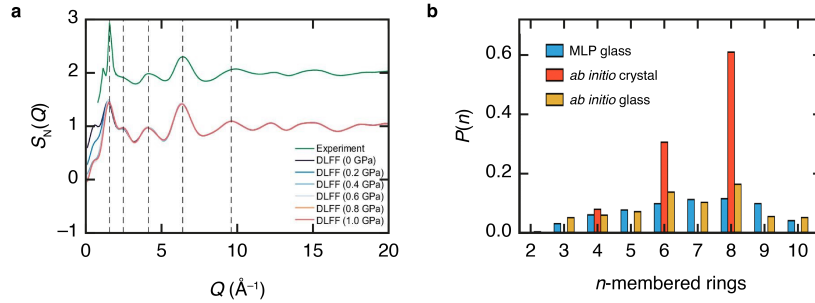


Figure 1.16: Modelling disordered ZIFs with ML potentials. **a**, Comparing the normalised neutron total scattering functions, $S_N(Q)$, computed from melt-quenched ZIF glass structures generated with a DeePMD ML potential reported in Ref. 194 at variable pressures, with the experimental function [80]. **b**, Comparing ring size distributions for ZIF glasses generated using AIMD [161] and the NequIP ML potential reported in Ref. 192. The crystalline ZIF-4 ring size distribution is shown in red. The figure in panel **a** is adapted with permission from Ref. 194. The figure in panel **b** is adapted with permission from Ref. 192.

*The *normalised* total scattering function, $S(Q)$, is derived from $F(Q)$ by removing the scaling factors associated with species concentrations and scattering lengths [199]: $S(Q) - 1 = F(Q) / (\sum_{i=1}^n c_i \bar{b}_i)^2$, where c_i represents the proportion of species i in the material, and $\bar{b}_i$ denotes the averaged neutron scattering length of species i .

in agreement, the relative shapes and intensities differed between the experimental data and the simulations, particularly at low Q ($< 1.55 \text{ \AA}^{-1}$) values.

The **NequIP** and **Allegro** models reported in Ref. 192 were used to simulate the ZIF-4 glass, starting from $2 \times 2 \times 2$ supercells of ZIF-4 (comprising 2,176 atoms) and following a melt-quench protocol. These models accurately reproduced structural properties, such as radial distribution functions and bond-angle distributions, that were previously obtained using AIMD, but at a fraction of the computational cost and with larger system sizes. Importantly, the models were large enough to study the framework topology of the disordered ZIF forms [Fig. 1.16(b)]. The ring size distribution was found to be broader compared to that of crystalline ZIF-4 and extended to slightly larger ring sizes.

1.2.3 The direct approach to model building

Hand-built models

While hand-built models may seem distant from the techniques used in this thesis, and indeed more generally in the field today, their use has been crucial in understanding the structures of amorphous tetrahedral materials. These models were originally developed for α -SiO₂ [200], and then later for α -Si [201], and provided foundational insights that paved the way for the development of more advanced computational methods and simulations.

The idea was simple. Wire models were constructed by a process representing the condensation from a fluid. Starting from a seed, ‘tetrapod’ (Greek: τετραπούς, meaning ‘four feet’) units were sequentially connected together by plastic tubes, where both moderate bending of the ‘bonds’ and rotation of tetrapods about the bonds were possible. In this way, topological disorder was introduced by the random relative orientation of tetrapods, while the fixed coordination number and small permitted departures from the preferred bond length and angles ensured physically-sensible solutions. The first physical model of α -SiO₂ (1964) built in this man-

ner is shown in Fig. 1.17(a) [200]. Based on this original model-building strategy, two new models of α -SiO₂ were generated in close succession [202, 203]. In one model [Fig. 1.17(b)], improvements were made by using bent springs for connecting tetrapods, and assigning bond angles according to a random number computer programme sampling from a Gaussian distribution centred at 163° [202]. The second model [Fig. 1.17(c)] similarly used random assignments of Si–O–Si angles [203]. The atomic coordinates were extracted from the model and used to compute the PDF, which gave a reasonable match to the experimental function [203].

The first amorphous model for α -Si was developed by Polk in 1971 [201]. Starting with a core of five- and six-membered rings, additional units were added to match the crystalline centre-to-centre distances. A local ‘minimisation’ procedure was applied, selecting sites requiring the least strain when filling unsatisfied (dangling) bonds. As with the α -SiO₂ models, the PDF computed from this model showed remarkable agreement with experiment [201, 204].

These models laid the groundwork for computer-based CRN modelling. Initially, computers were used to refine hand-built models [205–207] using simple empirical potentials like the Keating potential [132]. Eventually, model building became predominantly computational, though aspects of the earlier algorithms influenced subsequent approaches.

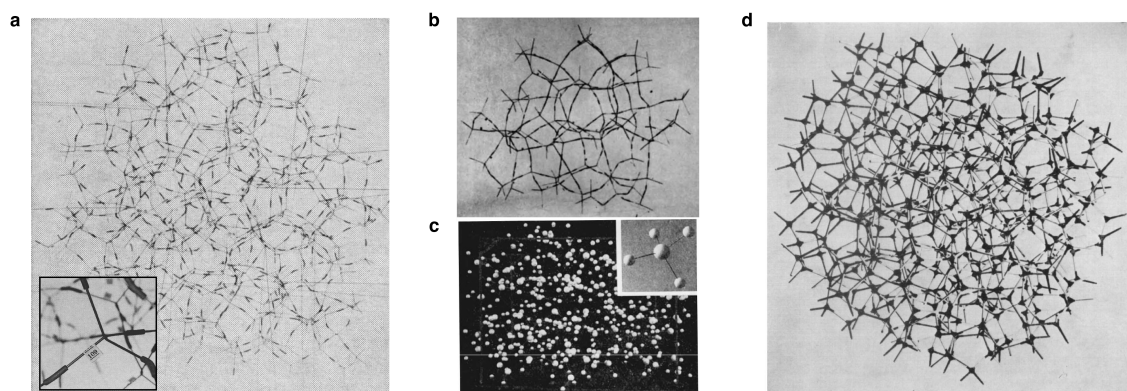


Figure 1.17: Hand-built models of CRNs. **a**, The first physical model for α -SiO₂ [200]. Panels **b** and **c** show two α -SiO₂ models built using bent springs and randomly assigned bond angles to improve the chemical sensibility [202, 203]. **d**, The first α -Si model [201].

Melt-quench simulations

One of the primary simulation protocols for generating amorphous structure models is ‘melt-quenching’. In this method, a well-equilibrated model of the liquid state is first obtained and then gradually cooled to form a disordered solid, analogous to the experimental glass transition process [105]. Due to its conceptual simplicity, melt-quenching is often the first technique attempted to create an amorphous model.

Melt-quenching is particularly effective when the structure of the liquid closely resembles that of the resulting glass. Silica [60, 208] and chalcogenide glasses [209] are prime examples. The reason for this success lies in the relatively minor changes in local- and intermediate-range order between the liquid and glass states, in contrast to materials like silicon. In other words, the fundamental structural units, or ‘building units’, between the two states are similar. Liquid silicon (*l*-Si) on the other hand transitions from a high-coordinate metallic liquid [210] to a tetrahedral-like amorphous state (*a*-Si), involving significant changes in atomic arrangements. The *l*-Si and *a*-Si building units are therefore very different, making the transition more challenging to describe.

It is important to note that the fastest computationally accessible quench rates are typically around 10^{11} K s^{-1} , which are orders of magnitude faster than those achieved in state-of-the-art sputtering experiments [2]. Moreover, accurately describing substantial changes in local atomic environments necessitates an accurate interatomic potential that describes many different atomic environments. This requirement explains why ML potentials have succeeded in generating high-quality *a*-Si models, whereas empirical potentials have not been as successful [184–186].

Bond-Switching Simulations

While melt-quenching is effective for some materials, generating high-quality amorphous models for others requires different approaches. Before ML-driven melt-quenches were introduced, the state-of-the-art models for *a*-Si were made with the ‘WWW’ or ‘W3’ technique, devolved by Wooten, Winer, and Weaire [111, 112].

The WWW method adopts a MC algorithm with ‘bond-switching’ moves, as shown in Fig. 1.18. An individual move proceeds as follows. First, for a bonded pair of atoms BC a pair of nearest neighbours, A and D are chosen, where A is the neighbour of B and not the neighbour of C, and D is the neighbour of C and not the neighbour of B. Second, the bonds AB and CD are broken (red bonds in Fig. 1.18, *left*) and new bonds AC and BD are created (green bonds in Fig. 1.18, *right*). Third, a partial structure relaxation using the Keating potential is performed, and the final system energy evaluated and assessed according to the Metropolis criteria.

In this way, the topological space of CRNs is navigated efficiently using quite physically-sensible moves [211]. The models that emerge from this method are in remarkable agreement with experiments probing the structure, electronic structure, and dynamics [111, 112, 212, 213].

Decorate and relax

In this section, I describe a simple model building strategy based on the concept of bond-decoration, as shown in Fig. 1.19(a) [214]. Starting with a four-coordinated *a*-Si (‘parent’) model, such as that produced by the WWW protocol, the Si–Si bonds are ‘decorated’ by inserting an oxygen atom at the centre of each bond, and rescaling the coordinates to the experimental density of the glass desired. The resultant (‘child’) structure is then relaxed, typically using an *ab initio* method [216, 217] or empirical potentials [218].

Further improvements to these decorated models can be achieved by refining the

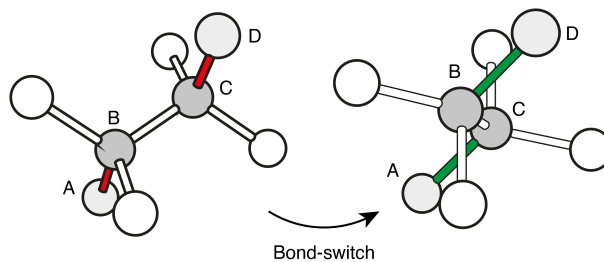


Figure 1.18: WWW bond-switch move.

relaxed configurations to experimental data using RMC [see later; Section 1.2.4] [219]. The additional refinement tends to ‘correct’ features intrinsic to interatomic potential models, such as too sharp intra-tetrahedral correlations.

The same idea can be applied to other tetrahedral networks. As shown in Fig. 1.19(a, *right*) silica models can be transformed into ZIF models. A model of α -SiO₂ can be modified by placing zinc atoms in place of silicon and Im molecules at the oxygen positions, oriented such that the C_2 axis of the Im molecule aligns with the Si–O–Si angle bisector, with the nitrogen atoms pointing towards the zinc atoms. This technique was used to generate the ZIF model illustrated in Fig. 1.19(b) [215]. The α -ZIF structure was derived from a 54-formula unit α -SiO₂ CRN. The atomic positions and cell size and shape were relaxed using DFT at the GGA level of theory [215]. While the local geometric features (e.g., bond lengths) in this model resembled those found in the crystalline polymorphs, the final structure had a density (3.0 Td nm⁻³) significantly lower than that of ZIF-4 (3.6 Td nm⁻³), which contradicts the experimentally observed increase in density upon amorphisation [76]. This inconsistency may arise from using a DFT GGA method without dispersion corrections [see Section 1.2.2], or because the strain-free prescribed topology is a poor representation of the material.

Nonetheless, the decorate and relax method offers a straightforward approach to

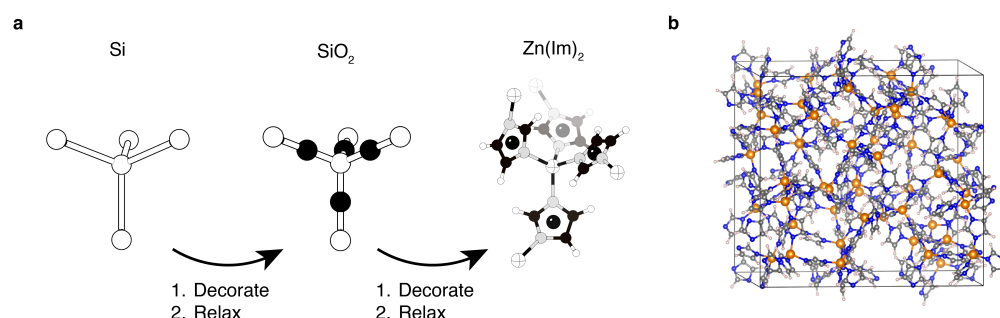


Figure 1.19: The decorate and relax method. **a**, Schematic of decorating Si to SiO₂ to Zn(Im)₂. **b**, Decorated and DFT-relaxed α -ZIF model based on an α -SiO₂ CRN [215]. The figure in panel **b** is adapted with permission from Ref. 215. Copyright 2016 American Chemical Society.

generating models that spare the need for expensive melt-quench simulations, while still achieving reasonable agreement with experiment. However, it necessitates the assumption that the intermediate-range order of the parent and child structures are similar.

Polymerisation algorithms

In cases where the change in building units between the liquid and glass renders melt-quenching unsuitable, and there is no parent structure to decorate, alternative modelling strategies are required. One approach is to model polymerisation events between the fundamental building units [220–222].

This method involves placing building units (monomers) with reactive sites in a low-density periodic box and then iteratively performing cycles of polymerisation and equilibration simulations. The process follows the following steps: (i) reactive atoms within a predefined distance cut-off form bonds; (ii) system energy minimisation is performed using an empirical force field; (iii) if bonding criteria are not met, MD simulation continues until bonding occurs; (iv) equilibration relaxes remaining stresses and facilitates structural rearrangement. This cycle repeats until all reactive sites are bonded or a set number of cycles is reached. A typical simulation is then carefully equilibrated to yield final structures with realistic and experimentally consistent densities [220, 223, 224].

Although originally devised for modelling polymers, this model has been successfully applied to model two amorphous MOFs: *a*-ZIF [221] and Fe-BTC, which features Fe^{3+} nodes linked by 1,3,5-benzenetricarboxylate linkers [222, 225]. The *a*-ZIF simulation employed a modified version of an empirical force field developed for ZIF-8 [226]. The ZIF glass model showed an intermediate density and porosity between ZIF-4 and ZIF-**zni**.

1.2.4 The inverse approach to model building

RMC refinement

A different approach to structural modelling to those introduced above is, rather than to construct a model based on interactions, to instead generate it using experimental data. This is the approach taken in RMC [227, 228].

In its simplest form, RMC refinement [227, 228] is a model-independent method for refining a configuration of atoms and/or spins to experimental data. RMC simulations proceed similarly to direct MC simulations described in Section 1.2.1: a series of moves are proposed sequentially and accepted or rejected according to the Metropolis criterion. The cost function in RMC is defined as the discrepancy between the calculated and experimental function. For a dataset i , it takes the form

$$\chi_i^2 = \sum_j \frac{[f_{i,\text{calc}}(x_j) - f_{i,\text{expt}}(x_j)]^2}{\sigma_{i,j}^2}, \quad (1.9)$$

where the sum runs over all data points x_j , weighted by $\sigma_{i,j}$, and $f_{i,\text{calc}}(x_j)$ and $f_{i,\text{expt}}(x_j)$ are the values of an observable computed from the model and obtained by experiment, respectively.

The datasets, $f_{i,\text{expt}}(x_j)$, are often diffraction data, but other experimentally accessible data have been used, such as EXAFS spectroscopy [229], single-crystal neutron, X-ray, and electron scattering data [230, 231], and more recently, X-ray Photoelectron Spectroscopy (XPS) [232].

The RMC method has achieved notable successes, particularly for systems described by pair potentials, such as liquid noble gases [228] and metallic glasses [233]. However, in general, more complex materials require additional constraints [234]. This need arises because experimental data, such as the PDF, do not directly provide information on higher-order correlation functions [235]. Additionally, RMC refinements are stochastic processes that produce the most disordered model consistent with the experimental data.

These challenges in RMC refinements are well-illustrated by different models of *a*-Si. A ‘naive’ RMC refinement of PDF data for *a*-Si yields an essentially perfect fit-to-data, but fails to reproduce the expected silicon atom coordination environment, instead giving highly-distorted and under-coordinated sites [236–238]. This failure is easily rationalised. An average over three- and five-fold coordination numbers can produce a similar PDF to a configuration which is entirely four-fold coordinated, such as those produced by the WWW method [111]. Such deficiencies must be addressed in order for the models produced by RMC to be useful.

Several approaches have been introduced to address these issues, which can be broadly divided into two categories: model-independent constraints and model-dependent constraints. The former seek to increase the ‘simplicity’ of the model, in a generalisable way, by reducing the structural variance between local atomic environments [239, 240]. There are numerous model-dependent constraints that aim to distinguish between the many highly discrepant configurations with similar satisfactory fits to the experimental data. The simplest examples are constraining the coordination number, bond length, and bond angle to expected values [236, 237]. By carefully selecting constraints for the system of interest, more reasonable structures can be obtained from RMC refinement.

In the following, I describe RMC models produced for ACC [241] and *a*-ZIF [76].

The ACC model refinements used configurations containing 1,620 formula units of $\text{CaCO}_3 \cdot \text{H}_2\text{O}$, which were refined to the experimental Q -weighted $F_X(Q)$ function shown in Fig. 1.10(a). The geometries of the carbonate molecules were maintained using bond-angle constraints, and a set of closest-approach restraints were used to prevent atoms from approaching one another too closely. H atoms were omitted during refinement due to their insignificant contribution to the experimental X-ray scattering function, and added at the end such that the H-bonding interactions were maximised. An additional soft data-based restraint on the shape of the PDF between 2 and 4 Å was used in order to prevent ‘spikiness’ in the partial PDFs that com-

monly results from closest-approach restraints [242]. A representative configuration produced is shown in Fig. 1.20(a) [241].

The models successfully reproduce the experimental total scattering data [241]. The most striking feature of the model is the emergence of two distinct components: a porous calcium-rich framework [red sticks in Fig. 1.20(a)] and a carbonate/water component that occupies the framework channels [blue surface representation in Fig. 1.20(a)] [241]. In the calcium-rich framework, the calcium atoms exhibit a broad coordination number distribution, averaging 5.3 and 5.8 for different starting configurations. A representative calcium environment is shown in Fig. 1.20(b). This calcium coordination environment differs from those in crystalline calcium carbonate phases, where there is a narrow distribution of coordination numbers in any single material, and the coordination numbers themselves are higher (6 in vaterite and calcite, 8 in monohydrocalcite, and 9 in aragonite). The calcium-poor regions contain some carbonate ions and water molecules left unbound by calcium [Fig. 1.20(c)]. The carbonate and water channels form connected pathways throughout the structure.

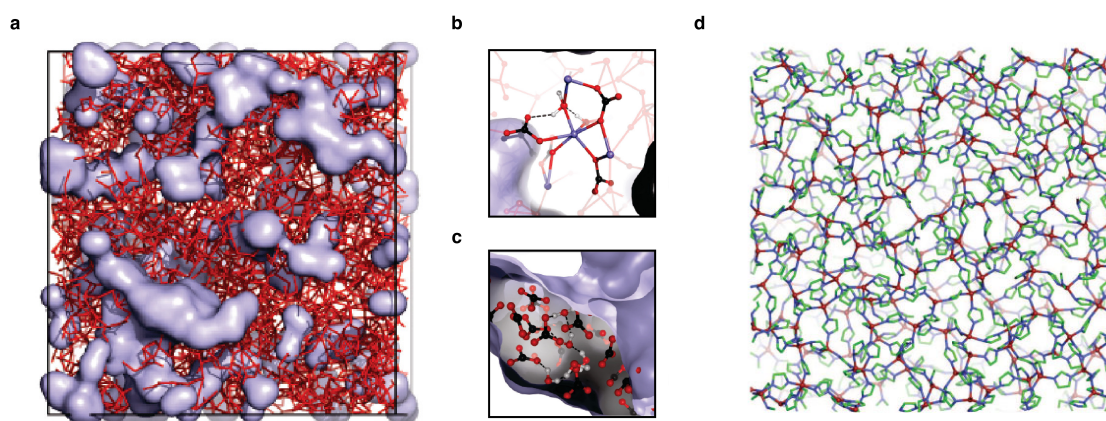


Figure 1.20: RMC models of ACC and α -ZIF. **a**, RMC model of ACC. The Ca-rich framework is shown with a red stick representation; the Ca-deficient channels are represented with a blue surface representation [241]. **b**, Representative Ca coordination environment from the ACC model [241]. **c**, Representative Ca-deficient channel featuring loosely-bound carbonate and water molecules [241]. **d**, RMC model of α -ZIF [76]. The figures in panels **a–c** are reproduced with permission from Ref. 241. Copyright 2010 American Chemical Society. The figure in panel **d** is reproduced with permission from Ref. 76. Copyright 2010 by the American Physical Society.

The ACC model shows pronounced interatomic correlations at 4 and 6 Å [*cf.* the effective PDF in Fig. 1.10(b) and the MD model described in Section 1.2.2], primarily arising from the Ca–Ca and Ca–O correlations. The two Ca–Ca separations are attributed to the two different ways that carbonate ions bridge calcium pairs: either by coordinating to the same oxygen atom or by bridging across the molecule [Fig. 1.20(b)]. Importantly, these features cannot be reproduced if the calcium atoms are distributed homogeneously throughout the ACC model [241].

The existence of a previously unrecognised channel network suggests mechanisms by which additives can be accommodated within the structure, providing temporary stabilisation and influencing the crystallisation process.

The starting models for the *a*-ZIF RMC refinements [76] were generated by decorating a CRN model of *a*-SiO₂ [Section 1.2.3], itself derived from an *a*-Si WWW [Section 1.2.3] model [111, 219]. The configuration was refined against the neutron $F_N(Q)$ and $G(r)$ data and the X-ray $F_X(Q)$ data [76]. The RMC refinement used additional constraints: connectivity constraints preserved the underlying network structure; molecular geometry for imidazolate ring geometries; and N–Zn–N bond angle restraints to ensure near-tetrahedral Zn coordination. A representative configuration is shown in Fig. 1.20(d) [76]. This model could reproduce both the neutron and X-ray experimental data, albeit with some discrepancy with $F_X(Q)$, particularly with respect to the intensity of the first peak. Notably, the CRN-based starting configuration could fit the experimental data, whereas refinements initiated with ZIF-4 or ZIF-**zni**—the bounding crystalline phases—could not [76]. This provided evidence that the topology of *a*-ZIF was disordered, rather than crystal-like.

HRMC refinement

More sophisticated approaches involve including both experimental and energetic information to construct the models. These approaches are referred to as hybrid Reverse Monte Carlo (HRMC) refinements. In general, the cost function to be

minimised is of the form

$$\chi^2 = \sum_i \chi_i^2 + \frac{\Phi_{\text{total}}}{k_{\text{B}}T}, \quad (1.10)$$

where the first term represents a sum over available experimental datasets [Equation 1.9], and the second is a contribution from an interatomic potential, Φ , weighted by $k_{\text{B}}T$, where k_{B} is the Boltzmann constant and T is the temperature.

HRMC refinement has been used to model *a*-C [243–245] and *a*-Si [246] using simple empirical potentials which formulate the atomic interactions in terms of a combination of two- and three-body terms [247–251]. In fact, a similar approach had been taken for modelling biomolecules back in 1978 [252]. The models produced by HRMC relieve many of the aforementioned issues with chemically unfeasible atomic environments, while still reproducing the experimental data.

EPSR

Another method that incorporates energetics during the refinement of experimental data is *Empirical Potential Structure Refinement* (EPSR) [253]. In EPSR, pairwise interatomic potentials are used to constrain the range of structural solutions. Unlike HRMC, the potentials themselves are refined against the diffraction data. This has two principal advantages. First, it introduces chemical information refinement which aids sensible structural solution as described above. Second, the potentials are differentiable and therefore MD-like collective moves for all particle coordinates can be performed, achieving faster convergence. It should be noted however that the potentials obtained in EPSR are not guaranteed to be physically meaningful. This can be important for systems with delicately balanced competing interactions.

For example, EPSR was used to generate a model of ACC [254]. A representative configuration is shown in Fig. 1.21(a). The starting model consisted of 1,000 formula units of $\text{CO}_3 \cdot \text{H}_2\text{O}$. The configuration was refined against neutron and X-ray total scattering data. The following additional constraints were applied: the molecular geometries of carbonate and water molecules were constrained with intramolecular harmonic bond length, bond angle, and torsional potentials; Lennard–Jones poten-

tials between all atoms to represent the short-range repulsive forces [Fig. 1.21(b)]; and pseudo-Coulomb potentials, to represent the attractive long-range forces [254].

The EPSR ACC model exhibits ordered anhydrous nano-domains [Fig. 1.21(a); orange surface] with an interstitial net-like matrix of water molecules [Fig. 1.21(a); ball and stick molecules], creating a heterogeneous mesoscale structure [254]. On a local scale, the bond length and coordination number distributions are in line with experiments [Section 1.1.4]. Considering the intermediate-range order, the characteristic peaks appear in the $g_{\text{Ca}}(r)$ pair correlation at 4 and 6 Å, but notably an additional peak is found at 8 Å, which is absent in the experimental PDF [95, 254].

An earlier model of Mg-stabilised ACC was generated using EPSR to fit neutron and X-ray total scattering data [255]. This model followed a similar protocol to the one described above, including the use of Lennard–Jones interatomic potential constraints. The composition for this model included 950 Ca, 50 Mg, 1000 CO_3 , and 250 H_2O units. Here, the discrepancy with the experimental X-ray PDF is most clear, as shown in Fig. 1.21(c).

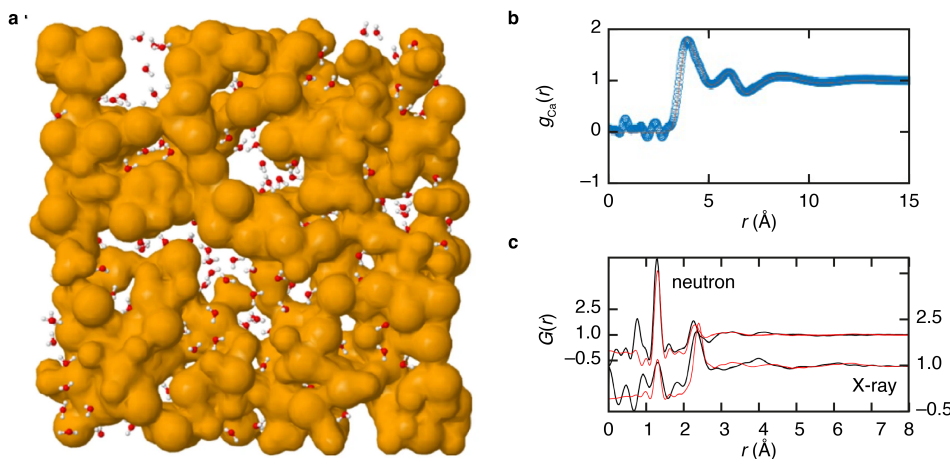


Figure 1.21: EPSR models of ACC. **a**, Representation of a converged structure of ACC obtained using EPSR [254]. **b**, The $g_{\text{Ca}}(r)$ pair correlation function extracted from the EPSR configurations [254]. It shows maxima at 4, 6, and 8 Å [254]. **c**, Comparison between the experimental (black) and simulated (red) total pair distribution functions $G(r)$ for the EPSR model reported in Ref. 255. This figures in panels **a** and **b** are adapted with permission from Ref. 254. Copyright 2022 Author(s) licensed under a Creative Commons Attribution 4.0 License. The figure in panel **c** is adapted with permission from Ref. 255.

1.3 Aims and structure of this thesis

This thesis has two primary aims. The first aim is to develop a modelling strategy that can generate the most ‘realistic’ models for amorphous materials. In this context, ‘realistic’ means that the model should be consistent with experimental data while also being energetically sensible. To achieve this, I implement an HRMC protocol that simultaneously minimises fits to experimental total scattering data and the system energy computed using high-quality force fields. Second, I aim to use these models to identify the key structural features and chemical interactions at play, thereby shedding light on the origins of structural complexity in these materials. Coarse-graining is a common theme throughout my analysis to aid in this understanding.

The thesis is structured as follows. In Chapter 2, I describe the HRMC implementation used throughout this thesis. Using accurate empirical potentials and experimental X-ray total scattering data for ACC, I demonstrate HRMC’s efficacy in generating high-quality structural models. ACC is an important material for understanding biomineralisation pathways. I further our understanding of the material by considering a coarse-grained interaction model that successfully captures the heterogeneous calcium distribution observed in the atomistic ACC model. Importantly, this model maps onto the Lennard-Jones–Gauss (LJG) phase space that provides context for the structural complexity in ACC.

The remainder of the thesis focuses on MOFs. In Chapter 3, I investigate various MOFs, all derived from vertex-sharing tetrahedral polyhedra, in terms of their local geometries. The aim is to understand the relationship between the local geometries of MOF building units and the topologies they support. Building on previous work described in Ref. 256, I develop automated structural coarse-graining routines and combine them with atom-density-based descriptors of local environments to ‘map out’ the structural diversity accessible to tetrahedral MOF frameworks. I describe the implementation, as released in an open-source software named CHIC

(Coarse-graining Hybrid and Inorganic Crystals) [257], and the structural analysis it supports.

Building on the success of the HRMC approach for modelling ACC and the relationship between chemical components and framework topology in MOFs, the final two chapters focus on modelling the structure and topology of *a*-ZIF-4. In Chapter 4, I develop a fast and accurate ML potential for ZIFs, trained using a carefully constructed dataset of structures. I introduce a new approach for expanding the dataset based on the HRMC protocol, enabling the exploration of configurations that would otherwise not be represented in the training data. I demonstrate that the ML potential is well-suited for modelling the *a*-ZIF form using both MD and HRMC simulations.

In Chapter 5, I show that the ML potential developed in the previous chapter enables the generation of high-quality *a*-ZIF-4 models. A simulated annealing process performed with MD allows the topology to rearrange into an energetically preferred arrangement, and subsequent HRMC refinement ensures the model aligns well with experimental X-ray and neutron total scattering data. This approach facilitates investigations of the structure of an amorphous MOF on multiple length scales and allows for a discussion of the nuanced landscape of CRNs accessible to different material families.

I conclude in Chapter 6 with a discussion of the work as a whole, offering suggestions for future research, and exploring how the tools and approaches developed in this thesis may be applicable to the broader field of materials modelling.

Chapter 2

The structure of ACC

2.1 Acknowledgements

I am grateful to Adam Stones (formerly University of Oxford) for providing the code for deriving pair potentials from pair correlation functions and particle coordinates, and for helpful discussions. I thank Adam Patel (formerly University of Oxford) for his contributions during a preliminary investigation of this project. I am grateful to Dirk Aarts (University of Oxford), F. Marc Michel (Virginia Tech), Richard Reeder (Stony Brook University), and Martin Dove (Queen Mary University of London) for useful discussions regarding the analysis and interpretation of the ACC structural model. The findings discussed in this chapter are published in Ref. 258.

2.2 Introduction

In Section 1.1.4, I introduced ACC, emphasising its significance as an amorphous precursor in the biomineralisation processes of marine organisms [5, 88, 89]. Key challenges include understanding the structure of ACC and rationalising its metastability as an amorphous phase. In Chapter 1, I presented three existing models for ACC [122, 241, 254], which I will briefly review here.

The first model, developed using RMC refinement [Section 1.2.4] [241], revealed a heterogeneous structure comprising a calcium-rich framework with interconnected channels filled with water and carbonate molecules. This heterogeneity was inferred from the experimental PDF, where pronounced interatomic correlations at 4 and 6 Å were attributed to the distribution of calcium atoms within the calcium-rich framework. If the calcium atoms were homogeneously distributed, these features would be absent, and the experimental PDF would not be accurately reproduced. This model suggests a significant thermodynamic cost associated with maintaining charge

separation due to the segregation into a cationic framework and anion-dominated pore network. Additionally, the calcium coordination number distributions were broader and shifted to lower values compared to crystalline structures.

The second model used a carefully parameterised force field for aqueous calcium carbonate [Section 1.2.2] to perform MD simulations to generate ACC nanoparticles [122]. These models produced coordination number distributions more aligned with those observed in crystalline polymorphs than the RMC model. Nanoparticles with water content between 1 and 1.5 molecules per formula unit exhibited Ca–Ca correlations similar to those in the RMC model and experimental PDF [95, 241]. Rapidly grown clusters exhibited drier cores compared to outer regions, leading to structural inhomogeneity, though less pronounced than in the RMC model. However, the investigation into intermediate-range order was restricted by the nanoparticle size (≈ 5 nm). Importantly, this model did not exhibit charge separation on a nanometre scale, unlike the RMC model.

The third model, developed using EPSR [Section 1.2.4] [254], depicted highly ordered anhydrous CaCO_3 nano-domains held together by an interstitial net-like matrix of water molecules. The local atomic environments were consistent with those in crystalline polymorphs and MD simulations [122]. However, the Ca–Ca partial pair correlation function in this model featured an additional correlation at 8 Å, which is inconsistent with the experimental PDF [95]. The discrepancy between the model and experiment may be due to the use of inappropriate constraints in the EPSR refinement. Rather than using interatomic potentials fitted specifically to model the ACC system, the potentials themselves are refined during the simulation and are therefore not guaranteed to be accurate. In turn, this might result in a model which misses key aspects of the structure, such as the correct distribution of calcium ions.

Considering these models together, three main observations emerge. First, refinement to experimental total scattering data alone is insufficient to generate a

chemically sensible model. This is evidenced by the incorporation of energetics in MD simulations and EPSR refinements, which preclude the charge segregation found in the RMC model. Second, in the absence of data-fitting, key aspects of intermediate-range order are not recovered, possibly due to the limited timescales accessible in MD simulations preventing appropriate exploration of configurational space [122]. Third, an accurate description of chemical energetics is crucial for generating a reliable structural model. Even with experimental data constraints, the EPSR model developed features that could possibly be attributed to the inappropriate functional form of interatomic potential constraints.

Motivated by these observations, in this chapter, I explore using HRMC refinements [Section 1.2.4] as a strategy to improve the quality of ACC structural models. By incorporating refinement to data and energy minimisation using accurate aqueous calcium carbonate potentials into the model-building protocol, I aim to recover the structural heterogeneity informed by the former and the chemical sensibility promoted by the latter.

This chapter is structured as follows. First, I describe the HRMC simulation protocol that I have implemented and used throughout this thesis. This protocol allows for the simultaneous refinement of a structural model to fit total scattering data while minimising the system’s energy using accurate force field methods. Second, I demonstrate how this approach significantly enhances the quality of the structural models for ACC, compared to models produced by MD simulation or RMC refinement. Finally, by focusing on the distribution of Ca atoms—the main contributor to intermediate-range order in the amorphous model—I rationalise the (meta-) stability of ACC. This is achieved by mapping the effective interactions within the Ca–Ca distribution onto a well-studied theoretical phase space known to support complex structures.

2.3 Methods

2.3.1 HRMC

My approach to carrying out HRMC refinements was modelled closely on the previous RMC refinements of Ref. 241, which were carried out using the `RMCTProfile` program [242]. However, to support evaluating the potential-energy using a variety of sophisticated force fields (beyond the simple harmonic bond and angle terms supported by `RMCTProfile`), I wrote a custom HRMC code in the Python coding language. The X-ray scattering calculations carried out by this code were exactly those employed by `RMCTProfile`, while the potential energy evaluations were delegated to the highly-parallelised Large-scale Atomic/Molecular Massively Parallel Simulator (`LAMMPS`) software [259], using the `PyLammpsMPI` interface [260]. `LAMMPS` is an open-source MD simulator which supports community-led contributions to the code. This brings the benefit that many new force field developments—including many ML potentials—are, and continue to be, implemented in `LAMMPS`. Furthermore, expensive calculations common to many force field approaches, such as evaluating Coulomb contributions to the energy, are highly optimised. On a practical level, in conjunction with high-performance computing infrastructures, this increases the accessibility of such advanced modelling strategies, which otherwise would be computationally prohibitively expensive.

2.3.2 Proposed moves

Individual moves at each step of the HRMC refinement involved either atomic displacements or, if specified, rigid-body translations and rotations. Support for rigid bodies allowed for a more efficient refinement strategy where internal molecular degrees of freedom are known to be unimportant, while also allowing the use of rigid-body force fields where molecular degrees of freedom may not be parameterised.

