Informed modeling of disorder

From spaghettis to meatballs

Phillip M. Maffettone
Corpus Christi College, Oxford



A thesis submitted for the degree of
Doctor of Philosophy

Department of Chemistry
University of Oxford

Trinity 2018

The work presented in this thesis was carried out between April 2015 and May 2018 in the Inorganic Chemistry Laboratory, University of Oxford under the supervision of Prof. Andrew L. Goodwin. All of the work is my own unless otherwise stated, and has not been submitted previously for any degree at this or any other university.

Phillip Michael Maffettone
May 2018

Informed modeling of disorder

Phillip M. Maffettone, Corpus Christi College
D.Phil., Trinity Term 2018

Understanding structure is paramount in developing structure-property relationships in all categories of materials, from hard to soft matter. Developing this understanding becomes particularly challenging in the realm of disorder, where the notion of periodicity breaks down. For systems of increasing disorder, the experimental probes of canonical crystallography contain diminishing information content, and traditional methods prove insufficient for structural solution. This thesis uses the methodology of *informed modeling* to gain insight into the structures of disordered materials, and investigate the limits of using reciprocal space information to derive meaningful real space structures.

Informed modeling procedures are data-driven approaches to extracting structural solutions from the total scattering pattern, that incorporate some instance of prior information. This inclusion of priors — which can from complementary experiments or from first principles — is an inherently Bayesian approach to modeling structure. To establish this methodology, I investigate a series of three distinct systems with increasingly severe levels of disorder, that also possess increasingly powerful prior information. This thesis is comprised of a series of monographs on structural modeling, tied together by a central inquiry, and spanning crystalline, amorphous, and solution phase materials.

The anomalous hydration behavior of ZrW_2O_8 is modeled using a set of local rules derived from experimental scattering and *ab initio* calculations. The propagation of deformations associated with these local rules throughout the structure is described by a series of spaghetti states which map onto the population of many negative thermal expansion phonons. The developed mixed modeling procedures effectively incorporate prior information from first principles, and cast a meaningful structure solution that is validated by total scattering and lattice dynamics.

Next, a system of unique topological disorder is considered. Total scattering data are combined with rigid local constraints to develop the first ever square-planar continuous random network (CRN), which is subsequently used to model an amorphous $Pd/Pt(im)_2$ phase. This model is refined against the total scattering data using reverse Monte Carlo (RMC).

Finally, an explicitly Bayesian approach is developed for refining solution phase protein structures against the pair distribution function (PDF). Although the protein PDF has very few features and limited information content, continuous prior probability distributions of local conformations are implemented in a Bayesian RMC. This proof of concept study vindicates the utility of the PDF in structure prediction, so long as a set of robust prior information is available.

The three systems are considered in concert to demonstrate when structure solution is tractable: as the data becomes less discriminating, more judicious prior information is necessary in order to approach a meaningful structural solution.

Acknowledgements

First and foremost I would like to thank my supervisor, Prof. Andrew Goodwin, for being so approachable and taking me on as a student after 6 months in an entirely different DPhil. It has often been the case that a 15 minute chat with Andrew proved more productive and thought provoking than a week's reading. I am truly fortunate to have found such a brilliant and compassionate mentor. Thank you for such a profound educational experience, and for tolerating my sporting and globe trotting.

I'd like to gratefully acknowledge all of my colleagues and collaborators for their contributions to the work presented in this thesis, and their insightful conversations regarding science in general. The ZrW_2O_8 work would not have been possible if not for Dr. Fabien Trousselet, Dr. F-X Coudert, and Mia Baise and their DFT calculations. I'm very appreciative of Prof. Jane Allison's insightful conversations regarding biological systems and molecular dynamics. So much experimental insight has been offered through collaboration, and I'm very grateful for the work of Dr. Nicholas Funnell with ZrW_2O_8, Dr. Andrew Cairns with $Pd(im)_2$, and Dr. Nicola Steinke and Dr. Sylvia McLain with the protein work. Beyond your contributions to scientific content, you have played pivotal roles in my personal development and understanding of crystallography.

