Data-driven understanding of disordered materials



Joe David Morrow

The Queen's College

A thesis submitted for the degree of Doctor of Philosophy

Supervisors: Prof. Volker Deringer & Prof. Andrew Goodwin

Inorganic Chemistry, University of Oxford

Declaration

The work described in this thesis was undertaken between October 2020 and November 2024 in the Inorganic Chemistry Laboratory, University of Oxford under the joint supervision of Prof. Volker Deringer and Prof. Andrew Goodwin. The thesis is the result of my original work; where it builds on the work of others, this is acknowledged through citations at the appropriate points in the text and in an acknowledgements section for each chapter. This thesis has not been submitted in whole or in part for a degree at this or any other institution.

Abstract

Machine learning (ML) interatomic potentials have become invaluable tools for atomistic simulations, achieving near quantum-chemical accuracy at extended length and time scales. For disordered materials, both slow quenching from the melt and large cells are precisely what is needed for realistic structural models, which make ML potentials especially useful for this application. This thesis describes methodological developments for further accelerating ML-driven simulations and their application to two important amorphous materials: silicon and molybdenum trisulfide.

First, we show how one ML potential model can be used to to train another, referred to as a teacher–student approach. We use an accurate, but computationally expensive model to generate training data for much faster potentials. Without the need for additional quantum-mechanical computations, extensive training datasets can be easily generated, and we find this improves the fast potentials.

Secondly, we use these fast potentials to prepare ultra-large structural models of amorphous silicon, through which we investigate rare coordination defects in the mostly fourfold random network using reliable statistics on thousands of individual defect examples. We combine structural descriptors and ML local energies to categorise over-coordinated defects into three types—a revision of the established 'floating-bond' model. We study defect-defect interactions and find that strain induced by defects on their neighbours leads to an energetic driving force for aggregation.

Thirdly, we parametrise an ML potential using a broad Random Structure Search database to model amorphous MoS_3, whose structure has long been debated. Constrained MD simulations discriminate between the main propositions for the structural units in MoS_3: trinuclear Mo clusters or extended Mo chains. Our results support the existence of triangular Mo_3 motifs, which agree better with total scattering data and are lower energy according to DFT than models based on chains. In large-scale MD simulations, we observe a tendency for triangles to align, which produces local nematic order in the amorphous phase. This could explain the easy decomposition of a-MoS_3 to form crystalline MoS_2.

Acknowledgements

I would like to thank my supervisors, Volker Deringer and Andrew Goodwin for their unfailing encouragement, enthusiasm, and support over the course of doctoral studies. Volker, it was a pleasure to be with the group '*ab initio*' and I am very appreciative of your direction and mentorship. I will enjoy following the excellent research coming from it in the coming years. Andrew, with you there was never a meeting that didn't fizz with ideas and suggestions with incredible breadth. You always made me feel as if my work and wider DPhil experience mattered.

I was privileged to work with Simon Cassidy during my time here, thank you for your patience, care, and attention in helping me carry out the experimental work on a-MoS_3 and for collecting X-ray data. Your invaluable expertise in data interpretation and ever-friendly manner made it productive and thoroughly enjoyable. I would also like to thank my collaborators for their scientific contributions and also advice and support: Chinonso Ugwumadu, Dave Drabold, Ed Pyzer-Knapp, Gabor Csányi, Luke Dicks, Michele Ceriotti, Raymond Chong, and Stephen Elliott.

The OxICFM CDT programme greatly enriched my DPhil experience, offering a friendly community and numerous opportunities for academic and personal growth. Especially, I want to thank Maitane for all her advice and assistance throughout.

To my colleagues and friends in the Deringer and Goodwin groups, thank you for the excellent company, stimulating discussions, and friendship. In particular, I would like to thank Amaarah and Muxue, who I had the pleasure of supervising for their excellent Part II years; Elodie, for your Gunrun-savvy and kindness to share it; Yuanbin, for a fun coding collaboration and thought-provoking discussion; Bianca, Litong, Andy, for being exemplary office-mates; Quentin, Louise, Natascia, for your energy and good humour; Tom, who was always interested in a discussion on disorder; Daniel, for your care and limitless interest in science and your fellow scientists; Zoe, for brightening up the office with your energy, games, and geniality; Zak, Yuxing, for your encouragement and company from the beginning, through to conferences, and the end; and John, for a friendship grounded in science, swimming, cycling, and stag. Long may it continue.

Outside of chemistry, I want to pay special thanks to the Oxford University Cross Country Club for filling my time here with yet more purpose, camaraderie, and, above all, a great group of friends. Heartfelt thanks are owed to Matt Seddon, friend and coach, for your care and conviction that have led to so much success for the club, and Vincent's club for financial support as well as providing, even more valuable, a home away from home populated by a wonderful community of sportspeople.

Finally, Kim, you have been a constant source of happiness and support. I can not wait for our next chapter.

Publications

The following publications contributed directly to this thesis:

- J. D. Morrow, V. L. Deringer;
  'Indirect learning and physically guided validation of interatomic potential models';
  J. Chem. Phys. **2022**, *157*, 10, 104105. (Ref. [1])

- J. D. Morrow, J. L. A. Gardner, V. L. Deringer;
  'How to validate machine-learned interatomic potentials';
  J. Chem. Phys. **2023**, *158*, 12, 121501. (Ref. [2])

- J. D. Morrow, C. Ugwumadu, D. A. Drabold, S. R. Elliot, A. L. Goodwin, V. L. Deringer;
  'Understanding Defects in Amorphous Silicon with Million-Atom Simulations and Machine Learning';
  Angew. Chem. Int. Ed. **2024**, *136*, 22, e202403842. (Ref. [3])

I was involved in the following publications that developed the validation of ML potential models discussed in the methods chapter:

- T. K. Stenczel, Z. El-Machachi, G. Liepuoniute, J. D. Morrow, A. P. Bartók, M. I. J. Probert, G. Csányi, V. L. Deringer;
  'Machine-learned acceleration for molecular dynamics in CASTEP';
  J. Chem. Phys. **2023**, *159*, 4, 044803. (Ref. [4])

- S. Chong, F. Grasselli, C. Ben Mahmoud, J. D. Morrow, V. L. Deringer, M. Ceriotti;
  'Robustness of Local Predictions in Atomistic Machine Learning Models';
  J. Chem. Theory Comput. **2023**, *19*, 22, 8020-8031. (Ref. [5])

Abbreviations

ACE atomic cluster expansion

ACSF atom-centred symmetry function

ADF angle distribution function

AIMD *ab initio* molecular dynamics

DFT density functional theory

FSDP first sharp diffraction Peak, e.g. in $S(Q)$

HAADF high-angle annular dark-field imaging

hcp hexagonal close-packed crystal structure

HRMC hybrid reverse monte carlo

GAP Gaussian approximation potential

GNN graph neural network

GST class of materials with varying compositions containing the elements Ge, Sb, and Te

ICOBI integrated crystal orbital bond index

KPCA kernel principal component analysis

LDA low density amorphous

LPR local prediction rigidity

MAE mean absolute error

MAS magic-angle spinning

ME mean error

ML machine learning

MLIP machine learning interatomic potential

MTP moment tensor potential

NMR nuclear magnetic resonance

NN neural network

pc polycrystalline

PES potential energy surface

PCA principal component analysis

QM quantum mechanics

RDF radial distribution function

RMC reverse monte carlo

RMSE root mean square error

sh simple hexagonal crystal structure

SOAP smooth overlap of atomic positions

STEM scanning transmission electron microscopy

SW Stillinger-Weber

t-SNE t-distributed stochastic neighbour embedding

VHDA very high density amorphous

VSEPR valence shell electron pair repulsion theory

XPS X-ray photoelectron spectroscopy

ZBL Ziegler–Biersack–Littmark (ZBL) screened nuclear repulsion

Contents

List of Figures **xii**

List of Tables **xv**

1 Introduction **1**

1.1 Atomistic modelling with ML potentials 3

1.2 Amorphous materials 6

1.2.1 a-Si 7

1.2.2 a-MoS_3 10

1.3 Research aims and thesis structure 11

2 Methods **13**

2.1 Acknowledgements 13

2.2 Introduction 14

2.3 Machine learning interatomic potentials 14

2.3.1 Data 14

2.3.2 Representation of chemical environments 23

2.3.3 Fitting 29

2.4 Density Functional Theory 44

2.4.1 DFT in periodic solids 47

2.4.2 Pseudopotentials 49

2.4.3 Dispersion corrections 50

2.5 Validation of ML potentials 52

2.5.1 Numerical errors: MAE, RMSE 53

2.5.2 Random Structure Search 59

2.6 Model building with molecular dynamics 61

2.6.1 Velocity Verlet integration 62

2.6.2 Thermostats and barostats 63
2.6.3 Constrained simulations 64
2.7 Dimensionality reduction 65
2.7.1 Principal Component Analysis 66
2.7.2 Kernel Principal Component Analysis 66
2.7.3 t-Distributed Stochastic Neighbour Embedding 67
2.8 Scattering experiments and radial distribution functions 68

3 Teacher–student models for accelerated simulations 73
3.1 Acknowledgements 73
3.2 Introduction: the teacher–student concept 73
3.3 Essential methodology 76
3.4 Synthetic database 79
3.5 Results 82
3.5.1 Numerical errors 83
3.5.2 Physically-guided validation of ML potentials 85
3.6 Compression benchmark 85
3.7 Student MTPs outperform directly-learned MTPs 89
3.8 Application: 1-million atom simulations of Si 89
3.9 Outlook 95
3.10 Further methodological details 96
3.10.1 **D1** database 96
3.10.2 Test configurations 97
3.10.3 MTP to MTP indirect learning 99

4 Understanding defects in a-Si with million-atom structures 101
4.1 Acknowledgements 101
4.2 Introduction 101

4.3 Results . . . 106
4.3.1 Revision of the floating-bond model . . . 109
4.3.2 Locality of defect environments . . . 113
4.4 Connection with defects in crystals . . . 117
4.5 Why do 5-fold defects cluster? . . . 118
4.6 Outlook . . . 122
4.7 Further methodological details . . . 122
4.7.1 Randomising cluster distributions . . . 122
4.7.2 Calculation of defect formation energy . . . 124
4.7.3 Molecular dynamics . . . 125

5 The hierarchical structure of amorphous MoS_3 127

5.1 Acknowledgements . . . 127
5.2 Introduction . . . 127
5.2.1 The experimental evidence . . . 129
5.2.2 The modelling evidence . . . 133
5.3 Structural models . . . 135
5.4 ACE fitting . . . 137
5.5 Modelling results . . . 137
5.6 Local structure of clusters . . . 143
5.7 Connections between clusters . . . 147
5.8 From connections to nematic order . . . 149
5.9 Pre-organisation for crystallisation . . . 152
5.10 Outlook . . . 153
5.11 Further methodological details . . . 154
5.11.1 Structural database . . . 154
5.11.2 Selection . . . 156

5.11.3 ACE potential fitting 159
5.11.4 Validation of potential for MoS_2 and MoS_3 161
5.11.5 Electronic Structure and Bonding Analysis 168
5.11.6 Synthesis 169
5.11.7 Characterisation 169
5.11.8 Computational Methods 170

6 Conclusions and future work **173**
6.1 ML potentials and their utility 173
6.2 Amorphous materials 174

Bibliography **177**

Appendices

A Understanding defects in a-Si with million-atom structures **208**

B The hierarchical structure of amorphous MoS_3 **213**

List of Figures

1.1 Machine Learning interatomic potentials 5

1.2 Total scattering data for amorphous Si. 9

2.1 Automated iterative RSS for fitting *de novo* potentials 21

2.2 Harmonic repulsions for RSS geometry optimisations 22

2.3 Descriptors for atomistic learning 25

2.4 Local Prediction Rigidity for silicon 36

2.5 Pseudopotentials for DFT 50

2.6 Disaggregated error reporting 55

2.7 Scaling of energy errors with system size 56

2.8 Physically-guided validation with an MD compression test 58

2.9 Physically-guided MD test in practice 59

2.10 Automated, system-agnostic validation with RSS test 60

2.11 Comparison of dimensionality reduction techniques 69

3.1 A teacher–student method for interatomic potentials 76

3.2 Composition of the ML potential fitting databases for teacher and student models 80

3.3 Typical failure mode during compression simulation 82

3.4 Numerical errors vs computational cost for directly learned and student potentials 84

3.5 Compression benchmark for validating potentials 87

3.6 Using pressure adjustment for benchmarking against known crystalline phases of Si 88

3.7 Systematic evaluation of MTPs via compression benchmark 90

3.8 Accuracy at $S(Q)$ prediction for selected MTP models vs. GAP-18 and experiment 91

3.9 Comparison of structural and energetic indicators for 1M-atom model with experiment 92

3.10 Application of accelerated student model to study a-Si at the 1-million atoms scale 94

3.11 MTP to MTP indirect learning 100

4.1 Defects in amorphous silicon from million-atom simulations 107

4.2 Logarithmic dependence of local-energy-derived bond strength on bond distance 109

4.3 Three categories of fivefold-connected defects 111

4.4 Locality of defect environments 115

4.5 Connection with defects in crystals 119

4.6 Clustering of 5-fold defects in amorphous silicon 120

4.7 Pair correlations for defect atoms 126

5.1 Proposed structural building blocks for MoS_3 128

5.2 A selection of known trinuclear molybdenum clusters 132

5.3 Schematic for constrained structure generation 136

5.4 Mo−S database structural map 139

5.5 Comparison of structural models with experimental data 140

5.6 Comparison of structural models with EXAFS data 141

5.7 Local structure of Mo_3 triangles 144

5.8 Classification of sulfur environments 145

5.9 Charge and bonding analysis 147

5.10 Calculation of nematic order 150

5.11 Medium-range correlations in a-MoS_3 151

5.12 Seeding of RSS structures with variable stoichiometry 155

5.13 Energy-based selection policy for RSS 158

5.14 MD failure due to cluster aggregation observed for the default ACE weighting scheme 160

5.15 ACE numerical errors for energy and force components 162

5.16 Illustrative RSS structures from energy error deciles 163

5.17 Contextualised dimer curves 164

5.18 Straining MoS_2 crystals. 165

5.19 Phonon dispersion curves for monolayer MoS_2 vs. DFT 165

5.20 Defect migration energy barriers in MoS_2 166

5.21 RSS validation via energy-volume distributions 167

5.22 Variable-temperature X-ray total scattering 170

A.1 Classification methods for defects 208

A.2 Absence of clustering of 3-fold defects in amorphous silicon 209

A.3 Comparison of clustering methods 209

A.4 Correlation of local energy with bond distances for 3-fold and 4-fold atoms 210

A.5 An alternative 1-million atom structure from rapid quenching and annealing. 211

A.6 Locality effects in the alternative 1-million atom structure. 212

B.1 Expanded version of $S(Q)$ in Fig. 5.5c 213

B.2 Additional $S(Q)$ results for alternative chain and triangle models . 213

B.3 Structural illustration of the chain models 214

B.4 Mo–Mo g(r) for alternating chain 215

B.5 Expanded version of nematic field in Fig. 5.11 216

List of Tables

3.1 Quality metrics for 100,000-atom LDA Si models produced by $10^{11}\,\mathrm{Ks}^{-1}$ quench simulations 91

3.2 General hyperparameters used for SOAP-based structural analysis 98

3.3 General hyperparameters used for all MTPs described in this chapter 98

3.4 Absolute timing information for the error analysis in Fig. 3.4 . . . 99

4.1 Calculation of defect formation energy 124

5.1 ACE fitting database composition by configuration type 138

5.2 Parameters for initial cell generation with keywords for `buildcell` 156

5.3 SOAP parameters for CUR selection and sketch map (Fig. 5.4) generation. 159

5.4 Basic ACE parameters for fitting during RSS. Where applicable, the same settings are applied to all bonds. 160

5.5 RSS ACE parameters, varying with generation as more data are accumulated. 160

1 | Introduction

At the time of writing, the Nobel prizes for physics and chemistry have just been awarded to those who developed technologies that underpin the work in this thesis; I subscribe to the belief that much of the scientific progress in the coming years will also derive from these advances. Neural networks (NNs), originally conceived of in the 1980s, only made a significant impact in 2012 with the debut of ImageNet for visual classification[6] and AlexNet for speech recognition,[7] both of which represented a leap in performance over other approaches and ultimately resulted in the prize for physics. At a similar time, Jörg Behler was developing NN-based interatomic potentials,[8] an area which had also seen pioneering early efforts in the 1990s to represent the potential energy surface (PES) of the H_2 molecule on a metal surface,[9,10] but languished due to a lack of computational power. The Behler-Parrinello NNs could be applied to systems with varying number of atoms and with much more generality than was possible previously, opening the door to simulations that could provide insights not previously available with *ab initio* methods, such as for phase–change material GeTe.[11] The Gaussian Approximation Potential (GAP) method,[12,13] shortly followed, which has been used to good effect for Machine learning (ML) driven simulations of amorphous solids such as silicon and carbon.[14,15]

In 2018, AlphaFold stunned at the 13$^{\text{th}}$ Critical Assessment of Structure Prediction (CASP-13) competition with its learned representation of the chemical interactions driving protein folding, a project which now accounts for the Nobel

prize for chemistry.[16] At the same time, the Atomic Cluster Expansion (ACE) descriptor[17] improved upon earlier encodings of atomic environments for ML interatomic potentials, such as the Behler-Parrinello atom-centred symmetry functions (ACSFs) and Smooth Overlap of Atomic Positions (SOAP) method, in a continuation of the quest for *ab initio* simulation accuracy at force-field cost. Now we see a blossoming of ML fitting frameworks with increasingly advanced architectures that blend symmetry-aware and chemically-natural graph-based input[18–20] with the high-performance GPU compute that has become synonymous, via scaling laws,[21] with the success of Large Language Models.

In a chemical context, and with ML potentials in particular, we seek not to learn the sum of human knowledge (as with language),[i] but only the interactions between atoms, which are governed by simple laws. Therefore, I am very optimistic about the promise of foundation models for chemistry, defined as ML models trained on large datasets that are capable of making predictions across virtually all of chemical space. In fact, several early iterations of such a concept have been released in the previous few months, produced by enterprises with very deep computational pockets, including Google's GNoME[20] and Microsoft's MatterSim.[24] This dramatic uptake in ML for chemistry has been fed by tremendous efforts to provide large open datasets of materials structures and associated computations, e.g. the Materials Project,[25] Novel Materials Discovery (NOMAD),[26] Automatic FLOW for Materials Discovery,[27] and the Open Quantum Materials Database.[28] Academic

[i]There are, however, language models that exist specifically for chemical applications, which seek to capitalise on the extensive back-catalogue of literature for, e.g. , organic synthesis.[22,23]

groups still remain competitive with corporations in the foundation model space, with MACE-OFF and AIMNet2 early examples of robust performance on most organic chemistry problems,[29,30] and MACE-MP-0 and DPA-2, more generally, for most elements in the periodic table.[31,32]

Despite the 'bitter lesson' of ML-based research over the previous years,[33] i.e. that simple algorithms which scale well with enormous compute consistently outperform carefully constructed algorithms imbued with human insights, I think there are nevertheless important contributions to be made by chemists with good intuition for validating and directing ML-driven efforts. A cautionary example can be found in limitations of high-throughput X-ray powder diffraction for conclusively identifying new phases, for which the autonomous efforts associated with GNoME[24,34] were criticised after publication.[35] More widely, a number of start-ups, many based in the UK, now seek to apply these data-driven technologies to solve some of the most pressing problems that society faces and only the chemical sciences can solve—whether that is designing enzymes for low-energy industrial reactions,[36] or discovering a carbon capture material capable of scaling us to net negative emissions.[37]

1.1 Atomistic modelling with ML potentials

The research in this thesis centres around the construction of ML interatomic potentials for use in materials chemistry simulations. ML is a group of statistical techniques for extracting useful information from large datasets with minimal prior knowledge of any laws generating the data. For our purposes, these data are

quantum-mechanical energy and force computations on collections of atoms, which certainly does derive from underlying laws—namely the Schrödinger equation. Why not just solve for Ψ_k if we know the physical laws that predict electronic structures and all the chemistry it causes? By learning from data with an ML potential, we seek to accurately approximate the quantum-mechanical potential energy surface (PES) in relevant regions of configuration space with a function that is, computationally, far cheaper to evaluate for a given structure—by a factor of around 10^6 in some cases. In this way, near-quantum-mechanical quality can be achieved for simulations on far longer timescales and lengthscales, which provides new answers, and new questions, for the design of functional materials. Molecular Dynamics (MD) simulations are perhaps the most intuitive way to study the behaviour of materials at the atomic scale: one replays how a collection of atoms evolves over time under almost any thermodynamic conditions, and, through this, can infer a vast array of properties, from the structures solids take, to their chemical reactions. One of the primary applications for ML potentials is accelerating MD simulations, which extends the range of materials and phenomena that can be studied compared to *ab initio* methods.

The three primary ingredients that go into developing an ML potential are outlined in Fig. 1.1, adapted from Ref. [38]. Step 1 is building the database: devising and selecting appropriate structures with enough chemical information to inform the fit, and calculating the energies and forces of these structures with an accurate reference method—most often Density Functional Theory (DFT). Step 2 involves encoding the local environment of atoms in the database in a mathematical

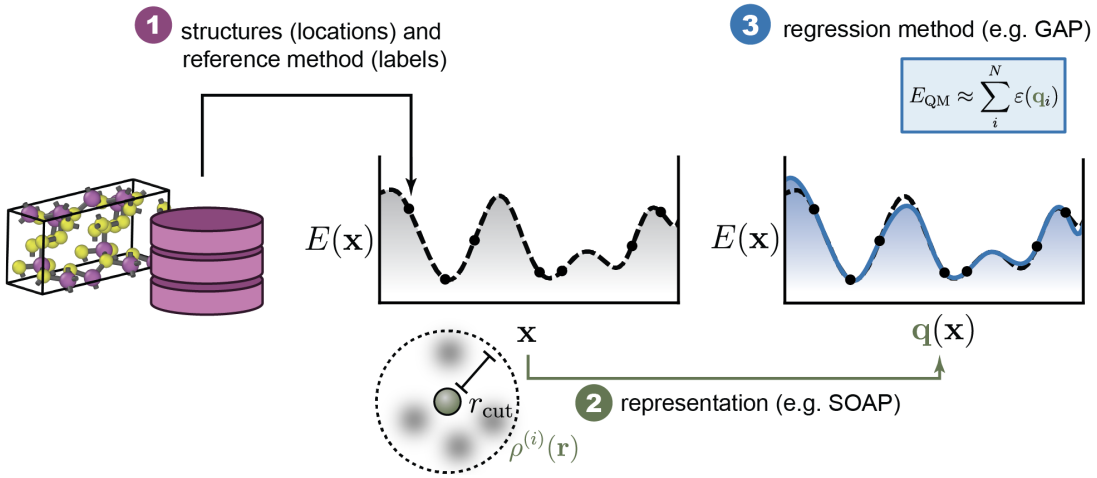


Figure 1.1: Machine Learning interatomic potentials. (**1**) Building the database, which consists of structural snapshots (black dots) that sample the PES (dashed black line). 'Structures' here means Cartesian coordinates along with QM forces and a total energy for each entry. The snapshot pictured is an MoS_3 configuration from a Random Structure Search (*vide infra*) (**2**) The representation: a mathematical encoding, $\mathbf{q}$, of Cartesian coordinates, $\mathbf{x}$, that is convenient for learning. This is usually a vector representing the local environment of each atom, which is invariant to translations, rotations, and permutations of atoms. (**3**) Fitting the data. The QM total energy is decomposed into atomic contributions, ε_i, which are a function of the local environments, and regression is performed (blue line) to minimise the error in energy/force predictions on the training database. Adapted from Ref. [38].

form that is convenient for Step 3: the fitting of a flexible function to the reference data. While I focus primarily on the development of Step 1, understanding and practising Steps 2 and 3 are also of critical importance for the performance of the resulting potentials.

There are a number of existing frameworks for Steps 2 and 3, which have been demonstrated to be highly successful in, for example, describing reactive organic molecules,[39] providing a general purpose model for the class of ternary phase-change materials based on Ge–Sb–Te (GST),[40] and for high entropy alloys.[41] Two methods of particular interest are the Gaussian Approximation Potential (GAP) framework,[12] which employs Gaussian process regression to make predictions

on 'unseen' structures based on their similarity to those in the training database, and the Smooth Overlap of Atomic Positions (SOAP) descriptor,[13] which represents atomic environments as an overlap integral, thereby defining this 'similarity'. In the next chapter, I introduce these methods in more detail, along with other such frameworks that I have used in the research of this thesis.

1.2 Amorphous materials

Disordered materials, those which can not be defined by a space group symmetry, lattice parameters, and special positions, have had less attention from large-scale efforts to discover new materials with ML-driven modelling. However, the advantages that ML potentials provide are, perhaps, uniquely suited to the needs of modelling disordered materials. Firstly, large simulation cells are required with a representative chunk of structure—to approximate well the lack of periodicity and to capture sufficient medium-range order in the structure.[42] Secondly, long simulation times are needed, most often in simulated cooling simulations that mimic the physical formation of the majority of amorphous minerals. The cooling must be slow enough—usually slower than, say, $10^{11}\,\mathrm{K\,s^{-1}}$ in the case of amorphous silicon—to allow the development of this medium-range order, if the hope is to obtain a structural model that is consistent with experimental samples.

In light of these advantages offered for disordered materials, it is not surprising that there have been numerous successful ML potentials specifically engineered for liquid and amorphous structures. Notable examples include models for the more structurally diverse elements, such as gallium,[43] phosphorus,[44] silicon,[45]

and carbon;[46,47] but also, increasingly, more chemically complex phases such as SiO_2,[48] GST,[40] and the class of battery electrolyte materials based on Li_2S–P_2S_5.[49]

1.2.1 a-Si

The modelling of disordered materials has a history about as long as computers,[50–54] but that started from the position, admitted by Zachariasen in 1932, that "we know practically nothing about the atomic arrangement in glasses".[55] He was chiefly concerned with amorphous silica (SiO_2), the major component of perhaps the most ubiquitous amorphous material (window glass), and introduced the 'Continuous Random Network' (CRN) to conceptualise amorphous materials as locally chemically ordered with bonds analogous to those in crystals, but with no crystalline grains or well-defined order beyond the first few coordination shells of each atom. For silica, this means [SiO_4] tetrahedral units, mostly sharing vertices, that are linked together in a fashion that lacks any long-range order.

Amorphous silicon is one of the canonical examples of covalent disordered materials and is the subject of the first two chapters of this thesis. Experimentally, we have, principally, the radial distribution function (RDF) which contains the average pairwise correlations between atoms, and can be derived from total scattering experiments via their Fourier transform. Three key observations from the $G(r)$ for a-Si, shown in Fig. 1.2b and taken from Ref. [56], are (a) the average Si−Si bond length, which is the same as that for crystalline Si, (b) the integral of the first peak, which implies a coordination number close to $N = 4$, and (c) the lack of well-defined order past about 8 Å. The latter is equivalent to observing no

Bragg peaks in the X-ray scattering data, and indicates an absence of long-range order. Theoretical background for the collection and interpretation of scattering data is provided in section 2.8.

Therefore, experiments indicate that, assuming most silicon atoms are in a similar environment, we have on average a network of fourfold-connected Si atoms that, in the first coordination shell, resemble crystalline Si but are disordered on any longer length scale. Furthermore, broad signals observed in magic-angle spinning (MAS) solid-state ^{19}Si nuclear magnetic resonance (NMR) experiments[57] are consistent with a broad range of coordination *geometries* within that $N = 4$ number. As for many amorphous materials,[58–60] different preparation methods and sample history can lead to different structures. With the recent increased popularity of ball-milling techniques,[61] new amorphous phases are being discovered in materials even as well-studied as water ice.[62] For a-Si, the purest samples are typically obtained by self-ion implantation of a crystal wafer,[56,63,64] whereas commercial a-Si used for photovoltaics is produced by chemical vapour deposition from SiH_4.[65] Annealing the as-implanted structure leads to subtle changes in the $S(Q)$ associated with dihedral angles, and an accompanying lowering of energy as measured by calorimetry,[66,67] without triggering crystallisation. The presence of defects in a-Si, especially atoms which aberr from $N = 4$ coordination, is the subject of chapter 4.

To model a-Si, early approaches used a bond-switching algorithm (named the WWW method, after the authors) to randomise the topology of the crystalline

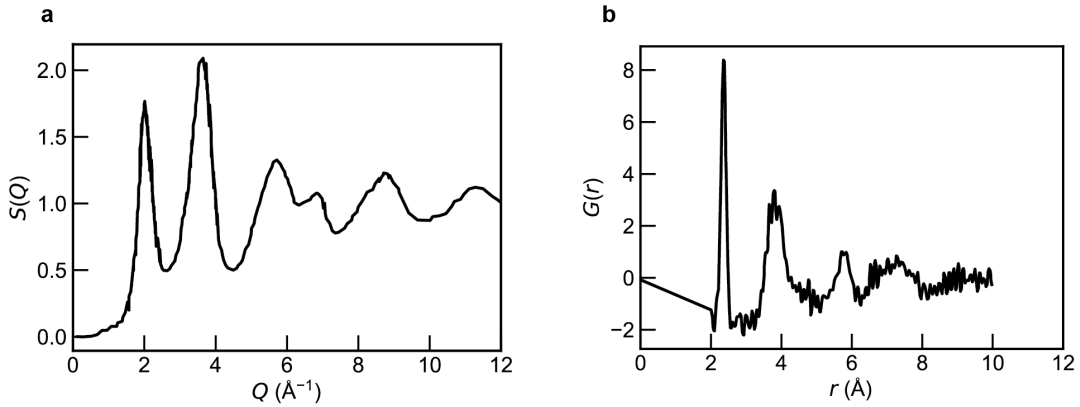


Figure 1.2: Total scattering, in the form of $S(Q)$, and accompanying Fourier transform as $G(r)$ for a-Si, drawn from data in Ref. [56]. Theoretical background for this data is provided in section 2.8.

a-Si network.[68,69] The empirical Keating potential,[70] which depends on bond lengths and angles with four defined neighbours, is able to relax the structure resulting from a bond-switch whilst maintaining the strict fourfold coordination. These so-called WWW structures provide a computational realisation of the CRN concept, with 'perfect' coordination. However, the bonds in these models are quite strained such that their energies, as calculated with either empirical potentials or, later, with *ab initio* methods, are higher than can be expected for experimental a-Si and do not remain defect-free after relaxation.[71]

An alternative 'energy-free' methodology for producing a model to fit the data is Reverse Monte Carlo (RMC).[72] In this approach, random moves are trialled and accepted with a Metropolis-like criterion that depends on whether they cause the model to agree better with the $S(Q)$.[50] For a-Si, the moves are usually nudges of atomic positions, but these can be more complex for other systems, e.g. molecular rotations or 'loop moves' in spin-ices.[73] Silicon provided an early indication that naive RMC could fit the $S(Q)$ *without* giving chemically sensible

structures (since any $S(Q)$ has no unique structure that gives rise to it). Modified versions of RMC, however, are more successful. These can be constrained via coordination number, bond angle distributions, or 'structural simplicity';[74] or can involve structural relaxations to achieve realistic structures.[75,76]

Direct simulation via quenching from molten Si provides a constraint-free approach, unlike WWW or RMC, to producing amorphous structural models. ML potentials have now provided a cost-effective means for conducting these simulations at high accuracy, first with Behler-Parrinello NNs,[8] and later with a GAP model.[45] It is with these direct MD simulations that I continue the discussion of amorphous silicon in chapter 4.

1.2.2 a-MoS_3

The compound with stoichiometry MoS_3 does not crystallise. Instead, it remains amorphous (a-MoS_3) until it undergoes desulfurisation above about 250 °C to the more familiar crystalline MoS_2, releasing sulfur vapour in the process. a-MoS_3 has attracted interest as a less-reactive substitute for elemental sulfur in candidate battery cathodes[77] and supercapacitors.[78] Additionally, its catalytic activity for the hydrogen evolution reaction (HER) is superior to that of MoS_2.[79] The preparation route to 'classic' a-MoS_3 requires careful thermal decomposition of $(NH_4)_2MoS_4$ under vacuum.[80] Alternative routes to what appears to be a similar amorphous structure include those via mechanochemical means,[81] precipitation from aqueous solution with HCl,[82] and electrodeposition as a thin film.[83,84]

There is uncertainty over the Mo−S structural units that predominate

in *a*-MoS_3, with different experiments and modelling approaches leading to two quite separate conclusions on the dominant motifs: metal-metal bonded Mo−Mo chains, or Mo_3 triangles.[82,85,86] The sulfur chemistry of trinuclear Mo in molecular complexes is itself a rich subject,[87,88] with much interest dedicated to biomimetic compounds related to the $[Mo_3S_4]^{4+}$ core of the cofactors in nitrogenase and HER enzymes. In the absence of clarity on its atomic-scale structure, however, the details of the catalytically-active sites or lithium migration pathways in the amorphous Mo−S phase remain obscure, so it is difficult to optimise the material systematically for its potential applications.

The combination of speed and accuracy offered by ML potentials is most advantageous when applied to systems where the nanoscale structure is of central interest, but quantum-mechanical accuracy is necessary to describe chemically- and electronically complex interactions: precisely the situation we find ourselves in with a-MoS_3. I explore this approach for modelling MoS_3 in chapter 5.

1.3 Research aims and thesis structure

The research of this thesis has two broad aims: firstly, to learn how to construct accurate and useful ML potentials, with a particular focus on the chemistry in the input data. Secondly, I aim to use ML potentials to understand the chemistry of amorphous materials, especially by generating realistic structural and analysing how they relate to familiar chemical principles. To achieve this, I primarily use MD simulations to produce large structural models, and a variety of structural analysis techniques to extract chemical features, including local energies, coarse-graining,

bonding analysis, and contributions to simulated total scattering. In the methods section, I outline the theoretical and experimental underpinnings of the research chapters. I focus on emphasising the similarities and differences between the three ML potential frameworks used—namely, GAP, MTP, and ACE—[17,89,90] by introducing them with a consistent notation. I also highlight the importance of the data that goes into fitting ML potentials—both the structures themselves (how much chemical information do they contain) and the quality of the reference 'labels' (i.e. the quantum-mechanical energies and forces). In chapter 3, I describe a method for further accelerating the predictions of ML potentials without a noticeable loss of accuracy. This enabled us to produce a 1-million atom structural model of amorphous silicon in close agreement with experimental data. In chapter 4, I present an investigation of the structural defects in this model, which makes use of the large system size to explore their typical structure and the interactions between defects. In chapter 5, I address the question of the local structure of a-MoS_3 by constructing an ML potential for the MoS_x phase diagram, and explore the consequences of the connections between motifs for medium range order in the structure. In each research chapter, I start by introducing the relevant literature and setting up the scientific questions we sought to address when carrying out the research. Similar to some journal article formats, I then present the main results and their interpretation, followed by the additional methodological material that, while not essential for understanding the scientific conclusions of the work, is necessary to validate our conclusions and provide enough detail to facilitate reproducibility.

2 | Methods

2.1 Acknowledgements

In this and subsequent chapters, figures, where declared 'adapted' in the caption from an article in which I was involved, are done so implicitly under the terms of the CC-BY 4.0 license, where details of the associated terms and conditions are available at https://creativecommons.org/licenses/by/4.0/.

The sections on validating ML potentials draw on the work of our Ref. [2], for which John Gardner designed the numerical errors plot (Fig. 2.7) and Volker Deringer drew Fig. 2.8. The derivation and conceptualisation of the Local Prediction Rigidity (LPR) was led by Sanggyu Chong and Michele Ceriotti. I contributed the choice of silicon and carbon structures as useful realistic applications of the LPR and to their generation, analysis, and chemical interpretation. Many of the heuristics used for Random Structure Search derive from Ref. [91], unless explicitly stated otherwise. Where snapshots of atomistic structures are shown, they are invariably drawn with the Ovito software package,[92] and I am indebted to other high-quality open source code for the manipulation and analysis of atomistic structures throughout the work described here. In particular, I highlight the Atomic Simulation Environment,[93] NetworkX (a graph-based analysis useful for the defects in chapter 4),[94] and Scikit-learn (for efficient implementations of many ML algorithms, including all those used for dimensionality reduction).[95] Other external software that I have used is cited alongside the relevant work.

2.2 Introduction

This chapter serves as a somewhat detailed reference for, and technical introduction to, the specific modelling and machine learning methodologies that are used in the results chapters to follow. In writing such a chapter, I aim to set out the tools used in this thesis in a single coherent, consistent manner to emphasise the connections between ML fitting methodologies, data acquisition approaches, and validation of the predictions of ML potentials. Essential details, however, are contained nearer the relevant results, such that the thesis can be read and understood excluding this chapter.

2.3 Machine learning interatomic potentials

Machine learning (ML) is the science of extracting information from large datasets. For my applications, the data consist of quantum-mechanical energies and forces for a set of periodic structures defined by coordinates, atomic numbers, and a unit cell. An ML *potential* is an approximate function that is fitted to a given potential energy surface (PES) whose parameters have been learned from reference data (the 'ground truth'). There are three main ingredients in the process of constructing an ML potential, which are outlined in Fig. 1.1 and discussed in detail in the following sections.

2.3.1 Data

The first step in developing an ML potential is building the reference database: contriving small-scale structural models that contain enough, and sufficiently

relevant, chemical environments to give the model high accuracy where property predictions are desired, as well as constraining the model for it to be valid for large regions of configuration space. These structures must be small so that we can afford to perform DFT (or higher-level) calculations for large numbers of them. Once representative structures (data 'locations') have been chosen, we calculate reference values using the ground-truth method: energies, forces, and stresses on the unit cell (data 'labels' in the ML jargon). Two aspects of the reference data are especially relevant for determining the quality of an ML potential, corresponding to the horizontal and vertical axes in Fig. 1.1.

Firstly, a shrewd choice of data locations is important, which can be guided by chemical knowledge, or by automated approaches for database building, such as active learning and random structure search.[91,96–99] Often, a combination of both works very well.[40,44] Even with large and widely-sampled datasets, there is still a risk of potentials extrapolating poorly outside the training domain and, consequently, displaying qualitatively incorrect behaviour.

Secondly, the quality of the data labels are important. Any ML potential can only, at best, reproduce the level of data on which it has been trained: for example, if the reference data labels have been computed with a DFT functional that is known to underestimate the dispersion interaction, so will the resulting ML potential. The role of the numerical quality and consistency of the data, e.g. that strict convergence of the k-point sampling is required (see sec. 2.4), was highlighted in Refs. [100–102], and has been more systematically analysed recently.[103]

Ab initio MD

Early ML potential developers would start with *ab initio* MD (AIMD) trajectories when fitting potentials: they are reliable, supported in a user-friendly way by most quantum chemistry codes, and require little prior knowledge of the chemistry beyond a choice of stoichiometry, density, and temperature. Under high temperature conditions, the melt was thought to sample a wide range of environments and thus provide the broad base on which to fit a robust ML potential.[104] However, AIMD is very expensive computationally and, perhaps surprisingly, high temperature trajectories turn out to sample quite a narrow region of configuration space.[105] This often leads to failures when a potential causes a configuration to wander anywhere even slightly away from such a region.[46,96] Examples of such failures are given in Figs. 3.7 and 5.14 in later chapters.

Domain-specific data

Often, researchers develop potentials for a chemical system with which they are already very familiar. Especially if they want to study a particular property of the system, such as a grain boundary or site defect, then examples of the relevant structure should be included in the database to maximise the quality of predictions for similar structures. Atoms should be randomly displaced by small amounts in these example structure, as non-zero forces on the atoms contain more information (as there are $3N$ of them) about the curvature of any local equilibria in the PES than does the single total energy. Handcrafting such models, however, is time-consuming and does not generalise well to new systems or permit high-throughput

studies. Work is under way to automatically generate relevant structures for common molecular and materials science applications, such as finding transition states[106] and studying defects.[107]

Iterative MD

One solution to both the sluggishness and the narrowness of AIMD is to let the ML potential take the reins as soon as possible. By fitting an early version of a potential to some sort of data—either a few frames of AIMD or even just the known crystal structures (which are readily accessible via resources such as The Materials Project[25])—it is possible to propagate dynamics for a short period, at which point the simulation can be stopped, a sample of snapshots extracted, and that batch of data added to the total database before fitting a new potential. Since the early, primitive potential will likely make chaotic predictions that sample a greater number of unusual structures (away from simple harmonic vibrations of bonds, for instance), these new data help to constrain future models to more sensible predictions. Proceeding like this in an iterative fashion to produce a boot-strapped ML potential ends up being faster than running a long AIMD trajectory, since ML potentials are orders of magnitude faster for MD, *and* providing better data coverage. The downside is that multiple stages of starting and stopping simulations cost valuable researcher-input time, and early versions of potentials can fail very early indeed, resulting rapidly in many useless structures, perhaps with atoms too close to converge their electronic structure with conventional DFT.

Synthetic data

In order to start fitting with iterative MD, we need some small database of initial structures. One approach to avoid AIMD entirely in acquiring this is to use structures generated via MD by some other potential. Not to be confused with data relating to chemical synthesis of molecules, 'synthetic data', in the ML jargon, refers to artificially generated data that are designed to mimic the distribution of the real-world data that we are trying to learn. In the chemical modelling context, this definition is somewhat ambiguous because almost all the data that we deal with is artificial to some extent,[i] whether that is coupled-cluster, DFT, or the predictions of empirical potentials. We use the term to refer to any data that are cheaper to acquire than the target property we are trying to learn, but sufficiently similar in distribution to serve as a useful surrogate target during a pre-training or fine-tuning step during fitting. As an example, it has been shown that using existing carbon data, with suitably rescaled bonds, can be useful for training a silicon model ('alchemical learning').[108]

Active learning

Instead of iterating through batches of MD simulations, why not refit continuously during a single MD trajectory and avoid having to restart calculations? The key question with this 'active learning' approach is to decide when an ML potential's predictions have become too unreliable to continue and a refit to new data is required. Various strategies have been developed for quantifying the uncertainty of

[i] with the exception of experimental data, which are not yet often incorporated into the fitting of potentials.[20]

the ML potential predictions. For neural networks, an ensemble of models trained from random initial parameters will show a high dispersion in their predictions far away from the training set, which can act as proxy for uncertainty.[97] As part of the Moment Tensor Potential (MTP) fitting framework, the early implementation of active learning, the 'extrapolation grade' was introduced, based on the distance to the volume spanned by the training set (in descriptor space, see section 2.3.2 for further explanation). Yet another useful proxy for the uncertainty would be the Bayesian predictive variance that is natural to Gaussian process regression models.[100]

A number of DFT codes now incorporate an active learning functionality as a sort of ML-accelerated simulation, where the details of fitting and database generation are dealt with 'under the hood' with little user input.[4,109] Producing potentials with this philosophy enables them to excel at a specific task, but typically they will the lack generality to take 'offline' and, importantly, are not necessarily scalable to system sizes that exceed the capability of DFT. More widely, active learning has been used as part of the database for successful general purpose potentials,[96,97,99] such as the example of Ref. [48] for the Si−O system, which can handle disorder, variable stoichiometry, and extremely high pressures.

Random Structure Search for *de novo* potentials

ML potentials have few physical constraints beyond rotation, translation, and permutation invariance. Therefore, far away from regions sampled in the training database, their predictions are quite likely to be unphysical. In producing many of

the successful general-purpose potentials currently available, choosing the structures to be included in the database was a time-consuming, iterative process. Early versions of a potential were, for the most part, based on snapshots from AIMD simulations to sample high temperature liquid and amorphous phases as outlined earlier. Separately, the known crystalline phases of a material with distortions were added to capture the energy and curvature of minima in the PES.[45,46,110] The performance of resulting potentials in predicting the energies of new structures was then evaluated, and any failings remedied using chemical intuition and trial and error, e.g. adding dimer data to capture the core-repulsion at very short separations, or adding small-scale models of crystalline defects to enable a better description of similar structures.[111]

Refs. [112] and [91] describe an alternative approach for exploring the PES in order to build databases. Instead of adding structures by hand, the principles of the Random Structure Search (RSS), introduced by Pickard *et al.* as a DFT-based crystal structure prediction algorithm,[113,114] were employed to sample configuration space automatically. In outline (and illustrated in Fig. 2.1), an early version of a potential is first initialised on a database of random 'sensible' structures (i.e. without close approaches of atoms and with some random symmetry operations). Next, this potential is used to relax a new set of random structures to their local minima. A diverse subset of these structures, including steps along each minimisation trajectory, is selected for DFT computations, and these are added to the database to refit a new potential. The previous step is repeated with the latest potential until predictions of satisfactory quality are achieved.

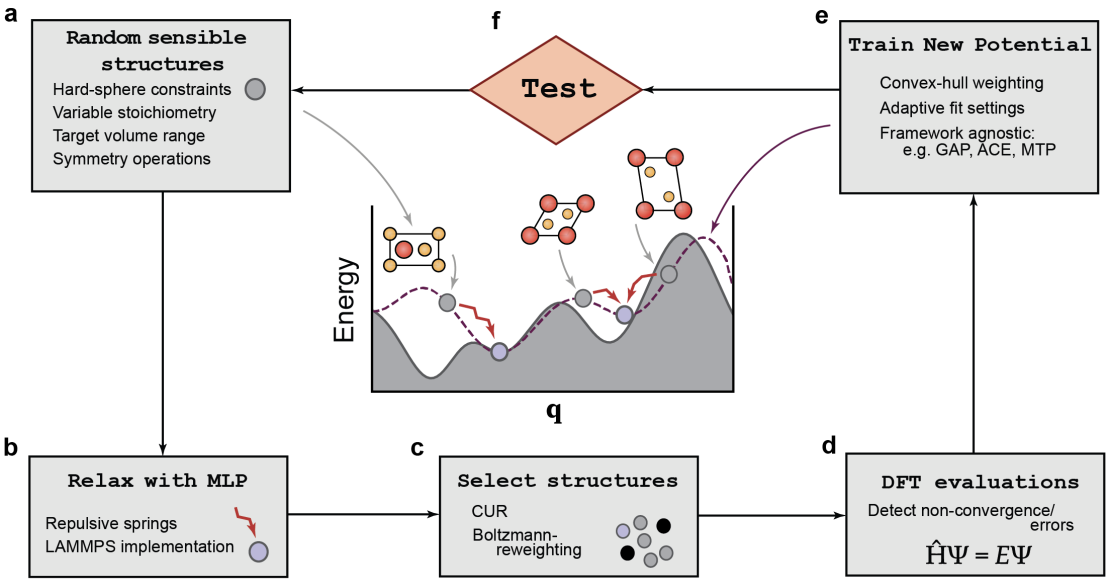


Figure 2.1: Automated iterative RSS for fitting *de novo* potentials. (**a**) Generation of initial 'sensible' random structures. (**b**) Relaxation with current iteration of machine learning potential. For iteration 0, there is no potential yet to use for geometry optimisation, so we simply take the random structures as-generated and continue. (**c**) Selection of structures. We adopt the two-stage procedure of Ref. [91]: CUR followed by Boltzmann-reweighting. (**d**) DFT computations. Extra care must be taken to avoid including poorly-converged results in the training data, which are more common when relaxing the unusual random geometries produced by RSS. (**e**) Refitting of the potential. (**f**) After sufficient accuracy on a test set is achieved, or a maximum computational budget exceeded, the iterations are terminated.

The main advantages of this approach are, firstly, that it is automatic: the most time-consuming step of potential development is the human input of designing the contents of the database, and, secondly, that it is robust: potentials are prone to predicting spuriously low energetic minima for unseen structures. For an example of such a problem, see the C-GAP-20 carbon potential of Ref. [115] which, after publication, was found to predict body-centred cubic as more stable than graphite. Through RSS, sufficient high-energy regions of configuration space, including an unbiased sample of crystal structures, are visited during training which ameliorates the possibility of 'holes' appearing in the potential's approximation of the PES. In other words, including high-energy structures is important because

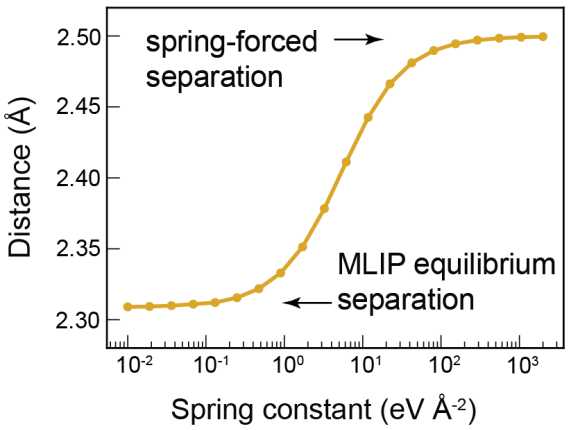


Figure 2.2: Training with stabilisers: harmonic repulsions for geometry optimisations with rough-and-ready RSS potentials. Below a cutoff, we apply a force of kx^2 to ensure the core overlap of atoms is avoided for potentials that have not yet learned this physical behaviour. This figure helps to calibrate the required spring constant for the Mo–Mo bond, with values of 10^1–10^2 eV Å^{-1} providing intermediate strengths that will enforce the constraint without producing troublesome large forces.

this bounds the region explored by a potential during its most common use case: MD simulations.

We use RSS extensively to produce the foundation of the dataset for our Mo–S potential, described in detail in chapter 5. We adopt many of the tricks developed by Pickard *et al.* for generating random 'sensible' structures,[114] such as applying random symmetry operations, targetting a volume range around the experimental quantity, biasing towards even numbers of atoms in the unit cell (as is observed in natural minerals[116]), and using hard-sphere constraints. We also develop heuristics of our own, including a two-step approach to achieving a broad range of S mole fractions for small cells whilst also focusing on the stoichiometric phases, and the use of harmonic repulsions during relaxations with early versions of the potential to avoid unstable optimisations.

2.3.2 Representation of chemical environments

The second step in developing an ML potential is to represent the local environments of the atoms in the database in a mathematical form that is amenable to learning. This representations is referred to as a descriptor in the community and is analogous to 'featurisation' in wider ML research. A descriptor should be invariant with respect to translations, rotations, and permutations of atoms (unlike, say, a list of Cartesian coordinates) in order to avoid requiring large quantities of data to teach the model these symmetries. Descriptors should also be as close to complete as possible, i.e. uniquely characterising every meaningfully different geometry, while remaining smooth and numerically efficient. One example of a descriptor is the Coulomb matrix, which is built from the pairwise distances between all atoms in the structure.[117] Its advantages are simplicity, interpretability, and invariances to translations and rotations, and it can work well for small molecules. However, it scales poorly ($\mathcal{O}(N^2)$) with number of atoms, is sensitive only to two-body correlations (incomplete), and is not smooth: for permutation invariance, the rows are sorted by the L2-norm, which can lead to discontinuities if distances change from structure to structure. Many established ML potential frameworks rely on carefully designed, physically-motivated descriptors based on radial and angular basis functions (not dissimilar to atomic orbital basis sets in conventional quantum chemistry).[13,17,118,119] Other approaches, usually with neural networks, learn the structural representation as part of the potential fit, after being supplied basic invariants such as bond lengths, angles, and dihedrals.[18,120–122] The zoo of structural descriptors for atomistic ML has been reviewed in Ref. [123], with

particular attention paid to the mathematical relationships between frameworks.

Smooth Overlap of Atomic Positions

The Smooth Overlap of Atomic Positions (SOAP) descriptor was originally developed in the context of ML potentials[13] and is widely employed in GAP fitting;[12] however, SOAP has has also been used for classifying and understanding complex structures across a wide range of chemical systems.[125] The SOAP formalism encodes the local environment of atom i as a neighbour density $\rho^{(i)}(\mathbf{r})$ by placing a 3D Gaussian function of width σ_{at} on atom i and each of its neighbours up to a (smoothed) cutoff,

$$\rho^{(i)}(\mathbf{r}) = \sum_j \exp\left[\frac{-|\mathbf{r}-\mathbf{r}_{ij}|^2}{2\sigma^2}\right] f_{\text{cut}}(r_{ij}) \ . \tag{2.1}$$

To enable efficient computations of integrals and rotational invariance, this density is expanded in a basis,

$$\begin{aligned} \rho^{(i)}(\mathbf{r}) &= \sum_{nlm} c^i_{nlm} R_n(r) Y^m_l(\hat{\mathbf{r}}) \\ c^i_{nlm} &= \int \mathrm{d}\mathbf{r} R_n(r)^* Y^m_l(\hat{\mathbf{r}})^* \rho_i(\mathbf{r}) \end{aligned} \tag{2.2}$$

where I have assumed only a single element for simplicity of presentation. The SOAP descriptor does extend to multiple elements; naively (having a separate channel for each element-element combination) this makes the size of the descriptor scale poorly with the number of elements in the system, but it is possible to define linear combinations of densities which pare this back without losing much information.[126] Rotational invariance is achieved by summing over m, to yield the 'power spectrum'

$$p^i_{nn'l} = \frac{1}{\sqrt{2l+1}} \sum_m (c^i_{nlm})^* c^i_{n'lm} \ . \tag{2.3}$$

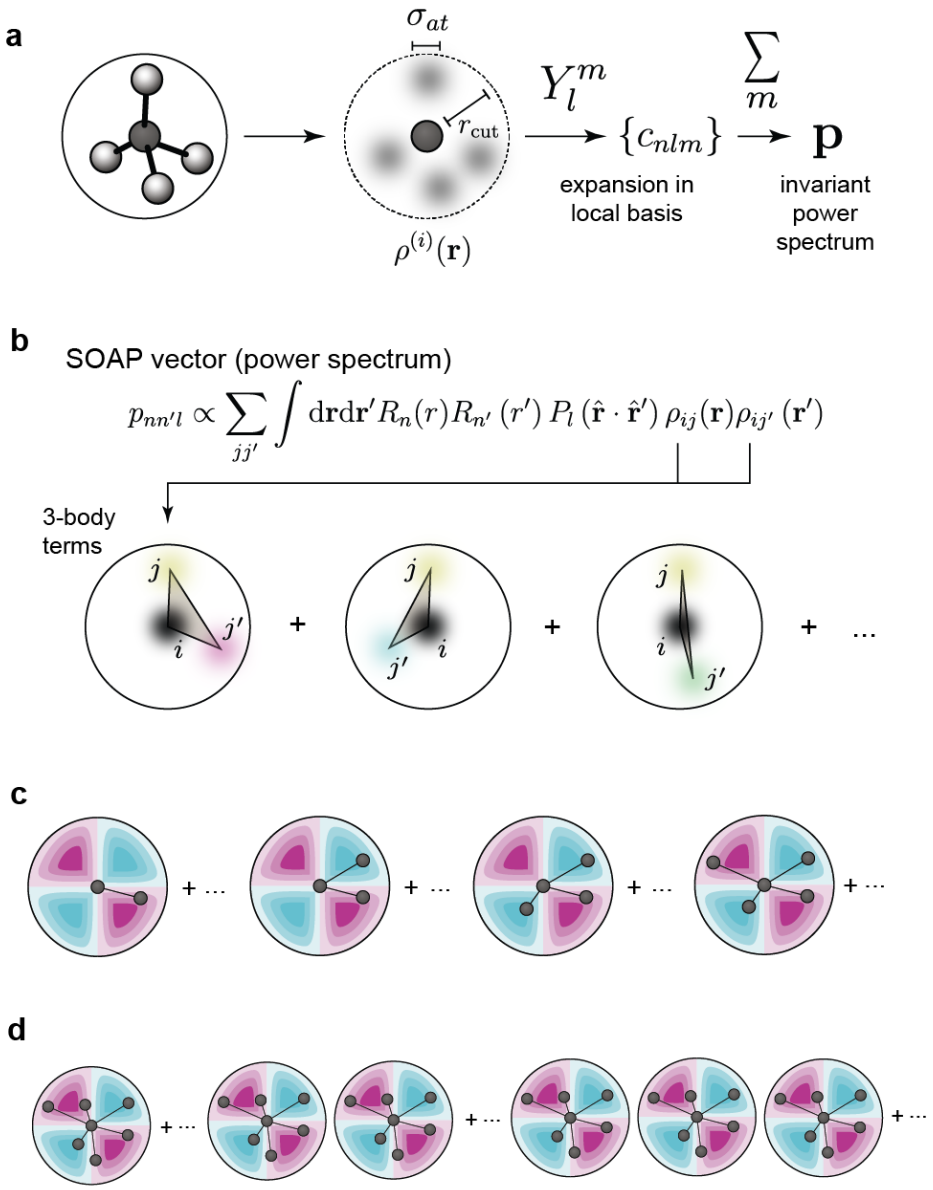


Figure 2.3: Descriptors for atomistic learning. (**a**) Schematic for the SOAP descriptor. The neighbour density, $\rho^{(i)}$, for atom i is permutationally invariant. After expansion in a local basis of spherical harmonics (Y_l^m) and radial functions (often spherical Bessel functions), we can sum over the m index to ensure the rotational invariance of the power spectrum, $\mathbf{p}$. This is analogous to the sum of directional p orbitals describing a spherical hydrogenic system. (**b**) Illustration of why the power spectrum is a three-body descriptor. (**c**) The atomic cluster expansion (ACE) as a sum of two-, three-, four-body, and higher order contributions (Eq. 2.3.2). (**d**) ACE rewritten as a polynomial of the atomic base terms, A_{iv_t} in Eq. 2.3.2, whose evaluation scales linearly with number of neighbours. The same atomic density is projected on to the basis functions in each term. The smoothness, in contrast with the SOAP σ parameter, is adjusted by the sensitivity of these basis functions. Adapted from Refs. [124] and [100].

A similarity measure (kernel) $k_{\mathrm{SOAP}}(i,j)$ between two environments is defined by evaluating the squared overlap of their respective densities, $\rho(r)$, integrated over all possible rotations, $\hat{R}$, of one of the local environments. The power spectrum makes this extremely convenient in the form of a simple inner product

$$\mathbf{p}_i \cdot \mathbf{p}_j \propto \int_{\hat{R}\in O_3} \mathrm{d}\hat{R} \left| \int \mathrm{d}\mathbf{r} \rho^i(\mathbf{r}) \rho^{i'}(\hat{R}\mathbf{r}) \right|^2 , \tag{2.4}$$

such that projecting on to the power spectra is the only time-consuming component of evaluating the SOAP kernel for structures. The SOAP descriptor was originally thought to be complete, but it has since been pointed-out that it is only sensitive to, at most, 3-body correlations.[127]

Moment Tensor Potentials

MTP models have been applied to a range of research questions, including phase diagrams,[128] reaction dynamics of small molecules,[129] and lithium ion conduction.[130] Full details of their construction are given in the original work by Shapeev and co-workers,[90,131] which we summarise briefly here. The descriptor of a local environment is defined as the moment tensor,

$$M_{\mu,\nu}(\mathbf{x}_i) = \sum_j f_\mu(|\mathbf{r}_{ij}|) \underbrace{\mathbf{r}_{ij} \otimes \cdots \otimes \mathbf{r}_{ij}}_{\nu \text{ times}}, \tag{2.5}$$

where $\mathbf{x}_i$ denotes the positions of the i^{th} atom and all neighbours j up to a cutoff, R_{cut}, which we set to 5 Å to be consistent with the descriptor used for GAP-18.[45] The radial part f_μ depends only on the position vectors $\mathbf{r}_{ij}$ of each j-th atom from the i-th one, and the angular information is encoded by tensors of rank ν produced from the repeated outer product of $\mathbf{r}_{ij}$ vectors. These $M_{\mu,\nu}$ are contracted to give

the basis functions B_α.

The number of basis functions defining the particular functional form of an MTP (and hence its flexibility and computational efficiency) grows exponentially with the important hyperparameter $\mathrm{lev}_{\mathrm{max}}$, where the level ("lev") is defined as

$$\mathrm{lev}(M_{\mu,\nu}) = 2 + 4\mu + \nu \tag{2.6a}$$

$$\mathrm{lev}(M_{\mu,\nu} : M_{\eta,\lambda}) = \mathrm{lev}(M_{\mu,\nu}) + \mathrm{lev}(M_{\eta,\lambda}) \tag{2.6b}$$

and ':' represents an appropriate tensor contraction operation. All possible $M_{\mu,\nu}$, and contractions of one or more $M_{\mu,\nu}$, are included as degrees of freedom in the functional form of an MTP, provided the resulting basis function satisfies $\mathrm{lev}(B_\alpha) \leq \mathrm{lev}_{\mathrm{max}}$. Like with SOAP, the evaluation of the M scales only linearly with the number of neighbours. The 'Moment' of the MTP acronym comes from the mechanical interpretation of the lower-order $M_{0,\nu}$: $M_{0,0}$ gives the number of atoms within distance r_{cut}, hence it is effectively the mass of the environment, $M_{0,1}$ gives the *centre* of mass, $M_{0,2}$ gives the second moments of inertia tensor, and so on. For $\mu > 0$, $M_{\mu,\nu}$ become weighted moments of inertia.