2.3.3 Cost function

The HRMC cost function, χ^2 , defined in Equation 1.10, to be minimised during refinement, contains two terms. The first represents the difference between the calculated and experimental functions for all experimental datasets, and is computed according to Equation 1.9. In this thesis, I use a constant weighting term, $\sigma_j = \sigma$, for all data points. This parameter therefore controls the relative weighting of dataset i to the overall loss function, χ^2 . Small values of σ place a greater weight on fitting experimental data i relative to other cost functions, and *vice versa*. The second term in Equation 1.10 represents the contribution from an interatomic potential, weighted by $k_B T$. I set the value of T to the temperature that the experiment was performed at. I chose the value of σ by performing a number of refinements with different values, on a logarithmic between 0.001 and 1.0, and selecting the value that gives a good compromise between data-fitting and energy minimisation.

All experimental datasets used for HRMC refinement in this thesis are total scattering functions, $F(Q)$ [199, 242]. In this chapter, I refine to the X-ray total scattering function, $F_X(Q)$.

X-ray total scattering calculations

In this section, I will describe the key equations used to compute $F_X(Q)$, following the formalism described in Ref.199. X-rays are scattered by the electrons surrounding each nucleus. Denoting the electron density about a given nucleus as a continuous distribution, $\rho_{\text{el}}(\mathbf{r})$, the X-ray atomic scattering factor for atom or ion i is given by

$$f_i(\mathbf{Q}) = \int \rho_{\text{el}}(\mathbf{r}) \exp(i\mathbf{Q} \cdot \mathbf{r}) \, d\mathbf{r}. \quad (2.1)$$

This form factor is therefore determined by the size of the electron shells and the number of electrons, and arises from the interference of scattered radiation from different parts of the electron shell. It has the effect of suppressing the intensity of diffraction at high- Q values. For all atoms and ions of interest, $\rho_{\text{el}}(\mathbf{r})$ can be com-

puted using quantum mechanics. In practice, there is not an analytical function for $f_i(\mathbf{Q})$, and so instead numeric values $f_i(Q)$ are fitted to a functional form, expressed as

$$f_i(Q) = \sum_{i=1}^4 [a_i \exp(-b_i s^2)] + c, \quad (2.2)$$

where $s = Q/4\pi$ and a_i and b_i are tabulated coefficients for each atom and ion [261, 262].

Now, considering a material with an average number density $\rho_0 = N/V$ (in atoms $\text{\AA}^{-3}$), and c_i as the proportion of species i in the material, the partial pair correlation functions, $g_{ij}(r)$, can be obtained as histograms

$$g_{ij}(r) = \frac{n_{ij}(r)}{4\pi r^2 dr \rho_j}, \quad (2.3)$$

where $n_{ij}(r)$ are the number of particles of type j between distances r and $r + dr$ from a particle of type i , and $\rho_j = c_j \rho_0$. The Faber–Ziman partial structure factors, $A_{ij}(Q)$, are related to $g_{ij}(r)$ as follows [31]

$$A_{ij}(Q) - 1 = \rho_0 \int_0^\infty 4\pi r^2 [g_{ij}(r) - 1] \frac{\sin(Qr)}{Qr} dr. \quad (2.4)$$

Then, $F_X(Q)$ is computed as

$$F_X(Q) = \rho_0 \sum_{i,j=1}^n c_i c_j f_{ij}(Q) [A_{ij}(Q) - 1], \quad (2.5)$$

where

$$f_{ij}(Q) = \frac{f_i(Q) f_j(Q)}{[\sum_{i=1}^n c_i f_i(Q)]^2} \quad (2.6)$$

are the Q -dependent X-ray partial pair form factors.

2.3.4 ACC HRMC refinement

I now describe in detail the HRMC refinements performed for ACC. The cost function for ACC HRMC refinements balanced two contributions. First, the fit to the

experimental Q -weighted X-ray total scattering function, $QF_X(Q)$ using the data reported in Refs. 95 and 241. This data spans the range of momentum transfers $1.2 \leq Q \leq 25.0 \text{ \AA}^{-1}$. Second, the configurational energy as obtained from the rigid-body force field for aqueous calcium carbonate of Ref. 121, which I described in Section 1.2.2. The simulation temperature was set to $T = 300 \text{ K}$.

Starting configurations for the ACC HRMC refinement were generated in the following manner. I began by placing 1620 Ca atoms randomly in a box of dimensions $52.8 \times 54.8 \times 45.2 \text{ \AA}^3$ —the same box size as that used in the original RMC study of Ref. 241. An initial energy minimisation was carried out with a repulsive potential between Ca atoms (a Lennard-Jones functional form, truncated at the zero-crossing distance), so as to prevent any Ca–Ca distances $< 3.0 \text{ \AA}$. I then added a stoichiometric number of rigid carbonate and water molecules with idealised geometries to this homogeneous distribution of Ca atoms. Carbonate molecules were fixed to adopt a C–O bond length of $r_{\text{CO}} = 1.284 \text{ \AA}$ —as observed in calcite and as defined for the relevant CaCO_3 potential [121]. Water molecules were treated using the four-site rigid model TIP4P-Ew [127] with the gas-phase geometry of the water monomer ($r_{\text{OH}} = 0.9572 \text{ \AA}$, $\theta_{\text{HOH}} = 104.52^\circ$). Short-range repulsive terms (as described above) were applied between the rigid bodies to prevent close contacts during a subsequent energy minimisation and a short NVT ensemble MD run, aiming to achieve a homogeneous distribution of particles (*cf.* the method of Ref. 263).

HRMC refinements were continued until the values of χ^2 and E were deemed to have converged [Fig. 2.1(a) and (b), respectively]. To check the stability of the HRMC ACC model, I performed a 1 ns MD simulation in the NVT ensemble at 300 K, starting from the final HRMC configuration. The simulation was stable with respect to NVT MD [Fig. 2.1(b–e)], by which I mean that no significant structural reorganisation occurred. During the MD simulation, it quickly decreased in energy by $\approx 0.3 \text{ eV/Ca}$ before reaching thermal equilibrium [Fig. 2.1(b)]. The negative pressure [Fig. 2.1(d)] indicates that the system is driven towards a more compact

atomic arrangement than implied by the experimental density. An MD simulation performed in the NPT ensemble at 300 K led to an increase in density from 2.43 to 2.91 g cm⁻³.

2.3.5 Coarse-grained simulations

The coarse-grained Lennard-Jones–Gauss (LJG) simulations discussed in Section 2.4.2 were performed using a custom Python program with potential-energy evaluations and MD protocols carried out using LAMMPS [259]. For each method, 12 starting configurations (24 in total) were generated by randomly placing 1620 Ca atoms in a box with cell dimensions identical to the HRMC configuration ($52.8 \times 54.8 \times 45.2$ Å³) and the simulation temperature set to $T = 300$ K [*cf.* HRMC simulations; Section 2.3.4]. For MC simulations, the maximum atomic displacement was 0.1 Å. MC simulations were terminated after 4 million moves were accepted, at which point convergence with respect to the ensemble energy was met.

An initial energy minimisation (using the the conjugate-gradient algorithm for 100 steps) was performed for the starting MD configurations to prevent close contacts rendering the simulation unstable. The MD simulations were run in the canon-

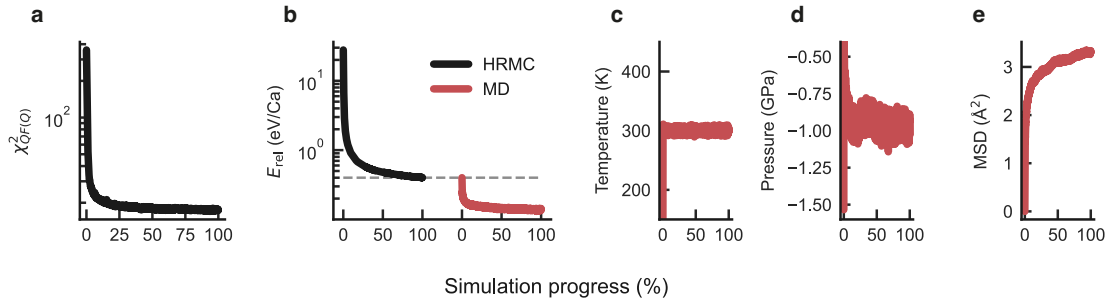


Figure 2.1: Evolution of system properties during HRMC refinement and MD simulation. **a**, The cost function relating to the X-ray scattering data during HRMC refinement. **b**, The change in potential energy of the simulated structural model during HRMC refinement (black) and subsequent MD simulation (red). Energies are reported relative to the monohydrocalcite polymorph (same stoichiometry as the HRMC configuration) which itself was equilibrated in an NVT MD simulation starting from a supercell of the experimental structure (following the same simulation protocol as defined for the HRMC configuration in the following section). Panels **c–e** show the temperature, pressure, and mean square displacement (MSD) of atoms during the HRMC MD simulation, respectively.

ical ensemble (constant volume and temperature; NVT) for 1 ns with a Nosé-Hoover thermostat and a temperature damping parameter of 50 fs [114, 115]. The timestep was 0.5 fs.

2.3.6 Structure analysis

All coordination statistics reported throughout this chapter use the radial cut-offs for determining neighbouring sites given in Table 2.1. These values were selected to match those used in the earlier RMC study [241], enabling a direct comparison of the structural analysis results.

Table 2.1: Radial cut-offs for determining local coordination environments.

Species	Atoms involved	Cut-off (Å)
Ca-CO ₃	Ca, O _C	2.80
Ca-H ₂ O	Ca, O _W	2.80
CO ₃ -H ₂ O	O _C , H	2.30
H ₂ O-H ₂ O	O _W , O _W	3.75

2.4 Results

2.4.1 The structure of ACC

I show in Fig. 2.2(a) the $QF_X(Q)$ function calculated from a representative HRMC ACC configuration, and compare it to the equivalent functions predicted using ‘pure’ RMC refinement [241] and unconstrained MD simulations with the potentials of Ref. 121. Both RMC and HRMC give very similar high-quality fits to experiment, whilst the MD simulation misses some aspects of the $QF_X(Q)$ function. By contrast, the HRMC and MD models give similar relative energies, whereas the RMC model is less stable than both by more than 800 kJ mol⁻¹ per formula unit. These results are summarised in Fig. 2.2(b).

I show the HRMC model in Fig. 2.2(c). The ACC structure consists of CaCO_3 -rich regions separated by a filamentary network of water molecules: I term this a ‘blue cheese’ model. Qualitatively similar descriptions were obtained in MD simulations [130], in EPSR refinements of combined neutron/X-ray total scattering data [Section 1.2.4] [254, 264], and also inferred from solid-state NMR measurements [Section 1.1.4] [265]. This is reflected by the fact that nearly all H_2O molecules are bound to Ca^{2+} (as implied by ^1H NMR; Ref. 266), but with neighbouring water molecules sufficiently close to form a network that percolates the ACC structure.

Structural analysis of the HRMC ACC configuration provides further evidence for the blue cheese model description. I show the RDFs relevant to the water-based interactions in Fig. 2.3(a,b). The presence of a peak at ≈ 1.8 Å for the $g_{\text{OCH}}(r)$ function, and its absence in the corresponding $g_{\text{OWH}}(r)$ function [Fig. 2.3(a) and (b), respectively], indicates a preferential H-bonding interaction between water and carbonate molecules, rather than between water molecules. This can be rationalised in terms of an energetic preference for electrostatic interactions between

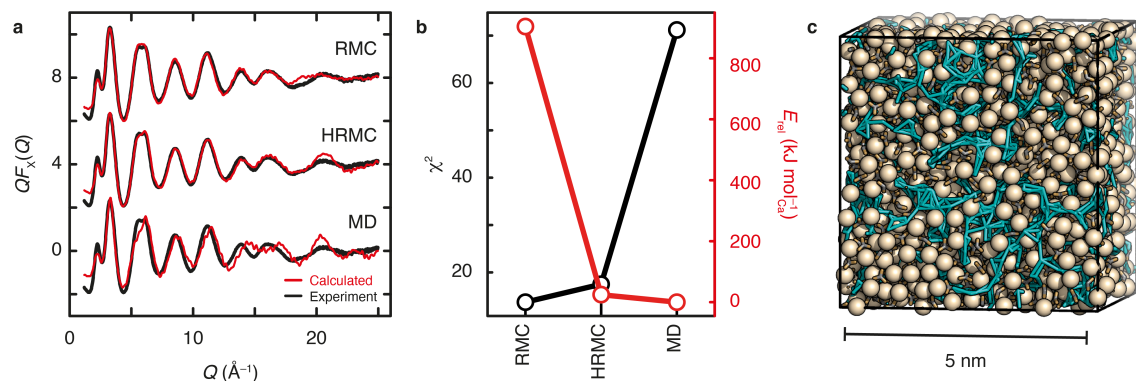


Figure 2.2: HRMC refinement for ACC. **a**, The experimental Q -weighted X-ray total scattering function $QF_X(Q)$ is well fitted by both RMC (*top*) and HRMC (*middle*) refinements, but is meaningfully different to that calculated using the interatomic potential of Ref. 121 (*bottom*). Experimental data are shown as black lines, and fits or calculations as red lines. **b**, Quality criteria for the different structural models. HRMC simultaneously optimises both goodness-of-fit to experimental data χ^2 and cohesive energy (E_{rel}). **c**, Representation of a converged structure of ACC obtained using HRMC. Water molecules connect to form filamentary strands (aquamarine strings) that separate calcium-carbonate-rich domains (Ca atoms as large beige spheres, carbonate ions shown as stick representation).

water molecules and the more highly charged carbonate molecules, compared to the weaker water–water interactions. Furthermore, the application of percolation theory to MD-derived atomistic models of ACC has suggested that the percolation transition of water through the structure occurs at a hydration level of $x \approx 0.8$ for $\text{CaCO}_3 \cdot x\text{H}_2\text{O}$ [263]. Indeed, the HRMC model ($x = 1.0$) exhibits percolating water channels with 1404 (of 1620) water molecules participating in a percolating cluster. In this calculation, I have assumed that water molecules are connected when $r_{\text{O}_\text{W}\text{O}_\text{W}} \leq 3.80 \text{ \AA}$, which is the same upper limit as that used in Ref. 263. The remaining (‘finite-size cluster’) water molecules are predominantly found as lone species; however, non-percolating clusters containing as many as 17 water molecules are found in the HRMC configuration [Fig. 2.3(c)].

Locally, the HRMC model shows coordination environments that are consistent with the consensus of recent computational and experimental studies. To quantify this, I will first report coordination statistics and show the key bond-length and bond-angle distributions.

The contributions to the Ca–O partial pair distribution function arising from nearest-neighbour Ca–O_W and Ca–O_C bonds are shown in Fig. 2.4(a). One key feature to emphasise—which I will return to in Section 2.4.3—is that there are two

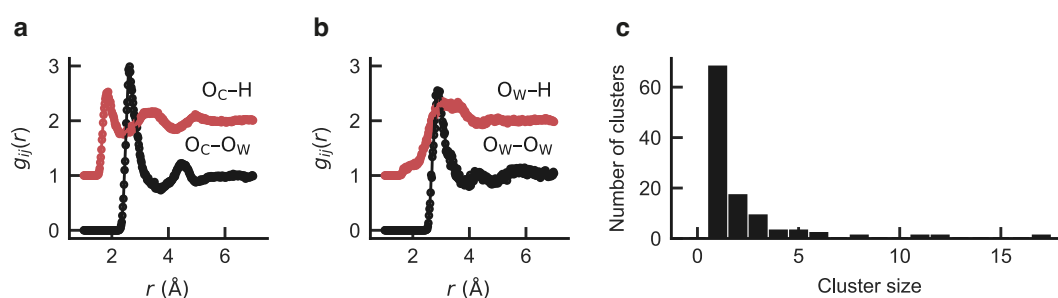


Figure 2.3: Characteristics of water distribution in the HRMC ACC model. **a**, RDFs for neighbours of carbonate oxygen atoms in the structure. **b**, RDFs for neighbours of oxygen atoms in water molecules (“O_W”) only. Intramolecular correlations are omitted for clarity. **c**, The distribution of cluster sizes for those water molecules that do *not* take part in the percolating water network. A cluster size of 1 means an isolated water molecule that is not within reach of any others according to the defined cut-off distances (Table 2.1).

possible ways that a carbonate ion can bridge calcium-ion pairs: either a pair of Ca^{2+} ions share a common carbonate oxygen neighbour, or they bind distinct oxygens and so are connected formally by a Ca-O-C-O-Ca pathway. This difference manifests in the distinctive bimodal distribution of Ca–Ca bond lengths shown in Fig. 2.4(b), at separations of ≈ 4 and $6 \text{ \AA}$. These preferred separations recur amongst crystalline calcium carbonates, as noted in Ref. 241.

Figures 2.4(c) and (d) illustrate the bond-angle distributions associated with key O–Ca–O triplets. In Fig. 2.4(c) I show the distribution of $\text{O}_\text{C}\text{--Ca--O}_\text{C}$ angles, decomposed into their contributions from bidentate and monodentate carbonate ligands. The former give rise to a sharp peak centred near 50° , whereas the latter correspond to a broad range of bond angles spanning $50\text{--}180^\circ$. Figure 2.4(d) shows the contributions arising from $\text{O}_\text{W}\text{--Ca--O}_\text{C}$ and $\text{O}_\text{W}\text{--Ca--O}_\text{W}$ triplets; these contributions are relatively similar in distribution to one another. The lower relative intensity for the

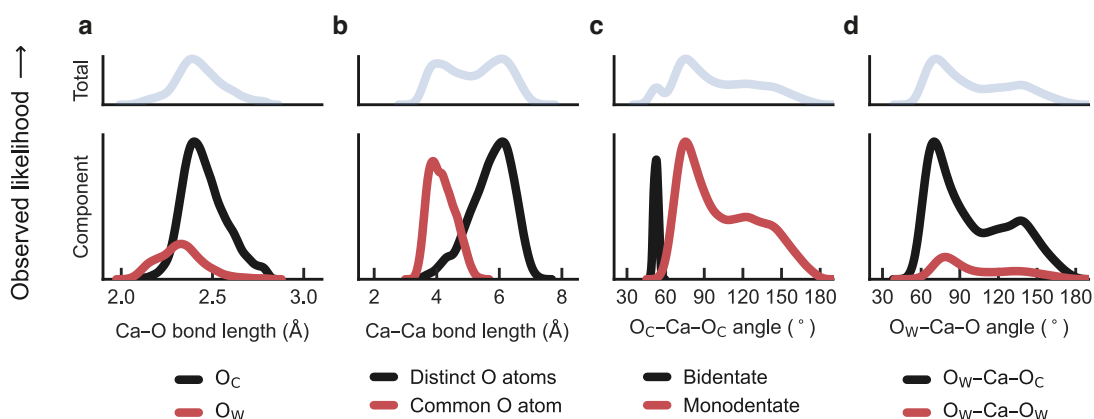


Figure 2.4: Key bond-length and bond-angle distributions in ACC. **a**, The bond-length distributions for Ca–O divided into Ca– O_C and Ca– O_W contributions. The average Ca– O_C bond length ($2.5 \pm 0.1 \text{ \AA}$) is shifted to larger r values relative to the average Ca– O_W bond length ($2.3 \pm 0.1 \text{ \AA}$). **b**, The interatomic distance between calcium atoms bound to the same carbonate can be divided into those which share a common oxygen atom and those which are connected *via* the carbonate molecule. These two coordination interactions give rise to the characteristic geometric frustration observed for ACC. **c**, Bond-angle distributions for $\text{O}_\text{C}\text{--Ca--O}_\text{C}$ triplets, decomposed into mono- and bi-dentate coordination modes. The bidentate coordination mode gives rise to a sharp peak centred at $\approx 50^\circ$. **d**, $\text{O}_\text{W}\text{--Ca--O}$ bond-angle distributions. The lower relative intensity of $\text{O}_\text{W}\text{--Ca--O}_\text{W}$ correlations reflects the low likelihood of finding multiple water molecules bound to the same calcium atom. The plots at the top show the corresponding summed distribution functions.

O_W -Ca- O_W correlation compared to the O_W -Ca- O_C correlation reflects the small number of calcium environments bound by two (or more) water molecules.

Considering now the local calcium and carbonate environments, I show in Fig. 2.5(a) and (b) the distribution of calcium and carbonate coordination numbers. The average Ca^{2+} coordination number of 7.0 is similar to that reported in Refs. 254, 264, 267. Whereas most Ca^{2+} bind either five or six distinct carbonate anions, the carbonates tend to bind one fewer Ca^{2+} ion each, reflecting a binding-mode distribution of $\sim 80:20$ monodentate:bidentate. Representative modal coordination geometries for Ca^{2+} and CO_3^{2-} are shown in Fig. 2.5(c) and (d). Note in particular the two distinct carbonate bridging motifs in Fig. 2.5(d) at the characteristic 4 and 6 Å distances.

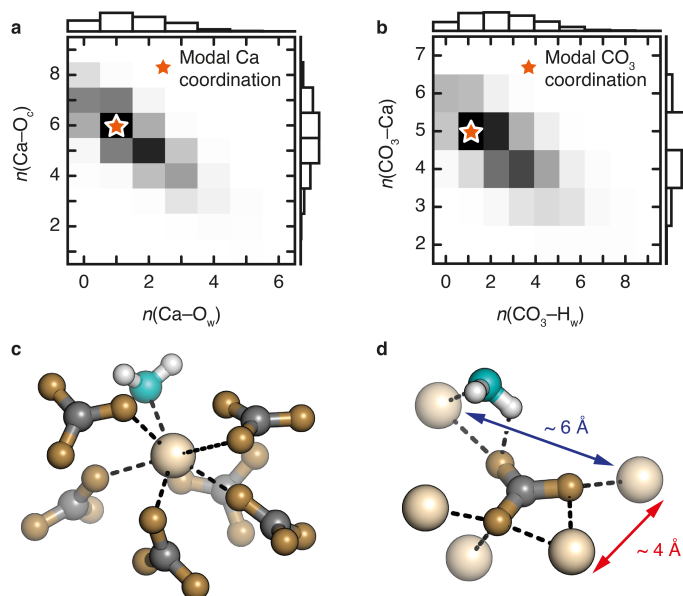


Figure 2.5: Local structure analysis of ACC. **a**, Coordination number histogram for Ca^{2+} environments, decomposed into contributions from carbonate and water oxygen donors (O_C and O_W , respectively). **b**, Coordination number histogram for CO_3^{2-} environments, decomposed into contributions from Ca^{2+} and water hydrogen donors (O_W). **c**, Representative Ca^{2+} coordination sphere for the modal coordination environment marked by a star in **a**. **d**, Representative CO_3^{2-} coordination sphere for the modal coordination environment marked by a star in **b**. Ca, C, O, and H atoms shown in beige, grey, aquamarine, and white, respectively.

2.4.2 Coarse-graining

In seeking to improve our understanding of the structure of ACC, I focused on the Ca-pair correlation function $g_{\text{Ca}}(r)$ because this is the contribution to the pair distribution function that exhibits the strongest persistent well-defined oscillations. In Fig. 2.6, I show this function as extracted from the HRMC configuration compared against that reported in the RMC study of Ref. 241. I also include the Fourier transform of the experimental $F_{\text{X}}(Q)$ function; this transform is an approximate (total) pair distribution function that emphasises contributions from Ca–Ca pairs as a consequence of the larger X-ray scattering cross-section for Ca relative to C, O, and H. While all three real-space correlation functions show maxima at $r \simeq 4$ and $6 \text{ \AA}$, it is the HRMC result that resolves these features most clearly.

Access to a smoothly-varying measure of $g_{\text{Ca}}(r)$, together with the HRMC configurations from which it is calculated, enables the use of a recently-developed approach for the direct measurement of effective pair potentials from particle-coordinate data [268]. The method works by equating the pair distribution functions measured

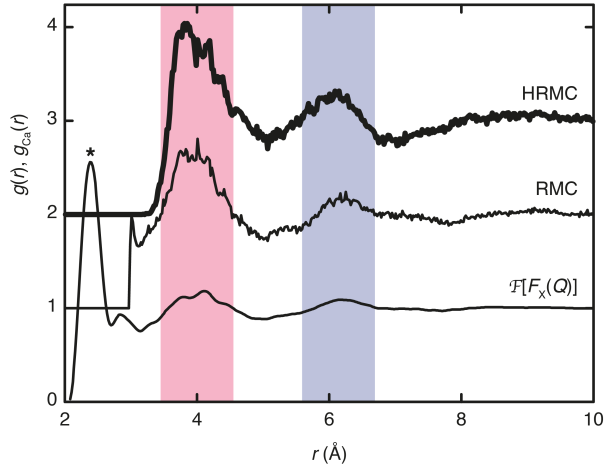


Figure 2.6: Ca-pair pair correlation functions in ACC. Ca-pair correlation functions $g_{\text{Ca}}(r)$ extracted from HRMC and RMC configurations, compared against the normalised Fourier transform of the experimental X-ray total scattering function (which includes contributions from all atom-pairs, e.g., the Ca–O peak at $2.4 \text{ \AA}$ marked with an asterisk). The two principal peaks common to all functions, indicated by red and blue shading, can be assigned to the two types of Ca^{2+} -ion pairs illustrated in Fig. 2.5d.

directly, on the one hand, and calculated using a test-particle insertion approach [269, 270], on the other hand [271]. By applying this methodology to the atomic coordinates in the HRMC-refined model, I extracted the effective Ca^{2+} -ion pair potential $u_{\text{Ca}}(r)$ shown in Fig. 2.7(a). This function has two distinct potential wells, the minima of which are centred near distances for which $g_{\text{Ca}}(r)$ has its maxima. The physical meaning of $u_{\text{Ca}}(r)$ is that it captures the effective two-body interactions between Ca–Ca pairs, mediated by carbonate and water, as required to account for the observed $g_{\text{Ca}}(r)$. That this interaction energy is minimised when Ca–Ca pairs are separated by distances corresponding to either Ca–O–Ca or Ca–O–C–O–Ca pathways makes intuitive sense; the longer (6 Å) separation appears energetically more favourable likely because it minimises the electrostatic repulsion between calcium ions bound to a common carbonate.

One assumption in using an isotropic effective pair potential to capture interactions mediated by anisotropic molecular species, such as H_2O and CO_3^{2-} , is that any

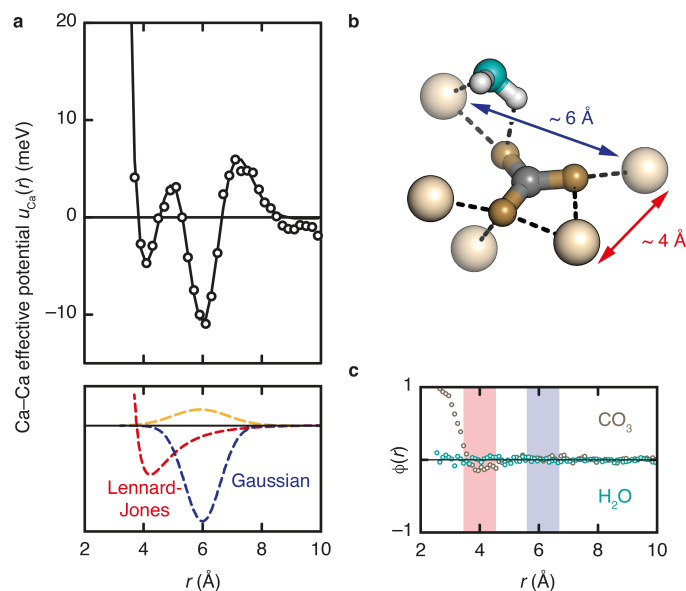


Figure 2.7: Effective $\text{Ca} \cdots \text{Ca}$ interactions in ACC. **a**, The effective $\text{Ca} \cdots \text{Ca}$ interatomic potential extracted from the HRMC configurations [268] (open circles) and least-squares fit using a modified LJG model (line) as described in the text. **b**, Illustration of the two key $\text{Ca} \cdots \text{Ca}$ distances and their origin in terms of coordination modes to carbonate molecules. **c**, The orientational correlation functions (Equation (2.7)), which vanish for distances relevant to the $\text{Ca} \cdots \text{Ca}$ separations.

local anisotropy is sufficiently short-ranged. I checked this point by calculating from the original (all-atom) HRMC configurations the orientational correlation functions

$$\phi(r) = \langle P_2(\mathbf{S}_i \cdot \mathbf{S}_j) \rangle_r, \quad (2.7)$$

where $P_2(x)$ is the second-order Legendre polynomial, and $\mathbf{S}_i$ and $\mathbf{S}_j$ are vectors parallel to a suitable local axis (e.g., C_3, C_2) of molecules i and j , separated by distance r [272]. The function $\phi(r)$ is essentially featureless—for both carbonate and water—for all but the very shortest intermolecular approaches, which occur for distances smaller than the nearest-neighbour $\text{Ca} \cdots \text{Ca}$ separation [Fig. 2.7(c)]. Hence the anisotropy of individual molecules is relevant only at a length-scale smaller than that over which the effective interactions between Ca^{2+} -ion pairs operate.

The double-well form of $u_{\text{Ca}}(r)$ is important for two main reasons. First, it explains why simple hard-sphere or LJ potentials fail to reproduce the experimental $g_{\text{Ca}}(r)$: if tuned to capture the 4 Å nearest-neighbour correlation, such potentials predict a second maximum just below 8 Å that is not observed in experiment. Second, multi-well potentials are known to support very complex phase behaviour (I will discuss this in more detail in Section 2.4.3) [273–278], suggesting a qualitative explanation for the complexity of ACC.

To test this hypothesis of a link between effective potential and structural complexity, I sought to map $u_{\text{Ca}}(r)$ onto a suitable multi-well potential for which the corresponding theory is already well established. I found that a double-well LJG interaction [279] described the observed functional form surprisingly well, so long as an additional broad, repulsive, Gaussian term was included to help capture the local maxima between and beyond the two minima of the LJG function. The behaviour of the LJG potential is characterised by three parameters— ϵ, r_0, σ —which describe, respectively, the depth, position, and width of the Gaussian well relative to the LJ component [274, 278]. A least-squares fit to $u_{\text{Ca}}(r)$ gave the values $\epsilon = 4.1$, $r_0 = 1.4$ and $\sigma = 0.14$ [Fig. 2.7(a)]; note that r_0 is particularly well defined because

it is closely related to the ratio of the $\text{Ca}\cdots\text{Ca}$ separations resulting from the two different carbonate bridging motifs ($\simeq 6 \text{ \AA} / 4 \text{ \AA}$).

To confirm that the combination of $g(r)$ inversion and potential parameter fitting does indeed result in a meaningful effective pair-potential, I carried out both direct MC and MD simulations (to confirm the results do not depend on the chosen simulation method), driven by the parameterised LJG potential.

The corresponding pair correlation function matches closely the HRMC $g_{\text{Ca}}(r)$, as expected [Fig. 2.8(a)], and the resulting structure [Fig. 2.8(b)] resembles the Ca distribution in the fully atomistic model [Fig. 2.8(c)]. The similarity between coarse-grained LJG-driven and HRMC Ca distributions can be quantified beyond pair correlations. The Voronoi cell volumes [Fig. 2.8(d)] and cell face counts [Fig. 2.8(e)], ring statistics [Fig. 2.8(f,g)], and a mixed radial/angular three-body distribution function, $g_3(r, \theta)$ [280], [Fig. 2.8(h,i)] are all almost identical for the HRMC and LJG-based calcium distributions. The Voronoi cell volumes and cell-face counts (neighbours) represent the local packing environments of Ca particles. These properties are (of course) strongly correlated to the density, which is equal for all configurations by construction. I note in particular the asymmetry to the Voronoi cell volume distributions, which is indicative of a common global structural heterogeneity. The ring sizes probe the structural similarity on longer length scales [Fig. 2.8(f,g)]. Calcium particles were designated as neighbours if their interatomic separation was less than $< 5.20 \text{ \AA}$, corresponding to the minimum after the first peak in $g_{\text{Ca}}(r)$. As a final check, I computed the three-body correlation functions, $g_3(r, \theta)$, which are proportional to the probability of finding a second neighbour, C, from a central atom, B, at a distance r_{BC} and angle θ_{ABC} , where atom A is the nearest neighbour of B [Fig. 2.8(h,i)]. The similarity between all of these higher-order correlation functions and properties lend further weight to the quality of the LJG description of effective $\text{Ca}\cdots\text{Ca}$ interactions.

A particular point of interest is that the MC configurations [Fig. 2.8(b)] contain

Ca-rich/-poor regions that are qualitatively similar to those observed in the original atomistic model [Fig. 2.8(c)]—the clear implication being that, at the density of ACC, an inhomogeneous Ca distribution is encoded within the effective $\text{Ca}\cdots\text{Ca}$ interaction itself.

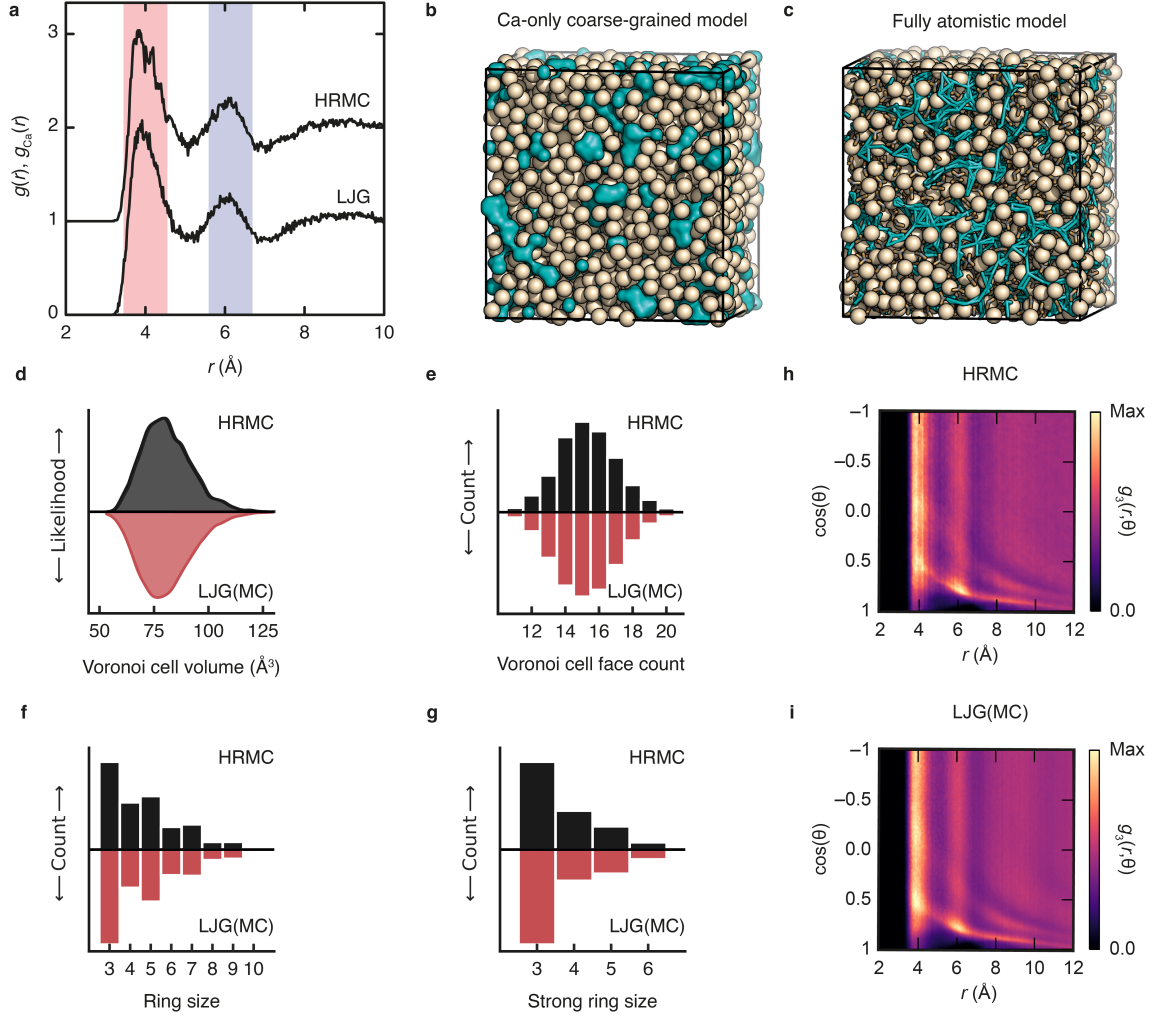


Figure 2.8: Comparing local-structure properties in the HRMC (black) and Monte Carlo-driven LJG (red) configurations. **a**, Ca-pair correlation functions $g_{\text{Ca}}(r)$ extracted from HRMC and LJG configurations. **b**, Representative MC configuration of Ca atoms (beige spheres) generated by the LJG potential, parameterised by the fit shown in Fig. 2.7(a). The Ca atoms are not uniformly distributed, but cluster to leave Ca-poor voids, shown as aquamarine surfaces. **c**, The fully-atomistic structure of ACC obtained using HRMC, included again for direct visual comparison with the LJG configuration in **b**. The **d** Voronoi volume, **e** cell face count, **f** ring size, and **g** strong ring distributions in both configurations are similar. The Ca-triplet three-body correlations $g_3(r, \theta)$ extracted from **h** HRMC and **i** LJG configurations.