The Goodwin group has been a consistent source of scientific stimulation and personal joy. Thank you Ali for being around whenever Fortran was wrong, and for high-five support when I got it right; you've been a great colleague and a better friend. Thank you to Mia and WJKF for being here from my start in the group, and kicking off some of my work. Joe has been a continual source of insight and critical thought, always opening new doors with cutting questions. Even after moving on with their careers, Matt and Andrew have been regularly helpful. I am so grateful for Emily, who read the majority of this thesis and helped me continually edit it. Beyond all of this, thank you to Harry, Phil, Emma, Chloe, Hanna, Peter, Josh,

Arkadiy, Carl, and Hamish for being part of such a phenomenal experience.

This thesis would not have been possible without the tangential experiences. I am grateful to Dr. Martin Galpin for the opportunity to teach in the department, as well as Jerome Luchesa-Smith for the formative experience as a project mentor. To the players and coaches of the OURLFC, thank you for 4 years of incredible sport and high spirits — everything I traded in physical health, you returned to me tenfold in happiness and wellbeing.

The support of my friends has been critical in this experience. I'm grateful for the gentlemen of the Living Room for providing a respite from politics and work-related science. Thank you to the friends at the Old Bookbinders for years of Tuesday nights, and a job which built community in such a transient place. Thank you Jamie and José for being the tropisk I've needed from afar, and keeping the Art in STEAM.

A special shout out to the Doctors Koch for supporting my foolish escapades across central Asia — though I'm not so sure that was constructive to this thesis — and providing the office with a coffee pot — which I am very sure served as the foundation for this thesis.

My time in the UK was supported financially by the Marshall Aid Commemoration Commission, for which I am very thankful. Though lackluster as far as a stipend is concerned, they have offered unprecedented intellectual freedom and opportunity.

On the topic of intellectual freedom, I would like to thank Prof. Susan Perkin. It was on her advice that I approached the Goodwin group, and I could not have been happier with the result.

Finally, I'd like to thank the person who will be most proud of this thesis, and will likely never read a word of it: thanks Mom for supporting me, whether I'm a barman or a barrister.