Atomic Cluster Expansion

The ACE descriptor[17,124,132] has been shown to unify a large subset of other density-based representations of atomic environments, including SOAP and MTPs, into a single theoretical framework, with those individual methods as special cases.[123] The form is inspired by the cluster formalism for alloys and the decades-old procedure of expanding the total energy into contributions from each atom, E_i, and these E_i

into many-body terms in the expression:[133]

$$\begin{aligned} E_i =& V_0\left(Z_i\right) + \sum_{j_1} V_1\left(\boldsymbol{r}_{ij_1}, Z_{j_1}; Z_i\right) + \sum_{j_1<j_2} V_2\left(\boldsymbol{r}_{ij_1}, Z_{j_1}, \boldsymbol{r}_{ij_2}, Z_{j_2}; Z_i\right) + \ldots \\ &+ \sum_{j_1<\cdots<j_N} V_N\left(\boldsymbol{r}_{ij_1}, Z_{j_1}, \ldots, \boldsymbol{r}_{ij_N}, Z_{j_N}; Z_i\right), \end{aligned} \tag{2.7}$$

where $(\boldsymbol{r}_i, Z_i)$ are the position and atomic number of atom i and $\mathbf{r}_{ij} = \mathbf{r}_i - \mathbf{r}_j$ are relative positions. The problem with this approach is the combinatorial scaling of the sums, $\sum_{j_1<\cdots<j_N}$, with body order, N, which becomes impractical for $N > 3$. Key to avoiding this scaling is the so-called 'density trick', also used by the SOAP and MTP formalisms. It is possible to expand the V_N terms in a body-ordered basis (suppressing the Z_i)

$$E_i = \sum_{\nu=1}^{\nu_{\max}} \sum_{\mathbf{v}} \boldsymbol{\theta}_{\mathbf{v}} \sum_{j_1,\ldots,j_\nu} \Phi_{\mathbf{v}}(\mathbf{r}_{j_1 i}, \ldots, \mathbf{r}_{j_\nu i}) \ , \tag{2.8}$$

where $\Phi_{\mathbf{v}}$ are ν-order basis functions (involving the coordinates of a cluster of ν neighbours), $\hat{\boldsymbol{\theta}}_{\mathbf{v}}$ are the model parameters obtained during fitting, and ν index the basis functions. Then, projecting the atomic density,

$$\rho_i(\mathbf{r}) = \sum_{j\neq i} \delta(\mathbf{r} - \mathbf{r}_{ij}) \ , \tag{2.9}$$

on to atom-pair basis functions $\phi_\nu(\mathbf{r})$ (more general than with SOAP) gives the atomic base

$$A_{iv} = \sum_{j\neq i} \phi_v(\mathbf{r}_{ij}) \ . \tag{2.10}$$

Choosing $\Phi_{\mathbf{v}}$ to be a basis formed from the tensor product of the atom-pair functions,

$$\Phi_{\mathbf{v}}(\mathbf{r}_{1i}, \ldots, \mathbf{r}_{\nu i}) = \prod_{t=1}^{\nu} \phi_{v_t}(\mathbf{r}_{ti}) \ , \tag{2.11}$$

leads to

$$\sum_{j_1,\ldots,j_r} \Phi_{\mathbf{v}}(\mathbf{r}_{j_1 i}, \ldots, \mathbf{r}_{j_r i}) = \prod_{t=1}^{\nu} A_{iv_t} \ . \tag{2.12}$$

The key here is that the combinatorially-increasing sum has been rewritten as a product of only ν terms, where the evaluation of each $A_{i\nu}$ only costs $\mathcal{O}(N)$ for N neighbours. It turns out that a recursive procedure can eliminate the ν-scaling entirely, leaving a highly efficient formalism that scales linearly with the number of neighbours. Further symmetrisation with respect to rotations with spherical harmonics and an explicit choice for the radial basis functions are still needed, but Eq. 2.3.2 outlines the advantage of ACE descriptors over SOAP and MTP equivalents: they are systematically improvable (i.e. achieve a higher body-order than SOAP, and are more stable and have fewer linear dependency issues than MTP),[134] and are extremely efficient to evaluate.

2.3.3 Fitting

The third step in ML potential development is to fit a flexible function (i.e. one with many parameters) to the reference data. Among the main classes of fitting approaches currently used for ML potentials, there are:

1. Kernel-based methods such as the Gaussian Approximation Potentials (GAP)[12] and symmetrised gradient-domain machine learning (sGDML) frameworks.[135,136]

2. Linear models including the Spectral Neighbor Analysis Potential (SNAP),[119] Moment Tensor Potential (MTP),[90] and Atomic Cluster Expansion (ACE) frameworks (the latter in its simplest flavour).[17]

3. Artificial neural network (NN) models, originally of the Behler–Parrinello-type,[8,104] and later related schemes such as ANI,,[137] DeepMD,[138,139] and

SchNet.[120] More recent developments have been based on message-passing graph NNs such as the NequIP,[18] MACE,[19] and AIMNet[121] schemes.

ML preliminaries

Before discussing the details of the fitting frameworks used in this thesis, it is helpful to define some ML-specific terminology (highlighted in bold below). Many ML models can nearly perfectly fit to the data on which they have been trained. A model's utility lies in its ability to generalise, so it is important to test on data points that have not been included in the training. This sets up the three sets of data used for fitting and assessing ML models:

1. **Training set**: the data points which the model weights are made to match by updating them according to a loss function.

2. **Validation set**: data points which have a similar distribution to the training set, but which do not feed into weight updates. The model predictions for these data are used as an unbiased estimator for the convergence of the model weights towards a local minimum in the loss surface, to ensure we do not overfit, and for tuning the hyperparameters.

3. **Test** (or **holdout**) set: this is comprised of data never used in training, which can provide an unbiased estimate of the *final* model performance.

If the training and validation set errors (the difference between model prediction and ground truth) diverge during fitting, this is a sign of **overfitting**—a flexible model that is memorising the precise training data and not smoothly interpolating

between points. It is important to ensure that the structures in the training set do not **leak** into the test set and therefore mislead on the generalisation ability of the model. In atomistic ML, this could easily happen, for instance by including snapshots from a single MD trajectory, which are nearby in time and therefore closely correlated.

Hyperparameters are parameters associated with a particular model architecture that are not adjusted, e.g. by gradient descent, when fitting to a dataset. Different settings can be tested and optimised to give the best quality of fit, with some examples being the Gaussian width, σ_{at}, for SOAP, the learning rate for NNs, and the maximum body order, ν, for ACE. Almost all ML potential frameworks will impose local atomic environments with a cutoff hyperparameter, r_{cut}, beyond which neighbours are no longer considered. Software packages now exist to perform efficient automated hyperparameter tuning for atomistic ML, such as XPOT.[140]

Regularisation is an important technique for ML with the aim of avoiding overfitting by penalising more 'complex' (i.e. less smooth) models.[141] In the GAP framework, this is formulated as an energy scale for the inherent noise in the reference data, e.g. due to convergence errors and those inherent in the DFT functional. In an NN context, two common approaches for regularisation are weight decay—adding a term to the loss function that penalises large weights—and dropout: freezing the weights of random neurons during each optimisation step to encourage multiple neurons to contribute to a prediction, which tends to reduce

overfitting. Implicit regularisation can also be achieved by **data augmentation**, adding more, finely-spaced data to an overfitted region of configuration space in order to smooth out the predictions (see chapter 3 for an example of this in practice).

Can we trust local energies?

Almost all ML potential frameworks make a foundational approximation that the total energy can be written as a sum of atomic contributions. Initially, this might appear to be a grave assumption because such extensive properties of matter arise from the collectively interaction of all atoms, with a single wavefunction, and can therefore only be rigorously defined as global quantities. However, even before ML potentials, there have been significant efforts towards developing quantum-chemical methods that can capitalise on quantum-mechanical nearsightedness[142] to perform a similar local decomposition.[143–148] This approach has two quite separate advantages: on the one hand, breaking the macroscopic observable down into simpler components allows drastic improvements in computational efficiency scaling and parallelisation; on the other hand, looking at complex interacting systems through simple building blocks is the cornerstone of the chemist's approach to a qualitative understanding of matter: be it functional groups in organic chemistry or the sharing of polyhedra in inorganic crystal structures. A decomposition of the total energy for the purposes of building ML potentials is useful in the same ways. It enables million-atom simulations for the work in chapter 3 of this thesis, and a qualitative understanding of defect behaviour in chapter 4. In Ref. [5], we showed, through the 'local prediction rigidity' (LPR), it is possible to

quantify how well-defined local energies are. By using variable stoichiometry and including a diverse range of structural types and defects, it is possible to increase the robustness of the decomposition for our carbon and silicon examples. This lends credence to the *post hoc* chemical interpretation of these quantities in the remainder of this thesis.

To get an overview of the meaning of the LPR, consider the loss function arising from minimising the difference between reference values Y_A for structure A (such as the total energy) and predictions $\hat{y}(A_i)$ given model weights $\boldsymbol{\theta}$

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2} \sum_{A \in \text{train}} \left[Y_A - \sum_{A_i \in A} \hat{y}(A_i|\boldsymbol{\theta}) \right]^2 \tag{2.13}$$

By then considering how sensitive the model is to a change, ϵ_k, in a local prediction associated with environment k we can assess the model's robustness. For this purpose, we can add a Lagrangian term to the loss to constrain the model prediction to respect this perturbation

$$\hat{\mathcal{L}}(\boldsymbol{\theta}) = \mathcal{L}(\boldsymbol{\theta}) + \lambda \left[\epsilon_k - (\boldsymbol{\phi}_k^o)^T (\boldsymbol{\theta} - \boldsymbol{\theta}^o) \right] \tag{2.14}$$

where $\boldsymbol{\phi}_k^o$ is a derivative from the Taylor expansion of the prediction about some unconstrained optimal model parameters $\boldsymbol{\theta}^o$,

$$\boldsymbol{\phi}_k^o \equiv \left. \frac{\partial \hat{y}(A_k|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right|_{\boldsymbol{\theta}^o} . \tag{2.15}$$

By minimisation of this new loss, following some algebraic manipulations described fully in Ref. [5], and using the 2$^{\text{nd}}$-order terms from the Taylor expansion, the new optimised loss can be written as

$$\hat{\mathcal{L}}^o(\epsilon_k) = \mathcal{L}^o + \frac{1}{2} \left. \frac{\partial^2 \hat{\mathcal{L}}^o}{\partial {\epsilon_k}^2} \right|_{\epsilon_k = 0} \epsilon_k^2 \tag{2.16}$$

where the term

$$\left.\frac{\partial^2 \hat{\mathcal{L}}^o}{\partial {\epsilon_k}^2}\right|_{\epsilon_k=0} \equiv \frac{1}{(\boldsymbol{\phi}_k^o)^T(\mathbf{H}^o)^{-1}\boldsymbol{\phi}_k^o} \tag{2.17}$$

describes how sensitive the model is to perturbations in a given local prediction, hence we termed it the 'local prediction rigidity' (LPR). $\mathbf{H}^o$ is the Hessian of the loss evaluated at θ^o

$$\begin{aligned} \mathbf{H}^o &= \sum_A \sum_{A_i,A_j \in A} \boldsymbol{\phi}_i^o (\boldsymbol{\phi}_j^o)^T + \sum_A (Y_A - \hat{Y}_A)\boldsymbol{\Psi}_A^o \\ [\boldsymbol{\Psi}_i^o]_{ab} &\equiv \left.\frac{\partial^2 \hat{y}(A_i|\boldsymbol{\theta})}{\partial\theta_a \partial\theta_b}\right|_{\boldsymbol{\theta}^o} . \end{aligned} \tag{2.18}$$

The LPR can be calculated exactly for linear and kernel-based models, and approximated for neural networks.

Some simple thought experiments provide intuition for the LPR: for single-component, highly-symmetric phases, the LPR will be higher. In the case of a crystal with a single Wyckoff position, such as diamond, then the local energies can be rigorously defined as Y_A/N because all atoms must have the same local energy by symmetry. In that case, LPR ≈ 1. Next, consider a dataset of structures for a two-component system with fixed stoichiometry, such as the zinc blende structure with two Wyckoff positions, one for each element. The relative contributions to the total energy from atoms of each element are completely arbitrary: the local predictions for either of the two elements can be freely made as the total energy can be recovered simply by adjusting the contribution from the other element. Addition of even a single structure with a different stoichiometry, however, will resolve the contributions from each element and lead to more robust predictions.

A pertinent practical application of the LPR is for amorphous silicon, for

which we study the local energies in detail in chapter 4; hence we would like to know we can rely on their chemical interpretability. In Fig. 2.4, adapted from Ref. [5], we show a 2D projection of the SOAP kernel distance matrix, such that the distance between points is related to their dissimilarity as measured by SOAP (there is more detail on this methodology around Fig. 2.11). Amorphous silicon is mostly made up of 4-fold connected atoms, with some 3-fold and 5-fold coordination defects at around 1 % concentration. The test set used to make Fig. 2.4 consists of structures containing a few examples of these defects, which separate clearly into clusters (the biggest contribution to the SOAP kernel effectively comes from the coordination number). These clusters are shown by the colouring according to coordination number in Fig. 2.4a. In fact, it is possible to prepare 'pristine' (i.e. defect-free) structural models for sufficiently small structure sizes—say, 64 atoms—by generating many such structures and selecting the ones that happen not to contain any defects. Training a SOAP-based kernel model on 500 of these defect-free structures leads to high values of the LPR for the 4-fold atoms in a test set. The 4-fold atoms tend to be quite homogenous in structure, hence they produce a tight range of local energy values which can be robustly partitioned. The closer the environment to the centre of the cluster, the higher the LPR, which can be rationalised through the greater self-similarity approaching the crystalline limit of maximal rigidity. However, the LPR takes low values for the defective atoms, as a lack of representation in the training data leads to some arbitrariness in the partitioning of energy between the defect and the wider 4-fold matrix (analogous to the case of entirely different elements). Inclusion of just a few defect-containing

structures in the training set raises the LPR by 2–3 orders of magnitude.

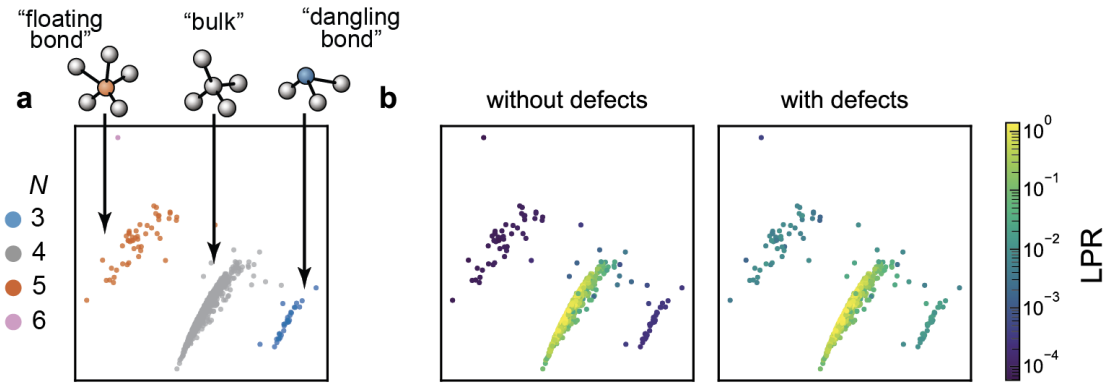


Figure 2.4: Local Prediction Rigidity for silicon with and without defects. (**a**) Kernel principal component analysis (KPCA) map with points colour-coded by the coordination number. Schematic diagrams illustrate the typical environment found for coordination numbers 3–5. (**b**) The same KPCA map, now coloured by the LPR value before (left) and after (right) the inclusion of defective structures in the training set Adapted from our Ref [5], originally conceptualised and produced by Raymond Chong.

The dataset used to train the models that predict local energies for the work in chapter 4 is very diverse, having been intended to produce a general-purpose potential;[45] there are examples of a-Si with defects, as well as snapshots from MD simulations of interstitials in diamond and a large number of other crystal structures. Hence, we expect the local energies for this model to be quite robust. For 'application-specific' models, like the ones trained on-the-fly as part of DFT codes, one might expect the local energies to be less predictable, since often only a single phase feeds into the training database.

Gaussian Approximation Potentials

The GAP approach uses kernel-fitting in the form of Gaussian Process Regression (GPR) to calculate a total energy as a sum of atomic energies,

$$E = \sum_i \sum_s^M \theta_s K(\mathbf{x}_i, \mathbf{x}_s) \ , \tag{2.19}$$

where i runs over the number of atoms in the system, K is a kernel basis function, $\mathbf{x}_i$ is the collection of relative position vectors of atoms in the neighbourhood of i, and s runs over M 'representative' atoms (also known as the 'sparse set'), which are selected from the training data and whose environments serve as a basis in which the potential is expanded. This latter choice is particular to the kernel-based methods; there is no necessity for the kernel basis functions to be located precisely on training data points, but this choice seems quite natural and has been shown to work well.[149] The kernel quantifies the similarity between environments, and with GAP models most often the SOAP kernel is chosen for this purpose, although it is perfectly possible to choose others, as in Refs. [46, 150, 151]. For the GAP model for silicon used in chapter 3, the following choice is made for the precise form of the SOAP kernel:

$$\begin{aligned} K(\mathbf{x}_i, \mathbf{x}_j) = \delta^2(\mathbf{p}_i \cdot \mathbf{p}_j)^\zeta = \delta^2 \left| \frac{\tilde{K}(\mathbf{x}_i, \mathbf{x}_j)}{\sqrt{\tilde{K}(\mathbf{x}_i, \mathbf{x}_i)\tilde{K}(\mathbf{x}_j, \mathbf{x}_j)}} \right|^\zeta \\ \tilde{K}(\mathbf{x}_i, \mathbf{x}_j) = \int_{\hat{R}\in O_3} \mathrm{d}\hat{R} \left| \int \mathrm{d}\mathbf{r} \rho^i(\mathbf{r}) \rho^{i'}(\hat{R}\mathbf{r}) \right|^2 , \end{aligned} \tag{2.20}$$

where the hyperparameter $\zeta = 4$ increases the sensitivity of the kernel and δ =3 eV sets the energy scale for the many-body interaction. The coefficients θ_s are minimised according to the following loss function:

$$\mathcal{L} = \sum_i^N \frac{[Y_i - \hat{Y}(\mathbf{x}_i)]^2}{\sigma_i^2} + R \, , \tag{2.21}$$

where Y_i are ground-truth labels, $\hat{Y}$ are model predictions, R is a regularisation term, and σ_i are weighting factors for individual structures (which can also be viewed as regularisation strength, by multiplying the whole expression by a constant σ for all environments). In GPR, Tikhonov regularisation is the most natural,

defined as

$$\sum_{j,j'} \theta_j k(\mathbf{x}_j, \mathbf{x}_{j'}) \theta_{j'} \ , \tag{2.22}$$

which forces the coefficients to remain small.

In matrix form, the loss function to be minimised can now be expressed as

$$\mathcal{L} = (\mathbf{Y} - \mathbf{K}_{IJ}\boldsymbol{\theta})^T \boldsymbol{\Sigma}^{-1} (\mathbf{Y} - \mathbf{K}_{IJ}\boldsymbol{\theta}) + \boldsymbol{\theta}^T \mathbf{K}_{IJ} \boldsymbol{\theta} \ , \tag{2.23}$$

where $[\mathbf{K}_{IJ}]_{ij} \equiv k(\mathbf{x}_i, \mathbf{x}_j)$, $\boldsymbol{\Sigma}_{ij} = \sigma_i^2 \delta_{ij}$ and remembering that $J << I$ for sparse GPR, as computational time for fitting scales as $\mathcal{O}(N^3)$. The analytical expression for the optimal coefficients in this problem is

$$\boldsymbol{\theta} = (\mathbf{K}_{JJ} + \mathbf{K}_{JI}\boldsymbol{\Sigma}^{-1}\mathbf{K}_{IJ})^{-1} \mathbf{K}_{JI} \boldsymbol{\Sigma}^{-1} \mathbf{y} \ . \tag{2.24}$$

In reality, an additional linear differential operator, L, of size $N_{\text{at}} \times D$, is required because we are trying to predict *local* energies from a total energy reference (we also have gradients, in the form of forces, that contribute much of the information). Hence, $\mathbf{Y} = \mathbf{L}\mathbf{Y}'$, where $\mathbf{Y}'$ is the vector with N_{at} components containing the unknown atomic energies and $\mathbf{Y}$ is a vector containing all D components of the input data: total energies, forces, and (possibly) stresses. The modified expression for the coefficients is then

$$\boldsymbol{\theta} = (\mathbf{K}_{JJ} + \mathbf{L}\mathbf{K}_{JI}\boldsymbol{\Sigma}^{-1}\mathbf{L}\mathbf{K}_{IJ})^{-1} \mathbf{L}\mathbf{K}_{JI} \boldsymbol{\Sigma}^{-1} \mathbf{y} \ . \tag{2.25}$$

To make this discussion of $\mathbf{L}$ more concrete, consider learning based on the covariance of two composite observations (i.e. total energies) Y_1 and Y_2:

$$\begin{aligned} \langle Y_1(\mathbf{X}_1), Y_2(\mathbf{X}_2) \rangle &= k(\mathbf{X}_1), \mathbf{X}_2) \\ \equiv k(\mathbf{x}_1, \mathbf{x}_2) + k(\mathbf{x}_1, \mathbf{x}_2') + k(\mathbf{x}_1', \mathbf{x}_2) + k(\mathbf{x}_1', \mathbf{x}_2') \end{aligned} \tag{2.26}$$

where $\mathbf{x}$ and $\mathbf{x}'$ are subsets of the total coordinates for atomic structures defined by $\mathbf{X}_1$ and $\mathbf{X}_2$, that is, local environments, such that

$$\mathbf{X}_1 = \mathbf{x}_1 \bigoplus \mathbf{x}_1' \ . \tag{2.27}$$

The simple sum of kernels leads to identity-operator-like structure in $\mathbf{L}$. It is also straightforward to extend this to any linear functional of the kernel functions (e.g. differentiation for forces):

$$\langle \nabla_\alpha Y(\mathbf{x}), \nabla_\beta Y(\mathbf{x}') \rangle = \frac{\partial}{\partial x_\alpha} \frac{\partial}{\partial x'_\beta} k(\mathbf{x}, \mathbf{x}') \ . \tag{2.28}$$

Moment Tensor Potentials

Training an MTP involves finding the regression coefficients θ_α that minimise a loss function based on the errors in predicted energies, forces, and stresses. Unlike with GAP, there is no known analytical solution to this problem, so the MTP authors employ a straightforward quasi-Newton optimisation method from randomly initialised coefficients—specifically, the BFGS algorithm.[152] The hope is, as with training neural networks, that the loss surface contains sufficient near-equivalent local minima such that it does not particularly matter which we end up in. Fortunately, the interchangeability of minima in the loss surface tends to be what is observed empirically; a common strategy is to train five MTPs using different random seeds, taking the best (lowest validation error) as the final model.[131] We found that all random models were of comparable quality when adopting the same approach. More formally, one can define the loss, $\mathcal{L}$, in terms

of the predicted MTP energy

$$E_{\mathrm{MTP}} = \sum_i \varepsilon_i \equiv \sum_i \sum_\alpha \theta_\alpha B_\alpha(\mathbf{x}_i) \ , \tag{2.29}$$

as well as forces (and, possibly, stresses), which can be expressed in terms of the first and second derivatives of E_{MTP} with respect to $\mathbf{r}_{ij}$, and the reference values (switching to Y for total energies to make the connection with the GAP loss clear):

$$\begin{aligned} \mathcal{L} = & \sum_{i=1}^{I} \Big[w_{\mathrm{e}} \left| \hat{Y}_i \left(\boldsymbol{\theta}; \boldsymbol{X}_i \right) - Y_i \left(\boldsymbol{X}_i \right) \right|^2 \\ & + w_{\mathrm{f}} \sum_\alpha \left| \nabla_\alpha \, \hat{Y}_i \left(\boldsymbol{\theta}; \boldsymbol{X}_i \right) - \nabla_\alpha Y_i \left(\boldsymbol{X}_i \right) \right|^2 \Big] \ , \end{aligned} \tag{2.30}$$

where the structure- or atom-wise weights, w, can be identified with $1/\sigma^2$ in the GAP formalism, but now lack the same Bayesian interpretation as expected errors and act as simple weighting factors. The B_α basis functions are those described above in Eq. 2.3.2. This simple, linear form makes evaluation of MTPs extremely efficient—orders of magnitude faster even than GAP. Unlike with ACE models, the B_α make no explicit priority use of the low-dimensional terms in the body-ordered expansion, which perhaps explains the empirical performance gap compared to ACE in transferability (extrapolation behaviour) and overall fitting accuracy.[153]

Atomic Cluster Expansion

Practically, the loss function used in the linear ACE framework for optimisation of the fitting coefficients, θ_{nl}, takes a similar form to that for MTPs:

$$\begin{aligned} \mathcal{L} = & \sum_{i=1}^{I} \Big[w_{\mathrm{e}} \left| \hat{Y}_i \left(\boldsymbol{\theta}; \boldsymbol{X}_i \right) - Y_i \left(\boldsymbol{X}_i \right) \right|^2 \\ & + w_{\mathrm{f}} \sum_\alpha \left| \nabla_\alpha \, \hat{Y}_i \left(\boldsymbol{\theta}; \boldsymbol{X}_i \right) - \nabla_\alpha Y_i \left(\boldsymbol{X}_i \right) \right|^2 \Big] \\ & + \beta_1 \sum_{K\mathbf{nl}} \left| \theta_{\mathbf{nl}}^{(K)} \right|^2 + \beta_2 \sum_{K\mathbf{nl}} \left| \theta_{\mathbf{nl}}^{(K)} \right|^2 \ , \end{aligned} \tag{2.31}$$

with additional regularisation terms and accompanying hyperparameters β_1 and β_2, which are typically set to small values such as 10^{-10} and 10^{-14} The n and l indices are left bold to emphasise that the θ can be collected based on body-order with implicit summation (see Ref. [17] for further details). The optimisation is also usually carried out hierarchically by fitting the lower-order basis functions first, i.e. one parametrises a two-body potential, then uses that as the initialisation for a combined two-body, three-body potential and so on to the maximum specified ν (a so-called 'ladder fit'). In the hyperparameter specification for the `PACEMAKER` implementation of ACE descriptors, the practitioner specifies maximum ν and the basis-set size via the `nradmax_by_orders` and `lmax_by_orders` keywords, where 'order' refers to ν (e.g. see table 5.5).[132]

As with GAP and MTP, the derivatives of the ACE basis functions take a simple analytical form and hence forces are efficient to calculate. The chain rule allows the expression of the symmetrised basis functions, B_{inl} in terms of the gradients of the raw tensor product basis, $\nabla_\alpha A_{inlm}$, which themselves are given, through the product rule, in terms of the elementary atom-pair functions (or 'one-bond functions', in the original ACE work):[17]

$$\nabla_\alpha A_{inlm} \nabla_\alpha \phi_{nlm}(\mathbf{r}_{\alpha i}) + (-1)^l \sum_j \nabla_\alpha \phi_{nlm}(\mathbf{r}_{ij}) \delta_{i\alpha} \ . \tag{2.32}$$

The body-ordered expansion of the ACE framework can be identified most closely, perhaps, with the MTP formalism described in parallel in the previous sections. However, a major contribution to the observed performance advantage of the ACE framework, along with the systematic ladder-fitting approach, is the physically-

motivated form for the total energy.[132] With non-linear ACEs, as used for the MoS_3 system in chapter 5 of this thesis, it is possible to construct models that converge to DFT energies much faster than is possible with the purely linear case. The tight binding (TB) approximation derived from DFT can be used to motivate the specific non-linear form used by the ACE framework as a surrogate for atomistic Hamiltonians.[154] TB starts with free atoms, which are spherical, and can capture the repulsive overlap of charge densities. The next term relaxes the isotropy and allows orbital hybridisation, hence bond formation. In applications to extended solids, bond formation can be understood as being driven by reductions in the band energy, which, for the simplest models (e.g. semi-infinite linear chains), depends on the square root of the density of states.[155] This loose physical motivation explains the original formulation of the non-linear local energy:

$$E_i = -\sqrt{\rho_i^{(1)}(1+\rho_i^{(2)})} + \rho_i^{(3)} \ , \tag{2.33}$$

where each ρ term is a different ACE-expansion of the local density with its own fitting parameters, which can be viewed as an extension of the Finnis-Sinclair potential to including many-body interactions in ρ.[156] For the ACE model used in chapter 5, we extend this form by adding more density terms because we found empirically that improved the quality of the fit. Full details of the hyperparameters used are given in table 5.4.

Baselines: dispersion correction and collision prevention

The success of very simple pairwise models, such as Lennard-Jones, in capturing the broad-brush behaviour of condensed matter (e.g. crystallisation, the liquid

phase, the qualitative shape of phonon dispersion etc.[157–159]) can be rationalised because of the decreasing relative energy scales with increasingly complex ν-bodiness of the interactions. Therefore, when constructing ML potentials, much of the interaction strength can be taken care of with simple expressions, with ML learning the difference ('delta-learning'[160]). The GAP framework makes extensive use of so-called baseline potentials (see Ref. [100] for an overview); the silicon model, on which chapter 3 builds, makes use of a repulsive core potential that is fitted to the DFT interactions of a pair of Si atoms approaching one another.[45] A general-purpose GAP model for phosphorus utilises both a repulsive overlay at short distances, and an attractive Lennard-Jones term, with $1/r^6$ distance scaling, to capture the van der Waals interactions as far out as 20 Å.[44] In these cases, the two-body models are learned from (possibly different subsets of) the data. Then, to fit the many-body model, e.g. using a SOAP descriptor, the two-body energy and forces are subtracted from the reference data, so that the many-body model is learning the energy difference, which is much smaller in magnitude. It is also common to use well-known empirical potentials as a baseline, e.g. the -D3 dispersion correction (described in section 2.4.3) is subtracted before fitting for the AIMNet2 model.[30] Unlike with DFT, however, the -D3 correction here by no means represents a trivial addition to the cost (approximately doubling the time per force evaluation). For the MTP and ACE frameworks, small distances are often prevented with a Ziegler–Biersack–Littmark (ZBL) screened nuclear repulsion, which is an empirical potential that was originally developed for describing high-energy collisions in ion bombardment applications, but which has found use in ML potentials for

preventing unphysical attractions at short range. Usually, there are no training data in the very short-range region, either because the structures are very high in energy or perhaps even impossible to converge DFT for if the pseudopotential regions overlap, so ML models are unlikely to make reliable predictions alone. Note that high accuracy in the sense of a low error metric is not required in this region of close-approaching atoms, instead only the right qualitative repulsion to prevent atoms crashing together. However, even with a ZBL repulsion term, problems can be encountered during geometry optimisations and MD simulations, sometimes due the steepness of its exponential wall creating enormous forces, and sometimes because the spuriously attractive ML term is strong enough to outweigh the ZBL term for some distance range at close approach. To address this issue for models that are trained purely on RSS data, we build on the work of Ref. [161] by adding additional harmonic repulsions to prevent, softly, the models from visiting regions with excessive forces until sufficient data have been supplied for predictions to be valid at close approach (Fig. 2.2).

2.4 Density Functional Theory

DFT is currently the primary tool for simulations of condensed matter.[162] It is formulated as an exact theory for a system of interacting quantum particles, based on the Hohenburg–Kohn theorems:

1. All properties are completely and uniquely determined by the ground state electron density $n_0(\mathbf{r})$.

2. The energy functional $E[n]$ is sufficient to determine the exact ground state

density and energy.

These can be generalised to spin-polarised systems by considering the spin density as well as particle density, which is important for magnetic properties.[ii] A 'functional' is defined as a function that itself takes a function as an argument—in this case that argument is the electron density, which itself is a function of spatial coordinates. The practical problem with these general theorems is that no known method exists for obtaining the form of $E[n]$, and in fact it can be shown that a closed form for $E[n]$ will be arbitrarily complex to write down.[164] Fortunately, there are many approximations to $E[n]$ that rely on calculating exactly its 'large' terms and replacing the smaller, more difficult exchange and correlation terms with a well-chosen ansatz. This is usually done within the Kohn–Sham approach, which assumes that the density of the true interacting system is the same as that of some non-interacting, more-easily solvable one, which leads to an effective Hamiltonian of the form

$$H^{\sigma}_{\mathrm{KS}}(\mathbf{r}) = -\frac{1}{2}\nabla^2 \;+\; V_{\mathrm{Hartree}}(\mathbf{r}) \;+\; V^{\sigma}_{\mathrm{xc}}(\mathbf{r}) \;. \tag{2.34}$$

The difficult terms are incorporated into the 'exchange-correlation potential' $V^{\sigma}_{\mathrm{xc}}(\mathbf{r})$, containing functionals of the density $n(\mathbf{r})$, which leads to eigenvalue equations that must be solved iteratively until the density, expressed in some basis, is self-consistent.[iii]

The local density approximation (LDA) is the simplest realistic class of

[ii] Excited states fit less naturally into this formalism, which ML potentials also struggle with. Hence magnetic properties are difficult to study, although efforts are underway to include spin as a variable in the representation.[163]

[iii] This convergence is not always easy to achieve, particularly in systems with unpaired spins or chemically unusual environments, as is discussed further in chapter 5 for the Mo–S system.

density functional, which approximates the exchange-correlation energy density, $\epsilon_{\mathrm{xc}}([n], \mathbf{r})$, at each point as that of a homogeneous electron gas with the same density. This was inspired by the observation that solids can often be considered close to the electron gas limit, since the ranges of exchange and correlation effects are short,[iv] which may explain the greater popularity of DFT amongst condensed matter scientists compared to those interested in molecular applications. Explicit analytic forms for the LDA can be fitted to experiments or highly accurate quantum Monte Carlo methods.[165] It is, perhaps, surprising that such a simple approximation is good enough to be useful,[46,166] which can be explained in part because its Hamiltonian obeys the correct sum rules for the size of the exchange-correlation hole. The most significant failure of the LDA is that it does not fully eliminate a spurious self-interaction of electrons, which leads to over-delocalisation and consequently a systematic underestimation of bond lengths.[167] The generalised gradient approximation (GGA) is a substantial improvement on the LDA, which provided sufficiently high accuracy to incentivise very wide adoption by the solid-state modelling community.[168] The GGA incorporates a Taylor expansion in the gradient of the density, with damping functions that modify the behaviour at large gradients to preserve the sum rules. For the studies of silicon in this thesis, we use the PW91 functional (an early example of a GGA) to ensure compatibility with the work on which we build.[45,169] For the study of the Mo−−S system, we employ the PBE[170] functional, which is something of an 'industry standard' for solid state calculations because of its improved accuracy for binding energies, geometries,

[iv] This short range is what enables us to decompose the total energy into atomic contributions in the construction of ML potentials without losing too much accuracy.

and reaction energies compared to PW91, and its affordability compared with higher-level methods.[171]

2.4.1 DFT in periodic solids

The translational symmetry of crystals imposes a specific form for any admissible wavefunction for them. Bloch's theorem gives this form as:

$$\psi_{n,\mathbf{k}}(\mathbf{r}) = e^{i\mathbf{k}\cdot\mathbf{r}}u(\mathbf{r}) \tag{2.35}$$

$$\mathbf{k} = \frac{n_a}{N_a}\mathbf{a}' + \frac{n_b}{N_b}\mathbf{b}' + \frac{n_c}{N_c}\mathbf{c}', \quad 0 < n_i < N_i, \tag{2.36}$$

where $u(\mathbf{r})$ is any periodic function, so that $u(\mathbf{r}) = u(\mathbf{r} + \mathbf{T})$ for any lattice translation $\mathbf{T}$. The wave vector $\mathbf{k}$ gives the phase change of the wavefunction between unit cells and is quantised to respect periodic boundary conditions, with the integer n_i taking values up to the number of unit cells along i. The band index, n, enumerates the independent eigenstates for a given $\mathbf{k}$. This introduces the reciprocal lattice, which is defined by the vectors

$$\mathbf{a}' = 2\pi\frac{\mathbf{b}\times\mathbf{c}}{|\mathbf{a}\cdot(\mathbf{b}\times\mathbf{c})|} \tag{2.37}$$

and its cyclic permutations for $\mathbf{b}'$, and $\mathbf{c}'$, where $\mathbf{a}, \mathbf{b}, \mathbf{c}$ are the real space crystal lattice vectors. All properties are determined by values of $\mathbf{k}$ in the first Brillouin zone, the smallest volume in reciprocal space enclosed by planes perpendicular to the reciprocal lattice vectors (the Wigner-Seitz cell), because addition of any reciprocal lattice vector $\mathbf{G}$ to $\mathbf{k}$ leaves equation 2.35 unchanged.

Plane waves are a natural choice for the basis functions in which to expand

the wavefunction, from which the electron density can be constructed,

$$\frac{1}{\sqrt{\Omega_{\mathrm{BZ}}}} \sum_{\mathbf{G}}^{E_{\mathrm{cut}}} C_{\mathbf{G}n\mathbf{k}} e^{i(\mathbf{G}+\mathbf{k})\cdot\mathbf{r}} \ , \tag{2.38}$$

normalised by the volume of the Brillouin zone Ω_{BZ}. The quality of the basis set is determined by the finite set of plane waves with energy less than E_{cut}, which have wave vectors satisfying

$$\frac{1}{2}|\mathbf{G}+\mathbf{k}|^2 < E_{\mathrm{cut}}. \tag{2.39}$$

For most properties, e.g. the total energy, it is necessary to sum over all **k**-points, which becomes an integral in the thermodynamic limit, because there are as many values of **k** as there are unit cells in a bulk crystal. A major parameter choice in setting up calculations is the sampling grid used for this quadrature, since every additional **k** point increases the number of basis functions and therefore the computational cost. Metals require very dense sampling compared to insulators because of their partially filled bands, which can go from being occupied to unoccupied over a small range of **k**. Larger supercells permit much sparser grids because the values of **k** describe the phase relationships between cells, which is less important when points in the cell are more distant in real space from their associated periodic image. Point group symmetry can also reduce the calculations, as it renders certain **k**-points equivalent. Practically, the energy and interatomic forces need to be converged both with respect to the **k** grid and E_{cut}. ML potentials are extremely sensitive to noise in the energies and forces in their training database, which has the practical consequence of requiring much tighter convergence settings for DFT calculations than is usually required for *ab initio* investigations (typically

to something like 10^{-7} eV in energy).[100]

2.4.2 Pseudopotentials

For an atom with a substantial number of core electrons (i.e. anything beyond the 2$^{\text{nd}}$ row), the true wavefunction varies rapidly in the neighbourhood of the nucleus to maintain orthogonality between the one-electron states. To capture this large curvature with plane wave basis functions, whilst simultaneously reproducing the much smoother behaviour in the interatomic region, it would be necessary to use a computationally intractable number of plane waves (i.e. a very large E_{cut}). However, since the core electrons are largely insensitive to their chemical environment and the wavefunction outside the core region can be made invariant to a change of the true potential, this problem can be solved by switching to a weaker *effective* potential that includes the core electrons implicitly (see Fig. 2.5 for a diagram). This is the basis of the pseudopotential approach, where only wavefunctions for the valence electrons outside a defined cut-out sphere (centred on the core region) are included explicitly. Highly accurate QMC calculations on isolated atoms are used initially to construct the pseudopotentials and these are made available as part of electronic structure codes. Different choices for the pseudopotentials lead to small differences in the calculated energies, which renders data calculated with different DFT codes incompatible for any single fitting database for an ML potential. Efforts are underway both to standardise the treatment of pseudopotentials across codes,[172–174] and design fitting architectures that can handle heterogeneous DFT data to avoid the duplicated computational

effort associated with relabelling.[175]

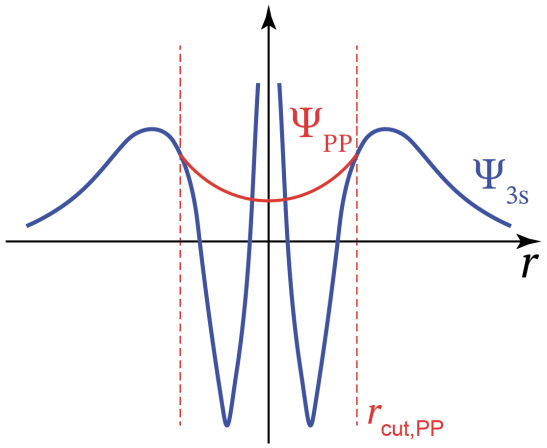


Figure 2.5: 3s wavefunction (blue) and associated smooth pseudopotential for the core region (red). Adapted from Ref. [162]

2.4.3 Dispersion corrections

The locality approximation inherent in the GGA and LDA, along with the neglect of virtual orbitals, leads to a systematic underestimator of the dispersion interaction, E_{disp}, which acts at long range ($\propto 1/r^6$). Reaction energies and activation barriers are typically predicted with errors of 25–35 kJ mol^{-1} with bare DFT across a range of functionals, which can be reduced by around 6 kJ mol^{-1} with inclusion of Grimme's 'DFT-D3' dispersion correction,[176] which is of the form

$$E_{\text{disp}} = -\frac{1}{2}\sum_{i=1}^{N_{at}}\sum_{j\neq i}^{N_{at}}\left(f_{d,6}\left(r_{ij}\right)\frac{C_{6ij}}{r_{ij}^{6}} + f_{d,8}\left(r_{ij}\right)\frac{C_{8ij}}{r_{ij}^{8}}\right). \tag{2.40}$$

The familiar C_6/r^6 term is included, which has a good theoretical underpinning as an asymptotic limit at large r; C_6 is derived from the polarisability. The extra C_8/r^8 term is also included in the expansion to provide more accuracy in the medium-range regime where there is interference from the short-range description of electron correlation by the density functional (e.g. PBE). For this reason, the C_6

and C_8 terms have to be parametrised for each density functional separately, so this method is considered semi-empirical, to match a trusted reference method—usually coupled cluster with singles, doubles, and perturbative triples, CCSD(T). The C_n terms are also environment-dependent (via the coordination number). Damping functions, $f_{d,n}$, are required to avoid near singularities for small r and double-counting effects for correlation at intermediate distances. We use the variant with Becke-Johnson damping,[177,178] which is of the form

$$f_{d,n}\left(r_{ij}\right) = \frac{s_n r_{ij}^n}{r_{ij}^n + \left(a_1 R_{0ij} + a_2\right)^n} \ , \tag{2.41}$$

with $R_{0ij} = \sqrt{\frac{C_{8ij}}{C_{6ij}}}$, $s_6 = 1$ and a_1, a_2, and s_8 being adjustable parameters.

There are now more advanced methods for correcting the missing dispersion interactions in DFT, such as the random phase approximation (RPA)[179,180] or symmetry-adapted perturbation theory (SAPT),[181] but these are prohibitively expensive for large systems and lack analytic gradients (i.e. forces) in their current early implementations. The many-body dispersion method (MDB) has recently become popular as an alternative to '-D3' with better accuracy and a less extreme increase in cost.[182] The main idea is to calculate C_6-like parameters from a partition of the electron density as calculated with DFT, which feed into a model system of interacting dipoles (Drude oscillators) whose interactions can be solved self-consistently. However, we found its current implementation to be ill-suited to the Mo−S system, because of instabilities during the self-consistent screening step of this method for the unusual geometries needed to fit our ML potentials.

For the Mo-S system, we elect to use PBE with the -D3 dispersion correction

as the best balance between accuracy, cost, and for maximising comparability with previous work on this material.[183–186] For MoS_2, PBE-D3 provides an excellent prediction of the interlayer spacing compared to experiment and to high-level RPA calculations (0.8 %).[187,188]

2.5 Validation of ML potentials

ML potentials are becoming increasingly widely-used. On the one hand, this makes it important to ensure careful validation of the fitting methodologies in their own right, which is typically done by the developers or experienced practitioners performing benchmarks. On the other hand, specific models developed for a particular task by researchers interested in applications should also be validated with the same level of rigour. This requirement is especially important for data-driven models, which do not have a great deal of physical information incorporated (that is by construction: to allow the expressiveness needed for accuracy in the fitting).

An influential article by Zuo *et al.* established an early benchmark in which the authors performed a detailed study of numerical energy and force prediction errors across different types of ML potentials.[189] Their focus was on comparing different fitting frameworks fairly in order to identify a Pareto front: i.e. the most accurate method (for given hyperparameter settings) at a each level of computational cost to establish a line giving the 'optimal models' in the 2D space of cost-accuracy trade-off. The validation test sets used by Zuo *et al.* included liquid and crystalline structures for a number of elements. An advantage of this protocol

is its generality: it can easily be applied to any elemental system. In contrast, a recent publication of a carbon ML potential includes various domain-specific tests, e.g. the formation energies of well-known topological defects in graphene,[190] and more complex tests were later carried out in other comparative studies.[47,191,192] Increasingly in the literature, the importance of validating ML potentials is being highlighted—not only based on numerical errors, but also on the predicted physical behaviour.[193–196]

2.5.1 Numerical errors: MAE, RMSE

The first test for any ML regression model, as described in section 2.3.3, is to generate a predicted label, $\hat{Y}_i$, for each entry in the database and compare it to the ground-truth label, Y_i. Numerical error metrics are designed to aggregate the disparity in these values across some test dataset to give a single measure by which we can assess model performance.

Two of the most commonly used metrics are the mean absolute error (MAE) and root mean square error (RMSE), defined as

$$\begin{aligned} \text{MAE} &= \frac{1}{N}\sum_i \left|Y_i - \hat{Y}_i\right| \text{ , and} \\ \text{RMSE} &= \sqrt{\frac{1}{N}\sum_i (Y_i - \hat{Y}_i)^2} \text{ .} \end{aligned} \tag{2.42}$$

While these metrics both measure some sort of average error, always RMSE $\geq$ MAE, and the RMSE is skewed upward by outlying high errors. Additionally, there is the mean error (ME) measure for the model bias, $\frac{1}{N}\sum_i (Y_i - \hat{Y}_i)$, which tends to zero if positive and negative deviations cancel (i.e. there is no average shift between predictions and ground truth). In ML potential applications, a significant

ME often derives from either inconsistent DFT data (i.e. with a different 'zero of energy' due to differences in basis set, k-point sampling etc.), or because a model is not learning a constant contribution to the total energy in the reference data, such as that of the isolated atoms.

Although a single error metric is convenient for quotation in tables that compare ML potential fitting frameworks, a lot of detail is lost in the aggregation. In Fig. 2.6, I show an example of disaggregated error reporting in the form of a parity plot and force error histogram. Additionally, the data are separated into two classes depending on the type of structure, which show strikingly different distributions. The 'phonon' data consist of near-perfect crystalline structures with small displacements, which are an example of domain-specific data that are included for higher accuracy in phonon calculations. The 'MD' set consists of snapshots extracted for testing during iterative MD simulations. The data used for this plot are a subset of what was used to fit the ACE potential for MoS_3, with full details in chapter 5. We see that, as might be anticipated, the MD simulations cover a much wider range of absolute force magnitudes. The same percentage error on these forces (compared to their magnitude) would give a much higher MAE, hence it is important to consider their distribution. Attempts to define a single percentage error metric, however, run into trouble with the high concentration of forces around zero for both data types. We also see a spike in the 'Phonon' error distribution around $0.2\,\mathrm{eV\,Å^{-1}}$. This comes from the corresponding structures using a greater-than-usual displacement parameter which samples larger forces, hence the model makes greater prediction errors for these data compared to the

other data focused on phonon predictions.

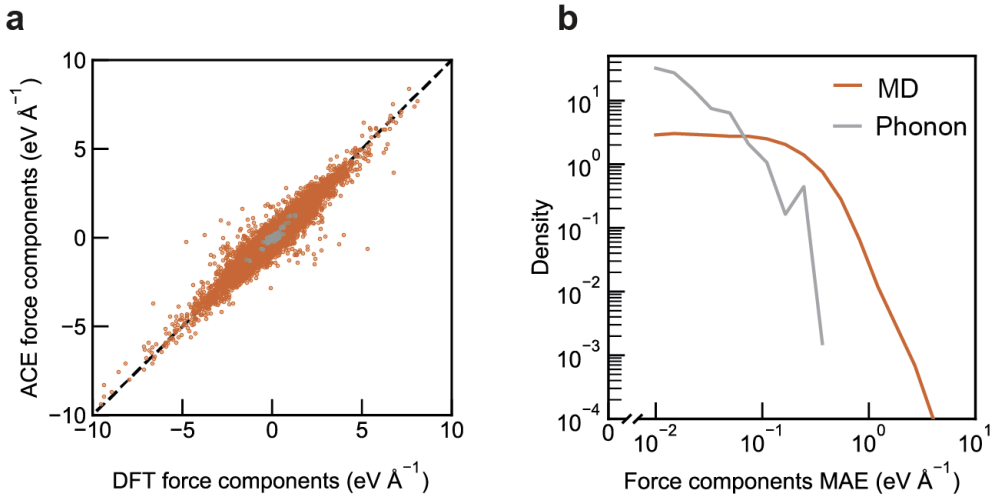


Figure 2.6: Disaggregated error reporting for the example of MoS_3, separately by data type (discussed in detail in chapter 5). 'Phonon' (orange) consists of crystalline structures with small displacements, 'MD' consists of snapshots sampled during iterative MD (**a**) Parity plot with reference method predictions (DFT) on the horizontal axis and ACE model predictions on the vertical axis for the force components. (**b**) Histograms of force component MAE. Notice the log-log scale.

When predicting per-atom properties, such as the Cartesian components of forces, there is a one-to-one mapping between DFT ground truth and ML-predicted values. Therefore, no further considerations regarding the number of atoms in the cell are required when reporting errors. However, for energies, DFT does not normally provide per atom values because only the total energy is a quantum-mechanical observable of the many-body wavefunction.[v] When predicting extensive properties, such as the total energy, N individual ML model predictions must be summed and compared to a single true label. In these cases, the system size does have an effect on the behaviour of the numerical errors. For the same reason that the variance of the estimated mean of any random variable scales as

[v] or, for DFT, the total electron density

$1/\sqrt{N}$, where N is the number of observations, so too does the 'mean' atomic energy from N atoms, which is, of course, related to the total energy by a simple constant (N). An extreme example of this phenomenon is given by the 1,000-atom structures in chapter 3 for teacher–student learning, where the level of agreement is within 1 meV per atom between student and the teacher reference due to the system size, despite the agreement for smaller structures being considerably worse. Figure 2.7 illustrates this concept in a systematic way based on the amorphous structures in the carbon dataset of Ref. [46].

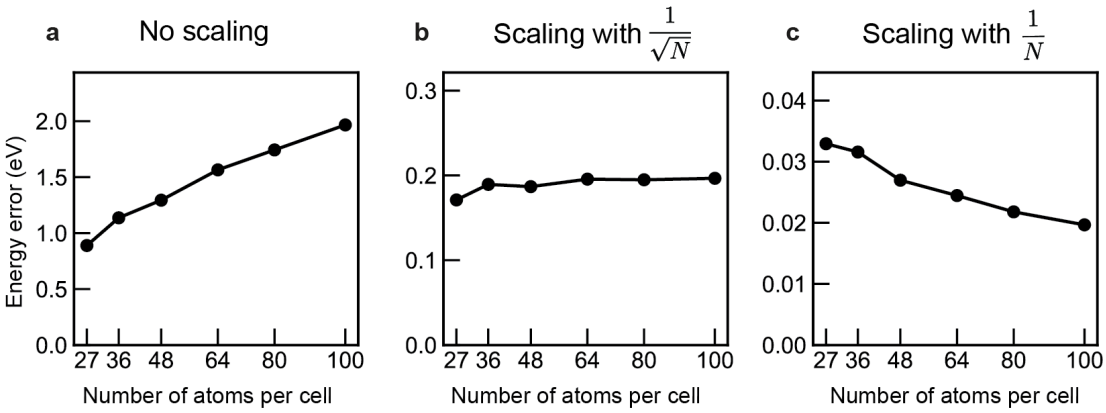


Figure 2.7: Scaling of energy errors with system size for a subset of the C-GAP-17 carbon dataset of Ref. [46] We plot errors from left to right for the total energy predicted by an existing ML potential,[46] normalised using (**a**) no scaling factor, (**b**) scaling of $1/\sqrt{N}$, and the commonly used scaling of $1/N$. Errors were evaluated for predictions made by the C-GAP-17 ML potential (as a sum of local energies) compared to ground-truth DFT total energies. A subset of the full C-GAP-17 reference database was selected for this exercise, such that the configuration type and energy distribution of structures is the same for all N to illustrate the trend more clearly. This figure is adapted from our Ref. [2].

The wider field reports energy errors per atom (particularly for extended solids and bulk liquids), so it is often desirable to include these figures as well for ease of comparison. However, it is important to keep in mind that this will cause predictions for larger structures to appear, unfairly, of higher accuracy. I

note that the MTP formalism also appreciates this square root dependence during *fitting* (but not error evaluation), by introducing a scaling in the loss function with the same $\sqrt{N}$ term when the resulting MTP is intended to predict accurate vibrational properties.[90]

Molecular Dynamics

MD simulations, carried out over a reasonably wide range of temperatures and pressures, provide a good test of the validity of an ML potential.[197–199] MD simulations *per se* also represent one of the most useful types of simulation for gaining insights on the material or molecule under the microscope: be it the design of new materials or drug discovery.[200,201] Therefore, any good potential ought to be able to propagate dynamics for the properties of interest with a good level of accuracy. Errors accumulate as the system explores configurations, which may lead to regions of phase space that are increasingly far from the training data. One example is the presence of spurious minima—predictions of low energy by the ML model for physically incorrect structures that should be high in energy as predicted by DFT. From these minima, the simulation cannot easily revert to other regions of the PES (even if those *are* correctly described). Furthermore, once in a poorly described region, it is unlikely that the potential has seen similar data, so the potential makes increasingly worse predictions that can lead to instabilities in the simulation, such as predictions of extreme forces that overwhelm the temperature and pressure controls.[202] Qualitatively, a potential with higher errors overall on a test set can therefore be much more useful if it avoids the pathologies described in panels b and c of Fig. 2.8.

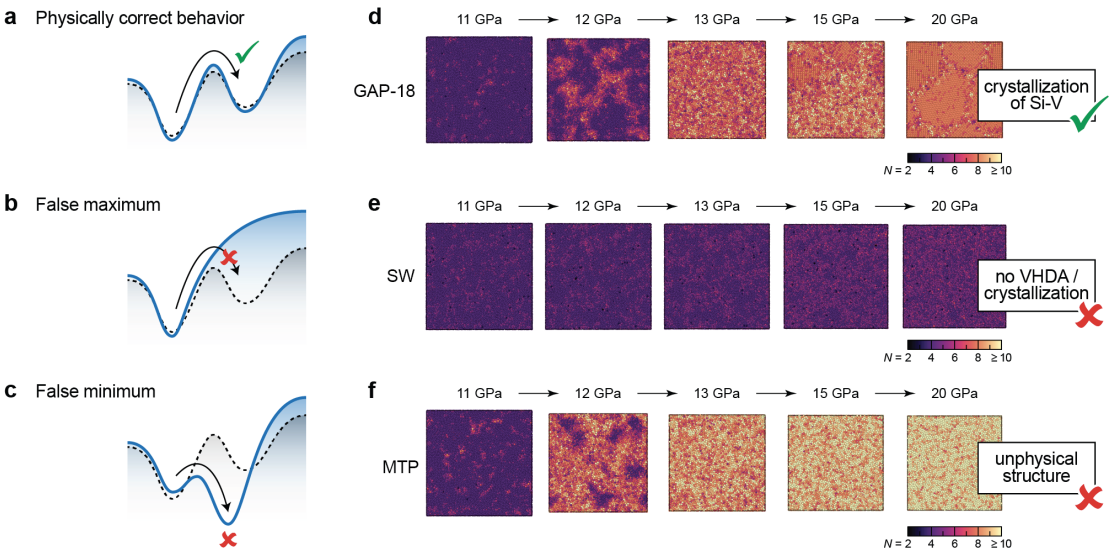


Figure 2.8: Qualitative physically-guided validation with an MD compression test. **(a)–(c)** Diagrams comparing a reference PES (dashed lines) and the potential that approximates it (solid blue lines). If a false maximum exists, the simulation is not able to reach the relevant local minimum. If a false minimum exists, the potential will readily visit a spurious minimum that corresponds to a physically incorrect structure. **(d)–(f)** Examples from compression simulations for silicon. The reference simulation here is from a recent study of the high-pressure structural changes.[203] The 'false maximum' is exemplified by a simulation of the same protocol with an empirical potential that does not access the very high-density amorphous (VHDA) phase. The 'false minimum' is exhibited by an MTP potential (one of the failed 'direct' models of chapter 3) that shows spurious formation of an unphysical highly coordinated structure. The severe detrimental effect of false minima for the quality ML PES approximations has also been connected with Boltzmann sampling.[204] This figure was reproduced from our Ref. [2]. Panels d and e were drawn with data from Ref. [203], and panel f with data from Ref. [1].

The high-dimensional space of coordinates means it is near-impossible to systematically remove all spurious minima by covering them with training data. Empirically, potentials trained with RSS data seem to achieve this quite efficiently, perhaps by 'filling in' the high-entropy basins in the PES from a wide array of initial conditions. In Fig. 2.8, we illustrate a compression test for Si that uses MD simulations to discriminate between potential models. The Stillinger-Weber (SW) empirical potential can only favour tetrahedral structures with its simple functional form (it was designed to model low-density a-Si),[205,206] hence it fails to form the

very-high density amorphous (VHDA) phase with higher coordination numbers. One of the failed direct MTPs from chapter 3 exhibits a 'false minimum'. As well as with compression, melt-quench simulations can be useful for checking whether a potential makes qualitatively correct predictions: both at high temperatures, and for the quite-highly-strained environments encountered at grain boundaries during crystallisation. Fig. 2.9 illustrates this for the emergence of poorly crystalline MoS_2 from the rapid cooling of the melt.

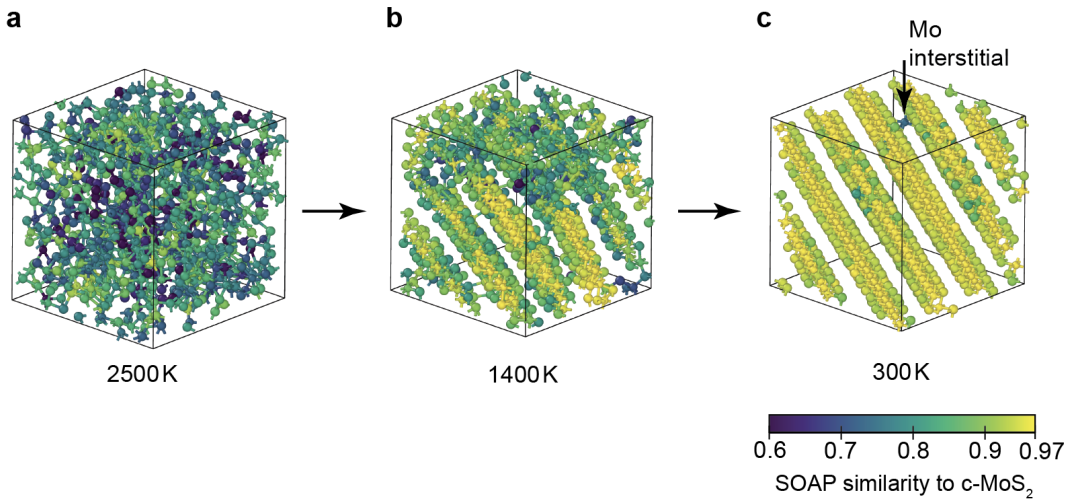


Figure 2.9: Qualitative physically-guided MD test in practice: quenching molten MoS_2. **(a)–(c)** Snapshots from a melt-quench trajectory for MoS_2 using the ACE model described in chapter 5, from 2500–300 K at 10^{13} K s^{-1}. Atoms are coloured according to their SOAP similarity (via the kernel discussed in section 2.3.2) to the crystalline MoS_2 phases. The correct qualitative chemical behaviour is observed: the formation of trigonal prismatic layers. The speed of the quench leads to turbostratic disorder, that is, a lack of registry between planes, and an Mo interstitial defect is also left over between layers. Both observations are chemically sensible for a rapid quench; the melting point of MoS_2 is somewhere around 1500–1900 K.[207,208]

2.5.2 Random Structure Search

We have already introduced RSS for the purposes of data generation in section 2.3.1. Yet another application of the method is the unbiased validation of interatomic potentials, as first used in Ref. [45]. To illustrate, in Fig. 2.10 I show the energies

and volumes of 10,000 initial random silicon structures that have been relaxed into local minima with:

(**a**) The general-purpose GAP-18 model taken from Ref. [45], which had previously been shown to reproduce AIRSS results quite closely.[114]

(**b**) A candidate (GAP → MTP, see chapter 3) that has been trained on GAP-18 data.[1]

(**c**) The empirical SW potential,[205,206] which has been widely used for modelling silicon and other semiconductors.