2.4.3 Geometric frustration in ACC

From the perspective of soft-matter theory, much of the interest in the LJG potential lies in the complexity of phase behaviour which it supports [273, 274, 278, 281]. This complexity arises because of competition between the structure-directing effects of LJ and Gaussian components, which operate at different length-scales. When the positions of the LJ and Gaussian minima are related by a factor $1.2 \leq r_0 \leq 1.6$, there is particularly strong geometric frustration that results in a general resistance to crystallisation [274] and the emergence of many competing low-symmetry ground-state structures [278]. This behaviour extends across a wide variety of relative well depths $\epsilon \geq 1$ for the corresponding relative Gaussian width $\sigma = 0.14$ [278]. Hence the empirical parameters I have determined to be relevant to effective $\text{Ca} \cdots \text{Ca}$ interactions in ACC—*viz.* $r_0 = 1.4$, $\epsilon = 4.1$, and $\sigma = 0.14$ —locate the system in one of the most complicated parts of the LJG phase diagram [Fig. 2.9] [278].

In developing this mapping onto the LJG model, I assume that the additional broad Gaussian term we have included in the parameterisation does not strongly influence phase behaviour (note that the value of r_0 does not vary significantly

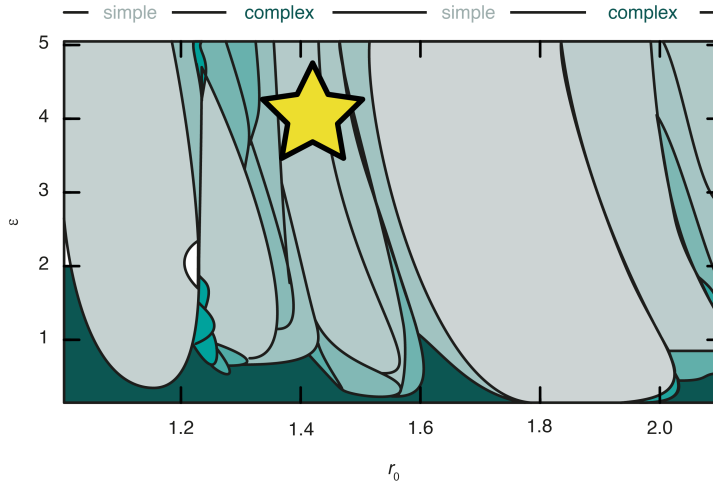


Figure 2.9: The complex phase space of the LJG multi-well potential. The approximate location of ACC effective LJG parameters, in the LJG phase space as reported in Ref. 278, is indicated with a star. Shaded regions correspond to the stability fields of different ground-state structures, and solid lines denote the approximate location of phase boundaries.

if the additional Gaussian term is omitted from the fit). Likewise, I cannot be certain of the effects of finite temperature and fixed density on LJG phase behaviour, since these have not yet been studied from theory. It is nevertheless a general phenomenon in frustrated systems that regions of strong geometric frustration are characterised by the existence of many competing ground states with suppressed ordering temperatures, above which the system is disordered but far from random [282, 283].

In the context of ACC, the implications are twofold. First, the different favoured $\text{Ca}\cdots\text{Ca}$ separations for carbonate-bridged Ca pairs—*viz.* 4 and 6 Å—are configurationally difficult to satisfy in a three-dimensionally periodic (i.e., crystalline) structure with the density of ACC and in the absence of anisotropic interactions. This rationalises, in general terms, why an amorphous form of CaCO_3 is so readily accessible. Second, the sensitivity of the LJG potential around $r_0 = 1.4$ hints—again in general terms—at why ACC might be directed to crystallise into a number of different polymorphs. Of course, as the water content of ACC varies during ageing of the material, both the effective $\text{Ca}\cdots\text{Ca}$ interactions and the bulk density must change, which makes extrapolation from the LJG effective potential of ACC difficult. Increased material density also strengthens the orientational correlations in molecular components, and so one expects the isotropic effective pair potential description to break down as crystallisation is approached. Nevertheless, at the density of ACC and in the absence of orientational order, the occurrence of an amorphous state is now more easily rationalised.

2.5 Conclusions

In this chapter, I have shown that incorporating accurate interatomic force fields within HRMC refinements allows for the generation of a high quality ACC structural model. The local atomic environments are in qualitative agreement with those obtained in MD simulations [122, 130], EPSR [254, 255], and solid-state NMR measurements [95], as would be expected by incorporating the interatomic potential

into refinements. Importantly, using potentials specifically developed for the aqueous CaCO_3 system allows the refinement to carefully balance the data-fitting and energies terms, and generate a model with intermediate-range order consistent with the experimental PDF.

Using the HRMC ACC model, I then provided a general explanation for the (meta-) stability of ACC. By extracting the effective interactions governing the $\text{Ca} \cdots \text{Ca}$ distribution and mapping them onto the phase space governed by the LJG pair potential, I showed that ACC is in a state of geometric frustration. This frustration arises from the two competing distances that Ca–Ca pairs bridge carbonate molecules, at 4 and 6 Å, which at the density of ACC are difficult to satisfy in a three-dimensionally periodic structure.

Chapter 3

Geometric Diversity in MOFs

3.1 Acknowledgements

This project was a collaboration with Davide Proserpio (University of Milan), Eugeny Alexandrov (Samara State Technical University), Vladislav Blatov (Samara State Technical University), and Alexander Shevchenko (Samara State Technical University). These collaborators validated the coarse-graining routines reported herein, and provided valuable advice and discussions about the data analysis. The findings discussed in this chapter are published in Ref. 284.

3.2 Introduction

The number of synthesised MOFs has grown rapidly in the last two decades and the structures are deposited in continually expanding databases. These databases are now reaching sufficient sizes (on the order of millions of structures) to provide an opportunity for extracting structure–property relationships based on data mining and ML, in principle [285–287]. Establishing these relationships, however, is a non-trivial task because there are multiple ways by which crystal structures can be represented, compared, and analysed. This challenge is particularly acute for MOFs where the diversity of both the metal centres (‘nodes’) and the organic linkers gives rise to considerable structural complexity [288–291].

One of the ways in which MOFs are commonly described is in terms of their topology; that is, the connectivity between nodes and linkers. Topological analysis routines are well-established and are implemented in automated computer packages, such as **ToposPro** [52] and **Systre** [292]. Topology descriptors have been found to be useful predictors for a number of material properties. For example, by considering the *deformability* afforded by a given topology (i.e., the ability of a framework to

distort geometrically without disrupting the net connectivity), one may predict the *rigidness* of a framework [293], the tendency to interpenetrate [294], and elastic properties [295, 296].

However, two MOFs with the same *topology* may have very different network *geometries*. By the latter term, I mean the spatial arrangement of local atomic environments in a MOF, including variations in bond lengths, angles, and longer-range ordering—all of which may affect material properties. For example, it was shown in Ref. 297 that two MOFs, zinc(ii) isonicotinate ($\text{Zn}(\text{ISN})_2$) and deuterium indium(iii) terephthalate ($\text{InD}(\text{BDC})_2$), that have the same doubly-interpenetrated **qtz** topology (denoted **2#qtz**), possess drastically different local geometries that lead to opposite mechanical responses with respect to heating. On heating, $\text{Zn}(\text{ISN})_2$ contracts in the (a,b) -plane and expands along the c -axis, whereas $\text{InD}(\text{BDC})_2$ expands in the (a,b) -plane but contracts along the c -axis.

Each established geometric descriptor, however, will cover only individual aspects of the structure; two MOFs with similar metal–linker distances might have different porosities, for example. Atom-density-based representations offer an alternative means of quantifying the geometric similarity of crystal structures [170, 298, 299]. In particular, the Smooth Overlap of Atomic Positions (SOAP) descriptor was developed initially in the context of fitting ML potentials [180]. Subsequently, it was shown how similarity kernels based on the SOAP formalism may be used to analyse the structural similarity for molecular and bulk periodic structures [300].

The development of ML models to predict MOF properties is becoming increasingly common [286, 287, 301]. For example, using the Computation-Ready, Experimental (CoRE) MOF database [302]—a collection of experimental MOF structures prepared for computational modelling—ML models have been developed to predict DFT-derived partial atomic charges with high accuracy and low computational cost [301]. These models represent MOF structures using a combination of one-hot encoding of chemical elements and atomic graphs encoding the bonds between atoms.

Beyond fundamental studies, ML models have also been developed to predict application-relevant properties of MOFs, such as band gaps [287]. These models trialled a range of descriptors, from simple compositional encodings that ignore MOF structure, to local geometric descriptions using SOAP [180], and graph-based representations [303]. Additionally, ‘structure maps’ were generated based on SOAP vectors to represent the diversity of local atomic environments, identifying regions in configurational space correlated with particularly small or large band gaps [287]. This approach can help identify the characteristics of an ideal MOF with a specific band gap, guiding the design of new materials.

However, it was noted that these structure maps were primarily influenced by the differences between organic linkers, overshadowing other critical features such as the metal centre. To address this, alternative descriptors were developed to explicitly account for metal chemistry, linker chemistry, and functional groups [304]. These descriptors separately encode the atomic graph of each component and incorporate heuristic properties such as Pauling electronegativity and nuclear charge. This approach enabled the identification of key features in different MOFs responsible for CO₂ and CH₄ adsorption properties, as well as providing insight into the inherent biases present when searching for new functional MOFs, such as exploring a limited diversity of metal species.

All of these approaches are inherently focused on the atomistic representation of MOFs and their associated properties. In this chapter, I demonstrate that a strategy based on structural coarse-graining and unsupervised machine learning can offer new insights into the geometric diversity of MOF structures and its relationship with network topology. Specifically, by focusing on AB₂ MOFs with fourfold-connected nodes and twofold-connected linkers, I assess how certain topologies support a broad range of local geometries, while others impose strict requirements on node and linker arrangements, thereby limiting these topologies to specific chemical units.

3.3 Methods

The main computational workflow used in this chapter is shown in Fig. 3.1. It can be separated into two stages: structure processing and structure analysis. The methodology is implemented in an open-source Python package, CHIC [257], which automates all of the steps, supporting application of the method to large datasets of complex materials.

3.3.1 Structure processing

The main structure-processing routines in CHIC are handled by a Python class called **Structure**. It uses Pymatgen [305] to parse and store information in the Crystallographic Information File (CIF) format. All subsequent structure-processing tasks are then callable as attributes of the **Structure** class, as follows.

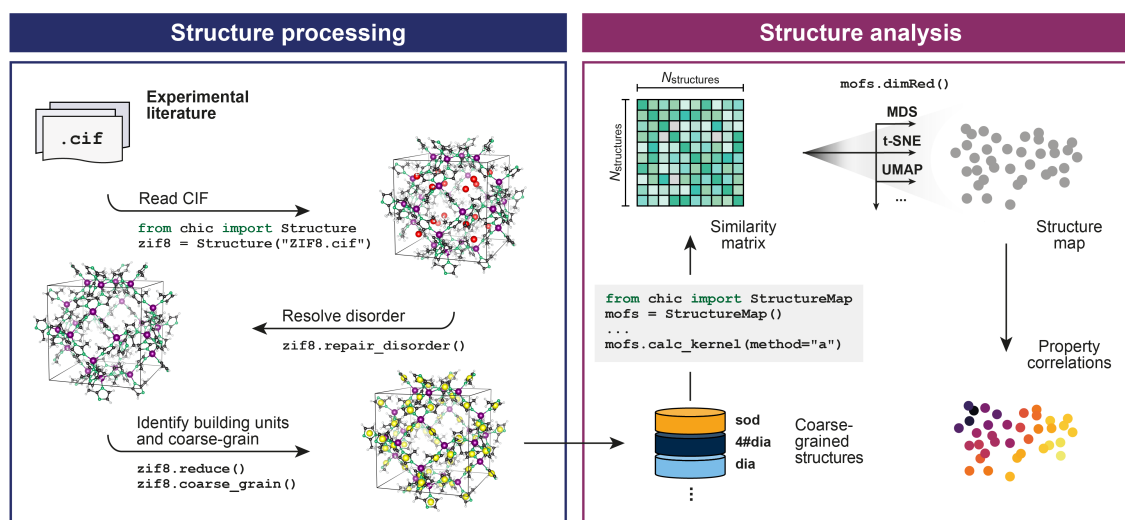


Figure 3.1: Methodology for processing and analysing MOF structure datasets using a cg-SOAP-based approach. *Left*, Structures from experimental literature may first need to be ‘cleaned-up’ for analysis. Here, ZIF-8 is used as an example: the H-atom disorder is resolved using average positions of half-occupancy sites, and guest atoms in the pores are removed. The structure is then reduced to its constituent building blocks, *viz.* 12 Zn^{2+} A sites (purple) and 24 methylimidazolate B sites; the latter are coarse-grained using connectivity graphs (here, identifying the five-membered imidazolate rings as the B-site centres; yellow). *Right*, The SOAP kernel is used to determine the similarity between all coarse-grained and rescaled structures in the database, and the result is visualised using dimensionality reduction. The resultant structure map (sketched here in a purely schematic way) may then be interpreted by correlating the locations of data points with structural and physical properties.

Sorting sites

To enable the automatic identification of individual nodes and linkers within the framework, the atomic species are classified into two categories: A (metal) sites and ‘other’ sites. By default, elements are sorted according to the IUPAC International Chemical Identifier classification [306]: metals are assigned to A sites, and non-metals to ‘other’. This study focuses on the geometric structure of the underlying nets, excluding non-framework species. Therefore, alkali and alkaline-earth metals (excluding Li and Be, which are known to form MOFs [307, 308]) are removed during sorting. B, Si, and P are sorted case-by-case, as they can function as either A or ‘other’ sites. For example, boron acts as an A site in boron imidazolate frameworks [309], but as a non-framework atom in BF_4^- anions. If an element is bound to O or N, indicating Lewis-acidic behaviour, it is assigned as an A site.

Removing disorder

Site disorder frequently occurs in MOFs and is often modelled crystallographically using split sites with partial occupancies [310]. To enable structure-processing tasks requiring discrete atomic positions, partially occupied or disordered sites are simplified using three algorithms. First, atoms of the same species within a given distance range (0–0.7 Å by default) are replaced by a single atom at the average position. The other routines address delocalised electron density from guest species in framework pores, typically modelled by clusters of fictitious oxygen atoms. These clusters are removed by identifying oxygen atoms that either only have oxygen nearest neighbours or have no neighbours within a set cut-off radius. Complete averaging of all disordered sites is unnecessary for the coarse-graining method; the minimum requirement is that individual molecules (e.g., organic ligands) do not overlap. However, unresolved disorder may affect the placement of the discrete place-holder atoms (e.g., by skewing the centroid of a given molecule).

Clustering

The success of the coarse-graining protocol depends on the correct identification of neighbours. This is a non-trivial task in the context of MOFs, which have atomic interactions on different length scales (for example, covalent, dispersion, hydrogen-bonding interactions). The default algorithm chosen for this work to construct the neighbour list, `CrystalNN` [311], uses Voronoi decomposition to assign weights to an initial list of neighbours within a hard cut-off radius, and then normalises and assigns probabilities to each unique weighting according to a smooth cut-off function. An adjacency matrix for each site type is constructed from this neighbour list, and decomposed into connected components [312]. This identifies all nodes and linkers (building units) in the structure.

Coarse-graining

The identified building units are then replaced by a single bead to generate the coarse-grained net. The placement of this bead is not uniquely defined but should reflect the fundamental interactions of the node or linker. The simplest method is to use the geometric centroid of all atomic positions in the building unit (not weighted by atomic masses). This approach, used in both crystal net determination [292] and coarse-grained molecular dynamics simulations [313], is used in this chapter. Alternative placements include the shortest path along the molecular graph and the geometric centroid of all rings within the building unit, the latter being particularly suitable for ZIF applications and is used in subsequent chapters in this thesis.

The building-unit connectivity is defined as the number of nearest-neighbour atoms of a different site-type connected to a given fragment. All fragments with connectivity greater than one (distinguishing non-framework species, such as solvents, from the B sites) are coarse-grained by placing a bead at the chosen bonding centre and removing all other atoms. The processed structures can then be output in CIF format, including the topology connectivity.

3.3.2 Structure analysis

Studying large, complex datasets requires generalised analysis routines. Here, I use the SOAP descriptor [180] to quantify the similarity between coarse-grained structural models.

Similarity metric

The SOAP kernel measures the similarity of pairs of atomic environments. Formally, for each atom, α , an atomic density, $\rho_\alpha(\mathbf{r})$, is constructed with a sum of Gaussians of broadness σ , centred on each neighbour, β , of α (as well as on α itself):

$$\rho_\alpha(\mathbf{r}) = \sum_{\beta} \exp \left[-\frac{(\mathbf{r}_\alpha - \mathbf{r}_{\alpha\beta})^2}{2\sigma^2} \right]. \quad (3.1)$$

The SOAP kernel is then defined as the overlap integral of any two neighbour densities, integrated over all three-dimensional rotations, $\hat{R}$,

$$k(\alpha, \beta) = \int d\hat{R} \left| \int \rho_\alpha(\mathbf{r}) \rho_\beta(\hat{R}\mathbf{r}) \right|^n d\mathbf{r}, \quad (3.2)$$

where the exponent is typically set to $n > 1$ to retain angular information [300]. In practice, it is computationally more efficient to expand the atomic density in a set of orthogonal radial basis functions and spherical harmonics up to a given $n_{\max}$ and $l_{\max}$. The resulting combination coefficients on their own do not yet ensure rotational invariance (because all spherical harmonics with $l > 0$ depend on the angular orientation) and are therefore collected into a power spectrum vector. The SOAP kernel may then be calculated by taking the normalised dot product of the two power spectrum vectors associated with each atomic environment, raised to an exponent ζ which serves to accentuate the distinction between the two environments [180].

As shown in Ref. 314 for elemental structures and in Ref. 256 for a range of inorganic and hybrid materials, geometric similarity may be assessed using SOAP

for uniformly rescaled structures, enabling direct comparison irrespective of characteristic A–B distances. I therefore rescale all coarse-grained structures to a uniform minimum $r(\text{A–B})$ distance.

To extend the similarity measure beyond comparing individual atomic environments, similarities between pairs of atoms in each of the crystal structures are calculated:

$$\bar{k}(\mathcal{A}_i, \mathcal{A}_j) = \frac{1}{N_{\mathcal{A}_i} N_{\mathcal{A}_j}} \sum_{\alpha \in \mathcal{A}_i} \sum_{\beta \in \mathcal{A}_j}^{N_{\mathcal{A}_j}} k(\alpha, \beta), \quad (3.3)$$

where α (β) runs over all atomic sites, $\mathcal{A}$, in the unit cell of structure i (j), respectively, and $N_{\mathcal{A}_i}$ ($N_{\mathcal{A}_j}$) denotes the number of $\mathcal{A}$ -site atoms in structure i (j), respectively. Variations of this method are also implemented where the atomic sites considered are restricted to a given site-type (e.g. A sites), thereby shifting the focus of the similarity analysis towards those particular sites. (In this case, information regarding the other site-types is still implicitly encoded through the neighbour densities.)

Dimensionality reduction and visualisation

To interpret (cg-) SOAP analysis results, the (dis)similarity between all structures in the database is often visualised as a two-dimensional projection (or ‘embedding’) [256, 300, 315]. In this chapter I compare three commonly used embedding schemes: multi-dimensional scaling (MDS) [316], t-distributed stochastic neighbour embedding (t-SNE) [317], and the uniform manifold approximation and projection (UMAP) [318].

To perform dimensionality reduction, the symmetric similarity matrix, $\mathbf{K}$, is computed using the per-cell averaged similarity (Equation 3.3), *viz.* $K_{i,j} = \bar{k}(\mathcal{A}_i, \mathcal{A}_j)$. A corresponding geometric distance matrix, $\mathbf{D}$, is then defined with elements

$$D_{i,j} = \sqrt{2 - 2K_{i,j}} \quad (3.4)$$

to satisfy the triangle inequality [300]. The role of an embedding scheme is then to

best preserve the distances encoded in $\mathbf{D}$, while reducing the data to two dimensions. The result is an embedding (or ‘structure map’) that carries the intuitive interpretation that structures that are similar appear close together, and structures that are dissimilar are further apart.

Here, I also define the A-site SOAP heterogeneity introduced in previous work [256], which measures the diversity of the A-site environments in a given structure using the SOAP similarity. It is calculated as

$$\mathcal{H}_A = \binom{N_A}{2}^{-1} \sum_{(\alpha, \beta) \in P} d(\alpha, \beta), \quad (3.5)$$

where P is the set of (unordered) pairs of distinct A-site environments in the structure, and $d = \sqrt{2 - 2k(\alpha, \beta)}$ is the SOAP distance [*cf.* Equation 3.4]. A value of zero means that all A-site environments are geometrically equivalent (up to the SOAP cut-off radius); a higher value indicates greater diversity.

Database details

Having now described the workflow, I detail the dataset of MOF structures used in this chapter. The dataset was curated for this work, and expands upon the AB_2 study reported in Ref. 256, now focusing on analysing the diversity in the wider set of AB_2 MOFs with twofold-connected ligands.

Primary data were selected from the sample that was prepared for Ref. 294, filtering for all structures in which the coordination formula was AB_2 and the ligand (B) was twofold-connected. The sample contained 1,160 crystal structures, to which 102 structures from the CSD 5.42 update 1 (Feb 2021) were added. The ‘Experimental literature’ to be processed were exported as CIFs from the CSD using the 1,262 entry RefCodes and processed unchanged.

3.3.3 Computational details

Code dependencies

CHIC imports functionality from `Pymatgen` (version 2020.12.31) [305], the `Atomic Simulation Environment` (ASE) (version 3.19.0) [319], and `NetworkX` (version 2.3) [320] for structure processing tasks. The SOAP implementation is imported from `Dscribe` (version 0.4.0) [321]. The `Scikit-learn` (version 0.21.3) [322] implementation of MDS [316] and t-SNE [317], as well as the stand-alone UMAP [318] implementation (version 0.4.2) are used for dimensionality reduction. The `Scikit-learn` [322] implementation of affinity propagation was used for cluster analysis.

SOAP parameters

SOAP vectors were computed using the polynomial basis functions implemented in `Dscribe`, with a radial cut-off of $r_{\text{cut}} = 2.5 \text{ \AA}$, smoothness of $\sigma = 0.2 \text{ \AA}$, and atomic neighbour density expansion up to $n_{\text{max}} = 10$, $l_{\text{max}} = 9$ (as in Ref. 256).

3.4 Results

3.4.1 Visualising geometric diversity

I start by showing the structure map produced by MDS in Fig. 3.2. In the context of MOFs, structures are generally reported with their topological identifier which uniquely identifies the underlying net. I therefore highlight the distribution of the three most commonly occurring topologies in the dataset—the diamond-like **dia** net (189 structures), the sodalite-like **sod** net (136 structures), and the four-fold interpenetrated **4#dia** net (116 structures)—to investigate the degree to which topological variation is represented in the cg-SOAP representation. Strikingly, all three topological families have structures dispersed over the map, with the **dia** topology being most widely distributed among the three (light blue). There are regions of overlapping points, particularly for **dia** and **sod** in the lower part of Fig. 3.2, indicating that topologically dissimilar MOFs might in fact show similarities

in terms of their geometric structure.

I further quantify the relative distributions of each topology in the map based on how far, on average, the structures of that group are from its respective centroid. These relative distributions are 0.77, 0.39, and 0.44 for **dia**, **sod**, and **4#dia**, respectively. For ease of comparison, I normalise the relative distributions such that the value for **dia** is unity (Table 3.1).

These results show that **dia** networks can be formed with a much larger variety of geometries compared to **sod**, in accordance with the conclusions derived from an analysis of the natural tiles (cages) [294]. The breadth of local geometries reflects a larger deformability of diamondoid structures and adaptability to building blocks with geometries spanning a wide range of volumes, lengths, and angles. This feature promotes the dominance of the **dia** topology in coordination networks [294]. The folding of the **dia** networks into interpenetrating arrays, however, significantly restricts the diversity of acceptable network geometries.

Cluster analysis can assist the interpretation of complex datasets by grouping



Figure 3.2: Geometric diversity within isorecticular groups of AB_2 MOFs. The graph shows a two-dimensional visualisation of the cg-SOAP-based structural distances using an MDS embedding: generally, the closer two points are, the more similar their coarse-grained and re-scaled geometric structures. The distribution of the three most commonly occurring topologies within the dataset (**dia**, **sod**, and fourfold-interpenetrated **dia**, denoted **4#dia**) is emphasised by lines that originate from the respective centroid. Dataset entries with different topologies than the three aforementioned ones are all represented by grey points.

data and identifying representative examples, from which patterns can more readily be identified. Affinity propagation is a clustering algorithm that views each data point as a node in a network, and recursively minimises the edge weights between nodes; the magnitude of each edge at a given time reflects the current affinity the point has for selecting the second point as its ‘exemplar’ [323]. Fig. 3.3(a) shows the **dia** structures (light blue data points in Fig. 3.2) now divided into four clusters, with the exemplar coarse-grained structures visualised. From these clusters, I investigate the distribution of structural properties in different regions of the map in order to appreciate the geometric diversity available within the **dia** topology.

In the context of MOFs, low metal densities are the simplest indication for the presence of void space: an important feature for catalytic applications [324]. I define the geometric density as the (unitless) density of the structures that have been coarse-grained and scaled (to unity minimum $r(\text{A-B})$ bond length). The *geometric* density is related to the *experimental* metal density by the characteristic framework bond length. I plot the distribution of the geometric density, average angular component of the Chau–Hardwick order parameter [325], S_g (a value of zero corresponds to ideal tetrahedral bond angles; a value of unity would correspond to

Table 3.1: Characteristics of commonly occurring AB₂ MOF topologies

Topology	Occurrences	Geometric density ^a (mean $\pm$ SD)	Relative distribution		
			MDS	t-SNE	UMAP
dia	189	125 $\pm$ 56	1 (<i>reference</i>)		
sod	136	80 $\pm$ 26	0.50	0.45	0.60
4#dia	116	261 $\pm$ 77	0.57	0.47	0.44
5#dia	95	271 $\pm$ 87	0.67	0.47	0.50
3#dia	88	206 $\pm$ 76	0.71	0.48	0.45
2#dia	79	182 $\pm$ 82	0.91	0.76	0.55
cds	41	221 $\pm$ 87	0.81	0.54	0.44
qzd	38	156 $\pm$ 15	0.12	0.18	0.26
6#dia	36	336 $\pm$ 137	0.74	0.55	0.47
3#dmp	33	292 $\pm$ 49	0.38	0.23	0.31

^aThe geometric density is here defined as $n(\text{A}) \times 1000 / V_{\text{scaled}}$, where $n(\text{A})$ is the number of A sites in the unit cell, and V_{scaled} is the volume of the scaled unit cell.

the extreme case where all four bonds are superimposed), and the average A–B–A angle in Fig. 3.3(b–d), respectively.

Figure 3.3(b) shows the distribution of geometric density values for each cluster. There is a relatively clear separation between each group, with cluster **2** (orange) containing notably more dense frameworks. It is interesting to note that clusters **2** and **4**, which have higher average geometric density, also seem to have a broader distribution across the MDS map, relative to the narrow distributions of clusters **1** and **3**. This could, in part, be explained by the relatively high deviation away from ideal tetrahedral bond angles about the A sites, as evident from the plot of the angular Chau-Hardwick tetrahedral order parameter, S_g [Fig. 3.3(c)]. Figure 3.3(d)

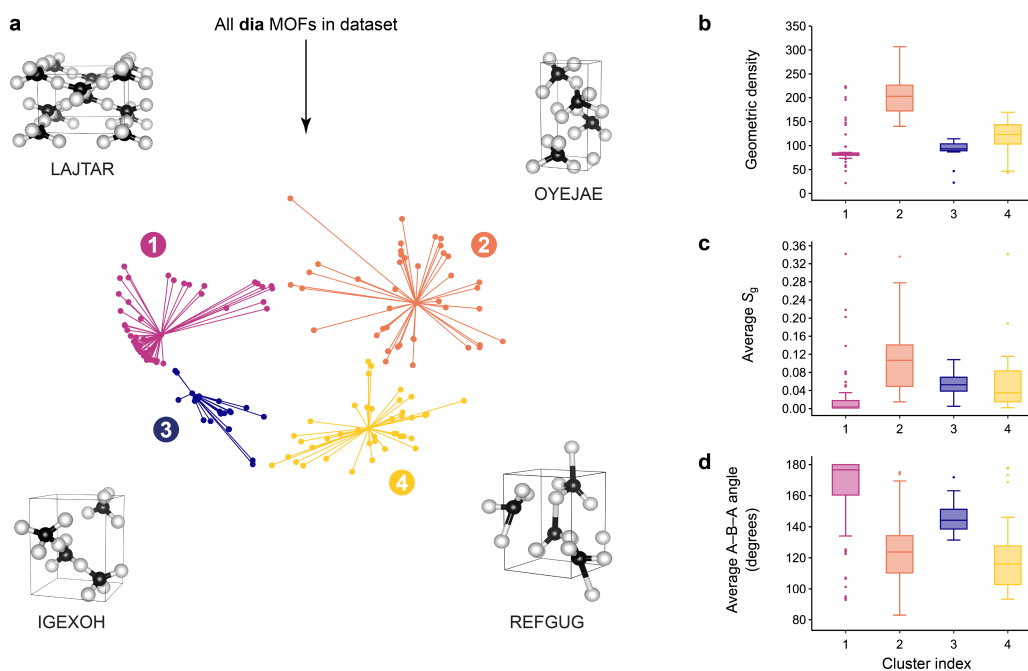


Figure 3.3: Geometric diversity within MOFs of dia topology. **a**, From the dataset characterised in Fig. 3.2, I isolate the **dia** entries, and analyse their distribution using the affinity propagation clustering algorithm [323]. For each cluster, the algorithm selects an ‘exemplar’ datapoint, and the corresponding coarse-grained structures and their CSD refcodes are shown. Distributions of local properties for each cluster are presented on the right-hand side: **b**, the geometric density; **c**, the average angular component of the Chau–Hardwick order parameter, S_g , and **d**, the average A–B–A angle. Box plots are drawn such that boxes range from the 25th to the 75th percentile, with the median indicated by a horizontal line; whiskers span ± 1.5 times the interquartile range, and points outside this range are plotted with circle markers.

shows the distribution of the average A–B–A bond angle; again, clusters **2** and **4** demonstrate broad distributions with lower average values.

Figure 3.3 emphasises that a single topological classification may give rise to a geometrically diverse set of structures; that is to say, one cannot necessarily predict the geometric features of a structure from its topological label alone. From Fig. 3.3, I therefore infer that the study of the latent *geometric* configuration space may provide insight into a different set of material properties (e.g., bulk modulus) compared to those which are accounted for by the topology (e.g., porosity [294]).

3.4.2 Embedding Schemes

The universal aim of dimensionality reduction algorithms is to capture meaningful structure in high-dimensional data when embedded into low dimensions. However, it is imperative to consider the algorithm methodology when interpreting the structure map. To illustrate this, I have visualised the dataset with three dimensionality reduction algorithms; namely, MDS [316], t-SNE [317], and UMAP [318]. The structure maps are shown in Fig. 3.4. My principal aim here is to demonstrate that the interpretation of the cg-SOAP representations is similar for different embedding schemes. In order to understand how each representation differs from one another, I colour-code the maps by the geometric density [Fig. 3.4(a)], and by the distributions of **dia**, **sod**, and **4#dia** topologies [Fig. 3.4(b)], and compare side-by-side how these properties are represented in the results of different embedding schemes.

All three embeddings demonstrate a strong correlation of the distribution of points with geometric density, and qualitatively similar trends in the relative distributions of topologies within each structure map [Table 3.1]. The detail, however, varies. MDS leads to a distinct region in the centre of the map where no structures are found. This perhaps suggests that this algorithm, which seeks to minimise a relatively simple loss function, is reaching its limits of usefulness for the size and complexity of the dataset considered here.

The ‘island-like’ features in both the t-SNE and UMAP representations—or more

specifically, the absence of islands in the MDS representation—further emphasise this point. The most prominent island (right hand-side of the t-SNE map and upper-right of the UMAP map) is a family of **qzd** MOFs, with the smallest relative distribution, among the ten most commonly occurring topologies in the dataset, for MDS and t-SNE [Table 3.1]. Manual inspection of the structures reveals a high frequency of similar CCDC reference codes (RefCodes). The RefCode system is designed to group together structures into families, such as the same compound having been crystallised and characterised under different conditions, or polymorphs of the same compound. There are two families of RefCodes within the **qzd** MOF cluster: LIWDEB (13 occurrences) and UKUVOL (21 occurrences) [326, 327]. All of these structures are one of two isomers of $\{\text{Cu}(\text{succinate})(4,4'\text{-bipyridine})\}_n$, isolated under different experimental conditions, and thus could be considered duplicates. The high geometric similarity between duplicates relative to the similarity between other structures in the dataset skews the representation towards creating an isolated

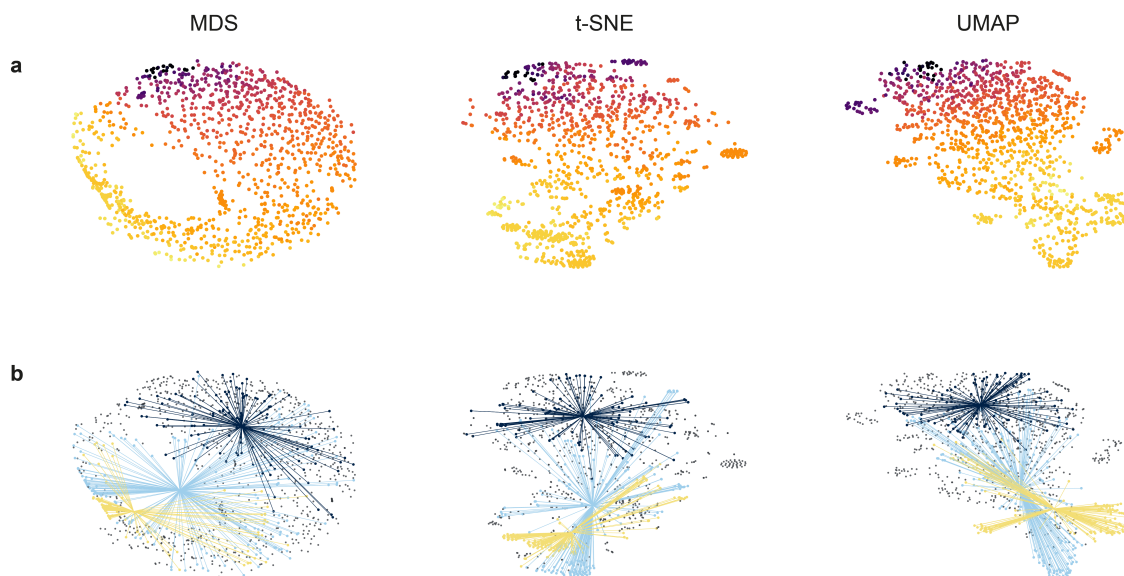


Figure 3.4: Visualising geometric diversity in MOFs using different embedding schemes. The plots compare the results of MDS (*left*), t-SNE (*centre*), and UMAP (*right*). For each embedding scheme, I show two-dimensional structure maps characterising the coarse-grained and scaled AB₂ MOF dataset, coloured **a**, by geometric density, and **b**, with the distribution of the three most commonly occurring topologies in the dataset, **dia**, **sod**, and **4#dia**, highlighting the respective centroids as in Fig. 3.2.

cluster.