Don't blame it on the sunshine

Don't blame it on the moonlight

Don't blame it on the good times

Blame it on the boogie

Contents

1 Introduction **1**

1.1 Theory of diffraction 4

1.2 The nature of disorder 15

1.3 Structural modeling and disorder 33

1.4 Structural modeling: an inverse problem 68

1.5 Aims and structure of this thesis 72

2 Methods **75**

2.1 Experimental X-ray scattering 75

2.2 DFT 79

2.3 Empirical simulations 83

2.4 Monte Carlo simulations 85

3 Negative hydration expansion in ZrW_2O_8 **92**

3.1 Background 93

3.2 Developing local constraints 105

3.3 Modeling a supercell 108

3.4 Results 112

3.5 Conclusions 117

4 Square-planar amorphous zeolitic imidazolate frameworks **119**

4.1 Amorphous metal-organic frameworks 120

4.2 Methods 126

4.3 Results and discussion 140

4.4 Conclusions and future work 153

5 Bayesian approach for information poor structure determination 156

5.1 Background 159

5.2 Algorithm 173

5.3 Results and discussion 183

5.4 Further developments 200

5.5 Conclusions 202

6 Conclusions and outlook 204

Bibliography 210

Publications 236

A Supplementary information: ZrW_2O_8 243

A.1 Density functional theory calculations 243

A.2 Detailed refinements 251

A.3 Configurational entropy of the spaghetti phase 253

B Supplementary data: square-planar CRNs 255

C Relating chemical shift and structure 264

C.1 Structure to chemical shift 264

C.2 Chemical shift to structure 268

D Melittin prior probability distributions 275

E PIPA side chain decoration 291

F Mixed media 296

List of Figures

1.1 Scattering of radiation 5

1.2 Schematic of increasing degrees of disorder 16

1.3 Ising model 20

1.4 Complex tiling patterns in liquid crystal honeycombs. 21

1.5 Square lattice compositional order, random and correlated disorder . 23

1.6 Diffuse scattering of higher order correlations on a square lattice. . . 25

1.7 Short-range correlations in Cu_3Au 26

1.8 Square ice. 28

1.9 Examples of scattering from topologically disordered states. 31

1.10 PDFs from disordered carbons 39

1.11 Single crystal diffuse scattering map for Cu_3Au. 42

1.12 Model 3D-ΔPDF maps. 44

1.13 Structures and diffuse scattering of stabilized cubic zirconias. 47

1.14 Thermal disorder in β-cristobalite 49

1.15 Diffuse scattering in MD simulated lead zirconate. 51

1.16 Early models of amorphous materials. 53

1.17 Amorphous silica CRN 56

1.18 PDF refinement in $Ga_{1-x}In_xAs$ 58

1.19 Competing models of $LaMnO_3$. 60

1.20 Cation disorder in Prussian blue analogs. 61

1.21 Comparison of different methods for modeling a-Si 63

1.22 A demonstrative energy landscape 66

1.23 Homometric system of points. 70

2.1 Powder averaging and experimental powder X-ray diffraction set up . 76
2.2 Simulated annealing flowchart 88

3.1 Plot of lattice parameter as a function of water content in $ZrW_2O_8{\cdot}H_2O$ 94
3.2 Select phases of ZrW_2O_8 96
3.3 Harmonic and anharmonic potential well. 99
3.4 Example NTE phonon modes 103
3.5 $ZrW_2O_8{\cdot}H_2O$ structures 106
3.6 Trajectory of ZrW_2O_8 hydration mechanism 108
3.7 Relaxed 5×5×5 supercell 111
3.8 Simulation behavior across supercell size. 112
3.9 ZrW_2O_8 diffraction and PDF 113
3.10 Probability isosurface of disordered $Pa\bar{3}$ $ZrW_2O_8{\cdot}H_2O$ 114
3.11 Linear combination of displacements from phonon projections. 116
3.12 Projection v Grüneisen parameters. 117

4.1 Comparative local geometry between $Zn(im)_2$ and silicates 121
4.2 ZIF-4 structure. 123
4.3 Scattering behavior of an amorphous tetrahedral ZIF 124
4.4 The crystal structure of $Pd(Him)_2(im)_2$ 125
4.5 WWW starting configuration for a-$Pd(im)_2$ generation. 129
4.6 WWW simulation flow chart 131
4.7 Illustrative bond swap. 132
4.8 Representation of decorating PdSO structure with imidazolate ions. . 137
4.9 Total scattering data for $Pd(im)_2$. 140
4.10 WWW results from three different thermostats. 143
4.11 Effect of thermostat choice on WWW. 145
4.12 Comparing WWW simulations at different temperatures. 146
4.13 Sampling from an adaptive thermostat WWW. 147
4.14 Slice of a square-planar CRN. 149

4.15 Comparison of $Pd(im)_2$ scattering data and fit. . . . 150
4.16 Structural solution of $Pd(im)_2$. . . . 151
4.17 One-dimensional spaghetti like substructure in square-planar CRNs . 153

5.1 Comparing one-dimensional PDFs across systems 161
5.2 Protein backbone structure 162
5.3 Typical Ramachandran plot. . . . 164
5.4 CryoEM single particle analysis. . . . 167
5.5 Blind structure prediction with CS-Rosetta. . . . 172
5.6 Standard amino acid backbone geometry 175
5.7 Defining a torsion angle. . . . 176
5.8 PIPA flowchart 183
5.9 Melittin structure 184
5.10 Prior probability Ramachandran distributions 185
5.11 Prior probability product distributions 187
5.12 Prior probability TALOS-N distributions 188
5.13 Threshold RMSD 190
5.14 Success rate for Ramachandran distribution sampling 191
5.15 Success rate for product distribution sampling 193
5.16 Success rate for TALOS-N distribution sampling 194
5.17 Success rate for incomplete data. . . . 195
5.18 Scattering only result. . . . 196
5.19 Ramachandran only result. . . . 197
5.20 TALOS-N only result. . . . 198
5.21 Prior information triangle 199
5.22 RMSD dependence 200
5.23 Larger protein structures generated by PIPA 202

A.1 DFT mode Grüneisen parameters. . . . 249
A.2 ZrW_2O_8 Rietveld refinements. . . . 251

A.3 ZrW_2O_8 PDF refinement. 252
A.4 $ZrW_2O_8{\cdot}H_2O$ PDF refinement. 252
A.5 $ZrW_2O_8{\cdot}H_2O$ PDF refinement. 253

B.1 Available four-coordinate nets. 256
B.2 Unphysical voids in a CRN. 257
B.3 RMC fit for different densities. 258
B.4 $Pd(im)_2$ geometry histograms for different densities. 259
B.5 Refined $Pd(im)_2$ at *a*-ZIF density. 260
B.6 Refined $Pd(im)_2$ at mid-range density. 261
B.7 Refined $Pd(im)_2$ at corrected $Pd(Him)_2(im)_2$ density density. 262
B.8 Detailed total scattering refinements for a D_{4h} model. 263
B.9 Detailed total scattering refinements for a T_d model. 263