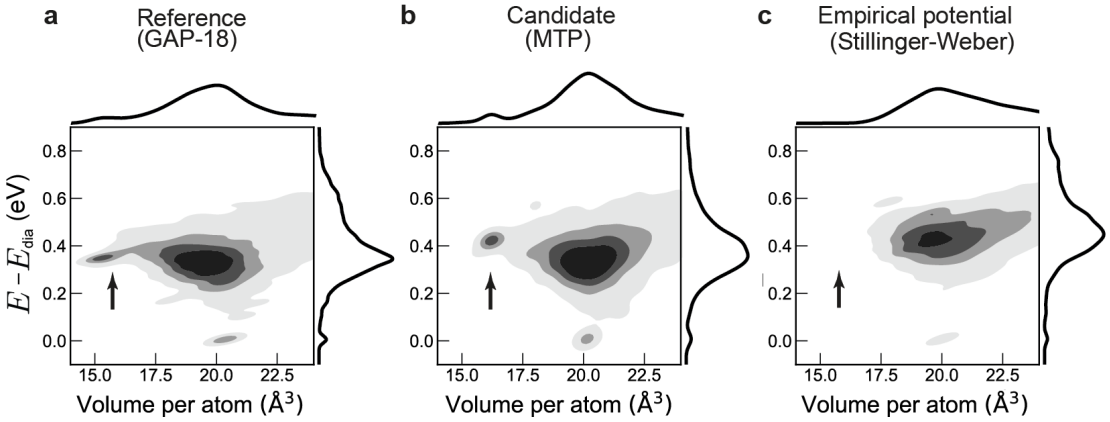


Figure 2.10: Automated, system-agnostic validation with RSS test. 10,000 random structures are relaxed with the following potential models: (**a**) GAP-18 model taken from Ref. [45]. (**b**) Student MTP potential from chapter 3. (**c**) Stillinger-Weber potential.[205,206] This figure is reproduced from our Ref. [2].

A basin in the PES at low volumes can be identified for both ML potentials (indicated by arrows in Fig. 2.10), which contains structures that are similar to simple hexagonal-like phases observed at high pressures. In contrast, the empirical SW potential fails to relax the same random structures into this chemically sensible local minimum because, again, it strongly favours tetrahedral (diamond-like), lower-density structures. This observation is consistent with the fact that SW predicts

no structural collapse under pressure in Fig. 2.8. All three potentials do find diamond-like minima, which we define by the relaxed structure having a SOAP similarity of ≥ 0.99 to the ideal crystalline form; however, the empirical SW potential finds considerably fewer: 112 (GAP-18), 135 (MTP), and 55 (SW).

Complementing the E-V distributions, the structures produced by an RSS run also constitute an unbiased and challenging set that can be useful for out-of-sample testing. Here, we mean both the fully relaxed structures (with zero forces according to the relaxing potential) and the points along the minimisation trajectory with non-zero forces. To this end, the RSS structures can be labelled with a reference method (e.g. DFT), and any candidate potential can be compared against this set. Using the ML potential to drive the structural relaxations is much cheaper computationally than relaxing the structures directly with DFT, which typically would be required when making an E-V comparison such as in Fig. 2.10. The numerical errors for an RSS test set are likely to be relatively high for what are rather unusual structures, which underlines the need to view absolute error metrics in the context of the test set composition. Examples of this approach in practice are given in chapters 3 and 5 for silicon and MoS_3 respectively.

2.6 Model building with molecular dynamics

To simulate amorphous materials, one of the most common methods used imitates the experimental preparation: melting followed by rapid cooling ('quenching') to freeze in the disorder. In MD simulations, the positions are updated according to Newton's equation of motion, $\mathbf{F} = m\mathbf{a}$, where $\mathbf{F}$ are the forces on atoms—e.g. as

calculated by an ML potential. Even the fastest experimental quenches, such as those involving microscopic laser heating,[209] are difficult to achieve in simulation: $10^{11}\,\mathrm{K\,s^{-1}}$ over a reasonable temperature range requires of the order of 10^6 force evaluations, which was about the limit of recent studies with ML potentials,[14] and is well beyond the reach of DFT. Other approaches to amorphous structure generation, such as sequential deposition[15,190] or ion bombardment,[210,211] have also been tried with success (all of which rely on MD to propagate the dynamics).

In this thesis, all MD simulations are carried out using the 'Large-scale Atomic/Molecular Massively Parallel Simulator' (LAMMPS) software package, which is written in fast C++ and optimised for parallelism across many CPU cores. It is the standard package for simulations of extended materials and has support for most currently used ML potential frameworks.[212,213] Below, I briefly summarise the algorithms behind the major parameter choices for these simulations.

2.6.1 Velocity Verlet integration

The velocity Verlet algorithm[214] for calculating the update in positions according to the current state is given by

$$\begin{aligned} \mathbf{x}(t+\Delta t) &= \mathbf{x}(t) + \mathbf{v}(t)\Delta t + \frac{1}{2}\mathbf{a}(t)\Delta t^2 \ , \\ \mathbf{v}(t+\Delta t) &= \mathbf{v}(t) + \frac{\mathbf{a}(t)+\mathbf{a}(t+\Delta t)}{2}\Delta t \ . \end{aligned} \tag{2.43}$$

The key parameter to set is the timestep, Δt.[215] It is desirable to choose as long a timestep as possible without sacrificing accuracy, e.g. by violating energy conservation, since this allows a faster sampling of phase space for the same number of computations. However, the less accurate the potential, and the stronger the

bonds or lighter the atoms (hence faster motion), then the shorter the suitable timestep. For systems containing light atoms such as H or C, usually 0.5 fs is most suitable. With ML potentials, it is rarely advisable to use a timestep longer than 1.0 fs for any system,[216] although with DFT it is possible to use 2–3 fs.[217] The integration scheme used for MD simulations in this thesis closely follows the velocity Verlet algorithm, with appropriate modifications introduced to ensure measure-preserving in the constant pressure and constant temperature ensembles, as described in Ref. [218].

2.6.2 Thermostats and barostats

I elect to use the Nosé–Hoover thermostat[219,220] to control the temperature. Conceptually, this involves coupling the system to a heat bath via an additional degree of freedom in the Hamiltonian, which introduces a parameter representing the timescale over which velocities are rescaled to match the desired temperature. Too slow, and the theromostatting is ineffective, too fast and unphysical temperature oscillations are produced. For the pressure control, the equations of motion we employ are those used in Ref. [221], which combines the hydrostatic equations of Ref. [222] and the strain energy terms proposed by Parinello and Rahman in Ref. [223]. A similar parameter is used for the barostat strength, where the analogy to the heat bath is the system volume acting as a piston, with fictitious 'mass' parameter, coupled to an external source of pressure, where volume adjustments are made to try and equalise the pressures.

2.6.3 Constrained simulations

Molecular dynamics samples according to the Boltzmann distribution. Often, the more chemically-interesting structures are those that are sampled very rarely under that distribution, such as transition states, or it is very hard to equilibrate the system fully if there are high activation barriers between wells in the PES. The techniques of enhanced sampling allow the exploration of these regions that are seldom visited during MD, but with an MD-like framework, by the addition of bias terms to the energy of the system based on collective variables to help with overcoming energetic and entropic barriers.[224] Some of the more popular methods include umbrella sampling,[225] metadynamics,[226] parallel tempering,[227] and replica exchange molecular dynamics.[228]

Separately, in molecular simulation there is a long history of coarse-graining molecular units to eliminate degrees of freedom and enable more efficient simulations, such as with the many proposed interatomic potentials for water.[229] In chapter 5 we combine the two by imposing collective variable constraints, enforced by energy penalties, to systematically explore two types of structure in amorphous MoS_3: triangular motifs and chains, without the dynamics leading to the destructive transformation to a nearby thermodynamic sink in crystalline MoS_2. To implement these whilst remaining within the efficient LAMMPS MD ecosystem, we use PLUMED as a plugin.[212,230]

2.7 Dimensionality reduction

In Machine Learning, data are most often represented as a 'feature vector' that encodes information in a format amenable to learning. In a chemistry context, we have seen how collections of atoms can be represented by fingerprint functions such as the SOAP, ACE, and MTP representations. They can also be represented by molecular connectivity, as with SMILES strings,[231] or properties, such as the melting point,[232] toxicity,[233] catalytic activity.[234,235] Whatever the featurisation, the vector representation will be high-dimensional (the length of the vector is greater than 3) and therefore it is difficult to visualise the valuable information contained within. In this section, I outline the three dimensionality reduction algorithms that are used in this thesis for analysing the information contained in the SOAP vector. For SOAP vectors, the axes in the reduced 2D space do not tend to be meaningful, so I plot them without axes throughout this thesis.

To perform dimensionality reduction on the SOAP descriptor, it is useful to first construct a distance matrix, $\mathbf{D}$, from the SOAP kernel matrix with elements $\mathrm{K}_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$ (recall section 2.3.2 for details of the kernel function) using the triangle inequality:

$$\mathrm{D}_{ij} = \sqrt{2 - 2\mathrm{K}_{ij}} \ . \tag{2.44}$$

By then reducing $\mathbf{D}$ into two dimensions, we can obtain a 'map' where the distance between points roughly corresponds to the distance in high-dimensional space, albeit with some information loss, of course. A useful analogy is the (non-linear) Mercator projection of the world map: it cannot be fully faithful to the full 3D

surface of the globe, e.g. the polar regions appear larger in area than they are in reality, but it is nonetheless useful for understanding world geography. Other projections, such as the Peter's, preserve area, but lines of constant bearing are no longer straight on the map.

2.7.1 Principal Component Analysis

In principal component analysis (PCA), the 'principal components' are constructed as linear combinations of the initial variables that capture as much of the variance in the data while remaining uncorrelated with each other (orthogonal).[236] More formally, they are the constructed recursively as the sequence of unit vectors where the i^{th} vector is the best-fit line (by least squares) to the data that is orthogonal to all the previous $i-1$ components. Therefore, the first component captures the most variance in the data, the second component the second most etc. It turns out the the principal components are the eigenvectors of the data's covariance matrix, so they can be calculated by singular value decomposition of it (for which there are many algorithmic implementations available). The advantages of PCA are that it is easy to calculate and deterministic—i.e. there is a single valid decomposition. However, it is a linear method and therefore is not best suited to non-linear data, which is often the case for atomistic descriptors.

2.7.2 Kernel Principal Component Analysis

KPCA is an extension of PCA that infuses some non-linearity into the decomposition via a kernel function, which makes this method particularly useful for clustering and is closely aligned with the spirit of Gaussian approximation potentials, which

also use kernel functions defined on the high-dimensional SOAP descriptor for an information-efficient fit.[237] The kernel function here is equivalent to the overlap integration (or the distance transformation of Eq. 2.7) over mutual rotations of two environments, so the projection for dimensionality reduction is from the $N \times N$ space, where N is the number of data points. The KPCA approach, as described, can be contrasted with projection from the $N \times M$ space, with M the number of components in the power spectrum, for a naive PCA on the SOAP vectors themselves.

2.7.3 t-Distributed Stochastic Neighbour Embedding

The t-SNE algorithm was developed specifically for *visualising* high-dimensional data.[238,239] Whereas PCA and KPCA focus on keeping the representation of dissimilar points far apart (in part due to the influence of the squared terms in making the best fit), t-SNE aims to keep the representation of very similar points close together. This leads to enhanced clustering, which enables the local and global structure of the data to be maintained, particularly if there are clusters at several scales. The drawbacks of t-SNE include relatively poor performance for intrinsically higher-dimensional data: the local nature of the algorithm, can produce spurious clusters in this case. The t-SNE algorithm is also not guaranteed to converge, and different embeddings are obtained depending on the random initial conditions.

The operation of t-SNE can be understood as optimising the initial coordinates based on a set of springs between map point y_i and all other map points

y_j. The spring between two points either repels or attracts depending on whether the distance between them is too small or too large to represent the similarities between high-dimensional datapoints according to the conditional probability that data point x_i would pick x_j as its neighbour if neighbours were picked based on the overlap of Gaussians centred on them:

$$p_{j|i} = \frac{\exp(-|x_i - x_j|^2/2\sigma_i^2)}{\sum_{k \neq j} \exp(-|x_i - x_k|^2/2\sigma_i^2)} \ , \tag{2.45}$$

where the similarity between map points is then defined as

$$q_{j|i} = \frac{\exp(-|y_i - y_j|^2)}{\sum_{k \neq j} \exp(-|y_i - y_k|^2)} \ . \tag{2.46}$$

The σ_i term for each point derives from the hyperparameter called the 'perplexity', which enables, for a single value of this parameter, the σ_i terms to vary depending on the local density. The perplexity can be interpreted as a smooth measure of the effective number of neighbours desired for the map.

In Fig. 2.11, I show the data used for the LPR evaluation of Si, but with PCA of the SOAP vectors, KPCA (via the SOAP kernel), and t-SNE.

2.8 Scattering experiments and radial distribution functions

Perhaps the primary, and most generally applicable, experimental technique for obtaining information about the structure of amorphous materials is by observing the total scattering—by X-rays or neutrons, and ideally both because of the complementary information that they contain. From the interference pattern of scattered particles it is possible to extract the two-body correlations between atoms

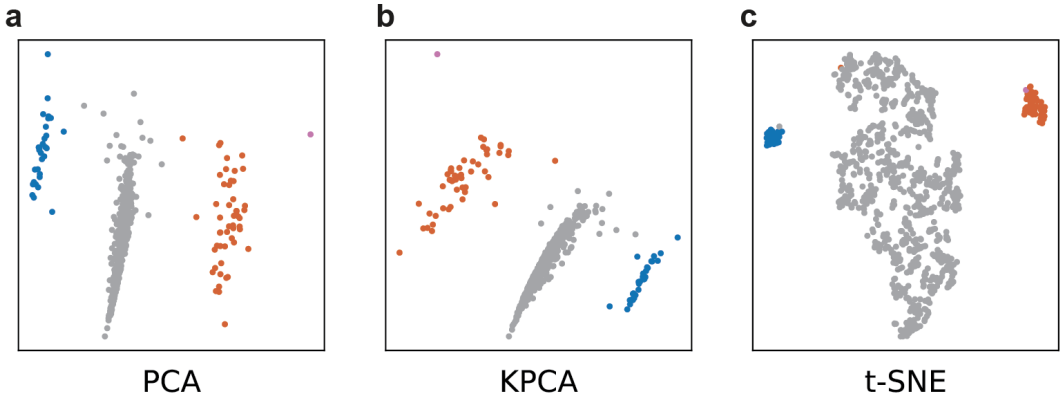


Figure 2.11: Comparison of dimensionality reduction techniques, using the silicon dataset of Ref. [5]. Points are coloured according to coordination number. (**a**) Principal component analysis. (**b**) Kernel principal component analysis. (**c**) t-Distributed Stochastic Neighbour embedding. The dimensions in panels a and b correspond to actual (linear combinations of) dimensions in the high-dimensional space. For t-SNE, these axes are not at all interpretable.

in the structure. I start with the (simpler) formalism for neutrons for isotropic materials. Taking after Ref. [240], for an isotropic material, the coherent part of the observed neutron cross section per particle per unit solid angle is, for a given particle energy window,

$$\frac{1}{N}\frac{\mathrm{d}\sigma}{\mathrm{d}\Omega} = \sum_{i,j=1}^{n} c_i c_j \bar{b}_i \bar{b}_j \left[A_{ij}(Q) - 1\right] + \sum_{i=1}^{n} c_i \bar{b}_i^2 \ , \tag{2.47}$$

where A_{ij} are the Faber-Ziman partial structure factors, c_i, $\bar{b}_i$ are the number concentration and coherent bound neutron scattering length of species i, respectively, and Q is the magnitude of scattering vector $\mathbf{k}-\mathbf{k}'$ of length $4\pi\sin\theta/\lambda$ for a neutron of wavelength λ scattered at an angle 2θ. These are related to the partial radial distribution functions (i.e. pair correlations), $g_{ij}(r)$, by the Fourier transform:

$$A_{ij}(Q) - 1 = \rho_0 \int_0^\infty 4\pi r^2 [g_{ij}(r) - 1] \frac{\sin Qr}{Qr} \mathrm{d}r \ , \tag{2.48}$$

and the $g_{ij}(r)$ are themselves defined as

$$g_{ij}(r) = \frac{n_{ij}(r)}{4\pi r^2 \mathrm{d}r \rho_j} \ , \tag{2.49}$$

where n_{ij} are the number of particles of type j between the shell r and $r + \mathrm{d}r$ centred on a particle of type i, and $\rho_j = c_j\rho_0$, with ρ_0 the total particle density of the material. The total structure factor, $S(Q)$, and total radial distribution function, $G(r)$ are then defined respectively as

$$S(Q) = 1 + \frac{1}{(\sum_{i=1}^{n} c_i\bar{b}_i)^2} \sum_{i,j=1}^{n} c_i c_j \bar{b}_i \bar{b}_j \left[A_{ij}(Q) - 1\right] \tag{2.50}$$

and

$$G(r) = \sum_{i,j=1}^{n} c_i c_j \bar{b}_i \bar{b}_j [g_{ij}(r) - 1] \ . \tag{2.51}$$

I also note here that the coordination number can be calculated by integrating radial distribution functions out to the first minimum in r via

$$\int_0^{r_{\min}} 4\pi r^2 c_j \rho_0 g_{ij}(r) \mathrm{d}r \ . \tag{2.52}$$

For X-rays, things are slightly more complicated because the equivalents of the scattering length ($\bar{b}$) terms are Q-dependent, a result of the lengthscale of the scattering object, the electron cloud, being commensurate with the radiation wavelength. The 'form factor' for X-ray scattering takes the place of the $\bar{b}$ terms, defined as

$$f_i(Q) = \sum_{i=1}^{4} a_i \mathrm{exp}\left[-b_i\left(\frac{Q}{4\pi}\right)^2\right] + c \ , \tag{2.53}$$

where a_i and b_i are tabulated coefficients for each atom/ion.[241,242]

Qualitatively, high-Q oscillations in the $S(Q)$ contain information on the shorter-scale, local structure in the material. Experiments must probe high Q ranges to achieve good spatial resolution (0.1 Å resolution requires at least 30 $\mathrm{Å}^{-1}$ in Q), and to reduce the termination ripples arising from the Fourier

transform from reciprocal space to real space. At low Q, there is often a first sharp diffraction peak (FSDP) in disordered materials. This precise interpretation of this feature in the $S(Q)$ has been the source of some debate over the years,[42,66,243] but, albeit rather vaguely, it can be agreed that the FSDP contains information on medium-range ordering in the structure. Layered structures,[244,245] cluster-like regions,[246,247] chemical ordering of voids in the structure,[248,249] and quasi-crystalline character[250,251] have been explored as the structural origins of FSDP. ML potentials, with their slow quench rates and large system sizes, are now allowing previously inaccessible studies of the order that contributes to the FSDP,[14,252] as is also exemplified in the work of this thesis.

In the very low-Q region, the limit $S(Q \to 0)$ provides a measure of the 'hyperuniformity' of an amorphous material,[253,254] which can be defined, loosely, by the absence of density fluctuations over long length scales. Hyperuniformity has relevance in optical properties for materials and is of fundamental interest in fields as diverse as photonics,[255] avian eye evolution,[256] and cosmology (the distribution of galaxies appears to be hyperuniform).[257] High quality experimental samples of a-Si approach perfect hyperuniformity, which is one of the reasons this material has acquired the status of canonical disordered material.[258] However, practical constraints associated with low-angle reflections and surface roughness make it difficult to probe accurately the $Q \approx 0$ region experimentally, and, equally, computational models would require an infinitely large simulation cell to achieve a simulation at $S(Q) = 0$. It is possible to make very large structures that nevertheless extrapolate with reasonable accuracy to the $Q = 0$ limit, which I

address in chapter 3 with a million-atom model of a-Si.

Any structural model, in order to be consistent with experiment, must match the $S(Q)$ reasonably closely in both high- and low-Q regions. However, there is no unique structure that can result in perfect agreement, since only the two-body correlations are contained in the $S(Q)$. Hence, reverse Monte Carlo methods (RMC)[72,259] that adjust the atomic positions to match the data are able to make very close fits to data using models that are chemically nonsensical. For the example of a-Si, even with bonding constraints, a high number of spurious triangular Si_3 units appear in RMC simulations unless specifically penalised.[260] There has been much research on the development of chemical,[261] structural,[74] and energetic[262] constraints to enable useful model building using RMC.[263]

3 | Teacher–student models for accelerated simulations

3.1 Acknowledgements

The work described in this chapter has been published as Ref. [1], from which any figures labelled as 'adapted' have been drawn. I thank Simon Batzner (Google Deepmind) and A. V. Shapeev (Skolkovo Institute of Science and Technology) for helpful comments on the original manuscript, and Louise Rosset for using the student potential produced in this work to study paracrystallinity in silicon, which has led to further insights on the efficacy of the teacher–student method.[264]

3.2 Introduction: the teacher–student concept

ML potentials are orders of magnitude faster than DFT. However, they incur a much larger computational cost than do efficient empirical interatomic potential models (which can routinely treat millions of atoms). In addition to developing increasingly accurate and fast potentials, there is a similarly important need for validating ML potentials. This is challenging because there is little physical information encoded into the model *a priori*, and because there is not always an unambiguous correlation between the numerical error of a potential (on a fixed test set) and its behaviour in simulations.[193] The importance of physically motivated tests has been demonstrated, e.g. in a comparative study of carbon interatomic potentials.[265] A popular benchmark study compared different ML potential fitting frameworks and their numerical performance in relation to the required computational cost.[189] Specifically for silicon (Si), various tests, including

random structure searching, have been devised and are outlined in chapter 2.[45,132]

Alternative ML fitting methodologies exist that are much faster still than GAP; one example is the Moment Tensor Potential (MTP) framework.[90,96] Many of these alternatives suffer from poorer extrapolation behaviour compared to GAP models for structures dissimilar to those encountered during training.[128] In this chapter, I describe work that we have published in Ref. [1] on a methodology using an accurate, but more computationally expensive GAP model to generate reference data—structures ('locations') and energies (known as 'labels' in the ML jargon)—for training a series of much faster MTP models. This GAP-generated data are referred to in the ML literature as 'synthetic data'[108] (not to be confused with *chemical* synthesis, as discussed in Section 2.3.1). For deep learning ML models, synthetic data are often much cheaper to acquire than 'real' data. A quite separate example of this is using video game footage for applications in self-driving cars[266]—the counterpart being footage of actual car journeys to quantum-mechanical calculations in this example, because of the higher computational expense of acquiring DFT data vs. those generated by ML potentials. We referred to the idea of using one ML potential to train another with the goal of capturing the original DFT as 'indirect learning'. This work is related to 'knowledge distillation' in ML research,[267] which seeks to transfer information from large 'teacher' to smaller 'student' models, enabling faster predictions. So far, distillation has been applied primarily to classification tasks for neural networks, such as image recognition,[267] although applications in physical modelling are becoming increasingly important,[268–270] especially with the arrival of foundation

models: i.e. neural networks that are trained on vast datasets, often based on transformer architectures with on the order of 10^{11} parameters.[271,272] Their size makes them prohibitively expensive at runtime for many applications, so it is attractive to distil their knowledge into smaller models for particular tasks, or fine-tune them to avoid highly-expensive retraining from scratch.[20,31,273] The term 'database distillation' was recently described as a way to condense information into smaller training sets;[274] we think of the term 'distillation' in a rather broad sense—to encapsulate the indirect learning of potentials via any combination of fitting frameworks, and the generation of new atomic *configurations* for training in addition to the energy and force data themselves. In other words, we use the teacher model to generate both the data locations and the data labels. Figure 3.1 outlines the teacher–student concept; we use the notation M′ for a directly-learned MTP, and M″ for a student one. The teacher GAP model is taken from Ref. [45], namely GAP-18. Without the need for additional quantum-mechanical computations, extensive databases can be easily generated with GAP-18, and we find that these improve the quality of fast potentials with less flexible functional forms. We test the technique on disordered silicon, including a simulation of vitrification and polycrystalline grain formation under pressure with a system size of a million atoms.

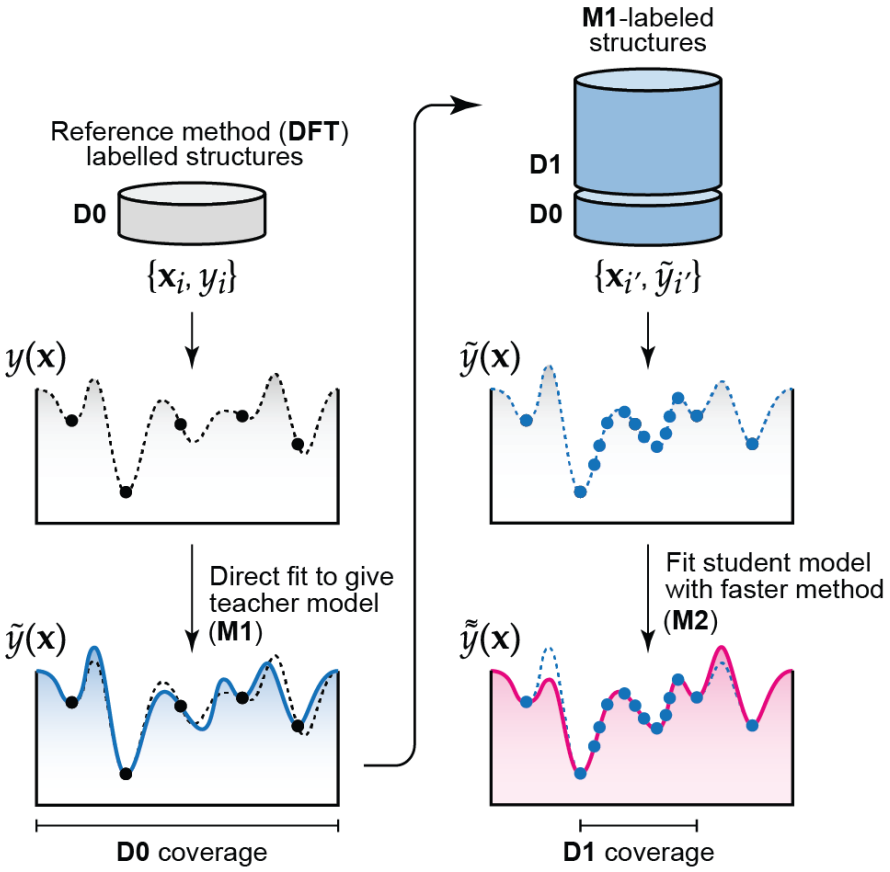


Figure 3.1: A teacher–student method for interatomic potentials. *Left:* Conventionally, to fit ML potentials, a PES is sampled by reference computations (typically at the DFT level) for a given number of structures, leading to the dataset **D0**. Fitting with the method **M1** yields an ML potential model (blue line). *Right:* We now generate a much larger set of structures using **M1**, and label them using the same method (i.e. the teacher model, not DFT). This 'synthetic' dataset, **D1**, covers a narrower structural space, but at more points. A second fit is then made with a different ML method, **M2**, to yield the student model. Reproduced from Ref. [1].

3.3 Essential methodology

ML potentials approximate a potential-energy surface (PES) based on reference data, typically from small-scale DFT computations:[100]

$$\{\mathbf{x}_i, y_i\}_{i=1..N} \xrightarrow{\mathbf{M1}} \tilde{y}(\mathbf{x}), \tag{3.1}$$

where $\mathbf{x}_i$ is a descriptor for the atomic environment,[123] y_i is one of N observations of the reference PES, and $\tilde{y}$ denotes the ML model fitted with a given method.[i] We refer to this method as '**M1**'.

Two fitting methodologies at different points on the Pareto front[ii] of accuracy and cost,[189] namely the GAP[12,100] and MTP[90] frameworks, can be combined to accelerate simulations of disordered Si. To do so, we train MTPs using an existing general-purpose ML potential for Si, referred to as GAP-18,[45] as the reference method. We use the MTP formalism (denoted **M2**) to approximate the GAP-18 PES (**M1**), which is itself approximating the DFT PES (**M0**). Hence, **M2** is *indirectly* learning quantum-mechanical reference data with **M1** as its teacher (Fig. 3.1a). Extending the notation of Eq. 3.1,

$$\{\mathbf{x}_i, \tilde{y}_i\}_{i=1..N_2} \xrightarrow{\mathbf{M2}} \tilde{\tilde{y}}(\mathbf{x}), \tag{3.2}$$

where $\tilde{\tilde{y}}$ is a student model, and $N_2 \gg N$. We use the notation M′ for a directly-learned MTP (against which to compare the student), and M″ for a student one. We note that the idea of mapping a kernel-based ML potential to a fast approximator has been explored previously: via cubic spline interpolation across a pre-calculated grid of points.[98,275] The idea of using one ML potential to train another had not been explored, to our knowledge, at the time of publication of our Ref. [1]. A similar approach has since been used in the deployment of the DPA-2 foundation model of Ref. [32].

[i] In practice, things are more complex: typically, atomic energies are machine-learned from total energies (and usually forces). Section 2.3.3 contains further details.[8,12]

[ii] The line denoting the model choice with optimal accuracy for a given cost.

There are two primary advantages to using GAP-18 as an intermediary rather than training directly on DFT-labelled data. Both hinge on the massively reduced computational cost of GAP compared to DFT. Firstly, we can build an almost arbitrarily large database of GAP-labelled training data, thus approximating the **M1** PES as faithfully as possible within the (flexibility and fitting) limitations of **M2**. Secondly, we can evaluate the fidelity of the fit via numerical errors on a wide-ranging set of out-of-sample testing data, and also test behaviour in large-scale molecular dynamics (MD) simulations. The practical advantages of our application of distillation rely on the observation that GAPs may be able to more accurately interpolate across a sparsely sampled PES compared to MTPs, which is not yet fully investigated.[128,189] We expect, however, that the same concept could be applied to any pair of fitting methodologies where a cost-accuracy disparity exists. The insights into database development and physically-guided validation of ML potentials are independent of the particular **M1** and **M2** methods chosen.

Moment tensor potentials

The formulation of MTPs is discussed in detail in Section 2.3.2. The complexity of the functional form of an MTP grows exponentially with the important hyperparameter $\mathrm{lev}_{\mathrm{max}}$, which determines the number of parameters in the model—hence the expressiveness and, typically, the amount of data required to avoid overfitting. In the work described in this chapter, we investigated the range $12 \leq \mathrm{lev}_{\mathrm{max}} \leq 26$. For brevity, I denote $\mathrm{lev}_{\mathrm{max}}$ as 'L' hereafter.

We found that the performance of MTPs was quite sensitive to the relative

weights of energies, forces, and stresses in the fit. After some testing, we settled on weighting ratios $10 : 1 : 0.01$ respectively. Our chosen force weight is a factor of 10 larger than the commonly-used default value, as we found that both direct and student MTPs suffered from unstable MD and poor predictions of the physical behaviour under pressure using the default force weight. Full details of all remaining MTP hyperparameter choices are given in Table 3.3 at the end of the chapter.

3.4 Synthetic database

Fig. 3.2 visualises the regions of structural space covered by the GAP-labelled, GAP-generated synthetic data (in red to orange) as compared to the original (**D0**) GAP-18 database that is common to all models in this chapter (blue). The distances between points corresponds to the distance between structures (i.e. their dissimilarity) defined by a 2D embedding of the SOAP distance matrix.[13,276] The SOAP kernel gives the similarity between two atomic environments on a scale from 0 (dissimilar) to 1 (identical up to rotations and translations), and is described in more detail in Section 2.3.2. In Fig. 3.2, at low pressure, a random starting structure, which rapidly transforms to approximate low-density amorphous (LDA) Si, is met with the good coverage provided by liquid and LDA structures in **D0**. Equally, the high-pressure region is densely sampled with distorted β-Sn-type and simple hexagonal (sh) structures.

The region straddling the low-pressure amorphous Si (lower left) and high-pressure crystalline Si (lower right) is notably sparse in the original training data, which perhaps explains why poorer extrapolators struggle in predicting the

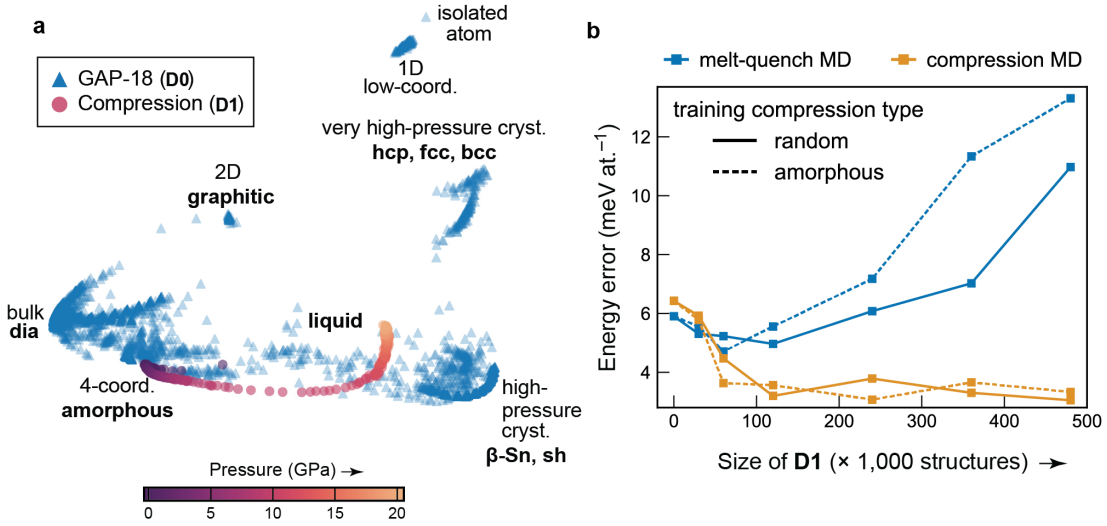


Figure 3.2: Composition of the ML potential fitting databases for teacher and student models. (**a**) A SOAP/KPCA map (cf. Ref. [276]) compares the structural space covered by the GAP-18 database,[45] **D0**, and the newly created indirect learning dataset, **D1**. The distance between points on this map is a 2D embedding of their structural dissimilarity as measured by SOAP.[13] The sizes of data points are proportional to the number of atoms in the corresponding training structures, and hence to the number of forces that contribute to fitting. (**b**) Mean absolute error of predicted energies versus quantity of compression MD training data of two types: starting separately from LDA structures (dashed lines) and randomly placed hard spheres (solid lines). 3 M''_{16} MTPs with the same settings were trained for each database and we report their average errors in energy prediction vs. GAP-18 on two independent validation sets: structures derived from melt-quench MD (blue) and isothermal compression MD (yellow). The median standard deviation in error across identically-trained MTPs, $\sigma_{\Delta E}/|\Delta E|$, is in the region of 10–45 meV. The MAEs show improvement with the addition of similar data to the training set, which is converged with respect to database size by 240 × 103 atomic environments. Adapted from Ref. [1].

transition at intermediate pressures around 10 GPa, despite them capturing both extrema well. The synthetic data of **D1** forms a 1D manifold within the distance embedding space, parametrised by the pressure, which now covers this region more densely. To improve interpolation for the high-pressure structures, we use a large training database (**D1** in Fig. 3.1a) containing 250 structures with 1,000 atoms each, which sample compression up to 20 GPa. Note that this kind of data would be practically inaccessible to label with DFT due the system size. However, in

order to have a simulation cell large enough to contain inter-grain regions that sample the kind of atomic environments prone to the failure seen in Fig. 3.3, 1,000 atoms is virtually the minimum size that can be used. Initial tests showed that including too few compression MD environments leads to erroneous behaviour in simulations (*vide infra*). The errors for melt-quench MD begin to degrade with the increased proportion of compression MD included during training (between 1.2×10^5 and 2.4×10^5 environments). Including even more such environments results in an unstable potential for MD simulations. The dependence of simulation quality on the quantity and type of training data in **D1** emphasises the need for the physically-guided validation of the resulting potentials—as with all machine-learned potentials—which we explore in the next section.

Figure 3.3 illustrates a typical failure mechanism of an ML potential during a compression-induced crystallisation simulation. The system (elemental Si) is initially prepared in amorphous form at ambient pressure. Ramping the pressure linearly at a rate of $10^{11}\,\mathrm{GPa\,s^{-1}}$ causes simple hexagonal (sh) nanocrystals to nucleate. As the nanocrystals begin to meet across the periodic boundary, the region impinged by two misaligned grains is placed under significant strain. Often, this strain produces an atomic environment that is poorly represented in the training set used to train the potential. If the extrapolation accuracy of the potential is poor and/or there is an insufficient prior that predicts high energy for such configurations, the potential could predict erroneously low energies for unphysical atomic geometries and get trapped 'down a hole'. In Fig. 3.3, the result is a point defect with an unreasonable coordination number that is tantamount to

overlapping atomic cores.

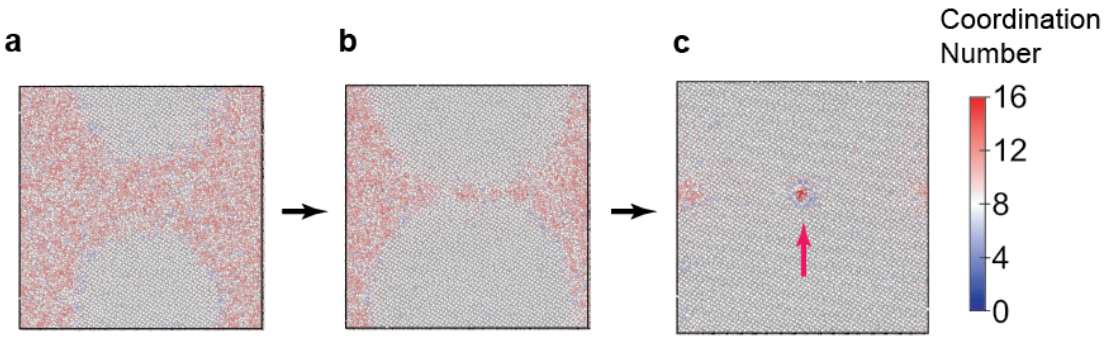


Figure 3.3: Typical failure mode during compression simulation. A crystal nucleus forms from an amorphous phase as the pressure increases. As it grows, the grain boundary region leads to a strained environment for which the direct MTP makes qualitatively incorrect predictions (it stabilises a very high coordination number).

The common failure mode shown in Fig. 3.3 is not suffered by the GAP-18 model, as is shown empirically in Ref. [203], where the predicted sequence of phases is in agreement with the available experiments.[277,278] In our particular application to Si, the avoidance of such a failure lends credence to our selection of GAP-18 as the teacher model, with a view to transferring this robustness to the student model whilst maintaining the computational speed of the MTP functional form. The conduit for this transfer of robustness is the synthetic dataset, which ought to cover the regions in which the MTP is more likely to fail (i.e. high pressures and grain boundaries).

3.5 Results

Once an ML potential has been trained, the immediate question is to what extent it is accurate and reliable. However, testing the predictions of a potential is not straightforward.[189,279] Developing robust strategies for doing so forms a central theme throughout this thesis through two primary means:

- Numerical errors: calculating a root-mean-square error in force and energy predictions on a testing set of structures not included during training is the simplest and most common means of evaluating a potential. The significance of such a metric depends heavily on which structures are included in the testing set, in particular how diverse they are (e.g. it is not challenging to model, with low errors, structures very similar to the known crystalline phases).

- Physically-guided validation: in this chapter, we test the predictions of phase transitions by potentials in MD simulations, which is known to be challenging because of the unusual non-equilibrium structures that appear along the pathways that link stable phases.[203]

3.5.1 Numerical errors

The accuracy of an ML potential is often characterised in the community by this error value versus its reference method,[189] typically evaluated on a test set not included during training. If the error is within the typical noise of the reference method, for instance, the amount by which different DFT functionals disagree (a few meV), then the ML potential is often considered of '*ab initio* accuracy'.[280] In this vein, Fig. 3.4 shows the mean absolute error on different test sets for a series of M′ and M″ models of increasing complexity and cost. We test on both RSS configurations, which are different from the fitting database, and also on snapshots from small-scale compression MD simulations in order to mirror the physical conditions under which the teacher–student method is intended to provide

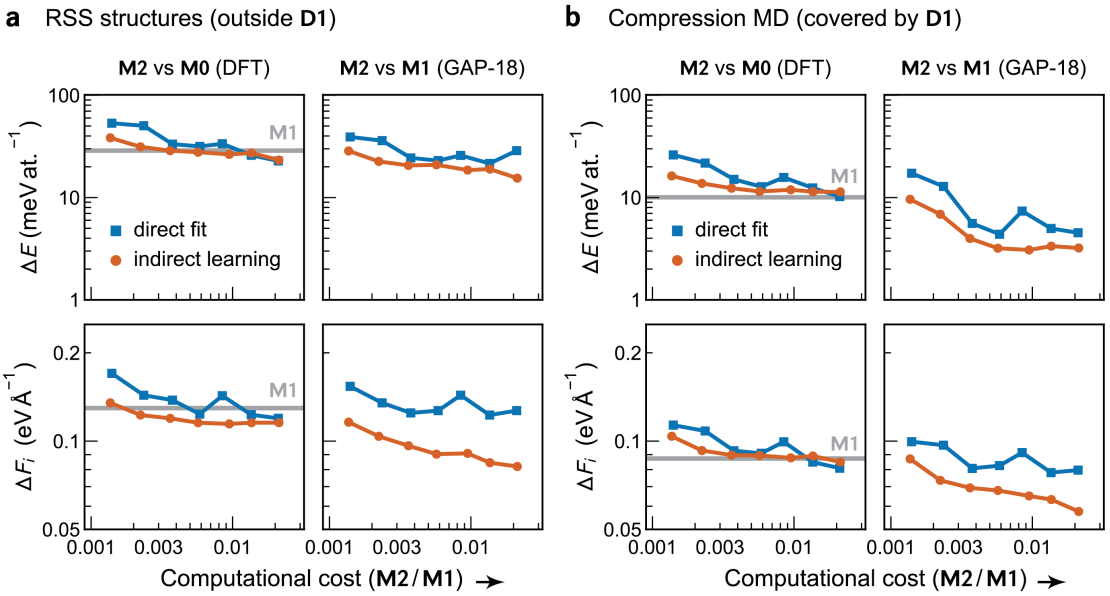


Figure 3.4: Numerical errors vs. computational cost for directly learned (blue) and student (red), or equivalently 'indirectly learned', potentials. This figure characterises MTPs with increasing number of fitting parameters, expressed via their respective compute-time per MD step relative to GAP-18 (**M1**). (**a**) Mean absolute error for random structures that were not part of the training. The test comprises 100 cells of 1–64 atoms produced using `buildcell`[114] and optimised with GAP-18 at pressures in the range of 0–32 GPa. For comparison to **M0** (**M1**), test-set structures were labelled with DFT, (GAP-18), respectively; grey lines report the prediction errors of GAP-18 vs DFT for those structures. This test conveys the trade-off between flexibility and efficiency in choosing the number of parameters in MTPs via the maximum level, L (here, set to values of 12–26 in increments of 2). (**b**) As before, but now for 64-atom compression MD snapshots: in this case, configurations of this type have been included in the indirect learning database, **D1**. Reproduced from Ref. [1].

a robustness advantage.

For both test sets, Fig. 3.4 provides encouraging results: comparing their predictions to DFT, student MTP models tend to at least match, if not improve upon, the corresponding directly-learned (**D0**-trained) MTPs. The force-component errors (lower panels in Fig. 3.4b-c) indicate that our student MTP models faithfully recover the predictions of **M1**—suggesting that the existing **M1** PES is smooth enough to be learned by a different method. More generally, the error analysis suggests that improvements to **M1** can be expected to transfer to

subsequent student models.

3.5.2 Physically-guided validation of ML potentials

While numerical errors provide some indication of performance, the correlation between both is unreliable enough to warrant the extensive validation of ML potentials based on physical behaviour—for example, assessing whether the structural characteristics of a reference simulation are reproduced. The SOAP kernel was introduced in Section 2.3.2 and used in Fig. 3.2a as a means to visualise the structural space of the training dataset.[13] SOAP has proven useful for analysing local structure in amorphous Si to evaluate how closely structural models agree.[3,203,281] We use it here to benchmark the progress of compression MD simulations compared to the trusted reference model. The computational speed of MTPs allows us to compare several different student potentials in large-scale (100,000-atom) simulations. Systems of even 10,000 atoms would successfully complete without encountering the failures described in Fig. 3.7, either because a failure event, by chance, did not occur, or a structural system-size effect came into play, e.g. suppression of grain boundary formation.

3.6 Compression benchmark

Figure 3.5 introduces the compression benchmark for Si. The colouring in panel **a** derives from the SOAP similarity of each *atomic* environment to the reference crystal structure (in this case, sh). Moving to panel **b**, we average the atomistic SOAP vectors for environments within each cell before evaluating the similarity to the reference crystals (diamond and sh), which provides a physical interpretation

of the average structure in a large simulation system.

The LDA phase from early in the trajectory is highly similar to diamond, the polycrystalline phase is similar to sh. The difference between dashed and solid lines in Fig. 3.5 comes from our methodological refinement for producing reference structures. We are interested in the structural changes that occur during MD simulations with increasing pressure beyond the trivial isotropic contraction of bond lengths. To subtract this effect from that of the SOAP similarity, we compare our simulation trajectory not only to ambient-pressure crystals ('fixed' in Fig. 3.5), but also to crystals with lattice parameters that are optimised as a function of pressure. At 1 GPa intervals, we relax the lattice parameters of the reference crystals using DFT whilst maintaining their space groups and interpolate between the parameters when calculating the similarity. That way, the reference crystal is under the same pressure as the amorphous phase as measured over each ps of MD simulation time ('scaled' in Fig. 3.5). More detailed inspection of the MD trajectory confirms the interpretation that the LDA phase does not change substantially beyond an isotropic contraction until it abruptly collapses to a very-high-density amorphous phase (VHDA) around 12 GPa,[203] seen as a sudden drop in the **dia** similarity. The VHDA phase has the closest similarity among tested crystals to the high-pressure Si-VI phase, also known as *Cmca*, which appears experimentally upon compression to about 40 GPa, after the formation of sh and before hexagonal close-packed (hcp) Si. At about 14 to 16 GPa, the nucleation of sh crystallites begins,[203] so the similarity to *Cmca* decays slightly and similarity to sh increases. At this point, characterising inhomogeneous polycrystalline sh (pc-sh) with any

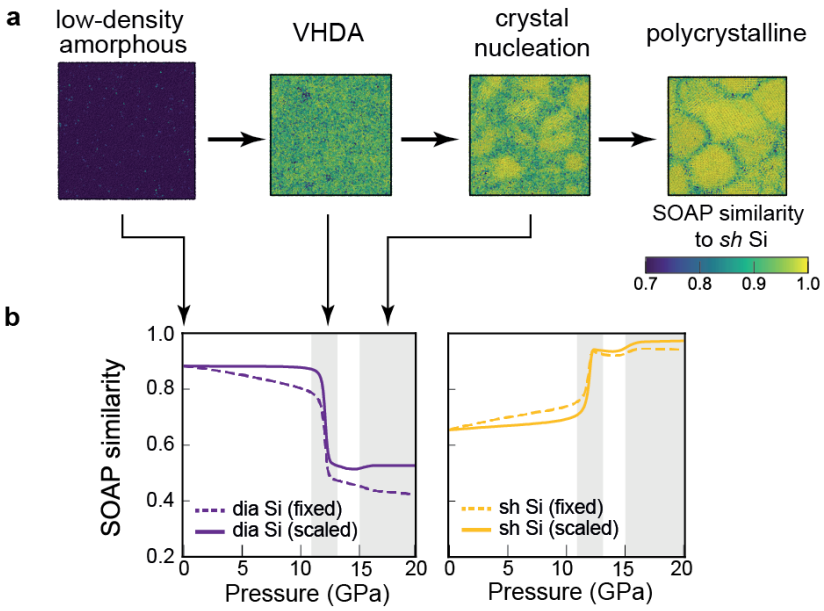


Figure 3.5: Compression benchmark for validating potentials. **(a)** A structural benchmark for high-pressure disordered Si. We evaluate the average SOAP similarity of the compression trajectory of Ref. [203] to selected crystalline structures. Dashed lines are for reference crystals with a fixed lattice optimised at 0 GPa with DFT. Solid lines are similarities to crystals with their lattice parameters optimised as a function of pressure. The structure images show representative snapshots, colour-coded according to SOAP similarity to the sh crystalline phase, with lighter colours indicating higher similarity.

single variable becomes fraught, which also explains why the *Cmca* similarity remains so high (along with β-Sn; see Fig. 3.6). The interfaces between sh grains, which are more similar to *Cmca* and β-Sn than they are to sh, occupy a reasonable fraction of the simulation-cell volume, hence contribute appreciably to the average similarity. Nevertheless, the analysis in Fig. 3.5 summarises the structural changes during compression in a comprehensive way.

More generally, we think that compression MD simulations starting from amorphous or disordered structures can provide a stringent and insightful test for the qualitative physical behaviour of ML potentials—especially if a large system size is used, which permits the frequent sampling of a range of configurations

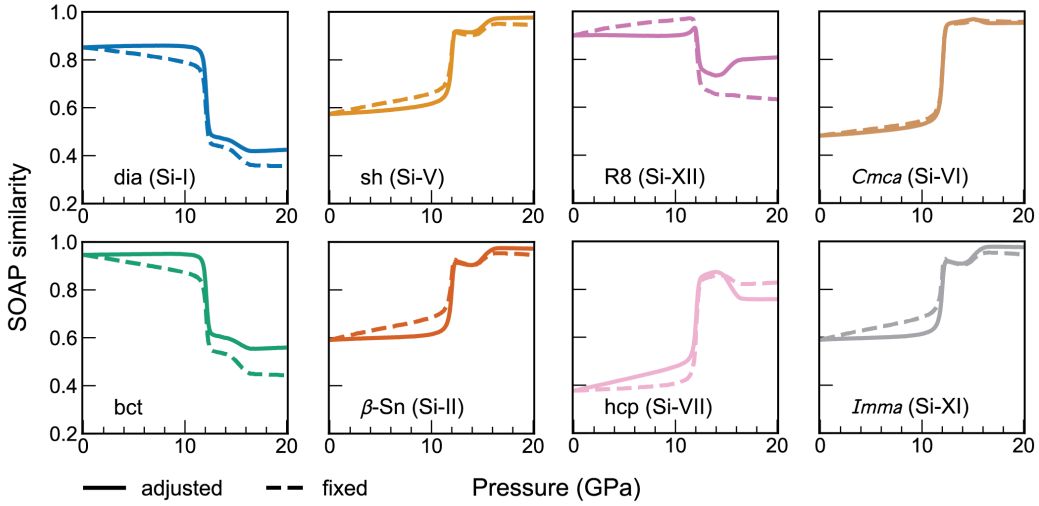


Figure 3.6: Using pressure adjustment for benchmarking against known crystalline phases of Si. Extension of Fig. 3.5 to include further crystal structures: comparing the average atomic environment during compression MD to reference structures. Dashed lines are for reference crystals with fixed structural parameters as optimised using DFT at 0 GPa. Solid lines indicate SOAP similarities to crystals with their lattice parameters optimised using DFT as a function of pressure. The rhombohedral phase, R8, results from decompressing β-Sn, which contains 5 and 6-membered rings of tetrahedrally-coordinated Si (see Ref. [282]). The bct phase is a hypothetical four-fold coordinated structure predicted to be stable under tensile stress.[283] The metallic phases commonly labelled with their orthorhombic space groups, *Imma* and *Cmca*, appear with increasing pressure respectively between β-Sn and sh and between sh and hexagonal close-packed (hcp) phases. Anzellini et al.[277] provide an introduction to the numerous Si phases and their nomenclature.

involving the close approach of atoms in various geometries. The increasing coverage and ready accessibility of structural databases, such as the Materials Project via API call,[25] mean that reference crystal structure data are available in an automatable way for many chemical systems, so structural similarity analyses such as those exemplified in Fig. 3.5 can be easily set up. Equivalent plots for the other known phases of Si in the Materials Project database are shown in Fig. 3.6.

3.7 Student MTPs outperform directly-learned MTPs

Figure 3.7 benchmarks the series of MTPs, directly-learned (blue; M'_L) or student (red; M''_L), with increasing quality and cost. All compression simulations start from the same low-density amorphous (LDA) structure from Ref. [203] and are run using LAMMPS.[212] We use the similarity to sh Si as quality measure, aiming to quantify how well a given model reproduces the behaviour of GAP-18 (**M1**). Overall, student MTPs perform much more reliably at the compression test than those directly fitted (to **D0**), with every M''_L model producing the expected sequence of phases,[203] that is, LDA $\longrightarrow$ very-high-density amorphous (VHDA) $\longrightarrow$ polycrystalline (pc) sh. In contrast, only 2 of the 5 directly-learned models perform acceptably. Such comparisons are only feasible because **M1** is efficient enough to produce a reference simulation (albeit with considerable effort), and **M2** is efficient enough to readily produce many of such simulations.

3.8 Application: 1-million atom simulations of Si

Finally, we test our student potentials for a task for which they were not specifically trained (in **D1**): the production of LDA structural models via simulated melt-quenching. Well-established indicators are available[14,284,285] to measure how well our student models reproduce the behaviour of GAP-18 (Table 3.1). We run quenching simulations with 100,000 atoms, as in Ref. [203], now comparing two representative levels of MTP fitting. In both cases, the student model outperforms a directly-learned one at the same level. Both M″ models come close to reproducing the defect count, with that of 3-fold bonded atoms being almost identical (0.68%,

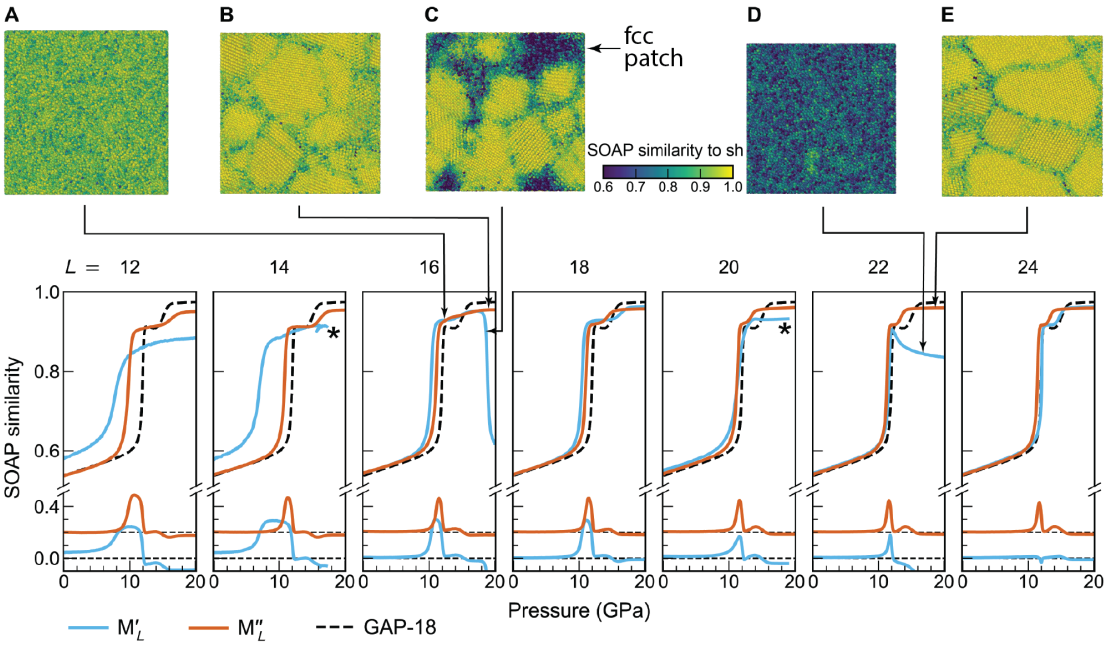


Figure 3.7: Systematic evaluation of MTP series via the compression benchmark. Points where MD simulations become numerically unstable (for M'_{14} and M'_{20}) are marked with asterisks. Performance of student learned models of increasing flexibility (controlled by the maximum level, L), in 100,000-atom compression simulations of amorphous Si. We report the average SOAP similarity to pressure-adjusted sh Si during compression (cf. Fig. S1), benchmarking MTPs (solid) vs GAP-18 (dashed). The difference for each MTP to GAP-18 is plotted underneath the absolute values, with offset. Representative snapshots during the run are displayed, coloured by atomistic SOAP similarity to sh (A)-(E) with arrows indicating approximate locations. Points where MD simulations become numerically unstable are marked with asterisks. Reproduced from Ref. [1].

0.69%, and 0.70% with M''_{16}, M''_{20}, and GAP-18, respectively), and that of 5-fold ones being somewhat more varied. There is not, however, a unambiguous advantage of M''_{20} over M''_{16}: specifically, the inverse height of the first sharp diffraction peak, H^{-1}, is much better described by the latter. Aside from H^{-1}, for which M''_{16} performs best of all models, MTPs largely recover the same, highly-accurate $S(Q)$ compared to experiment as does GAP-18 (Fig. 3.8). This agreement with the $S(Q)$ leads to a good match between the predictions of M''_{16} and the experimental $G(r)$ and the ring size distribution of GAP-18 (shown in Fig. 3.9), which gives us confidence in analysing the 1M-atom LDA structure in detail in the subsequent

chapter.

Table 3.1: Quality metrics for 100,000-atom LDA Si models produced by 10^{11} Ks^{-1} quench simulations: proportion of 3- and 5-fold connected atoms (N_3 and N_5), determined using a cutoff of 2.85 Å; inverse height of the first sharp diffraction peak (H^{-1});[253] mean ($\bar{\theta}$) and width ($\Delta\theta$) of the bond-angle distribution; and energy above diamond-type Si calculated with GAP-18.

Model	N_3 (%)	N_5 (%)	H^{-1}	$\bar{\theta}$ (deg)	$\Delta\theta$ (deg)	ΔE (eV)
M'_{16}	0.57	2.37	0.603	108.99	10.85	0.156
M''_{16}	0.68	1.64	0.560	109.15	10.23	0.152
M'_{20}	0.47	3.65	0.498	108.93	10.91	0.146
M''_{20}	0.69	1.33	0.504	109.16	9.93	0.144
GAP-18	0.70	0.95	0.564	109.17	9.99	0.145

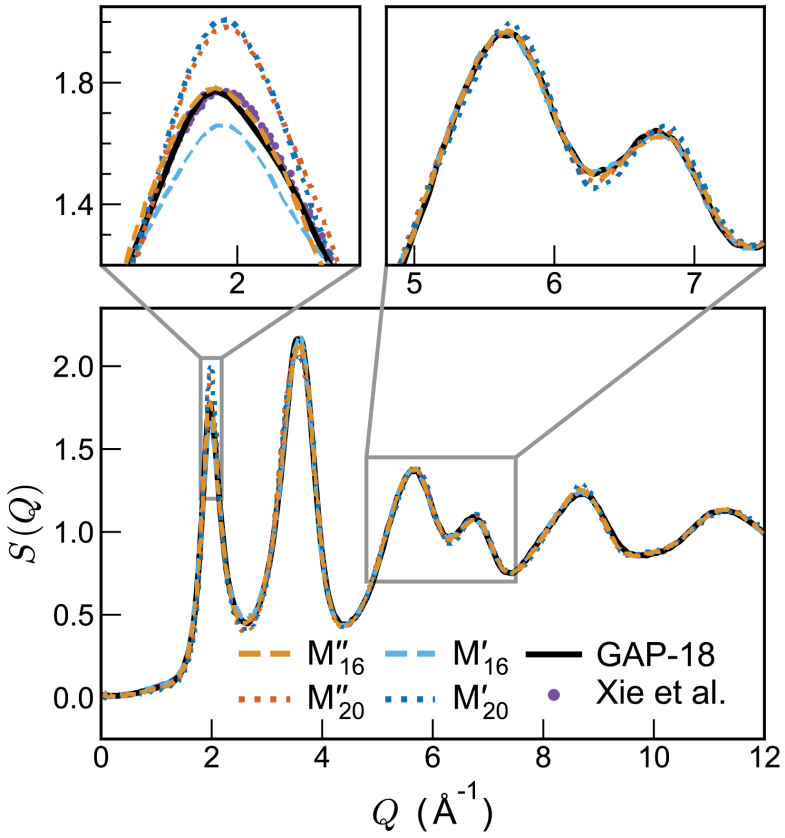


Figure 3.8: $S(Q)$ prediction for selected MTP models vs. GAP-18 and experiment. The model produced by M''_{16} agrees best with GAP-18 and experiment[253] in the FSDP region (top left inset).