3.4.3 Islands and duplicates

The occurrence of ‘islands’ of structures, separated out near the edge of the t-SNE and UMAP representations, requires a more subtle interpretation. For example, one of the islands to the left of the UMAP map corresponds to structures with **zni** and **coi** topologies (α and β polymorphs of $\text{Zn}(\text{Im})_2$, respectively), predominantly classified within the IMIDZB refcode family. ‘Duplicates’ cannot necessarily be identified as those with a common RefCode; however, cg-SOAP screening can help to automate this procedure. To illustrate this point, I isolate the IMIDZB family of structures from the database, create a structure map using t-SNE (because this algorithm generally achieves clearer clustering of data into distinct regions, albeit at the cost of meaningful inter-cluster distances), and analyse the representation using affinity propagation, as shown in Fig. 3.5.

The IMIDZB refcode family contains different polymorphs of $\text{Zn}(\text{Im})_2$ with different characteristic geometries and topologies, which therefore separate in the cg-SOAP map. By analysing the geometric diversity in this smaller configuration space, an automated ‘duplicate’ structure identification procedure is proposed. The **zni** topology is the densest, most stable crystalline polymorph of $\text{Zn}(\text{Im})_2$, and all structures of this connectivity are found in the same cluster, labelled as **1**. One structure (e.g., the cluster exemplar selected by the affinity propagation algorithm) could be taken as representative of this particular polymorph. The distribution of **cag** frameworks across two clusters (yellow and blue), however, is an example where frameworks with identical composition, connectivity, and space-group display different geometries. Whereas IMIDZB11 corresponds to the desolvated ZIF-4 framework under ambient conditions (298 K, 1 atm), IMIDZB12 and IMIDZB15 are the same framework after decreasing temperature (80 K) and increasing pressure (0.15 GPa), respectively, exemplifying the ‘breathing’ effect in the frameworks [75, 328]. With the change of external stimuli, the frameworks become more dense: the average

A–B–A angle decreases and S_g increases, corresponding to a lowered ‘tetrahedrality’ around the Zn (A) sites. On these grounds, it may be desirable to keep one structure from each cluster, i.e., IMIDZB10 (cluster **2**) and IMIDZB12 *or* IMIDZB15 (cluster **3**), in order to capture the geometric diversity fully.

More generally, one could propose an algorithm that classifies duplicates by considering the RefCode, topology, and cg-SOAP similarity as an automated approach that makes it possible to preserve the subtle geometric diversity that arises from varying experimental conditions. The screening of duplicates is expected to be helpful (and indeed required) for moving to very large databases in the future, as demonstrated in a recent analysis of DFT-optimised datasets of MOFs [329].

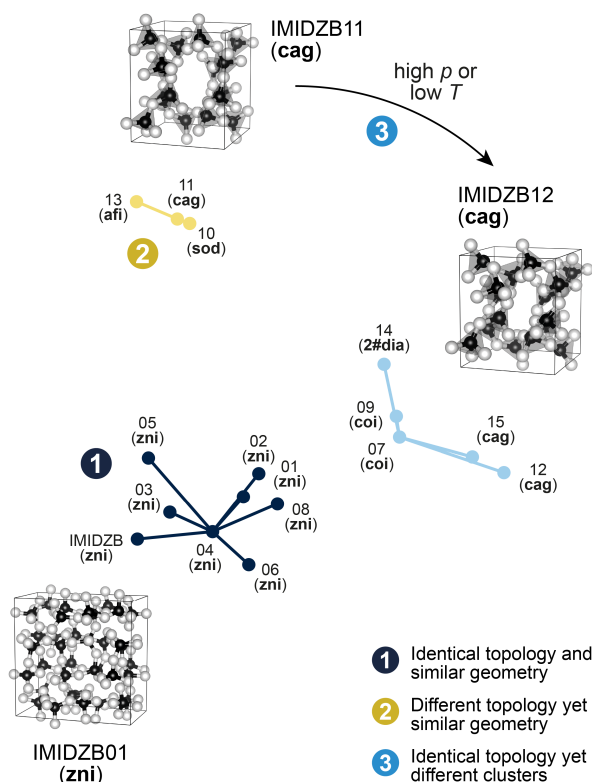


Figure 3.5: Using cg-SOAP analysis to identify relationships between individual CSD entries with a single RefCode. The example case here is given by the polymorphs of $\text{Zn}(\text{Im})_2$, with CSD RefCodes IMIDZB, suffixed with numbers representing a running index. The geometric diversity within the IMIDZB refcode family is visualised using t-SNE embedding and a cluster analysis by affinity propagation. The structures are labelled according to topology. The **zni** polymorphs cluster together; it could therefore be beneficial to choose a *representative* structure from this cluster. The spread of the **cag** topology into two distinct clusters illustrates the complexity of ‘duplicate’ detection, as discussed in the text.

3.4.4 Quantifying geometric diversity

The coarse-grained approach I have developed also enables quantitative investigation of local structural properties, and how they are distributed for different categories of structures (e.g., topology). For example, interpenetration is commonly found in MOFs, and it holds implications for the potential functionality of a given compound, because it is closely related to porosity [330]. Generally, porous materials minimise the energy of the framework through optimal filling of void space, and thus in cases where void space is of sufficient size, interpenetration may be observed. Controlling the degree of interpenetration has been explored using subtle changes to the synthetic methodology, such as varying reaction conditions, templating agents, and ligand design [331, 332]. Given the increasing number of MOF crystal structures reported, the question arises as to whether we can post-rationalise the extent to which the local geometry influences the tendency to interpenetrate.

In Fig. 3.6, I show the distributions of local properties for each structure, for different degrees of interpenetration of the diamond-like net ($Z = 1$ corresponds to **dia** MOFs, $Z = 2$ to **2#dia**, and so on). In Fig. 3.6(a), I show the distribution of A-site heterogeneity values [Equation 3.5]. Figure 3.6(b) and 3.6(c) illustrate the distributions of established local-property descriptors: namely, S_g and the average A–B–A angle, respectively.

I note that the majority of **dia** MOFs have locally homogeneous A sites, and this homogeneity does not appear to substantially depend on the degree of interpenetration [Fig. 3.6(a)]. Similarly, the distribution of the average S_g does not show a clear correlation with the degree of interpenetration [Fig. 3.6(b)]. There is, by contrast, a much stronger correlation with the A–B–A angle: as the degree of interpenetration increases, the average A–B–A angle tends toward 180° [Fig. 3.6(c)]. This makes intuitive sense when one considers that longer ‘rod-like’ ligands give rise to greater void space and therefore enable a greater degree of interpenetration, and is consistent with the synthetic approach of employing longer spacer ligands

to target higher degrees of interpenetration [333–335]. It may be inferred from these distributions that the A sites maintain a similar environment, irrespective of the degree of interpenetration, whereas the linker geometry plays a crucial role in determining this property.

Finally, I extend this quantitative analysis to all topologies that occur at least five times in the dataset, plotting the distribution of A-site heterogeneity values in the respective structures in Fig. 3.7.

The **qzd** structures have a distinctly narrow distribution of A-site heterogeneity

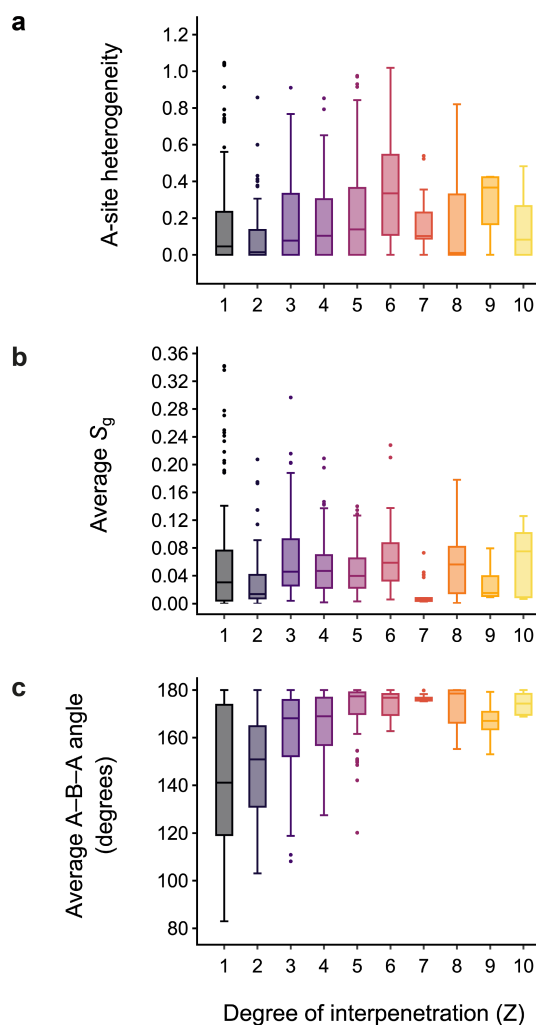


Figure 3.6: Distributions of geometric property indicators for increasing degrees of interpenetration, Z , of the dia-based MOFs in the dataset. a, The A-site heterogeneity (eq 3.5). **b,** The average angular component, S_g , of the Chau–Hardwick order parameter. **c,** The average A–B–A angle.

values, which reinforces the hypothesis that the separation from the main body of structures in the t-SNE and UMAP embeddings [Fig. 3.4] is a skewing of the visualisation as a result of duplicate structures. It is also interesting to note that the **zni** topology has a narrow A-site heterogeneity distribution at relatively high average value, which likely contributes to the separation of the **zni** structures.

Figure 3.7 highlights examples of topologies with particularly low (**qzd**) and particularly high (**4T13**, **pts**, and **2#pts**) A-site heterogeneity, for which illustrative coarse-grained and scaled crystal structures are visualised. The structures with **pts** and **2#pts** topology (which are related by increasing from a single to a two-fold interpenetrated net) all contain both tetrahedral and square planar geometries about different A sites, typically by combining Zn (tetrahedral geometric preference) with any of Ni, Cu, Pt, or Pd (square-planar geometric preference [336]). Some frameworks have Cu in both the tetrahedral and square-planar sites of the **pts** framework [337]. When combined with Au, Cu/Ag occupy the tetrahedral sites in the **2#pts** framework [338]. One might consider attempting to target these **pts**

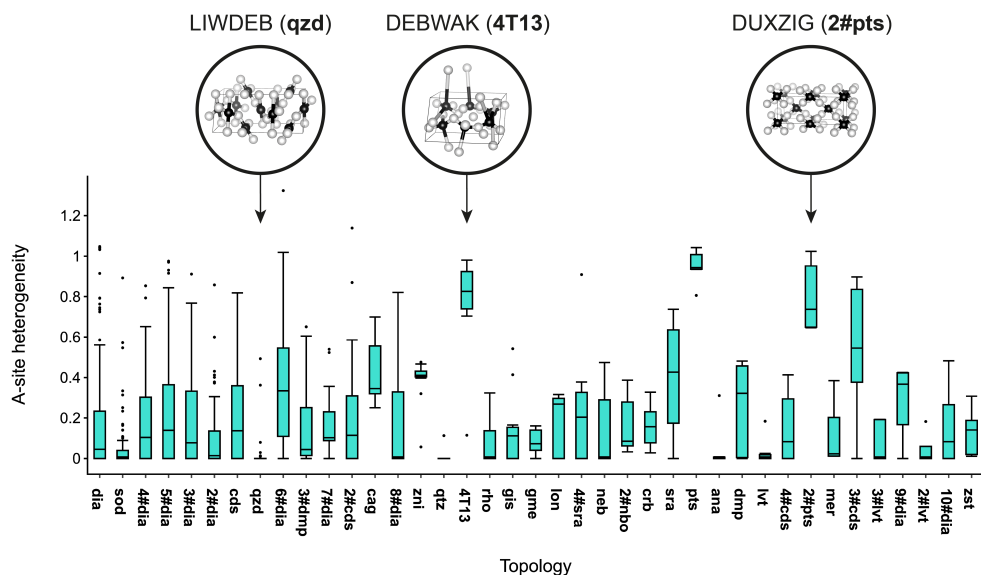


Figure 3.7: Distributions of A-site heterogeneity (Eq. 3.5) for all structures in the dataset for which the topology occurs ≥ 5 times. Three example MOFs are highlighted with coarse-grained structural representations, CSD refcodes, and the corresponding topology symbol, and are discussed in the text.

topologies, therefore, by selecting metals with the appropriate geometric preferences demanded by the framework.

The high degree of heterogeneity in the **4T13** frameworks can be ascribed to a tension between the differing lengths of two organic linkers, on the one hand, and the additional flexibility afforded by using two linkers, on the other hand (cf. DEBWAK in Fig. 3.7). The **4T13** frameworks have a short linker (e.g., isophthalate in DEBWAK) and a long-chain, flexible organic linker (e.g., *N,N'*-bis(pyridin-4-yl)-2,2'-bipyridine-5,5'-dicarboxamide in DEBWAK) [339, 340]. The resultant framework has one-dimensional helical chains that cross each other to create a two-dimensional molecular braid with geometrically distinct Zn sites. Automated, quantitative analyses such as those exemplified in Fig. 3.6 and 3.7 should be a helpful part of the methodology used in future work for understanding the geometric and structural diversity in databases of MOF structures.

3.5 Conclusions

In this chapter I have studied the structures of AB_2 MOFs containing a diverse set of twofold-connected organic linkers. By coupling a cg-SOAP approach to different embedding schemes, I visualised and analysed the geometric diversity in a database of MOF structures. With the aid of cluster analysis, the structure maps of the AB_2 MOF configuration space can be better understood.

Here, I have focused on clustering within the low-dimensional embedding, and demonstrated how the location of structures in the map was consistent with the grouping of structures with similar structural properties. I anticipate that the methodology and CHIC code developed in this chapter will be useful in visualising, analysing, and understanding the geometric diversity in larger MOF datasets.

Chapter 4

An ML potential for ZIFs

4.1 Acknowledgements

I am grateful to Daniel Thomas du Toit (University of Oxford) for advice on fitting ACE ML models and for providing the `XPOT` code for optimising the ACE model hyperparameters. I thank Linus Erhard (Technical University of Darmstadt) for helpful discussions regarding the preparation of α -SiO₂ structure models. The methodology for generating hypothetical ZIF structures, and a subset of the ‘iter-0’ training configurations described in this chapter were used and published in Ref. 341.

4.2 Introduction

MOFs have been proposed for a variety of applications due to their high surface areas, tunable pore sizes, and adjustable internal surface properties [342]. Most MOF studies have focused on highly ordered crystalline structures, partly because these structures can be readily solved using common laboratory diffraction equipment and crystallographic analysis. This is evidenced by large databases containing millions of MOF crystal structures [174]. However, the inapplicability of crystallographic techniques and the necessity for central facilities to perform total scattering experiments up to high momentum transfers have left the structural characterisation of amorphous MOFs restricted to a small number of studies [343, 344].

Among these, α -ZIFs have emerged as the most widely studied amorphous MOFs [Section 1.1.3] [7, 76, 78, 345]. ZIF-4, one of the 18 crystalline Zn(Im)₂ polymorphs [66, 75], can be amorphised *via* several different routes, including heating, ball milling, melt-quenching, and applying high pressures [343, 344]. In most cases, the crystalline topology breaks down, giving way to topologically disordered α -ZIF forms [76].

To gain atomistic insight into the structure of *a*-ZIF, various modelling approaches have been employed, each offering different advantages [Section 1.2]. Here, I briefly review two simulation methods which are particularly relevant to this chapter; namely, RMC and AIMD.

One of the first approaches used was RMC modelling, applied to study the *a*-ZIF form that emerges when ZIF-4 is heated to 300 °C [Section 1.2.4]. By adapting a CRN model of *a*-SiO₂ to create a topologically-disordered ZIF for refinement, the final RMC configuration successfully reproduced key aspects of the experimental X-ray and neutron total scattering data [76]. This finding established the network topology in *a*-ZIF as disordered. Although the *a*-SiO₂ CRN likely had different ring statistics compared to *a*-ZIF (notably, ZIF-4 and ZIF-**zni** networks both contain a relatively large number of four-membered rings, whereas the *a*-SiO₂ model contained none), RMC modelling did not exhibit sufficient sensitivity to these nuances. It should be noted, however, that the fit to the X-ray data, which is predominantly influenced by Zn-based correlations, was significantly worse than the fit to the neutron data, especially regarding the intensity of the first peak.

In contrast, AIMD simulations, despite being computationally intensive, have provided critical insights into the mechanisms of melting and vitrification in ZIF frameworks with high chemical accuracy [Section 1.2.2] [80, 159, 161]. These simulations highlight the critical role of transiently under-coordinated zinc sites in facilitating topological rearrangements. These sites are key features of the ZIF liquid state, as evidenced by qualitative comparisons of the computed and experimental X-ray total scattering functions during the melting process [80]. Upon vitrification, the ZIF glass structures exhibit high degrees of local order, with bond length and bond angle distributions closely resembling those of their crystalline counterparts [161]. These simulations also suggest a shift towards topologically-disordered glass structures with broadened ring size distributions relative to the crystalline distributions, however, the small system sizes used limit the statistical accuracy of the

models for analysing intermediate-range structural ordering.

To combine the strengths of both RMC and AIMD approaches, this chapter explores the coupling of HRMC refinement with ML potentials. The motivation for using an ML potential is their suitability for describing diverse chemical environments, such as those encountered when ZIF frameworks rearrange, while maintaining high chemical accuracy by training on energies and atomic forces computed using DFT [166]. At the start of this project, no ML potentials were available for ZIFs, prompting the development of one in this thesis. As discussed in Section 1.2.2, three ZIF ML potentials have since been published demonstrating the suitability of ML potentials for this task [192–194].

In this chapter, I discuss my approach for developing an ML potential for ZIFs. There are two distinguishing features of the approach taken here, compared to those in Refs. 192–194. First, I use databases of AB_2 networks with a wide diversity of topologies to generate hypothetical ZIF frameworks. This ensures comprehensive coverage of possible configurations that could be encountered when using the ML potential to model the structure of *a*-ZIF. Second, I introduce an HRMC-based active learning protocol that iteratively adds configurations refined to experimental data to the training dataset. This approach enhances the accuracy and robustness of the ML potential by ensuring that the diverse environments encountered during HRMC are well-sampled, thus making HRMC simulations using the ML potential more reliable.

The chapter is structured as follows. I will first provide an overview of the ML potential fitting methodologies used [Section 4.3]. Although aspects of this methodology rely on a description of the components of the training dataset to which the models are fitted, I will delay the description of the dataset until Section 4.4. This is because the training dataset developed in this chapter is a key result of this work. Throughout the Methods section, I will cross-reference the relevant results section to clearly link the fitting methodology with the relevant dataset components.

In the remainder of the Methods section, I will describe the DFT calculations used to obtain the dataset labels (i.e., the energies and atomic forces to be ‘learnt’ by the ML model), as well as the HRMC and MD simulations, and topology analysis used in this chapter. In the Results section, I will first provide a comprehensive overview of the entire dataset, followed by a description of the individual steps taken to build it. To conclude the Results section, I will explain the optimisation process of the final ACE models, focusing on achieving both speed and accuracy.

4.3 Methods

4.3.1 ML potential fitting overview

I used two ML potential fitting architectures: first, the Gaussian Approximation Potential (GAP) framework [166, 169, 180]; second, the Atomic Cluster Expansion (ACE) framework [178, 346]. GAP was chosen originally because it generally achieves more reliable MD behaviour in the earlier stages of ML training, compared to other fitting frameworks, which is particularly beneficial when building the training dataset from scratch—as I did here. The motivation for changing architecture was practical: ACE provided a speed-up that made the final simulations feasible [Section 4.4.8]. Rather than describing the full theoretical details of the ML fitting architectures themselves, I will concentrate on the key features that enabled the work discussed in this chapter.

First, I will discuss the functional form of the descriptors ($\mathbf{q}$) used in each framework, followed by an explanation of the hyperparameter selection process. In the context of ACE, I will also introduce the concept of uncertainty quantification for atomic environments, in terms of the extrapolation grade, γ . I use γ in order to select structures that are not yet well-represented in my training dataset during HRMC-iterative training, so that they can be added.

4.3.2 GAP model fitting

Initial fits were performed with GAP, a kernel-based method that evaluates the resemblance between atomic environments using similarity measures [166, 169, 180]. This allows for the assessment of how closely a new environment matches those in the training dataset. Many GAP models employ a combination of descriptors, with the relative weights, δ , of each descriptor being crucial hyperparameters.

GAP model hyperparameter selection

After finalising the iter-0 database, I fitted the first GAP ML potential, initiating the process of hyperparameter optimisation. I incrementally added higher body-order descriptors, setting the weight, δ , for each descriptor as the root-mean-square error (RMSE) of the energy per atom of the model before incorporating that specific body-order descriptor [347]. In this way, δ are scaling parameters, and each corresponds to the distribution of energy contributions that a given interaction term (descriptor) has to represent. The energy error was evaluated on the training dataset.

The local atomic energy [*cf.* Equation 1.7] corresponding to each descriptor $d \in \{\text{two-body, three-body, SOAP}\}$ is given by a linear combination of kernel functions [169]

$$\epsilon^{(d)}(\mathbf{q}^{(d)}) = \sum_{t=1}^{N_t^{(d)}} \alpha_t^{(d)} K^{(d)}(\mathbf{q}, \mathbf{q}_t), \quad (4.1)$$

where t denotes one of N_t training configurations $\mathbf{q}_t$ each of which attains a weighting coefficient α_t during fitting, and K is a covariance kernel which quantifies how similar the input configuration $\mathbf{q}$ is to the t^{th} training configuration $\mathbf{q}_t$.

For two- and three-body descriptors, a squared exponential kernel is used

$$K^{(d)}(\mathbf{q}_i, \mathbf{q}_t) = \exp \left[-\frac{1}{2} \sum_{\xi} \frac{(q_{\xi,i}^{(d)} - q_{\xi,t}^{(d)})^2}{\theta_{\xi}^2} \right] \quad (4.2)$$

where ξ is an index running over the components of the descriptor vector $\mathbf{x}^{(d)}$, and θ is a hyperparameter to be chosen.

Initially, I created a two-body-only model with all radial cut-offs, r_{cut} set to 2.4 Å, which included the first and second nearest neighbour atoms. This cut-off was later increased to 6.0 Å for GAP-iter-4 during iterative refinement, as this adjustment improved the numerical accuracy of the model (discussed further in Section 4.4.7).

Next, I incorporated intramolecular three-body descriptors: {H, C, N}, {C, N, N}, {H, C, C}, and {C, C, N}, with $r_{\text{cut}} = 1.8$ Å. Each set of curly braces denotes a unique descriptor, including all triplets of the specified species within the cut-off distance. Additionally, I added intermolecular three-body descriptors with $r_{\text{cut}} = 2.2$ Å: {C, N, Zn} and {N, N, Zn}.

Finally, I included a many-body SOAP descriptor [*cf.* Section 3.3.2] for each chemical species [180]. I used the **TurboSOAP** implementation of SOAP [348], which enabled the optimisation of the per-species smoothing parameters, σ_{at} . I briefly summarise the pertinent features of SOAP here. SOAP starts with a neighbourhood density of a given atom i , defined as [169]

$$\rho_i(\mathbf{r}) = \sum_j \exp \left[-\frac{(\mathbf{r}_i - \mathbf{r}_{ij})^2}{2\sigma_{\text{at},\mu}^2} \right] \times f_{\text{cut}}(r_{ij}) \quad (4.3)$$

where the sum is over neighbouring atoms j , and the cut-off function, f_{cut} , goes smoothly to zero at r_{cut} over a characteristic width, r_{Δ} . The parameter $\sigma_{\text{at},\mu}$ controls the smoothness of the potential, and can be set per species μ , for each SOAP descriptor. For example, a SOAP descriptor centred on zinc atoms which includes zinc, carbon, and nitrogen atoms in the neighbourhood atomic density, $\sigma_{\text{at,Zn}}$, $\sigma_{\text{at,N}}$, and $\sigma_{\text{at,C}}$ can be tuned independently. The neighbour density is expanded into a local basis of orthogonal radial basis functions, g_n and spherical harmonics, Y_{lm} ,

$$\rho_i(\mathbf{r}) = \sum_{nlm} c_{nlm}^{(i)} g_n(r) Y_{lm}(\hat{\mathbf{r}}), \quad (4.4)$$

from which the expansion coefficients $c_{nlm}^{(i)}$ are used to form the power spectrum

$$p_{nn'l}^{(i)} = \sqrt{\frac{8\pi^2}{2l+1}} \sum_m \left(c_{nlm}^{(a)} \right)^* c_{n'lm}^{(a)}. \quad (4.5)$$

The many-body SOAP descriptor vector, $\mathbf{q}_i^{(\text{MB})}$, is defined as the normalised components of $p_{nn'l}^{(i)}$, up to a given n_{max} and l_{max} for the truncation of the radial and angular basis functions, respectively. The kernel function is then

$$K^{(\text{MB})}(\mathbf{q}_i^{(\text{MB})}, \mathbf{q}_t^{(\text{MB})}) = \left| \mathbf{q}_i^{(\text{MB})} \cdot \mathbf{q}_t^{(\text{MB})} \right|^\zeta, \quad (4.6)$$

where ζ accentuates the distinction between different atomic environments.

The values of each $\sigma_{\text{at},\mu}$ were determined for each species μ by varying each parameter in turn on a linear scale between 0.1 and 1.0 Å with a step size of 0.05 Å, keeping all other $\sigma_{\text{at},\mu}$ fixed, and selecting the value that resulted in the smallest energy training error.

The MPI-parallelised version of `gap_fit` was used for fitting the final GAP models (i.e., after the MD iterative training described in Section 4.4.3) [169, 180, 349, 350]. This enabled the distribution of a total memory requirement of 15 TB across 32 nodes (4,096 CPU cores) of the ARCHER2 UK national supercomputing facility [351]. These final fits are computationally very demanding. For this reason, the hyperparameters could no longer be optimised.

GAP model sparsification

During the GAP model fitting procedure, rather than using all N_{total} atomic environments in the dataset, N_t representative training points are selected for each descriptor to describe the dataset [$N_t^{(d)}$ in Equation 4.1]. This process is referred to as ‘sparsification’ [162]. I used different numbers of representative points from various parts of the dataset to reflect the diversity of atomic environments in each section and the relative importance of those configurations for the domain-application.

For sparsification of the two- and three-body terms, a uniform grid of N_{sparse}

basis function locations was used [169]. For the SOAP kernel, sparsification was achieved using CUR decomposition, a matrix factorisation technique that selects a subset of representative columns and rows [352]. This method allows for setting different numbers of sparse points for various structure categories. Alongside the details of each category, if I specifically allocate above or below the average number of sparse points, I will provide the rationale behind the decision.

I show in Table 4.1 the final GAP fitting parameters selected during the process described in Section 4.3.2, as well as the sparsification categories used, and the number of representative sparse points for each descriptor (N_{sparse}). I will briefly describe these categories below, and indicate in which results section they are described in more detail.

Crystalline. All DFT geometry optimised training structures (and distorted variants) sampled along the hydrostatic energy–volume curve for ZIF-4 (high- and low-temperature polymorphs), ZIF-**zni**, and ZIF-**coi** [Section 4.4.2].

Table 4.1: The final GAP ML potential fitting parameters.

	2-body	3-body		TurboSOAP			
		Zn	Im	Zn	N	C	H
δ (eV)	1.0	0.087	0.087	0.053	0.053	0.053	0.053
θ	2.0	1.0	1.0				
r_{cut} (Å)	6.0	2.2	1.8	6.0	6.0	6.0	6.0
r_{Δ} (Å)				0.5	0.5	0.5	0.5
σ_{at} (Zn) (Å)				0.80	0.25	0.45	
σ_{at} (N) (Å)				0.30	0.65	0.30	0.25
σ_{at} (C) (Å)				0.25	0.35	0.50	0.25
σ_{at} (H) (Å)					0.60	0.35	0.25
$n_{\text{max}}, l_{\text{max}}$				8, 4	8, 4	8, 4	8, 4
ζ				4	4	4	4
Sparsification	Uniform	Uniform		CUR			
N_{sparse} (crystalline)				1,000	1,000	1,000	500
N_{sparse} (topology)				500	500	500	250
N_{sparse} (heat)				2,500	2,000	2,000	1,500
N_{sparse} (pressure)				1,500	1,000	1,000	500
N_{sparse} (liquid)				1,000	1,000	1,000	1,000
N_{sparse} (buffer)				1,000	1,000	1,000	1,000
N_{sparse} (total)	50×8	150×4	150×2	7,500	6,500	6,500	4,750

Topology. Hypothetical ZIF (*h*-ZIF) structures produced by decorating AB_2 networks [Section 4.4.2]. I allocated fewer SOAP sparse points to this category because, although it is the most topologically diverse subset, sampling different geometries accessible to various topologies [*cf.* Chapter 3], it was the least relevant set of configurations for specifically studying the amorphisation process.

Heat. High-temperature MD simulations, using the MOF-FF for ZIFs force field in iter-0 [Section 4.4.2], and different iterations of the GAP ML potential during MD iterative training [Section 4.4.3]. These configurations were allocated the most SOAP sparse points because typically they lead to more diverse, disordered local environments.

Pressure. MD simulations where the pressure was varied, using the MOF-FF for ZIFs force field in iter-0 [Section 4.4.2], and different iterations of the GAP ML potential during MD iterative training [Section 4.4.3]. I specifically allocated more SOAP sparse points for the zinc environments here because the pressure simulations explored varying network geometries, which affects the zinc tetrahedral environment the most.

Liquid. High-temperature MD simulations performed with the final iter-6 GAP ML potential [Section 4.4.3]. These configurations are distinct from those in the **Heat** category because the iter-6 GAP ML potential was sufficiently stable between 1500 and 1800 K to enable longer simulations whereby the ZIF framework degraded.

Buffer. Highly-distorted structures exhibiting unphysical behaviour, such as imidazolate rings breaking open [Section 4.4.3].

GAP model regularisation

In the context of GAP ML models, regularisation helps to control the ‘expected error’ of the input data by applying parameters that penalise overfitting, thus ensuring the model generalises well to unseen data. Regularisation parameters can be specified for different categories of structures, as shown in Table 4.2.

The regularisation parameters σ_{energy} , σ_{force} , and σ_{stress} represent the expected

uncertainty in the energy, force, and stress predictions, respectively. Lower values indicate higher confidence in the accuracy of the input data, while higher values suggest greater uncertainty and flexibility in fitting. For example, crystalline structures, which have well-defined properties, use lower regularisation values, reflecting higher confidence in the data. In contrast, more diverse categories like liquid and buffer use higher regularisation values, indicating a higher tolerance for variation and uncertainty in these data types.

4.3.3 ACE model fitting

ACE basis vectors

The final ML potential models were trained using the ACE framework [178, 346, 353]. ACE decomposes the system energy into body-ordered contributions, which are then summed into local atomic energies (i.e., ϵ in Equation 1.7). It does this by first projecting the atomic neighbour density onto one-particle basis functions, ϕ_{ν} (i.e., a single sum over neighbours, centred on one atom). The product of these projections containing ν factors gives a $(\nu+1)$ -body ‘ A -basis’ function. By averaging the A -basis function over the three-dimensional rotation group, a rotationally-invariant ‘ B -basis’ is formed. This is the key descriptor in ACE, which I shall henceforth refer to as the ‘ACE vector’. I now describe the relationship between the ACE vector and local

Table 4.2: GAP regularisation parameters.

Category	Regularisation Parameters		
	σ_{energy} (meV/atom)	σ_{force} (eV/Å)	σ_{stress} (eV)
Crystalline	0.2	0.002	0.002
Topology	5	0.2	0.2
Heat	5	0.2	0.2
Pressure	5	0.2	0.2
Liquid	5	0.2	0.2
Buffer	50	0.2	0.2

energy. First, atomic properties φ_i are expanded as follows

$$\varphi_i = \sum_{\nu} c_{\nu}^{(p)} \mathbf{B}_{i\nu}, \quad (4.7)$$

where $c_{\nu}^{(p)}$ are the expansion coefficients and $\mathbf{B}_{i\nu}$ are the ACE vectors [178]. The number of functions used in the expansion can control the *accuracy* of the representation.

ACE atomic properties

The atomic energy can be expressed as a function of P atomic properties:

$$E_i = \mathcal{F} \left(\varphi_i^{(1)}, \dots, \varphi_i^{(P)} \right), \quad (4.8)$$

where $\mathcal{F}$ is a general non-linear embedding function. This partitioning into different contributions can improve the convergence of the energy. For example, the Finnis–Sinclair potential [354] is expressed in the form:

$$E_i = \varphi_i^{(1)} + \sqrt{\varphi_i^{(2)}}. \quad (4.9)$$

Here, the linear term, $\varphi_i^{(1)}$, represents the repulsive core-core interaction, and the square-root term, $\sqrt{\varphi_i^{(2)}}$, represents the cohesive, many-body nature of metallic bonding. Incorporating additional atomic properties introduces *flexibility* into the functional form [187, 353].

By varying the number of functions used in the expansion of the basis vectors [cf. Equation 4.7], and the number of atomic properties [cf. Equation 4.8], in Section 4.4.8 I will explore different trade-offs between speed and accuracy in the final ACE models.

ACE model hyperparameter selection

Similarly to the GAP models, there are key hyperparameters to optimise [178, 353]. In practice, I found most of the default parameters implemented in **pacemaker** gave reasonable results [353]. For completeness, I describe the main hyperparameters for the ACE model fitting that I used for the initial ACE fitting, and during HRMC-driven iterative training.

As discussed above, the complexity of an ACE model is governed by the type of embedding function $\mathcal{F}$, the number P of atomic properties $\varphi_i^{(p)}$, and the number and body-order of basis functions and their corresponding expansion coefficients [353]. I used a Finnis–Sinclair type embedding ($P = 2$). The number of basis functions in **pacemaker** is specified in terms of number of functions per element, which I set to 1200, and the maximum body-order was set to $\nu = 6$.

For all species, the atomic neighbour density was determined up to a cut-off of $r_{\text{cut}} = 6.0 \text{ \AA}$, to maintain consistency with the GAP models. The one-particle basis functions, ϕ_ν , are expanded into a local basis of orthogonal radial basis functions and spherical harmonics [*cf.* the SOAP descriptor; Equation 4.4]. I used the simplified spherical Bessel radial basis functions multiplied by a cut-off function that goes smoothly to zero over the width $r_\Delta = 0.01 \text{ \AA}$.

The loss function used in **pacemaker** uses a weighted mean square error of the difference between energies and forces in the reference dataset and those predicted by ACE:

$$\mathcal{L} = (1 - \kappa)\Delta_E^2 + \kappa\Delta_F^2 + \Delta_{\text{coeff}} + \Delta_{\text{rad}}, \quad (4.10)$$

where the relative contribution of the errors of the energies Δ_E^2 and forces Δ_F^2 is weighted by the parameter κ . For each ACE model, I performed four consecutive fits using $\kappa = 0.95, 0.1, 0.01$, and 0.001 gradually increasing the relative weighting of the energies to the loss function. Δ_{coeff} and Δ_{rad} are regularisation contributions which ensure the smoothness of the PES and the radial basis functions, respectively. These were both set to a value of 10^{-8} . The importance of each structure to the

loss function can also be controlled using weighting policies. I used the energy-based weighting scheme, which assigns greater importance to structures closer to the energy convex hull.