D.1 Neighbor dependent Ramachandran distributions for melittin residues. 276
D.2 Neighbor dependent Ramachandran distributions for melittin residues. 277
D.3 Neighbor dependent Ramachandran distributions for melittin residues. 278
D.4 As produced distributions from TALOS-N for melittin residues. . . . 279
D.5 As produced distributions from TALOS-N for melittin residues. . . . 280
D.6 As produced distributions from TALOS-N for melittin residues. . . . 281
D.7 As produced distributions from DANGLE for melittin residues. . . . 282
D.8 As produced distributions from DANGLE for melittin residues. . . . 283
D.9 As produced distributions from DANGLE for melittin residues. . . . 284
D.10 Combined TALOS-N and Ramachandran distributions. 285
D.11 Combined TALOS-N and Ramachandran distributions. 286
D.12 Combined TALOS-N and Ramachandran distributions. 287
D.13 Combined DANGLE and Ramachandran distributions. 288
D.14 Combined DANGLE and Ramachandran distributions. 289
D.15 Combined DANGLE and Ramachandran distributions. 290

E.1 Flowchart for the side chain decoration program 292
E.2 Oriented polytope basis vectors. 294
E.3 Building bounding volumes. 295

List of Tables

1.1 Limiting behavior of total-scattering functions 14

1.2 Simulation of amorphous silicon using bond-switching methods 55

3.1 $ZrW_2O_8{\cdot}H_2O$ DFT coordinates 107

3.2 ZrW_2O_8 empirical potentials 110

4.1 WWW simulation parameters 142

4.2 Square-planar CRN statistics 144

5.1 Ranges of (ϕ, ψ) values associated with secondary structures 165

5.2 Available annealing schedules in PIPA 182

5.3 Likelihood quotients 186

A.1 DFT $ZrW_2O_8{\cdot}H_2O$ structural coordinates. 245

A.2 DFT based W–O distances. 246

A.3 Structural and energetic comparison for DFT configurations. 247

A.4 Thermal expansion coefficients for anhydrous and hydrated structures. 250

C.1 Comparison of chemical shift based structure predictors. 274

Abbreviations

1D	one-dimensional
2D	two-dimensional
2θ	scattering angle (diffraction)
3D	three-dimensional
3D-ΔPDF	3D difference pair distribution function
ADP	atomic displacement parameter
a-	amorphous
ANN	artificial neural network
ART	activation-relaxation technique
BFGS	Broyden–Fletcher–Goldfarb–Shanno algorithm
BM	Barkema-Mousseau
BZ	Brillouin zone
CMR	colossal magneto resistance
CRN	continuous random network
CovFeFe	covalent iron–iron interactions
CPU	central processing unit
DFT	density functional theory
EXAFS	extended X-ray absorption fine structure
EM	electron microscopy
FEM	fluctuation electron microscopy
FFT	fast Fourier transform
GGA	generalized gradient approximation
GULP	General Utility Lattice Program
LDA	local density approximation
LJ	Lennard-Jones
MC	Monte Carlo

MCMC	Markov chain Monte Carlo
MD	molecular dynamics
MHMC	Metropolis-Hastings Monte Carlo
MOF	metal organic framework
NHE	negative hydration expansion
NOE	nuclear Overhauser effect
NTE	negative thermal expansion
PDB	Protein Data Bank
PDF	pair distribution function
PIA	pressure induced amorphization
PIPA	Prior Information Protein folding Algorithm
RCSB	Research Collaboratory for Structural Bioinformatics
RDC	residual dipolar coupling
RMSD	root mean squared deviation
RUM	rigid unit mode
SA	simulated annealing
SAS	small angle scattering
SRO	short-range order
SW	Stillinger-Weber
TDS	thermal diffuse scattering
TS	total scattering
WWW	Wooten, Winer, and Weaire method
XRD	X-ray diffraction
XFEL	X-ray free electron laser
ZIF	zeolitic imidazolate framework

Chapter 1

Introduction

It is difficult to overstate the significance of structure in the physical sciences. The properties of solids are not solely dependent on their chemical composition, but the arrangement of their constituent atoms in three-dimensional space. The prototypical polymorphism of carbon illustrates this: the difference between coal, diamond, and graphite lies in their respective structures [1]. It is the atomic structure of a mineral that determines its macroscopic crystal habit [2]. This same structure determines the electronic properties of metals and semiconductors. The amorphous structure of glasses determines their tensile response, viscosity, and optical transparency [3]. Outside of the solid state, protein biological activity is governed by three-dimensional structure of a finite variety of repeat units [4]. So I attest with no fear of becoming the scientific equivalent of a broken record: structure determines function!