[iii] available at https://github.com/MorrowChem/RingsStatisticsMatter.jl

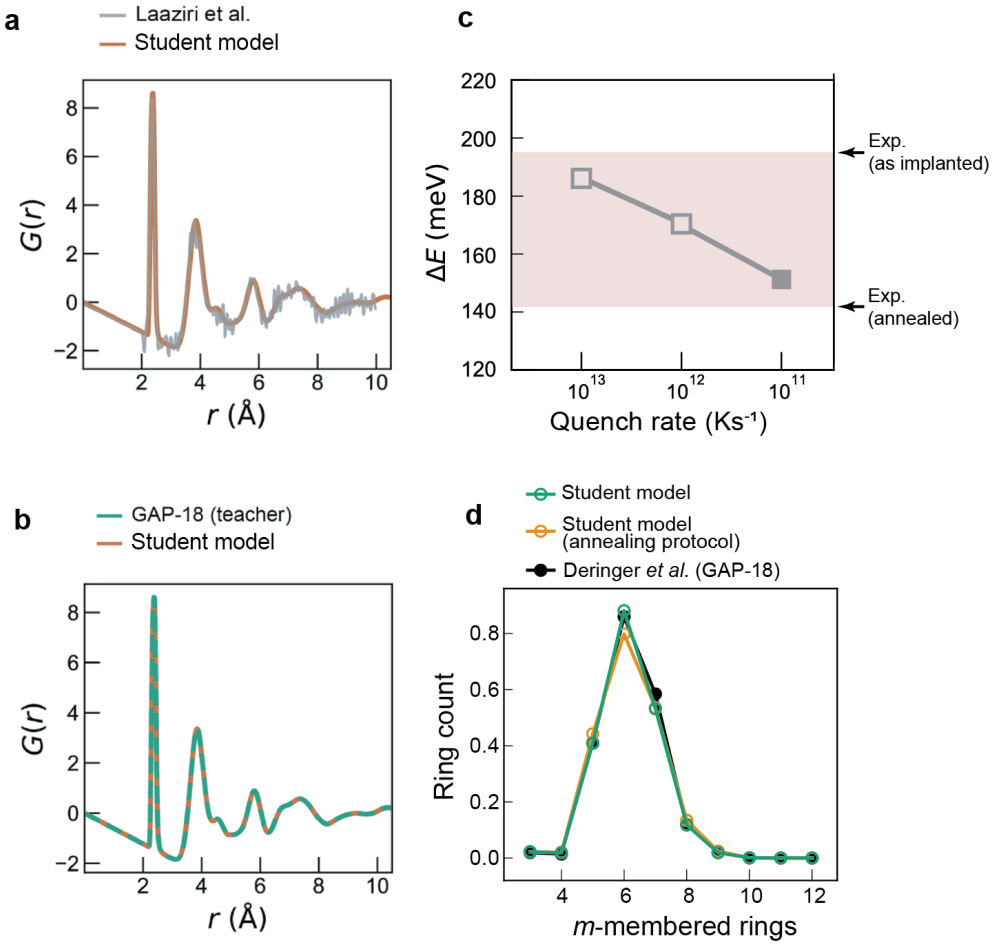


Figure 3.9: Comparison of structural and energetic indicators for 1M-atom model with experiment. **a** Total radial distribution function, $G(r)$, for the 1M-atom model of chapter 3 compared to the experiment from Ref. [56]. **b** $G(r)$ for the same structural model compared to the 100,000-atom model produced by GAP-18,[45] which is the 'teacher' potential as described in Ref. [1]. **c** Energy above diamond-type Si as a function of quench rate for 1M-atom student-derived structural models. The filled square corresponds to the primary model analysed in the next chapter. The pink region corresponds to the experimentally-determined excess energies of a-Si compared to the crystal for samples produced initially by self-ion implantation (upper) and then after annealing (lower).[66] **d** Ring statistics for the 1M-atom models compared to the 100,000-atom model produced by GAP-18. Ring counts per atom are displayed for both 1M-atom structural models: the one discussed in this chapter and an alternative annealed structure shown in Fig. A.5 (orange). The results are compared to those for a 4,096-atom reference structure (black) from Deringer et al. (Ref. [14]; obtained with the GAP-18 potential). They have very similar ring statistics, providing further validation of the quality of the 1M-atom model. The algorithm used to obtain ring counts is described in Ref. [286], which I have implemented[iii] efficiently enough to analyse 1M-atom structures.

To demonstrate its usefulness in practice, we employ our approach to generate million-atom structural models of LDA and pc-sh Si. The nanoscale grain structure of pc-sh invites the use of larger-scale models for more extended studies of the grain-boundary region and the size and orientation of crystallites. The efficiency of MTPs permits fast simulations on long timescales without having to compromise on system size. We perform melt-quench (Fig. 3.10a) and compression MD (Fig. 3.10b) simulations of a 1M-atom system, using M''_{16}[iv].

The LDA structure, visualised in Fig. 3.10a, has a bond-angle distribution closely similar to that for the 100k-atom GAP-18-generated structure (Fig. 3.10c). However, a more detailed analysis of the environments of the $N = 5$ atoms (coordination defects) is possible with larger system sizes, as is evident from the much lower scatter in Fig. 3.10c. The histogram bin widths are fixed at 0.25° for fair comparison—sufficiently narrow for analysis of small deviations away from the ideal tetrahedral angle. The broad distribution of bond angles in $N = 5$ atoms is consistent with their wide spread in local structure, as characterised in Ref. [281] for much smaller simulation cells. A detailed analysis of these defects is the subject of chapter 4. The structure factor and radial distribution functions are in very close agreement amongst experiment, teacher-derived, and student-derived structural models and the energetics are similar between student and teacher structural models for defects and 4-fold atoms, as shown in Fig. 3.9.

The expected pressure-induced collapse of the largely 4-fold connected structure into a VHDA phase[203,289] is reproduced by M''_{16}, the onset occurring

[iv] For recent, large-scale atomistic simulations with ML potentials, see, e.g. Refs. [287] or [288].

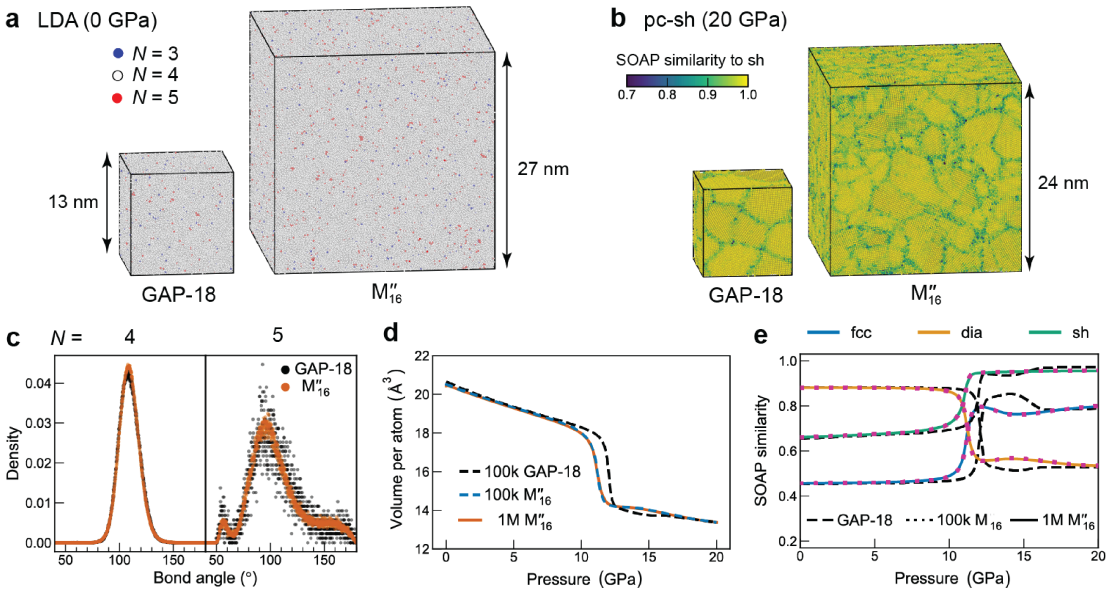


Figure 3.10: Application of accelerated student model to study a-Si at the 1-million atom scale. The importance of such a large scale will become apparent in the study of defects in a-Si described in chapter 4. (**a**) Structural models of low-density amorphous (LDA) Si, obtained by melt-quenching at $10^{11}\,\mathrm{K\,s^{-1}}$ to 500 K before relaxation. Colour-coding indicates coordination numbers, N. The structure on the left is the 100,000-atom system generated in an earlier GAP-18 MD simulation (taken from Ref. [203]); the structure on the right contains 1M atoms and was obtained with the M''_{16} potential. (**b**) Structural models as before, now for a polycrystalline simple-hexagonal (pc-sh) phase produced by compression to 20 GPa, and colour-coded by the per-atom SOAP similarity to single-crystalline sh Si. Close inspection of the 1M-atom structure reveals a mixture of thick, amorphous interfaces and coherent grain boundaries. (**c**) Bond-angle distribution function of the LDA structure (cf. panel a) compared to GAP-18. The larger system-size permits a detailed analysis of $N = 5$ coordination defects. (**d**)–(**e**) Pressure–volume and pressure–similarity curves during compression of LDA → VHDA → pc-sh.

at only slightly lower pressure than the reference GAP-18 simulation. There are no discernible differences between the volume–pressure curves of 100k-atom and 1M-atom simulations performed using the same student model (Fig. 3.10d), making system-size effects for VHDA *formation* unlikely. However, system-size effects in the nucleation of sh crystallites are clearly still important even in 100k-atom systems. The top face of the GAP-18 structure in Fig. 3.10b shows a crystallite extending all across the cell, hence infinitely in that direction under periodic

boundary conditions. The 1M-atom structure has similar grain sizes, but the larger cell ensures no grain approaches the length of the cell itself.

The differences in the volume–pressure curves between GAP-18 and M''_{16} (Fig. 3.10d) are of a similar magnitude to those between GAPs trained using different reference methods in Ref. [203]. The crystallisation and volume of the pc-sh phase are well reproduced, as is the behaviour as characterised using SOAP (Fig. 3.10e). Figure 3.10 therefore suggests that our student M''_{16} model *can*, to a large extent, match predictions of GAP-18. This way, more detailed analyses become possible—for example, of the grain structure in pc-sh Si, as well as a closer study of the LDA phase which is the subject of chapter 4.

3.9 Outlook

One takeaway from the work of this chapter is the importance of the database underpinning ML potentials—the adage 'garbage in, garbage out' does not, in my opinion, do justice to the converse framing: what constitutes 'quality' input? By supplying quite different fitting methods with carefully-chosen data types, we were able to bridge the performance gap between methods. A similar paradigm is the pre-training/fine-tuning approach that is currently highly popular for language models,[290] and is beginning to see adoption in chemistry.[108,291,292] In the community that develops fitting frameworks, database construction has been overlooked, although this is beginning to change, as in Ref. [279], where it is observed that active learning does not produce general-purpose data that work well for more than one fitting methodology. However, the use of more diverse, or 'difficult',

training data tends to degrade the absolute value of the numerical prediction errors of potentials as a side-effect of, in my view, more important improvements in robustness and in getting the right qualitative physical behaviour over a wide range of conditions.

The question of how to use physically-guided validation methods has been addressed in this chapter, and will form the basis of the ultimate evaluation of the quality of potentials in subsequent chapters as well. We find that using large-scale structures and the application of pressure are some useful strategies for this validation. Finally, when choosing which scientific questions to address with cutting-edge ML potentials, it is worth considering where there are advantages to be had over more established methods: DFT is still, at the time of writing, more transferable and reliable than the best fitting methodologies trained on even higher-level quantum-mechanical data. Large system sizes and long timescales, however, are completely inaccessible with DFT, hence disordered materials—which need large cells to represent and long melt-quench simulations to prepare—and rare defects are promising targets that form the basis for two case studies in the subsequent chapters.

3.10 Further methodological details

3.10.1 D1 database

The reference configurations for the indirect learning step were obtained with an MD simulation protocol of 50 ps thermalisation of a random initial structure at 500 K, followed by a constant pressure ramp to 20 GPa over 200 ps (as used

in Ref. [203]). In Fig. 3.1b, we study errors for compression MD with respect to quantity of compression training data, which we find to converge by around 2×10^5 atomic environments. Meanwhile, the errors for melt-quench MD start to *diverge* at this point (becoming worse with larger **D1**), suggesting that over-weighting environments encountered during compression MD comes at the expense of performance in other regions of configuration space. Based on these results, and our observations of unstable MD for compression datasets much greater than 2×10^5 environments, we found potentials trained on 2.4×10^5 environments from random starting points to be suitable and therefore use a database of this size for validation.

3.10.2 Test configurations

The 'RSS' (random structure searching) test set, the results for which are shown in Fig. 3.4, consists of 100 structures produced using the `buildcell` code of the Ab Initio Random Structure Searching (AIRSS) software[113,114] and optimised with **M1** at pressures in the range 0–32 GPa. The initial geometries of the random structures were constrained to have a minimum interatomic separation of 2 Å, a volume of 20(5) Å^3 per atom, 1–4 random symmetry operations, and 1–64 atoms in the unit cell. The corresponding `buildcell` settings were `NATOM=1-16, VARVOL=20, MINSEP=2.0, SYMMOPS=1-4`. Each structure was optimised at a random pressure taken from the (positive) exponential distribution centred at 0 GPa with width 5 GPa. See Ref. [114] for details of the AIRSS framework. The 'compression MD' set (Fig. 3.4) consists of structures from 64-atom compression runs of

Table 3.2: General hyperparameters used for SOAP-based structural analysis, carried out using the QUIP/GAP code as interfaced to the `quippy` Python package (see https://github.com/libAtoms/QUIP). SOAP vectors were configuration-averaged (`average=T` in `quippy`) for each MD frame before their transformation to the power spectrum and then kernel via a dot product with the reference crystal. The dot product was raised to a power of $\zeta = 4$.

σ_{at}	0.5 Å
R_{cut}	5.0 Å
l_{max}	6
n_{max}	10
`cutoff_transition_width`	2.0 Å
ζ	4

Table 3.3: General hyperparameters used for all MTPs described in this chapter. The default `radial_basis_size` for each value of $\mathrm{lev}_{\mathrm{max}}$ was used throughout, as was the case for any parameters not listed below.

`min_dist`	1.5 Å
`max_dist`	5.0 Å
`radial_basis_type`	`RBChebyshev`
`energy-weight`	1.0
`force-weight`	0.1
`stress-weight`	0.001
MTP version	1.1.0

LDA Si (snapshots taken between 5 and 20 GPa). Both test sets were evaluated in single-point DFT computations with settings corresponding to Ref. [45] (cf. https://doi.org/10.5281/zenodo.1250555), using the PW91 functional,[170] a 250 eV energy cut-off (with automatic finite basis set correction), 0.05 eV Fermi–Dirac smearing, and an on-the-fly pseudopotential, as implemented in `CASTEP`.[293]

Computational cost

Timing data (used to construct the horizontal axis in Fig. 3.4) were acquired by averaging across 9 MD runs of 5,000 timesteps with 10,000 atoms, parallelised across 48 CPU cores using LAMMPS. The protocol, initial structure, and random

Table 3.4: Absolute timing information for the error analysis in Fig. 3.4. Average timings and standard deviations (σ) over 9 runs are reported separately for student and direct MTPs. Timings are in timesteps (τ) per second. The value for GAP-18 simulations was $0.161(1)\,\tau\,\mathrm{s}^{-1}$. Calculations were performed using LAMMPS and parallelised over 48 cores.

$\mathrm{lev}_{\max}(L)$	$\tau\,s^{-1}$ (M'_{L})	σ	$\tau\,s^{-1}$ ($\mathrm{M}''_{\mathrm{L}}$)	σ
12	117.5	4.4	114.7	6.7
14	71.9	3.2	68.2	1.8
16	44.4	1.2	42.7	0.9
18	28.2	1.2	27.4	1.2
20	17.0	0.6	19.0	0.2
22	11.8	0.3	11.9	0.4
24	7.6	0.1	7.6	0.1
26	4.6	0.1	4.5	0.1

seed used were the same for each potential.

The 1-million atom compression simulation was parallelised across 768 CPU cores, requiring a total wall-time of 15.5 hours for 0.25 ns of simulation time: 11.9×10^3 CPU hours, 2.50×10^5 timesteps, 5.85×10^3 timesteps atom^{-1} core-second^{-1}, where forces are evaluated at each 1 fs timestep. The 1-million atom melt-quench simulation was parallelised across 384 CPU cores, requiring a total wall-time of 195.6 hours for 2.22 ns of simulation time: 75.1×10^3 CPU hours, 2.22×10^6 timesteps, 8.2×10^3 timesteps atom^{-1} core-second^{-1}.

3.10.3 MTP to MTP indirect learning

To further probe the teacher–student concept, we used MTPs with varying flexibility as **M1** *and* **M2**, i.e. with direct MTPs as teacher and indirect MTPs as student models. Figure 3.11 presents the success of the resulting potentials for the compression test of Fig. 3.7 in the main text. The results are somewhat mixed, but broadly only the more flexible MTPs are capable of acting as a satisfactory

M1 for smaller models. However, even M'_{24}, the best direct MTP produced in this chapter, leads to a problematic M''_{22} . When using GAP-18 as **M1**, all tested M''_L models performed acceptably.

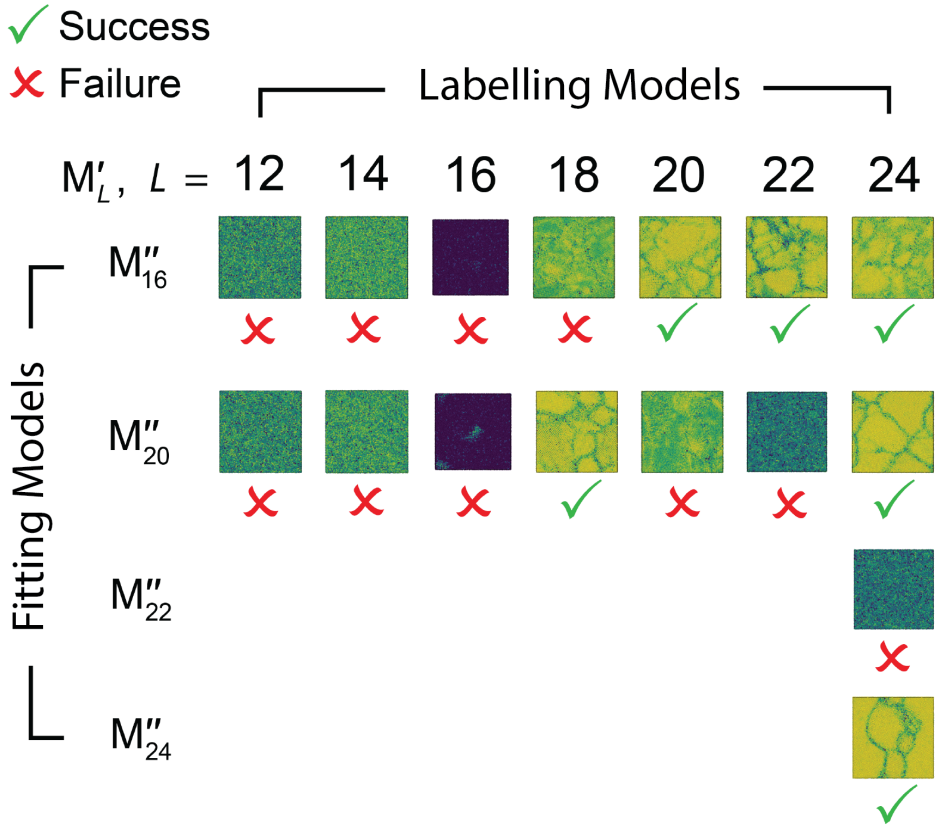


Figure 3.11: MTP to MTP indirect learning. Compression MD test for a set of 16 student MTPs trained on **D1** (rows corresponding to levels L), but with labels (energies and forces) calculated using direct MTPs of varying levels (columns). The test is the same as is carried out in Fig 3.7. We show the final MD frame, coloured according to similarity to simple hexagonal Si as in Fig. 3.7 to illustrate the degree of success/failure. We characterise a success (marked with a green tick under the structural snapshot) as the correct prediction of the polycrystalline simple-hexagonal phase.

4 | Understanding defects in a-Si with million-atom structures

4.1 Acknowledgements

The work described in this chapter has been published as Ref. [3], from which any figures labelled as 'adapted' have been drawn. I am grateful to Prof. Stephen Elliott (Visiting Professor, University of Oxford) for insights on medium-range order in a-Si and for discussions on the implications of our simulation results in the context of the experimental evidence for the structure of a-Si. I thank Prof. David Drabold and Chinonso Ephrahaim (University of Ohio) for performing electronic Density of States calculations on our large structures and for insights on the electronic nature of defects throughout the writing of the article. I depended on the Ovito python infrastructure for efficient geometric calculations on large datasets, which I strongly recommend.[92] Details of the code used for plotting and analysis are available at https://github.com/MorrowChem/understanding_defects/releases/tag/v0.1.0.

4.2 Introduction

The student MTPs described in the previous chapter achieve around a 1,000-fold acceleration in the computational time required to perform MD simulations compared to GAP-18, whilst maintaining effectively the same quality in their description of the PES. Our access to longer lengthscales and timescales can be put to good use in the study of defects in amorphous silicon (a-Si), which are embedded at around the 1 % level, hence require extremely large structural models to study systematically at realistic concentrations.

Amorphous silicon (a-Si) is perhaps *the* textbook example of a disordered material. Its structure resembles Zachariasen's idea of a 'continuous random network' (CRN) of covalently bound atoms.[55,68,294] Amorphous silicon has been the subject of practical and fundamental interest for decades, most notably as the most popular photovoltaic material,[295] but also for the hyperuniform nature of its disordered network,[253] the transition between silicon's highly-coordinated metallic liquid and the fourfold-connected structure,[296] its complex phase behaviour under pressure,[203,278,297] and a tension-compression asymmetry in its mechanical properties.[298]

While the structure of a-Si is dominated by fourfold-connected atoms ($N = 4$, where N denotes the number of bonded nearest neighbours for an atom), there is particular interest in 'defective' atoms: those that have fewer ($N = 3$) or more ($N = 5$) neighbours.[67,299–302] There is experimental evidence—for example, from spectroscopy—for the presence of point defects in a-Si, such that its structure is likely more complex than an idealised all-fourfold CRN description.[66,67,300,303–306] In well-relaxed structural models of a-Si, N is determined by a clear minimum in the radial distribution function that separates the first (bonded) peak from the second (non-bonded) one at about 2.85 Å. Defining neighbours, and therefore coordination defects, is straightforward if the minimum in the distribution approaches zero; it becomes more ambiguous otherwise, which is most relevant to $N = 5$ atoms.

a-Si has been a prominent example, since the early days of computation, of the important role for simulations in studying amorphous materials.[307] In the 1980s,

models of a-Si were created via melt-quench MD with the empirical Stillinger-Weber potential,[284,308] which has a functional form based on bond lengths and penalising bond angles that stray from tetrahedral:

$$E = \sum_i \sum_{j>i} \phi_2(r_{ij}) + \sum_i \sum_{j\neq i} \sum_{k>j} \phi_3(r_{ij}, r_{ik}, \theta_{ijk}), \tag{4.1}$$

where ϕ_2 is a two-body term and ϕ_3 is a three-body term.[i] The large concentration of coordination defects in the models obtained this way ($\approx 20\%$) permitted a discussion of the average structure of $N = 5$ defects, as well as an analysis of the energetics predicted by the potential. Later studies, with increasingly fast computers, extended system sizes to hundreds of thousands of atoms with empirical potentials.[284,311,312] Equally, DFT based MD studies on much smaller systems have provided important insights.[313–316] Nevertheless, predictive DFT simulations of a-Si beyond the few-nm length scale remain out of reach, due to the cubic scaling of the computational cost of DFT with the number of electrons (i.e. system size). As an illustration, a recent 'optimal' DFT-MD-based a-Si structure contains 215 atoms,[315] and the limit, as of writing, for such types of MD simulations appears to be about 1,000 atoms.[316] There remains debate also over the relative abundance of $N = 3$ and $N = 5$ defects, with the spectroscopic evidence ambiguous[299,317] and computational methodologies, such as *ab initio* MD[313] and empirical potentials,[206]

[i]The full expressions for ϕ_2 and ϕ_3 are:

$$\begin{aligned} \phi_2(r_{ij}) &= A_{ij}\epsilon_{ij}\left[B_{ij}\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{p_{ij}} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{q_{ij}}\right]\exp\left(\frac{\sigma_{ij}}{r_{ij}-a_{ij}\sigma_{ij}}\right) \\ \phi_3(r_{ij}, r_{ik}, \theta_{ijk}) &= \lambda_{ijk}\epsilon_{ijk}\left[\cos\theta_{ijk} - \cos\theta_{0ijk}\right]^2 \exp\left(\frac{\gamma_{ij}\sigma_{ij}}{r_{ij}-a_{ij}\sigma_{ij}}\right)\exp\left(\frac{\gamma_{ik}\sigma_{ik}}{r_{ik}-a_{ik}\sigma_{ik}}\right) \end{aligned}, \tag{4.2}$$

where parameters were originally fitted to crystalline and liquid phases of silicon,[205] but many modifications exist in the literature, e.g. tailored to the a-Si phase,[206] or more chemically complex semiconductors such as the III-V materials[309] and even MoS_2.[310]

tending to predict more 5-fold defects. Integrating the experimental $G(r)$ out to the first minimum yields an average coordination number of 3.8, which may seem to suggest a prevalence of 3-fold defects.[66] However, the existence of nanovoids within experimental samples has been posited as an alternative explanation, and the agreement between the first $G(r)$ peaks of our structural model (Fig. 3.9), and previous models,[14,313] suggest that discrepancies in densities could indeed be the cause.[318,319] In any case, an average coordination number of 3.8 implies a 12% density of dangling bonds, which is much too high compared to the couple of percent (at most) figure measured by hydrogen solubility and electron paramagnetic resonance.[320–322]

The previous chapter introduced recent developments in ML potentials which have made it possible to prepare realistic structural models of a-Si with a quality that rivals DFT-based simulations, but with a much larger size:[14,203,281,323] our million-atom model reached a cell length of 27 nm.[1] Perhaps the most aggressive assumption made with most current ML potentials is that the total energy of a system of atoms (which is a quantum-mechanical observable) can be decomposed into a sum of local atomic energies (which are not observables).[8,12] The upshot of making this locality approximation is the ability to 'machine-learn' energetics that are independent of the system size, which allows a linear scaling of computational cost with the number of simulated atoms. Although initially this purpose alone motivated such a decomposition, there is increasing evidence that local energies have a physical relevance that can be useful for chemical interpretation aside from the computational efficiency they provide.[281,323] We have interrogated the

robustness of local energy predictions, with a-Si as a specific example, in Ref. [5] (summarised in Section 2.3.3) and shown that the consistency of local energy predictions across different ML models and training datasets can be quantified. For other systems, it has been shown that ML atomic energies in amorphous graphene can be used both to drive structural exploration and to analyse the resulting structures,[324] and that the density of ML atomic-energy contributions can yield insights into solid-state ion conductors.[325] The local energy contributions assigned by neural-network potentials have been explored as well.[326–328] The idea of considering local contributions to the total energy in silicon has already been pioneered based on empirical potentials,[308] and our ML-based approach in this chapter extends this type of thinking and places it on a more quantitative, DFT-accurate basis.

We identify three structural prototypes for over-coordinated atoms in a-Si, enabling us to define a 'taxonomy' of defects in this canonical disordered material. Further, we uncover and explain a tendency for fivefold defects to aggregate by means of strain that they induce on their fourfold atomic neighbours. Million-atom simulation cells, which contain tens of thousands of examples of threefold- and fivefold-connected defective atoms, are necessary to study defects that occur at the few-percent level in a systematic way, and, crucially, to investigate 'second-order' processes, such as the interaction of defects with one another that leads to aggregation. Beyond silicon, the research of this chapter establishes a general protocol for routine, quantum-quality, million-atom scale, ML-driven simulations of rare events in functional materials.

4.3 Results

In Fig. 4.1a, we sketch the main coordination defects in a-Si and their first coordination shells. Figure 4.1b presents the ML atomic energies in the 1M-atom model from Ref. [1], separated by coordination number for the central atom. It confirms earlier findings that $N = 3$ atoms have decidedly higher energies on average than $N = 5$ ones.[281] However, the distributions in Fig 4.1b result from a much larger dataset than in Ref. [281] and do not include any histogram broadening.

Figure 4.1c contains analysis of the defects from a somewhat orthogonal, purely structural, perspective by binning the distances to the individual neighbours of each atom in the structure: i.e. for atom i, we find its nearest neighbour and add this to distribution 1 (purple), then we find its second-nearest and add this to distribution 2 (blue), and so on up to $N \in \{3, 4, 5\}$ for atom i, which we subsequently repeat for all atoms in the structure. Note that all these neighbours are directly bonded to the central atom ($d < 2.85\,\text{Å}$), so are all in the first *coordination shell*. As the coordination number of the central atom increases from 3 to 5, so too do the distances between bonded atoms (the distributions shift to the right), as is typical when considering any bonding centre in coordination chemistry. The well-defined 5^{th} peak, and the even spacing between peaks for the 5-fold distribution, indicate that a majority of 5-fold defects are best described as truly 5-coordinated, rather than, alternatively, as [4+1] with a mostly non-bonded 5^{th} neighbour. [4+1] defects are what is observed in the cogener of a-Si: tetrahedral amorphous carbon.[330] There is, however, a longer tail observed for the $n = 5$

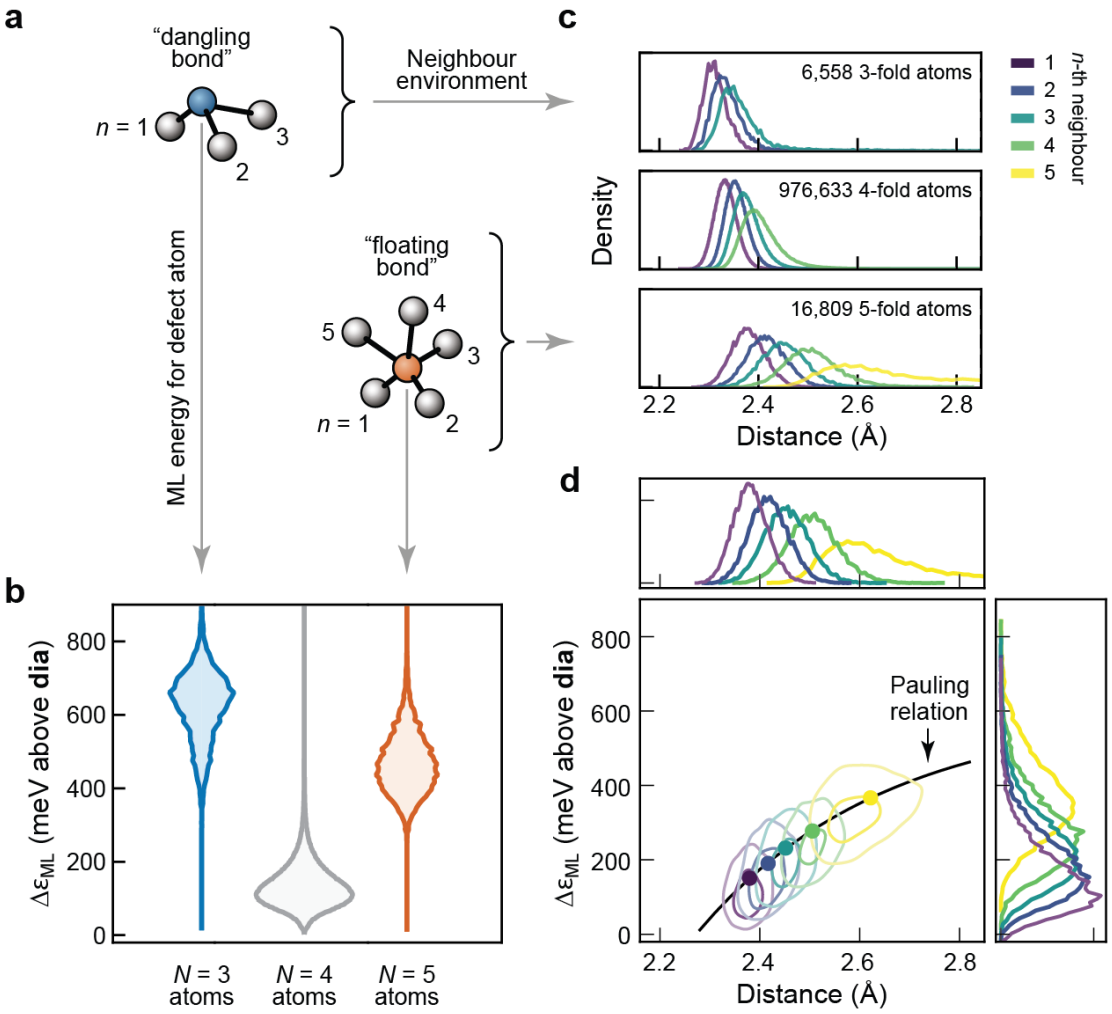


Figure 4.1: Defects in amorphous silicon from million-atom simulations. (**a**) Schematic drawing of 3- and 5-fold-connected atoms, with neighbouring atoms numbered in order of their distance from the central atom. (**b**) Distributions of the ML-predicted atomic energies, $\varepsilon_{\mathrm{ML}}$, of defects in the million-atom model, shown separately for different nearest-neighbour coordination numbers, N. Values are referenced to crystalline diamond-type silicon, which is set as $E = 0$. (**c**) Neighbour distributions, resolved on separate axes according to 3-, 4-, and 5-fold-connected central atoms and their individual immediate neighbours, sorted by distance within the first coordination shell. (**d**) 2D correlation plot of the neighbour density for 5-coordinated defects (as in panel c) versus energy (as in panel b), given separately for the immediate neighbours (n = 1–5, purple to yellow). The black line describes a fit to a Pauling-like relation.[329] Further details on the fitting function and choice of parameters are given in Fig. 4.2. Adapted from Ref. [3].

distribution (yellow on the lowest plot in Fig. 4.1c), which contains those more distant 5$^{\text{th}}$ neighbours that are indeed in this minor group of [4+1] environments.

In Figure 4.1d, we examine the correlation between the structural and energetic characteristics of 5-fold defects: via the distance of the neighbours from the central atom on the x-axis and the local energy of the neighbours of the 5-fold atom (not the 5-folds themselves!) on the y-axis, respectively. Notably, we find a logarithmic relationship between these two quantities: the elevated local energy at 5-fold defect centres (previously noticed for empirical and ML potentials[14,308]) is also delocalised onto the surrounding atoms with longer bonds to the defect. On reflection, it is comforting that ML local energies, the physical meaning of which has been a source of debate, reproduce a well-founded result from valence theory: the exponential dependence of bond strength on bond length. The equivalent correlation plots for $N = 3$ and $N = 4$ atoms (Fig A.4 in further methods Section A.4) show a much tighter distribution of energies for their respective atomic neighbours. We used a modified version of Pauling's relation between atomic radii and interatomic distances in metals to obtain the line in Fig. 4.1d.[329] The modified equation to produce these fits is

$$\log n(E) = A(r_{\text{c}} - r) \ , \tag{4.3}$$

where $n(E)$ is the bond strength defined in terms of neighbour local energies as

$$\frac{E_{\text{max}} - E}{E_{\text{max}} - E_{\text{min}}} \ , \tag{4.4}$$

$r_{\text{c}} = 2.38\,\text{Å}$ is a characteristic radius (taken to be the mean minimum bond length per atom in the structure), r is the bond distance, and A is a fitting parameter

with optimal value 1.17 derived from linear regression on all 5-folds in the structure. In the definition of bond strength, we use approximately the highest and lowest energy neighbours of 5-folds (99^{th} percentile) to transform local energies to a dimensionless bond-strength measure between 0 (no bond) and 1 (strongest bond). In Fig. 4.2a, we use $E_{\text{max}} = 600\,\text{meV}$ and $E_{\text{min}} = 230\,\text{meV}$. The quality of the fit is not sensitive to the precise values used for E_{min}, as is demonstrated in Fig. 4.2b. The sum of the mean bond strengths for the five bonds to a fivefold atom is 3.97 using this definition, which suggests that the additional bonds do not make up for their increased weakness. Directly fitting to the bond-energy data, rather than to a bond strength measure, with an additional parameter, B, gives a similar quality of fit using the equation $\log(E) = A(r_{\text{c}} - r) + B$.

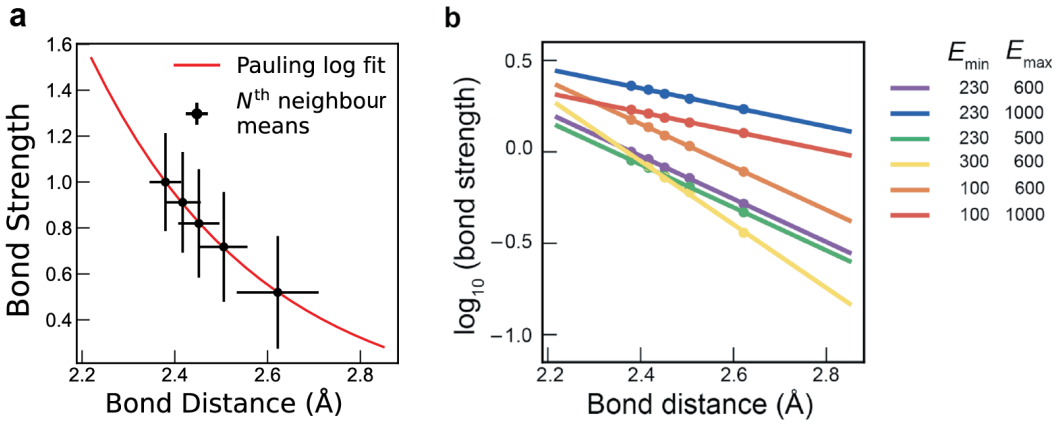


Figure 4.2: Logarithmic dependence of local-energy-derived bond strength on bond distance. (**a**) The mean energy (circle) and standard deviation (bars) of each of the N^{th} neighbours in Fig. 4.1d is fitted to equation 4.3. (**b**) We show fitted equations for a range of energy values for normalisation: E_{min} over 100–300 meV and E_{max} over 500–1000 meV.

4.3.1 Revision of the floating-bond model

The presence of mostly relatively short fifth-neighbour contacts at distances below about 2.6 Å (Fig. 4.1c) begs the question of whether there is a 'typical' geometric

structure of 5-fold defects in a-Si. Debate over this has a long history in the field. The term 'floating bonds' was originally used to describe defects consisting of silicon atoms with five bonded neighbours, introduced by Pantelides in 1986 as part of the early recognition of the importance of over-coordinated atoms in a-Si.[299] The canonical structure of such a defect was described as a perfect tetrahedron with an extra fifth atom bonded directly opposite to one of the existing (previously tetrahedral) bonds. However, Pantelides noted himself that, in the amorphous phase, bond distances and angles can vary considerably from such an idealised geometry, which is also supported by the MAS-NMR data for ^{19}Si.[57]

In Fig. 4.3, we address this question using the common prototypes for 5-coordinated atoms known from structural inorganic chemistry. On the left of Fig. 4.3a, we sketch the trigonal bipyramidal (TBP) and square pyramidal environments predicted by the valence shell electron-pair repulsion (VSEPR) model[332,333] (familiar from undergraduate chemistry textbooks), and also the idealised floating-bond environment, with a fifth bond directly opposite one of the bonds in a tetrahedron (Fig. 4.3a).[299] We took the 5-fold defects in the 1M-atom structural model and evaluated their structural similarity to the three respective prototypes on a scale of 0 (dissimilar) to 1 (identical) via the SOAP kernel,[13] as detailed in Section 4.7 (further methodology). In this analysis, because we aim to discriminate between environments based on their bond angles and their respective similarity to the prototypical VSEPR configurations, we rescaled all nearest-neighbour distances to 2.5 Å, such that we examine only the angular (and higher order many-body) distribution of atoms around each defect, independent of

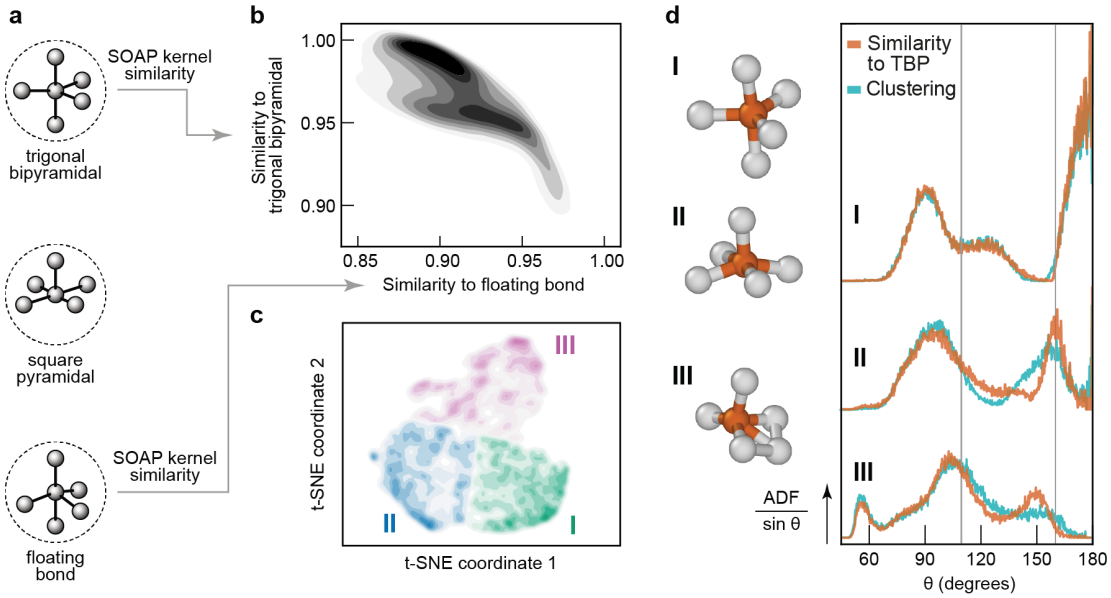


Figure 4.3: Three categories of fivefold-connected defects. (**a**) Schematic drawing of idealised trigonal bipyramidal (TBP), square pyramidal, and 'floating-bond' environments. (**b**) 2D plot of the SOAP kernel similarity of $N = 5$ atoms in the a-Si model to the idealised TBP and floating-bond environments, respectively. The distribution of values for individual atoms is shown as a heat map. (**c**) Unsupervised classification of 5-fold atoms. The full distance matrix, $\mathbf{D} = \sqrt{2 - 2\mathbf{K}_{\mathrm{SOAP}}}$, is embedded in 2D with the dimensionality reduction algorithm t-SNE,[239] where $\mathbf{K}$ is the kernel matrix built from the similarity of each 5-fold atom with every other 5-fold atom. The bisecting k-means algorithm is used to identify clusters I–III, which we use to colour the point densities.[331] (**d**) Bond-angle distribution function (BADF), scaled by $\sin\theta$, for atomic triples centred on all 5-fold coordinated atoms. The BADFs are plotted separately for the three distinct categories of $N = 5$ defects related to idealised structures as illustrated with selected examples of such configurations from the a-Si model alongside. The 5-fold atoms are separated into these categories via two methods but with similar results: by comparison to the idealised structures of panels a and b via SOAP similarity (orange) and from unsupervised clustering (cyan). Vertical lines at the tetrahedral angle (109.5°) and at 160° are guides for the eye. Adapted from Ref. [3].

distance variations. What results is a multimodal distribution of similarity values, as displayed in Fig. 4.3b. We arrive at a classification of the 5-fold defects in a-Si by partitioning this distribution into three regions: as TBP-like (I), square-pyramidal-like (II), or "floating bond"-like (III). A very similar classification is obtained using unsupervised ML[334,335] in Fig. 4.3c via dimensionality reduction followed by clustering. In this plot, the distance between points corresponds to

their structural dissimilarity as measured by SOAP. The three-lobed distribution, with higher densities of points at the outer edges, supports the identification of exactly three major classes of 5-fold atoms. Alternative presentations of the same data are included in Fig. A.3 at the end of the chapter.

To aid the structural interpretation of the SOAP results, we show bond-angle distribution functions (BADFs) and accompanying representative example structures taken from the model in Fig. 4.3d. Note that the BADFs in a-Si have been closely linked to the Raman transverse-optic peak width,[336] to the exponential tails in optical-absorption band edges,[337] and to the through-bond (topological) distance to over- and under-coordinated atoms,[338] so the angular structure of these defects has broader physical significance.[339] Our million-atom simulations allow us to derive finely resolved BADF plots, which immediately reveal that the categories have distinctly different shapes to one another. The BADF for category I in Figure 2c directly reflects the angles in an idealised TBP environment: 180° between the axial atoms, 120° between equatorial atoms, and 90° between axial and equatorial atoms. The square-pyramidal BADF (category II) has a fairly sharp distribution of angles at 160° and a broader peak at ≈ 100° angles. Our designation 'floating bond' (III), maximally dissimilar from the TBP (I), is less well-defined structurally, but can be understood as originating from an ideal tetrahedron with an additional 5$^{\text{th}}$ atom approaching a face or edge, similar to what was suggested by Pantelides. However, fewer than half of all $N = 5$ defects adopt a structure similar to the floating-bond description. In the BADF, this arrangement manifests in the following features:

1. The occurrence of $\approx 60°$ angles between the 5^{th} atom and those neighbours that it approaches most closely;

2. A corresponding closing of the ideal tetrahedral angle as these neighbours are displaced away from the 5^{th} atom; and

3. Larger angles between the 5-th atom and those on the opposite side of the tetrahedron.

The boundaries between the classes of 5-fold defects are blurred to a degree, which suggests an overlap of types of defect structures, rather than three entirely separate categories. We believe this is to be expected for such a strongly disordered amorphous structure, and this observation is consistent with the low energy barriers between structural minima that are observed for 5-fold complexes in molecular inorganic chemistry, which often exhibit fluxionality.[340–342] Further discussion of where we draw the boundaries between categories, along with partial pair correlation functions for defects, can be found in Figs. A.1 and 4.7 in the Appendix.

4.3.2 Locality of defect environments

The correlation between bond distance and local energy in Fig. 4.1f implies that the structural disturbance associated with 5-fold defects extends over a longer range than the radius of the defect centre itself. To what extent does this extend even further? Figure 4.4 provides a more detailed analysis of the atomic energies and their degree of 'locality'—that is, their dependence on nearest-, next-nearest-,

and further neighbours. The locality of physical properties is of general interest for the development of atomistic ML models, because it limits the information that can be captured by finite-range models.[100] To systematise our discussion, we introduce some notation to describe the topological neighbourhood of defective atoms, as sketched schematically in Fig. 4.4a: 4-fold coordinated atoms that are directly bonded to a defect centre (i.e., to an under- or over-coordinated atom) are referred to as $4'$; 4-coordinate atoms directly bonded to a $4'$ atom, but not to a defect centre, are referred to as $4''$; and those bonded to a $4''$ atom, but not to any closer defect, are referred to as $4'''$. Atoms which are more topologically distant than the third neighbour shell, $4'''$, we designate as 'bulk'.

Fig. 4.4b demonstrates how, upon moving away from an $N = 3$ defect atom, the atomic-energy distributions quickly approach those of the bulk; the directly adjacent atoms of a 3-fold atom, $4'$, have a distribution that is similar to that of the bulk atoms (shown in light grey). In other words, an $N = 3$ atom does not appear to significantly affect the atomic energies of its bonded neighbours. In contrast, Fig. 4.4c shows that the $4'$ atoms connected to a $N = 5$ defect are much higher in energy on average. This suggests that the disturbance induced by 5-fold defects is less localised, so has a longer-range effect. In both cases, the energy distributions for the $4''$ atoms and beyond are not significantly affected by the presence of nearby defects. There is a parallel here with the (widely discussed) electronic structure of defects in a-Si:[299,313,337] the electronic states associated with 5-fold defects are thought to be more delocalised than those of 3-fold defects, for which the latter can be conceptualised as an unpaired electron occupying an sp^3 hybrid.

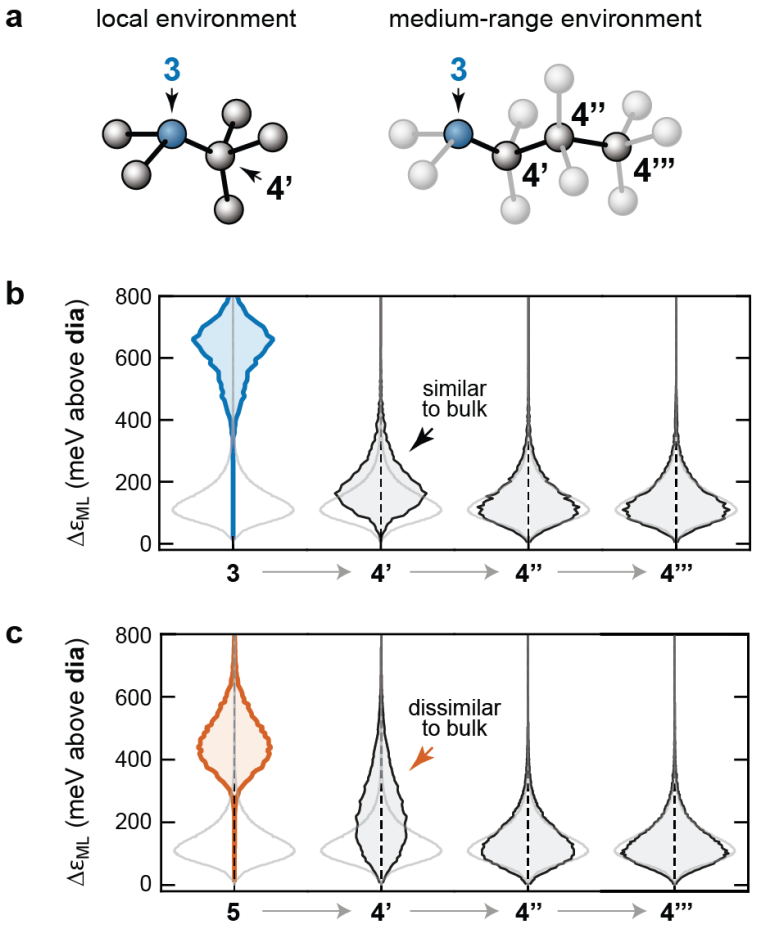


Figure 4.4: Locality of defect environments. (**a**) Schematic of the labelling scheme for local and medium-range environments, based on the bond topology. A fourfold atom directly connected to a defect (here, to a 3-coordinate atom) is labelled as $4'$, a fourfold atom connected to that one is labelled as $4''$, and so on. (**b**) Distribution of ML local energies for $N = 3$ atoms and their surroundings. The distribution for bulk a-Si is shown in light grey. (**c**) Same but for $N = 5$ atoms and their surroundings. Adapted from Ref. [3].

As with the connection to valence implied by the Pauling relation of Eq. 4.3, it is encouraging that the ML local energies appear to contain implicit physics despite their formulation which is 'without the electrons'.[12] Furthermore, the difference in the locality of the 5-fold and 3-fold defects has an interesting consequence: when the environment of each defect is considered, the locally averaged energy for 3-fold and 5-fold defects is 726 meV and 749 meV above the bulk, respectively, which

indicates that 3-fold defects are predicted to have a marginally *lower* formation energy, in contrast to what was found previously for the energies of just the defect atoms themselves, and apparently confounding expectations based on the observed relative abundance of defect types.[281] Note that these sums take into account the clustering of 5-fold defects described in Fig. 4.6, which reduces the number of $4'$ atoms per 5-fold defect.

Why, then, are more 5-fold defects observed in MD, despite their higher energy? This seemingly contradictory result can be rationalised as a consequence of either entropy or kinetics. The 5-fold centres have considerably more flexibility than the relatively rigid pyramidal structure of 3-fold defects, which implies 5-fold defects may have higher entropy and hence a lower free energy of formation. 5-fold defects are more similar to the structure of molten Si than are 3-folds, so are likely to be more kinetically accessible during quenching. The difference in the local strain amongst 5-folds and 3-folds also leads us to advise care in interpreting a single local energy in the context of defects, as highlighted previously in the context of amorphous graphene.[324] Local averages, which include physical strain induced by the defect and smooth out the assignment of local energies, appear to be more appropriate in drawing physical interpretations. After all, in the limit of averaging over the entire structure, naturally the quantum-mechanically-observable total energy is recovered (within the error of the ML prediction).

4.4 Connection with defects in crystals

The structural and energetic landscapes of defects, even for simple crystalline materials, can be deceptively complex, and understanding those defects requires advanced computational techniques. For example, it was shown recently that relaxations from ideal or symmetric small-scale defect models often determine incorrect (only metastable) defect geometries, and that a more comprehensive exploration of the associated structural and energetic landscape (e.g. implemented in the 'ShakeNBreak' code[343]) is required even for seemingly simple inorganic crystals.[344–346] ML techniques have begun to be applied to defects in crystalline materials in addition to these DFT-based methods.[347–349] Here, we probe the connection between defects in crystalline silicon, which are well-studied, and the atomic environments of defects in the amorphous phase.

It has previously been shown that 2D scatter plots of structural properties (such as 'diamond-likeness') and atomic energies are useful for understanding defects in a-Si.[281] We extend this methodology by using it to compare defects in the amorphous form of silicon to different types of defects in the crystalline phase and present the results in Fig. 4.5. What we find is broadly intuitive: the 3-fold defects are most similar to vacancies (blue in Fig. 4.5a), 5-fold defects are most similar to crystalline interstitials (orange in Fig. 4.5c), and neither defect is similar to the perfect crystal, as measured by the SOAP-kernel on the horizontal axis. The bimodal distribution in the similarity of 5-fold defects to vacancies suggests that a minority of such defects, likely those with two long bonds, are related to vacancies

and, in fact, there is an ambiguity of classification here. The crystalline interstitial example here was produced by equilibrating an idealised (10-fold coordinated) interstitial site in diamond-type Si at 500 K. These simulations show that the structure of the crystalline interstitial, like the amorphous 5-fold defect, is highly fluxional and therefore is difficult to understand via even several typical structures. Fluxionality of 5-coordinate complexes is common in molecular systems because of the similar energies of different conformations, so it is perhaps unsurprising to observe a similar effect in an extended solid. This observation highlights the advantage of having many examples of individual defects in the million-atom model available for comparison, as we have discussed in the context of Fig. 4.4. This can be thought of as a form of 'spatial averaging', which is easily parallelised, and hence amortises the cost of an equivalent (extremely long) time-averaging performed on a small system size (which cannot be parallelised). Our analysis in Fig. 4.5 extends the long-standing assumption that amorphous materials contain structural building blocks of the corresponding crystalline phases.[55,350] We therefore suggest that amorphous phases can be thought of as containing building blocks, not only of crystalline phases, but also of defects in crystals.

4.5 Why do 5-fold defects cluster?

Although the majority of over-coordinated defects in a-Si exist as isolated $N = 5$ atoms, we find a tendency for those atoms to occur close together beyond that which would be expected for a random placement of defects in a CRN—i.e. the clustering level that would be observed if 5-fold defects were placed randomly

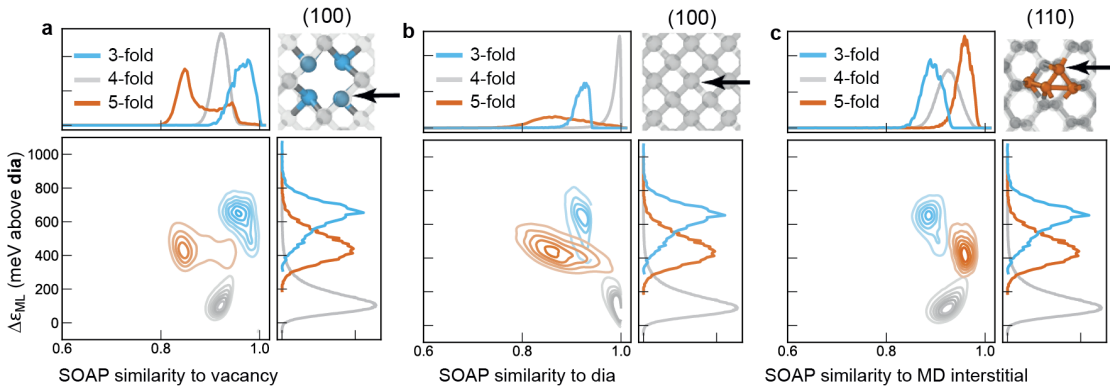


Figure 4.5: Connection with defects in crystals. (**a**) A 2D SOAP similarity–energy map for the structure obtained at a quench rate of $10^{11}\,\mathrm{K\,s^{-1}}$, separately for 3-fold, 4-fold, and 5-fold coordinated atoms. The horizontal axis shows the structural similarity to the relaxed vacancy. The $N = 3$ atoms (blue) are structurally the most similar to the vacancy, the $N = 5$ atoms (orange) the least so. The vertical axis gives the ML atomic energy. The structure is visualised in the top right part, with the viewing direction given, and the arrow indicating the atom used for comparison. (**b**) Same for the perfect crystalline diamond structure, as in Ref. [281]. (**c**) Same for a snapshot from an MD simulation of an interstitial in the crystalline structure. Adapted from Ref. [3].

without the influence of energetics. The energy distributions in Fig. 4.6, evaluated for defects and their local environments up to $4'''$ sites, show that such clustering reduces the total average energy associated collectively with defect atoms and their immediate neighbourhoods. This effect is largely due to clustered defects having fewer nearest neighbours per 5-fold defect—recall that the neighbours of defects, on average, have an elevated energy compared to bulk a-Si, as we showed in Fig 4.4. To avoid the extensivity of the excess energy with surface area when referenced to the ground-state crystal energy, we reference the energies in Fig. 4.6a to the average value for bulk a-Si. In this way, the neighbours of a defect will only increase the energy of the defect cluster if they have an average energy above that of bulk a-Si. Hence, we can converge the prediction of the cluster excess energy, so that it becomes independent of the number of neighbours considered, by including neighbours that are sufficiently distant from the defect centre as to be typical of

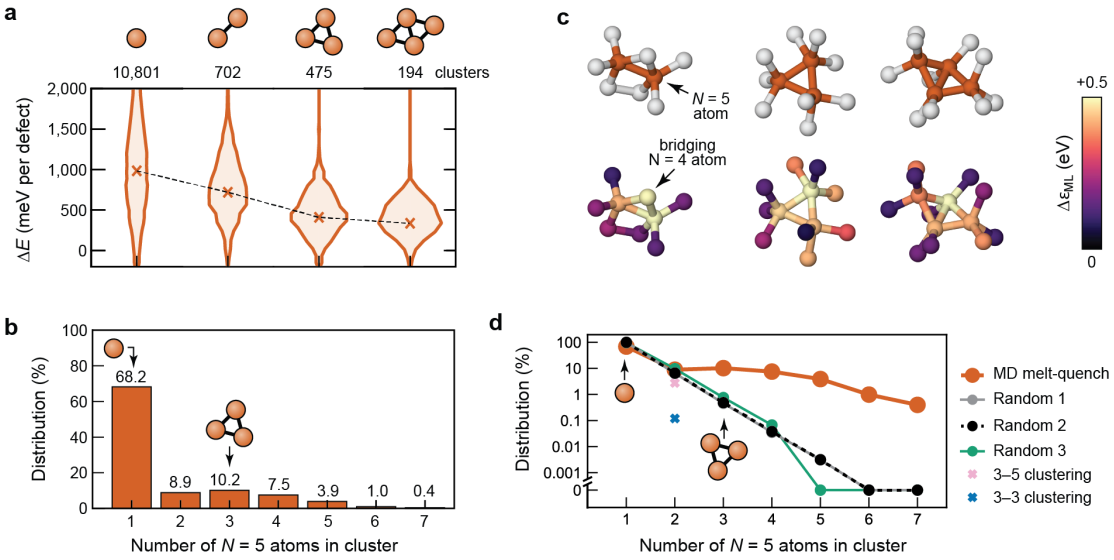


Figure 4.6: Clustering of 5-fold defects in amorphous silicon. (**a**) ML-predicted energy distributions for the most common 5-fold defect clusters, with the number of occurrences of each structure directly above each histogram. Defect cluster energies are calculated by summing the individual atomic energies of defect cores and their immediate topological neighbours up to 3 bonds away, relative to the mean a-Si energy, and are reported per-coordination defect. (**b**) Statistics for the number of occurrences of clustered 5-coordination defects. (**c**) Examples of clustered $N = 5$ defects of different sizes, colour-coded by coordination numbers of the atoms (left) and by their ML local energy above diamond-type Si, $\Delta\epsilon_{\mathrm{ML}}$ (right). Note the high atomic energy of the bridging $N = 4$ atom in the top part of the panel, indicated by an arrow. (**d**) Comparison of observed clustering of 5-fold defects in the MD-derived a-Si model (orange, same data as in b) with three different idealizations of a random clustering distribution and the occurrence of 3-fold–5-fold bonded pairs (pink cross) and 3-fold–3-fold pairs (blue cross). Protocol 1 (grey) randomly labels 1.68% of atoms (the 5-fold defect concentration) in a perfect diamond structure as 'defective'. Protocol 2 (black) randomly labels 1.68% of atoms on the 4-fold network of the a-Si model as 'defective'. Protocol 3 employs a bond-switching algorithm to equilibrate the MD-derived clustering distribution by moving 5-fold centres randomly across the graph formed by atoms (nodes) and bonds (edges). This accounts for an increased likelihood of clustering purely due to the 5 bonds at 5-fold-coordinate atoms compared to the 4 bonds for most atoms. More details of the idealisations are given in Section 4.7. Adapted from Ref. [3].

the bulk. In practice, this is achieved by including the 4‴ atoms via the sum

$$\Delta E = \sum_i (E_i - \bar{E})/n_5 \ , \tag{4.5}$$

where i indicates the topological classification and runs over 5, 4′, 4″, and 4‴; and n_5 is the number of 5-fold atoms in the cluster. The widths of the distributions in Figure 5a get narrower with cluster size because larger clusters have a greater total number of 4′, 4″ etc. neighbours (although fewer per 5-fold atom). The relationship between probability and cluster size is approximately exponential (Fig. 4.6b), with the exception of clusters of size 2, which are disfavoured compared to 3-membered clusters. 3-fold defects are far less likely to cluster than 5-fold defects: only 19 examples of 3–3 bonds were found and none were found containing more than one 3–3 bond. A possible qualitative explanation for this difference is provided by the examples in Fig. 4.6c: in the first case, a pair of $N = 5$ atoms is connected via a bridging $N = 4$ atom, which in itself has a rather high atomic energy (highlighted by an arrow), consistent with the strain formed in the associated three-membered ring. In the larger fragments shown in Fig. 4.6c, there are still three-membered rings, but they are now formed by three and four $N = 5$ atoms clustering together, respectively. In these cases, the directly-connected $N = 4$ atoms have lower energies. The clustering of 5-coordinate defects can be interpreted as causing a reduction in the defective 'surface area', which leads to aggregation with similar physics to, say, surfactant bubbles coalescing.