ACE extrapolation grade

In Section 1.2.2, I introduced the DP-GEN active learning protocol [355] used in the fitting of a neural-network potential for ZIFs [194]. This protocol identifies structures that are poorly described by the ML potential by training an ensemble of randomly initialised ML potentials, each of which will give different predictions (e.g., energies, forces) for the dataset it was trained on, and selecting the structures with the greatest variability in predictions between the different models.

An alternative strategy, and the one implemented in the **pacemaker** within the ACE framework, is based on the D-optimality criterion [176, 356]. This gives an ‘extrapolation grade’, γ , which determines whether a given atomic environment can be regarded as an interpolation or extrapolation of the reference training dataset.

For each atom i of species μ in a training dataset, the ACE vector $\mathbf{B}_i$ is computed. If there are N_μ atomic environments for species μ , these vectors can be written in matrix form, $\hat{B}_\mu$, where each row i corresponds to $\mathbf{B}_i$. Then, a subset of vectors $\mathbf{B}_1^A, \mathbf{B}_2^A, \dots, \mathbf{B}_{n_\nu}^A$ are selected from $\hat{B}$ that have the largest absolute value of the determinant [357, 358]. These vectors form the ‘active set’ matrix, $\hat{A}_\mu$. This implies that all vectors $\mathbf{B}_i$ ($i = 1, \dots, N_\mu$) can be represented as linear combinations of the vectors in the active set

$$\mathbf{B}_i = \sum_{k=1}^{n_\nu} \gamma_k^{(i)} \mathbf{B}_k^A, \quad (4.11)$$

where $|\gamma_k^{(i)}| \leq 1$. The extrapolation grade for atom i of type μ is then defined as

$$\gamma_i = \max_k (|\gamma_k^{(i)}|) = \max \left| \mathbf{B}_i \cdot \hat{A}_\mu^{-1} \right|. \quad (4.12)$$

Active sets were generated using the full non-linear embedding of B-projections, as implemented in **pacemaker** [176, 353].

4.3.4 DFT reference data

I performed periodic DFT calculations using the Vienna Ab initio Simulation Package (VASP) [359–361]. The electron–ion interactions were described using the projector-augmented wave (PAW) method [362]. For the single-point calculations, the Kohn–Sham equations were solved self-consistently until the total free energy change and the band-structure-energy change converged to within 10^{-8} eV.

The Strongly Constrained and Appropriately Normed Semilocal Density (SCAN) meta-GGA functional was used [363]. The plane wave cut-off energy was set to 1000 eV, and partial occupancies were treated using Gaussian smearing with $\sigma = 0.2$ eV. Due to the large unit cell sizes considered, the Brillouin-zone sampling was restricted to the Γ -point.

For the DFT structure optimisations of the crystalline structures, the force tolerance for breaking the ionic relaxation loop was set to when the norms of all of the forces were smaller than $0.01 \text{ eV } \text{\AA}^{-1}$.

4.3.5 MD simulations

All MD simulations discussed in this chapter were carried out using the LAMMPS software [259]. For NVT simulations, I used a Nosé–Hoover thermostat [114, 115]. For NPT simulations, I added a Nosé–Hoover barostat [116]. Different timesteps, temperature damping constants, and pressure damping constants were used for different protocols. I will therefore give the appropriate details in each section.

4.3.6 HRMC refinements

In this chapter, I used HRMC refinements in order to generate training configurations for the ML potential fitting. This was to ensure atomic environments likely to be encountered when modelling disordered ZIFs are well represented in the training dataset. In the following, I describe these HRMC refinements, including any additional functionality implemented in the code that was not previously discussed in Chapter 2.

The HRMC refinements followed the same approach, and used the same custom code, described in Chapter 2. The cost function [Equation 1.10] for *a*-ZIF HRMC refinements balanced three contributions. The first contribution was the fit to the experimental neutron total scattering function, $F_N(Q)$, reported in Ref. 76. This data spans the range of momentum transfers $0.4 \leq Q \leq 50.0 \text{ \AA}^{-1}$, and was collected at 300°C using the GEM diffractometer at the ISIS spallation source [364]. I will describe the method for computing $F_N(Q)$ below. The second contribution was the fit to the experimental X-ray total scattering function, $F_X(Q)$, also using the data reported in Ref. 76. This data spans the range of momentum transfers $0.7 \leq Q \leq 8.0 \text{ \AA}^{-1}$, and was collected on an *a*-ZIF sample recovered from the neutron experiment. The X-ray data was collected at room temperature using a PANalytical Ag-source X’Pert Pro MPD lab diffractometer [365]. $F_X(Q)$ was calculated from the structure models as described in Section 2.3.3. The third contribution to the cost function was the system energy, computed using the most-recent iteration of the ACE ML potential [Section 4.4.4].

Neutron total scattering calculations

Here, I describe the key equations used to compute $F_N(Q)$, briefly explaining the principal differences between X-ray and neutron scattering. My aim here is to provide context for why they are complementary techniques [366–368].

Whereas X-rays interact with the electrons surrounding the atomic nuclei, neutrons interact with the atomic nuclei through the Strong Nuclear Force. Although the explicit form of Strong Nuclear Force is unknown, it is known to be of very short range, far shorter than the wavelength of the incident neutron. Considering then an incident neutron described by a plane-wave $\exp(i\mathbf{k} \cdot \mathbf{r})$ at position $\mathbf{r}$ interacting with nuclei j at position $\mathbf{R}_j$, the total potential of the system is well-approximated by a sum of Fermi pseudopotentials [369] centred at each nucleus:

$$\Phi_{\text{Fermi}}(\mathbf{r}) = \frac{2\pi\hbar^2}{m_N} \sum_j b_j \delta(\mathbf{r} - \mathbf{R}_j), \quad (4.13)$$

where m_N is the mass of a neutron and b_j is the scattering length of nuclei j . Then, the probability of a system transitioning from state λ to λ' , where the scattered neutron has initial and final wavevectors $\mathbf{k}$ and $\mathbf{k}'$ (giving a momentum transfer $\mathbf{Q} = \mathbf{k} - \mathbf{k}'$), respectively, can be expressed through Fermi's Golden Rule and is proportional to

$$\langle \mathbf{k}'\lambda' | \Phi_{\text{Fermi}} | \mathbf{k}\lambda \rangle = \frac{2\pi\hbar^2}{m_N} \sum_j b_j \exp(i\mathbf{Q} \cdot \mathbf{R}_j), \quad (4.14)$$

Formally, Equation 4.14 shows that the isotropic scattering of neutrons interacting with atomic nuclei depends on a single parameter, b , and does not depend on Q . The scattering length is a property of the respective nucleus and is therefore isotope-dependent, and does not have a clear pattern with respect to position in the periodic table. Neutrons are able to scatter from light elements just as well as (and in the case of hydrogen far more intensely than) heavier elements.

As the neutron has a spin of $\frac{1}{2}$, there is an additional effect that arises when the atomic nuclei have a non-zero nuclear spin I . The interaction depends on the relative orientation of the neutron spin (+ or $-$) relative to the nucleus, giving rise to two configurations with total spins $L = I + \frac{1}{2}$ and $L = I - \frac{1}{2}$ for b_+ and b_- , respectively. Generally, the neutron beam hitting a sample contains neutron with both + and $-$ spins, and the nuclei in the sample also have random orientation. Hence, the orientation of the neutron spin is arbitrary relative to the nuclear spin of the scattering nucleus, and both combinations, $L_{\pm} = I \pm \frac{1}{2}$ occur. The associated probabilities of each total spin is then given by

$$p_+ = \frac{I + 1}{2I + 1} \quad (4.15)$$

$$p_- = \frac{I}{2I + 1} \quad (4.16)$$

which can be used to define the average scattering lengths, $\bar{b}$, of a given atomic nucleus

$$\bar{b} = \left(\frac{I+1}{2I+1} \right) b_+ + \left(\frac{I}{2I+1} \right) b_-. \quad (4.17)$$

These average scattering lengths, $\bar{b}$, are tabulated for each isotope, and effectively replace the X-ray atomic form factors in Equation 2.5, giving the corresponding equation for the neutron total scattering structure factor as

$$F_N(Q) = \sum_{i,j=1}^n c_i c_j \bar{b}_i \bar{b}_j [A_{ij}(Q) - 1], \quad (4.18)$$

where A_{ij} are defined in Equation 2.4. A neutron total radial distribution, $G_N(r)$, is defined as [199]

$$G_N(r) = \sum_{i,j=1}^n c_i c_j \bar{b}_i \bar{b}_j [g_{ij}(r) - 1], \quad (4.19)$$

where $g_{ij}(r)$ are the partial pair correlation functions defined in Equation 2.3. I do not refine explicitly to the $G_N(r)$ function during HRMC simulations, however I do use it to parameterise the smoothing functions applied to $g_{ij}(r)$, described below.

Rigid body correlations

I kept the imidazolate molecules rigid and applied collective translations and rotations during the HRMC refinements. This approach reduced the degrees of freedom in the simulation, thereby decreasing the number of MC moves needed to attain equilibrium. The use of rigid bodies can lead to artificially sharp intramolecular peaks in the pair correlation functions [Fig. 4.1(a)]. This, in turn, can lead to the introduction of Fourier components in the corresponding structure factor $F(Q)$ that decay too slowly. This effect is most prominent in the case of neutron scattering, where the neutron atomic form factor does not decrease with increasing Q , unlike X-ray atomic form factors [Section 2.3.3].

Gaussian filters were therefore applied to the partial pair correlation functions, $g_{ij}(r)$, associated with intramolecular correlations to improve the agreement between the computed and experimental $G_N(r)$. I used the `gaussian_filter1d` function implemented in the `scipy` library [370]. The Gaussian filter smooths the $g_{ij}(r)$ functions by convolving them with a Gaussian kernel, where the width of the kernel is parameterised by its standard deviation, σ . The σ value controls the degree of smoothing: a larger σ results in broader smoothing, reducing the prominence of finer details in the function.

First, I optimised the Gaussian width of a filter applied to the for $g_{\text{CH}}(r)$ function. The σ value was determined by minimising the difference between the computed and experimental $G_N(r)$, specifically targeting the peak between 0.92 and 1.22 Å [Fig. 4.1]. This Gaussian filter was then applied uniformly to all C–H, N–H, and H–H partial PDFs. An example fit to $G_N(r)$ is shown in Fig. 4.1(b).

Next, a second Gaussian filter was applied to the C–N, C–C, and N–N partial PDFs. The σ value for this filter was optimised by minimising the difference between the computed and experimental $G_N(r)$ over the range 0 to 3 Å. An example fit with all Gaussian filters applied is shown in Fig. 4.1(c).

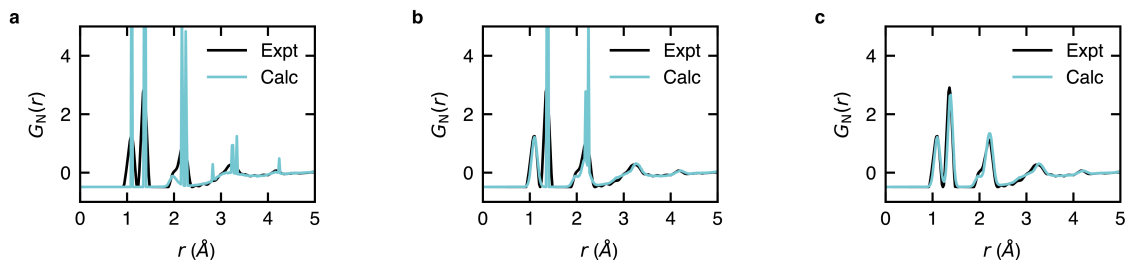


Figure 4.1: Rigid body effects in PDF calculations. **a**, Fixing the imidazolate molecules rigid creates sharp features in the PDF. **b**, The computed $G_N(r)$ after applying Gaussian filters to C–H, N–H, and H–H partial PDFs. **c**, The computed $G_N(r)$ after applying Gaussian filters to all partial PDFs involved in intramolecular interactions.

4.3.7 Topology analysis

I use topology analysis to investigate the structural diversity in the training dataset [*cf.* Chapter 2]. To determine the topology of each framework, the connectivity between zinc nodes and imidazoles must first be determined (i.e., the network graph). To construct graphs for ZIF structures, I first determined the Zn–Im connectivity using a Zn–N cut-off of $r_{\text{cut}} = 2.5 \text{ \AA}$ [80, 192]. I then assigned the edges of the graph as between zinc nodes that are linked by a common imidazole molecule.

As a simple proxy for topological disorder, I used the `CrystalNets.jl` software [371] to classify the nets of ZIF structures in my ML potential training dataset. If a corresponding periodic network could not be identified by `CrystalNets.jl` automatically—either suggesting the topology had not been found before, or, more likely, the framework was sufficiently distorted or topologically disordered for it to be given a label—then I classified the structure as disordered.

4.4 Results

4.4.1 Training dataset summary

My strategy for developing the training dataset began with broad coverage of the configurational space accessible to ZIFs, gradually specialising towards simulations tailored to the specific needs of the application. This progression is illustrated schematically in Fig. 4.2(a). The dataset development involved three main stages: generation of an initial training dataset (‘iter-0’); iterative improvement of the ML model using MD; and iterative training through HRMC refinements.

The overall composition of the final dataset is visualised in Fig. 4.2(b). In this two-dimensional structure map, the distance between any two points indicates the similarity between those configurations [*cf.* Chapter 3], as determined by the ACE B-basis vectors [178], averaged over all zinc environments in each structure. Structures that are close in the map are similar; structures that are far apart are dissimilar.

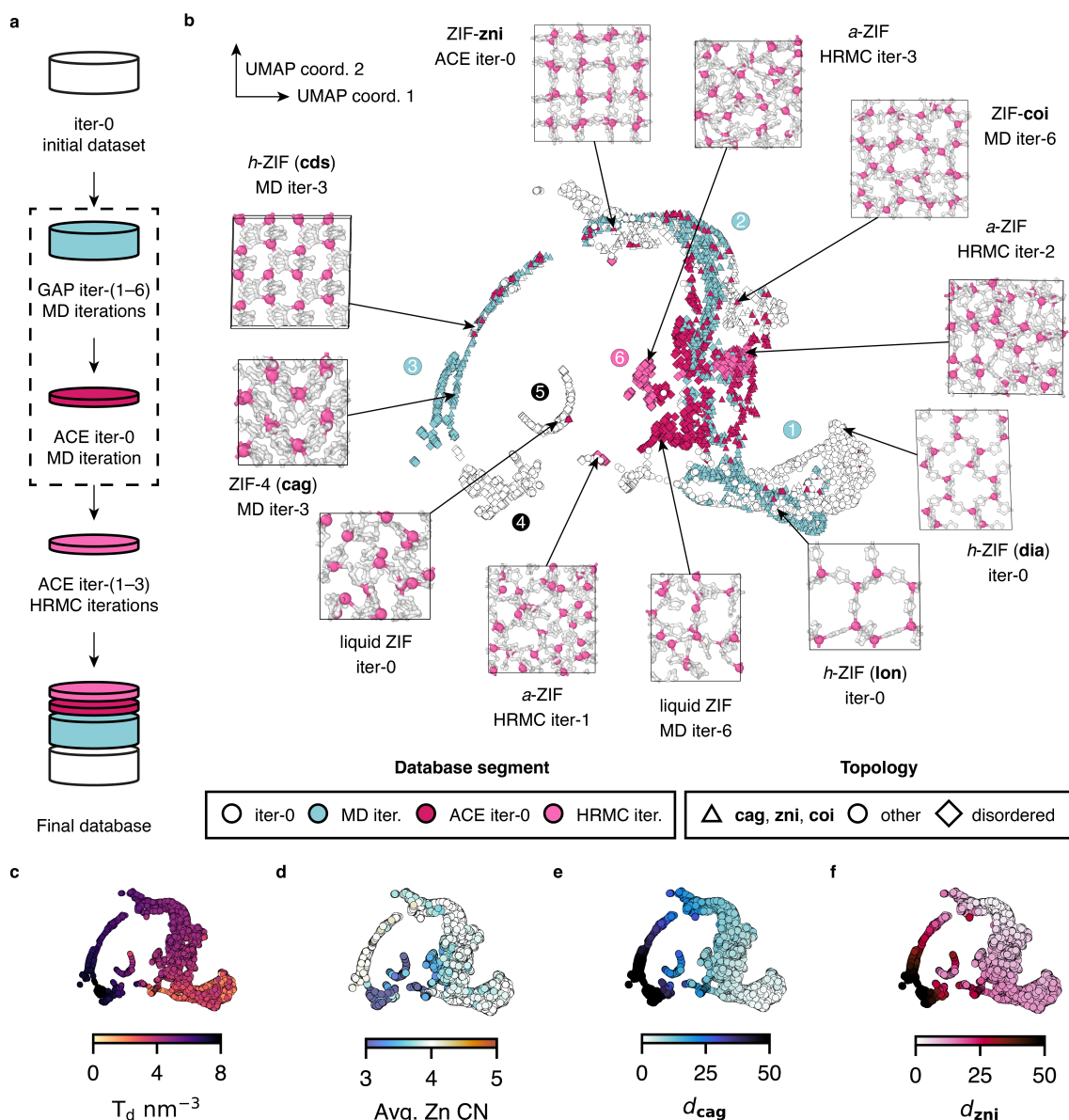


Figure 4.2: A structurally diverse ZIF dataset for fitting ML potentials. **a**, Three main components of the ZIF ML potential dataset: ‘iter-0’ configurations; configurations generated during MD-based iterations; and configurations generated during HRMC-based iterations. **b**, A two-dimensional structure map representing the dataset. The structure map was generated by computing the per-structure average ACE vector, centred on zinc atoms, for the entire dataset. Dimensionality reduction was then performed on these vectors using UMAP [318]. The map is colour-coded by dataset segment. The crystalline ZIF topologies most relevant to this work, **cag**, **zni**, and **coi**, are indicated with triangle markers. All other ‘detectable’ topologies are indicated with circle markers [371]. The remaining structures are considered topologically disordered and indicated with diamond markers. To illustrate the diversity of training configurations, the map is colour-coded with different structure metrics in **c–f**. From *left to right*: framework density, average zinc coordination number, the dissimilarity to the ZIF-4 (**cag**) structure, and the dissimilarity to the ZIF-**zni** structure. The dissimilarity metrics are computed as the Euclidean distance between the ACE vector of the reference structure and each structure in the dataset. A value of zero corresponds to an identical average zinc environment, up to the descriptor cut-off. Structure images were created using OVITO [372], with hydrogen atoms omitted for clarity.

The dataset is designed to encompass a wide range of configurations. Regions labelled (1) and (2) predominantly represent crystalline structures similar to ZIF-4 (**cag**) and ZIF-**zni**, respectively. These regions were particularly explored in the early stages of MD iterative training, as indicated by the high density of blue markers. Many crystalline hypothetical ZIF (*h*-ZIF) structures are also located in these regions. Region (3) includes dense frameworks, mostly explored during high-pressure iterative MD training. Region (4) also contains dense frameworks generated during iter-0 but is distinct from region (3) due to highly-distorted local environments, resulting from decorating networks with topologies that do not readily support the $\text{Zn}(\text{Im})_2$ structural motif. Region (5) contains ‘liquid’-like structures with many under-coordinated zinc sites. Region (6) hosts *a*-ZIF snapshots, with most configurations generated during HRMC iterative training. Interestingly, this disordered region of configuration space bridges regions (1) and (2), mirroring the transformation of ZIF-4 to ZIF-**zni** *via a*-ZIF [76].

I visualise the diversity in configurations by colour-coding the map with structure metrics in Fig. 4.2(c-f). The dataset spans a wide range of densities, encompassing all known experimental ZIF values [Fig. 4.2(c)] [68]. The highest-energy structures generally feature defects such as under-coordinated sites, as found in the liquid-like configurations in regions (4) and (5) [Fig. 4.2(d)]. ZIF-4 and ZIF-**zni** are important structural markers in the ZIF configuration space. I compute the dissimilarity of all structures with ZIF-4 and ZIF-**zni** in Fig. 4.2(e) and (f), respectively. The dissimilarity is defined as the Euclidean distance between the averaged zinc ACE vectors of each structure and the reference structure. Smaller distances (lighter shades) indicate more similar structures, and *vice versa*. Many disordered configurations, including *a*-ZIF representative snapshots, are found at intermediate dissimilarities between these two crystalline phases.

4.4.2 Iter-0 dataset generation

Crystalline structures

The training dataset included crystal structures of ZIF-4, ZIF-**zni**, and ZIF-**coi**. For ZIF-4, both high-temperature and low-temperature structures were used [328]. Reference structures for each polymorph were retrieved from the Cambridge Structural Database (CSD) [174]. For each structure, I performed DFT geometry optimisations, sampling 20 structures along the hydrostatic energy-volume curve.

Additionally, four training structures were generated for each configuration by applying random displacements to the atomic positions. This approach samples configurations above the minimum of the potential energy surface [60, 373]. Two sets were generated with random displacements drawn from a normal distribution with $\sigma = 0.01$ Å, and two further sets with larger displacements with $\sigma = 0.05$ Å.

Topologically-diverse hypothetical ZIFs

A large set of *h*-ZIF frameworks was generated from AB₂ framework structures [341], using a ‘decorate and relax’ procedure similar to that described in Section 1.2.3. These AB₂ structures were obtained from two existing databases of fully-connected AB₂ network structures: the coarse-grained AB₂ MOF nets described in Chapter 3, and the experimentally-realised zeolite framework types approved by the Structure Commission of the International Zeolite Association (IZA-SC) [374].

I will briefly describe the decorate and relax procedure I used to generate *h*-ZIFs. First, the connectivity between A and B sites was determined. Second, the structure was rescaled so that the minimum A–B bond length was 3.5 Å. Third, all A sites were replaced with zinc atoms and all B sites were replaced with imidazolate molecules. Imidazolate molecules were positioned such that the centre of the ring coincided with the B site coordinates, and the *C*₂ symmetry axis of the molecule aligned with the vector bisecting the A–B–A bond angle. Finally, the structure was geometry optimised using the MOF-FF for ZIFs empirical force field [133],

which parameterises the system energy in terms of bonding, angular, torsional, van der Waals, and electrostatic terms. Although the force field was originally fitted for the ZIF-8 framework, the explicit incorporation of bonding terms ensured that most hypothetical structures relaxed to chemically reasonable geometries, without overlapping atoms or excessive bond distortions.

After geometry optimisations, any structures with pairwise distances (including hydrogen atoms) less than 1.0 (0.6) Å were removed. This selection criterion ensured that highly-distorted frameworks with non-physical structures were omitted for the dataset, to prevent a skewing of the fitting to structures that were not relevant.

For all hypothetical structures, three additional training structures were generated by applying random strains to the unit cells and random displacements to the atomic positions, a process colloquially referred to as ‘rattling’:

- (i) The cell lengths (angles) were randomly varied by up to 5 % (2.5°) of the relaxed values, with random displacements applied to the atomic positions, having magnitudes drawn from a normal distribution with $\sigma = 0.01$ Å.
- (ii) The cell lengths (angles) were randomly varied by up to 10 % (5.0°) of the relaxed values, with random displacements applied to the atomic positions, having magnitudes drawn from a normal distribution with $\sigma = 0.1$ Å.
- (iii) The structures were coarse-grained to AB_2 nets; the cell lengths (angles) were randomly varied by up to 5 % (2.5°), and random displacements were applied to the coarse-grained bead positions, with magnitudes drawn from a normal distribution with $\sigma = 0.01$ Å; and the perturbed coarse-grained structures were then decorated back to ZIF chemical units.

Empirical potential trajectories

Additional training structures were generated using the MOF-FF empirical potential for ZIFs [133]. These simulations aimed to broadly sample the potential energy surface rather than model specific transitions with high accuracy.

First, MD simulations were conducted for the compression of ZIF-4 up to 2.0 GPa and ZIF-**zni** up to 10.0 GPa, as well as for the expansion of ZIF-**coi** under 5 GPa negative pressure. These pressures allowed for the full range of framework contraction and expansion. Training structures were extracted from these trajectories by taking evenly-spaced snapshots between the minimum and maximum volume snapshots.

Second, ‘melted’ training structures were generated by removing the bonded interactions between zinc atoms and imidazolate molecules, leaving only van der Waals and electrostatic interactions. For each of the ZIF-4, ZIF-**zni**, and ZIF-**coi** structures, simulations were performed at external pressures of 0.0 and 1.0 GPa, and at temperatures ranging from 50 to 300 K in 10 K increments. The systems were equilibrated for 10 ps, and the final trajectory snapshot was labelled using DFT and added to the training dataset.

All simulations were performed with a 0.001 fs timestep in the NPT ensemble at 300 K, with pressures rapidly ramped over 5 ps.

4.4.3 MD iterative training

I explored the structural space using iterative training [162]. The goal of this approach was to generate reference data efficiently to develop a ‘converged’ potential suitable for the desired application without further refinements. My focus was on achieving stable high-pressure and high-temperature MD starting from the ZIF-4, ZIF-**coi**, and ZIF-**zni** structures.

All simulations were performed in the NPT ensemble on single unit cells containing 272, 512, and 512 atoms for ZIF-4, ZIF-**coi**, and ZIF-**zni**, respectively. The timestep was set to 1 fs, and the temperature and pressure damping constants were set to 100 fs. Velocities were initialised randomly at 300 K. The structures were equilibrated at the initial conditions for 50 ps before heating or pressure were introduced.

For GAP MD iterations 1–5, I conducted both heating and pressure simulations. Heating simulations were carried out between 300 and 1500 K at three pressures:

−0.01, 1, and 0.01 GPa. The temperature was ramped up over 100 ps, followed by equilibrating the structure for 150 ps. Pressure simulations were conducted between 0 and 1 GPa for ZIF-4, and 0 to 10 GPa for ZIF-**coi** and ZIF-**zni**. The pressure was ramped over 50 ps, and the structures were equilibrated for 200 ps.

For GAP MD iterations 4 and 5, additional ‘expansion’ simulations were performed for ZIF-4 and ZIF-**coi**. In these simulations, a negative pressure from 0 to −10 GPa was applied. For ZIF-4, this allowed sampling of strained framework geometries. For ZIF-**coi**, the simulations imitated the reverse of the experimentally known pressure-induced single-crystal-to-single-crystal transition from ZIF-**zni** to ZIF-**coi** [375].

In the final GAP MD iteration, iter-6, I focused on heating ZIF-4 from 300 to 1800 K. Although these temperatures were higher than those used for production runs (1500 K), they provided a ‘buffer’ to ensure that high-energy configurations were well sampled. Ten trajectories were generated with densities on a linear scale between that of crystalline ZIF-4 ($3.6 \text{ T}_d \text{ nm}^{-3}$) and ZIF-**zni** ($4.7 \text{ T}_d \text{ nm}^{-3}$). The structures were heated to 1800 K over 15 ps and equilibrated at 1800 K for up to 250 ps. Snapshots were extracted from the equilibration period up to the point where unphysical ring-opening events occurred. Snapshots corresponding to the onset of ring-opening events were also included in the training dataset.

As already stated in Section 4.3.1, while the GAP model performed well, it was too computationally expensive. Consequently, I refitted the model using the ACE framework [178]. To ensure the ACE model maintained the same level of stability and accuracy, I performed one round of MD iterative training. This involved conducting heating and pressure simulations as previously described, including the 1800 K melting simulations for ZIF-4. Additionally, I included distorted structures where the imidazolate rings were compressed to varying extents perpendicular and parallel to the molecular C_2 axis to sample particularly short interatomic distances. While GAP-based ML models often include dimer curves, I found it more effective

to use many-body sampling of close-approach configurations in the ACE model. This difference in behaviour between the two frameworks likely originates from the explicit inclusion of a two-body term in GAP before adding the many-body terms as a correction, whereas ACE fits these terms simultaneously. This many-body sampling approach better captured the interactions encountered in simulations.

4.4.4 ML-driven HRMC iterative training

After developing an ML model with accurate energy predictions, I implemented ML-driven HRMC refinements for *a*-ZIF. The workflow for generating starting configurations is illustrated in Fig. 4.3(a).

I first generated 32 formula unit *a*-SiO₂ models using melt-quenched MD with the

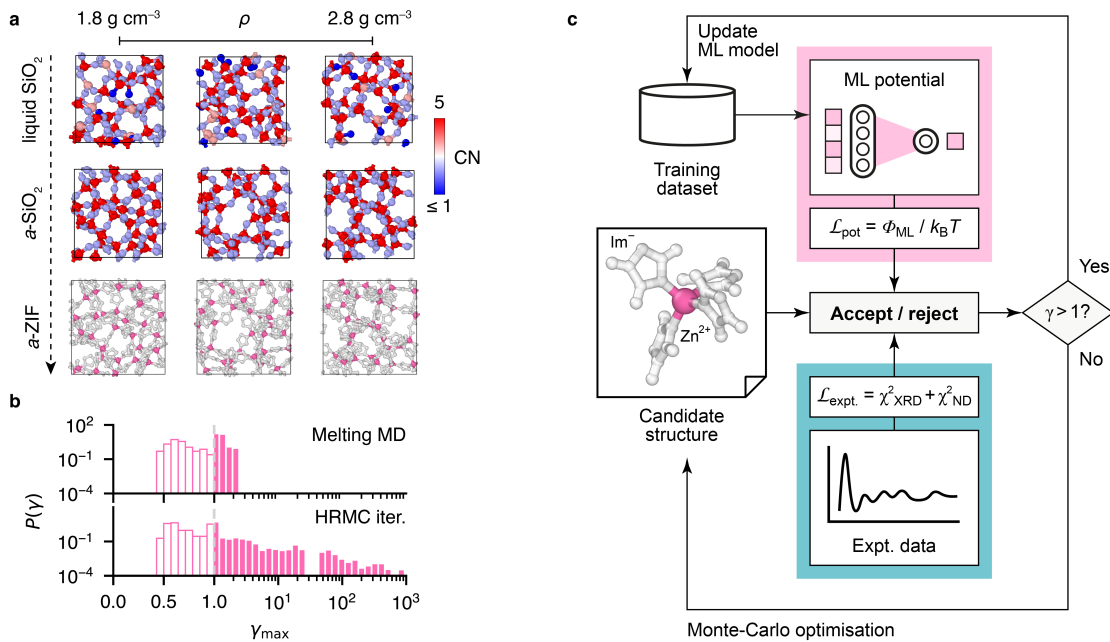


Figure 4.3: A data-driven protocol for modelling amorphous structures. **a**, Protocol for generating starting configurations for HRMC refinement of ZIFs. **b**, I distinguish ‘interpolative’ (empty bars) and ‘extrapolative’ (filled bars) behaviour using a threshold of $\gamma = 1$ [176]. While high-temperature (1500 K) MD simulations for crystalline structures remained predominantly interpolative (*top*), HRMC enables systematic exploration of different regions of amorphous configuration space (*middle*). **c**, Overview of the approach. HRMC refinements minimise a loss function with two contributions: the difference between experimental and computed data ($\mathcal{L}_{\text{expt}}$, turquoise) and the ML-predicted potential energy ($\mathcal{L}_{\text{pot}}$, pink). Moves are accepted or rejected according to the Metropolis–Hastings criterion. By computing the uncertainty, γ [176], of the ML prediction during HRMC, I can select structures to re-evaluate and add to the training set.

CHIK potential [120] and the simulation protocol from Ref. 60. Briefly, 32 silicon and 64 oxygen atoms were randomly added to the simulation box, further randomised at 6000 K in the NVT ensemble for 10 ps, and then held at 4000 K in the NPT ensemble for 100 ps. This melt was quenched to 300 K at a rate of 10^{11} K s⁻¹ and equilibrated for 10 ps, with a timestep of 1 fs throughout. Finally, the *a*-SiO₂ structures were decorated to ZIF frameworks using the protocol described in Section 4.4.2. These configurations served as the starting point for HRMC refinements.

I found that the HRMC refinements explored configuration space beyond the structures produced by MD. This exploration, beyond the scope of the ML training set, can be quantified using the ACE extrapolation grade, γ [176]. This metric distinguishes between ‘interpolative’ environments and ‘extrapolative’ behaviour using a threshold of $\gamma = 1$ [176]. Figure 4.3(b) (*top*) shows that heating all of the crystalline Zn(Im)₂ polymorphs to 1500 K using MD is well described by the ML model, as indicated by the majority of $\gamma_{\text{max}} < 1$. However, increasing the weighting of the data-matching term in the HRMC loss function causes γ_{max} to increase, as shown in Fig. 4.3(b) (*bottom*).

To ensure the ML potential’s stability for HRMC refinements, I performed three rounds of HRMC-driven iterative training, illustrated schematically in Fig. 4.3(c). By increasing the weighting of the data-matching term, $\mathcal{L}_{\text{expt.}}$, in the overall loss function, γ_{max} increases. Then, by selecting structures with extrapolative atomic environments and adding them to the training dataset, I ensured the model adequately sampled these regions of the potential energy surface, thereby constraining the HRMC refinements to chemically reasonable structures.

For each round of HRMC iterative training, ten starting configurations were produced using the *a*-SiO₂ CHIK and decoration protocol. To ensure the amorphous networks differed in their intermediate-range order, the densities of the initial random SiO₂ configurations were varied on a linear scale between 1.8 and 2.8 g cm⁻³ [60]. This resulted in ZIF structures with different ring statistics and associated

pore shapes and sizes [Fig. 4.3(a); *bottom row*]. Notably, although the ML potentials allowed for bond breaking and formation, such events tended not to occur during these HRMC refinements due to the contributions of local order to both the energy term and the data-fitting term. Consequently, the topology of the framework remained unchanged. For each input configuration, I ran three HRMC refinements with $\sigma = 0.1, 0.01, 0.001$ to control the relative weight of the data-fitting terms compared to the energy. Smaller values of σ gave stronger weighting to the data-fitting, promoting more extrapolative atomic environments.

4.4.5 Final dataset summary

For completeness, I give a summary of the full composition of the dataset in Table 4.3. I have shown, for each stage of the dataset development, the number of structures (‘Cells’) and the corresponding number of atoms. To give an indication of the diversity of atomic environments, I have also given the average and standard deviation of the relative DFT energies, ΔE , and DFT atomic force magnitudes, F , in each subset of the dataset. The energy value is given relative to the energy of

Table 4.3: Summary of the composition of the ML potential reference dataset.

		Database Size		ΔE (meV/atom)		F (eV/Å)	
		Cells	Atoms	μ	σ	μ	σ
iter-0	Crystal	374	101,728	96.9	79.1	10.3	9.7
	Topology	1,182	425,425	269.9	362.0	10.7	13.5
	MOF-FF	285	132,192	165.3	97.2	11.2	9.6
GAP iter-1	MD iter.	156	70,176	135.1	70.5	7.0	2.3
GAP iter-2	MD iter.	231	94,112	118.7	60.6	6.2	2.4
GAP iter-3	MD iter.	282	122,400	99.9	51.4	6.3	2.6
GAP iter-4	MD iter.	327	124,848	79.6	40.5	5.9	2.1
GAP iter-5	MD iter.	321	141,712	143.3	86.4	7.3	3.5
GAP iter-6	MD iter.	96	26,112	231.8	20.1	10.5	2.0
ACE iter-0	MD iter.	525	158,304	244.7	131.5	11.0	5.7
ACE iter-1	HRMC iter.	55	29,920	174.9	132.1	8.3	10.1
ACE iter-2	HRMC iter.	98	53,312	102.7	62.6	6.5	3.9
ACE iter-3	HRMC iter.	151	82,144	74.6	24.7	3.3	0.7
Total		4,083	1,562,385	149.0	93.7	8.0	5.2

the DFT geometry optimised ZIF-**zni** structure, the thermodynamically most stable ZIF at ambient conditions.

4.4.6 Testing set

Many ML models can nearly perfectly fit the data they have been trained on. Therefore, it is crucial to test a model on data points that were not included in the training set to evaluate its generalisation capability [167]. To achieve this, I developed a separate testing set of configurations representative of those anticipated in the application domain.