It is thus the imperative of the structural scientist to determine this three-dimensional arrangement, and develop subsequent relationships to the myriad of functionalities observed in the physical sciences. This has been the task of crystallographers since von Laue's discovery of diffraction of X-rays by crystalline materials, and Bragg's seminal work in using the diffraction patterns of X-rays to determine the crystal structures of sodium chloride, potassium chloride, potassium bromide [5], and that of diamond [6, 7]. Even 100 years later, our strongest microscopes are just beginning to attain atomic resolution in two dimensions [8] or topographically [9, 10]. So diffraction experiments have emerged and remain the key experimental tool for determining the structure of materials.

Conventional crystallography depends on the notion of periodicity: a crystal consisting of a three-dimensional array of identical repeat units. This is neatly cast in the notion of Bragg diffraction (the diffraction associated with sets of evenly

spaced atomic planes described by Bragg's law). Effectively, Bragg diffraction offers us average one-body correlations, including atomic positions, thermal ellipsoids and site occupancies [11]. An average, however, does Nature a disservice — to average a zebra's coat would leave us with an out of place ass — and limits our understanding of the essential structure underpinning the material. There is in fact rich two-body information underneath the Bragg peaks, stemming from diffuse scattering [12]. These two-body correlations can help us elicit how atoms and molecules in a material interact. In practice, we expect any deviation from the ideal crystal lattice to manifest in the diffuse scattering. Together, these profiles make up the total scattering, and represent the Fourier transform of the real space structure.

This comes with a compelling caveat: the experimentally observed scattering is in fact the squared modulus of the Fourier transform. All information about the phase of the Fourier transform is lost, and a trivial inversion becomes impossible. Known as the 'phase problem' in crystallography, this has been the impetus for 100 years of work dedicated to solving it [13]. This century of scientific fortitude has shown that classic crystallographic techniques can indeed be applied to determine structures of ordered matter. From minerals to magnets and from proteins to elemental solids, the framework of canonical crystallography has demonstrated abundant success across disciplines. Even so, the challenge of extracting real space positions from reciprocal space intensities becomes intimately more challenging as we stray from assumptions of order, and venture into regions of disorder. Herein lies the central question of this thesis. As we examine more disordered materials, in what scenarios can we meaningfully invert the diffraction pattern — and the one- and two-body correlation functions it contains — into a structural model?

Despite the challenges in comprehending their atomic structure, disordered materials are nonetheless ubiquitous in technology and daily use. Metallic glasses — amorphous materials that produce no Bragg diffraction — make use of unique deformation mechanics and easy processing [14], whereas covalent glasses are inescapable

to the chemist [15]. Biominerals range in composition but form well structured amorphous phases from bones to shells [16, 17]. Household technology has been impacted by amorphous chalcogenide based materials used for repeatable phase changes in nonvolatile memory [18, 19], liquid crystals in display screens [20], and amorphous semiconductors [21]. $BaTiO_3$ is commonly used as a ferroelectric/dielectric, and is widely studied for its complicated structure [22].

Contrary to the common expectation that impurities or defects are to be avoided, there is growing appreciation that the unique and exploitable properties in these materials actually arise from disorder [23–25]. Anomalous thermal expansion has been cast as a consequence of structural disorder in ZrW_2O_8 [26, 27]. Some zeolitic imidazolate frameworks undergo a pressure induced amorphization [28, 29] amongst other anomalous mechanical properties [30–32], due in part to their structural flexibility. Geometric frustration is at the root of colossal magnetoresistance in manganites [33, 34], orientational correlations in plastic crystals [35], and collective transport properties of superionics and elements [36, 37]. Residual entropy in 'ice'-like systems impacts the magnetism of $Ho_2Ti_2O_7$ [38, 39] and has been suggested as a mechanism for the lattice dynamics governing colossal negative thermal expansion in $Cd(CN)_2$ [40]. This entropy will dominate thermodynamic and lattice dynamical phenomena, producing disorder in phonons that is exploited in thermal conductivity [41]. Lastly, correlated disorder is also responsible for the stabilization of polar nanoregions in relaxor ferroelectrics, producing their advantageous dielectric properties [42, 43]. These are just a few of the innumerable cases of disorder in fascinating and functional materials.