4.6 Outlook

Our study has more general implications for materials modelling beyond silicon. Defects and impurities in solids occur at a concentration of only a couple of percent at most, yet they are highly consequential for the performance of functional materials. The structure and chemical bonding of these defects likely differ markedly from the bulk, which requires a quantum-mechanically accurate description of the PES associated with defect formation, ruling out empirical potentials as reliable parametrisations of the PES. We showcased how ML interatomic potentials can be used to study rare events in a systematic way, and how ML-predicted atomic energies can explain the stability of defect environments in one of the canonical amorphous materials: a-Si. Qualitatively, this means a step away from idealised CRN models or small-scale simulation systems which necessarily can contain only a handful of defects, which moves us towards a fully realistic description of the amorphous state and device-scale models for applications of materials based on them. In the future, I envisage similar ML-driven studies of defects and defect complexes being performed for a wide range of functional materials, with returns to scale coming in the form of observations of interesting unexpected physics—similar to the extended influence of 5-fold defects in Si resulting in a surface tension effect.

4.7 Further methodological details

4.7.1 Randomising cluster distributions

The clustering distributions of the following models are included in Fig. 4.6 to illustrate the increased probability of 5-folds clustering.

Random protocol 1: 16,809 atoms in a 1M-atom perfect diamond structure (the same defect concentration as in the MD structure) were labelled as 'defects', without disturbing their positions or bonding topology.

Random protocol 2: 16,809 random atoms in the amorphous 1M-atom structure were relabelled as 'defects' using the existing bonding topology defined by a 2.85 Å cutoff. The majority of these are, statistically, 4-fold connected atoms.

Random protocol 3: a bond creation/deletion procedure takes the original clustering distribution of the MD-derived model and randomises it by walking 5-fold defects across the graph initialised by the bonds in the structure (2.85 Å cutoff). The atomic positions are ignored, with only the bonding topology modified at each step as follows: a 5-fold connected atom is selected at random for movement (atom i). A 4-fold connected target is selected at random to become the new 5-fold one under the constraint that it must be exactly three bonds away from atom i (atom l). Atoms j and k are selected randomly to form a path from atom i to atom l under the constraint that both must not be neighbours of either i or l. The bond i–j is deleted and bond j–l is formed, which moves the 5-fold across the graph by three bonds. This procedure can be repeated until the cluster distribution is equilibrated, giving a random distribution in green in Fig. 4.6 after 1.65×10^6 moves (approx. 100 moves per 5-fold centre).

4.7.2 Calculation of defect formation energy

A rough estimate of the energy of a 5-fold defect, including its environment, can be made using the mean values of distributions in Fig. 4.4, after subtracting the bulk mean energy, and counting the approximate number of neighbours (neglecting the possibility of shared neighbours), which is itemised in Table 4.1. A similar

Table 4.1: Calculation of defect formation energy.

Species	**Number**	**E (meV)**
5-fold	1	331
$4'$	5	104
$4''$	15	7
$4'''$	45	0.07
Total		964

calculation for 3-fold defects gives 772 meV. Summations of the local-energy data using the following pair of methods are consistent with this estimate. Figure 4.6 includes 5-fold defects among the 10,801 'isolated' cases provided there are no direct bonds between two 5-folds. We then construct an isolated 5-fold atom's environment by searching for all surrounding $4'$ atoms whilst ignoring any other defects, then $4''$s, and finally $4'''$s. This encompasses cases such as a bonded 5–$4'$–5 triple, where the second 5-fold would be ignored. The average energy of 'isolated' 5-fold atoms under this definition is 984 meV. An alternative definition of 'isolated' considers only 5-fold atoms for which no other defects occur up to the surrounding $4'''$ shell so no other 5-folds are found within 3 bonds. There are fewer (3,953) of such 5-fold defects in the 1M-atom structure, with an average energy of 950 meV. Across almost all defects ($> 97\%$), using the second definition, the average energy

of 5-folds and their environment, including those that cluster, is 749 meV. This value compares with average energies for 3-fold atoms of 726 meV. The remaining 3% were excluded because these rare cases involve, for example, 2-fold and 6-fold connected atoms, or more complicated clusters involving 3- and 5-fold coordination which are complex to handle systematically.

4.7.3 Molecular dynamics

MD simulations and local-energy evaluations were carried out using the LAMMPS software,[212] interfaced to the MTP[90] and GAP[12] codes, respectively. The timestep was 1 fs. Melt-quench simulations were conducted using the same variable-rate protocol as described in Ref. [14], that is:

1. Melting a random initial structure at 2500 K for 20 ps
2. Equilibrating the resulting liquid at 1500 K for 100 ps
3. Quenching at a variable rate: 10^{13} K s^{-1} between 1500–1250 K, 10^{11} K s^{-1} between 1250–1050 K over which vitrification takes place, back to 10^{13} K s^{-1} between 1050–500 K for a total cooling time of 2.08 ns, before a relaxation with conjugate-gradient (CG) descent.

The same type of analysis of defects performed on the structure cooled at 10^{11} K s^{-1} (taken from Ref. [1]) was performed on a separate 1M-atom a-Si structure produced by annealing a rapidly quenched structure (10^{13} K s^{-1} throughout the temperature range 1500–500 K) at 850 K for 3 ns before cooling back close to 0 K at 10^{13} K s^{-1} and relaxing by CG descent. The results for an alternative structure produced via

a different melt-quench protocol ('annealed'), which are equivalent to those of this chapter, may be found in the appendix at Figs. A.5 and A.6, and show that our conclusions are robust with respect to the precise number of defects and the path to reaching them.

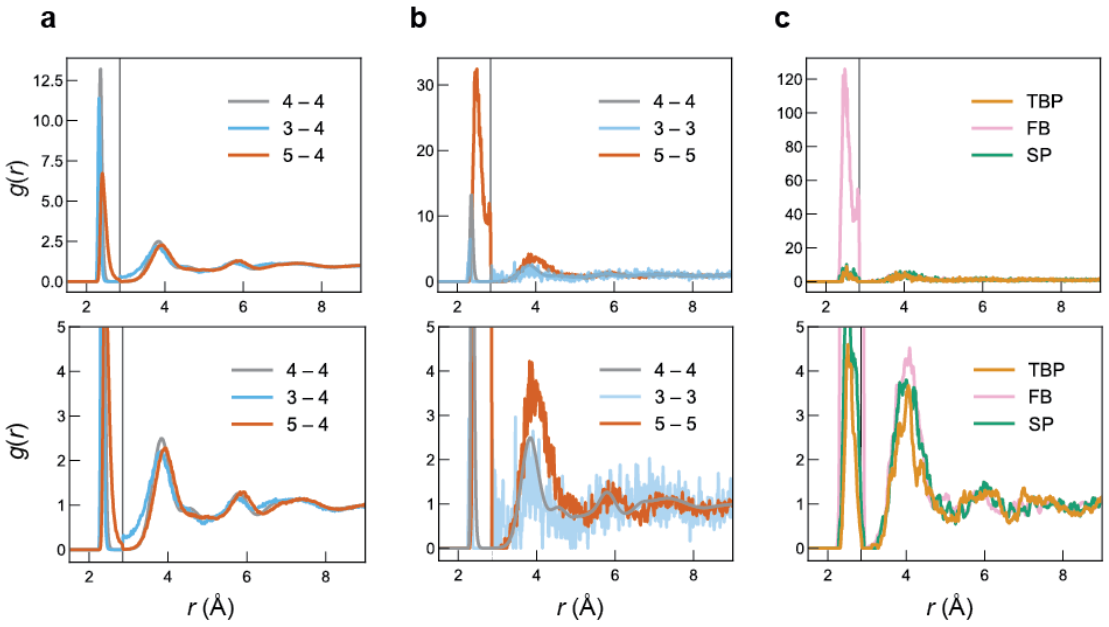


Figure 4.7: Pair correlations for defect atoms. In all columns, we show partial radial distribution functions, $g(r)$, in full (upper panels) and magnified (lower panels). (**a**) Correlations between each of 3-, 4-, and 5-fold connected atoms and 4-fold connected atoms. (**b**) Correlations between atoms of the same connectivity. (**c**) Correlations between 5-fold connected defects, separated by classification into trigonal-bipyramid-like (TBP), 'floating bond'-like (FB), and square-pyramidal-like (SP) as defined in Fig. 4.3. Vertical lines indicate the radial cut-off used to define coordination numbers.

5 | The hierarchical structure of amorphous MoS_3

5.1 Acknowledgements

First, I must thank Prof. Simon Clarke for access to laboratory facilities, and especially Dr. Simon Cassidy for his practical leadership and assistance in the synthesis and characterisation of our a-MoS_3 samples. Simon procured the starting material, tutored me through glove box practices and the Schlenk line heating setup, ran the X-ray total scattering experiments, and provided invaluable scientific insight into the interpretation of the data—it would have been impossible without his help. I also owe thanks to Prof. David Keen and Elodie Harbourne for their help in reducing the total scattering data, and instrument scientist Dr. Paul Henry for his help in collecting the neutron data.

5.2 Introduction

The compound MoS_3 has only been observed to be amorphous until above 250 °C where it decomposes to crystalline MoS_2 and elemental sulfur.[85] a-MoS_3 appears to be a promising material for a number of applications: most notably a substitute for sulfur in battery cathodes[77] and supercapacitors with less problematic reactivity than the elemental form,[78] and a catalyst for hydrogen evolution or nitrogen activation superior to MoS_2.[79]

In fundamental studies, there remains uncertainty over the Mo–S structural units that predominate in a-MoS_3. Originally, the identity of MoS_3 as a true single

compound was even doubted, with it thought to be a mixture of MoS_2 and noncrystalline sulfur.[351] Later, based on X-ray radial distribution functions that suggested MoS_3 likely was, in fact, a genuine compound, a model was proposed that featured Mo chains possessing alternating long and short bonds and trigonal prismatic coordination by sulfur (see Fig. 5.1 for an illustration).[85,352,353]

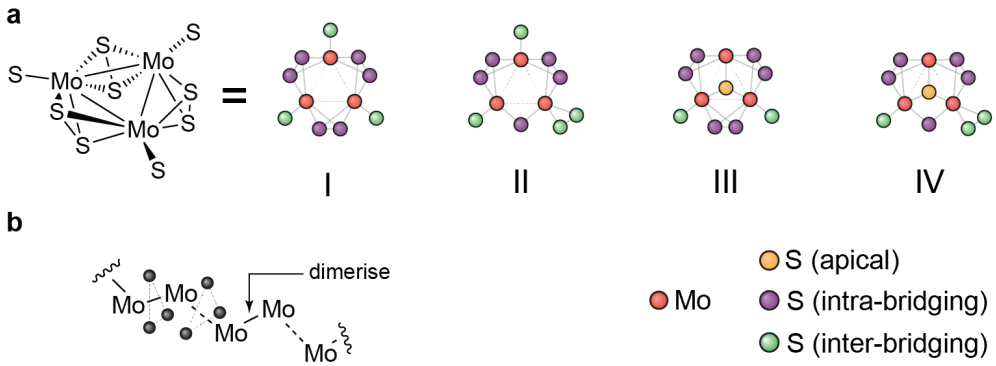


Figure 5.1: Proposed structural building blocks for MoS_3. (**a**) Triangles, with the four fragments proposed by Weber *et al.* labelled I–IV. My sulfur categories are shown in different colours. In my nomenclature, apical sulfur (S^{2-}) sits on top of an Mo_3 triangle, 'intra-bridging' sulfur connects two Mo atoms from the same triangle, and 'inter-bridging' sulfur connects entirely separate Mo units. (**b**) Chains, with alternating long and short Mo-Mo bonds.

Shortly after the chain proposal, Weber *et al.* introduced a model consisting of triangular trinuclear Mo_3 clusters. The new proposal was based on chemical extrusion experiments and a reinterpretation of the X-ray photoelectron spectroscopy (XPS), vibrational, and extended X-ray absorption fine structure (EXAFS) evidence that was initially relied on by Liang *et al.* and others for the chain model.[82] Later still, Hibble *et al.* used reverse Monte Carlo (RMC) simulations to fit a structural model to neutron total scattering data, concluding, contrary to Weber *et al.*, that chains were in fact the dominant motif after all.[86,354]

Since these publications, to our knowledge, further experimental studies on a-MoS_3 have focused on characterising its catalytic activity,[79,355] tribology,[356] and capacitive properties[77,78,357] without revisiting the structural debate.

In this chapter, I describe the development of an ML potential for compositions MoS_x where $2 \leq x < \infty$. Combining constrained Molecular Dynamics (MD) simulations and new total scattering experiments on our freshly-prepared samples, I explore the hierarchical structure of MoS_3 and its consequences for the decomposition to poorly-crystalline MoS_2. We find that, in agreement with Ref. [82], structures containing triangular trinuclear Mo_3 clusters agree best with the available experimental data and provide the lowest-energy MoS_3 structural models. In contrast to Ref. [82], the configurations of our proposed clusters differ substantially from, and the linkages between clusters are more complex than, what was originally thought. We see that the sulfur bridges between clusters result in a tendency for the planes of connected triangles to align. This imposes medium-range orientational order on the amorphous phase, which can be compared to nematic order in liquid crystals. The pre-organised alignment of triangles provides an easy route to crystal nucleation that could explain the facile decomposition of a-MoS_3 to layered MoS_2.

5.2.1 The experimental evidence

Before analysing the results of our simulations, it is important to assess the existing evidence for the structure of a-MoS_3. The original experimental evidence behind the chain model, originally proposed by Liang *et al.*,[85] can be summarised as

follows:

1. Their EXAFS data indicated a first Mo–Mo distance of about 2.75 Å, an average Mo–S bond length of 2.44 Å, and a coordination number of six for Mo by S. They also found a second Mo–Mo distance at about 3.1 Å.[353,358] However, we note that Hibble *et al.* and others found no such extra distance in other EXAFS experiments.[80,359] Both sets of EXAFS data are compared with our structural models later in Fig. 5.6.

2. XPS appeared to show two types of sulfur in a 1:2 ratio, with formal oxidation states given by $Mo^{V}(S_2)_{1/2}S_2$. Again, this interpretation has been questioned (e.g. by Weber *et al.*[82]).

3. Magnetic susceptibility measurements implied a diamagnetic compound (i.e. negligible unpaired electrons on Mo or S).

With one unpaired electron remaining for oxidation state Mo^{V}, it seems natural to pair these via metal-metal bond formation. *Direct* evidence for chains, however, is lacking. Furthermore, there is an established tendency towards electron-precise trinuclear clusters for Mo (see Cotton's classic *Advanced Inorganic Chemistry* for an excellent summary).[360,361] Notable examples include the $[Mo_3O_4]^{4+}$ core present in Mo^{IV} aqua ions and many other oxo complexes.[362] The sulfur chemistry of trinuclear Mo is similarly well-established,[87,88] with much interest dedicated to biomimetic compounds related to the $[Mo_3S_4]^{4+}$ core of the cofactors in nitrogenase and HER enzymes.[360,363,364] Polymeric forms of such cluster cores, linked by halide ions in $\{Mo_3X_7Y_4\}_n$ (X=$\{S, Se\}$, Y=$\{Cl, Br\}$), are also well-known.[365] Including

the Mo – Mo bonds and allowing for an extra sulfur atom bridging clusters, each metal atom is 9-coordinate in these cores and hence involves all valence orbitals in bonding. This electronic structure goes some way to explaining the kinetic inertness of such clusters, even under oxidising conditions where metal-metal bonds rarely survive.[361] Fig. 5.2 illustrates a few examples of notable trinuclear molybdenum sulfide clusters.

The major new piece of experimental evidence added by Weber *et al.* in support of the triangular proposal was the chemical extrusion of derived compounds of these clusters by reactions with, separately, NH_3, KCN, or NEt_4. It was claimed that the Mo^V-based chains could not explain their formation. A similar study observed no intermingling of isotopes during extrusion reactions of a physical mixture of $^{92}MoS_3/^{100}MoS_3$,[368] which supports Weber's underlying assumption of rapid extraction of pre-existing Mo_3, rather than rearrangement of chains to produce the extrusion products. The binding energy implied by the Mo $3d_{5/2}$ band was also argued to be consistent with a formal oxidation state of Mo^{IV}, rather than Mo^V as previously thought. The sulfur bands, with overlapping binding energies for apical S_2^-, bridging S_2^{2-}, and terminal S_2^{2-} (from XPS), make it difficult to accurately quantify the relative ratio of these species. The vibrational (infrared and Raman) signatures of isolated cluster compounds and a-MoS_3 show similarities that further support the cluster model. Additional, albeit indirect, evidence for the triangular model comes from HAADF-STEM images of highly-dispersed MoS_3 on carbon surfaces, which show oligomeric aggregations of triangular Mo_3 units.[369]

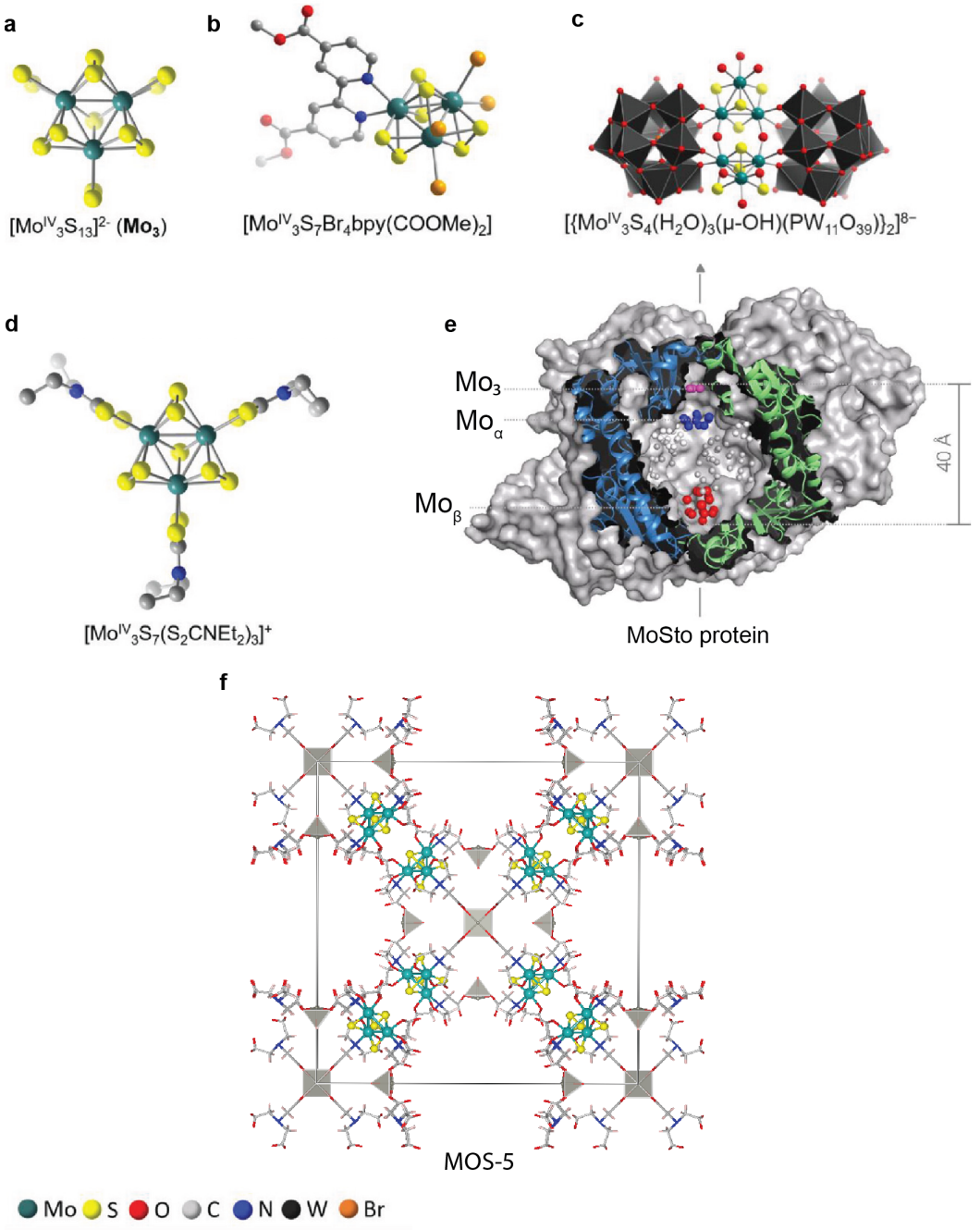


Figure 5.2: A selection of known trinuclear molybdenum clusters. (**a**–**d**) Promising molecular HER catalysts, reproduced from the review of Ref. [88] under a CCBY license. Panel c shows two $[Mo_3S_4(H_2O)_3(\mu\text{-OH})]^{3+}$ species that are stabilised with a polyoxometalate (POM) cluster for light-driven, homogeneous HER. (**e**) Molybdenum storage protein (MoSto), which can harbour in excess 100 Mo or W atoms as discrete POM clusters. An Mo_3 cluster plays a key role at the binding site. Reproduced with permission from Ref. [366], copyright 2014 Elsevier Inc. (**f**) The MOS-5 metal organic framework (MOF), drawn with crystallographic data from Ref. [367]. Trinuculear Mo_3S_7 cores are joined together with the nitrilotriacetic acid (NTA) linker to form a molecular analogue of the a-MoS_3 structure, which is also competent at hydrogen evolution.

5.2.2 The modelling evidence

The first significant theoretical investigation of a-MoS_3 was conducted in Ref. [354]. Therein, Hibble *et al.*[354] agree that the oxidation of state Mo^{IV} is most appropriate for a-MoS_3, but observe that the fragments as described by Weber *et al.* do not result in Mo^{IV} when linked together with the required stoichiometry, which we also find to be a problem with the Weber models. Hibble *et al.* employed reverse Monte Carlo (RMC) simulations to compare models based on chain and triangle motifs to their neutron total scattering data, concluding that only their chain models are able to fit the $G(r)$. However, it is well known that there is no particular structural model that can uniquely reproduce the experimental $G(r)$. In fact, the inverse problem is sufficiently underdetermined that chemical constraints are almost always required to maintain sensible structures during refinement based on RMC simulations.[370] The choice of such constraints places a strong inductive bias on the model. In this case, the connectivity of their triangle model was constrained internally (i.e. all Mo−Mo, Mo−S, and S−S bonds were fixed within distance ranges) to be precisely that of Weber's cluster I in Fig. 5.1a, and, externally (i.e. connections between triangles), to match the connectivity of carbon fullerenes. The latter design choice appears to have been one of convenience for initial structure generation, rather than because of any implied structural similarity of a-MoS_3 and fullerenes. Such restrictions, we argue, prevented the triangle model from being sufficiently flexible to fit the data as well as their chain model did. We shall see from constrained MD simulations later in this chapter that triangle models *do* agree with the scattering data when allowed to evolve with looser bonding

constraints, steered instead by the energetics of an ML potential.

A separate, more recent *ab initio* study examined the vibrational and energetic characteristics of small triangle and chain models for a-MoS_3.[371] The authors of Ref. [371] found that the fragments proposed by Weber *et al.* reconstruct significantly under DFT-driven geometry optimisation, and that the interconnections between clusters are likely complex in that they involve more sulfur than Weber envisioned, similar to the findings we will come to present. Fragments with oxidation state Mo^{IV} were found to be much more stable than Mo^{V}; chains based on Mo^{IV} were found to have a single typical Mo−Mo bond length.

This study provided[371] a comparison between the energetics of chains and triangles, which showed no particular preference for either model—in fact, contrary to our findings, they conclude that a mixture of the two motifs is most likely contained in a-MoS_3. For the realisations of each model in Ref. [371], either triangles *or* chains can be slightly energetically preferred over one another depending on the precise system-size and the structural linkages between units. However, it was not possible to access bulk-like structures with DFT because of the computational cost—the maximum size studied in Ref. [371] was 64 atoms, at which point a plateau in energy was not yet achieved—and were limited to studying predetermined structures based on either an MoS_2-like lattice, or manually-created quasi-1D structures.

Some preference for aligned triangles was also observed for a 'patch model' of MoS_3 in Ref. [371], which was produced by decorating an MoS_2-like lattice

with connected triangular units. This model was found to be more stable than corresponding linear oligomers of triangles,and was more stable than any chain model. The energetic stability of aligned triangle models here foreshadows our discussion of nematic order in Fig. 5.11 later. The use of large-lengthscale, long-timescale simulations with ML potentials allows us to overcome these limitations and discuss the bulk structure of a-MoS_3 on a system-size-independent basis.

5.3 Structural models

Amorphous structural models are produced computationally most often via simulated slow cooling.[313,372,373] For a-MoS_3, crystalline MoS_2 provides a deep thermodynamic sink that precludes this approach; slow cooling leads to phase separation, which, we emphasise, is consistent with the experimental observations for MoS_3 under heating. To enable us to explore the amorphous phase, we overlay soft bonding constraints on top of the predictions of a quantum-accurate ML potential to compare hypotheses on the local motifs present in a-MoS_3, i.e. chains, triangles, or mixtures thereof. The constraints prevent the unwanted crystallisation that occurs due to both the elevated temperatures that we need to sample structural space widely, and slow cooling rates that we need to produce a sufficiently stable amorphous phase. Crucially, we enforce only the bonds between Mo atoms (including for various chain lengths), so that the arrangements of sulfur are free to evolve according to their lowest free energy states.

Practically, we prepare starting structures according the the schematic in Fig. 5.3, before melting and cooling with the same protocol to produce structural

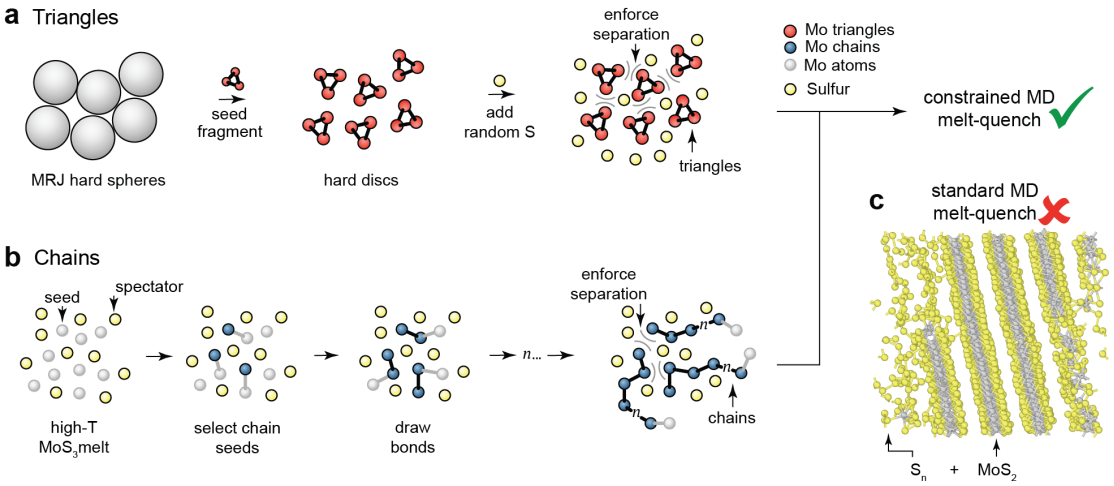


Figure 5.3: Schematic for constrained structure generation. (**a**) Triangle protocol (**b**) Chains protocol (**c**) Inaccessibility of a-MoS_3 to unconstrained melt-quench simulations. There is a strong driving force for the thermal decomposition of metastable a-MoS_3 to MoS_2 and elemental sulfur, so any melt-quench simulation that is both large and slow enough to produce a reasonably ordered amorphous phase instead phase-separates.

models via constrained MD. We minimise all structures with the constraints lifted before analysis, except for the analysis of triangle–triangle interactions to make coarse-graining simpler to achieve. For our triangle model, we start by generating a maximally-jammed random packing of hard spheres using the code of Ref. [374]. We place bare Mo_3 units on the centres of these spheres with random orientations such that $r_{\mathrm{Mo-Mo}} \geq 2.5\,\text{Å}$, which effectively gives us a hard disc model. Then, sulfur atoms are randomly sequentially added into the remaining gaps in the structure such that $r_{\mathrm{Mo-S}}, r_{\mathrm{S-S}} \geq 1.5\,\text{Å}$ until the stoichiometry reaches MoS_3. The ACE potential (the fitting of which is described in Section 5.4) is used to propagate MD simulations under constraints that penalise the break-up of triangle units via an attractive harmonic potential starting at $r_{\mathrm{Mo-Mo}} = 2.9\,\text{Å}$ and, additionally, enforce an Mo−Mo coordination number of two.

For the chains, we start from a high temperature melt of MoS_3 and draw

'bonds' on this template using a breadth-first[375] greedy algorithm: N seed atoms are chosen to form the base of each of N chains. The nearest neighbours of each seed are added to the growing chains and become the new living 'end', and so on until all Mo atoms are members of chains, or the chain length reaches the desired length according to a target distribution.

5.4 ACE fitting

We use approximately 9,000 structures to fit an ML potential for the MoS_x system, with a base level of coverage provided by *de novo* RSS, and additional structures added to improve accuracy specifically for the triangle- and chain-based MD simulations that we later use to discriminate between the models. The database is summarised in Fig. 5.4 and Table 5.1; further details of the fitting and validation for a variety of physical tests involving MoS_2 and MoS_3 are left for the methodology section (5.11) at the end of the chapter.

5.5 Modelling results

Our structural model built with triangles is approximately $15\,\mathrm{kJ\,mol^{-1}}$ more stable than the corresponding chain model. Both are metastable with respect to decomposition to MoS_2 and elemental sulfur, as is observed experimentally. To reduce any uncertainty in our estimation of the total energy of our models due to the accuracy of the ML potential, in Fig. 5.5a we compare the DFT energies for structures of the largest size for which we can evaluate them: approximately 4,000 atoms. For the chain model discussed in this chapter, we use $N = 1$ seeds, i.e. a single chain looping throughout the entire structure. Similar results are obtained

Table 5.1: ACE fitting database composition by configuration type. 'Displaced crystals' include pure Mo and pure S, as well as other experimentally-observed Mo–S phases contained in the Materials Project database.[25] Nudged elastic band trajectories contain migration paths for several types of point defect in MoS_2: S and Mo interlayer interstitials, the S vacancy, and the Mo split interstitial. 'Phonon' structures consist of idealised MoS_2 monolayers with small displacements to individual atoms, created with the `Phonopy` software package.[376] 'Exfoliation' structures consist of MoS_2 layers sampled over a range of interlayer separations.

Structure type	**Quantity (structures)**	**Quantity (atoms)**
RSS generations (minima)	2,026	27,447
RSS generations (trajectory points)	4,357	43,006
Displaced crystals	485	7,327
Defect nudged elastic band	234	19,300
Displacements for phonons	341	20,640
Exfoliation	53	318
MD snapshots (chains)	584	116,800
MD snapshots (triangles)	389	36,330
MD snapshots (free)	517	124,087
Totals	8,986	326,843

for chains of length 5–1000 for 20,000-atom simulations, which are included in the Appendix along with results for mixtures of chains and triangles, and mixtures of triangles and free Mo. None of these models agree better with the scattering data than a model based purely on triangles.

We prepared an experimental sample of a-MoS_3 according to the method of Ref. [80] and measured its X-ray and neutron total scattering, which we compare to existing data[354] in Fig. 5.5c. Our experimental scattering data are broadly consistent with previous experiments, suggesting that our sample of a-MoS_3 has a similar structure to other samples,[354] but with a larger Q range for the neutron data and higher resolution for the X-ray data due to improvements in technology and facilities since the original work. Structurally, the triangle model also compares

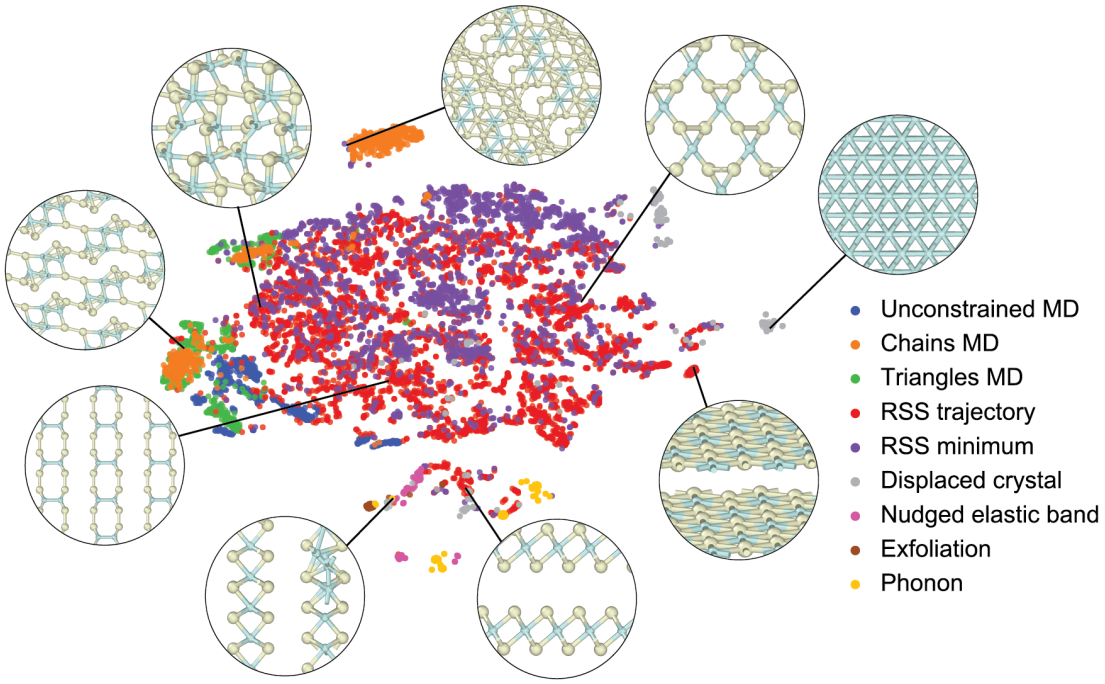


Figure 5.4: Mo–S database sketch map. To illustrate the fitting database used for the ACE potential, we use the t-SNE dimensionality reduction technique on an Mo-centred SOAP distance matrix. Parameters for the SOAP descriptor are given at the end of the chapter. Around the outside, we highlight particular regions of the map with illustrative structure sketches, with Mo in blue and S in yellow. RSS structures provide very broad coverage of both MD-like and crystal-like structures.

more favourably than the chain model with this total scattering data. The $S(Q)$ agrees more closely for the triangle model, particularly in the high-Q region, which is influenced primarily by the local structure, and also the region 10–12 $Å^{-1}$. This difference in the local structure is evident in the pair correlation function, $G(r)$, where the chain model has somewhat less well-defined Mo–Mo distance and longer Mo–S distances than are present in the experimental data. The height of the FSDP, for any of our models, does not quite reach that of the experimental data, which may suggest increased medium-range order in experimental samples than is present in our models or, alternatively, could be a result of difficulties in the absolute normalisation of the experimental database. This order must be distinct from the

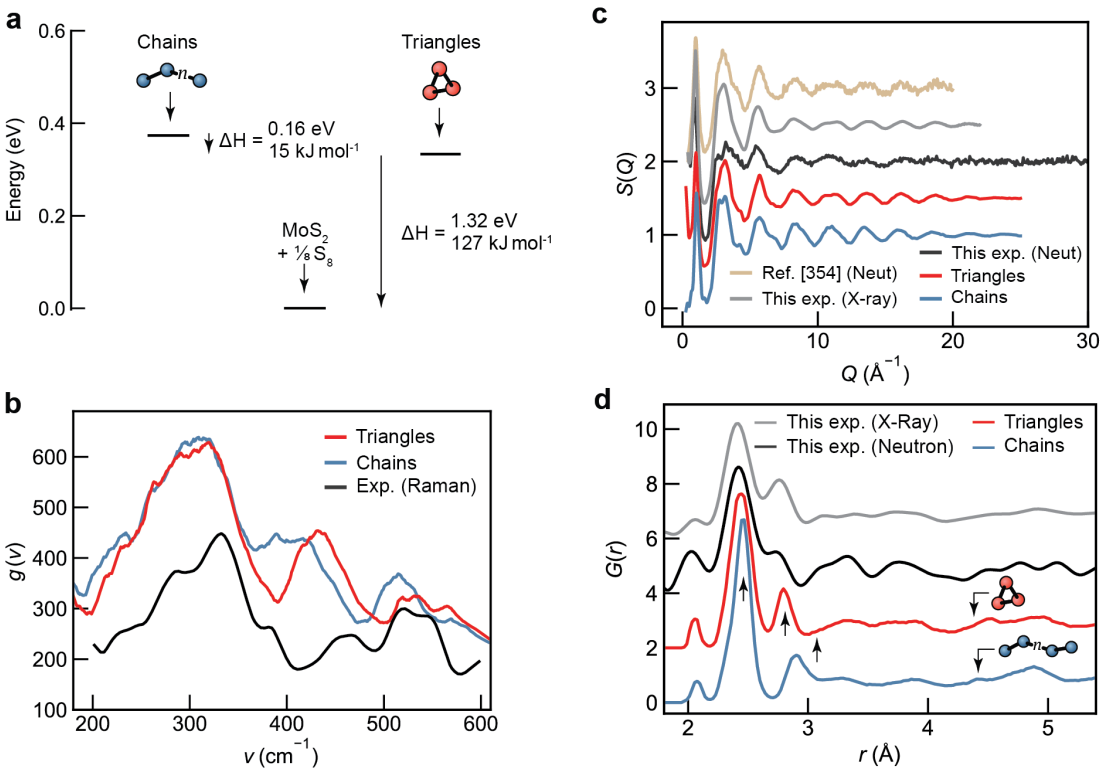


Figure 5.5: Comparison of structural models with experimental data. (**a**) Energetics. We produce structural models containing 4,006 atoms from melt-quench MD simulations under the constraints described in panels a and b and evaluate their energies with DFT. (**b**) Simulated vibrational spectra from velocity-velocity autocorrelation functions compared to Raman data from Ref. [82] (**c**) Structure factors from our diffraction experiments, X-ray (grey) and neutron (black) for the same sample, compared to calculated values from the constrained MD simulations (neutron weightings) described in panel a and to a previous experiment (light brown).[354] Arrows indicate regions where the short-range order disagrees with experiment for the chain model, but agrees for the triangle model. (**d**) Total radial distribution function via the Fourier transform of the experimental $S(Q)$, compared with direct calculation (with neutron weightings). Broadening by convolution with a tophat function of width 0.1 Å has been applied to the simulated data.

Bragg-like peaks associated with MoS_2 formation, which occur at substantially different Q values compared to the FSDP. We emphasise that no fitting to the scattering data was performed in the production of our models, in contrast to RMC methods, yet the quality of the match to data remains comparable to that of RMC.

Figure 5.6 compares the partial RDFs of our structural models to the

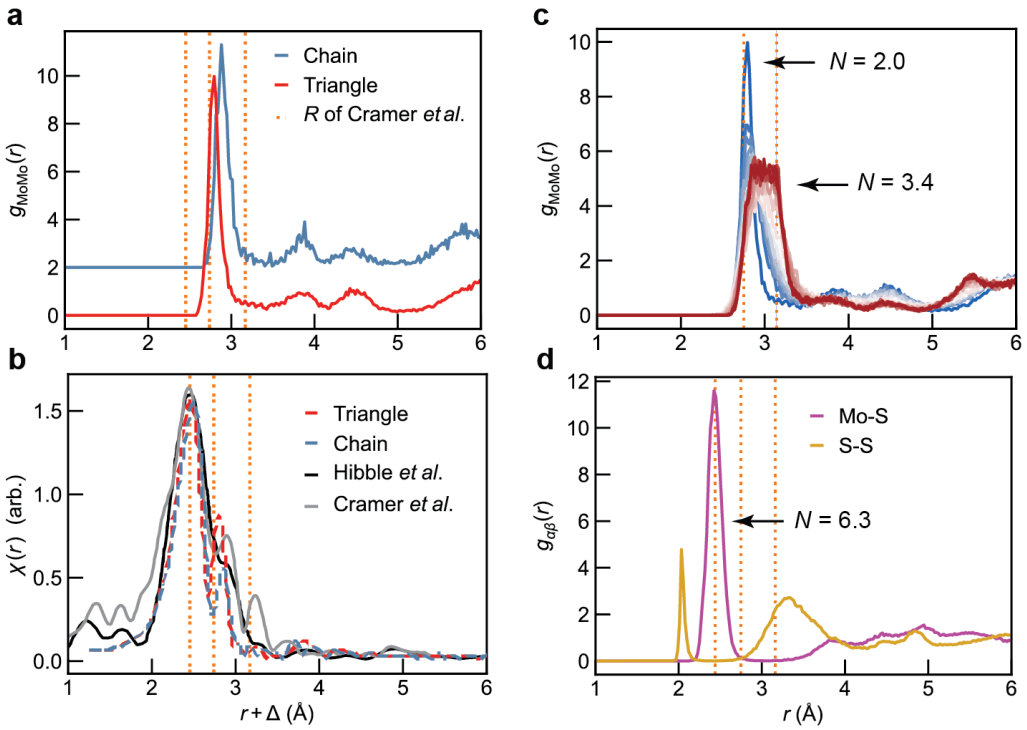


Figure 5.6: Comparison of structural models with EXAFS data. (**a**) Mo–Mo partial RDF for our structural models based on chains (blue) and triangles (red). Vertical dashed lines mark the distances fitted to EXAFS data by Cramer *et al.*[353] (**b**) Experimental EXAFS data taken from Cramer *et al.*[353] and Hibble *et al.*[80] and compared to simulated curves using the code of Ref. [377]. Curves are shifted in r via constants that are chosen to align the positions of the first peak and normalised to match the heights of the first peak. (**c**) Partial Mo–Mo RDF as a function of heating to 1400 K, which triggers the partial formation of MoS_2-like layers (described fully in Fig. 5.11). N values indicate the coordination number, which is determined by integration of the corresponding peak. (**d**) Partial RDFs for Mo–S and S–S for our triangle model.

available EXAFS data. In Fig. 5.6b, I have marked the key first-neighbour distances fitted and reported by Cramer *et al.* with vertical orange dashes.[353] Note that $\chi(r)$ can not be directly interpreted as a radial distribution function due to a phase shift that appears in the EXAFS absorption coefficient.[378] The theoretical curves in Fig. 5.6b are calculated from the Fourier transform of a simulated EXAFS $\chi(k)$ signal using the `FEFF9` code of Ref. [377]. We use a k^3 weighting and a Gaussian window function with width 8 Å for the transform. Both structural models agree well for the Mo–S distance and the coordination number of S around

Mo ($N = 6.3$), as estimated with EXAFS ($6 \pm 25\%$).[353] Our triangle model agrees better for the primary Mo–Mo distance (middle orange dash from left); neither model predicts any significant second bonded Mo–Mo distance even though both models are relaxed without bonding constraints after quenching. We see, from a comparison of the experimental EXAFS data of Cramer[353] and Hibble,[80] that the presence of this second Mo–Mo distance within the Fourier ripples is perhaps questionable. Hibble also suggested that any second Mo–Mo distance could have been a result of the presence of some MoS_2 in the sample. The production of MoS_2 via simulated heating in our models appears to support this conclusion, where Fig. 5.6c shows that the $g_{\mathrm{Mo-Mo}}(r)$ of a mixed MoS_2-MoS_3 decomposition product tends towards a peak at exactly this distance, which corresponds to non-bonded Mo–Mo separations in the MoS_2 plane.

Figure 5.5b shows that the vibrational properties of the triangle model appear a better match for the experimental data than the chain model, particularly in the region between 400–500 cm^{-1} that has been associated with μ_1-S_1 species and the region 500–600 cm^{-1} corresponding to S–S stretches for different hapticities.[371] The predicted spectra in Fig. 5.5b should be viewed with caution, however, as they are calculated via the vibrational density of states alone, which does not take into account contributions to the intensity arising from the wavefunction, such as the Franck-Condon overlap or any selection rules.

5.6 Local structure of clusters

If we accept that a-MoS_3 is mostly made up of triangle-like motifs, what are the typical structures of such fragments? The chemistry of trinuclear molybdenum clusters is well-studied, with the stoichiometries [Mo_3S_7] and [Mo_3S_4] providing examples of particularly stable cluster cores.[379,380] The apical (or 'capping') μ_3 sulfur atom is common to most isolable molecular clusters,[82,381] and is present in the vibrational spectra of a-MoS_3.[82,371] We find that the cluster cores which predominate in our triangle-based structural models can be placed on a series from [Mo_3S_4] to [Mo_3S_7]. This series is traversed by sequentially replacing bridging monosulfides for disfulfides. The relative energy of the clusters, obtained by averaging the ML-predicted atomic energies of their Mo and S atoms, tends to be lower for the more abundant clusters, and lower for clusters with more S^{2-} species. In contrast to the fragments proposed by Weber *et al.*,[82] apical sulfur atoms are present in almost all our Mo_3 clusters. We also find a minority of bicapped clusters (i.e. those containing two apical sulfurs), which are also known in molecular clusters.[382] Their relative abundances are shown in Fig. 5.7b. A more detailed categorisation of all the different types of sulfur species found in our models is presented in Fig. 5.8.

Mo^{IV} is the most dominant oxidation state for Mo in our models, as indicated by the Löwdin charges for a set of 20 small-scale structural models which compare most closely with MoS_2 (shown in Fig. 5.9 at the end of the chapter). An alternative estimation for the formal Mo oxidation state can be obtained from

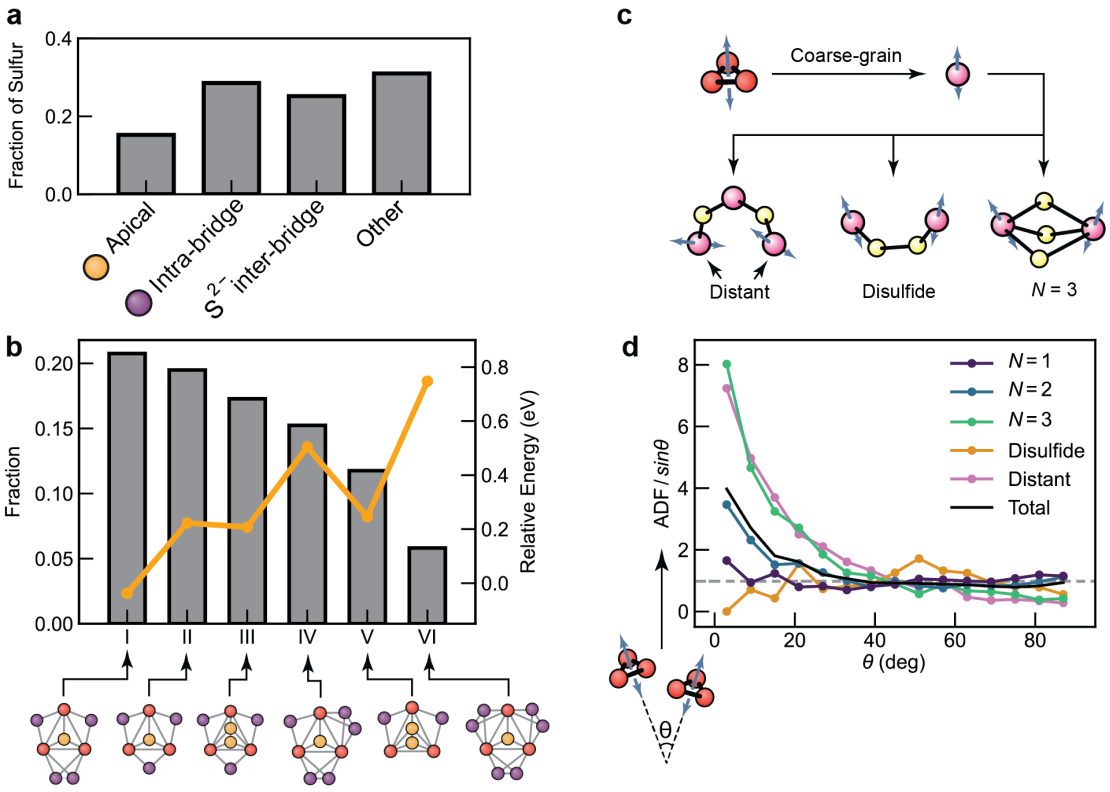


Figure 5.7: Local structure of Mo_3 triangles. (**a**) Classification of sulfur environment type. Apical S^{2-} (orange in panel b) sit above the centroid of Mo_3 units. Edge S^{2-} and $S_2{}^{2-}$ (purple in panel b) mostly sit below the bonds linking Mo atoms in each triangle. Bridging S^{2-} interconnect different Mo_3 triangles. 'Other' species include disulfide or polysulfide bridges and sulfides that are a hybrid of the 'edge' and 'bridge' categories. (**b**) Most common $[Mo_3S_n]$ cores, defined by fragments consisting of directly-bonded Mo atoms along with any S atoms that possess at least two bonds to these Mo atoms. Fragments are sketched underneath with Mo atoms coloured red, apical S atoms orange, and bridging S atoms purple. The average of the per-atom energies of the Mo and S atoms for these fragments are overlaid on the frequency histograms. (**c**) Schematic for the coarse-graining and topological classification in panel d. We define each Mo_3 unit's orientation by the normal to its triangular plane (blue). The triangles are replaced by beads, with the sulfur connections between beads preserved. The type of connection between beads is defined first by the topological distance and then, if the topological distance is one, the number of edges linking them (N). (**d**) Angular distribution functions (ADFs) between beads (triangles) separated by less than 7 Å, subdivided according to their topological link-type and normalised by $\sin\theta$. The black line shows the total ADF for all triangles within 7 Å.

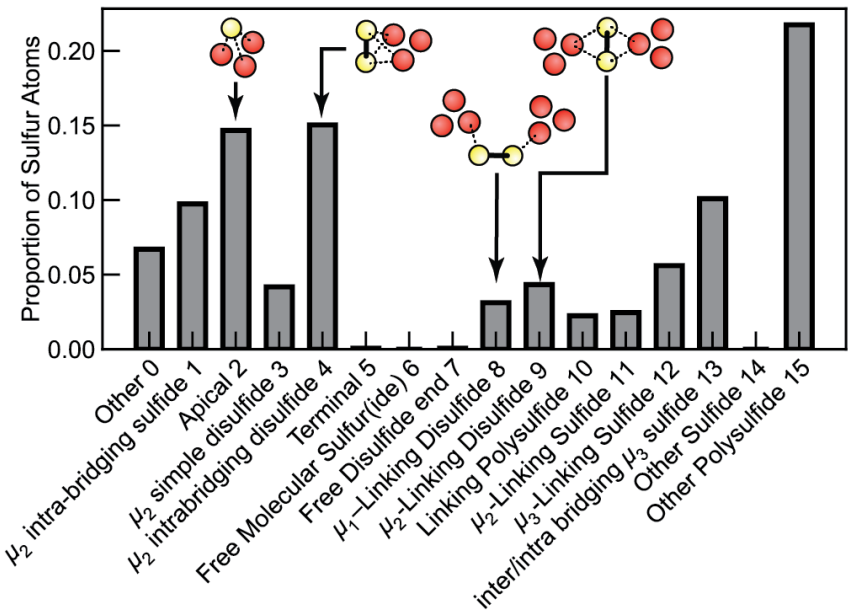


Figure 5.8: Relative abundance of different environments for sulfur, classified by the number of S and Mo contacts within bonding range (2.7 Å). 'Linking' and 'inter bridging' (di)sulfides connect different triangular units. Polysulfides consist of at least three sulfur atoms bound together with single bonds. There is a wide array of different sulfur environments, many of which are difficult to classify systematically (hence the large population in the category 'other polysulfide').

the S–S coordination numbers, which are 43 %, 52 %, and 5 % for zero, one, and two sulfur neighbours, respectively, using a cutoff of 2.8 Å. Assuming charges of –2, –1, and 0 for these sulfur species, we arrive at an Mo oxidation state of +4.11. This result is consistent with the most recent available XPS data, and the conclusions of most previous work,[81,82,86,371] which distinguish Mo^{IV} from Mo^{V} via the binding energy of the Mo3$d_{3/2}$ and Mo3$d_{5/2}$ doublet, which is red-shifted for Mo^{IV} compared to Mo^{V}. This band in a-MoS_3 matches better with MoS_2 and reference compounds such as $(NH_4)_2[Mo_3S] \cdot H_2O$ compared to Mo^{V}-containing oxides.[383] The chemical bonding in triangles is best described with single σ bonds between Mo atoms, according to the bonding analysis of the wavefunctions of the small-scale structural models (Fig. 5.9b). The Mo–Mo distance is consistent with

single bonds in Mo^{IV} cluster compounds.[384] The crystal orbital bond index (COBI) method quantifies the covalent bonding in solid-state materials.[385] For comparison, the ICOBI in diamond-type carbon is 0.95, in α-quartz SiO_2 is 0.76, and 0.09 in NaCl.[385] Metallic-like character, as well as ionic bonding, can lower the ICOBI, as in the $MoSi_2$ intermetallic, which has an ICOBI for Mo–Si bonds of 0.34 despite the small electronegativity difference between the elements.[386] In Fig. 5.9b, the Mo–Mo ICOBI values centred around 0.5 suggest a single covalent bond between atoms is the most sensible description, which is slightly stronger than the bonds in our chain model and is broadly in line with the strength of bonding predicted for the isolated fragments of Weber *et al.*[82]

The nature of the sulfur atoms in a-MoS_3 has been a source of debate. We find an extremely diverse range of environments for sulfur, with about half the total occupied in cluster 'cores' at apical sites and bridging internal Mo–Mo bonds (labelled as 'intra-bridging' in Fig. 5.7a). The remainder are primarily involved in bridging between clusters, with a mixture of S^{2-}, ${S_2}^{2-}$, and longer S_n chains in various bonding modes. Sahu *et al.* interpret their simulated vibrational spectra as evidence for the existence of terminal μ^1-S_1.[371] However, the the frequency region in experimental Raman data around $500\,\mathrm{cm^{-1}}$, the area in which DFT predicts the Mo–S stretching modes for terminal sulfides, actually appears to coincide with a minimum in intensity. We also see small numbers of these species in our models, but their number is sufficiently small as to reproduce this minimum around $500\,\mathrm{cm^{-1}}$. Separately, the authors' DFT-MD simulations also produce μ^3-species that bridge three different triangular units, which were not previously proposed

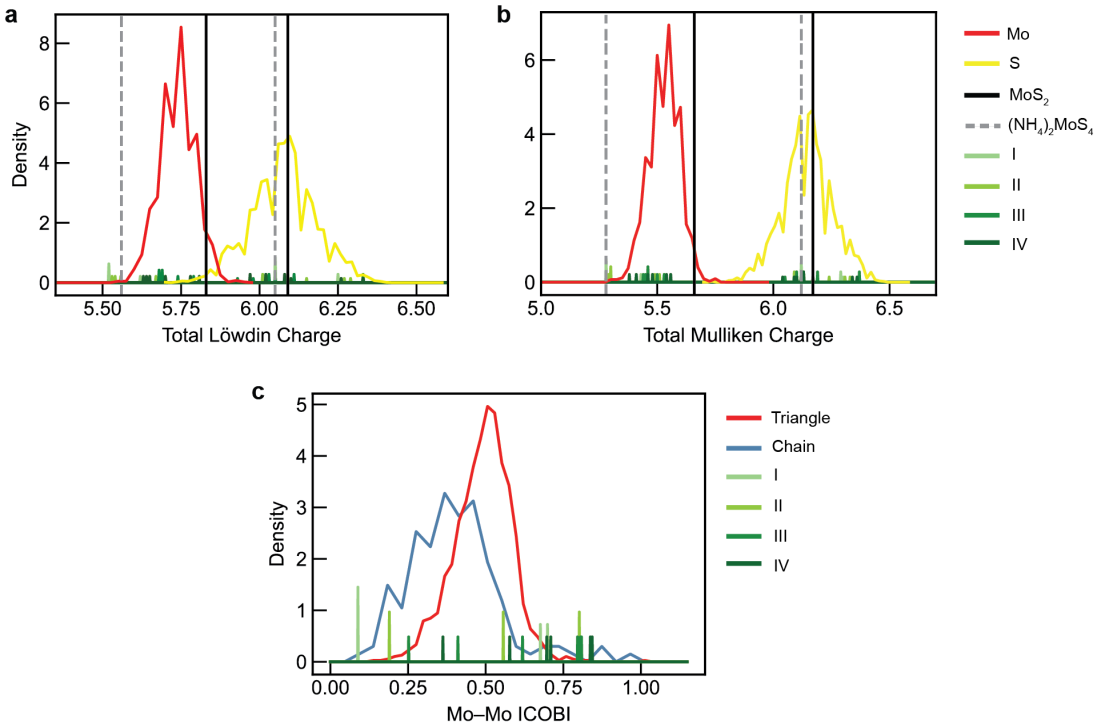


Figure 5.9: Charge and bonding analysis. (**a**) Löwdin and (**b**) Mulliken charge distributions for 20 small-scale (360-atom) structures produced with the triangle simulation protocol and evaluated with DFT and LOBSTER.[387] Black vertical lines show the charge partitioning for crystalline MoS_2; green sticks indicate the charge partitioning for the 4 isolated fragments proposed by Weber *et al.*;[82] grey dashed lines indicate the partitioning for the thiomolybdate starting material used in a-MoS_3 synthesis, which has a nominal Mo^{VI} oxidation state. The charges indicate that Mo has an oxidation state best described as Mo^{IV}, despite the apparent elevated ratio of S^{2-} to $S_2{}^{2-}$. (**b**) Integrated crystal orbital bond index[385] for Mo–Mo bonds in triangle (blue), chain (red), and idealised fragment (green) models.

by Weber *et al.* This is one of the more common modes in our models as well, as shown in Fig. 5.8.

5.7 Connections between clusters

Because there are many different ways in which Mo_3 fragments are connected via sulfur atoms in our models, to categorise them systematically we adopt a coarse-grained, topological approach illustrated diagrammatically in Fig. 5.7c. This simplifies a triangular centre to a single bead with a headless director which

describes the orientation of its plane, whilst preserving at the bead any bonds to sulfur that existed with any of the triangle's Mo atoms prior to coarse-graining. For example, end-on μ_1 'disulfide' linkages between triangles have a distance of three because there are three through-bond hops that link Mo atoms in the chain Mo–S–S–Mo. 'Distant' linkages include all beads that have a greater topological distance than three that separates them.

We see that the type of link between triangles strongly influences the orientational ordering of adjacent triangles. Where more sulfides directly connect two beads, there is an increased tendency for their normals to align (see Fig. 5.7d). Since more highly-connected triangles exhibit greater alignment, the sulfur connections between them may mediate the nematic order discussed in the subsequent section. Triangles that are spatially, but topologically remote, also tend to align their planes, which reflects a geometric constraint: close-approaching triangles that are not directly bonded through sulfur must align to avoid a steric clash from their meridional sulfur decoration. Hence these 'distant' linkages can be thought of partly as 'stacked' triangles. The stacking goes some way towards explaining the intense first sharp diffraction peak in the $S(Q)$, which has been interpreted, in the context of a-MoS_3, as characteristic of layering.[82] We limit this analysis to relatively close triangles (in space) to ensure that the classes of linkage have a comparable range of distances, because there is also a strong relationship between through-space distance and angular correlation.

5.8 From connections to nematic order

We argue that the influence of the local bonding between clusters on the alignment of triangles leads to orientational ordering in a-MoS_3 on a medium-range length scale. In this way, the order in a-MoS_3 can be likened to nematic ordering in liquid crystals. Fig. 5.10c shows the value of the nematic order parameter, S, averaged over the simulation cell for various window sizes. At each point, S is calculated by finding the eigenvalues of a local nematic tensor, $Q_{\alpha\beta}(\mathbf{r})$, of which the largest eigenvalue is equal to $\frac{2}{3}S$. We adopt the following definition[i] for Q:[388]

$$Q_{\alpha\beta}(\mathbf{r}) = \frac{1}{N}\sum_{i=1}^{N}(u_{i\alpha}u_{i\beta} - \frac{1}{3}\delta_{\alpha\beta}) \ , \tag{5.1}$$

where $u_{i\alpha}$ are the components of a vector normal to the plane of each triangle included within a cutoff radius of the point, $\mathbf{r}$, which can be any point in space. The *product*, $u_{i\alpha}u_{i\beta}$, ensures that the tensor is independent of the direction chosen for the normal. The local director, $\mathbf{n}(\mathbf{r})$, is the eigenvector associated with the local eigenvalue $\frac{2}{3}S(\mathbf{r})$. S itself indicates the strength of the alignment of the triangles in the vicinity of $\mathbf{r}$. The absolute value of S, bounded from above by 1 and below by 0,[ii] depends on the size of the averaging window (i.e. the which triangles are included in the sum of equation 5.8). Fig. 5.10c shows that the nematic order decays with distance, as is intuitive for orientational correlations

[i] If the director is known ahead of time, we can use

$$S = \frac{1}{2}\int_{\mathcal{B}}(3\cos^2\beta - 1) \ ,$$

where β is the angle to the director, $\mathbf{n}$.