The testing set was generated by performing MD simulations using the ACE model fitted after the final round of HRMC-iterative training. For each crystalline $\text{Zn}(\text{Im})_2$ polymorph with topologies **cag**, **coi**, **crb**, **dft**, **gis**, **mer**, **neb**, **nog**, **sod**, **zni**, ten starting configurations were created by rescaling the densities on a linear scale between the experimental density of each polymorph and 1.63 g cm^{-3} , which is the density reported for ZIF glasses [80]. This approach samples the full range of expected densities likely to be explored.

Each structure was then subjected to a melting simulation. The simulation protocol involved performing MD simulations in the NVT ensemble in two stages: an equilibration stage and a melting stage. For the equilibration stage, each structure was equilibrated at 20 K for 5 ps with a temperature damping constant of 0.01 ps and a timestep of 0.25 fs. The temperature was then ramped from 20 K to 300 K over 224 ps with a temperature damping constant of 0.1 ps. The structure was subsequently held at 300 K for 5 ps. For the melting stage, the temperature was incrementally increased from 300 K to 1500 K in 12 stages, each involving 200 ps of ramping followed by 200 ps of equilibration, using a timestep of 0.1 fs and a temperature damping constant of 0.04 ps. At the end of each stage, the configuration was extracted, labelled with DFT energies and forces, and added to the testing set.

To include topologically-disordered structures, ten *a*- SiO_2 configurations containing 64 formula units were generated following the melt-quench protocol described in

Section 4.4.4, and decorated to form ZIF structures. The disordered ZIF structures were then subjected to the same equilibration and melting simulations described above. However, because these structures contain 1,088 atoms each, DFT calculations are considerably more expensive than those for the crystalline ZIF structures (which typically contain between 272 and 544 atoms per unit cell). As such, only the 300 K and 1500 K equilibrated configurations were labelled with DFT.

4.4.7 Numerical validation

The energy and force errors calculated for the external testing set are shown in Fig. 4.4(a) and (b), respectively. These plots show that MD iterative training gradually improved the numerical performance of the GAP ML model. The significant reduction in errors between iter-3 and iter-4 is due to increasing the radial cut-off for the two-body and SOAP descriptors to 6.0 Å, which improved the numerical errors and stability of MD simulations. For iter-5 and iter-6, more disordered high-energy structures were accessible through high-temperature MD. This is evident in Table 4.3, which details the number of structures, relative energies compared to ZIF-zni, and forces in each dataset subset. The increased errors in iter-5 likely arose from the model being skewed to describe more challenging structures.

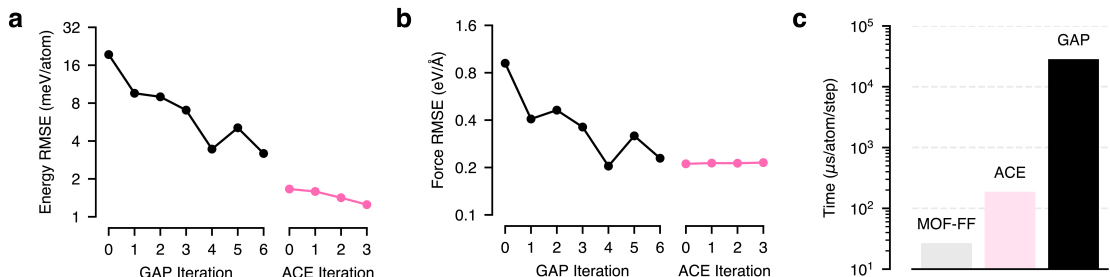


Figure 4.4: Numerical errors during iterative training. The **a**, energy and **b**, forces root-mean-square errors (RMSE) through the MD-iterative training of the GAP model and the subsequent HRMC-iterative training of the ACE model. **c**, Timing comparison between the final GAP iter-6 model and the initial ACE iter-0 model. The MOF-FF (non-reactive) empirical potential is included for reference [133]. The speed-up of the ACE framework, while maintaining the numerical accuracy of the GAP model, motivated the change in the ML potential framework for application in ML-driven HRMC refinements.

Figure 4.4 also demonstrates that switching to the ACE framework did not compromise the numerical accuracy of the ML model; on the contrary, the energy errors improved while the force errors remained comparable to the GAP iter-6 model. The relative speed of the two models is shown in Fig. 4.4(c), alongside MOF-FF for ZIFs as a reference empirical potential. The speed-up provided by the ACE framework, combined with its good numerical accuracy, justified the shift in ML framework, making ML-driven HRMC refinements feasible in practice. The energy errors improved during the HRMC iterations, while the force errors remained approximately the same, likely because HRMC refinements are driven primarily by energy differences.

4.4.8 Hyperparameter optimisation

Having finalised the dataset of training configurations, I optimised the hyperparameters of the ACE model using Bayesian optimisation [376]. Bayesian optimisation consists of two main components: a statistical model for modelling the objective (‘loss’) function and an acquisition function for deciding where next to sample [377]. By initiating an optimisation with typically wide-spread hyperparameters across a pre-set range and recording the loss value, the optimiser can begin to explore the loss landscape efficiently. This is the approach I used to optimise my final ACE models, as implemented in the XPOT software [378].

By varying the size of the basis set functions and the number of atomic properties, P , [see Section 4.3.3], I generated six ACE models, as described in Table 4.4. For all models, the functional form of the first two atomic properties were set to the Finnis–Sinclair form [Equation 4.9]. For models with more than two atomic properties, the exponents of the additional atomic properties ($f_{i>2}$) were optimised. I also optimised the hyperparameter `dcut`, which controls the range over which the radial components smoothly transition from one to zero, and `wlow`, which adjusts the relative weighting of low-to-high energy structures in the dataset. Additionally, I optimised the choice of radial basis function and associated parameters, with all

models ultimately refining to use simplified spherical Bessel functions. The final optimised parameters for each model are shown in Table 4.4.

The motivation for optimising these six models lies in balancing accuracy and speed. I show the energy and force errors in Fig. 4.5(a) and (b), respectively, and the relative timings in Fig. 4.5(c). While increasing the size of the basis set improves accuracy, particularly for force errors, the relative improvement is not substantial. Quantitatively, the difference between the models with the minimum and maximum RMSE for energies and forces are 0.61 meV/atom and 0.05 eV/Å, respectively. However, models with smaller basis set sizes were significantly faster [Fig. 4.5(c)].

This timing comparison also includes two empirical potentials: MOF-FF for ZIFs [133] and *ReaxFF* [142]. The fastest ACE model (model 1) is only approximately four times slower than MOF-FF for ZIFs. This is impressive given that MOF-FF for ZIFs requires a fixed topology whereas the ACE model does not. A fairer comparison is between the ACE models and *ReaxFF*, since both can describe reactive events, albeit at different relative accuracies. Notably, ACE model 1 approximately 1.1 times faster than *ReaxFF*.

Ultimately, the goal of the ML potential developed here is to provide an accurate description of the ZIF PES for HRMC refinements. HRMC refinements typically involve millions of moves, each requiring the calculation of the system’s energy, making computational efficiency a priority. Accordingly, I selected ACE model 3, which has the smallest basis set (600 functions per element) and includes 6 atomic

Table 4.4: Optimised ACE hyperparameters.

Model	nfunc	P	$f_{i>2}$	dcut	wlow
1	600	2		0.01	0.86
2	600	4	0.025, 0.883	0.73	0.99
3	600	6	0.975, 0.635, 0.025, 0.540	1.00	0.85
4	900	2		0.01	0.87
5	900	4	0.025, 0.760	0.11	0.76
6	1200	2		0.97	0.85

properties, for the final ACE ML model.

With the final dataset and ML model hyperparameters converged, I confirmed that the model is ‘saturated’ with training data by fitting progressively smaller numbers of training structures. The learning curves in Fig. 4.6 show the resulting errors, providing an indication of how well the model is ‘saturated’ with data. For each training set size, I performed five independent fits, each initialised with different random seeds. The average and standard deviation of the errors across each model are represented by the solid line and shaded area, respectively. The learning curves indicate that the ML model predictions were close to convergence with the amount of training data used. Increasing the number of structures from 2,048 to 4,096 resulted in only marginal decreases in RMSE for energies and forces, specifically 0.33 meV/atom and 0.02 eV/Å, respectively.

4.4.9 Comparison to other ML potentials for ZIFs

In Section 1.2.2 I discussed three recently released ZIF ML potentials and their key findings for the structure of *a*-ZIF [192–194]. Below, I compare the approach and quality of the ML potential developed herein to those three ML potentials. It is important to note, however, that a detailed comparison between ML potentials is

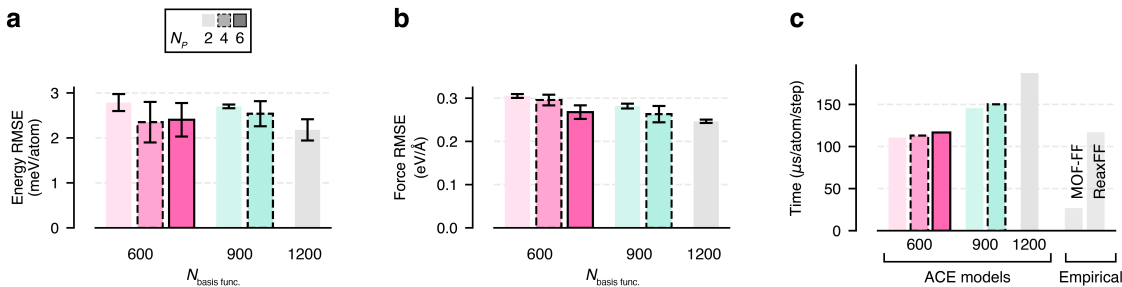


Figure 4.5: Hyperparameter optimisation of the ACE model. I optimised six ACE models in total, with an aim of achieving different trade-offs between speed and accuracy. This was achieved by varying the number of basis functions used (indicated with different colours), and number of additional atomic properties (indicated with different shades). I report the **a** energy errors, **b** force errors, and **c** timings. In panel **c**, I include the timings for two empirical potentials: MOF-FF for ZIFs [133] and *ReaxFF* [142]. Each model fit was repeated five times. The bar heights indicate the average value and the whiskers indicate the standard deviation.

challenging due to differences in architectures, training datasets, testing sets, and the digital experiments used for validation. Nonetheless, it is possible to make some general remarks. Here, I focus on the training dataset sizes and compositions, numerical errors, and efficiency. The results are summarised in Table 4.5.

The dataset developed in this work is similar in size to those used for the **NequIP** and **Allegro** models, but significantly smaller than those required for the **DeepMD** models [Table 4.5]. This difference likely stems from the choice of ML potential framework. Models like **GAP**, **ACE**, **NequIP**, and **Allegro** use more information-rich atomistic descriptors compared to **DeepMD** [see Section 1.2.2], allowing them to achieve good accuracy with a smaller dataset. Reducing training dataset size is crucial for lowering computational costs, particularly as obtaining reference data with DFT is often the most resource-intensive step in developing an ML potential.

In terms of dataset composition, the approach taken in this work differs markedly from other ZIF ML potentials. Rather than relying on AIMD trajectories, I leveraged the availability of various **AB₂** networks and the relatively inexpensive decorate-and-relax method to efficiently generate a structurally and geometrically diverse dataset [*cf.* Chapter 3]. While some configurations may be less relevant to the specific study of *a*-ZIF-4, I anticipate this approach will yield a more robust dataset

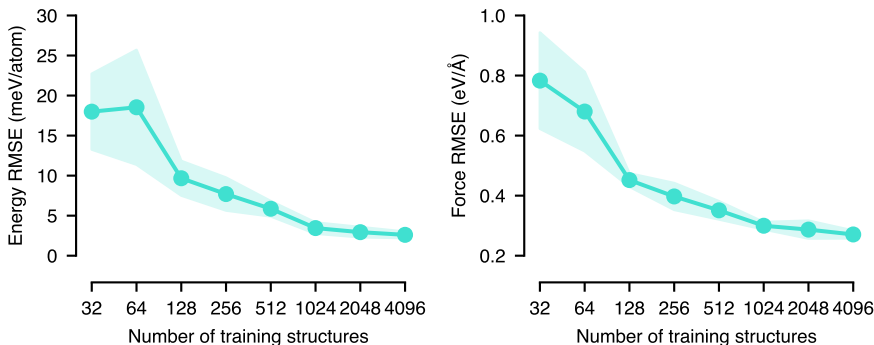


Figure 4.6: Learning curves for the final ACE ML potential. Using the final ('iter-3') dataset and the ACE model 3 hyperparameters described in Table 4.4, I fit to progressively smaller numbers of training structures. The resulting errors indicate how well the model is 'saturated' with data. Each fit was performed five times with different random seeds. The line indicates the average and the shading indicates the standard deviation.

for future research on the role of topology in stabilising different ZIF polymorphs [66, 69, 74]. Moreover, this methodology can be easily adapted to develop datasets with different ligands. For instance, it is known that the glass-forming ability of ZIFs can be experimentally modulated by adjusting the relative ratio of imidazolate linkers with various substituents [379, 380]. As the number and diversity of linker–linker interactions increase, the need for efficient generation of representative configurations will become increasingly critical.

When comparing the numerical accuracy of the different potentials, it is important to highlight that the error metrics for the ML potential developed in this work were evaluated using an external test set [see Section 4.4.6], unlike the literature potentials, which used reserved configurations from their training data generation protocols. Additionally, the error metrics in Ref. 192 are reported in terms of the mean absolute error (MAE), rather than RMSE. The RMSE is always greater than or equal to MAE and is skewed upward by higher errors [167]. While the energy errors for the ACE model 3 are comparable to those reported for other ML potentials, the force errors are generally higher [Table 4.5]. This discrepancy may be attributed to the higher accuracy achievable by the highly flexible form of neural network potentials, as well as the more diverse (and therefore more challenging) testing set used in this work [*cf.* the discussion of numerical performance in Ref. 187].

Finally, in terms of computational efficiency, the ACE model 3 demonstrates an intermediate performance level compared to the other ML potentials [Table 4.5].

Table 4.5: Comparing different ZIF ML potentials.

ML Potential	Dataset Size	Energy Error	Force Error	Efficiency
	Cells	(meV/atom)	(eV/Å)	(μ s/atom/call)
DeepMD [193]	>200,000	4.4	1.2	12
DeepMD [194]	39,512	1.4	0.1	-
NequIP [192]	2,375	0.6 (MAE)	0.015 (MAE)	282
Allegro [192]	2,375	0.7 (MAE)	0.03 (MAE)	347
ACE Model 3	4,083	2.6	0.27	117

It should be noted that this comparison does not account for the fact that the literature ML potentials were evaluated using graphics processing units, whereas I used central processing units for large-scale simulations. Although not as fast as the DeepMD models, it is significantly more efficient than the NequIP and Allegro models. All the models have demonstrated the capability to simulate large systems ($>17,408$ atoms), such as a $4 \times 4 \times 4$ supercell of ZIF-4, over timescales of hundreds of picoseconds. This indicates that each of these potentials is technically capable of modelling disordered ZIF forms effectively.

4.5 Conclusions

In this chapter, I described the development of a diverse dataset for the $\text{Zn}(\text{Im})_2$ material, labelled with quantum mechanical energy and forces. This dataset captures both the structural diversity of the crystalline polymorphs and their complex responsive behaviour, as well as the structure of disordered and defective forms of ZIFs. By optimising ACE models with varying basis set sizes and numbers of atomic properties, I selected a final model that is both accurate and efficient to evaluate. This will be crucial for generating large-scale *a*-ZIF-4 models in Chapter 5.

A key aspect of the methodology was the introduction of an iterative HRMC refinement protocol to produce relevant amorphous training configurations for fitting ML potentials. This protocol is likely to be generalisable for fitting other ML models for amorphous materials, particularly when there are limited alternative methods for generating training sets [171].

Chapter 5

The structure of *a*-ZIF-4

5.1 Acknowledgements

I am grateful to Louise Rosset (University of Oxford) for preparing large-scale *a*-SiO₂ structure models, which served as templates for starting configurations for my simulations of *a*-ZIF-4. I also thank Davide Proserpio (University of Milan) for invaluable discussions on topological metrics and for performing topological analysis with ToposPro to assist with the analysis and comparison of the network structures.

5.2 Introduction

In this chapter, I use the ML potential developed in Chapter 4 to generate high-quality models of *a*-ZIF-4 through a two-stage simulation protocol. First, a melt-quench MD simulation, guided by the ML potential, was used to facilitate the topological rearrangement of the disordered network. Second, the quenched configuration was refined against experimental X-ray and neutron total scattering data using HRMC coupled with the ML potential. The resulting *a*-ZIF-4 model is validated by comparing its structural features with those of crystalline ZIF frameworks.

I then explore how the topological features of *a*-ZIF-4 differ from those of prototypical amorphous tetrahedral networks such as *a*-Si and *a*-SiO₂. To quantify the (dis)similarities between the different networks, I propose a nomenclature based on well-established topological metrics used for crystalline frameworks, aimed at distinguishing between various amorphous networks.

5.3 Methods

5.3.1 MD simulations

First, I describe my approach for generating topologically-disordered ZIF configurations for the MD simulations. The starting point was 1,728 formula unit α -SiO₂ structures, generated using a melt-quench MD protocol introduced in Ref. 60. This protocol uses the CHIK empirical potential for the melt-quench, as described in Section 4.4.4. An additional annealing step is then performed on the melt-quenched structure using the GAP ML potential introduced in Ref. 60. This ‘hybrid’ protocol has been reported to generate the highest-quality α -SiO₂ models [60]. The use of the GAP ML potential, in conjunction with empirical potentials, ensures that the intermediate-range order of the α -SiO₂ models is consistent with experimental diffraction data, while also being substantially more energetically favourable than structures produced by the CHIK potential alone. The annealed α -SiO₂ configurations were ‘decorated’ to a ZIF structure following the protocol described in Section 4.4.2, and used as the starting configurations for MD in this chapter.

All MD simulations discussed in this chapter were carried out using the LAMMPS software [259]. The NVT ensemble was used with a Nosé–Hoover thermostat and a temperature damping parameter of 100 fs, unless otherwise specified [114, 115]. In all MD simulations for α -ZIF, the imidazolate molecules were held rigid using the LAMMPS `rigid/small` fix.

Before the melt-quench MD simulation, the α -ZIF-4 structures were equilibrated at 300 K following the ‘preparation’ protocol described in Ref. 134. This protocol was as follows. Initially, the system was relaxed at 20 K for 50 ps with a temperature damping parameter of 40 fs. This was followed by an additional relaxation period at the same temperature for 5 ps with a temperature damping parameter of 10 fs. The system was then heated to 300 K at a rate of 5 K ps⁻¹, with a temperature damping parameter of 100 fs, and finally equilibrated at 300 K for 5 ps.

5.3.2 HRMC simulations

HRMC simulations were performed using the same code and protocols as introduced in Section 4.3.6. Briefly, the cost function [Equation 1.10] for the HRMC refinements balanced three contributions: the X-ray and neutron total scattering functions reported in Ref. 76, and the configurational energy as determined by the ACE ML model [model 3; Section 4.4.8] described in the previous chapter.

The relative weighting of each experimental dataset i in the overall HRMC cost function is controlled by the parameter σ_i [Section 2.3.3]. For simplicity, I used the same σ_i value for both the X-ray and neutron datasets. To decide what weighting to use, I performed a series of HRMC refinements starting from the RMC configuration reported in Ref. 76, and varied the value of σ on a logarithmic scale between $0.001 \leq \sigma \leq 1.0$. Each refinement was run for one million proposed moves, with the simulation temperature $T = 573$ K. The energies (relative to the lowest energy HRMC configuration generated) and fits to data are shown in Fig. 5.1. From this analysis, a value of $\sigma = 0.01$ was deemed to provide an optimal balance between fitting the data and maintaining the system's energy.

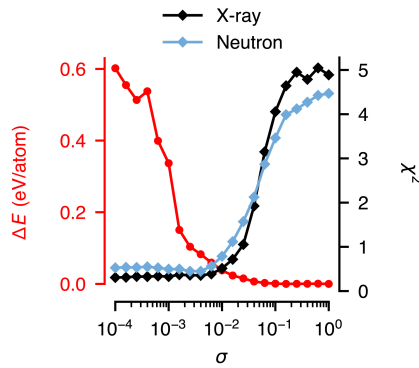


Figure 5.1: Selecting the optimal weighting between data-fitting and energy minimisation in HRMC refinements. As σ is increased, the importance of minimising the system increases. Accordingly, the final system energy (red) decreases, while the fits to X-ray (black) and neutron (blue) total scattering data get worse.

5.3.3 Pore size distribution calculations

Pore size distributions were calculated using the `PorosityPlus` software [381]. The radii of each element were set to the van der Waals radii (Zn: 1.3 Å; N: 1.55 Å; C: 1.70 Å; H: 1.20 Å). The probe particle radius was set to 1.2 Å.

5.3.4 Ring statistics calculations

Ring statistics analysis were performed using a custom Python interface to the `PeriodicGraphs.jl` Julia package [371, 382]. The package computes rings up to a specified maximum size, given the edges of the graph. For ZIF structures, Zn–Im connectivity was determined using a Zn–N cut-off of $r_{\text{cut}} = 2.5$ Å [80, 192], and zinc nodes connected by a common imidazolate molecule were linked. For *a*-Si, a Si–Si cut-off of $r_{\text{cut}} = 2.85$ Å [184] was used. For *a*-SiO₂, Si–O connectivity was determined with a cut-off of $r_{\text{cut}} = 2.0$ Å [60], and silicon nodes connected by a common oxygen atom were linked. The topology calculations were cross-checked against outputs from `ToposPro` [52] to ensure consistency in ring enumeration routines.

5.4 Results

5.4.1 Simulation protocol

First, I describe the procedure for generating the *a*-ZIF-4 models. The initial configurations were generated following the method detailed in Section 5.3.1. These configurations were then subjected to a melt-quench protocol. The temperature profile for this protocol is shown in Fig. 5.2. First, the system was heated to a temperature sufficiently high to allow the network to reorganise over 500 ps, allowing zinc–imidazolate bonds to break and reform. The system was annealed at this temperature for 5 ns, before quenching to 600 K at a rate of 10^{12} K s⁻¹. The system was then equilibrated at 600 K for 1 ns.

The melt-quenched model was then used as the starting configuration for HRMC refinement. The refinements were performed in three stages. Initially, an HRMC

refinement for five million proposed moves was conducted with imidazolate molecules held rigid, at a temperature of 300 K. Subsequently, an HRMC refinement for five million proposed moves was conducted at 300 K without the rigid molecule constraint. Finally, an HRMC refinement at 573 K was performed without the rigid body constraint for ten million proposed moves.

5.4.2 The structure of *a*-ZIF-4

I show the final fits to the X-ray and neutron total scattering functions in Fig. 5.3(a) and (b), respectively. Both are in excellent agreement with the experiment [76]. Notably, the fit to the X-ray data improves upon that reported in the RMC study of Ref. 76. I attribute this improvement to relaxing the fixed-topology constraint, allowing the network to rearrange during the MD melt-quench stage. The significance of the *a*-ZIF-4 topology will be discussed in greater detail in Section 5.4.3.

I show the partial PDFs extracted from the *a*-ZIF-4 models in Fig. 5.3(c). The PDFs reflect a high degree of local order. In particular, the Zn–N first peak at 2.0 Å and Zn–Zn peak at 6.0 Å are consistent with an extended network of Zn(Im)₄ tetrahedra. A slice of a representative *a*-ZIF-4 structure model is shown in Fig. 5.3(d).

Further evidence supporting the chemical validity of the model is provided in Fig. 5.4(a) and (b), comparing the *a*-ZIF-4 bond-length and bond-angle distribu-

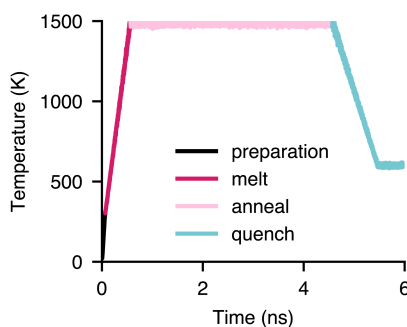


Figure 5.2: Temperature profile of melt-quench protocol for *a*-ZIF-4 models The generation of input configurations for HRMC consisted of four MD stages: (i) a preparatory stage where the system is relaxed at 300 K; (ii) a melting stage where the temperature is gradually increased; (iii) a 5 ns annealing stage at high temperature; (iv) followed by a quench and annealing step at 600 K.

tions, respectively, to those found in the crystalline ZIF polymorphs. The *a*-ZIF-4 model shows broadened distributions centred at the same values as the crystalline structures. The crystalline ZIF polymorphs all have an ‘ideal’ coordination, defined as fourfold-connected zinc atoms and twofold-connected imidazolate linkers. The *a*-ZIF-4 models have a small fraction of randomly distributed ‘defective’ sites. Figure 5.4(c) shows the fraction of under-coordinated (blue) and over-coordinated (red) sites for zinc atoms (three- and five-coordinate) and imidazolate linkers (one- and three-coordinate). The existence of defective sites in *a*-ZIFs has been proposed based on experimental NMR results [83].

Figure 5.4(d) shows the simulated pore size distributions for ZIF-4, ZIF-**zni**, and *a*-ZIF-4. Consistent with positron annihilation lifetime spectroscopy and *a*-ZIF-4 models constructed using *Polymatic*, both reported in Ref. 221, the *a*-ZIF-4 structure has a broader pore size distribution than ZIF-4 and ZIF-**zni**. The unimodal distribution of pore sizes in *a*-ZIF-4 is consistent with a homogeneous network.

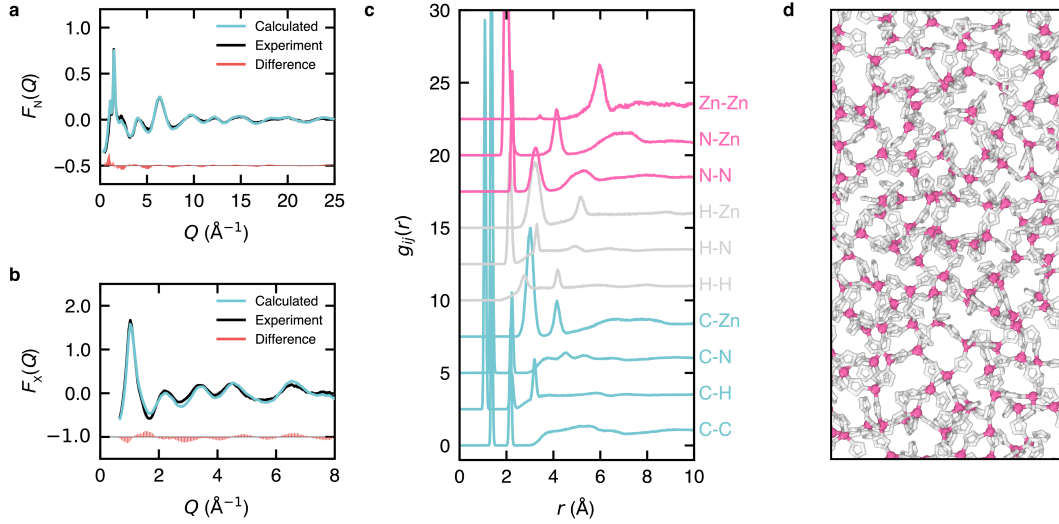


Figure 5.3: High-quality models of *a*-ZIF-4. The total scattering functions computed from the structure models (blue) are in excellent agreement with the experimental [76] (black) **a**, neutron total scattering structure factor, and **b**, X-ray total scattering structure function [199]. The difference between the computed and experiment functions are shown in pink, offset by -0.5 and -1 for the neutron and X-ray functions, respectively. **c**, Partial PDFs extracted from the *a*-ZIF-4 model. The PDFs are computed with a bin width of $\Delta r = 0.02$ Å. **d**, A slice of a representative *a*-ZIF-4 model. Hydrogen atoms have been omitted for clarity. The structure image was created using OVITO [372].

To probe the *a*-ZIF-4 structural order on longer length scales, all rings in the *a*-ZIF-4 models were enumerated. An *n*-membered ring is defined as a ring with *n* zinc nodes. Figure 5.4(e) shows the ring size distribution in *a*-ZIF-4 (*top*) alongside the crystalline ZIF distributions (*bottom*). *a*-ZIF-4 shows a broad range of ring sizes from *n* = 3 to 15, encompassing the full range of ring sizes observed in the crystalline polymorphs. Additionally, a small fraction of three-membered rings, not found in any crystalline polymorphs, is present in the amorphous models. Examples of different ring sizes extracted from an *a*-ZIF-4 model are illustrated above the distribution plots.

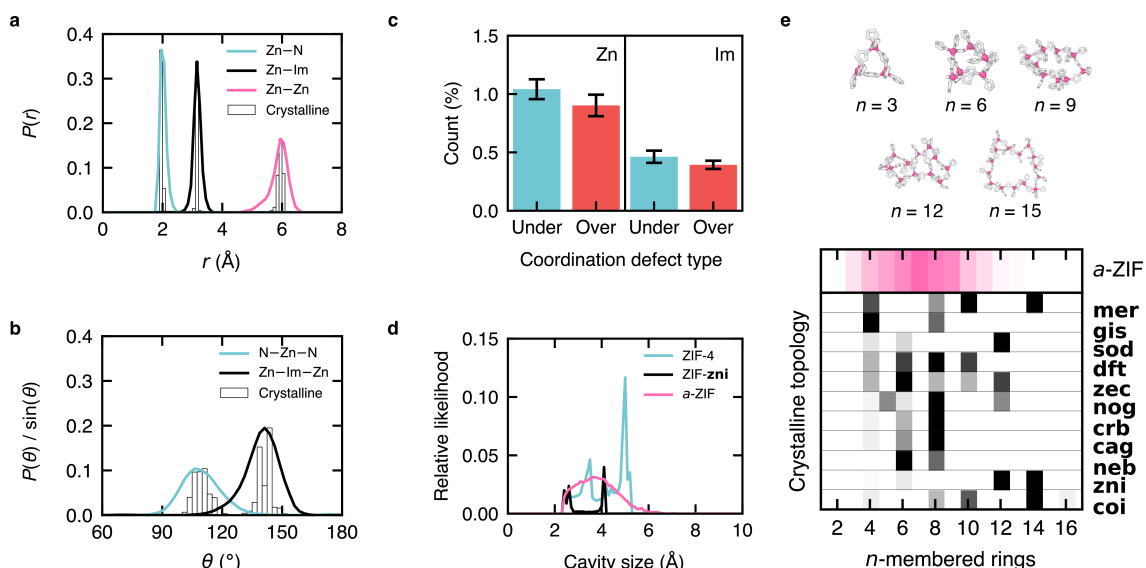


Figure 5.4: The structure of *a*-ZIF-4. **a**, Bond-length and **b**, bond-angle distributions for *a*-ZIF-4 models (continuous lines) compared to broadened distributions of those quantities for the crystalline ZIFs (white bars). Zn–N bonds are defined with a cut-off of $r_{\text{cut}} = 2.5$ Å [80, 192]. Zn–Im bonds were assigned if either N atom was bound to the Zn atom, and Zn–Zn bonds are defined as Zn atoms both bound to the same Im. Bond-length distributions were computed with a bin width of $\Delta r = 0.1$ Å, and bond-angle distributions with a bin width of $\Delta \theta = 2.5^\circ$. The bond-length and bond-angle crystalline distributions were rescaled to match the maximum values of the corresponding distributions for the amorphous models. **c**, Fractions of defect sites in the *a*-ZIF-4 models. An under-coordinated site is defined as zinc with a coordination number of 3 or imidazolate with 1; an over-coordinated site is defined as zinc with a coordination number of 5 or imidazolate with 3. **d**, Pore size distributions for ZIF-4, ZIF-zni, and *a*-ZIF, simulated using the PorosityPlus software [381]. **e**, Ring size distribution for *a*-ZIF-4 (pink) and crystalline ZIF topologies (grey). The pink line indicates the average distribution across the five *a*-ZIF-4 models, with light pink shading indicating the standard deviation. An *n*-membered ring has *n* zinc nodes. I show examples of different ring sizes in the *a*-ZIF-4 model above.

5.4.3 Tetrahedral CRNs

To gain deeper insights into the intermediate-range order of *a*-ZIF-4, it is helpful to compare its structure to other tetrahedral amorphous networks. In particular, *a*-Si and *a*-SiO₂ have been widely studied in the context of their topologies [10].

In Fig. 5.5(a–c), I show the underlying framework for structure models of *a*-Si [184], *a*-SiO₂ [60], and *a*-ZIF-4, respectively. Each node is colour-coded by the average ring size of all rings that involve that node. Nodes with an average ring size of ≤ 6 are blue; nodes with an average ring size of ≥ 7 are red. This illustration makes it immediately clear that the three amorphous materials have very different

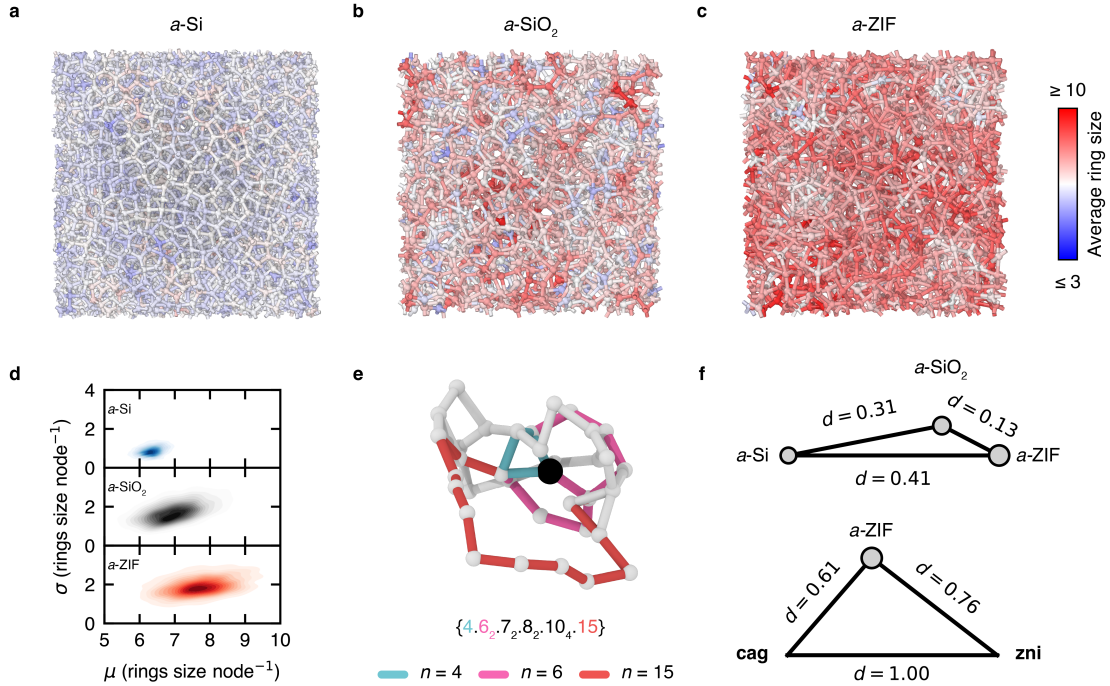


Figure 5.5: Comparing AB₂ CRNs of *a*-Si, *a*-SiO₂, and *a*-ZIF-4. The underlying net of **a**, *a*-Si, **b**, *a*-SiO₂, and **c**, *a*-ZIF-4, with each node colour-coded by the average ring size including that node. Note, the networks have very different tetrahedral densities of 48.9, 22.0, and 4.2 T_d nm⁻³ for *a*-Si, *a*-SiO₂, and *a*-ZIF-4, respectively. **d**, The average number of rings of size n per node plotted against the standard deviation of the same distribution for *a*-Si, top, *a*-SiO₂ (middle), and *a*-ZIF-4 (bottom). **e**, An illustration of the information encoded in a local topology metric. The number of rings of a size n about a central node (black) are enumerated. **f**, The topological distance metric, as discussed in the main text, can be used to plot the (dis)similarity between the topologies of amorphous (top) and crystalline (bottom) networks alike. Here, the size of each point is proportional to the collective standard deviation [Equation 5.3].

intermediate-range order. Compared to *a*-Si, both *a*-SiO₂ and *a*-ZIF-4 shift towards larger ring counts. This is shown quantitatively in Fig. 5.5(d) by plotting the average ($\mu(n)$) versus the standard deviation ($\sigma(n)$) of the ring size about a given node. There is a weak positive correlation between $\mu(n)$ and $\sigma(n)$, indicative of disassortative mixing (i.e., small rings are likely to neighbour large rings) [383].