[ii] The lower bound is reached when the triangles are randomly oriented. For particles with a special direction for their principle axis, a tendency towards anti-alignment can result in negative values for S.

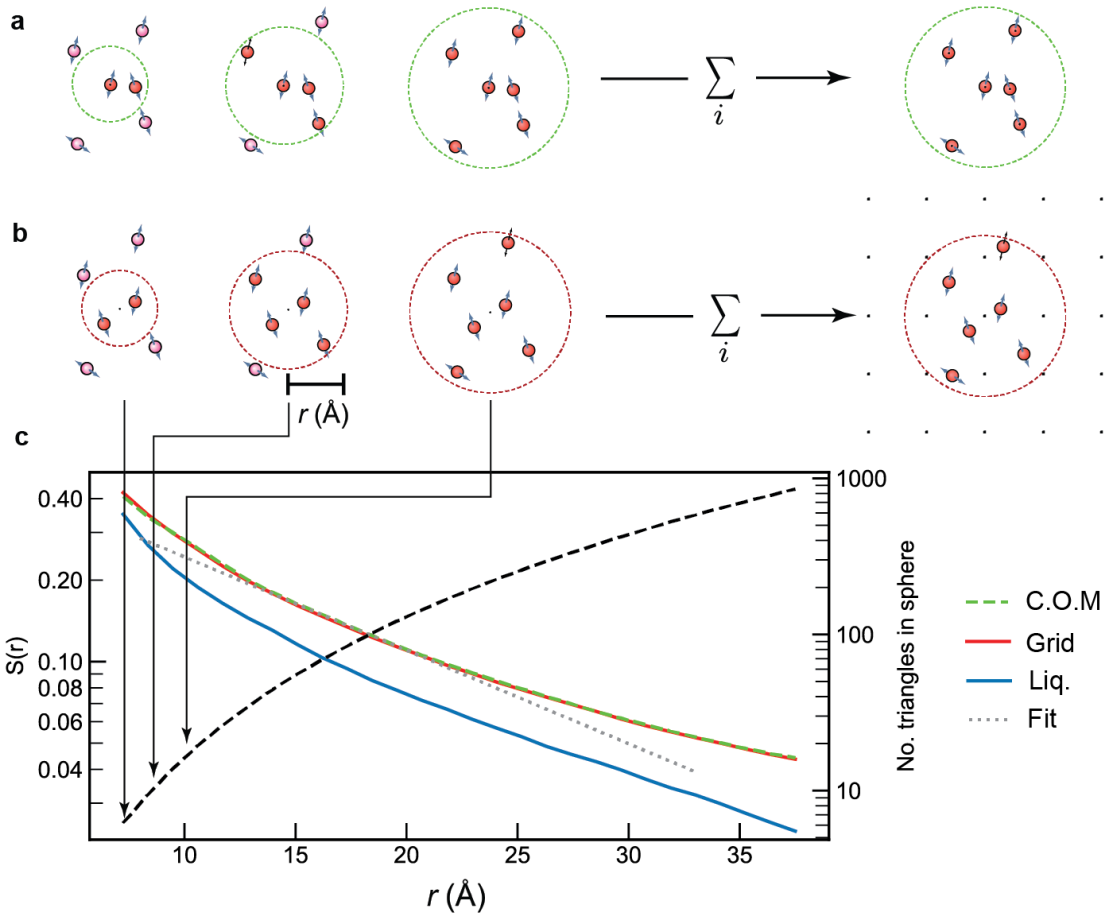


Figure 5.10: Calculation of nematic order with three different approaches. (**a**) The triangle-centred approach (yellow dash). For each triangle's centre of mass (COM), we find all other COMs within a radius, r, and construct a local nematic tensor, $\mathbf{Q}$, for this point. For each point, we extract the largest eigenvalue to calculate S, then average over all such points to give an overall $S(r)$ for the characteristic radius, r. (**b**) Same as panel b, except that we use a regular Cartesian grid to define the points at which each S is calculated. The two methods give very similar results. (**c**) Plot of the nematic order as a function of averaging window radius for the methods in panels a and b compared to high-temperature liquid MoS_3. The black dashed line gives the average number of triangles included in a window of radius r.

that also decay with distance. The rate of decay is lower, and the absolute values larger, for the amorphous phase than it is for the high-temperature liquid phase, which suggests that the amorphous phase has a higher degree of orientational order with a characteristic distance on the order of 12 Å.

Taking a slice through the scalar field $S(\mathbf{r})$ and vector field $\mathbf{n}(\mathbf{r})$, as

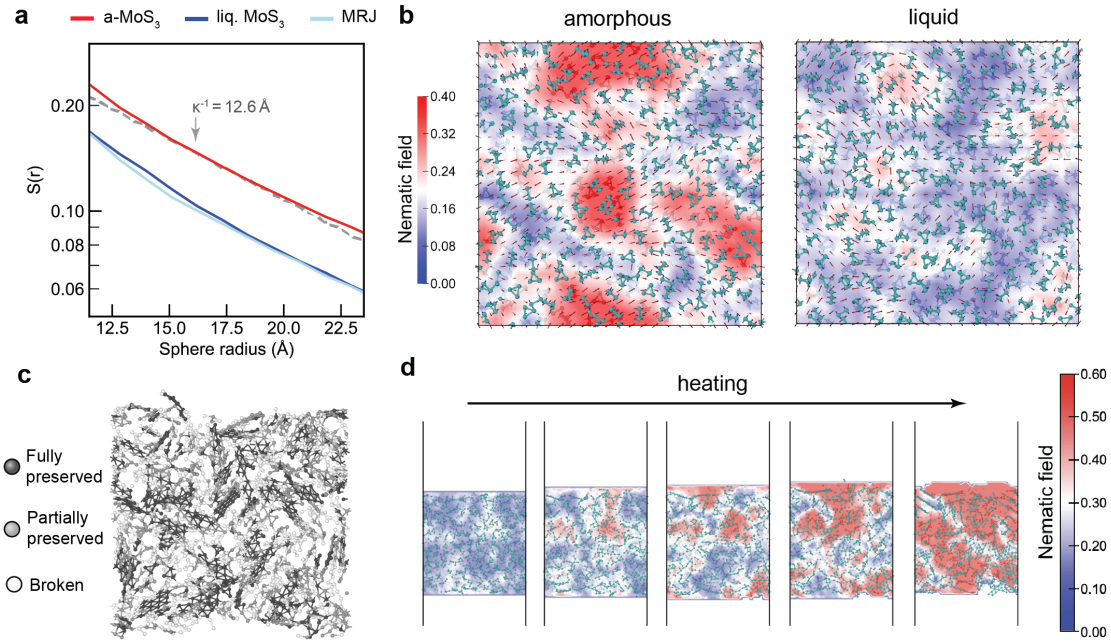


Figure 5.11: Medium-range correlations in a-MoS_3. (**a**) Nematic order parameter, S, for the triangle model (red) as a function of averaging sphere radius, r, compared to a high-temperature liquid MoS_3 (dark blue) and random MRJ model (light blue). The orientation of a triangle is defined by the headless vector[iii], $\mathbf{u}_i$, normal to the plane of its three constituent atoms. A triangle's position is defined by its centroid. A decay constant of 1/12.6 Å can be fitted to the region $r \geq 12$ Å (grey dashed line), although the decay is super-exponential. (**b**) Nematic field for the amorphous structural model (left) and high-temperature liquid (right) for an averaging radius of 12.6 Å. There are domains of strong ordering separated by more disordered regions. (**c**) Illustration of surviving triangles during the rapid heat-induced decomposition to MoS_2 layers. The darkest atoms indicate that the triangle is intact after heating to 1200 K. (**d**) Progress of thermal decomposition of MoS_3 to nanocrystalline MoS_2 and S_n rings, with colouring indicating the relation of nematic order to crystallite nucleation.

displayed in Fig. 5.11b, we see substantial domains of local order in our model of the amorphous phase. In contrast, the local value of $S(\mathbf{r})$ is much smaller for high temperature *liquid* MoS_3—roughly equivalent to that in the initial MRJ hard disk model. The fact that the ordering observed in the amorphous phase is noticeably stronger than in both the liquid phase and a random packing suggests that the measure introduced in equation 5.8 is picking up on genuine structural correlations in the amorphous phase. A visual inspection of the nematic director field in the

slice in the left panel of Fig. 5.11b is consistent with the values of the scalar field, $S(\mathbf{r})$, with discernable regions of swirling alignment that map closely on to the triangular units contained therein.

5.9 Pre-organisation for crystallisation

To gain insights into the mechanism of the decomposition, we simulate the thermal decomposition to MoS_2 by heating the amorphous phase with all constraints lifted. We use the NVT ensemble with vacuum above an a-MoS_3 slab to allow local volume fluctuations and to give the sulfur atoms flexibility to rearrange. The majority of triangles persist through the annealing intact to become the nearest-neighbour Mo atoms in the MoS_2-like layers, as depicted in Fig. 5.11c as the darkest-coloured atoms. This indicates that a *rearrangement* of existing cluster cores, rather than their decomposition and reformation, is the dominant process by which the transition occurs. Therefore, the nematic order pre-existing in the amorphous phase appears to provide nuclei for the formation of MoS_2 platelets. Fig. 5.11d charts the progress of the heating simulation, with MoS_2 layers growing from the areas of elevated nematic order, which appear white in the first panel of Fig. 5.11d (note the change in scale from Fig. 5.11b to Fig. 5.11d). The nuclei for this growth also determine the eventual location of boundaries where incompatible growth directions for the layers meet. We posit that the nematic order of pre-assembled units of the crystal in the amorphous phase explains the facile *kinetic* route in a low-temperature solid-state reaction to the thermodynamically stable phase: MoS_2. We notice that sulfur atoms are much more mobile in our simulations

than Mo, which may permit the efficient mass transport at the particle edges that is needed to eject sulfur during the formation of MoS_2 layers. The nematic order sets in most strongly near the surfaces of the slab, implying that MoS_3 particle surfaces may be the first areas to begin decomposition to MoS_2.

5.10 Outlook

Computationally, it appears that the sample history has an effect on the level of nematic order in the system: in the frames between the initial amorphous phase and the quite layered MoS_2-like endpoint, there are structures with intermediate levels of preorganisation. This begs the question as to whether similar treatment of experimental samples could enhance any nematic order, *without* irreversibly producing MoS_2, for example, via repeated annealing steps at lower temperatures, or by mechanochemical means. The latter method has already been used to synthesise MoS_3,[81,389,390] with some authors concluding that the a-MoS_3 so obtained differs structurally from that derived from the thermal decomposition route. Would this treatment have an impact on the catalytic properties? We have only analysed the bulk structure of MoS_3 in detail, with the heating simulation only hinting at a preference for MoS_2 to form near the surface, where catalysis must occur in experimental samples. One avenue for future work could involve a study of this surface structure, including whether triangles adopt preferred orientations and *post hoc* DFT analyses of the reactions of H_2 at these surfaces. Finally, it may be surprising that a compositionally simple, extended inorganic solid could harbour order related to liquid crystals (for which most known examples are

molecular). Perhaps other amorphous systems, especially those composed of anisotropic building blocks, are waiting to be found that could exhibit similar correlated disorder.

5.11 Further methodological details

5.11.1 Structural database

We built the dataset using an iterative random structure search (RSS) approach (described in Section 2.3.1) for compositions between MoS_2 and Mo_3S_{23}, drawing on Ref. [91], implemented in a form that was later merged in to our `autoplex` software.[391] There were a total of 16 generations of RSS. The number of structures per iteration, the numbers of atoms per structure, and the number of ACE parameters were increased with increasing generation as described in table 5.5. In short, 10,000 cells were generated and optimised (except the zeroth generation) at each iteration for up to 200 steps. Between 200 (at 1^{st} gen.) and 1,000 (from 6^{th} gen. onwards) structures were selected for DFT evaluation by Boltzmann-weighted histogram sampling and CUR selection based on SOAP vectors.[91] The energies for the Boltzmann-weighting are given by the distance to the three-dimensional convex hull consisting of volume, sulfur mole fraction, and energy. We also weighted structures according to these values during fitting. This selection procedure is illustrated in Fig. 5.13.

Stoichiometries from MoS_2 to Mo_3S_{23}, with cells containing up to 60 atoms, were generated randomly with the `buildcell` code before optimisation.[114] Building on the approach of Ref. [161], we use harmonic repulsions between atoms to prevent

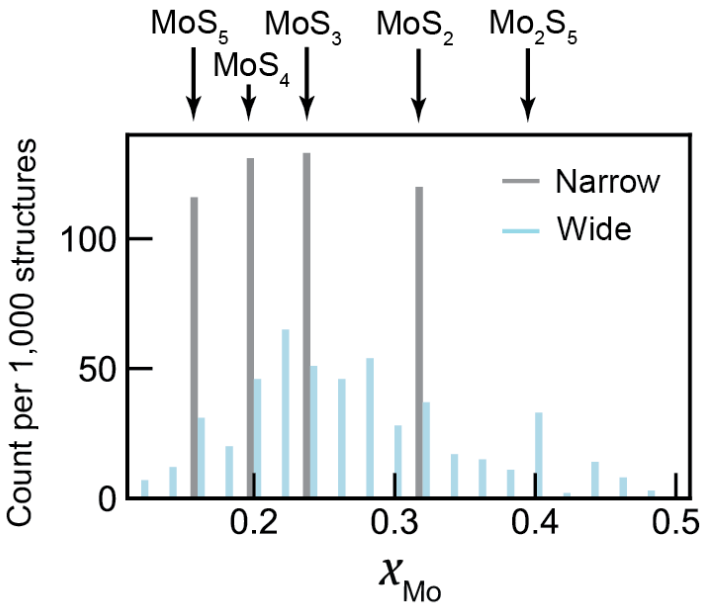


Figure 5.12: Seeding of RSS structures with variable stoichiometry. Two sets of parameters are employed to give a wide, but shallower, coverage of general stoichiometries (light blue), with enhanced data concentration at particular key values (grey). The `buildcell` commands used to generate these ranges are given in Table 5.2.

ACE potentials in the low-data regime from erroneously producing structures with extremely short distances. These begin acting at 1.4 Å, much shorter than any realistic bond in this system, with a spring constant of 100 eV Å^{-1} (cf. the typical bond lengths S=S ≈1.89 Å, Mo−S ≈2.0 Å, and Mo ≡ Mo ≈2.2 Å).

We use potentials trained on this RSS data to generate domain-specific structures to augment the dataset, which are described in more detail in table 5.1. For minimisations, we apply pressures sampled from the exponential distribution.[392] For the first five generations we use decay constants of 1, 1, 2, 4, and 8 GPa, and then 0.01 GPa thereafter. The rationalisation for this is to, first, build stability against unphysical structural collapse for the initial generations with increasing robustness (i.e. pressures) as the potential quality improves, before concentrating

Table 5.2: Parameters for initial cell generation with keywords for `buildcell`. Stoichiometry and system size are controlled by the `SPECIES` and `NFORM` parameters, respectively, where the former sets the formula unit and the latter decides the number of such units per structure. 'Wide' and 'Narrow' indicate the parameter sets referred to in Fig. 5.12. Where numeric ranges are displayed, integers are chosen randomly from within that range (inclusive) for each structure.

Parameter	**Value**
r_{Mo}	1.45 Å
r_{S}	1.00 Å
r_{char}	$1/2(r_{\mathrm{Mo}} + r_{\mathrm{S}})$
V_{fac}	14.5 Å^3
`NFORM` (wide)	`{1,2,3}`
`NFORM` (narrow)	`{1,2,3,4,5,6}`
`SPECIES` (wide)	`Mo%NUM=3-6, S%NUM=6-20`
`SPECIES` (narrow)	`Mo%NUM=1, S%NUM=2-5`
`VARVOL`	$r_{\mathrm{char}}^3 \times V_{\mathrm{fac}} = 26.655$ Å^3
`VARVOL_RANGE`	0.75–1.25
`SYMMOPS`	1-8
`SLACK`	0.25
`OVERLAP`	0.1
`COMPACT`	`T`
`MINSEP`	`Mo-Mo=2.5 S-S=2.5 Mo-S=2.0` (Å)

on the structures of interest for this study, which are all prepared at ambient pressure. We use LAMMPS[212] to perform the minimisations in batches parallelised over 48 cores, which means we can optimise all 10,000 structures for a given iteration within a few minutes.

5.11.2 Selection

Following Ref. [91], we follow a two-stage selection protocol to diversely sample the structures produced during the RSS minimisations. This is necessary because each iteration results in up to 2.5×10^6 snapshots to choose from, which is more than can be afforded routinely with DFT. Furthermore, many of the structures are similar because consecutive minimisation steps are correlated, but also because

low-energy structures (such as MoS_2-like phases) are frequently sampled. These structures do not provide significant new information, so do not improve the fit. We first select approximately 1% of structures using an unsupervised approach: CUR on the SOAP kernel matrix.[393] CUR is a decomposition approach which selects columns that best span the vector space defined by an $M \times N$ matrix. In our case, this matrix consists of the pairwise similarity between all structures generated in the current RSS iteration *and* those already generated in previous generations, where the similarity is given by the SOAP kernel discussed in Section 2.3.2 using parameters detailed in Table 5.3. We need to select entire columns because they correspond to particular structures, whereas other more common matrix decomposition algorithms, such as singular value decomposition (SVD), mix the columns during decomposition but it is not straightforward for us to generate 'mixed' structures that are linear combinations of examples in the dataset. Evaluation of the SOAP kernel matrix scales as $\mathcal{O}(N^2)$ in memory, which becomes prohibitive if we use all minimisation steps, hence we keep only steps 1, 10, and 100, as well as the final minimum before selecting with CUR.

In the second stage, we aim to sample according to a Boltzmann distribution (a heuristic that was found to work well in Ref. [91]). To achieve this, we use the energies predicted by the current version of the potential, which are converted to enthalpies using the pressure under which they were optimised. Naively sampling the observed enthalpic distribution, however, would not favour sufficiently the few low-energy structures that are produced by a typical random structure search, even using Boltzmann-weighted probabilities, so we include an additional factor

such that the probability of selecting a structure is inversely proportional to its histogram weight. Therefore, all energy intervals are sampled evenly independent of how many structures appear in each. This ensures that the selected distribution matches a Boltzmann profile, for which Fig. 5.13 shows an example: a typical Mo−−S RSS enthalpy histogram in black, and its transformation, according to this procedure, to the red distribution. For the Boltzmann biasing factor, we use temperatures of 500 K for the first generation and 300 K for all subsequent ones.

In targeting a range of stoichiometries, the systematic variation in absolute energy with the mole fraction of Mo requires us to extend the approach of Ref. [91] to avoid overrepresentation of Mo-rich structures in the database. We use the enthalpies above the convex hull, defined by the all the structures currently accumulated, in the Boltzmann weighting. The RSS dataset provides a foundation

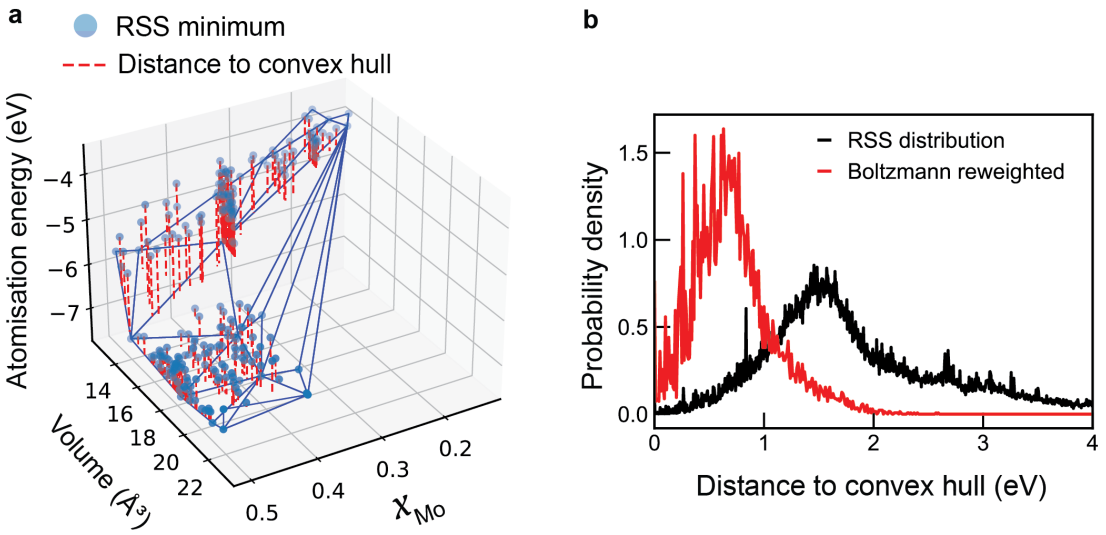


Figure 5.13: Energy-based selection policy for RSS. (**a**) Convex hull scheme, including volume dimension, used for weighting RSS structures during fitting. (**b**) Boltzmann-reweighted histogram used to downsample the $\approx 10^6$ structures minimised with the ACE potential during each RSS generation into $\approx 10^2$ structures for labelling with DFT. The ratio of those numbers corresponds to a ratio of approximately 1:10 in walltime for optimising structures vs. DFT evaluation.

Table 5.3: SOAP parameters for CUR selection and sketch map (Fig. 5.4) generation.

Parameter name	**Selection**	**Map**
`soap_average`	T	T
$l_{\max}$	6	6
$n_{\max}$	12	12
σ_{at}	0.5	0.5
r_c	5.0	5.0
n_Z	2	1
Z	`{16 42}`	`{42}`
n_{species}	2	2
`species_Z`	`{16 42}`	`{16 42}`

for the potential, which we supplement with domain-specific data to improve performance at import tasks as detailed in Table 5.1.

5.11.3 ACE potential fitting

We use the ACE framework as implemented in the PACEMAKER code[17,124,394] to fit non-linear ACE potentials according to the parameters in Table 5.4. This framework results in a good level of accuracy and expressiveness compared to other models, whilst maintaining excellent efficiency in the parallelisation of long MD simulations (as compared to GNN-based models such as MACE and NeQUIP[18,19]). We find that weighting the structures in the fit according to their distance to the convex hull (defined by stoichiometry *and* volume) performs better than either the stoichiometry hull alone. Fig. 5.14 illustrates an example of a problem encountered, though not obvious in the aggregated error metrics, when attempting MD simulations in each case. The weights are set according to

$$w = \begin{cases} 1/\Delta E^2 & \text{if } 0.1 \leq \Delta E \leq 1.0\text{eV} \\ 0.1 & \text{if } \Delta E < 0.1 \\ 1.0 & \text{if } \Delta E > 1.0 \ , \end{cases} \tag{5.2}$$

where ΔE is the distance to the convex hull.

Table 5.4: Basic ACE parameters for fitting during RSS. Where applicable, the same settings are applied to all bonds.

Parameter name	RSS	Final model
`repulsion`	1.3	auto
`kappa`	auto	0.8
`ladder_type`	`power_order`	`power_order`
`custom_weights`	T	T
`rcut`	6.5	6.5
`cutoff function`	`cos`	cos
`radbase`	`SBessel`	`ChebExpCos`
`radparameters`	5.25	5.25
`rcut`	6.5	6.5
`ndensity`	2	5
`fs_parameters`	{1}×3, 0.5	{1}×3, {0.5}×7
`batch_size`	100	100
`npot`	`FinnisSinclairShiftedScaled`	

Table 5.5: RSS ACE parameters, varying with generation as more data are accumulated.

Parameter name	Gen. 0	Gen. 1-5	Gen. 6-16	Final model
`maxiter`	250	300	300	400
`ladder_step`	200	200	300	500
`nradmax_by_orders`	8,4	8,4,4	8,6,4	10,6,4,3
`lmax_by_orders`	0,3	0,3,2	0,3,3	0,3,3,2

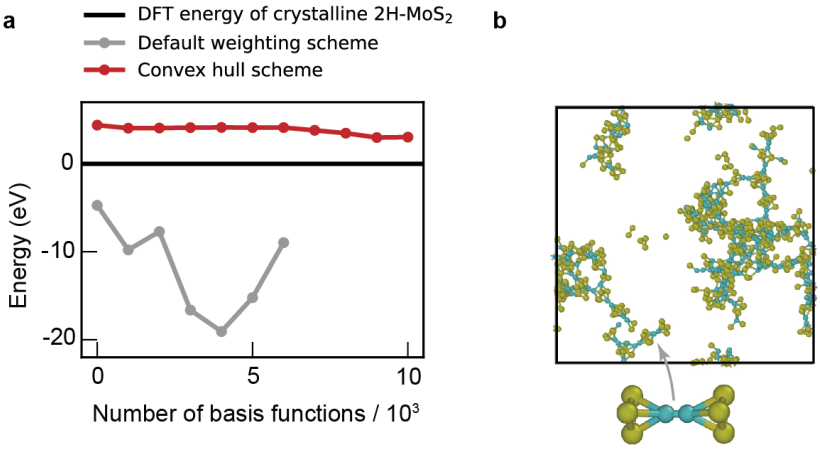


Figure 5.14: MD failure due to cluster aggregation observed for the default ACE weighting scheme. (**a**) Energy of unphysical Mo_2S_6 cluster produced during MD sampling for models fitted using the default weighting scheme (grey) and the energy-volume convex hull weights (red). Using the default weighting scheme predicts energies (erroneously) much lower than the ground state crystal structure, because the strength of Mo–Mo bonds is overestimated. (**b**) Snapshot illustrating typical failure of MD simulation ('explosion'), highlighting the Mo_2S_6 cluster for which the energy is evaluated in panel a.

5.11.4 Validation of potential for MoS_2 and MoS_3

We intend for the ML potential described in this chapter to be useful for MoS_2 as well as a-MoS_3, hence we included significant additional domain-specific training data for the former composition. We also describe numerical and physically-guided validation tests to benchmark its performance for this task in this section. In Fig. 5.15, we show the numerical errors—MAE on energies (left) and forces (right)—separately by configuration type. For 'easy' structures, such as those with small finite displacements used for phonon calculations, we have very small energy errors (i.e. all under $0.2\,\mathrm{meV\,atom^{-1}}$), whereas the RSS minima and trajectory points are much more difficult to capture accurately.

Fig. 5.16 displays a random selection of three snapshots each from the 1^{st}, 4^{th}, and 8^{th} error decile to aid chemical intuition on the types of structures explored during RSS. We find that the higher-energy error structures often have Mo–Mo metal-metal bonds (see Fig. 5.16g-i), which have a more complex electronic structure, or states that are likely to have significant multi-reference character (e.g. unpaired spins) as in Fig. 5.16h. There is an inherent limitation in the ability of the ML potential, which has no concept of spin built in, to capture these types of structures with the same kind of accuracy as for simple crystals. We note, however, that the absolute errors even on the 8^{th} percentile structures are within a level close to what is typically referred to as '*ab initio* accurate': 8 meV, or under $2\,\mathrm{kcal\,mol^{-1}}$ (note the logarithmic scales in Fig. 5.15). During the progress of RSS generations, the structures produced become increasingly low in energy vs. DFT,

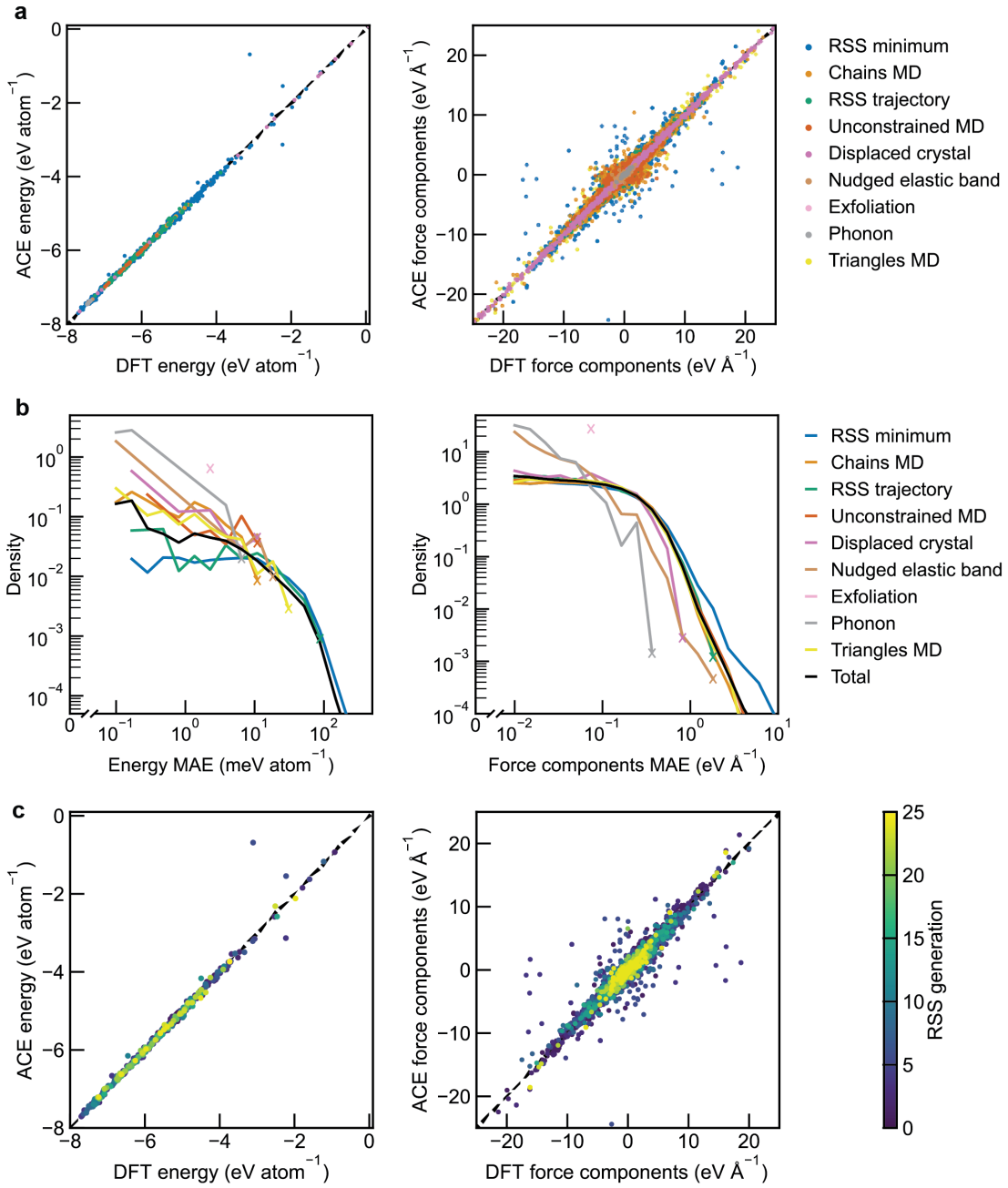


Figure 5.15: Numerical errors for energy (left) and force components (right). (**a**) Parity plot for the full dataset, coloured by configuration type. (**b**) MAE distribution curve (black) separated by configuration type (colours). Crosses indicate where distributions terminate (i.e. the largest error on a displaced crystal is approx. 10 meV). These show that the mean test set errors are dominated by the RSS structures (many of which are very high in energy). (**c**) Parity plot showing the improvement in prediction accuracy with greater number of RSS generations.

as shown by the narrowness of the yellow region in Fig. 5.15c compared to the darker colours corresponding to the early generations.

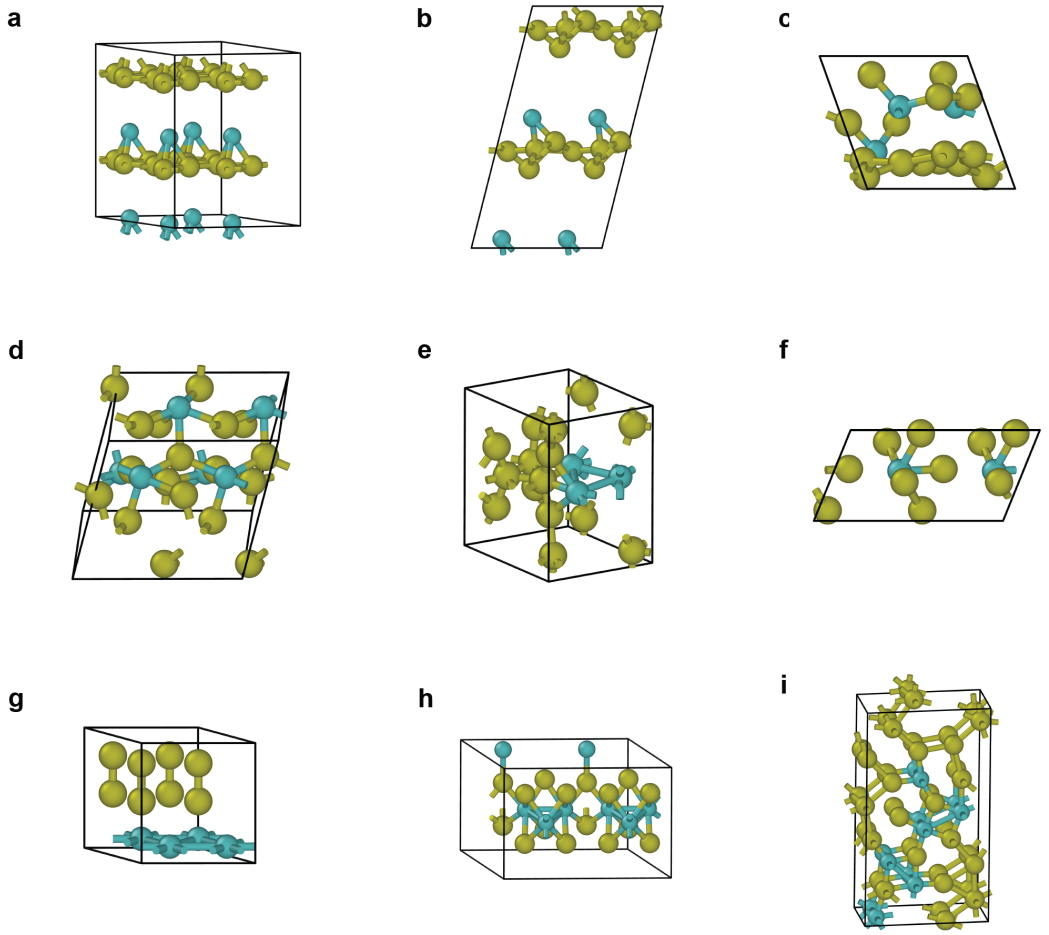


Figure 5.16: Illustrative RSS structures from energy error deciles. (**a-c**) 1st decile (lower error). (**d-f**) 4th decile. (**g-i**) 8th decile (higher error).

It is common to show isolated dimer curves when validating ML potentials:[46,394] that is, the energy as a function of bringing all possible pairs of chemical species from infinite separation to within bonding distance. Instead, we show a related plot of 'in-context' dimer curves, where two atoms are brought closer within the crystal lattice of MoS_2. Performance at this task is more relevant for avoiding the structural collapses that occur during MD of bulk systems. We find that including

isolated dimer curves for this system degrades performance of the ML potential elsewhere, on the one hand, because isolated Mo atoms bond extremely strongly (formally, a hextuple bond), and, on the other, because we struggled to converge DFT calculations for isolated S atoms approaching one another.

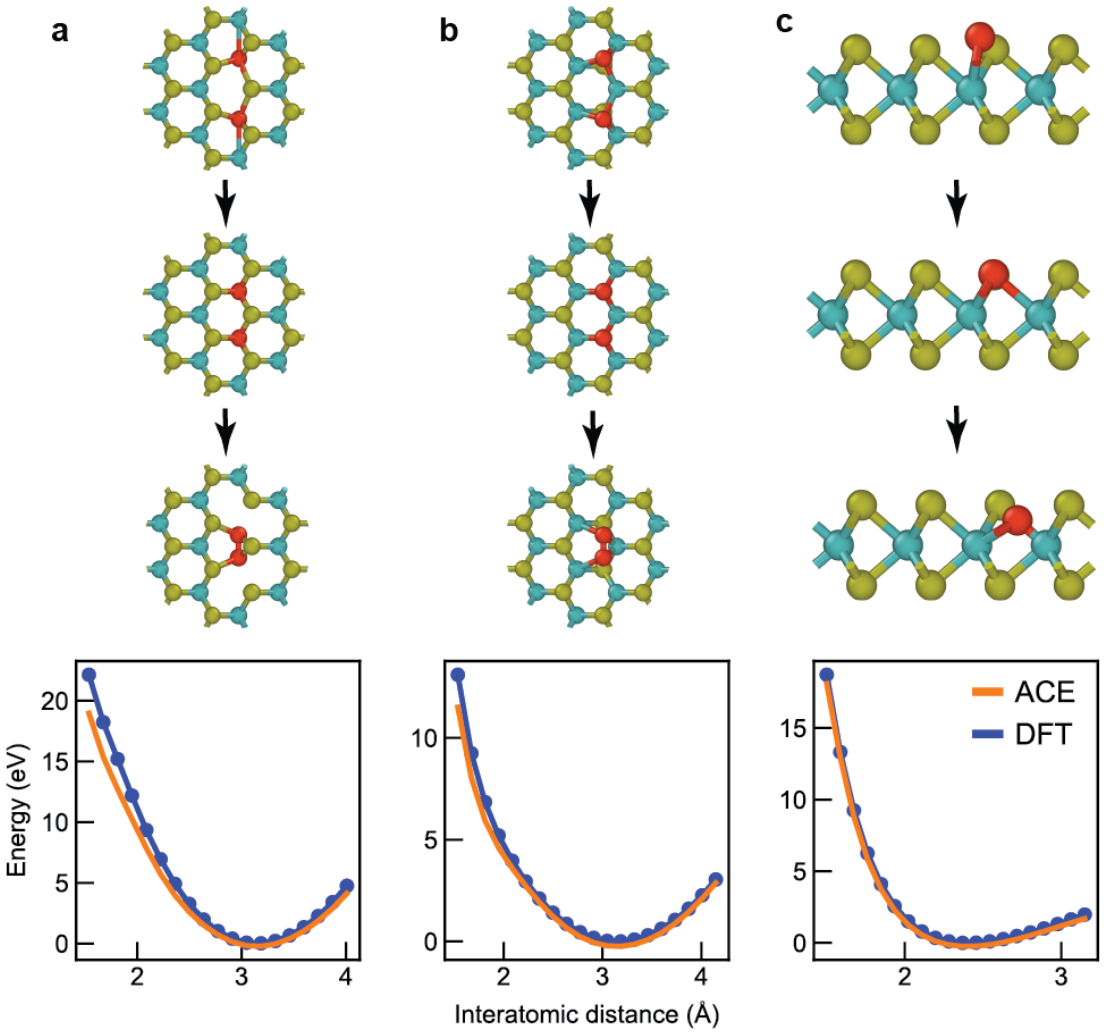


Figure 5.17: Contextualised dimer curves. For each of Mo–Mo (**a**), S–S (**b**), and Mo–S (**c**) pairs, we evaluate the total energy of a monolayer of crystalline MoS_2 where the members of such pairs are moved closer along their interatomic vector. Snapshots of the structure with the relevant atoms highlighted in red are above the plots. Other atoms are coloured teal (Mo) and yellow (S).

The accuracy of these dimer curves at small displacements, and the bulk tensile/compressive strain of Fig. 5.18, is reflected in the phonon dispersion curves

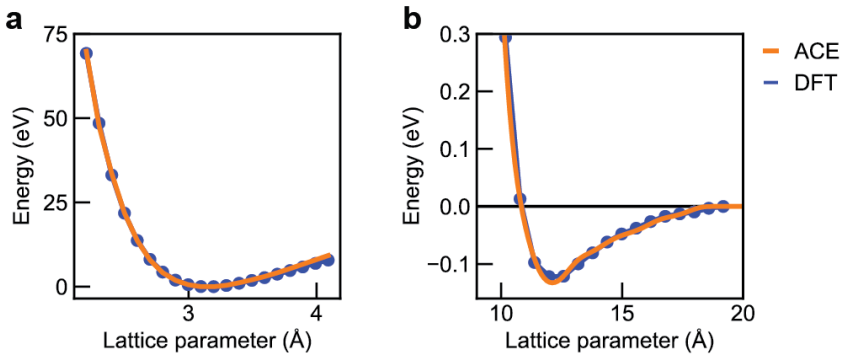


Figure 5.18: Straining MoS_2 crystals. (**a**) Energy-area curve strain within the MoS_2 plane of ACE (orange) vs DFT (blue). (**b**) Exfoliation curve from separating MoS_2 layers.

for monolayer MoS_2 (Fig. 5.19), which are competitive with other ML potentials, such as Ref. [395], and Stillinger-Weber-type potentials[396] that were parametrised specifically for the thermal properties of the MoS_2 system.

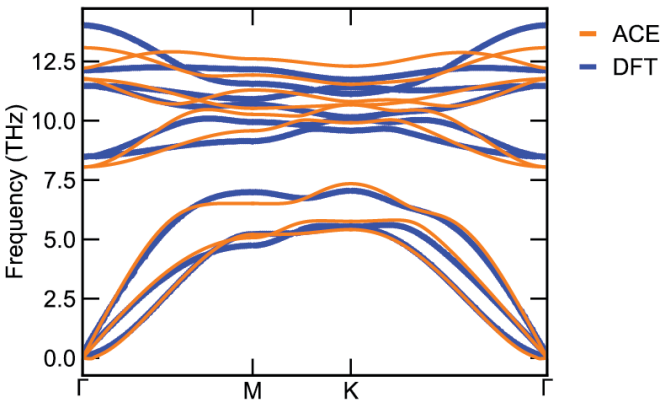


Figure 5.19: Phonon dispersion curves for monolayer MoS_2 vs. DFT (from Ref. [397]) along high symmetry lines.

The defect migration paths, as calculated with the Nudged Elastic Band method with the ACE potential,[398] are less well-captured (shown in Fig. 5.20). We show calculations for two of the most prominent defect types, described in Refs. [187, 399], which are also observed experimentally[400,401] and are the point defects with

lowest formation energies for S and Mo, respectively. In both cases, the height of the activation barrier is underestimated, although the defect formation energies are better described, with errors of 0.6 meV and 1.2 meV per atom respectively for sulfur vacancy and Mo split-interstitials, comparing favourably with a recent ML potential fitted for this purpose alone, which has an error of 0.9 meV per atom for the sulfur vacancy and did not report calculations for the split interstitial. The ACE model has seen NEB data for these migration paths, but only those produced with earlier versions of the model which *had not* seen similar data (rather than paths generated with DFT). It is likely that, with further iterations of NEB path searching, the quality of the model predictions will improve.

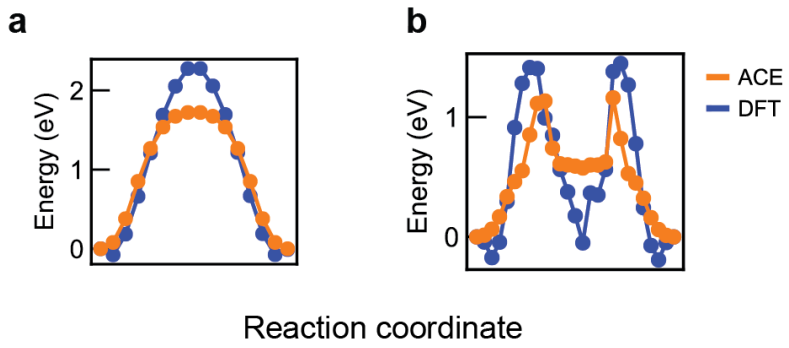


Figure 5.20: Defect migration energy barriers in MoS_2 from the Nudged Elastic Band method for (**a**) the sulfur vacancy and (**b**) an Mo split-interstitial.

A more general approach for validating ML potentials, which requires little systemic-specific know-how, can be provided by a standardised RSS protocol as described in our work in Ref. [2]. The task of relaxing random structures into their local minima, which are often high in energy, provides a stringent test for an interatomic potential. This type of test has been introduced for silicon,[45] where a range of widely-used empirical potentials fail to reproduce the distribution of the

RSS minima (in energy-volume space) as closely as an ML potential. In Fig. 5.21, we apply a similar test in which 10,000 structures are relaxed from random initial points generated by the `buildcell` code.[114] In this case, we separately relax the

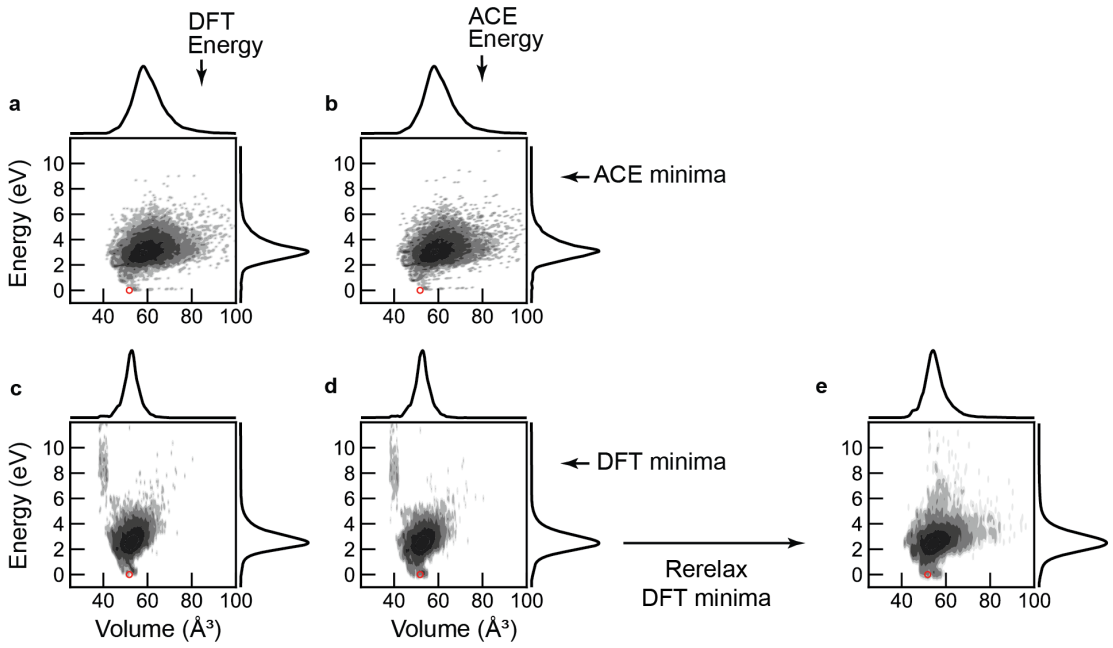


Figure 5.21: RSS validation via energy-volume distributions. (**a-d**) The upper row shows the energy and volume of RSS structures relaxed from random starting points with the ACE model, with energies evaluated with DFT (left column) and the ACE model (right column). The lower row shows similar information for DFT minima. ACE and DFT agree well on the energies of all minima, however the distributions obtained with each are quite different. The high energy, low volume region for the DFT minima is characterised by extremely short Mo–Mo bond lengths which are a result of poorly converged SCF in the DFT, so do not present a reliable ground truth. (**e**) Re-relaxing the DFT minima with the ACE model does not substantially change the distribution, except for repelling those short Mo–Mo contacts.

same initial points with the ACE potential (panels **a** and **b**) and with DFT (panels **c** and **d**), before evaluating each the single-point energies of these minima with both methods. We notice that the ACE models fail to find the high energy, low volume region and have a higher-volume tail than the DFT distribution. However, in both cases, the energies of the minima are consistent across methods, which implies that a difference in force predictions at early points in the trajectory are

likely what is behind the divergence in the final distribution of energy wells. We confirm this interpretation by 're-relaxing' the DFT minima with the ACE model (panel **e**), which shows little change compared to the original DFT distribution with the notable exception of the aforementioned high energy, low volume region. Closer inspection of the structures in this region reveals extremely short Mo – Mo bonds. These are the result of poor convergence in the DFT, which do not represent a reliable ground truth for these cases.

5.11.5 Electronic Structure and Bonding Analysis

The Löwdin population analysis[402] uses the occupied atom-centred orbitals that contribute towards the expansion of the wavefunction to estimate the number of electrons associated with each nucleus (hence, the atomic charge). Since we use a plane-wave basis to calculate the ground-state electron density, we must first project these basis functions on to an alternative atom-centred basis. We perform this with the LOBSTER code,[403] using a basis of, for Mo, 5s and 4d orbitals and, for S, 3s and 3p orbitals to match the explicit electrons in the pseudopotentials. Although improved over Mulliken charges, the Löwdin charges are still basis-set dependent, so the absolute values should be interpreted with caution. Note that there are six electrons in both 'neutral' Mo and S atoms, with the rest of the core electrons subsumed into the pseudopotential. Therefore, any deviation from six indicates the sign of partial charges on each atom, with higher being more negative. The resulting charges are shown in Fig. 5.9a, which indicate that the oxidation state of Mo in the triangle models is similar to that in crystalline MoS_2,

i.e. Mo^{IV} more than Mo^{V}. However, Mo is slightly more oxidised in our amorphous structural models than it is in MoS_2, which agrees with the formal oxidation state calculated based on the sulfur coordination numbers. The metal-metal bonds, not present in MoS_2, will also delocalise electron density into the internuclear region and hence reduce the amount projected on to Mo.

5.11.6 Synthesis

We synthesised samples of a-MoS_3 as described in Ref. [80] by heating ground $(NH_4)_2MoS_4$ (2.14 g, Sigma-Aldrich) to 150 °C at 6 °C min^{-1}, then to 215 °C at 3 °C min^{-1}, in a silica tube under dynamic vacuum for 6 hr. 1.52 g of product was recovered, compared to a theoretical yield of MoS_3 of 1.58 g. A yellow ring formed near the quartz wool stopper at the top of the tube, indicating sulfur loss at around 140 °C, at which point the powder sample was observed to move. The sample was then stored under argon to prevent oxidation.

5.11.7 Characterisation

Synchrotron X-ray and neutron diffraction confirmed that the sample was amorphous. The X-ray measurements were performed using the I15-1 beamline of the Diamond Light Source operating with an X-ray wavelength of 0.823 Å. The neutron measurements were performed using the POLARIS diffractometer at the ISIS spallation source at the Rutherford Appleton Laboratory. We also performed a variable-temperature X-ray diffraction experiment on the I11 beamline of the Diamond Light Source, the results of which are presented in Fig. 5.22. Data correction and normalisation was performed with the `GudrunN` and `GudrunX`

software packages.[404] We use a density of 3.1 g cm^{-3} and the composition MoS_3 when reducing the data.

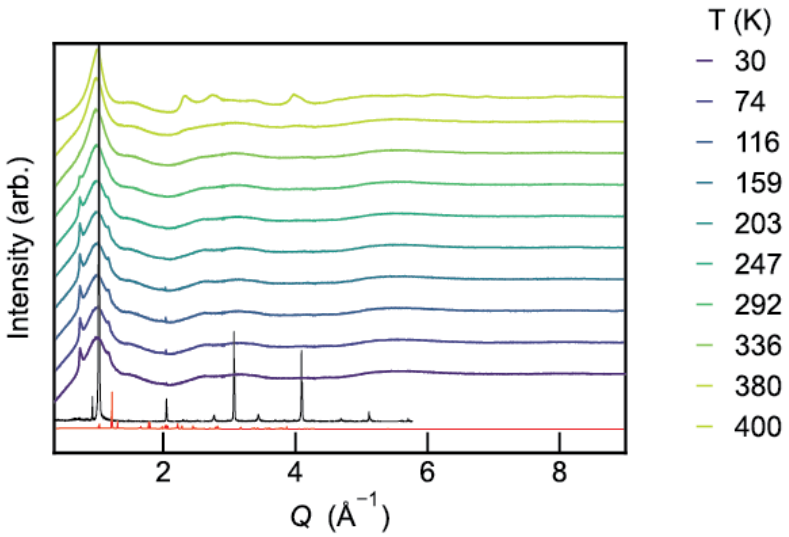


Figure 5.22: Variable-temperature X-ray total scattering. We start from crystalline $(NH_4)_2MoS_4$ and heat at a rate of 5 K min^{-1}. The powder patterns of 2H-MoS_2 (black, from Ref. [405]) and the starting material (red) are included for comparison.

5.11.8 Computational Methods

We performed MD simulations with the LAMMPS[213] software interfaced to the PACE code[394] for ML potentials and to the PLUMED code version 2.8.2[230] to enforce bonding and coordination constraints. We use the Vienna ab initio simulation package (VASP) version 6.3.1 for DFT reference calculations with the Perdew-Burke-Ernzerhof functional and -D3(BJ) dispersion correction.[170,406,407] The cutoff for plane waves was 500 eV; the Gaussian smearing width, convergence tolerance for self-consistent field iterations, and KSPACING value for automatically generating k-point grids were set to 0.1 eV, 10^{-7} eV, and 0.15 Å^{-1} respectively. The -D3 empirical dispersion correction was applied with `IVDW=12`.[407]

We prepared initial structures as illustrated schematically in Fig. 5.1. Maximally random jammed hard sphere models were produced using the code of Ref. [374]. We produced amorphous structural models with the following protocol:

- Anneal the initial structures in the NVE ensemble for 0.1 ps whilst all constraints are gradually switched on, additionally using a viscous damping force with coefficient 0.1 on the velocities.
- Melt at 2500 K for 50 ps.
- Quench for 1 ns to 300 K.
- Optimise the geometry using the conjugate-gradient algorithm as implemented in LAMMPS.

We use a Nosé–Hoover thermostat and barostat to control temperature with coefficient characteristic time 0.1 ps. The timestep was 1 fs throughout. The final structural model that we analysed in Fig. 5.11 contained 20,004 atoms (1,667 triangle units). PLUMED restraints were defined programmatically with settings for bonded Mo−Mo pairs via a `DISTANCE` command and restraints such as: `UPPER_WALLS AT=2.9 KAPPA=5.848` and `LOWER_WALLS AT=2.0 KAPPA=5.848`. The coordination number was imposed for all Mo species with a command like:

```
COORDINATIONNUMBER SWITCH=EXP D_0=3.3 R_0=0.05 D_MAX=6
MORE_THAN=RATIONAL R_0=2.0 NN=6 MM=12 D_0=0
LESS_THAN=RATIONAL R_0=2.0 NN=6 MM=12 D_0=0
```

and scheduled to switch on during equilibration with:

```
MOVINGRESTRAINT STEP0=0 AT0=0 KAPPA0=0.0 STEP1=10000 AT1=0
KAPPA1=10.0/Nat*400 STEP2=20000 AT2=0 KAPPA2=20.0/Nat*400
```

where `Nat` is the total number of atoms in the simulation.

6 | Conclusions and future work

6.1 ML potentials and their utility

During the course of my DPhil research, the lead time for developing ML potentials has been, by far, the biggest barrier to understanding more materials and more application cases. What is more, adding chemical complexity—whether through doping, or wanting to understand any reactions a material may catalyse—is trivial for DFT to handle, yet necessitates the retraining of most ML potential, or fine-tuning for foundation models, and requires reams more training data to achieve the same level of accuracy. At the same time, differences in the precise details of the implementation of DFT when labelling reference data leads to a general problem with lack of reuse of data within the field, so this development time is often duplicated even for groups in contact working on similar materials.

Nevertheless, the future is bright. Foundational ML potentials are promising to speed up initial investigations and to provide rapid routes to high-quality synthetic data. Perhaps, with enough data (and distillation), and with frameworks such as our `autoplex`, the development of ML potentials can largely be automated and massively accelerated, allowing more time for researchers to think about the chemical consequences of their models' predictions—which, after all, is the whole point. I hope that the kind of physically-motivated validation approaches for ML potentials that we have applied throughout this work will be used to ensure that the utility of these models is just as good as what any numerical error benchmarks might imply.

As crystals become ever easier to study in a high-throughput way, the design of disordered materials can also become more systematic. This could include investigations of ultra-large structural models to understand the interactions and reactions of rare defects, the use of RSS to automatically build databases with both the coverage and the specificity needed for the diverse range of environments encountered during amorphisation, and the use of modified or constrained, ML-driven MD simulations to obtain these metastable phases by circumventing the formation of stable crystals.

6.2 Amorphous materials

For silicon, work is underway in the group to study the more chemically complex, but far more industrially-relevant, hydrogenated form of the low-density amorphous phase. The details of how hydrogen could passivate the defects, and whether this modifies the fundamental chemistry of their interactions with each other (especially their aggregation), will be a fascinating development. There is also the question of paracrystallinity in amorphous silicon:[258] a fusion of the early conceptualisation of the CRN by Zachariasen with the competing theory that amorphous solids consist of nanocrystals which are too small to reveal themselves in diffraction experiments. With ML potentials, we have accessed sufficiently slow quench rates and large system sizes to observe crystal nucleation as a minor component of the LDA phase of silicon. By quenching over a range of slower rates and exploring the effect of annealing, it becomes possible to quantify the agreement that such models may have with the available experimental data. Local energies would, in that

case, permit an analysis of the strain associated with the amorphous/crystalline interface in pursuit of an understanding of why some materials support a glass phase, whereas most only crystallise.

For a-MoS_3, one can envisage future work on simulating ball milling (e.g. via shear shocks, as Ref. [62] used to model metastable amorphous water ice), and experimental investigations of repeated heat treatment to observe any modifications to the nematic order, although the question of how to measure the latter is not straightforward. If our prediction of local anisotropy in a-MoS_3 turns out to be correct, then it may be possible to build upon the surface dispersal synthesis of Ref. [369] to produce a thin film of MoS_3 with a preferred global orientation for triangles. This could enable measurements of the anisotropy, e.g. via polarisation-dependent EXAFS and in-plane/out-of-plane X-ray scattering experiments, as has been used to link structural anisotropy with magnetic anisotropy in other amorphous thin films.[408–410] Further computational studies could examine the small-scale structural models of a-MoS_3 produced in this work to understand the active sites for hydrogen activation, and, in the manner of Ref. [411], could investigate the lithium capacity in the structure, including any of the proposed rearrangements that are induced by the incorporation of Li into the structure.[358] The rapid developments in foundational models outlined above may well, in a timescale of just a couple of years, enable these chemically-complex investigations to be performed on the full 20,000-atom structure used to probe the nematic order at close to *ab initio* accuracy, in which case even mass transport effects can begin to be considered.

ML potentials have proved their utility for atomistic modelling across the periodic table. In the coming years, I expect to see their adoption continue to increase in the community—perhaps nowhere more than for studying disordered structures, where high-quality amorphous models can help us to broaden the scope of functional materials design, going beyond idealised crystalline phases.

Bibliography

(1) Morrow, J. D.; Deringer, V. L. Indirect Learning and Physically Guided Validation of Interatomic Potential Models. *J. Chem. Phys.* **2022**, *157* (10), 104105.

(2) Morrow, J. D.; Gardner, J. L. A.; Deringer, V. L. How to Validate Machine-Learned Interatomic Potentials. *J. Chem. Phys.* **2023**, *158* (12), 121501.

(3) Morrow, J. D.; Ugwumadu, C.; Drabold, D. A.; Elliott, S. R.; Goodwin, A. L.; Deringer, V. L. Understanding Defects in Amorphous Silicon with Million-Atom Simulations and Machine Learning. *Angew. Chem. Int. Ed.* **2024**, *136* (22), e202403842.

(4) Stenczel, T. K.; El-Machachi, Z.; Liepuoniute, G.; Morrow, J. D.; Bartók, A. P.; Probert, M. I.; Csányi, G.; Deringer, V. L. Machine-Learned Acceleration for Molecular Dynamics in CASTEP. *J. Chem. Phys.* **2023**, *159* (4), 044803.

(5) Chong, S.; Grasselli, F.; Ben Mahmoud, C.; Morrow, J. D.; Deringer, V. L.; Ceriotti, M. Robustness of Local Predictions in Atomistic Machine Learning Models. *J. Chem. Theory Comput.* **2023**, *19* (22), 8020–8031.

(6) Krizhevsky, A.; Sutskever, I.; Hinton, G. E., *ImageNet Classification with Deep Convolutional Neural Networks* In *Advances in Neural Information Processing Systems*, Curran Associates Inc.: 2012; Vol. 25.

(7) Graves, A.; Mohamed, A. R.; Hinton, G., *Speech Recognition with Deep Recurrent Neural Networks* In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp 6645–6649.

(8) Behler, J.; Parrinello, M. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Phys. Rev. Lett.* **2007**, *98* (14), 146401.

(9) Sumpter, B. G.; Noid, D. W. Potential Energy Surfaces for Macromolecules. A Neural Network Technique. *Chem. Phys. Lett.* **1992**, *192* (5), 455–462.

(10) Blank, T. B.; Brown, S. D.; Calhoun, A. W.; Doren, D. J. Neural Network Models of Potential Energy Surfaces. *J. Chem. Phys.* **1995**, *103*, 4129–4137.

(11) Sosso, G. C.; Miceli, G.; Caravati, S.; Behler, J.; Bernasconi, M. Neural Network Interatomic Potential for the Phase Change Material GeTe. *Phys. Rev. B* **2012**, *85* (17), 174103.

(12) Bartók, A. P.; Payne, M. C.; Kondor, R.; Csányi, G. Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, without the Electrons. *Phys. Rev. Lett.* **2010**, *104* (13), 136403.

(13) Bartók, A. P.; Kondor, R.; Csányi, G. On Representing Chemical Environments. *Phys. Rev. B* **2013**, *87* (18), 184115.