Given the usefulness of comparing and navigating materials databases using geometric descriptors, as explored in Chapter 3, I developed an analogous approach for network topologies. This approach is based on topology descriptors commonly used to distinguish periodic nets [38, 51]. Conceptually, I assume that the network consists of a single ‘type’ of node characterised by an average set of topological features (i.e., coordination number and local ring size distribution). Each node in the network can then be viewed as a sample from a statistical distribution centred around this average node.

In practice, for each node j , I compute a set of coefficients $a_j(n)$, representing the number of rings of size n . An example of enumerating the rings about a central (black) node is shown in Fig. 5.5(e). The symbol illustrated therein indicates there is one four-membered ring, two six-membered rings, two seven-membered rings, and so on. This is similar to a topological ‘point symbol’.* The coefficients are assembled into a vector

$$\mathbf{r}_j = [a_j(1), a_j(2), \dots, a_j(n_{\max})]. \quad (5.1)$$

I then compute a per-structure topology metric by averaging over all nodes in the structure,

$$\mathbf{r} = \langle \mathbf{r}_j \rangle = [\langle a_j(1) \rangle, \langle a_j(2) \rangle, \dots, \langle a_j(n_{\max}) \rangle]. \quad (5.2)$$

Additionally, I define the collective standard deviation as

$$\sigma = \sqrt{\frac{\sum_j |\mathbf{r}_j - \langle \mathbf{r}_j \rangle|^2}{N}}. \quad (5.3)$$

*The primary difference in my definition compared to a point symbol is that I am using *rings*, whereas point symbols use *shortest circuits* [51].

This metric provides an intuitive view of the network topology by quantifying both the average local environment of the nodes and the variability of these environments within the structure. By capturing this information, the descriptor allows for a detailed comparison of different network topologies, irrespective of the chemistry.

I computed these descriptors for *a*-Si, *a*-SiO₂, and *a*-ZIF-4, and show the labels obtained when applying a threshold fraction of 1.0 for a given ring size, in Table 5.1. The collective standard deviation is shown as a subscript at the end of each symbol. The descriptors highlight the principal differences between the three amorphous networks. First, from *a*-Si to *a*-SiO₂ to *a*-ZIF-4, the ring size distributions broaden and extend to higher ring counts. Second, the diversity of local topologies increases, indicated by increasing values of σ .

I also define a topological distance metric as the Euclidean distance between any two normalised topological vectors, $d(\alpha, \beta) = |\tilde{\mathbf{r}}(\alpha) - \tilde{\mathbf{r}}(\beta)|$, which holds the intuitive meaning that topologically-similar networks will have small distances, while topologically-dissimilar networks will have large distances. Taking the triplet of *a*-Si, *a*-SiO₂, and *a*-ZIF-4 as an example, I compute the topological distances and plot these in Fig. 5.5(f, *top*). As expected, *a*-Si is further away from both *a*-SiO₂ and *a*-ZIF-4.

This approach can equally be applied to crystalline networks. The ZIF-4 **cag** and ZIF-**zni** topologies have one unique type of node each, and are represented by the symbols in Table 5.1. By definition, $\sigma = 0$ for uninodal nets, and so this is omitted from the symbol. Computing the distance from the *a*-ZIF-4 models gives

Table 5.1: Topology metric for crystalline and amorphous materials.

	Topology metric
<i>a</i> -Si	$\{5_{2.0}.6_{5.2}.7_{4.1}.8_{1.0}\}_{\sigma=3.5}$
<i>a</i> -SiO ₂	$\{5_{1.6}.6_{2.3}.7_{2.6}.8_{2.1}.9_{1.5}\}_{\sigma=4.2}$
<i>a</i> -ZIF-4	$\{5_{1.3}.6_{2.0}.7_{3.0}.8_{3.0}.9_{3.1}.10_{1.8}.11_{1.1}\}_{\sigma=5.7}$
cag	$\{4.6_6.8_{16}\}$
zni	$\{4.6_9.12_{120}.14_{140}\}$

the distance triangle embedding plotted in Fig. 5.5(f, *bottom*). As expected, the topology of *a*-ZIF-4 is intermediate between these two end-members, with a closer match to ZIF-4. This is likely because ZIF-**zni** is strongly shifted to high ring counts of $n = 12$ and 14, whereas *a*-ZIF-4 has a broader distribution centred around $n = 7$ [Fig. 5.4]. A similar broadening of the ring size distribution and shift to higher ring sizes, with respect to the crystalline ZIF-4 ring size distribution, was also reported for the NequIP and Allegro *a*-ZIF-4 models in Ref. 192.

5.5 Conclusions

I generated structural models of *a*-ZIF-4 using MD melt-quench simulations and HRMC refinements, guided by the ML potential developed in Chapter 4. These models show excellent agreement with X-ray and neutron diffraction experiments, exhibiting sensible local coordination geometries. By allowing the topology to rearrange as governed by the ML potential, the models provide valuable insights into the intermediate-range order of *a*-ZIF-4.

The topological order in *a*-ZIF-4 differs notably from that of *a*-Si and *a*-SiO₂, showing a shift towards larger ring sizes and a broader distribution. This explains why previous RMC models of *a*-ZIF-4 struggled to reproduce the total scattering functions: they were constrained by a topology derived from *a*-Si, incompatible with the energetic preferences of the ZIF network [76, 111, 219]. This may also explain why the DeepMD ML potential in Ref. 194, which allowed for topology relaxation at the appropriate density, was able to reproduce the experimental $S_N(Q)$ peak positions without refinement [80]. The glass model from that study could likely be further refined to match the experimental $S_N(Q)$ function using an ML potential-driven HRMC approach.

Finally, to facilitate the comparison between different amorphous networks, I introduced a topology metric based on descriptors commonly used to analyse crystalline networks. This metric enables a detailed and quantitative comparison of the intermediate-range order and variability within different network structures.

Chapter 6

Conclusion and outlook

In Section 1.3, I identified two primary aims for this thesis: (i) to develop a general modelling strategy for generating ‘realistic’ models of amorphous materials; and (ii) to characterise and understand the structural disorder observed in ACC and *a*-ZIF-4. In this concluding chapter, I assess the success in achieving these aims. I begin by discussing the HRMC approach, and possible future applications. I will then discuss conclusions specific to ACC, MOFs, and *a*-ZIF-4 in turn, before summarising with some general remarks.

6.1 HRMC simulations

The HRMC refinement strategy is very successful in producing high-quality amorphous models. I generated models for ACC (Chapter 2) and *a*-ZIF-4 (Chapter 5) that were capable of reproducing the experimental total scattering functions while also being energetically sensible. I confirmed the chemical sensibility of both models by comparing local descriptions of the amorphous atomic environments to those in the analogous crystalline compounds, and by confirming qualitative consistency with other experimental data (such as NMR). Based on these criteria, the models obtained in this thesis using HRMC are likely the most accurate structural solutions of these amorphous materials generated to date.

A prerequisite for the HRMC approach is access to an accurate force field for the system of interest. The availability of such force fields is expected to increase significantly with the progress in the field of ML atomistic modelling [384]. ‘Universal models’ are beginning to be introduced as potentials that are able to describe broad regions of configuration space [171, 385, 386]. These universal potentials can then be specialised for the study of specific materials, structural phases, or properties. Efficient strategies for generating domain-specific configurations will therefore be

required in order to ensure the final potential is trained on a suitable dataset [171]. The HRMC-driven iterative training protocol introduced in Chapter 4 can provide an efficient approach to targeting specific phases of interest by refining to the relevant experimental data.

A particular strength of the (H)RMC method is its ability to incorporate multiple constraints. Recent work has shown it is possible to develop ML models that can efficiently and accurately predict other measurable properties, such as XPS [232, 387] and NMR [388]. Incorporating these as constraints would provide complementary information that enhances the quality and interpretation of the models produced. For instance, NMR may offer additional sensitivity to material defects to which diffraction techniques are weakly sensitive.

6.2 ACC

The high-quality ACC model allowed me to establish a coarse-grained mapping between the effective interactions governing the distribution of calcium cations—the most ordered component of the system—and the theoretical phase-space of the LJG potential. This mapping locates ACC in a particularly complex region of phase-space, with many competing low-symmetry ground-state structures. This helps rationalise why an amorphous form of calcium carbonate exists, and how ACC can transform into various crystalline polymorphs.

Developing bio-inspired material design strategies where synthetic control over phase and morphology selection is possible requires a full understanding of the crystallisation mechanism. Having established a model for ACC, a possible route to studying these mechanisms could be refining models to *in-situ* X-ray diffraction data. In this way, one could obtain ‘snapshots’ of the evolving structure as it changes form.

Similarly, the incorporation of dopants such as Mg^{2+} , PO_4^{3-} , or organic components are known to drastically affect the stability and transition pathways of ACC, and are often found in biomineral samples of ACC. Empirical potentials have

emerged in the past couple of years which incorporate additional metal species [124]. Provided the relevant experimental data can be obtained, an HRMC approach would be an attractive option for investigating the role of counter-ions in stabilising and guiding such transformations to particular polymorphs and morphologies.

On a fundamental level, this study is also important because it presents an example of a material governed by isotropic multi-well potentials due to the presence of molecular ions. Molecular ions tend to exhibit a range of coordination modes which must give rise to multiple preferred distances between the counterions they coordinate. Whenever these distances are geometrically frustrated, as in ACC, one expects that the corresponding effective potential must contain multiple wells so as to stabilise both separations at once. Through suitable choice of chemical components, one might hope to navigate the complicated phase space associated with such unconventional potentials. Doing so may allow the targeted design of other complex structures through self-assembly.

6.3 MOFs

The scope of this thesis regarding MOFs is broader. I began by exploring the relationship between local geometries and the topologies that support them for AB_2 MOFs in Chapter 3. This involved developing an automated coarse-graining routine for MOFs and applying it to a curated dataset of experimental structures. By simplifying the structures to a representation that captured the underlying framework geometries, I demonstrated how a combination of atom-density-based descriptors and dimensionality reduction and clustering algorithms can assist in identifying key structure-property relationships.

A key result was the quantitative assessment of the breadth of network geometries afforded by a given topology. Understanding the likelihood of a given topology being adopted as a function of its building units (nodes and linkers) is invaluable because topology is an important indicator of material properties, such as porosity. This analysis also highlighted that changing the chemical building blocks in a material

will likely influence the intermediate-range order, even if those building blocks share a common local geometric preference.

This key insight provided the motivation for the subsequent approach taken in developing an ML potential for ZIFs and modelling *a*-ZIF, which forms the focus of the final two results chapters of this thesis. In Chapter 4, I built on the curated coarse-grained database from Chapter 3 to construct a topologically diverse dataset of hypothetical ZIF structures that sampled the potential energy surface broadly. After labelling the configurations with DFT energies and forces and improving the ML potential with MD-driven iterative training, I converged on an accurate ML potential capable of performing high-temperature MD simulations suitable for sampling disordered ZIF forms. To ensure the ML potential performs reliably and accurately during HRMC, I introduced a new ML potential training strategy based on HRMC refinements, which enabled the exploration of configurations that would otherwise not have been included in the training dataset. Ultimately, the ML potential developed was accurate, efficient, and suitable for reliably modelling *a*-ZIF.

I demonstrate this in Chapter 5 by generating a model for *a*-ZIF-4. This model is important for two main reasons. First, it achieves consistency with experimental total scattering data and is energetically sensible. Second, the *a*-ZIF-4 topology emerged from the simulation governed by the ML potential, rather than serving as a simulation constraint. Together, these two points allow us, for the first time, to understand the structure and topology of an amorphous MOF. An important finding is that, like *a*-Si and *a*-SiO₂, *a*-ZIF-4 adopts a CRN based on vertex-sharing tetrahedra, but there are meaningful differences between the topologies of the materials. I introduce a topology symbol that captures these differences, aiming to facilitate a general comparison between amorphous materials as is so successfully achieved with their crystalline counterparts.

6.4 Final remarks and future directions

There are several possible research directions that could be pursued based on these results. First, the coarse-grained approach developed in Chapter 3 could be extended for supervised ML tasks. The coarse-grained representation would likely complement other atomic property descriptors (e.g., ionisation potential, electronegativity) commonly used to predict the application properties of MOFs such as gas adsorption [389, 390]. Coarse-grained ML potentials are another exciting avenue [391–393]. We recently demonstrated with a proof-of-concept study [341] that local energies in ZIFs are predicted with good accuracy using coarse-grained AB_2 representations. Extending this approach to a full coarse-grained ML potential (i.e., learning energies and forces) might facilitate larger time- and length-scale simulations of ZIFs. This could circumvent the requirement for very high-temperature MD simulations, such as those used in this thesis, to probe topological rearrangements. In this way, mechanistic insight into the complex phase transitions occurring in MOFs might become computationally accessible.

Second, the topology descriptor introduced for amorphous materials in Chapter 5 is entirely general to any system for which bond connectivity can be defined. It could, therefore, provide a concise and informative description of transformation processes such as crystallisation events from amorphous forms.

Third, and more fundamentally, the work developed in Chapter 5 reveals that the nuances of the CRN landscape may be a fertile ground for exploration as a function of different chemistries. It is known, for example, that by substituting the imidazolate ring in ZIFs with different chemical groups, the susceptibility to amorphisation and vitrification changes [394]. How do these different linkers affect the topological features of the network? In turn, can we design the topology of amorphous networks through judicious choice of functional groups? Can these be enhanced by considering composite materials, known to have different topological preferences?

In summary, combining refinement to experimental data and accurate force field methods provides a robust technique for producing high-quality amorphous models which can be used as a general tool for atomistic modelling.

Bibliography

- [1] M. Thorpe, L. Tichý, *Properties and Applications of Amorphous Materials*, NATO Science Series II: Mathematics, Physics and Chemistry (Springer Netherlands, 2012).
- [2] S. Elliott, *Physics of Amorphous Materials* (John Wiley & Sons, Incorporated, 1986).
- [3] S. Shekhar, W. Bogaerts, L. Chrostowski, *et al.*, *Nat. Commun.* **15**, 751 (2024).
- [4] D. Drabold, *Eur. Phys. J. B* **68**, 1 (2009).
- [5] S. Weiner, I. Sagi, L. Addadi, *Science* **309**, 1027 (2005).
- [6] H. Furukawa, K. E. Cordova, M. O’Keeffe, O. M. Yaghi, *Science* **341**, 1230444 (2013).
- [7] T. D. Bennett, A. K. Cheetham, *Acc. Chem. Res.* **47**, 1555 (2014).
- [8] C. Giacovazzo, 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, 2011).
- [9] W. H. Zachariasen, *J. Am. Chem. Soc.* **54**, 3841 (1932).
- [10] S. R. Elliott, *Nature* **354**, 445 (1991).
- [11] T. Egami, S. J. L. Billinge, *Underneath the Bragg Peaks: Structural Analysis of Complex Materials* (Pergamon, Oxford, 2003).
- [12] A. C. Hannon, *Basic Concepts of Network Glass Structure* (John Wiley & Sons, Ltd, 2021), chap. 2.1, pp. 129–140.
- [13] P. Biswas, R. Atta-Fynn, D. A. Drabold, *Phys. Rev. B* **69**, 195207 (2004).
- [14] J. T. Randall, P. Rooksby, B. S. Cooper, *Z. Kristallogr. - Cryst. Mater.* **75**, 196 (1930).
- [15] K. Laaziri, S. Kycia, S. Roorda, M. Chicoine, J. L. Robertson, J. Wang, S. C. Moss, *Phys. Rev. B* **60**, 13520 (1999).
- [16] K. Laaziri, S. Kycia, S. Roorda, M. Chicoine, J. L. Robertson, J. Wang, S. C. Moss, *Phys. Rev. Lett.* **82**, 3460 (1999).
- [17] W.-L. Shao, J. Shinar, B. C. Gerstein, F. Li, J. S. Lannin, *Phys. Rev. B* **41**, 9491 (1990).

- [18] P. J. Heaney, C. T. Prewitt, G. V. Gibbs, eds., *Silica: Physical Behavior, Geochemistry, and Materials Applications* (Springer, Berlin, Boston, 1994).
- [19] J. S. Danel, G. Delapierre, *J. Micromech. Microeng.* **1**, 187 (1991).
- [20] J. Weitkamp, *Solid State Ion.* **131**, 175 (2000).
- [21] W. G. Cady, *Piezoelectricity: An Introduction to the Theory and Applications of Electromechanical Phenomena in Crystals* (McGraw-Hill Book Company, Inc., 1946), first edn.
- [22] A. Yeganeh-Haeri, D. J. Weidner, J. B. Parise, *Science* **257**, 650 (1992).
- [23] B. Yilmaz, U. Müller, *Top. Catal.* **52**, 888 (2009).
- [24] W. Blanc, B. Dussardier, *J. Opt.* **45**, 247 (2016).
- [25] A. Zeidler, K. Wezka, R. F. Rowlands, D. A. J. Whittaker, P. S. Salmon, A. Polidori, J. W. E. Drewitt, S. Klotz, H. E. Fischer, M. C. Wilding, C. L. Bull, M. G. Tucker, M. Wilson, *Phys. Rev. Lett.* **113**, 135501 (2014).
- [26] Y. Onodera, *et al.*, *NPG Asia Mater.* **12**, 85 (2020).
- [27] H. Masai, S. Kohara, T. Wakihara, Y. Shibazaki, Y. Onodera, A. Masuno, S. Sukenaga, K. Ohara, Y. Sakai, J. Haines, C. Levelut, P. Hébert, A. Isambert, D. A. Keen, M. Azuma, *Commun. Chem.* **6**, 269 (2023).
- [28] B. E. Warren, H. Krutter, O. Morningstar, *J. Am. Ceram. Soc.* **19**, 202 (1936).
- [29] B. E. Warren, J. Biscece, *J. Am. Ceram. Soc.* **21**, 49 (1938).
- [30] J. Konnert, P. D’Antonio, J. Karle, *J. Non-Cryst. Solids* **53**, 135 (1982).
- [31] T. E. Faber, J. M. Ziman, *Philos. Mag.* **11**, 153 (1965).
- [32] J. E. Enderby, D. M. North, P. A. Egelstaff, *Philos. Mag.* **14**, 961 (1966).
- [33] Q. Mei, C. J. Benmore, S. Sen, R. Sharma, J. L. Yarger, *Phys. Rev. B* **78**, 144204 (2008).
- [34] M. M. J. Treacy, K. H. Randall, S. Rao, J. A. Perry, D. J. Chadi, *Z. Kristallogr. - Cryst. Mater.* **212**, 768 (1997).
- [35] O. D. Friedrichs, A. W. M. Dress, D. H. Huson, J. Klinowski, A. L. Mackay, *Nature* **400**, 644 (1999).
- [36] M. M. J. Treacy, I. Rivin, E. Balkovsky, K. H. Randall, M. D. Foster, *Micro-porous Mesoporous Mater.* **74**, 121 (2004).
- [37] A. W. M. Dress, D. H. Huson, E. Molnár, *Acta Crystallogr. A* **49**, 806 (1993).
- [38] V. A. Blatov, O. Delgado-Friedrichs, M. O’Keeffe, D. M. Proserpio, *Acta Crystallogr. A* **63**, 418 (2007).
- [39] C. Baerlocher, L. B. McCusker, D. H. Olson, *Atlas of zeolite framework types* (Elsevier, 2007), 6th edn.

- [40] J. V. Smith, *Tetrahedral Frameworks of Zeolites, Clathrates and Related Materials*, vol. 14A of *Landolt-Börnstein* (Springer, Berlin, 2000).
- [41] R. A. Curtis, M. W. Deem, *J. Phys. Chem. B* **107**, 8612 (2003).
- [42] X. Li, M. W. Deem, *J. Phys. Chem. C* **118**, 15835 (2014).
- [43] M. Foster, I. Rivin, M. Treacy, O. Delgado Friedrichs, *Microporous Mesoporous Mater.* **90**, 32 (2006).
- [44] O. Delgado-Friedrichs, webGavrog (version 0.0.0), <https://github.com/odf/webGavrog> (2022).
- [45] E. M. Flanigen, *Studies in Surface Science and Catalysis*, H. van Bekkum, E. Flanigen, J. Jansen, eds. (Elsevier, 1991), vol. 58, chap. Introduction to Zeolite Science and Practice, pp. 13–34.
- [46] S. Abate, K. Barbera, G. Centi, P. Lanza fame, S. Perathoner, *Catal. Sci. Technol.* **6**, 2485 (2016).
- [47] K. Tanabe, W. F. Hölderich, *Appl. Catal., A* **181**, 399 (1999).
- [48] P. Sidheswaran, A. Bhat, *Thermochim. Acta* **298**, 55 (1997).
- [49] L. Čurković, Š. Cerjan-Stefanović, T. Filipan, *Wat. Res.* **31**, 1379 (1997).
- [50] M. E. Davis, R. F. Lobo, *Chem. Mater.* **4**, 756 (1992).
- [51] V. A. Blatov, M. O’Keeffe, D. M. Proserpio, *CrystEngComm* **12**, 44 (2010).
- [52] V. A. Blatov, A. P. Shevchenko, D. M. Proserpio, *Cryst. Growth Des.* **14**, 3576 (2014).
- [53] G. O. Brunner, F. Laves, *Wiss. Z. Tech. Univ. Dresden* **20**, 387 (1971).
- [54] M. M. J. Treacy, M. D. Foster, K. H. Randall, *Microporous Mesoporous Mater.* **87**, 255 (2006).
- [55] R. W. Grosse-Kunstleve, G. O. Brunner, N. J. A. Sloane, *Acta Crystallogr., Sect. A* **52**, 879 (1996).
- [56] D. Akporiaye, G. Price, *Zeolites* **9**, 321 (1989).
- [57] G. Brunner, *J. Solid State Chem.* **29**, 41 (1979).
- [58] D. Barthomeuf, *J. Phys. Chem.* **97**, 10092 (1993).
- [59] L. Öhrström, *Crystals* **5**, 154 (2015).
- [60] L. C. Erhard, J. Rohrer, K. Albe, V. L. Deringer, *npj Comput. Mater.* **8**, 90 (2022).
- [61] O. M. Yaghi, *ACS Cent. Sci.* **5**, 1295 (2019).
- [62] H. Jiang, D. Alezi, M. Eddaoudi, *Nat. Rev. Mater.* **6**, 466 (2021).

- [63] S. R. Batten, N. R. Champness, X.-M. Chen, J. Garcia-Martinez, S. Kitagawa, L. Öhrström, M. O’Keeffe, M. Paik Suh, J. Reedijk, *Pure Appl. Chem.* **85**, 1715 (2013).
- [64] M. Sturm, F. Brandl, D. Engel, W. Hoppe, *Acta Crystallogr., Sect. B* **31**, 2369 (1975).
- [65] R. Lehnert, F. Seel, *Z. Anorg. Allg. Chem.* **464**, 187 (1980).
- [66] 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, *Proc. Natl. Acad. Sci. USA* **103**, 10186 (2006).
- [67] A. Phan, C. J. Doonan, F. J. Uribe-Romo, C. B. Knobler, M. O’Keeffe, O. M. Yaghi, *Acc. Chem. Res.* **43**, 58 (2010).
- [68] Z. Zheng, Z. Rong, H. L. Nguyen, O. M. Yaghi, *Inorg. Chem.* **62**, 20861 (2023).
- [69] V. A. Blatov, G. D. Ilyushin, D. M. Proserpio, *Chem. Mater.* **25**, 412 (2013).
- [70] Z. Li, M. V. Nevitt, S. Ghose, *Appl. Phys. Lett.* **55**, 1730 (1989).
- [71] K. W. Chapman, G. J. Halder, P. J. Chupas, *J. Am. Chem. Soc.* **131**, 17546 (2009).
- [72] J.-C. Tan, B. Civalleri, C.-C. Lin, L. Valenzano, R. Galvelis, P.-F. Chen, T. D. Bennett, C. Mellot-Draznieks, C. M. Zicovich-Wilson, A. K. Cheetham, *Phys. Rev. Lett.* **108**, 095502 (2012).
- [73] J. J. Williams, K. E. Evans, R. I. Walton, *Appl. Phys. Lett.* **88**, 021914 (2006).
- [74] R. Banerjee, A. Phan, B. Wang, C. Knobler, H. Furukawa, M. O’Keeffe, O. M. Yaghi, *Science* **319**, 939 (2008).
- [75] R. N. Widmer, G. I. Lampronti, S. Chibani, C. W. Wilson, S. Anzellini, S. Farsang, A. K. Kleppe, N. P. M. Casati, S. G. MacLeod, S. A. T. Redfern, F.-X. Coudert, T. D. Bennett, *J. Am. Chem. Soc.* **141**, 9330 (2019).
- [76] 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, *Phys. Rev. Lett.* **104**, 115503 (2010).
- [77] D. W. Lewis, A. R. Ruiz-Salvador, A. Gómez, L. M. Rodriguez-Albelo, F.-X. Coudert, B. Slater, A. K. Cheetham, C. Mellot-Draznieks, *CrystEngComm* **11**, 2272 (2009).
- [78] T. D. Bennett, P. Simoncic, S. A. Moggach, F. Gozzo, P. Macchi, D. A. Keen, J.-C. Tan, A. K. Cheetham, *Chem. Commun.* **47**, 7983 (2011).
- [79] T. D. Bennett, J. C. Tan, Y. Yue, *et al.*, *Nat. Commun.* **6**, 8079 (2015).
- [80] R. Gaillac, P. Pullumbi, K. A. Beyer, K. W. Chapman, D. A. Keen, T. D. Bennett, F.-X. Coudert, *Nat. Mater.* **16**, 1149 (2017).
- [81] A. Kentgens, *Geoderma* **80**, 271 (1997).

- [82] A. Sutrisno, V. V. Terskikh, Q. Shi, Z. Song, J. Dong, S. Y. Ding, W. Wang, B. R. Provost, T. D. Daff, T. K. Woo, Y. Huang, *Chem. - Eur. J.* **18**, 12251 (2012).
- [83] E. F. Baxter, T. D. Bennett, C. Mellot-Draznieks, C. Gervais, F. Blanc, A. K. Cheetham, *Phys. Chem. Chem. Phys.* **17**, 25191 (2015).
- [84] R. S. K. Madsen, A. Qiao, J. Sen, I. Hung, K. Chen, Z. Gan, S. Sen, Y. Yue, *Science* **367**, 1473 (2020).
- [85] D. Gebauer, A. Völkel, H. Cölfen, *Science* **322**, 1819 (2008).
- [86] J. Aizenberg, G. Lambert, S. Weiner, L. Addadi, *J. Am. Chem. Soc.* **124**, 32 (2002).
- [87] S. Raz, P. C. Hamilton, F. H. Wilt, S. Weiner, L. Addadi, *Adv. Funct. Mater.* **13**, 480 (2003).
- [88] J. Aizenberg, *Adv. Mater.* **16**, 1295 (2004).
- [89] L. Addadi, S. Raz, S. Weiner, *Adv. Mater.* **15**, 959 (2003).
- [90] V. C. Farmer, *The Infrared Spectra of Minerals* (Mineralogical Society of Great Britain and Ireland, 1974).
- [91] S. Raz, O. Testeniere, A. Hecker, S. Weiner, G. Luquet, *Biol. Bull.* **203**, 269 (2002).
- [92] J. Aufort, R. Demichelis, *Cryst. Growth Des.* **20**, 8028 (2020).
- [93] Y.-B. Hu, D. A. Wolf-Gladrow, G. S. Dieckmann, C. Völker, G. Nehrke, *Mar. Chem.* **162**, 10 (2014).
- [94] Z. Zou, W. J. E. M. Habraken, G. Matveeva, A. C. S. Jensen, L. Bertinetti, M. A. Hood, C.-y. Sun, P. U. P. A. Gilbert, I. Polishchuk, B. Pokroy, J. Mahamid, Y. Politi, S. Weiner, P. Werner, S. Bette, R. Dinnebier, U. Kolb, E. Zolotoyabko, P. Fratzl, *Science* **363**, 396 (2019).
- [95] F. M. Michel, J. MacDonald, J. Feng, P. B. L., L. Ehm, C. Tarabrella, J. B. Parise, R. J. Reeder, *Chem. Mater.* **20**, 4720 (2008).
- [96] S. Susman, D. Price, K. Volin, R. Dejus, D. Montague, *J. Non-Cryst. Solids* **106**, 26 (1988).
- [97] Y. Levi-Kalisman, S. Raz, S. Weiner, L. Addadi, I. Sagi, *Adv. Funct. Mater.* **12**, 43 (2002).
- [98] A. Becker, U. Bismayer, M. Eppele, H. Fabritius, B. Hasse, J. Shi, A. Ziegler, *Dalton Trans.* pp. 551–555 (2003).
- [99] C. Günther, A. Becker, G. Wolf, M. Eppele, *Z. Anorg. Allg. Chem.* **631**, 2830 (2005).
- [100] M. Neumann, M. Eppele, *Eur. J. Inorg. Chem.* **2007**, 1953 (2007).
- [101] A. G. Christy, *Cryst. Growth Des.* **17**, 3567 (2017).

- [102] E. Makovicky, *Am. Mineral.* **101**, 1636 (2016).
- [103] J. Wang, F. Zhang, J. Zhang, R. C. Ewing, U. Becker, Z. Cai, *J. Cryst. Growth* **407**, 78 (2014).
- [104] A. Le Bail, S. Ouhenia, D. Chateigner, *Powder Diff.* **26**, 16 (2011).
- [105] M. P. Allen, D. J. Tildesley, *Computer Simulation of Liquids* (Oxford University Press, Oxford, 1987).
- [106] D. Frenkel, B. Smit, *Understanding Molecular Simulation: From Algorithms to Applications* (Academic Press, San Diego, 2001), second edn.
- [107] W. L. Jorgensen, *Advances in Chemical Physics* **70**, 469 (1988).
- [108] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, E. Teller, *Journal of Chemical Physics* **21**, 1087 (1953).
- [109] M. T. Dove, *Semin. SEM* **4**, 7 (2008).
- [110] M. C. Warren, M. T. Dove, E. R. Myers, A. Bosenick, E. J. Palin, C. I. Sainz-Diaz, B. S. Guiton, S. A. T. Redfern, *Mineral. Mag.* **65**, 221 (2001).
- [111] F. Wooten, K. Winer, D. Weaire, *Phys. Rev. Lett.* **54**, 1392 (1985).
- [112] B. R. Djordjević, M. F. Thorpe, F. Wooten, *Phys. Rev. B* **52**, 5685 (1995).
- [113] D. O. Morley, M. Wilson, *J. Phys. Condens. Matter* **30**, 50LT02 (2018).
- [114] S. Nosé, *Mol. Phys.* **52**, 255 (1984).
- [115] W. G. Hoover, *Phys. Rev. A* **31**, 1695 (1985).
- [116] W. Shinoda, M. Shiga, M. Mikami, *Phys. Rev. B* **69**, 134103 (2004).
- [117] M. Born, J. E. Mayer, *Z. Phys.* **75**, 1 (1932).
- [118] B. W. H. van Beest, G. J. Kramer, R. A. van Santen, *Phys. Rev. Lett.* **64**, 1955 (1990).
- [119] E. Flikkema, S. T. Bromley, *Chem. Phys. Lett.* **378**, 622 (2003).
- [120] A. Carré, J. Horbach, S. Ispas, W. Kob, *Europhys. Lett.* **82**, 17001 (2008).
- [121] P. Raiteri, J. D. Gale, D. Quigley, P. M. Rodger, *J. Phys. Chem. C* **114**, 5997 (2010).
- [122] P. Raiteri, J. D. Gale, *J. Am. Chem. Soc.* **132**, 17623 (2010).
- [123] P. Raiteri, R. Demichelis, J. D. Gale, *J. Phys. Chem. C* **119**, 24447 (2015).
- [124] B. Armstrong, A. Silvestri, R. Demichelis, P. Raiteri, J. D. Gale, *Philos. Trans. R. Soc., A* **381**, 20220250 (2023).
- [125] G. Wolf, J. Lerchner, H. e. a. Schmidt, *J. Therm. Anal.* **46**, 353 (1996).
- [126] G. Wolf, E. Königsberger, H. e. a. Schmidt, *J. Therm. Anal. Calorim.* **60**, 463 (2000).

- [127] H. W. Horn, W. C. Swope, J. W. Pitera, J. D. Madura, T. J. Dick, G. L. Hura, T. Head-Gordon, *J. Chem. Phys.* **120**, 9665 (2004).
- [128] W. S. Benedict, N. Gailar, E. K. Plyler, *J. Chem. Phys.* **24**, 1139 (1956).
- [129] Y. Wu, H. L. Tepper, G. A. Voth, *J. Chem. Phys.* **124**, 024503 (2006).
- [130] R. Demichelis, P. Raiteri, J. D. Gale, D. Quigley, D. Gebauer, *Nat. Commun.* **2**, 590 (2011).
- [131] D. A. Drabold, P. Stumm, P. A. Fedders, *Phys. Rev. Lett.* **72**, 2666 (1994).
- [132] P. N. Keating, *Phys. Rev.* **145**, 637 (1966).
- [133] J. P. Dürholt, G. Fraux, F.-X. Coudert, R. Schmid, *J. Chem. Theory Comput.* **15**, 2420 (2019).
- [134] N. Castel, F.-X. Coudert, *J. Phys. Chem. C* **126**, 19532 (2022).
- [135] N. L. Allinger, K. Chen, J.-H. Lii, *J. Comp. Chem.* **17**, 642 (1996).
- [136] A. K. Rappé, C. J. Casewit, K. Colwell, W. A. Goddard III, W. M. Skiff, *J. Am. Chem. Soc.* **114**, 10024 (1992).
- [137] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, P. A. Kollman, *J. Am. Chem. Soc.* **117**, 5179 (1995).
- [138] H. Sun, *J. Phys. Chem. B* **102**, 7338 (1998).
- [139] A. Schneemann, V. Bon, I. Schwedler, I. Senkovska, S. Kaskel, R. A. Fischer, *Chem. Soc. Rev.* **43**, 6062 (2014).
- [140] T. Senftle, S. Hong, M. Islam, et al., *npj Comput. Mater.* **2**, 15011 (2016).
- [141] A. C. T. van Duin, S. Dasgupta, F. Lorant, W. A. Goddard, *J. Phys. Chem. A* **105**, 9396 (2001).
- [142] Y. Yang, Y. K. Shin, S. Li, T. D. Bennett, A. C. T. van Duin, J. C. Mauro, *J. Phys. Chem. B* **122**, 9616 (2018).
- [143] S. S. Sørensen, M. B. Østergaard, M. Stepniewska, H. Johra, Y. Yue, M. M. Smedskjaer, *ACS Appl. Mater. Interfaces* **12**, 18893 (2020).
- [144] T. To, S. S. Sørensen, M. Stepniewska, et al., *Nat. Commun.* **11**, 2593 (2020).
- [145] T. Du, S. S. Sørensen, Q. Zhou, M. Bauchy, M. M. Smedskjaer, *Chem. Mater.* **34**, 8749 (2022).
- [146] D. S. Sholl, J. A. Steckel, *Density Functional Theory: A Practical Introduction* (Wiley, 2009).
- [147] M. C. Payne, M. P. Teter, D. C. Allan, T. A. Arias, J. D. Joannopoulos, *Rev. Mod. Phys.* **64**, 1045 (1992).
- [148] M. D. Segall, P. J. D. Lindan, M. J. Probert, C. J. Pickard, P. J. Hasnip, S. J. Clark, M. C. Payne, *J. Phys.: Condens. Matter* **14**, 2717 (2002).