(14) Deringer, V. L.; Bernstein, N.; Bartók, A. P.; Cliffe, M. J.; Kerber, R. N.; Marbella, L. E.; Grey, C. P.; Elliott, S. R.; Csányi, G. Realistic Atomistic

Structure of Amorphous Silicon from Machine-Learning-Driven Molecular Dynamics. *J. Phys. Chem. Lett.* **2018**, *9* (11), 2879–2885.

(15) Deringer, V. L.; Caro, M. A.; Jana, R.; Aarva, A.; Elliott, S. R.; Laurila, T.; Csányi, G.; Pastewka, L. Computational Surface Chemistry of Tetrahedral Amorphous Carbon by Combining Machine Learning and Density Functional Theory. *Chem. Mater.* **2018**, *30* (21), 7438–7445.

(16) Jumper, J. et al. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596*, 583–589.

(17) Drautz, R. Atomic Cluster Expansion for Accurate and Transferable Interatomic Potentials. *Phys. Rev. B* **2019**, *99* (1), 014104.

(18) Batzner, S.; Musaelian, A.; Sun, L.; Geiger, M.; Mailoa, J. P.; Kornbluth, M.; Molinari, N.; Smidt, T. E.; Kozinsky, B. E(3)-Equivariant Graph Neural Networks for Data-Efficient and Accurate Interatomic Potentials. *Nat. Commun.* **2022**, *13*, 2453.

(19) Batatia, I.; Kovacs, D. P.; Simm, G.; Ortner, C.; Csanyi, G., *MACE: Higher Order Equivariant Message Passing Neural Networks for Fast and Accurate Force Fields* In *35th Advances in Neural Information Processing Systems*, ed. by Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; Oh, A., Curran Associates Inc.: 2022; Vol. 35, pp 11423–11436.

(20) Yang, H. et al. MatterSim: A Deep Learning Atomistic Model Across Elements, Temperatures and Pressures. *ArXiv* **2024**, (arXiv:2405.04967).

(21) Kaplan, J.; McCandlish, S.; Henighan, T.; Brown, T. B.; Chess, B.; Child, R.; Gray, S.; Radford, A.; Wu, J.; Amodei, D. Scaling Laws for Neural Language Models. *ArXiv* **2020**, (arXiv:2001.08361).

(22) Ross, J.; Belgodere, B.; Chenthamarakshan, V.; Padhi, I.; Mroueh, Y.; Das, P. Large-Scale Chemical Language Representations Capture Molecular Structure and Properties. *Nat. Mach. Intell.* **2022**, *4*, 1256–1264.

(23) Chang, J.; Ye, J. C. Bidirectional Generation of Structure and Properties through a Single Molecular Foundation Model. *Nat. Commun.* **2024**, *15*, 2323.

(24) Merchant, A.; Batzner, S.; Schoenholz, S. S.; Aykol, M.; Cheon, G.; Cubuk, E. D. Scaling Deep Learning for Materials Discovery. *Nature* **2023**, *624*, 80–85.

(25) Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; a. Persson, K. Commentary: The Materials Project: A Materials Genome Approach to Accelerating Materials Innovation. *APL Mater.* **2013**, *1* (011002).

(26) Scheidgen, M. et al. NOMAD: A Distributed Web-Based Platform for Managing Materials Science Research Data. *J. Open Source Softw.* **2023**, *8* (90), 5388.

(27) Curtarolo, S.; Setyawan, W.; Hart, G. L. W.; Jahnatek, M.; Chepulskii, R. V.; Taylor, R. H.; Wang, S.; Xue, J.; Yang, K.; Levy, O.; Mehl, M. J.;

Stokes, H. T.; Demchenko, D. O.; Morgan, D. AFLOW: An Automatic Framework for High-Throughput Materials Discovery. *Comput. Mater. Sci.* **2012**, *58*, 218–226.

(28) Kirklin, S.; Saal, J. E.; Meredig, B.; Thompson, A.; Doak, J. W.; Aykol, M.; Rühl, S.; Wolverton, C. The Open Quantum Materials Database (OQMD): Assessing the Accuracy of DFT Formation Energies. *npj Comput. Mater.* **2015**, *1* (15010).

(29) Kovács, D. P.; Moore, J. H.; Browning, N. J.; Batatia, I.; Horton, J. T.; Kapil, V.; Witt, W. C.; Magdău, I.-B.; Cole, D. J.; Csányi, G. MACE-OFF23: Transferable Machine Learning Force Fields for Organic Molecules. *ArXiv* **2023**, (arXiv:2312.15211).

(30) Anstine, D.; Zubatyuk, R.; Isayev, O. AIMNet2: A Neural Network Potential to Meet Your Neutral, Charged, Organic, and Elemental-Organic Needs. **2023**, DOI: `10.26434/chemrxiv-2023-296ch`.

(31) Batatia, I. et al. A Foundation Model for Atomistic Materials Chemistry. *ArXiv* **2024**, (arXiv:2401.00096).

(32) Zhang, D. et al. DPA-2: A Large Atomic Model as a Multi-Task Learner. *ArXiv* **2024**, (arXiv:2312.15492).

(33) Sutton, R. The Bitter Lesson `http://www.incompleteideas.net/IncIdeas/BitterLesson.html` (accessed 10/20/2024).

(34) Szymanski, N. J.; Rendy, B.; Fei, Y.; Kumar, R. E.; He, T.; Milsted, D.; McDermott, M. J.; Gallant, M.; Cubuk, E. D.; Merchant, A.; Kim, H.; Jain, A.; Bartel, C. J.; Persson, K.; Zeng, Y.; Ceder, G. An Autonomous Laboratory for the Accelerated Synthesis of Novel Materials. *Nature* **2023**, *624*, 86–91.

(35) Leeman, J.; Liu, Y.; Stiles, J.; Lee, S. B.; Bhatt, P.; Schoop, L. M.; Palgrave, R. G. Challenges in High-Throughput Inorganic Materials Prediction and Autonomous Synthesis. *PRX Energy* **2024**, *3* (1), 011002.

(36) Xyme AI `https://www.xyme.ai/` (accessed 11/05/2024).

(37) Cusp AI `https://www.cusp.ai/` (accessed 10/23/2024).

(38) Deringer, V. L.; Caro, M. A.; Csányi, G. Machine Learning Interatomic Potentials as Emerging Tools for Materials Science. *Adv. Mater.* **2019**, *31* (46), 1902765.

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

(40) Zhou, Y.; Zhang, W.; Ma, E.; Deringer, V. L. Device-Scale Atomistic Modelling of Phase-Change Memory Materials. *Nat. Electron.* **2023**, *6*, 746–754.

(41) Kostiuchenko, T.; Körmann, F.; Neugebauer, J.; Shapeev, A. Impact of Lattice Relaxations on Phase Transitions in a High-Entropy Alloy Studied by Machine-Learning Potentials. *npj Comput. Mater.* **2019**, *5*, 55.

(42) Elliott, S. R. Medium-Range Structural Order in Covalent Amorphous Solids. *Nature* **1991**, *354*, 445–452.

(43) Niu, H.; Bonati, L.; Piaggi, P. M.; Parrinello, M. Ab Initio Phase Diagram and Nucleation of Gallium. *Nat. Commun.* **2020**, *11*, 2654.

(44) Deringer, V. L.; Caro, M. A.; Csányi, G. A General-Purpose Machine-Learning Force Field for Bulk and Nanostructured Phosphorus. *Nat. Commun.* **2020**, *11*, 5461.

(45) Bartók, A. P.; Kermode, J.; Bernstein, N.; Csányi, G. Machine Learning a General-Purpose Interatomic Potential for Silicon. *Phys. Rev. X* **2018**, *8* (4), 041048.

(46) Deringer, V. L.; Csányi, G. Machine Learning Based Interatomic Potential for Amorphous Carbon. *Phys. Rev. B* **2017**, *95* (9), 094203.

(47) Qamar, M.; Mrovec, M.; Lysogorskiy, Y.; Bochkarev, A.; Drautz, R. Atomic Cluster Expansion for Quantum-Accurate Large-Scale Simulations of Carbon. *J. Chem. Theory Comput.* **2023**, *19* (15), 5151–5167.

(48) Erhard, L. C.; Rohrer, J.; Albe, K.; Deringer, V. L. A Machine-Learned Interatomic Potential for Silica and Its Relation to Empirical Models. *npj Comput. Mater.* **2022**, *8*, 90.

(49) Staacke, C. G.; Huss, T.; Margraf, J. T.; Reuter, K.; Scheurer, C. Tackling Structural Complexity in Li2S-P2S5 Solid-State Electrolytes Using Machine Learning Potentials. *Nanomater.* **2022**, *12* (17), 2950.

(50) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. Equation of State Calculations by Fast Computing Machines. *J. Chem. Phys.* **1953**, *21* (6), 1087–1092.

(51) Polk, D. E. Structural Model for Amorphous Silicon and Germanium. *J. Non-Cryst. Solids* **1971**, *5* (5), 365–376.

(52) Bennett, C. H. Serially Deposited Amorphous Aggregates of Hard Spheres. *J. Appl. Phys.* **1972**, *43* (6), 2727–2734.

(53) Duffy, M. G.; Boudreaux, D. S.; Polk, D. E. Systematic Generation of Random Networks. *J. Non-Cryst. Solids* **1974**, *15* (3), 435–454.

(54) Barker, J. A.; Hoare, M. R.; Finney, J. L. Relaxation of the Bernal Model. *Nature* **1975**, *257*, 120–122.

(55) Zachariasen, W. H. The Atomic Arrangement in Glass. *J. Am. Chem. Soc.* **1932**, *54* (10), 3841–3851.

(56) Laaziri, K.; Kycia, S.; Roorda, S.; Chicoine, M.; Robertson, J. L.; Wang, J.; Moss, S. C. High-Energy x-Ray Diffraction Study of Pure Amorphous Silicon. *Phys. Rev. B* **1999**, *60* (19), 13520–13533.

(57) Shao, W.-L.; Shinar, J.; Gerstein, B. C.; Li, F.; Lannin, J. S. Magic-Angle Spinning ^{29}Si NMR Study of Short-Range Order in a-Si. *Phys. Rev. B* **1990**, *41* (13), 9491–9494.

(58) Zeng, Z.; Yang, L.; Zeng, Q.; Lou, H.; Sheng, H.; Wen, J.; Miller, D. J.; Meng, Y.; Yang, W.; Mao, W. L.; Mao, H.-k. Synthesis of Quenchable Amorphous Diamond. *Nat. Commun.* **2017**, *8*, 322.

(59) Shang, Y.; Liu, Z.; Dong, J.; Yao, M.; Yang, Z.; Li, Q.; Zhai, C.; Shen, F.; Hou, X.; Wang, L.; Zhang, N.; Zhang, W.; Fu, R.; Ji, J.; Zhang, X.; Lin, H.; Fei, Y.; Sundqvist, B.; Wang, W.; Liu, B. Ultrahard Bulk Amorphous Carbon from Collapsed Fullerene. *Nature* **2021**, *599*, 599–604.

(60) Zhou, Y.; Elliott, S. R.; Deringer, V. L. Structure and Bonding in Amorphous Red Phosphorus. *Angew. Chem. Int. Ed.* **2023**, *62* (24), e202216658.

(61) Auvray, T.; Friščić, T. Shaking Things from the Ground-Up: A Systematic Overview of the Mechanochemistry of Hard and High-Melting Inorganic Materials. *Molecules* **2023**, *28* (2), 897.

(62) Rosu-Finsen, A.; Davies, M. B.; Amon, A.; Wu, H.; Sella, A.; Michaelides, A.; Salzmann, C. G. Medium-Density Amorphous Ice. *Science* **2023**, *379* (6631), 474–478.

(63) Gorecki, A.; Liu, A. C.; Petersen, T. C. High-Resolution Radial Distribution Function of Pure Ion-Implanted Amorphous Silicon Measured Using Tilted-Illumination Selected-Area Electron Diffraction. *Microsc. Microanal.* **2014**, *20* (1), 50–54.

(64) Laaziri, K.; Kycia, S.; Roorda, S.; Chicoine, M.; Robertson, J. L.; Wang, J.; Moss, S. C. High Resolution Radial Distribution Function of Pure Amorphous Silicon. *Phys. Rev. Lett.* **1999**, *82* (17), 3460–3463.

(65) Filtvedt, W. O.; Holt, A.; Ramachandran, P. A.; Melaaen, M. C. Chemical Vapor Deposition of Silicon from Silane: Review of Growth Mechanisms and Modeling/Scaleup of Fluidized Bed Reactors. *Sol. Energy Mater. Sol. Cells* **2012**, *107*, 188–200.

(66) Roorda, S.; Sinke, W. C.; Poate, J. M.; Jacobson, D. C.; Dierker, S.; Dennis, B. S.; Eaglesham, D. J.; Spaepen, F.; Fuoss, P. Structural Relaxation and Defect Annihilation in Pure Amorphous Silicon. *Phys. Rev. B* **1991**, *44* (8), 3702–3725.

(67) Urli, X.; Dias, C. L.; Lewis, L. J.; Roorda, S. Point Defects in Pure Amorphous Silicon and Their Role in Structural Relaxation: A Tight-Binding Molecular-Dynamics Study. *Phys. Rev. B* **2008**, *77* (15), 155204.

(68) Wooten, F.; Winer, K.; Weaire, D. Computer Generation of Structural Models of Amorphous Si and Ge. *Phys. Rev. Lett.* **1985**, *54* (13), 1392–1395.

(69) Wooten, F.; Weaire, D. In *Solid State Physics*; Academic Press: New York, 1987; Vol. 40, pp 1–42.

(70) Keating, P. N. Effect of Invariance Requirements on the Elastic Strain Energy of Crystals with Application to the Diamond Structure. *Phys. Rev.* **1966**, *145* (2), 637–645.

(71) Barkema, G. T.; Mousseau, N. High-Quality Continuous Random Networks. *Phys. Rev. B* **2000**, *62* (8), 4985–4990.

(72) McGreevy, R. L.; Pusztai, L. Reverse Monte Carlo Simulation: A New Technique for the Determination of Disordered Structures. *Mol. Sim.* **1988**, *1* (6), 359–367.

(73) Coates, C. S.; Baise, M.; Schmutzler, A.; Simonov, A.; Makepeace, J. W.; Seel, A. G.; Smith, R. I.; Playford, H. Y.; Keen, D. A.; Siegel, R.; Senker, J.; Slater, B.; Goodwin, A. L. Spin-Ice Physics in Cadmium Cyanide. *Nat. Commun.* **2021**, *12*, 2272.

(74) Cliffe, M. J.; Bartók, A. P.; Kerber, R. N.; Grey, C. P.; Csányi, G.; Goodwin, A. L. Structural Simplicity as a Restraint on the Structure of Amorphous Silicon. *Phys. Rev. B* **2017**, *95* (22), 224108.

(75) Soper, A. K. Partial Structure Factors from Disordered Materials Diffraction Data: An Approach Using Empirical Potential Structure Refinement. *Phys. Rev. B* **2005**, *72* (10), 104204.

(76) Pandey, A.; Biswas, P.; Drabold, D. A. Force-Enhanced Atomic Refinement: Structural Modeling with Interatomic Forces in a Reverse Monte Carlo Approach Applied to Amorphous Si and SiO_2. *Phys. Rev. B* **2015**, *92* (15), 155205.

(77) Ye, H.; Ma, L.; Zhou, Y.; Wang, L.; Han, N.; Zhao, F.; Deng, J.; Wu, T.; Li, Y.; Lu, J. Amorphous MoS_3 as the Sulfur-Equivalent Cathode Material for Room-Temperature Li–S and Na–S Batteries. *Proc. Natl. Acad. Sci. U.S. A.* **2017**, *114* (50), 13091–13096.

(78) Pazhamalai, P.; Krishnamoorthy, K.; Sahoo, S.; Mariappan, V. K.; Kim, S. J. Supercapacitive Properties of Amorphous MoS_3 and Crystalline MoS_2 Nanosheets in an Organic Electrolyte. *Inorg. Chem. Front.* **2019**, *6* (9), 2387–2395.

(79) Merki, D.; Fierro, S.; Vrubel, H.; Hu, X. Amorphous Molybdenum Sulfide Films as Catalysts for Electrochemical Hydrogen Production in Water. *Chem. Sci.* **2011**, *2* (7), 1262–1267.

(80) Hibble, S. J.; Rice, D. A.; Pickup, D. M.; Beer, M. P. Mo K-edge EXAFS and S K-edge Absorption Studies of the Amorphous Molybdenum Sulfides $MoS_{4.7}$, MoS_3, and $MoS_3 \cdot nH_2O$ ($n \approx 2$). *Inorg. Chem.* **1995**, *34* (21), 5109–5113.

(81) Shirota, G.; Nasu, A.; Deguchi, M.; Sakuda, A.; Tatsumisago, M.; Hayashi, A. Electrode Performance of Amorphous MoS_3 in All-Solid-State Sodium Secondary Batteries. *J. Power Sources Adv.* **2021**, *10*, 100061.

(82) Weber, T.; Muijsers, J. C.; Niemantsverdriet, J. W. Structure of Amorphous MoS_3. *J. Phys. Chem.* **1995**, *99* (22), 9194–9200.

(83) Bhattacharya, R. N.; Lee, C. Y.; Pollak, F. H.; Schleich, D. M. Optical Study of Amorphous MoS_3: Determination of the Fundamental Energy Gap. *J. Non-Cryst. Solids* **1987**, *91* (2), 235–242.

(84) Zhang, L.; Wu, L.; Li, J.; Lei, J. Electrodeposition of Amorphous Molybdenum Sulfide Thin Film for Electrochemical Hydrogen Evolution Reaction. *BMC chem.* **2019**, *13*, 88.

(85) Liang, K. S.; Cramer, S. P.; Johnston, D. C.; Chang, C. H.; Jacobson, A. J.; deNeufville, J. P.; Chianelli, R. R. Amorphous MoS_3 and WS_3. *J. Non-Cryst. Solids* **1980**, *42* (1-3), 345–356.

(86) Hibble, S. J.; Walton, R. I.; Pickup, D. M.; Hannon, A. C. Amorphous MoS_3: Clusters or Chains? The Structural Evidence. *J. Non-Cryst. Solids* **1998**, *232–234*, 434–439.

(87) Müller, A.; Sarkar, S.; Bhattacharyya, R. G.; Pohl, S.; Dartmann, M. Directed Synthesis of $[Mo_3S_{13}]^{2-}$, an Isolated Cluster Containing Sulfur Atoms in Three Different States of Bonding. *Angew. Chem. Int. Ed.* **1978**, *17* (7), 535–535.

(88) Batool, S.; Langer, M.; Myakala, S. N.; Heiland, M.; Eder, D.; Streb, C.; Cherevan, A. Thiomolybdate Clusters: From Homogeneous Catalysis to Heterogenization and Active Sites. *Adv. Mater.* **2024**, *36* (7), 2305730.

(89) Bartók, A. P.; Csányi, G. Gaussian Approximation Potentials: A Brief Tutorial Introduction. *Int. J. Quantum Chem.* **2015**, *115* (16), 1051–1057.

(90) Shapeev, A. Moment Tensor Potentials: A Class of Systematically Improvable Interatomic Potentials. *Multiscale Model. Simul.* **2016**, *14* (3), 1153–1173.

(91) Bernstein, N.; Csányi, G.; Deringer, V. L. De Novo Exploration and Self-Guided Learning of Potential-Energy Surfaces. *npj Comput. Mater.* **2019**, *5*, 99.

(92) Stukowski, A. Visualization and Analysis of Atomistic Simulation Data with OVITO–the Open Visualization Tool. *Model. Simul. Mater. Sci. Eng.* **2009**, *18*, 015012.

(93) Larsen, A. H. et al. The Atomic Simulation Environment—a Python Library for Working with Atoms. *J. Phys.: Condens. Matter* **2017**, *29* (27), 273002.

(94) Hagberg, A. A.; Schult, D. A.; Swart, P. J., *Exploring Network Structure, Dynamics, and Function Using NetworkX* In *Proceedings of the 7th Python in Science Conference*, ed. by Varoquaux, G.; Vaught, T.; Millman, J., Pasadena, CA USA, 2008, pp 11–15.

(95) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

(96) Podryabinkin, E. V.; Shapeev, A. V. Active Learning of Linearly Parametrized Interatomic Potentials. *Comput. Mater. Sci.* **2017**, *140*, 171–180.

(97) Zhang, L.; Lin, D.-Y.; Wang, H.; Car, R.; E, W. Active Learning of Uniformly Accurate Interatomic Potentials for Materials Simulation. *Phys. Rev. Mater.* **2019**, *3* (2), 023804.

(98) Vandermause, J.; Torrisi, S. B.; Batzner, S.; Xie, Y.; Sun, L.; Kolpak, A. M.; Kozinsky, B. On-the-Fly Active Learning of Interpretable Bayesian Force Fields for Atomistic Rare Events. *npj Comput. Mater.* **2020**, *6*, 20.

(99) van der Oord, C.; Sachs, M.; Kovács, D. P.; Ortner, C.; Csányi, G. Hyperactive Learning for Data-Driven Interatomic Potentials. *npj Comput. Mater.* **2023**, *9*, 168.

(100) Deringer, V. L.; Bartók, A. P.; Bernstein, N.; Wilkins, D. M.; Ceriotti, M.; Csányi, G. Gaussian Process Regression for Materials and Molecules. *Chem. Rev.* **2021**, *121* (16), 10073–10141.

(101) Morgan, W. S.; Jorgensen, J. J.; Hess, B. C.; Hart, G. L. W. Efficiency of Generalized Regular K-Point Grids. *Comput. Mater. Sci.* **2018**, *153*, 424–430.

(102) Goswami, S.; Käser, S.; Bemish, R. J.; Meuwly, M. Effects of Aleatoric and Epistemic Errors in Reference Data on the Learnability and Quality of NN-based Potential Energy Surfaces. *Artificial Intelligence Chemistry* **2024**, *2* (1), 100033.

(103) Bayerl, D.; Andolina, C. M.; Dwaraknath, S.; Saidi, W. A. Convergence Acceleration in Machine Learning Potentials for Atomistic Simulations. *Digital Discovery* **2022**, *1*, 61–69.

(104) Behler, J. Neural Network Potential-Energy Surfaces in Chemistry: A Tool for Large-Scale Simulations. *Phys. Chem. Chem. Phys.* **2011**, *13* (40), 17930–17955.

(105) Kang, P.-L.; Shang, C.; Liu, Z.-P. Large-Scale Atomic Simulation via Machine Learning Potentials Constructed by Global Potential Energy Surface Exploration. *Acc. Chem. Res.* **2020**, *53* (10), 2119–2129.

(106) Dicks, L.; Pyzer-Knapp, E. O. TopSearch: A Python Package for Topographical Analysis of Machine Learning Models and Physical Systems. *J. Open Source Softw.* **2024**, *9* (99), 6711.

(107) Mosquera-Lois, I.; Kavanagh, S. R.; Ganose, A. M.; Walsh, A. Machine-Learning Structural Reconstructions for Accelerated Point Defect Calculations. *npj Comput. Mater.* **2024**, *10*, 121.

(108) A. Gardner, J. L.; Beaulieu, Z. F.; L. Deringer, V. Synthetic Data Enable Experiments in Atomistic Machine Learning. *Digital Discovery* **2023**, *2* (3), 651–662.

(109) Jinnouchi, R.; Miwa, K.; Karsai, F.; Kresse, G.; Asahi, R. On-the-Fly Active Learning of Interatomic Potentials for Large-Scale Atomistic Simulations. *J. Phys. Chem. Lett.* **2020**, *11*, 6946–6955.

(110) Bichelmaier, S.; Carrete, J.; Madsen, G. K. H. Neural Network Enabled Molecular Dynamics Study of HfO_2 Phase Transitions. *Phys. Rev. B* **2024**, *110* (17), 174105.

(111) Goryaeva, A. M.; Dérès, J.; Lapointe, C.; Grigorev, P.; Swinburne, T. D.; Kermode, J. R.; Ventelon, L.; Baima, J.; Marinica, M.-C. Efficient and Transferable Machine Learning Potentials for the Simulation of Crystal Defects in Bcc Fe and W. *Phys. Rev. Mater.* **2021**, *5* (10), 103803.

(112) Deringer, V. L.; Proserpio, D. M.; Csányi, G.; Pickard, C. J. Data-Driven Learning and Prediction of Inorganic Crystal Structures. *Faraday Discuss.* **2018**, *211*, 45–59.

(113) Pickard, C. J.; Needs, R. J. High-Pressure Phases of Silane. *Phys. Rev. Lett.* **2006**, *97*, 045504.

(114) Pickard, C. J.; Needs, R. J. Ab Initio Random Structure Searching. *J. Phys.: Condens. Matter* **2011**, *23*, 053201.

(115) Rowe, P.; Deringer, V. L.; Gasparotto, P.; Csányi, G.; Michaelides, A. Erratum: "An Accurate and Transferable Machine Learning Potential for Carbon" [J. Chem. Phys. 153, 034702 (2020)]. *J. Chem. Phys.* **2022**, *156* (15), 159901.

(116) Sikka, S. K.; Vohra, Y. K.; Chidambaram, R. Omega Phase in Materials. *Prog. Mater. Sci.* **1982**, *27* (3), 245–310.

(117) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Phys. Rev. Lett.* **2012**, *108* (5), 058301.

(118) Behler, J. Atom-Centered Symmetry Functions for Constructing High-Dimensional Neural Network Potentials. *J. Chem. Phys.* **2011**, *134* (7), 074106.

(119) Thompson, A. P.; Swiler, L. P.; Trott, C. R.; Foiles, S. M.; Tucker, G. J. Spectral Neighbor Analysis Method for Automated Generation of Quantum-Accurate Interatomic Potentials. *J. Comput. Phys.* **2015**, *285*, 316–330.

(120) Schütt, K. T.; Sauceda, H. E.; Kindermans, P.-J.; Tkatchenko, A.; Müller, K.-R. SchNet – A Deep Learning Architecture for Molecules and Materials. *J. Chem. Phys.* **2018**, *148* (24), 241722.

(121) Zubatyuk, R.; Smith, J. S.; Leszczynski, J.; Isayev, O. Accurate and Transferable Multitask Prediction of Chemical Properties with an Atoms-in-Molecules Neural Network. *Sci. Adv.* **2019**, *5* (8), eaav6490.

(122) Chen, C.; Ong, S. P. A Universal Graph Deep Learning Interatomic Potential for the Periodic Table. *Nat. Comput. Sci.* **2022**, *2*, 718–728.

(123) Musil, F.; Grisafi, A.; Bartók, A. P.; Ortner, C.; Csányi, G.; Ceriotti, M. Physics-Inspired Structural Representations for Molecules and Materials. *Chem. Rev.* **2021**, *121* (16), 9759–9815.

(124) Drautz, R. Atomic Cluster Expansion of Scalar, Vectorial, and Tensorial Properties Including Magnetism and Charge Transfer. *Phys. Rev. B* **2020**, *102* (2), 024104.

(125) De, S.; Bartók, A. P.; Csányi, G.; Ceriotti, M. Comparing Molecules and Solids across Structural and Alchemical Space. *Phys. Chem. Chem. Phys.* **2016**, *18* (20), 13754–13769.

(126) Darby, J. P.; Kermode, J. R.; Csányi, G. Compressing Local Atomic Neighbourhood Descriptors. *npj Comput. Mater.* **2022**, *8*, 166.

(127) Pozdnyakov, S. N.; Willatt, M. J.; Bartók, A. P.; Ortner, C.; Csányi, G.; Ceriotti, M. Incompleteness of Atomic Structure Representations. *Phys. Rev. Lett.* **2020**, *125* (16), 166001.

(128) Rosenbrock, C. W.; Gubaev, K.; Shapeev, A. V.; Pártay, L. B.; Bernstein, N.; Csányi, G.; Hart, G. L. W. Machine-Learned Interatomic Potentials for Alloys and Alloy Phase Diagrams. *npj Comput. Mater.* **2021**, *7*, 24.

(129) Novikov, I. S.; Suleimanov, Y. V.; Shapeev, A. V. Automated Calculation of Thermal Rate Coefficients Using Ring Polymer Molecular Dynamics and Machine-Learning Interatomic Potentials with Active Learning. *Phys. Chem. Chem. Phys.* **2018**, *20* (46), 29503–29512.

(130) Wang, C.; Aoyagi, K.; Wisesa, P.; Mueller, T. Lithium Ion Conduction in Cathode Coating Materials from On-the-Fly Machine Learning. *Chem. Mater.* **2020**, *32* (9), 3741–3752.

(131) Novikov, I. S.; Gubaev, K.; Podryabinkin, E. V.; Shapeev, A. V. The MLIP Package: Moment Tensor Potentials with MPI and Active Learning. *Mach. Learn.: Sci. Technol.* **2020**, *2* (2), 025002.

(132) Lysogorskiy, Y.; van der Oord, C.; Bochkarev, A.; Menon, S.; Rinaldi, M.; Hammerschmidt, T.; Mrovec, M.; Thompson, A.; Csányi, G.; Ortner, C.; Drautz, R. Performant Implementation of the Atomic Cluster Expansion (PACE) and Application to Copper and Silicon. *npj Comput. Mater.* **2021**, *7*, 97.

(133) Sanchez, J. M.; Ducastelle, F.; Gratias, D. Generalized Cluster Description of Multicomponent Systems. *Physica A Stat.* **1984**, *128* (1), 334–350.

(134) Dusson, G.; Bachmayr, M.; Csányi, G.; Drautz, R.; Etter, S.; van der Oord, C.; Ortner, C. Atomic Cluster Expansion: Completeness, Efficiency and Stability. *J. Comput. Phys.* **2022**, *454*, 110946.

(135) Chmiela, S.; Tkatchenko, A.; Sauceda, H. E.; Poltavsky, I.; Schütt, K. T.; Müller, K.-R. Machine Learning of Accurate Energy-Conserving Molecular Force Fields. *Sci. Adv.* **2017**, *3* (5), e1603015.

(136) Chmiela, S.; Sauceda, H. E.; Müller, K.-R.; Tkatchenko, A. Towards Exact Molecular Dynamics Simulations with Machine-Learned Force Fields. *Nat. Commun.* **2018**, *9*, 3887.

(137) Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1: An Extensible Neural Network Potential with DFT Accuracy at Force Field Computational Cost. *Chem. Sci.* **2017**, *8* (4), 3192–3203.

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

(139) Zeng, J. et al. DeePMD-kit v2: A Software Package for Deep Potential Models. *J. Chem. Phys.* **2023**, *159* (5), 054801.

(140) Thomas du Toit, D. F.; Deringer, V. L. Cross-Platform Hyperparameter Optimization for Machine Learning Interatomic Potentials. *J. Chem. Phys.* **2023**, *159* (2), 024803.

(141) Tian, Y.; Zhang, Y. A Comprehensive Survey on Regularization Strategies in Machine Learning. *Information Fusion* **2022**, *80*, 146–166.

(142) Prodan, E.; Kohn, W. Nearsightedness of Electronic Matter. *Proc. Natl. Acad. Sci. U.S. A.* **2005**, *102* (33), 11635–11638.

(143) Yang, W. Direct Calculation of Electron Density in Density-Functional Theory. *Phys. Rev. Lett.* **1991**, *66* (11), 1438–1441.

(144) Kohn, W. Density Functional and Density Matrix Method Scaling Linearly with the Number of Atoms. *Phys. Rev. Lett.* **1996**, *76* (17), 3168–3171.

(145) White, C. A.; Johnson, B. G.; Gill, P. M. W.; Head-Gordon, M. Linear Scaling Density Functional Calculations via the Continuous Fast Multipole Method. *Chem. Phys. Lett.* **1996**, *253* (3), 268–278.

(146) Baer, R.; Head-Gordon, M. Chebyshev Expansion Methods for Electronic Structure Calculations on Large Molecular Systems. *J. Chem. Phys.* **1997**, *107* (23), 10003–10013.

(147) Ochsenfeld, C.; White, C. A.; Head-Gordon, M. Linear and Sublinear Scaling Formation of Hartree–Fock-type Exchange Matrices. *J. Chem. Phys.* **1998**, *109* (5), 1663–1669.

(148) Goedecker, S. Linear Scaling Electronic Structure Methods. *Rev. Mod. Phys.* **1999**, *71* (4), 1085–1123.

(149) Klawohn, S.; Darby, J. P.; Kermode, J. R.; Csányi, G.; Caro, M. A.; Bartók, A. P. Gaussian Approximation Potentials: Theory, Software Implementation and Application Examples. *J. Chem. Phys.* **2023**, *159* (17), 174108.

(150) Rowe, P.; Csányi, G.; Alfè, D.; Michaelides, A. Development of a Machine Learning Potential for Graphene. *Phys. Rev. B* **2018**, *97* (5), 054303.

(151) Gillan, M. J.; Alfè, D.; Bartók, A. P.; Csányi, G. First-Principles Energetics of Water Clusters and Ice: A Many-Body Analysis. *J. Chem. Phys.* **2013**, *139* (24), 244504.

(152) Fletcher, R. In *Practical Methods of Optimization*; John Wiley & Sons, Ltd: New York, 2000; Chapter 3, pp 44–79.

(153) Van der Oord, C.; Dusson, G.; Csányi, G.; Ortner, C. Regularised Atomic Body-Ordered Permutation-Invariant Polynomials for the Construction of Interatomic Potentials. *Mach. Learn.: Sci. Technol.* **2020**, *1* (1), 015004.

(154) Drautz, R.; Hammerschmidt, T.; Čák, M.; Pettifor, D. G. Bond-Order Potentials: Derivation and Parameterization for Refractory Elements. *Modelling Simul. Mater. Sci. Eng.* **2015**, *23* (7), 074004.

(155) Heine, V.; Robertson, I. J.; Payne, M. C.; Murrell, J. N.; Phillips, J. C.; Weaire, D. L.; Haydock, R.; Inglesfield, J. E.; Pendry, J. B. Many-Atom Interactions in Solids. *Philos. Trans. R. Soc. A* **1997**, *334* (1635), 393–405.

(156) Finnis, M. W.; Sinclair, J. E. A Simple Empirical N-Body Potential for Transition Metals. *Philos. Mag. A* **1984**, *50* (1), 45–55.

(157) Broughton, J. Q.; Gilmer, G. H.; Jackson, K. A. Crystallization Rates of a Lennard-Jones Liquid. *Phys. Rev. Lett.* **1982**, *49* (20), 1496–1500.

(158) Zhang, Y. N.; Wang, L.; Morioka, S.; Wang, W. M. Modified Lennard-Jones Potentials for Cu and Ag Based on the Dense Gaslike Model of Viscosity for Liquid Metals. *Phys. Rev. B* **2007**, *75* (1), 014106.

(159) Turney, J. E.; Landry, E. S.; McGaughey, A. J. H.; Amon, C. H. Predicting Phonon Properties and Thermal Conductivity from Anharmonic Lattice Dynamics Calculations and Molecular Dynamics Simulations. *Phys. Rev. B* **2009**, *79* (6), 064301.

(160) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Big Data Meets Quantum Chemistry Approximations: The Δ-Machine Learning Approach. *J. Chem. Theory Comput.* **2015**, *11* (5), 2087–2096.

(161) Kang, S.; Jeong, W.; Hong, C.; Hwang, S.; Yoon, Y.; Han, S. Accelerated Identification of Equilibrium Structures of Multicomponent Inorganic Crystals Using Machine Learning Potentials. *npj Comput. Mater.* **2022**, *8*, 108.

(162) Martin, R., *Electronic Structure: Basic Theory and Practical Methods*; Cambridge University Press: Cambridge, 2004.

(163) Novikov, I.; Grabowski, B.; Körmann, F.; Shapeev, A. Magnetic Moment Tensor Potentials for Collinear Spin-Polarized Materials Reproduce Different Magnetic States of Bcc Fe. *npj Comput. Mater.* **2022**, *8*, 13.

(164) Kirkpatrick, J.; McMorrow, B.; Turban, D. H. P.; Gaunt, A. L.; Spencer, J. S.; Matthews, A. G. D. G.; Obika, A.; Thiry, L.; Fortunato, M.; Pfau, D.; Castellanos, L. R.; Petersen, S.; Nelson, A. W. R.; Kohli, P.; Mori-Sánchez, P.; Hassabis, D.; Cohen, A. J. Pushing the Frontiers of Density Functionals by Solving the Fractional Electron Problem. *Science* **2021**, *374* (6573), 1385–1389.

(165) Kim, J. et al. QMCPACK: An Open Source Ab Initio Quantum Monte Carlo Package for the Electronic Structure of Atoms, Molecules and Solids. *J. Phys.: Condens. Matter* **2018**, *30* (19), 195901.

(166) Hepting, M. et al. Electronic Structure of the Parent Compound of Superconducting Infinite-Layer Nickelates. *Nat. Mater.* **2020**, *19* (4), 381–385.

(167) Liu, H.-R.; Xiang, H.; Gong, X. G. First Principles Study of Adsorption of O2 on Al Surface with Hybrid Functionals. *J. Chem. Phys.* **2011**, *135* (21), 214702.

(168) Burke, K.; Perdew, J. P.; Ernzerhof, M. Why the Generalized Gradient Approximation Works and How to Go beyond It. *Int. J. Quantum Chem.* **1997**, *61* (2), 287–293.

(169) Perdew, J. P.; Wang, Y. Accurate and Simple Analytic Representation of the Electron-Gas Correlation Energy. *Phys. Rev. B* **1992**, *45* (23), 13244–13249.

(170) Perdew, J. P.; Chevary, J. A.; Vosko, S. H.; Jackson, K. A.; Pederson, M. R.; Singh, D. J.; Fiolhais, C. Atoms, Molecules, Solids, and Surfaces: Applications of the Generalized Gradient Approximation for Exchange and Correlation. *Phys. Rev. B* **1992**, *46* (11), 6671–6687.

(171) Mardirossian, N.; Head-Gordon, M. Thirty Years of Density Functional Theory in Computational Chemistry: An Overview and Extensive Assessment of 200 Density Functionals. *Mol. Phys.* **2017**, *115* (19), 2315–2372.

(172) Lejaeghere, K.; Van Speybroeck, V.; Van Oost, G.; Cottenier, S. Error Estimates for Solid-State Density-Functional Theory Predictions: An Overview by Means of the Ground-State Elemental Crystals. *Crit. Rev. Solid State Mater. Sci.* **2014**, *39* (1), 1–24.

(173) Bosoni, E. et al. How to Verify the Precision of Density-Functional-Theory Implementations via Reproducible and Universal Workflows. *Nat. Rev. Phys.* **2024**, *6*, 45–58.

(174) Lejaeghere, K. et al. Reproducibility in Density Functional Theory Calculations of Solids. *Science* **2016**, *351* (6280), aad3000.

(175) Dral, P. O.; Owens, A.; Dral, A.; Csányi, G. Hierarchical Machine Learning of Potential Energy Surfaces. *J. Chem. Phys.* **2020**, *152* (20), 204110.

(176) Grimme, S. Density Functional Theory with London Dispersion Corrections. *WIREs Comput. Mol. Sci.* **2011**, *1* (2), 211–228.

(177) Becke, A. D.; Johnson, E. R. A Density-Functional Model of the Dispersion Interaction. *J. Chem. Phys.* **2005**, *123* (15), 154101.

(178) Grimme, S.; Ehrlich, S.; Goerigk, L. Effect of the Damping Function in Dispersion Corrected Density Functional Theory. *J. Comput. Chem.* **2011**, *32* (7), 1456–1465.

(179) Nguyen, H.-V.; de Gironcoli, S. Efficient Calculation of Exact Exchange and RPA Correlation Energies in the Adiabatic-Connection Fluctuation-Dissipation Theory. *Phys. Rev. B* **2009**, *79* (20), 205114.

(180) Harl, J.; Kresse, G. Accurate Bulk Properties from Approximate Many-Body Techniques. *Phys. Rev. Lett.* **2009**, *103* (5), 056401.

(181) Szalewicz, K. Symmetry-Adapted Perturbation Theory of Intermolecular Forces. *WIREs Comput. Mol. Sci.* **2012**, *2* (2), 254–272.

(182) Tkatchenko, A.; DiStasio, R. A.; Car, R.; Scheffler, M. Accurate and Efficient Method for Many-Body van Der Waals Interactions. *Phys. Rev. Lett.* **2012**, *108* (23), 236402.

(183) Zhao, B.; Shang, C.; Qi, N.; Chen, Z. Y.; Chen, Z. Q. Stability of Defects in Monolayer MoS_2 and Their Interaction with O_2 Molecule: A First-Principles Study. *Appl. Surf. Sci.* **2017**, *412*, 385–393.

(184) Zhang, X.; Wang, S.; Lee, C.-K.; Cheng, C.-M.; Lan, J.-C.; Li, X.; Qiao, J.; Tao, X. Unravelling the Effect of Sulfur Vacancies on the Electronic Structure of the MoS_2 Crystal. *Phys. Chem. Chem. Phys.* **2020**, *22* (38), 21776–21783.

(185) Wang, J.; Dongare, A. M. Density Functional Theory Study of Electronic Structure of Defects and the Role on the Strain Relaxation Behavior of MoS_2 Bilayer Structures. *J. Mater. Sci.* **2018**, *53* (12), 9064–9075.

(186) Gusakov, V.; Gusakova, J.; Tay, B. K. Formation and Diffusion of Intrinsic Point Defects in Bulk and Monolayer MoS_2: First-Principles Study. *Phys. Status Solidi B* **2022**, *259* (8), 2100479.

(187) Komsa, H.-P.; Krasheninnikov, A. V. Native Defects in Bulk and Monolayer MoS_2 from First Principles. *Phys. Rev. B* **2015**, *91* (12), 125304.

(188) Schönfeld, B.; Huang, J. J.; Moss, S. C. Anisotropic Mean-Square Displacements (MSD) in Single-Crystals of 2H- and 3R-MoS_2. *Acta Cryst. B* **1983**, *39* (4), 404–407.

(189) Zuo, Y.; Chen, C.; Li, X.; Deng, Z.; Chen, Y.; Behler, J.; Csányi, G.; Shapeev, A. V.; Thompson, A. P.; Wood, M. A.; Ong, S. P. Performance and Cost Assessment of Machine Learning Interatomic Potentials. *J. Phys. Chem. A* **2020**, *124* (4), 731–745.

(190) Rowe, P.; Deringer, V. L.; Gasparotto, P.; Csányi, G.; Michaelides, A. An Accurate and Transferable Machine Learning Potential for Carbon. *J. Chem. Phys.* **2020**, *153* (3), 034702.

(191) Qian, C.; McLean, B.; Hedman, D.; Ding, F. A Comprehensive Assessment of Empirical Potentials for Carbon Materials. *APL Mater.* **2021**, *9* (6), 061102.

(192) Aghajamali, A.; Karton, A. Comparative Study of Carbon Force Fields for the Simulation of Carbon Onions. *Aust. J. Chem.* **2021**, *74*, 709–714.

(193) George, J.; Hautier, G.; Bartók, A. P.; Csányi, G.; Deringer, V. L. Combining Phonon Accuracy with High Transferability in Gaussian Approximation Potential Models. *J. Chem. Phys.* **2020**, *153*, 044104.

(194) Kovács, D. P.; van der Oord, C.; Kucera, J.; Allen, A. E. A.; Cole, D. J.; Ortner, C.; Csányi, G. Linear Atomic Cluster Expansion Force Fields for Organic Molecules: Beyond RMSE. *J. Chem. Theory Comput.* **2021**, *17* (12), 7696–7711.

(195) Fu, X.; Wu, Z.; Wang, W.; Xie, T.; Keten, S.; Gomez-Bombarelli, R.; Jaakkola, T. S. Forces Are Not Enough: Benchmark and Critical Evaluation for Machine Learning Force Fields with Molecular Simulations. *Transactions on Machine Learning Research* **2023**.

(196) Stocker, S.; Gasteiger, J.; Becker, F.; Günnemann, S.; Margraf, J. T. How Robust Are Modern Graph Neural Network Potentials in Long and Hot Molecular Dynamics Simulations? *Mach. Learn.: Sci. Technol.* **2022**, *3* (4), 045010.

(197) Poltavsky, I. et al. Crash Testing Machine Learning Force Fields for Molecules, Materials, and Interfaces: Model Analysis in the TEA Challenge 2023. **2024**, DOI: 10.26434/chemrxiv-2024-ctdm3.

(198) Stocker, S.; Gasteiger, J.; Becker, F.; Günnemann, S.; Margraf, J. T. How Robust Are Modern Graph Neural Network Potentials in Long and Hot Molecular Dynamics Simulations? *Mach. Learn.: Sci. Technol.* **2022**, *3* (4), 045010.

(199) Botu, V.; Batra, R.; Chapman, J.; Ramprasad, R. Machine Learning Force Fields: Construction, Validation, and Outlook. *J. Phys. Chem. C* **2017**, *121* (1), 511–522.

(200) Zhong, M. et al. Accelerated Discovery of CO_2 Electrocatalysts Using Active Machine Learning. *Nature* **2020**, *581*, 178–183.

(201) Artrith, N.; Urban, A.; Ceder, G. Efficient and Accurate Machine-Learning Interpolation of Atomic Energies in Compositions with Many Species. *Phys. Rev. B* **2017**, *96* (1), 014112.

(202) A. Young, T.; Johnston-Wood, T.; L. Deringer, V.; Duarte, F. A Transferable Active-Learning Strategy for Reactive Molecular Force Fields. *Chem. Sci.* **2021**, *12* (32), 10944–10955.

(203) Deringer, V. L.; Bernstein, N.; Csányi, G.; Ben Mahmoud, C.; Ceriotti, M.; Wilson, M.; Drabold, D. A.; Elliott, S. R. Origins of Structural and Electronic Transitions in Disordered Silicon. *Nature* **2021**, *589* (7840), 59–64.

(204) Behler, J.; Csányi, G. Machine Learning Potentials for Extended Systems: A Perspective. *Eur. Phys. J. B* **2021**, *94*, 142.

(205) Stillinger, F. H.; Weber, T. A. Computer Simulation of Local Order in Condensed Phases of Silicon. *Phys. Rev. B* **1985**, *31* (8), 5262–5271.

(206) Vink, R. L. C.; Barkema, G. T.; van der Weg, W. F.; Mousseau, N. Fitting the Stillinger–Weber Potential to Amorphous Silicon. *J. Non-Cryst. Solids* **2001**, *282* (2), 248–255.

(207) Haynes, W. M., *CRC Handbook of Chemistry and Physics*, 97th ed.; CRC Press: Digital, 2016.

(208) Wypych, G. In *Handbook of Fillers (Fourth Edition)*, Wypych, G., Ed.; ChemTec Publishing: Digital, 2016, pp 13–266.

(209) Sameshima, T.; Usui, S. Pulsed Laser-induced Melting Followed by Quenching of Silicon Films. *J. Appl. Phys.* **1993**, *74* (11), 6592–6598.

(210) Kateb, M.; Gudmundsson, J. T.; Brault, P.; Manolescu, A.; Ingvarsson, S. On the Role of Ion Potential Energy in Low Energy HiPIMS Deposition: An Atomistic Simulation. *Surf. Coat. Technol.* **2021**, *426*, 127726.

(211) Gasparotto, P.; Fischer, M.; Scopece, D.; Liedke, M. O.; Butterling, M.; Wagner, A.; Yildirim, O.; Trant, M.; Passerone, D.; Hug, H. J.; Pignedoli, C. A. Mapping the Structure of Oxygen-Doped Wurtzite Aluminum Nitride Coatings from Ab Initio Random Structure Search and Experiments. *ACS Appl. Mater. Interfaces* **2021**, *13* (4), 5762–5771.

(212) Plimpton, S. Fast Parallel Algorithms for Short-Range Molecular Dynamics. *J. Comput. Phys.* **1995**, *117* (1), 1–19.

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

(214) Swope, W. C.; Andersen, H. C.; Berens, P. H.; Wilson, K. R. A Computer Simulation Method for the Calculation of Equilibrium Constants for the Formation of Physical Clusters of Molecules: Application to Small Water Clusters. *J. Chem. Phys.* **1982**, *76* (1), 637–649.

(215) Fincham, D. Choice of Timestep in Molecular Dynamics Simulation. *Comput. Phys. Commun.* **1986**, *40* (2), 263–269.

(216) Martin-Barrios, R.; Navas-Conyedo, E.; Zhang, X.; Chen, Y.; Gulín-González, J. An Overview about Neural Networks Potentials in Molecular Dynamics Simulation. *Int. J. Quantum Chem.* **2024**, *124* (11), e27389.

(217) Liberatore, E.; Meli, R.; Rothlisberger, U. A Versatile Multiple Time Step Scheme for Efficient Ab Initio Molecular Dynamics Simulations. *J. Chem. Theory Comput.* **2018**, *14* (6), 2834–2842.

(218) Tuckerman, M. E.; Alejandre, J.; López-Rendón, R.; Jochim, A. L.; Martyna, G. J. A Liouville-operator Derived Measure-Preserving Integrator for Molecular Dynamics Simulations in the Isothermal–Isobaric Ensemble. *J. Phys. A: Math. Gen.* **2006**, *39* (19), 5629.

(219) Nosé, S. A Unified Formulation of the Constant Temperature Molecular Dynamics Methods. *J. Chem. Phys.* **1984**, *81* (1), 511–519.

(220) Hoover, W. G. Canonical Dynamics: Equilibrium Phase-Space Distributions. *Phys. Rev. A* **1985**, *31* (3), 1695–1697.

(221) Shinoda, W.; Shiga, M.; Mikami, M. Rapid Estimation of Elastic Constants by Molecular Dynamics Simulation under Constant Stress. *Phys. Rev. B* **2004**, *69* (13), 134103.

(222) Martyna, G. J.; Tobias, D. J.; Klein, M. L. Constant Pressure Molecular Dynamics Algorithms. *J. Chem. Phys.* **1994**, *101* (5), 4177–4189.

(223) Parrinello, M.; Rahman, A. Polymorphic Transitions in Single Crystals: A New Molecular Dynamics Method. *J. Appl. Phys.* **1981**, *52* (12), 7182–7190.

(224) Yang, Y. I.; Shao, Q.; Zhang, J.; Yang, L.; Gao, Y. Q. Enhanced Sampling in Molecular Dynamics. *J. Chem. Phys.* **2019**, *151* (7), 070902.

(225) Torrie, G. M.; Valleau, J. P. Nonphysical Sampling Distributions in Monte Carlo Free-Energy Estimation: Umbrella Sampling. *J. Comput. Phys.* **1977**, *23* (2), 187–199.

(226) Laio, A.; Parrinello, M. Escaping Free-Energy Minima. *Proc. Natl. Acad. Sci. U.S. A.* **2002**, *99* (20), 12562–12566.

(227) Swendsen, R. H.; Wang, J.-S. Replica Monte Carlo Simulation of Spin-Glasses. *Phys. Rev. Lett.* **1986**, *57* (21), 2607–2609.

(228) Sugita, Y.; Okamoto, Y. Replica-Exchange Molecular Dynamics Method for Protein Folding. *Chem. Phys. Lett.* **1999**, *314* (1-2), 141–151.

(229) González, M. A.; Abascal, J. L. F. A Flexible Model for Water Based on TIP4P/2005. *J. Chem. Phys.* **2011**, *135* (22), 224516.

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

(231) Weininger, D.; Weininger, A.; Weininger, J. L. SMILES. 2. Algorithm for Generation of Unique SMILES Notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29* (2), 97–101.

(232) Sivaraman, G.; Jackson, N. E.; Sanchez-Lengeling, B.; Vázquez-Mayagoitia, Á.; Aspuru-Guzik, A.; Vishwanath, V.; de Pablo, J. J. A Machine Learning Workflow for Molecular Analysis: Application to Melting Points. *Mach. Learn.: Sci. Technol.* **2020**, *1* (2), 025015.

(233) Mayr, A.; Klambauer, G.; Unterthiner, T.; Hochreiter, S. DeepTox: Toxicity Prediction Using Deep Learning. *Front. Environ. Sci.* **2016**, *3*.

(234) Elton, D. C.; Boukouvalas, Z.; Butrico, M. S.; Fuge, M. D.; Chung, P. W. Applying Machine Learning Techniques to Predict the Properties of Energetic Materials. *Sci. Rep.* **2018**, *8*, 9059.

(235) Gierlich, C.; Palkovits, S. Featurizing Chemistry for Machine Learning — Methods and a Coded Example. *Curr. Opin. Chem. Eng.* **2022**, *37*, 100840.

(236) Jolliffe, I. T.; Cadima, J. Principal Component Analysis: A Review and Recent Developments. *Philos. Trans. A Math. Phys. Eng. Sci.* **2016**, *374* (2065), 20150202.

(237) Cao, L. J.; Chua, K. S.; Chong, W. K.; Lee, H. P.; Gu, Q. M. A Comparison of PCA, KPCA and ICA for Dimensionality Reduction in Support Vector Machine. *Neurocomputing* **2003**, *55* (1-2), 321–336.

(238) Hinton, G. E.; Roweis, S., *Stochastic Neighbor Embedding* In *Advances in Neural Information Processing Systems*, MIT Press: 2002; Vol. 15.

(239) Van der Maaten, L.; Hinton, G. Visualizing Data Using T-SNE. *J. Mach. Learn. Res.* **2008**, *9* (86), 2579–2605.

(240) Keen, D. A. A Comparison of Various Commonly Used Correlation Functions for Describing Total Scattering. *J. Appl. Crystallogr.* **2001**, *34* (2), 172–177.

(241) Chantler, C. T. Theoretical Form Factor, Attenuation, and Scattering Tabulation for Z=1–92 from E=1–10 eV to E=0.4–1.0 MeV. *J. Phys. Chem. Ref. Data* **1995**, *24* (1), 71–643.

(242) Chantler, C. T. Detailed Tabulation of Atomic Form Factors, Photoelectric Absorption and Scattering Cross Section, and Mass Attenuation Coefficients in the Vicinity of Absorption Edges in the Soft X-ray (Z= 30–36, Z= 60–89, E= 0.1–10 keV)–Addressing Convergence Issues of Earlier Work. *J. Synchrotron Radiat.* **2001**, *8* (4), 1124–1124.

(243) Katayama, Y.; Mizutani, T.; Utsumi, W.; Shimomura, O.; Yamakata, M.; Funakoshi, K.-i. A First-Order Liquid–Liquid Phase Transition in Phosphorus. *Nature* **2000**, *403* (6766), 170–173.

(244) Phillips, J. C. Topology of Covalent Non-Crystalline Solids II: Medium-range Order in Chalcogenide Alloys and A-Si(Ge). *J. Non-Cryst. Solids* **1981**, *43* (1), 37–77.

(245) Busse, L. E.; Nagel, S. R. Temperature Dependence of the Structure Factor of As_2Se_3 Glass Up to the Glass Transition. *Phys. Rev. Lett.* **1981**, *47* (25), 1848–1851.

(246) Wright, A. C.; Sinclair, R. N.; Leadbetter, A. J. Effect of Preparation Method on the Structure of Amorphous Solids in the System As–S. *J. Non-Cryst. Solids* **1985**, *71* (1), 295–302.

(247) Červinka, L. Medium-Range Order in Amorphous Materials. *J. Non-Cryst. Solids* **1988**, *106* (1), 291–300.

(248) Blétry, J. Sphere and Distance Models for Binary Disordered Systems. *Philos. Mag. B* **1990**, *62* (5), 469–508.

(249) Elliott, S. R. Origin of the First Sharp Diffraction Peak in the Structure Factor of Covalent Glasses. *Phys. Rev. Lett.* **1991**, *67* (6), 711–714.

(250) Gaskell, P. H.; Wallis, D. J. Medium-Range Order in Silica, the Canonical Network Glass. *Phys. Rev. Lett.* **1996**, *76* (1), 66–69.

(251) Christie, J. K.; Taraskin, S. N.; Elliott, S. R. Structural Characteristics of Positionally Disordered Lattices: Relation to the First Sharp Diffraction Peak in Glasses. *Phys. Rev. B* **2004**, *70* (13), 134207.

(252) Kobayashi, K.; Okumura, M.; Nakamura, H.; Itakura, M.; Machida, M.; Urata, S.; Suzuya, K. Machine Learning Molecular Dynamics Reveals the

Structural Origin of the First Sharp Diffraction Peak in High-Density Silica Glasses. *Sci. Rep.* **2023**, *13*, 18721.

(253) Xie, R.; Long, G. G.; Weigand, S. J.; Moss, S. C.; Carvalho, T.; Roorda, S.; Hejna, M.; Torquato, S.; Steinhardt, P. J. Hyperuniformity in Amorphous Silicon Based on the Measurement of the Infinite-Wavelength Limit of the Structure Factor. *Proc. Natl. Acad. Sci. U.S. A.* **2013**, *110* (33), 13250–13254.

(254) Torquato, S.; Haslach, H. J. Random Heterogeneous Materials: Microstructure and Macroscopic Properties. *Appl. Mech. Rev.* **2002**, *55* (4), B62–B63.

(255) Florescu, M.; Torquato, S.; Steinhardt, P. J. Designer Disordered Materials with Large, Complete Photonic Band Gaps. *Proc. Natl. Acad. Sci. U.S. A.* **2009**, *106* (49), 20658–20663.

(256) Jiao, Y.; Lau, T.; Hatzikirou, H.; Meyer-Hermann, M.; Joseph C. Corbo; Torquato, S. Avian Photoreceptor Patterns Represent a Disordered Hyperuniform Solution to a Multiscale Packing Problem. *Phys. Rev. E* **2014**, *89* (2), 022721.

(257) Philcox, O. H. E.; Torquato, S. Disordered Heterogeneous Universe: Galaxy Distribution and Clustering across Length Scales. *Phys. Rev. X* **2023**, *13* (1), 011038.

(258) Treacy, M. M. J.; Borisenko, K. B. The Local Structure of Amorphous Silicon. *Science* **2012**, *335* (6071), 950–953.

(259) Tucker, M. G.; Keen, D. A.; Dove, M. T.; Goodwin, A. L.; Hui, Q. RMCProfile: Reverse Monte Carlo for Polycrystalline Materials. *J. Phys.: Condens. Matter* **2007**, *19* (33), 335218.

(260) Kugler, S.; Pusztai, L.; Rosta, L.; Chieux, P.; Bellissent, R. Structure of Evaporated Pure Amorphous Silicon: Neutron-diffraction and Reverse Monte Carlo Investigations. *Phys. Rev. B* **1993**, *48* (10), 7685–7688.

(261) Aoun, B. Fullrmc, a Rigid Body Reverse Monte Carlo Modeling Package Enabled with Machine Learning and Artificial Intelligence. *J. Comput. Chem.* **2016**, *37* (12), 1102–1111.

(262) Opletal, G.; Petersen, T.; O'Malley, B.; Snook, I.; McCulloch, D. G.; Marks, N. A.; Yarovsky, I. Hybrid Approach for Generating Realistic Amorphous Carbon Structure Using Metropolis and Reverse Monte Carlo. *Mol. Sim.* **2002**, *28* (10-11), 927–938.

(263) A. Young, C.; L. Goodwin, A. Applications of Pair Distribution Function Methods to Contemporary Problems in Materials Chemistry. *J. Mater. Chem.* **2011**, *21* (18), 6464–6476.

(264) Rosset, L. A. M.; Drabold, D. A.; Deringer, V. L. Signatures of Paracrystallinity in Amorphous Silicon. *ArXiv* **2024**, (arXiv:2407.16681), DOI: 10.48550/arXiv.2407.16681.

(265) De Tomas, C.; Suarez-Martinez, I.; Marks, N. A. Graphitization of Amorphous Carbons: A Comparative Study of Interatomic Potentials. *Carbon* **2016**, *109*, 681–693.

(266) Richter, S. R.; Hayder, Z.; Koltun, V. Playing for Benchmarks. *ArXiv* **2017**, (arXiv:1709.07322), DOI: 10.48550/arXiv.1709.07322.

(267) Hinton, G.; Vinyals, O.; Dean, J. Distilling the Knowledge in a Neural Network. *ArXiv* **2015**, (arXiv:1503.02531).

(268) Wang, J.; Olsson, S.; Wehmeyer, C.; Pérez, A.; Charron, N. E.; de Fabritiis, G.; Noé, F.; Clementi, C. Machine Learning of Coarse-Grained Molecular Dynamics Force Fields. *ACS Cent. Sci.* **2019**, *5* (5), 755–767.

(269) Ekström Kelvinius, F.; Georgiev, D.; Toshev, A.; Gasteiger, J. Accelerating Molecular Graph Neural Networks via Knowledge Distillation. *Advances in Neural Information Processing Systems* **2023**, *36*, 25761–25792.

(270) Zhou, P.; Wang, J.; Li, C.; Wang, Z.; Liu, Y.; Liu, C.; Sun, S.; Lin, J.; Wei, L.; Cai, X.; Lai, H.; Liu, W.; Wang, L.; Zeng, X.; Li, K. Instruction Multi-Constraint Molecular Generation Using a Teacher-Student Large Language Model. *ArXiv* **2024**, (arXiv:2403.13244).

(271) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; ukasz Kaiser, Ł.; Polosukhin, I., *Attention Is All You Need* In *Advances in Neural Information Processing Systems*, Curran Associates, Inc.: 2017; Vol. 30.

(272) Yenduri, G.; Ramalingam, M.; Selvi, G. C.; Supriya, Y.; Srivastava, G.; Maddikunta, P. K. R.; Raj, G. D.; Jhaveri, R. H.; Prabadevi, B.; Wang, W.; Vasilakos, A. V.; Gadekallu, T. R. GPT (Generative Pre-Trained Transformer)— A Comprehensive Review on Enabling Technologies, Potential Applications, Emerging Challenges, and Future Directions. *IEEE Access* **2024**, *12*, 54608–54649.

(273) King-Smith, E. Transfer Learning for a Foundational Chemistry Model. *Chem. Sci.* **2024**, *15* (14), 5143–5151.

(274) Nguyen, T.; Chen, Z.; Lee, J. Dataset Meta-Learning from Kernel Ridge-Regression. *ArXiv* **2021**, (arXiv:2011.00050).

(275) Glielmo, A.; Zeni, C.; De Vita, A. Efficient Nonparametric N-Body Force Fields from Machine Learning. *Phys. Rev. B* **2018**, *97* (18), 184307.

(276) Cheng, B.; Griffiths, R.-R.; Wengert, S.; Kunkel, C.; Stenczel, T.; Zhu, B.; Deringer, V. L.; Bernstein, N.; Margraf, J. T.; Reuter, K.; Csanyi, G. Mapping Materials and Molecules. *Acc. Chem. Res.* **2020**, *53*, 1981–1991.

(277) Anzellini, S.; Wharmby, M. T.; Miozzi, F.; Kleppe, A.; Daisenberger, D.; Wilhelm, H. Quasi-Hydrostatic Equation of State of Silicon up to 1 Megabar at Ambient Temperature. *Sci. Rep.* **2019**, *9*, 15537.

(278) Deb, S. K.; Wilding, M.; Somayazulu, M.; McMillan, P. F. Pressure-Induced Amorphization and an Amorphous–Amorphous Transition in Densified Porous Silicon. *Nature* **2001**, *414*, 528–530.

(279) Niblett, S. P.; Kourtis, P.; Magdău, I.-B.; Grey, C. P.; Csányi, G. Transferability of Datasets between Machine-Learning Interaction Potentials. *ArXiv* **2024**, (arXiv:2409.05590).

(280) Unke, O. T.; Chmiela, S.; Sauceda, H. E.; Gastegger, M.; Poltavsky, I.; Schütt, K. T.; Tkatchenko, A.; Müller, K.-R. Machine Learning Force Fields. *Chem. Rev.* **2021**, *121* (16), 10142–10186.

(281) Bernstein, N.; Bhattarai, B.; Csányi, G.; Drabold, D. A.; Elliott, S. R.; Deringer, V. L. Quantifying Chemical Structure and Machine-Learned Atomic Energies in Amorphous and Liquid Silicon. *Angew. Chem. Int. Ed.* **2019**, *58* (21), 7057–7061.

(282) Crain, J.; Ackland, G. J.; Maclean, J. R.; Piltz, R. O.; Hatton, P. D.; Pawley, G. S. Reversible Pressure-Induced Structural Transitions between Metastable Phases of Silicon. *Phys. Rev. B* **1994**, *50* (17), 13043–13046.

(283) Fujimoto, Y.; Koretsune, T.; Saito, S.; Miyake, T.; Oshiyama, A. A New Crystalline Phase of Four-Fold Coordinated Silicon and Germanium. *New J. Phys.* **2008**, *10* (8), 083001.

(284) Atta-Fynn, R.; Biswas, P. Nearly Defect-Free Dynamical Models of Disordered Solids: The Case of Amorphous Silicon. *J. Chem. Phys.* **2018**, *148* (20), 204503.

(285) Limbu, D. K.; Elliott, S. R.; Atta-Fynn, R.; Biswas, P. Disorder by Design: A Data-Driven Approach to Amorphous Semiconductors without Total-Energy Functionals. *Sci. Rep.* **2020**, *10* (1), 7742.

(286) Yuan, X.; Cormack, A. N. Efficient Algorithm for Primitive Ring Statistics in Topological Networks. *Comput. Mater. Sci.* **2002**, *24* (3), 343–360.

(287) Smith, J. S.; Nebgen, B.; Mathew, N.; Chen, J.; Lubbers, N.; Burakovsky, L.; Tretiak, S.; Nam, H. A.; Germann, T.; Fensin, S.; Barros, K. Automated Discovery of a Robust Interatomic Potential for Aluminum. *Nat. Commun.* **2021**, *12*, 1257.

(288) Kozinsky, B.; Musaelian, A.; Johansson, A.; Batzner, S., *Scaling the Leading Accuracy of Deep Equivariant Models to Biomolecular Simulations of Realistic Size* In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, Association for Computing Machinery: New York, NY, USA, 2023.

(289) Durandurdu, M.; Drabold, D. A. Ab Initio Simulation of First-Order Amorphous-to-Amorphous Phase Transition of Silicon. *Phys. Rev. B* **2001**, *64* (1), 014101.

(290) Brown, T. B. et al. Language Models Are Few-Shot Learners. *ArXiv* **2020**, (arXiv:2005.14165).

(291) Scherbela, M.; Gerard, L.; Grohs, P. Variational Monte Carlo on a Budget — Fine-tuning Pre-Trained Neural Wavefunctions. *Advances in Neural Information Processing Systems* **2023**, *36*, 23902–23920.

(292) Xie, Z.; Evangelopoulos, X.; H. Omar, Ö.; Troisi, A.; I. Cooper, A.; Chen, L. Fine-Tuning GPT-3 for Machine Learning Electronic and Functional Properties of Organic Molecules. *Chem. Sci.* **2024**, *15* (2), 500–510.

(293) Clark, S. J.; Segall, M. D.; Pickard, C. J.; Hasnip, P. J.; Probert, M. J.; Refson, K.; Payne, M. C. First Principles Methods Using CASTEP. *Z. Krist.* **2005**, *220* (5–6), 567–570.

(294) Elliott, S. R. In *Physics of Amorphous Materials*, 2nd ed.; Longman Scientific & Technical: Harlow, Essex, England, 1990.

(295) Solar Photovoltaic Cell Basics `https://www.energy.gov/eere/solar/solar-photovoltaic-cell-basics` (accessed 09/26/2024).

(296) Hedler, A.; Klaumünzer, S. L.; Wesch, W. Amorphous Silicon Exhibits a Glass Transition. *Nat. Mater.* **2004**, *3*, 804–809.

(297) McMillan, P. F.; Wilson, M.; Daisenberger, D.; Machon, D. A Density-Driven Phase Transition between Semiconducting and Metallic Polyamorphs of Silicon. *Nat. Mater.* **2005**, *4*, 680–684.

(298) Wang, Y.; Ding, J.; Fan, Z.; Tian, L.; Li, M.; Lu, H.; Zhang, Y.; Ma, E.; Li, J.; Shan, Z. Tension–Compression Asymmetry in Amorphous Silicon. *Nat. Mater.* **2021**, *20*, 1371–1377.

(299) Pantelides, S. T. Defects in Amorphous Silicon: A New Perspective. *Phys. Rev. Lett.* **1986**, *57* (23), 2979–2982.

(300) Stathis, J. H. Analysis of the Superhyperfine Structure and the g Tensor of Defects in Amorphous Silicon. *Phys. Rev. B* **1989**, *40* (2), 1232–1237.

(301) Zafar, S.; Schiff, E. A. Hydrogen and Defects in Amorphous Silicon. *Phys. Rev. Lett.* **1991**, *66* (11), 1493–1496.

(302) Williamson, D. L.; Roorda, S.; Chicoine, M.; Tabti, R.; Stolk, P. A.; Acco, S.; Saris, F. W. On the Nanostructure of Pure Amorphous Silicon. *Appl. Phys. Lett.* **1995**, *67* (2), 226–228.

(303) Polman, A.; Jacobson, D. C.; Coffa, S.; Poate, J. M.; Roorda, S.; Sinke, W. C. Defect States of Amorphous Si Probed by the Diffusion and Solubility of Cu. *Appl. Phys. Lett.* **1990**, *57* (12), 1230–1232.

(304) van den Hoven, G. N.; Liang, Z. N.; Niesen, L.; Custer, J. S. Evidence for Vacancies in Amorphous Silicon. *Phys. Rev. Lett.* **1992**, *68* (25), 3714–3717.

(305) Knief, S.; von Niessen, W.; Koslowski, T. Defects in A-Si and a-Si:H: A Numerical Study. *Phys. Rev. B* **1998**, *58* (8), 4459.

(306) Drabold, D. A.; Abtew, T. A. In *Theory of Defects in Semiconductors*, Drabold, D. A., Estreicher, S. K., Eds.; Springer-Verlag Berlin: Berlin, 2007; Vol. 104, pp 245–268.

(307) Lewis, L. J. Fifty Years of Amorphous Silicon Models: The End of the Story? *J. Non-Cryst. Solids* **2022**, *580*, 121383.

(308) Luedtke, W. D.; Landman, U. Preparation, Structure, Dynamics, and Energetics of Amorphous Silicon: A Molecular-Dynamics Study. *Phys. Rev. B* **1989**, *40* (2), 1164–1174.

(309) Wang, Z. Q.; Stroud, D.; Markworth, A. J. Monte Carlo Study of the Liquid CdTe Surface. *Phys. Rev. B* **1989**, *40* (5), 3129–3132.

(310) Wen, M.; Shirodkar, S. N.; Plecháč, P.; Kaxiras, E.; Elliott, R. S.; Tadmor, E. B. A Force-Matching Stillinger-Weber Potential for MoS_2: Parameterization and Fisher Information Theory Based Sensitivity Analysis. *J. Appl. Phys.* **2017**, *122* (24), 244301.

(311) Dahal, D.; Atta-Fynn, R.; Elliott, S. R.; Biswas, P. Hyperuniformity and Static Structure Factor of Amorphous Silicon in the Infinite-Wavelength Limit. *J. Phys. Conf. Ser.* **2019**, *1252* (1), 012003.

(312) Dahal, D.; Elliott, S. R.; Biswas, P. Extended-Range Order in Tetrahedral Amorphous Semiconductors: The Case of Amorphous Silicon. *Phys. Rev. B* **2022**, *105* (11), 115203.

(313) Car, R.; Parrinello, M. Structural, Dymanical, and Electronic Properties of Amorphous Silicon: An Ab Initio Molecular-Dynamics Study. *Phys. Rev. Lett.* **1988**, *60* (3), 204–207.

(314) Štich, I.; Car, R.; Parrinello, M. Amorphous Silicon Studied by Ab Initio Molecular Dynamics: Preparation, Structure, and Properties. *Phys. Rev. B* **1991**, *44* (20), 11092–11104.

(315) Pedersen, A.; Pizzagalli, L.; Jónsson, H. Optimal Atomic Structure of Amorphous Silicon Obtained from Density Functional Theory Calculations. *New J. Phys.* **2017**, *19* (6), 063018.

(316) Xu, Y.; Zhou, Y.; Wang, X.-D.; Zhang, W.; Ma, E.; Deringer, V. L.; Mazzarello, R. Unraveling Crystallization Mechanisms and Electronic Structure of Phase-Change Materials by Large-Scale Ab Initio Simulations. *Adv. Mater.* **2022**, *34* (11), 2109139.

(317) Biegelsen, D. K.; Stutzmann, M. Hyperfine Studies of Dangling Bonds in Amorphous Silicon. *Phys. Rev. B* **1986**, *33* (5), 3006–3011.

(318) Moss, S. C.; Graczyk, J. F. Evidence of Voids Within the As-Deposited Structure of Glassy Silicon. *Phys. Rev. Lett.* **1969**, *23* (20), 1167–1171.

(319) Roorda, S.; Lewis, L. J. Comment on "The Local Structure of Amorphous Silicon". *Science* **2012**, *338* (6114), 1539–1539.

(320) Coffa, S.; Poate, J. M. Hydrogen Induced Detrapping of Transition Metals in Amorphous Silicon. *Appl. Phys. Lett.* **1991**, *59* (18), 2296–2298.

(321) Acco, S.; Williamson, D. L.; Stolk, P. A.; Saris, F. W.; van den Boogaard, M. J.; Sinke, W. C.; van der Weg, W. F.; Roorda, S.; Zalm, P. C. Hydrogen Solubility and Network Stability in Amorphous Silicon. *Phys. Rev. B* **1996**, *53* (8), 4415–4427.

(322) Hiroyama, Y.; Motooka, T.; Suzuki, R.; Hirano, Y.; Sato, F. Effects of Hydrogen on Structural Relaxation and Defect Evolution in Amorphous Silicon. *Appl. Phys. Lett.* **1996**, *68* (22), 3126–3128.

(323) Wang, Y.; Fan, Z.; Qian, P.; Caro, M. A.; Ala-Nissila, T. Quantum-Corrected Thickness-Dependent Thermal Conductivity in Amorphous Silicon Predicted by Machine Learning Molecular Dynamics Simulations. *Phys. Rev. B* **2023**, *107* (5), 054303.

(324) El-Machachi, Z.; Wilson, M.; Deringer, V. L. Exploring the Configurational Space of Amorphous Graphene with Machine-Learned Atomic Energies. *Chem. Sci.* **2022**, *13* (46), 13720–13731.

(325) Wang, S.; Liu, Y.; Mo, Y. Frustration in Super-Ionic Conductors Unraveled by the Density of Atomistic States. *Angew. Chem. Int. Ed.* **2023**, *62* (15), e202215544.

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

(327) Chen, X.; Jørgensen, M. S.; Li, J.; Hammer, B. Atomic Energies from a Convolutional Neural Network. *J. Chem. Theory Comput.* **2018**, *14* (7), 3933–3942.

(328) Eckhoff, M.; Behler, J. From Molecular Fragments to the Bulk: Development of a Neural Network Potential for MOF-5. *J. Chem. Theory Comput.* **2019**, *15* (6), 3793–3809.

(329) Pauling, L. Atomic Radii and Interatomic Distances in Metals. *J. Am. Chem. Soc.* **1947**, *69* (3), 542–553.

(330) Caro, M. A.; Csányi, G.; Laurila, T.; Deringer, V. L. Machine Learning Driven Simulated Deposition of Carbon Films: From Low-Density to Diamondlike Amorphous Carbon. *Phys. Rev. B* **2020**, *102* (17), 174201.

(331) Steinbach, M.; Karypis, G.; Kumar, V. *A Comparison of Document Clustering Techniques*; Report; 2000.

(332) Gillespie, R. J.; Nyholm, R. S. Inorganic Stereochemistry. *Q. Rev., Chem. Soc.* **1957**, *11* (4), 339–380.

(333) Gillespie, R. J. The VSEPR Model Revisited. *Chem. Soc. Rev.* **1992**, *21* (1), 59–69.

(334) Ceriotti, M. Unsupervised Machine Learning in Atomistic Simulations, between Predictions and Understanding. *J. Chem. Phys.* **2019**, *150* (15), 150901.

(335) Glielmo, A.; Husic, B. E.; Rodriguez, A.; Clementi, C.; Noé, F.; Laio, A. Unsupervised Learning Methods for Molecular Simulation Data. *Chem. Rev.* **2021**, *121* (16), 9722–9758.

(336) Beeman, D.; Tsu, R.; Thorpe, M. F. Structural Information from the Raman Spectrum of Amorphous Silicon. *Phys. Rev. B* **1985**, *32* (2), 874–878.

(337) Drabold, D. A.; Li, Y.; Cai, B.; Zhang, M. Urbach Tails of Amorphous Silicon. *Phys. Rev. B* **2011**, *83* (4), 045201.

(338) Pan, Y.; Zhang, M.; Drabold, D. A. Topological and Topological-Electronic Correlations in Amorphous Silicon. *J. Non-Cryst. Solids* **2008**, *354* (29), 3480–3485.

(339) Pan, Y.; Inam, F.; Zhang, M.; Drabold, D. A. Atomistic Origin of Urbach Tails in Amorphous Silicon. *Phys. Rev. Lett.* **2008**, *100* (20), 206403.

(340) Berry, R. S. Correlation of Rates of Intramolecular Tunneling Processes, with Application to Some Group V Compounds. *J. Chem. Phys.* **1960**, *32* (3), 933–938.

(341) Couzijn, E. P. A.; Slootweg, J. C.; Ehlers, A. W.; Lammertsma, K. Stereomutation of Pentavalent Compounds: Validating the Berry Pseudorotation, Redressing Ugi's Turnstile Rotation, and Revealing the Two- and Three-Arm Turnstiles. *J. Am. Chem. Soc.* **2010**, *132* (51), 18127–18140.

(342) Moberg, C. Stereomutation in Trigonal-Bipyramidal Systems: A Unified Picture. *Angew. Chem. Int. Ed.* **2011**, *50* (44), 10290.

(343) Mosquera-Lois, I.; Kavanagh, S. R.; Walsh, A.; Scanlon, D. O. ShakeNBreak: Navigating the Defect Configurational Landscape. *J. Open Source Softw.* **2022**, *7* (80), 4817.

(344) Kavanagh, S. R.; Walsh, A.; Scanlon, D. O. Rapid Recombination by Cadmium Vacancies in CdTe. *ACS Energy Lett.* **2021**, *6* (4), 1392–1398.

(345) Mosquera-Lois, I.; Kavanagh, S. R. In Search of Hidden Defects. *Matter* **2021**, *4* (8), 2602–2605.

(346) Mosquera-Lois, I.; Kavanagh, S. R.; Walsh, A.; Scanlon, D. O. Identifying the Ground State Structures of Point Defects in Solids. *npj Comput. Mater.* **2023**, *9*, 25.

(347) Frey, N. C.; Akinwande, D.; Jariwala, D.; Shenoy, V. B. Machine Learning-Enabled Design of Point Defects in 2D Materials for Quantum and Neuromorphic Information Processing. *ACS Nano* **2020**, *14* (10), 13406–13417.

(348) Arrigoni, M.; Madsen, G. K. H. Evolutionary Computing and Machine Learning for Discovering of Low-Energy Defect Configurations. *npj Comput. Mater.* **2021**, *7*, 71.

(349) Witman, M. D.; Goyal, A.; Ogitsu, T.; McDaniel, A. H.; Lany, S. Defect Graph Neural Networks for Materials Discovery in High-Temperature Clean-Energy Applications. *Nat. Comput. Sci.* **2023**, *3*, 675–686.

(350) Mavračić, J.; Mocanu, F. C.; Deringer, V. L.; Csányi, G.; Elliott, S. R. Similarity Between Amorphous and Crystalline Phases: The Case of TiO_2. *J. Phys. Chem. Lett.* **2018**, *9* (11), 2985–2990.

(351) Ratnasamy, P.; Rodrique, L.; Leonard, A. J. Structural and Textural Studies in Molybdenum Sulfide Systems. *J. Phys. Chem.* **1973**, *77* (18), 2242–2245.

(352) Diemann, E. Radiale Verteilungsfunktionen. V. Strukturuntersuchungen an nichtkristallinem Molybdäntrisulfid, Wolframtrisulfid und Molybdäntriselenid. *Z. Anorg. Allg. Chem.* **1977**, *432* (1), 127–135.

(353) Cramer, S. P.; Liang, K. S.; Jacobson, A. J.; Chang, C. H.; Chianelli, R. R. EXAFS Studies of Amorphous Molybdenum and Tungsten Trisulfides and Triselenides. *Inorg. Chem.* **1984**, *23* (9), 1215–1221.

(354) Hibble, S. J.; Wood, G. B. Modeling the Structure of Amorphous MoS_3: A Neutron Diffraction and Reverse Monte Carlo Study. *J. Am. Chem. Soc.* **2004**, *126* (3), 959–965.

(355) Chu, K.; Nan, H.; Li, Q.; Guo, Y.; Tian, Y.; Liu, W. Amorphous MoS_3 Enriched with Sulfur Vacancies for Efficient Electrocatalytic Nitrogen Reduction. *J. Energy Chem.* **2021**, *53*, 132–138.

(356) Lince, J. R.; Pluntze, A. M.; Jackson, S. A.; Radhakrishnan, G.; Adams, P. M. Tribochemistry of MoS_3 Nanoparticle Coatings. *Tribol. Lett.* **2014**, *53* (3), 543–554.

(357) Yu, J.; Xiao, J.; Li, A.; Yang, Z.; Zeng, L.; Zhang, Q.; Zhu, Y.; Guo, L. Enhanced Multiple Anchoring and Catalytic Conversion of Polysulfides by Amorphous MoS_3 Nanoboxes for High-Performance Li-S Batteries. *Angew. Chem. Int. Ed.* **2020**, *59* (31), 13071–13078.

(358) Scott, R. A.; Jacobson, A. J.; Chianelli, R. R.; Pan, W. H.; Stiefel, E. I.; Hodgson, K. O.; Cramer, S. P. Reactions of Molybdenum Trisulfide, Tungsten Trisulfide, Tungsten Triselenide, and Niobium Triselenide with Lithium. Metal Cluster Rearrangement Revealed by EXAFS. *Inorg. Chem.* **1986**, *25* (9), 1461–1466.

(359) De Boer, M.; Van Dillen, A. J.; Koningsberger, D. C.; Geus, J. W. The Structure of Highly Dispersed SiO_2-Supported Molybdenum Oxide Catalysts during Sulfidation. *J. Phys. Chem.* **1994**, *98* (32), 7862–7870.

(360) Müller, A.; Jostes, R.; Cotton, F. A. Trinuclear Clusters of the Early Transition Elements. *Angew. Chem. Int. Ed.* **1980**, *19* (11), 875–882.

(361) Cotton, F. A.; Wilkinson, G.; Murillo, C. A.; Bochmann, M. In *Advanced Inorganic Chemistry*, 6th ed.; John Wiley & Sons: 1999, pp 920–973.

(362) Hibble, S. J.; Fawcett, I. D. Local Order and Metal-Metal Bonding in Li_2MoO_3, $Li_4Mo_3O_8$, $LiMoO_2$, and H_2MoO_3, Determined from EXAFS Studies. *Inorg. Chem.* **1995**, *34* (2), 500–508.

(363) Rees, D. C.; Chan, M. K.; Kim, J. In *Advances in Inorganic Chemistry*, Sykes, A. G., Ed.; Academic Press: 1993; Vol. 40, pp 89–119.

(364) Hernández-Molina, R.; Sykes, A. G. Chalcogenide-Bridged Cuboidal Clusters with M_4Q_4 (M = Mo, W; Q = S, Se, Te) Cores. *J. Chem. Soc., Dalton Trans.* **1999**, (18), 3137–3148.

(365) Feliz, M.; Guillamón, E.; Llusar, R.; Vicent, C.; Stiriba, S.-E.; Pérez-Prieto, J.; Barberis, M. Unprecedented Stereoselective Synthesis of Catalytically Active Chiral Mo_3CuS_4 Clusters. *Chem. Eur. J.* **2006**, *12* (5), 1486–1492.

(366) Poppe, J.; Warkentin, E.; Demmer, U.; Kowalewski, B.; Dierks, T.; Schneider, K.; Ermler, U. Structural Diversity of Polyoxomolybdate Clusters along the Three-Fold Axis of the Molybdenum Storage Protein. *J. Inorg. Biochem.* **2014**, *138*, 122–128.

(367) Ji, Z.; Trickett, C.; Pei, X.; Yaghi, O. M. Linking Molybdenum–Sulfur Clusters for Electrocatalytic Hydrogen Evolution. *J. Am. Chem. Soc.* **2018**, *140* (42), 13618–13622.

(368) Müller, A.; Fedin, V.; Hegetschweiler, K.; Amrein, W. Characterization of Amorphous Substances by Studying Isotopically Labelled Compounds with FAB–MS: Evidence for Extrusion of Triangular Mo_3^{IV} clusters from a mixture of $^{92}MoS_3$ and $^{100}MoS_3$ by reaction with OH^-. *J. Chem. Soc., Chem. Commun.* **1992**, *0* (24), 1795–1796.

(369) Tran, P. D.; Tran, T. V.; Orio, M.; Torelli, S.; Truong, Q. D.; Nayuki, K.; Sasaki, Y.; Chiam, S. Y.; Yi, R.; Honma, I.; Barber, J.; Artero, V. Coordination Polymer Structure and Revisited Hydrogen Evolution Catalytic Mechanism for Amorphous Molybdenum Sulfide. *Nat. Mater.* **2016**, *15*, 640–646.

(370) Gereben, O.; Pusztai, L. Understanding the Structure of Molecular Liquids via Combinations of Molecular Dynamics Simulations and Reverse Monte Carlo Modeling: Handling Information Deficiency. *J. Non-Cryst. Solids* **2015**, *407*, 213–219.

(371) Sahu, A.; Steinmann, S. N.; Raybaud, P. Size-Dependent Structural, Energetic, and Spectroscopic Properties of MoS_3 Polymorphs. *Cryst. Growth Des.* **2020**, *20* (12), 7750–7760.

(372) Wilson, M. Model Investigations of Network-Forming Materials. *Phys. Chem. Chem. Phys.* **2012**, *14* (37), 12701–12714.

(373) McMillan, P. F.; Wilson, M.; Wilding, M. C.; Daisenberger, D.; Mezouar, M.; Greaves, G. N. Polyamorphism and Liquid–Liquid Phase Transitions: Challenges for Experiment and Theory. *J. Phys.: Condens. Matter* **2007**, *19* (41), 415101.

(374) Skoge, M.; Donev, A.; Stillinger, F. H.; Torquato, S. Packing Hyperspheres in High-Dimensional Euclidean Spaces. *Phys. Rev. E* **2006**, *74* (4), 041127.

(375) Bundy, A.; Wallen, L. In *Catalogue of Artificial Intelligence Tools*, Bundy, A., Wallen, L., Eds.; Springer: Berlin, Heidelberg, 1984, pp 13–13.

(376) Togo, A. First-Principles Phonon Calculations with Phonopy and Phono3py. *J. Phys. Soc. Jpn.* **2023**, *92* (1), 012001.

(377) Rehr, J. J.; Kas, J. J.; Vila, F. D.; Prange, M. P.; Jorissen, K. Parameter-Free Calculations of X-ray Spectra with FEFF9. *Phys. Chem. Chem. Phys.* **2010**, *12* (21), 5503–5513.

(378) Bunker, B. A.; Stern, E. A. Phase Factor in Extended X-Ray-Absorption Fine Structure. *Phys. Rev. B* **1983**, *27* (2), 1017–1027.

(379) Fedorenko, A. D.; Semushkina, G. I.; Peregudova, N. N.; Lavrukhina, S. A.; Gushchin, A. L.; Fomenko, Y. S.; Sokolov, M. N.; Gusel′nikov, A. V.; Kalinkin, A. V.; Nikolenko, A. D.; Ivlyushkin, D. V.; Zavertkin, P. S.; Mazalov, L. N. Studying the Electronic Structure of Trinuclear Molybdenum Cluster Sulfides with Mo_3S_4 and Mo_3S_7 Cores by X-ray Spectroscopy. *J. Struct. Chem.* **2021**, *62* (6), 853–864.

(380) Kibsgaard, J.; Jaramillo, T. F.; Besenbacher, F. Building an Appropriate Active-Site Motif into a Hydrogen-Evolution Catalyst with Thiomolybdate $[Mo_3S_{13}]^{2-}$ Clusters. *Nat. Chem.* **2014**, *6*, 248–253.

(381) Gushchin, A. L.; Laricheva, Y. A.; Abramov, P. A.; Virovets, A. V.; Vicent, C.; Sokolov, M. N.; Llusar, R. Homoleptic Molybdenum Cluster Sulfides Functionalized with Noninnocent Diimine Ligands: Synthesis, Structure, and Redox Behavior. *Eur. J. Inorg. Chem.* **2014**, *2014* (25), 4093–4100.

(382) Tsuge, K.; Imoto, H.; Saito, T. Synthesis, Structure, and Molecular Orbital Calculation of the Bicapped Triangular Molybdenum Cluster Complex $[Mo_3S_5(PMe_3)_6]$. *Inorg. Chem.* **1995**, *34* (13), 3404–3409.

(383) Spevack, P. A.; McIntyre, N. S. Thermal Reduction of Molybdenum Trioxide. *J. Phys. Chem.* **1992**, *96* (22), 9029–9035.

(384) Llusar, R.; Vicent, C. Trinuclear Molybdenum Cluster Sulfides Coordinated to Dithiolene Ligands and Their Use in the Development of Molecular Conductors. *Coordination Chemistry Reviews* **2010**, *254* (13), 1534–1548.

(385) Müller, P. C.; Ertural, C.; Hempelmann, J.; Dronskowski, R. Crystal Orbital Bond Index: Covalent Bond Orders in Solids. *J. Phys. Chem. C* **2021**, *125* (14), 7959–7970.

(386) Reitz, L. S.; Hempelmann, J.; Müller, P. C.; Dronskowski, R.; Steinberg, S. Bonding Analyses in the Broad Realm of Intermetallics: Understanding the Role of Chemical Bonding in the Design of Novel Materials. *Chem. Mater.* **2024**, *36* (14), 6791–6804.

(387) Maintz, S.; Deringer, V. L.; Tchougréeff, A. L.; Dronskowski, R. LOBSTER: A Tool to Extract Chemical Bonding from Plane-Wave Based DFT. *J. Comput. Chem.* **2016**, *37* (11), 1030–1035.

(388) Andrienko, D. Introduction to Liquid Crystals. *J. Mol. Liq.* **2018**, *267*, 520–541.

(389) Alonso, G.; Del Valle, M.; Cruz, J.; Petranovskii, V.; Licea-Claverie, A.; Fuentes, S. Preparation of MoS_2 Catalysts by in Situ Decomposition of Tetraalkylammonium Thiomolybdates. *Catalysis Today* **1998**, *43* (1), 117–122.

(390) Matsuyama, T.; Hayashi, A.; Ozaki, T.; Mori, S.; Tatsumisago, M. Electrochemical Properties of All-Solid-State Lithium Batteries with Amorphous MoS_3 Electrodes Prepared by Mechanical Milling. *J. Mater. Chem. A* **2015**, *3* (27), 14142–14147.

(391) Autoatml/Autoplex, autoatml, 2024.

(392) Deringer, V. L.; Pickard, C. J.; Csányi, G. Data-Driven Learning of Total and Local Energies in Elemental Boron. *Phys. Rev. Lett.* **2018**, *120* (15), 156001.

(393) Mahoney, M. W.; Drineas, P. CUR Matrix Decompositions for Improved Data Analysis. *Proc. Natl. Acad. Sci., U. S. A.* **2009**, *106* (3), 697–702.

(394) Bochkarev, A.; Lysogorskiy, Y.; Menon, S.; Qamar, M.; Mrovec, M.; Drautz, R. Efficient Parametrization of the Atomic Cluster Expansion. *Phys. Rev. Mater.* **2022**, *6* (1), 013804.

(395) Liu, X.; Wang, B.; Jia, K.; Wang, Q.; Wang, D.; Xiong, Y. First-Principles-Based Machine Learning Interatomic Potential for Molecular Dynamics Simulations of 2D Lateral MoS_2/WS_2 Heterostructures. *J. Appl. Phys.* **2024**, *135* (20), 205107.

(396) Jiang, J.-W.; Park, H. S.; Rabczuk, T. Molecular Dynamics Simulations of Single-Layer Molybdenum Disulphide (MoS_2): Stillinger-Weber Parametrization, Mechanical Properties, and Thermal Conductivity. *J. Appl. Phys.* **2013**, *114* (6), 064307.

(397) Kaasbjerg, K.; Thygesen, K. S.; Jacobsen, K. W. Phonon-Limited Mobility in n-type single-layer MoS_2 from First Principles. *Phys. Rev. B* **2012**, *85* (11), 115317.

(398) Ásgeirsson, V.; Birgisson, B. O.; Bjornsson, R.; Becker, U.; Neese, F.; Riplinger, C.; Jónsson, H. Nudged Elastic Band Method for Molecular Reactions Using Energy-Weighted Springs Combined with Eigenvector Following. *J. Chem. Theory Comput.* **2021**, *17* (8), 4929–4945.

(399) Noh, J.-Y.; Kim, H.; Kim, Y.-S. Stability and Electronic Structures of Native Defects in Single-Layer MoS_2. *Phys. Rev. B* **2014**, *89* (20), 205417.

(400) Qiu, H.; Xu, T.; Wang, Z.; Ren, W.; Nan, H.; Ni, Z.; Chen, Q.; Yuan, S.; Miao, F.; Song, F.; Long, G.; Shi, Y.; Sun, L.; Wang, J.; Wang, X. Hopping Transport through Defect-Induced Localized States in Molybdenum Disulphide. *Nat. Commun.* **2013**, *4* (1), 2642.

(401) Coelho, P. M.; Komsa, H.-P.; Coy Diaz, H.; Ma, Y.; Krasheninnikov, A. V.; Batzill, M. Post-Synthesis Modifications of Two-Dimensional $MoSe_2$ or $MoTe_2$ by Incorporation of Excess Metal Atoms into the Crystal Structure. *ACS Nano* **2018**, *12* (4), 3975–3984.

(402) Löwdin, P.-O. On the Non-Orthogonality Problem Connected with the Use of Atomic Wave Functions in the Theory of Molecules and Crystals. *J. Chem. Phys.* **1950**, *18* (3), 365–375.

(403) Maintz, S.; Deringer, V. L.; Tchougréeff, A. L.; Dronskowski, R. Analytic Projection from Plane-Wave and PAW Wavefunctions and Application to Chemical-Bonding Analysis in Solids. *J. Comput. Chem.* **2013**, *34* (29), 2557–2567.

(404) Soper, A. K. GudrunN and GudrunX : Programs for Correcting Raw Neutron and X-ray Diffraction Data to Differential Scattering Cross Section. *Rutherford Appleton Laboratory Technical Reports* **2011**.

(405) Lafuente, B.; Downs, R. T.; Yang, H.; Stone, N. In *Highlights in Mineralogical Crystallography*, Armbruster, T., Danisi, R. M., Eds.; De Gruyter (O): 2015, pp 1–30.

(406) Kresse, G.; Furthmüller, J. Efficient Iterative Schemes for Ab Initio Total-Energy Calculations Using a Plane-Wave Basis Set. *Phys. Rev. B* **1996**, *54* (16), 11169–11186.

(407) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A Consistent and Accurate Ab Initio Parametrization of Density Functional Dispersion Correction (DFT-D) for the 94 Elements H-Pu. *J. Chem. Phys.* **2010**, *132* (15), 154104.

(408) Harris, V. G.; Aylesworth, K. D.; Das, B. N.; Elam, W. T.; Koon, N. C. Structural Origins of Magnetic Anisotropy in Sputtered Amorphous Tb-Fe Films. *Phys. Rev. Lett.* **1992**, *69* (13), 1939–1942.

(409) Hufnagel, T. C.; Brennan, S.; Zschack, P.; Clemens, B. M. Structural Anisotropy in Amorphous Fe-Tb Thin Films. *Phys. Rev. B* **1996**, *53* (18), 12024–12030.

(410) Shuleiko, D. V.; Kashaev, F. V.; Potemkin, F. V.; Zabotnov, S. V.; Zoteev, A. V.; Presnov, D. E.; Parkhomenko, I. N.; Romanov, I. A. Structural Anisotropy of Amorphous Silicon Films Modified by Femtosecond Laser Pulses. *Opt. Spectrosc.* **2018**, *124* (6), 801–807.

(411) Fujikake, S.; Deringer, V. L.; Lee, T. H.; Krynski, M.; Elliott, S. R.; Csányi, G. Gaussian Approximation Potential Modeling of Lithium Intercalation in Carbon Nanostructures. *J. Chem. Phys.* **2018**, *148* (24), 241714.

Appendices

A | Understanding defects in a-Si with million-atom structures

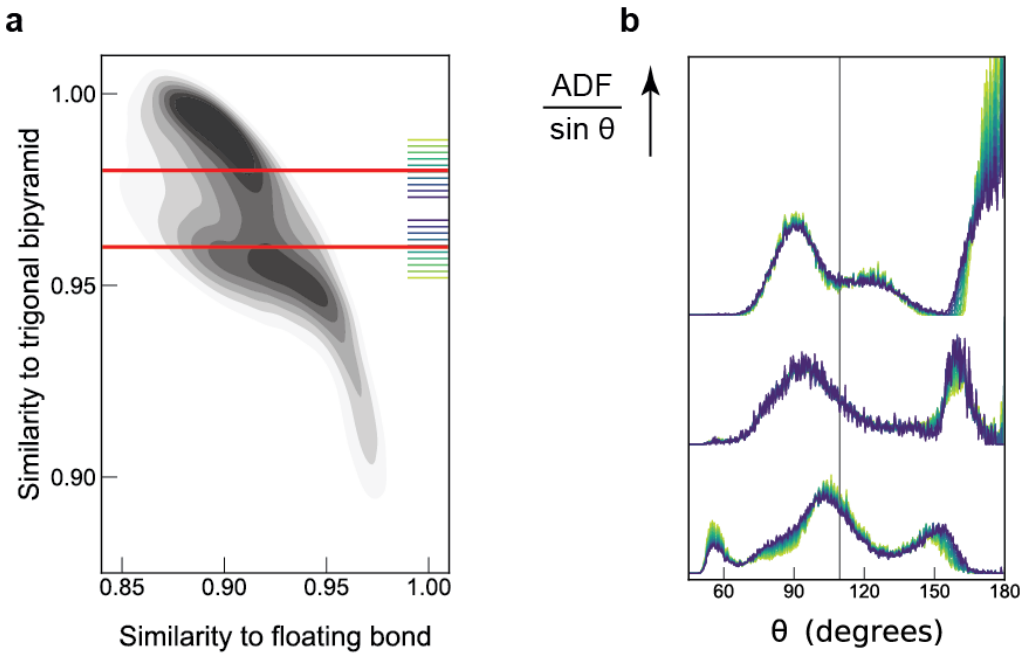


Figure A.1: Classification methods for defects. (**a**) SOAP similarity values used for defining the three classes of fivefold defect (red lines) in Fig. 4.3. Structures with a similarity greater than the upper red line are categorised as trigonal-bipyramid-like, those between the lines as square-pyramidal-like, and those below as 'floating bond'-like. (**b**) Sensitivity of ADFs to similarity cut-off values. ADFs are shown in colours purple to yellow corresponding to a range of similarity values in panel a of the same colour. The form of the ADF is quite insensitive over a wide range of similarity values; therefore the results in Fig. 4.3 are not strongly affected by the exact cut-off value for the similarity to distinguish classes.

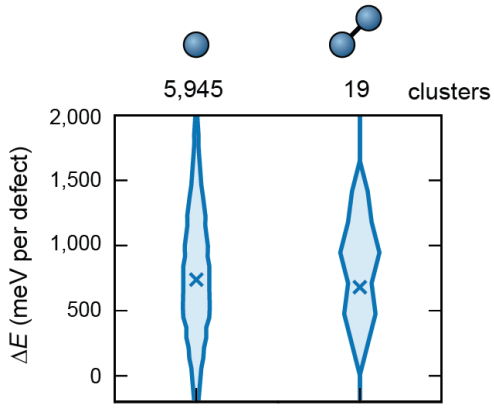


Figure A.2: Absence of substantial clustering of 3-fold defects in amorphous silicon. Similar to Fig. 4.6, local energy distributions for the largest defect cluster containing only 3-fold atoms are displayed next to the corresponding distribution for an isolated 3-fold defect. As with Fig. 4.6, the defect cluster energies are calculated by summing the individual atomic energies of defect cores and their immediate topological neighbours up to 3 bonds away relative to the mean a-Si energy (excluding other defects) and are reported per-coordination defect. Crosses indicate the mean of each distribution.

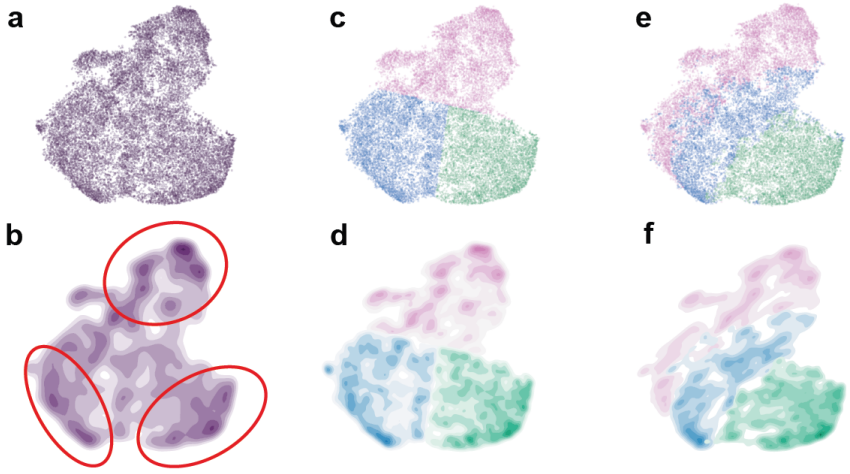


Figure A.3: Comparison of clustering methods. Scatter (upper) and gaussian KDE (lower) plots are shown for each of the following: (**a**–**b**) the t-SNE projection of the kernel matrix between SOAP vectors for all 5-fold atoms; (**c**–**d**) colouring by classification label derived from bisecting k-means clustering—an automated method to draw the class boundaries; (**e**–**f**) colouring by classification label derived from comparison with idealised trigonal bipyramidal and floating bond geometries as described in Fig. 4.3. The three most densely populated regions of the map are consistently identified by the supervised and unsupervised classification approaches (indicated by red rings). The areas throughout the centre of the map and in the lower left corner are more uncertain. 74% of environments are put in the same category by both methods.

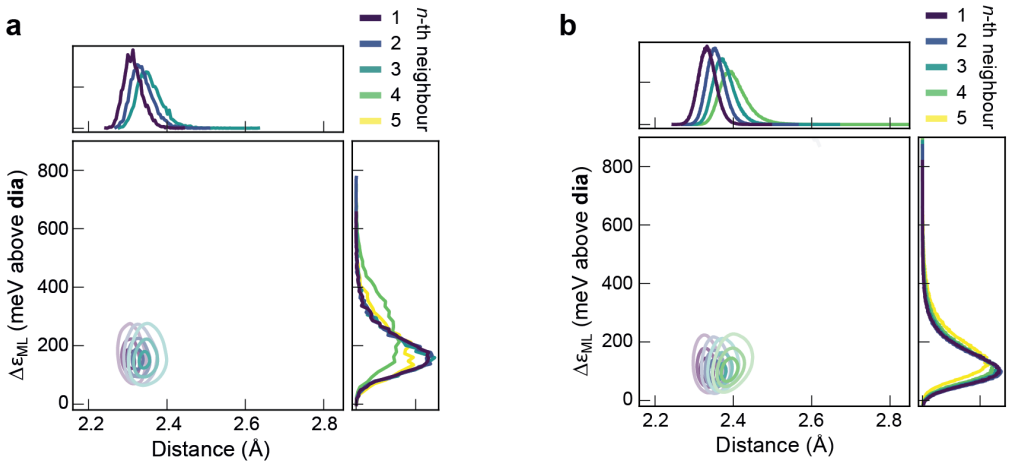


Figure A.4: 2D Correlation of local energy with enumerated bond distances for (**a**) 3-fold and (**b**) 4-fold atoms, as in Fig. 4.1d.

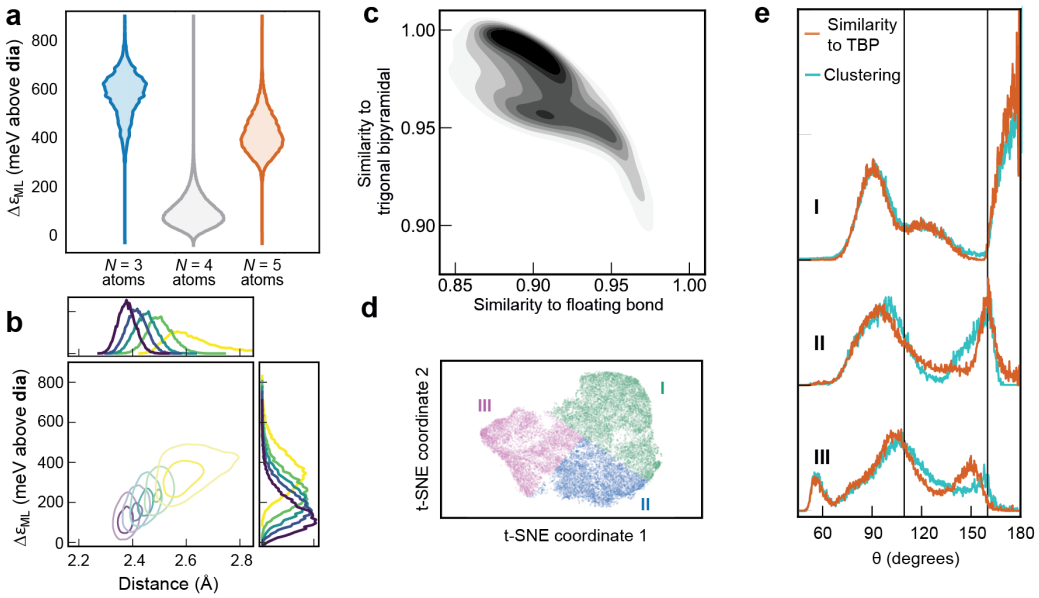


Figure A.5: An alternative 1-million atom structure from rapid quenching and annealing. We repeat analyses as in the main text for a separate, entirely uncorrelated 1M atom structure, which was produced with a slightly different protocol: rapid cooling at $10 \times 10^{13}\,\mathrm{K\,s^{-1}}$ followed by a long annealing period of 2 ns at 840 K. This structure has a similar number of 5-fold atoms as the structure discussed in chapter 4, which was derived from slower cooling at $10 \times 10^{11}\,\mathrm{K\,s^{-1}}$. (**a**) Distributions of the ML atomic energies, $\varepsilon_{\mathrm{ML}}$, of defects in the model, shown separately for different nearest-neighbour coordination numbers, N. (**b**) Two-dimensional correlation plot of the neighbour density for 5-coordinated defects (as in panel e). (**c**) 2D plot of SOAP kernel similarity of $N = 5$ atoms in the a-Si model to the idealised TBP and floating-bond environments, respectively. The distribution of values for individual atoms is shown as a heat map. (**d**) Unsupervised classification of 5-fold atoms. The full distance matrix is embedded in 2D with the dimensionality-reduction algorithm t-SNE, where **K** is the kernel matrix built from the similarity of each 5-fold atom with every other 5-fold atom. Bisecting k-means is used to identify clusters **I**–**III**. (**e**) Bond-angle distribution function (BADF) plots, scaled by $\sin\theta$, for atomic triples centred on all 5-fold coordinated atoms. The BADFs are plotted separately for the three distinct categories of 5-fold defects related to idealised structures respectively from top to bottom (as illustrated with selected examples of such configurations from the a-Si model). The 5-fold atoms are separated into these categories via two methods but with similar results: by comparison to the idealised structures of panels a and b via SOAP similarity (orange) and from unsupervised clustering (cyan). Vertical black lines at the ideal tetrahedral angle (109.5°) and at 160° are guides for the eye.

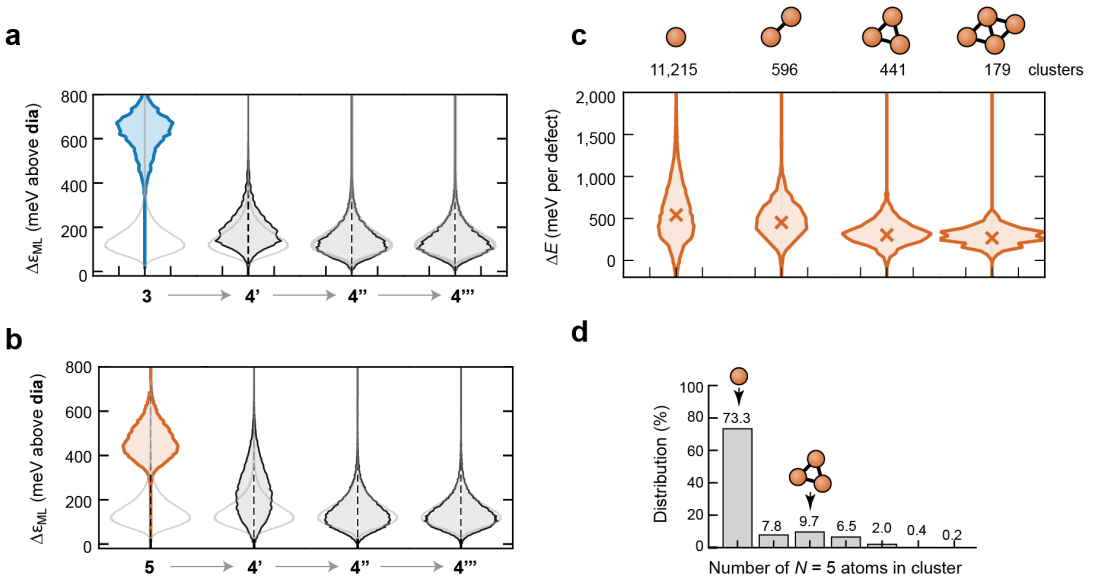


Figure A.6: Locality effects in the alternative 1-million atom structure. Further analysis of the locality of defects for the annealed 1-million atom structure of Fig. A.5. (**a**) Distribution of ML local energies for $N = 3$ atoms and their surroundings. The distribution for bulk a-Si is shown in light grey. (**b**) Same but for $N = 5$ atoms and their surroundings. (**c**) ML-predicted energy distributions for the most common 5-fold defect clusters, with the number of occurrences of each structure directly above each bar. Defect cluster energies are calculated by summing the individual atomic energies of defect cores and their immediate topological neighbors up to 3 bonds away relative to the mean a-Si energy and are reported per-coordination defect, as described in chapter 4. Crosses indicate the mean of each distribution. (**d**) Statistics for the number of occurrences of clustered 5-coordination defects.

B | The hierarchical structure of amorphous MoS_3

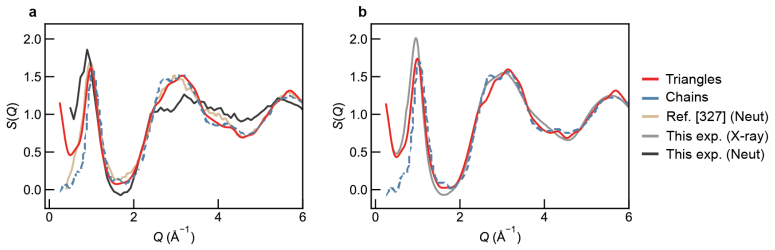


Figure B.1: Expanded version of $S(Q)$ in Fig. 5.5c.

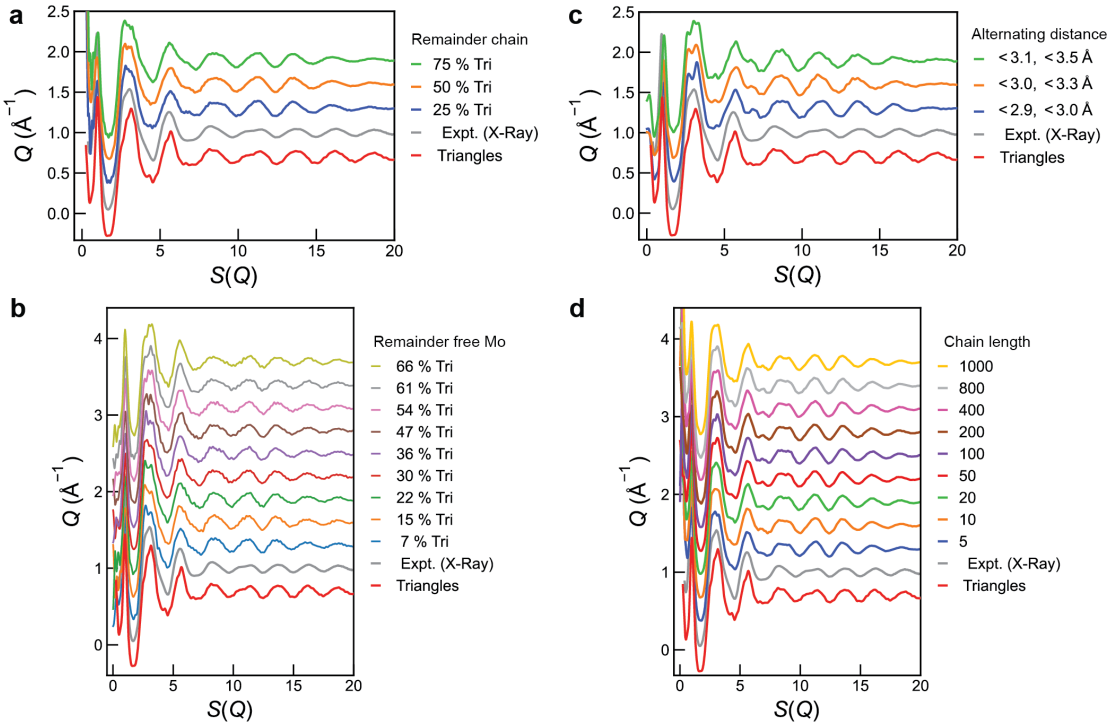


Figure B.2: Additional $S(Q)$ results for alternative chain and triangle models. In each case, I include our experimental X-ray $S(Q)$ and the main triangle model of Fig. 5.5c for comparison. (**a**) $S(Q)$s for models with a percentage of Mo atoms in chains. The legend gives the proportion of Mo atoms that are bound together in triangles. (**b**) Same as a, but with a percentage of Mo atoms allowed to evolve with no constraints ('free Mo'). For both a and b, the high Q region disagrees with experiment when either chains or free Mo are included in the structure in any proportion. (**c**) Chain models where distances between Mo atoms in chains are forced to alternate, with upper limits for each characteristic bond length given in the legend. (**d**) Chain models for a series of characteristic chain lengths given in the legend.

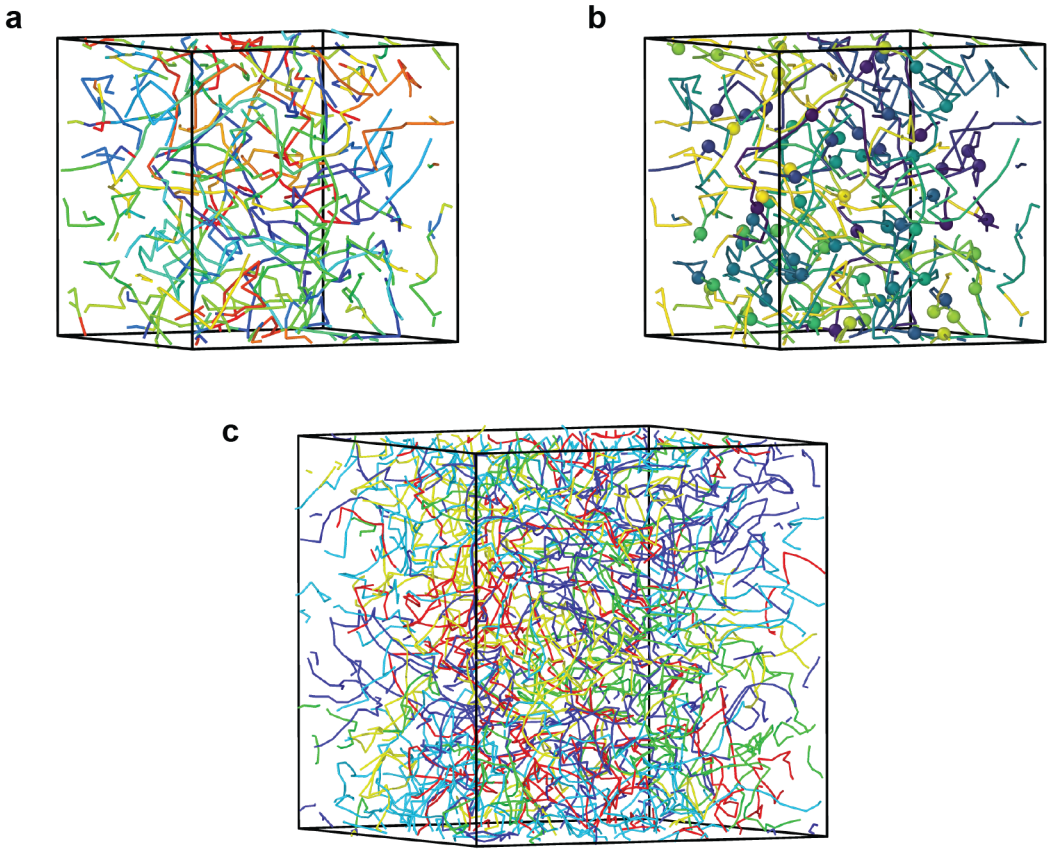


Figure B.3: Structural illustration of the chain models. In each case, all sulfur atoms are hidden and only Mo–Mo bonds under 3.1 Å are displayed. (**a**) Segments of chains highlighted. I randomly sample the continuous chain of the main model discussed in chapter 5 (4,000 atoms) for sections of length 50 and colour each segment randomly to guide the eye. (**b**) Continuous chain highlighted. Same as panel a, except the chain is coloured continuously for contiguous atoms. Every 10^{th} atom is highlighted with a ball. (**c**) Separate 20,000 atom chain model. I colour each of the separate chains (each of length ≈ 200) differently for the model in Fig. B.2d labelled '200'.

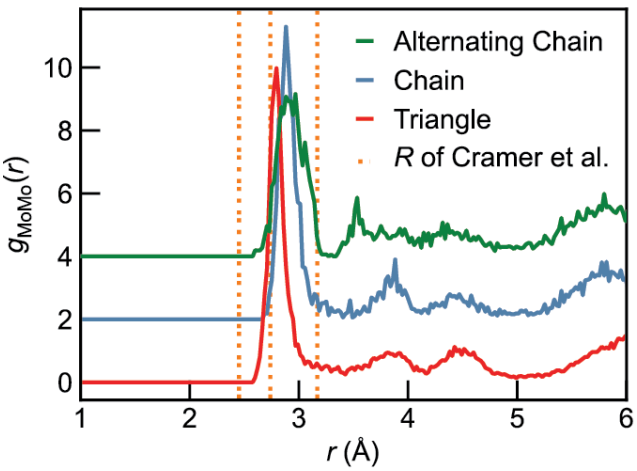


Figure B.4: Mo–Mo g(r) for alternating chain, corresponding to the < 3.0, < 3.3Å model in Fig. B.2

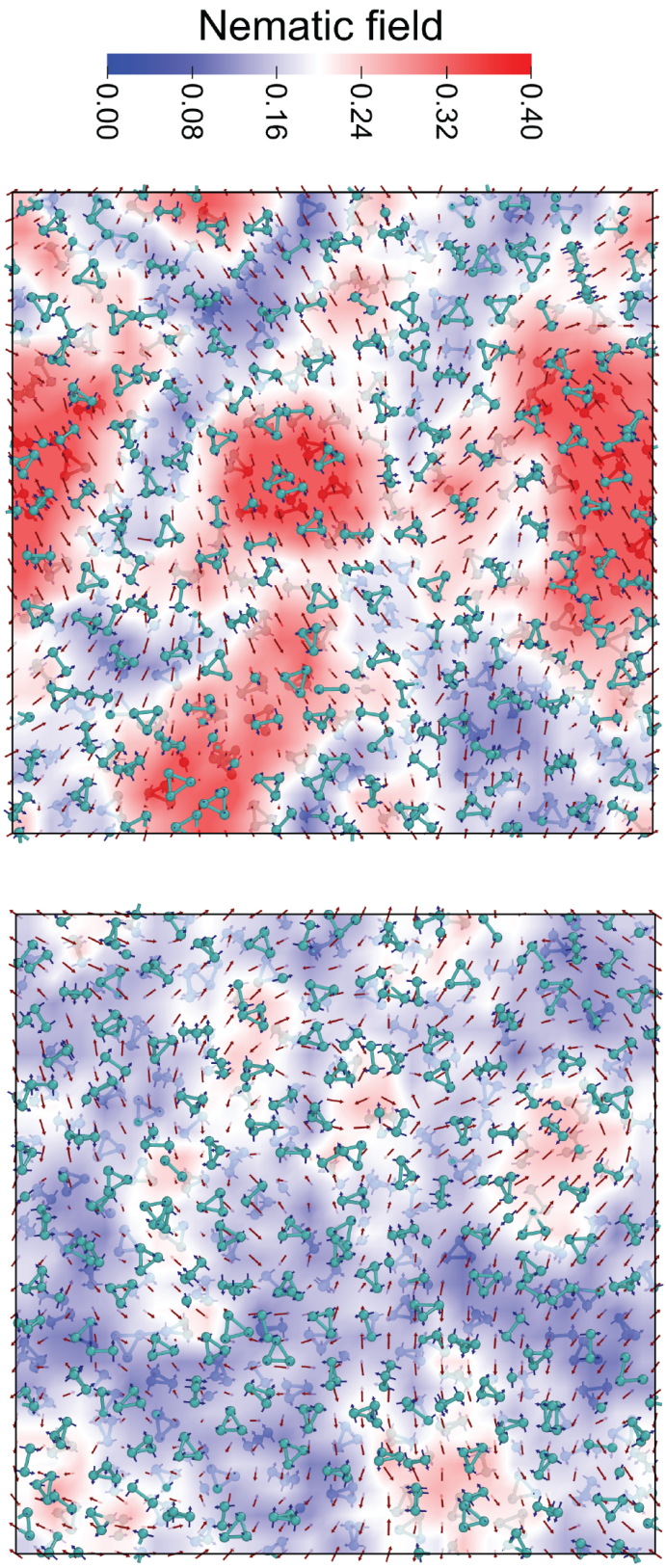


Figure B.5: Enlarged image size for the same nematic fields that are shown in Fig. 5.11.