- [149] P. Hohenberg, W. Kohn, *Phys. Rev.* **136**, B864 (1964).
- [150] W. Kohn, L. J. Sham, *Phys. Rev.* **140**, A1133 (1965).
- [151] J. P. Perdew, K. Schmidt, *AIP Conf. Proc.* **577**, 1 (2001).
- [152] D. C. Langreth, M. J. Mehl, *Phys. Rev. B* **28**, 1809 (1983).
- [153] A. D. Becke, *Phys. Rev. A* **38**, 3098 (1988).
- [154] C. J. Cramer, *Essentials of Computational Chemistry* (John Wiley & Sons, Chichester, UK, 2004), second edn.
- [155] S. Grimme, A. Hansen, J. G. Brandenburg, C. Bannwarth, *Chem. Rev.* **116**, 5105 (2016).
- [156] R. Car, M. Parrinello, *Phys. Rev. Lett.* **55**, 2471 (1985).
- [157] T. D. Kühne, *WIREs Comput. Mol. Sci.* **4**, 391 (2014).
- [158] M. Born, J. Oppenheimer, *Ann. Phys.* **84**, 457 (1927).
- [159] R. Gaillac, P. Pullumbi, F.-X. Coudert, *J. Phys. Chem. C* **122**, 6730 (2018).
- [160] R. N. Widmer, G. I. Lampronti, S. Anzellini, et al., *Nat. Mater.* **18**, 370 (2019).
- [161] R. Gaillac, P. Pullumbi, T. D. Bennett, F.-X. Coudert, *Chem. Mater.* **32**, 8004 (2020).
- [162] V. L. Deringer, M. A. Caro, G. Csányi, *Adv. Mater.* **31**, 1902765 (2019).
- [163] F. Noé, A. Tkatchenko, K.-R. Müller, C. Clementi, *Annu. Rev. Phys. Chem.* **71**, 361 (2020).
- [164] P. Friederich, F. Häse, J. Proppe, et al., *Nat. Mater.* **20**, 750 (2021).
- [165] O. T. Unke, S. Chmiela, H. E. Sauceda, M. Gastegger, I. Poltavsky, K. T. Schütt, A. Tkatchenko, K.-R. Müller, *Chem. Rev.* **121**, 10142 (2021).
- [166] V. L. Deringer, A. P. Bartók, N. Bernstein, D. M. Wilkins, M. Ceriotti, G. Csányi, *Chem. Rev.* **121**, 10073 (2021).
- [167] J. D. Morrow, J. L. A. Gardner, V. L. Deringer, *J. Chem. Phys.* **158**, 121501 (2023).
- [168] J. Behler, M. Parrinello, *Phys. Rev. Lett.* **98**, 146401 (2007).
- [169] A. P. Bartók, M. C. Payne, R. Kondor, G. Csányi, *Phys. Rev. Lett.* **104**, 136403 (2010).
- [170] F. Musil, A. Grisafi, A. P. Bartók, C. Ortner, G. Csányi, M. Ceriotti, *Chem. Rev.* **121**, 9759 (2021).
- [171] C. Ben Mahmoud, J. L. A. Gardner, V. L. Deringer, *Nat. Comput. Sci.* (2024).
- [172] J. Hoja, L. Medrano Sandonas, B. G. Ernst, et al., *Sci. Data* **8**, 43 (2021).

- [173] National Institute of Standards and Technology, Nist inorganic crystal structure database, nist standard reference database number 3, <https://doi.org/10.18434/M32147>. Gaithersburg MD, 20899.
- [174] C. R. Groom, I. J. Bruno, M. P. Lightfoot, S. C. Ward, *Acta Cryst. B* **72**, 171 (2016).
- [175] I. S. Novikov, K. Gubaev, E. V. Podryabinkin, A. V. Shapeev, *Mach. Learn.: Sci. Technol.* **2**, 025002 (2020).
- [176] Y. Lysogorskiy, A. Bochkarev, M. Mrovec, R. Drautz, *Phys. Rev. Mater.* **7**, 043801 (2023).
- [177] J. Westermayr, J. Gilkes, R. Barrett, *et al.*, *Nat. Comput. Sci.* **3**, 139 (2023).
- [178] R. Drautz, *Phys. Rev. B* **99**, 014104 (2019).
- [179] M. J. Willatt, F. Musil, M. Ceriotti, *J. Chem. Phys.* **150**, 154110 (2019).
- [180] A. P. Bartók, R. Kondor, G. Csányi, *Phys. Rev. B* **87**, 184115 (2013).
- [181] N. Artrith, T. Morawietz, J. Behler, *Phys. Rev. B* **83**, 153101 (2011).
- [182] K. T. Schütt, P. Kessel, M. Gastegger, K. A. Nicoli, A. Tkatchenko, K.-R. Müller, *J. Chem. Theory Comput.* **15**, 448 (2019).
- [183] C. E. Rasmussen, C. K. I. Williams, *Gaussian Processes for Machine Learning*, Adaptive Computation and Machine Learning (MIT Press, Cambridge, Mass, 2006).
- [184] 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, *J. Phys. Chem. Lett.* **9**, 2879 (2018).
- [185] V. L. Deringer, N. Bernstein, G. Csányi, *et al.*, *Nature* **589**, 59 (2021).
- [186] J. D. Morrow, C. Ugwumadu, D. A. Drabold, S. R. Elliott, A. L. Goodwin, V. L. Deringer, *Angew. Chem.* **136**, e202403842 (2024).
- [187] L. Erhard, J. Rohrer, K. Albe, *et al.*, *Nat. Commun.* **15**, 1927 (2024).
- [188] M. Eckhoff, J. Behler, *J. Chem. Theory Comput.* **15**, 3793 (2019).
- [189] S. Vandenhoute, M. Cools-Ceuppens, S. DeKeyser, *et al.*, *npj Comput. Mater.* **9**, 19 (2023).
- [190] R. Goeminne, L. Vanduyfhuys, V. Van Speybroeck, T. Verstraelen, *J. Chem. Theory Comput.* **19**, 6313 (2023).
- [191] S. Liu, R. Dupuis, D. Fan, S. Benzaria, M. Bonneau, P. Bhatt, M. Eddaoudi, G. Maurin, *Chem. Sci.* **15**, 5294 (2024).
- [192] N. Castel, D. André, C. Edwards, J. D. Evans, F.-X. Coudert, *Digital Discovery* **3**, 355 (2024).
- [193] Z. Shi, B. Liu, Y. Yue, A. Arramel, N. Li, *J. Am. Ceram. Soc.* **107**, 3845 (2024).

- [194] T. Du, S. Li, S. Ganisetti, M. Bauchy, Y. Yue, M. M. Smedskjaer, *Natl. Sci. Rev.* **11**, nwae023 (2024).
- [195] S. Batzner, A. Musaelian, L. Sun, et al., *Nat. Commun.* **13**, 2453 (2022).
- [196] J. Zeng, et al., *J. Chem. Phys.* **159**, 054801 (2023).
- [197] A. Musaelian, S. Batzner, A. Johansson, et al., *Nat. Commun.* **14**, 579 (2023).
- [198] J. Han, L. Zhang, R. Car, W. E, *Commun. Comput. Phys.* **23**, 629 (2018).
- [199] D. A. Keen, *J. Appl. Crystallogr.* **34**, 172 (2001).
- [200] F. Ordway, *Science* **143**, 800 (1964).
- [201] D. Polk, *J. Non-Cryst. Solids* **5**, 365 (1971).
- [202] D. Evans, S. King, *Nature* **212**, 1353 (1966).
- [203] R. Bell, P. Dean, *Nature* **212**, 1354 (1966).
- [204] S. C. Moss, J. F. Graczyk, *Phys. Rev. Lett.* **23**, 1167 (1969).
- [205] D. E. Polk, D. S. Boudreaux, *Phys. Rev. Lett.* **31**, 92 (1973).
- [206] P. Steinhardt, R. Alben, M. G. Duffy, D. E. Polk, *Phys. Rev. B* **8**, 6021 (1973).
- [207] P. H. Gaskell, I. D. Tarrant, *Philos. Mag. B* **42**, 265 (1980).
- [208] J. Sarnthein, A. Pasquarello, R. Car, *Phys. Rev. Lett.* **74**, 4682 (1995).
- [209] C. Massobrio, A. Pasquarello, R. Car, *Phys. Rev. B* **64**, 144205 (2001).
- [210] G. Fabricius, E. Artacho, D. Sánchez-Portal, P. Ordejón, D. A. Drabold, J. M. Soler, *Phys. Rev. B* **60**, R16283 (1999).
- [211] N. Mousseau, G. T. Barkema, *Phys. Rev. B* **61**, 1898 (2000).
- [212] X. Liu, B. E. White, Jr., R. O. Pohl, E. Iwanizcko, K. M. Jones, A. H. Mahan, B. N. Nelson, R. S. Crandall, S. Veprek, *Phys. Rev. Lett.* **78**, 4418 (1997).
- [213] S. Roorda, W. C. Sinke, J. M. Poate, D. C. Jacobson, S. Dierker, B. S. Dennis, D. J. Eaglesham, F. Spaepen, P. Fuoss, *Phys. Rev. B* **44**, 3702 (1991).
- [214] M. F. Thorpe, S. W. de Leeuw, *Phys. Rev. B* **33**, 8490 (1986).
- [215] P. Adhikari, M. Xiong, N. Li, X. Zhao, P. Rulis, W.-Y. Ching, *J. Phys. Chem. C* **120**, 15362 (2016).
- [216] D. N. Tafen, D. A. Drabold, *Phys. Rev. B* **68**, 165208 (2003).
- [217] D. N. Tafen, D. A. Drabold, *Phys. Rev. B* **71**, 054206 (2005).
- [218] K. O. Trachenko, M. T. Dove, M. J. Harris, V. Heine, *J. Phys. Condens. Matter* **12**, 8041 (2000).
- [219] M. G. Tucker, D. A. Keen, M. T. Dove, K. Trachenko, *J. Phys. Condens. Matter* **17**, S67 (2005).

- [220] L. J. Abbott, K. E. Hart, C. M. Colina, *Theor. Chem. Acc.* **132**, 1334 (2013).
- [221] A. W. Thornton, K. E. Jelfs, K. Konstas, C. M. Doherty, A. J. Hill, A. K. Cheetham, T. D. Bennett, *Chem. Commun.* **52**, 3750 (2016).
- [222] I. Bechis, A. F. Sapnik, A. Tarzia, E. H. Wolpert, M. A. Addicoat, D. A. Keen, T. D. Bennett, K. E. Jelfs, *Chem. Mater.* **34**, 9042 (2022).
- [223] D. Hofmann, L. Fritz, J. Ulbrich, C. Schepers, M. Böhning, *Macromol. Theory Simul.* **9**, 293 (2000).
- [224] G. S. Larsen, P. Lin, K. E. Hart, C. M. Colina, *Macromolecules* **44**, 6944 (2011).
- [225] A. F. Sapnik, I. Bechis, S. M. Collins, et al., *Nat. Commun.* **12**, 2062 (2021).
- [226] Z. Hu, Y. Chen, J. Jiang, *J. Chem. Phys.* **134**, 134705 (2011).
- [227] R. Kaplow, T. A. Rowe, B. L. Averbach, *Phys. Rev.* **168**, 1068 (1968).
- [228] R. L. McGreevy, L. Pusztai, *Mol. Simul.* **1**, 359 (1988).
- [229] V. Krayzman, I. Levin, M. G. Tucker, *J. Appl. Crystallogr.* **41**, 705 (2008).
- [230] V. M. Nield, D. A. Keen, R. L. McGreevy, *Acta Crystallogr. Sect. A Found. Crystallogr.* **51**, 763 (1995).
- [231] S. J. Gurman, R. L. McGreevy, *J. Phys. Condens. Matter* **2**, 9463 (1990).
- [232] T. Zarrouk, R. Ibragimova, A. P. Bartók, M. A. Caro, *J. Am. Chem. Soc.* **146**, 14645 (2024).
- [233] H. Sheng, W. Luo, F. Alamgir, et al., *Nature* **439**, 419 (2006).
- [234] R. L. McGreevy, *J. Phys.: Condens. Matter* **13**, R877 (2001).
- [235] T. R. Welberry, B. D. Butler, *J. Appl. Crystallogr.* **27**, 205 (1994).
- [236] S. Kugler, L. Pusztai, L. Rosta, P. Chieux, R. Bellissent, *Phys. Rev. B* **48**, 7685 (1993).
- [237] O. Gereben, L. Pusztai, *Phys. Rev. B* **50**, 14136 (1994).
- [238] M. J. Cliffe, M. T. Dove, D. Drabold, A. L. Goodwin, *Phys. Rev. Lett.* **104**, 125501 (2010).
- [239] M. J. Cliffe, A. L. Goodwin, *J. Phys. Condens. Matter* **25**, 454218 (2013).
- [240] M. J. Cliffe, A. L. Goodwin, *Phys. Status Solidi B* **250**, 949 (2013).
- [241] A. L. Goodwin, F. M. Michel, B. L. Phillips, D. A. Keen, M. T. Dove, R. J. Reeder, *Chem. Mater.* **22**, 3197 (2010).
- [242] M. G. Tucker, D. A. Keen, M. T. Dove, A. L. Goodwin, Q. Hui, *J. Phys.: Condens. Matter* **19**, 335218 (2007).
- [243] G. Opletal, T. Petersen, B. O'Malley, I. Snook, D. G. McCulloch, N. A. Marks, I. Yarovsky, *Mol. Simul.* **28**, 927 (2002).

- [244] T. Petersen, I. Yarovsky, I. Snook, D. G. McCulloch, G. Opletal, *Carbon* **41**, 2403 (2003).
- [245] G. Opletal, T. C. Petersen, D. G. McCulloch, I. K. Snook, I. Yarovsky, *J. Phys. Condens. Matter* **17**, 2605 (2005).
- [246] G. Opletal, T. Petersen, I. Snook, S. Russo, *Comput. Phys. Commun.* **184**, 1946 (2013).
- [247] N. A. Marks, *Phys. Rev. B* **63**, 035401 (2000).
- [248] M. Z. Bazant, E. Kaxiras, J. F. Justo, *Phys. Rev. B* **56**, 8542 (1997).
- [249] R. Vink, G. Barkema, W. van der Weg, N. Mousseau, *J. Non-Cryst. Solids* **282**, 248 (2001).
- [250] F. H. Stillinger, T. A. Weber, *Phys. Rev. B* **31**, 5262 (1985).
- [251] M. Posselt, A. Gabriel, *Phys. Rev. B* **80**, 045202 (2009).
- [252] A. Jack, M. Levitt, *Acta Crystallogr. Sect. A* **34**, 931 (1978).
- [253] A. K. Soper, *Phys. Rev. B* **72**, 104204 (2005).
- [254] S. M. Clark, B. Colas, D. E. Jacob, J. C. Neufeind, H.-W. Wang, K. L. Page, A. K. Soper, P. I. Schodder, P. Duchstein, B. Apeleo Zubiri, T. Yokosawa, V. Pipich, D. Zahn, E. Spiecker, S. E. Wolf, *Sci. Rep.* **12**, 6870 (2022).
- [255] G. Cobourne, G. Mountjoy, J. D. Rodriguez-Blanco, L. G. Benning, A. C. Hannon, J. R. Plaisier, *J. Non-Cryst. Sol.* **401**, 154 (2014).
- [256] T. C. Nicholas, A. L. Goodwin, V. L. Deringer, *Chem. Sci.* **11**, 12580 (2020).
- [257] T. Nicholas, CHIC: Coarse-graining hybrid iorganic crystals, <https://github.com/tcnicholas/chic> (2024). GitHub repository.
- [258] 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, *Nat. Chem.* **16**, 36 (2024).
- [259] 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, *Comp. Phys. Commun.* **271**, 108171 (2022).
- [260] J. Janssen, PyLammpsMPI (version 0.0.8), <https://github.com/pyiron/pylammpsmpi> (2022).
- [261] C. T. Chantler, *J. Phys. Chem. Ref. Data* **24**, 71 (1995).
- [262] C. Chantler, *J. Synchrotron Radiat.* **8**, 1124 (2001).
- [263] Y. G. Bushuev, A. R. Finney, P. M. Rodger, *Cryst. Growth Des.* **15**, 5269 (2015).
- [264] A. C. S. Jensen, S. Imberti, S. F. Parker, E. Schneck, Y. Politi, P. Fratzl, L. Bertinetti, W. J. E. M. Habraken, *J. Phys. Chem. C* **122**, 3591 (2018).

- [265] J. Ihli, W. C. Wong, N. E. H., Y.-Y. Kim, A. N. Kulak, H. K. Christenson, M. J. Duer, F. C. Meldrum, *Nat. Commun.* **5**, 3169 (2014).
- [266] H. Nebel, M. Neumann, C. Mayer, M. Epple, *Inorg. Chem.* **47**, 7874 (2008).
- [267] J. W. Singer, A. Ö. Yazaydin, R. J. Kirkpatrick, G. M. Bowers, *Chem. Mater.* **24**, 1828 (2012).
- [268] A. E. Stones, R. P. A. Dullens, D. G. A. L. Aarts, *Phys. Rev. Lett.* **123**, 098002 (2019).
- [269] J. R. Henderson, *Mol. Phys.* **48**, 389 (1983).
- [270] A. E. Stones, R. P. A. Dullens, D. G. A. L. Aarts, *J. Chem. Phys.* **148**, 241102 (2018).
- [271] B. Widom, *J. Chem. Phys.* **39**, 2808 (1963).
- [272] G. Cinacchi, S. Torquato, *Phys. Rev. E* **100**, 062902 (2019).
- [273] M. Engel, H.-R. Trebin, *Phys. Rev. Lett.* **98**, 225505 (2007).
- [274] M. Engel, H.-R. Trebin, *Z. Kristallogr.* **223**, 721 (2008).
- [275] M. Elenius, F. H. M. Zetterling, M. Dzugutov, D. C. Fredrickson, S. Lidin, *Phys. Rev. B* **79**, 144201 (2009).
- [276] M. Mihalkovič, C. L. Henley, *Phys. Rev. B* **85**, 092102 (2012).
- [277] P. Zhou, J. C. Proctor, G. van Anders, S. C. Glotzer, *Mol. Phys.* **117**, 3968 (2019).
- [278] J. Dshemuchadse, P. F. Damasceno, C. L. Phillips, M. Engel, S. C. Glotzer, *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2024034118 (2021).
- [279] M. Rechtsman, F. Stillinger, S. Torquato, *Phys. Rev. E* **73**, 011406 (2006).
- [280] S. V. Sukhomlinov, M. H. Müser, *J. Chem. Phys.* **152**, 194502 (2020).
- [281] A. Suematsu, A. Yoshimori, M. Saiki, J. Matsui, T. Odagaki, *J. Mol. Liq.* **200**, 12 (2014).
- [282] J. M. Ziman, *Models of disorder. The theoretical physics of homogeneously disordered systems* (Cambridge University Press, Cambridge, 1979).
- [283] R. Moessner, A. P. Ramirez, *Phys. Today* **59**, 24 (2006).
- [284] T. C. Nicholas, E. V. Alexandrov, V. A. Blatov, A. P. Shevchenko, D. M. Proserpio, A. L. Goodwin, V. L. Deringer, *Chem. Mater.* **33**, 8289 (2021).
- [285] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, A. Walsh, *Nature* **559**, 547 (2018).
- [286] K. M. Jablonka, D. Ongari, S. M. Moosavi, B. Smit, *Chem. Rev.* **120**, 8066 (2020).
- [287] A. S. Rosen, S. M. Iyer, D. Ray, Z. Yao, A. Aspuru-Guzik, L. Gagliardi, J. M. Notestein, R. Q. Snurr, *Matter* **4**, 1578 (2021).

- [288] M. Eddaoudi, D. B. Moler, H. Li, B. Chen, T. M. Reineke, M. O’Keeffe, O. M. Yaghi, *Acc. Chem. Res.* **34**, 319 (2001).
- [289] O. M. Yaghi, M. O’Keeffe, N. W. Ockwig, H. K. Chae, M. Eddaoudi, J. Kim, *Nature* **423**, 705 (2003).
- [290] J. R. Long, O. M. Yaghi, *Chem. Soc. Rev.* **38**, 1213 (2009).
- [291] D. J. Tranchemontagne, J. L. Mendoza-Cortés, M. O’Keeffe, O. M. Yaghi, *Chem. Soc. Rev.* **38**, 1257 (2009).
- [292] O. Delgado-Friedrichs, M. O’Keeffe, *Acta Crystallogr., Sect. A* **59**, 351 (2003).
- [293] F. M. Amombo Noa, E. Svensson Grape, S. M. Brülls, O. Cheung, P. Malmberg, A. K. Inge, C. J. McKenzie, J. Mårtensson, L. Öhrström, *J. Am. Chem. Soc.* **142**, 9471 (2020).
- [294] A. P. Shevchenko, E. V. Alexandrov, A. A. Golov, O. A. Blatova, A. S. Duyunova, V. A. Blatov, *Chem. Commun.* **56**, 9616 (2020).
- [295] A. U. Ortiz, A. Boutin, A. H. Fuchs, F.-X. Coudert, *Phys. Rev. Lett.* **109**, 195502 (2012).
- [296] A. U. Ortiz, A. Boutin, A. H. Fuchs, F.-X. Coudert, *J. Chem. Phys.* **138**, 174703 (2013).
- [297] I. E. Collings, M. G. Tucker, D. A. Keen, A. L. Goodwin, *CrystEngComm* **16**, 3498 (2014).
- [298] A. Sadeghi, S. A. Ghasemi, B. Schaefer, S. Mohr, M. A. Lill, S. Goedecker, *J. Chem. Phys.* **139**, 184118 (2013).
- [299] L. Zhu, M. Amsler, T. Fuhrer, B. Schaefer, S. Faraji, S. Rostami, S. A. Ghasemi, A. Sadeghi, M. Grauzinyte, C. Wolverton, S. Goedecker, *J. Chem. Phys.* **144**, 034203 (2016).
- [300] S. De, A. P. Bartók, G. Csányi, M. Ceriotti, *Phys. Chem. Chem. Phys.* **18**, 13754 (2016).
- [301] A. Raza, A. Sturluson, C. M. Simon, X. Fern, *J. Phys. Chem. C* **124**, 19070 (2020).
- [302] Y. G. Chung, J. Camp, M. Haranczyk, B. J. Sikora, W. Bury, V. Krungleviciute, T. Yildirim, O. K. Farha, D. S. Sholl, R. Q. Snurr, *Chem. Mater.* **26**, 6185 (2014).
- [303] T. Xie, J. C. Grossman, *Phys. Rev. Lett.* **120** (2018). Cited by: 1285; All Open Access, Green Open Access.
- [304] S. M. Moosavi, A. Nandy, K. M. Jablonka, D. Ongari, J. P. Janet, P. G. Boyd, Y. Lee, B. Smit, H. J. Kulik, *Nat. Commun.* **11**, 4068 (2020).
- [305] S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson, G. Ceder, *Comp. Mater. Sci.* **68**, 314 (2013).

- [306] S. R. Heller, A. McNaught, I. Pletnev, S. Stein, D. Tchekhovskoi, *J. Cheminf.* **7**, 23 (2015).
- [307] V. I. Isaeva, K. E. Papathanasiou, L. M. Kustov, *Crystals* **10**, 617 (2020).
- [308] H.-X. Zhang, M. Liu, T. Wen, J. Zhang, *Coord. Chem. Rev.* **307**, 255 (2016).
- [309] J. Zhang, T. Wu, C. Zhou, S. Chen, P. Feng, X. Bu, *Angew. Chem., Int. Ed.* **48**, 2542 (2009).
- [310] O. M. Yaghi, G. Li, H. Li, *Nature* **378**, 703 (1995).
- [311] N. E. R. Zimmermann, A. Jain, *RSC Adv.* **10**, 6063 (2020).
- [312] D. J. Pearce, *Victoria University, Wellington, NZ, Tech. Rep* (2005).
- [313] S. Y. Joshi, S. A. Deshmukh, *Mol. Simul.* pp. 1–18 (2020).
- [314] N. Bernstein, G. Csányi, V. L. Deringer, *npj Comput. Mater.* **5**, 99 (2019).
- [315] B. Cheng, R.-R. Griffiths, S. Wengert, C. Kunkel, T. Stenczel, B. Zhu, V. L. Deringer, N. Bernstein, J. T. Margraf, K. Reuter, G. Csanyi, *Acc. Chem. Res.* **53**, 1981 (2020).
- [316] W. S. Torgerson, *Psychometrika* **17**, 401 (1952).
- [317] L. v. d. Maaten, G. Hinton, *J. Mach. Learn. Res.* **9**, 2579 (2008).
- [318] L. McInnes, J. Healy, N. Saul, L. Großberger, *J. Open Source Softw.* **3**, 861 (2018).
- [319] A. H. Larsen, *et al.*, *J. Phys.: Condens. Matter* **29**, 273002 (2017).
- [320] A. A. Hagberg, D. A. Schult, P. J. Swart, *Proceedings of the 7th Python in Science Conference* (Pasadena, CA USA, 2008), pp. 11–15.
- [321] L. Himanen, M. O. J. Jäger, E. V. Morooka, F. Federici Canova, Y. S. Ranawat, D. Z. Gao, P. Rinke, A. S. Foster, *Comput. Phys. Commun.* **247**, 106949 (2020).
- [322] 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, *J. Mach. Learn. Res.* **12**, 2825 (2011).
- [323] B. J. Frey, D. Dueck, *Science* **315**, 972 (2007).
- [324] K. Berijani, A. Morsali, *Inorg. Chem. Front.* **8**, 3618 (2021).
- [325] P. Chau, A. J. Hardwick, *Mol. Phys.* **93**, 511 (1998).
- [326] S.-M. Ying, J.-G. Mao, Y.-Q. Sun, H.-Y. Zeng, Z.-C. Dong, *Polyhedron* **22**, 3097 (2003).
- [327] S.-T. Wu, Y.-R. Wu, Q.-Q. Kang, H. Zhang, L.-S. Long, Z. Zheng, R.-B. Huang, L.-S. Zheng, *Angew. Chem., Int. Ed.* **46**, 8475 (2007).

- [328] M. T. Wharmby, S. Henke, T. D. Bennett, S. R. Bajpe, I. Schwedler, S. P. Thompson, F. Gozzo, P. Simoncic, C. Mellot-Draznieks, H. Tao, Y. Yue, A. K. Cheetham, *Angew. Chem., Int. Ed.* **54**, 6447 (2015).
- [329] S. Barthel, E. V. Alexandrov, D. M. Proserpio, B. Smit, *Cryst. Growth Des.* **18**, 1738 (2018).
- [330] H.-L. Jiang, T. A. Makal, H.-C. Zhou, *Coord. Chem. Rev.* **257**, 2232 (2013).
- [331] G. Verma, S. Butikofer, S. Kumar, S. Ma, *Top. Curr. Chem.* **378**, 4 (2019).
- [332] M. Gupta, J. J. Vittal, *Coord. Chem. Rev.* **435**, 213789 (2021).
- [333] L. Carlucci, G. Ciani, D. M. Proserpio, S. Rizzato, *Chem. – Eur. J.* **8**, 1519 (2002).
- [334] A.-X. Tian, J. Ying, J. Peng, J.-Q. Sha, H.-J. Pang, P.-P. Zhang, Y. Chen, M. Zhu, Z.-M. Su, *Cryst. Growth Des.* **8**, 3717 (2008).
- [335] Y.-P. He, Y.-X. Tan, J. Zhang, *CrystEngComm* **14**, 6359 (2012).
- [336] N. N. Greenwood, A. Earnshaw, *Chemistry of the Elements* (Butterworth-Heinemann, Oxford, 1997), second edn.
- [337] G. A. van Albada, M. Ghazzali, K. Al-Farhan, I. Mutikainen, J. Reedijk, *Inorg. Chem. Commun.* **14**, 162 (2011).
- [338] J. S. Ovens, D. B. Leznoff, *Inorg. Chem.* **56**, 7332 (2017).
- [339] T. Jacobs, R. Clowes, A. I. Cooper, M. J. Hardie, *Angew. Chem. Int. Ed.* **51**, 5192 (2012).
- [340] F. Luo, M.-S. Wang, M.-B. Luo, G.-M. Sun, Y.-M. Song, P.-X. Li, G.-C. Guo, *Chem. Commun.* **48**, 5989 (2012).
- [341] Z. Faure Beaulieu, T. C. Nicholas, J. L. A. Gardner, A. L. Goodwin, V. L. Deringer, *Chem. Commun.* **59**, 11405 (2023).
- [342] H.-C. Zhou, J. R. Long, O. M. Yaghi, *Chem. Rev.* **112**, 673 (2012).
- [343] T. D. Bennett, S. Horike, *Nat. Rev. Mater.* **3**, 431 (2018).
- [344] J. Fonseca, T. Gong, L. Jiao, H.-L. Jiang, *J. Mater. Chem. A* **9**, 10562 (2021).
- [345] T. D. Bennett, S. Cao, J. C. Tan, D. A. Keen, E. G. Bithell, P. J. Beldon, T. Friscic, A. K. Cheetham, *J. Am. Chem. Soc.* **133**, 14546 (2011).
- [346] Y. Lysogorskiy, C. van der Oord, A. Bochkarev, S. Menon, M. Rinaldi, T. Hammerschmidt, M. Mrovec, A. Thompson, G. Csányi, C. Ortner, R. Drautz, *npj Comp. Mater.* **7**, 97 (2021).
- [347] V. L. Deringer, G. Csányi, *Phys. Rev. B* **95**, 094203 (2017).
- [348] M. A. Caro, *Phys. Rev. B* **100**, 024112 (2019).
- [349] G. Csányi, S. Winfield, J. R. Kermode, A. De Vita, A. Comisso, N. Bernstein, M. C. Payne, *IoP Comput. Phys. Newsletter* p. Spring 2007 (2007).

- [350] S. Klawohn, J. R. Kermode, A. P. Bartók, *Mach. Learn.: Sci. Technol.* **4**, 015020 (2023).
- [351] ARCHER2 UK National Supercomputing Service, ARCHER2 UK National Supercomputing Service, <https://www.archer2.ac.uk>. Accessed: 2024-06-09.
- [352] M. W. Mahoney, P. Drineas, *Proc. Natl. Acad. Sci. USA* **106**, 697 (2009).
- [353] A. Bochkarev, Y. Lysogorskiy, S. Menon, M. Qamar, M. Mrovec, R. Drautz, *Phys. Rev. Mater.* **6**, 013804 (2022).
- [354] M. W. Finnis, J. E. Sinclair, *Philos. Mag. A* **50**, 45 (1984).
- [355] Y. Zhang, H. Wang, W. Chen, J. Zeng, L. Zhang, H. Wang, W. E, *Comput. Phys. Commun.* **253**, 107206 (2020).
- [356] E. V. Podryabinkin, A. V. Shapeev, *Comput. Mater. Sci.* **140**, 171 (2017).
- [357] S. A. Goreinov, I. V. Oseledets, D. V. Savostyanov, E. E. Tyrtysnikov, N. L. Zamarashkin, *How to find a good submatrix* (World Scientific, Singapore, 2010), pp. 247–256.
- [358] A. Fonarev, A. Mikhalev, P. Serdyukov, G. Gusev, I. Oseledets, *Proceedings of the 2016 IEEE 16th International Conference on Data Mining (ICDM)* (IEEE, Piscataway, NJ, 2016), pp. 141–150.
- [359] G. Kresse, J. Hafner, *Phys. Rev. B* **47**, 558 (1993).
- [360] G. Kresse, J. Furthmüller, *Comp. Mater. Sci.* **6**, 15 (1996).
- [361] G. Kresse, J. Furthmüller, *Phys. Rev. B* **54**, 11169 (1996).
- [362] G. Kresse, D. Joubert, *Phys. Rev. B* **59**, 1758 (1999).
- [363] J. Sun, A. Ruzsinszky, J. P. Perdew, *Phys. Rev. Lett.* **115**, 036402 (2015).
- [364] A. C. Hannon, *Nucl. Instrum. Methods Phys. Res. A* **551**, 88 (2005).
- [365] J. te Nijenhuis, M. Gateshki, M. J. Fransen, *Possibilities and limitations of X-ray diffraction using high-energy X-rays on a laboratory system* (Oldenbourg Wissenschaftsverlag, München, 2009), pp. 163–170.
- [366] S. Lovesey, *Theory of Neutron Scattering from Condensed Matter, Volume 1* (Oxford Scientific Publications, Oxford, 1984).
- [367] G. Squires, *Introduction to the Theory of Thermal Neutron Scattering* (Dover, New York, 1978).
- [368] D. Sivia, *Elementary Scattering Theory* (Oxford University Press, Oxford, 2011).
- [369] K. Kehr, *Streumethoden zur Untersuchung kondensierter Materie*, R. Hölzle, ed. (Forschungszentrum Jülich, Jülich, 1996), pp. A3.1–A3.38.
- [370] P. Virtanen, *et al.*, *Nat. Methods* **17**, 261 (2020).

- [371] L. Zoubritzky, F.-X. Coudert, *SciPost Chem.* **1**, 005 (2022).
- [372] A. Stukowski, *Modell. Simul. Mater. Sci. Eng.* **18**, 015012 (2009).
- [373] J. George, G. Hautier, A. P. Bartók, G. Csányi, V. L. Deringer, *J. Chem. Phys.* **153**, 044104 (2020).
- [374] C. Baerlocher, L. B. McCusker, Database of zeolite structures, <http://www.iza-structure.org/databases/>.
- [375] E. C. Spencer, R. J. Angel, N. L. Ross, B. E. Hanson, J. A. K. Howard, *J. Am. Chem. Soc.* **131**, 4022 (2009).
- [376] D. F. Thomas du Toit, V. L. Deringer, *J. Chem. Phys.* **159**, 024803 (2023).
- [377] P. I. Frazier, A tutorial on bayesian optimization (2018).
- [378] D. F. Thomas du Toit, V. L. Deringer, *J. Chem. Phys.* **159**, 024803 (2023).
- [379] J. Song, L. Frentzel-Beyme, R. Pallach, P. Kolodzeiski, A. Koutsianos, W.-L. Xue, R. Schmid, S. Henke, *J. Am. Chem. Soc.* **145**, 9273 (2023).
- [380] N. Ma, S. Horike, *Chem. Rev.* **122**, 4163 (2022).
- [381] G. Opletal, T. C. Petersen, S. P. Russo, A. S. Barnard, *J. Phys.: Mater.* **1**, 016002 (2018).
- [382] L. Zoubritzky, Liozou/periodicgraphs.jl: Extend graphs.jl to manipulate periodic graphs.
- [383] M. E. J. Newman, *Phys. Rev. Lett.* **89**, 208701 (2002).
- [384] T. Ko, S. Ong, *Nat. Comput. Sci.* **3**, 998 (2023).
- [385] D. Zhang, *et al.*, Dpa-2: Towards a universal large atomic model for molecular and material simulation (2023). DOI: 10.48550/arXiv.2312.15492.
- [386] I. Batatia, *et al.*, A foundation model for atomistic materials chemistry (2024). DOI: 10.48550/arXiv.2401.00096.
- [387] D. Golze, M. Hirvensalo, P. Hernández-León, A. Aarva, J. Etula, T. Susi, P. Rinke, T. Laurila, M. A. Caro, *Chem. Mater.* **34**, 6240 (2022).
- [388] Z. Chaker, M. Salanne, J.-M. Delaye, T. Charpentier, *Phys. Chem. Chem. Phys.* **21**, 21709 (2019).
- [389] R. Gurnani, Z. Yu, C. Kim, D. S. Sholl, R. Ramprasad, *Chem. Mater.* **33**, 3543 (2021).
- [390] T. Bailey, A. Jackson, R.-A. Berbece, K. Wu, N. Hondow, E. Martin, *J. Chem. Inf. Model.* **63**, 4545 (2023).
- [391] K. K. Bejagam, S. Singh, Y. An, S. A. Deshmukh, *J. Phys. Chem. Lett.* **9**, 4667 (2018).
- [392] H. Ye, W. Xian, Y. Li, *ACS Omega* **6**, 1758 (2021).
- [393] W. G. Noid, *J. Phys. Chem. B* **127**, 4174 (2023).
- [394] C. Zhou, L. Longley, A. Krajnc, *et al.*, *Nat. Commun.* **9**, 5042 (2018).