The research presented in this thesis will begin to press the boundaries of our capability to derive meaningful structural models of disordered materials. By studying a series of systems with increasingly severe levels of disorder, I will grasp at the fundamental limits of the scattering information content, whilst developing novel data-driven procedures for overcoming its inherent limitations. In contributing to

a community's quest for any panacea, it becomes necessary to explore a breadth of subject matter. As classical crystallography has reached across the disciplinary silos, so too will this thesis. Contained herein is a series of monographs on structural modeling, tied together by a central inquiry, and spanning crystalline, amorphous, and solution phase materials. To aid our exploration, each chapter will be prefaced by an introduction of scientific and historic content unique to the system of interest. The introductory material presented in this chapter will cover the relevant background that lies at the foundation of this thesis: we discuss the theoretical principles of diffraction, the nature of and catalogue of disorder, the previous approaches to understanding disordered structures, and finally a perspective on these challenging inverse problems in the environment of information poverty.

1.1 Theory of diffraction

It has become clear that our primary experimental tool for tackling structural models is diffraction. More specifically, this thesis will make extensive use of *total scattering* (TS) and the *pair distribution function* (PDF). The following is by no means an exhaustive development of modern diffraction, and will draw extensively from the text of Dove [44] and the notation of Keen [45]. More comprehensive discussions of crystallography [13] and the PDF [12, 25, 46] can be found elsewhere.

We begin with the wave equation

$$\psi(\mathbf{x}) = A \exp(\mathrm{i}\mathbf{k} \cdot \mathbf{x}) \tag{1.1}$$

and the scattering of radiation from two particles separated by vector $\mathbf{r}$, where $\mathbf{k}_\mathrm{i}$ is the incident wave vector and $\mathbf{k}_\mathrm{s}$ is the scattered wave vector [Fig. 1.1]. The phase difference between the two waves scattered by the two particles is given by the dot product

$$\mathbf{Q} \cdot \mathbf{r}, \tag{1.2}$$

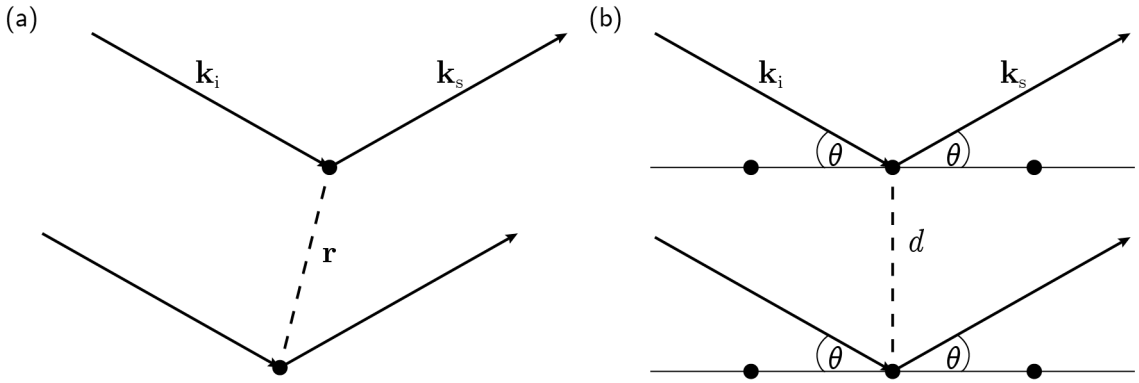


Figure 1.1: Scattering of radiation. Shown from (a) two particles in free space and (b) planes of atoms.

where we define a scattering vector $\mathbf{Q}$ as

$$\mathbf{Q} = \mathbf{k}_\mathrm{i} - \mathbf{k}_\mathrm{s}. \tag{1.3}$$

Thus, the waves scattered from the particles have the wave equations ψ_1 and ψ_2 respectively:

$$\psi_1(\mathbf{x}) = \exp(\mathrm{i}\mathbf{k}_\mathrm{s} \cdot \mathbf{x}), \tag{1.4}$$

$$\psi_2(\mathbf{x}) = \exp(\mathrm{i}\mathbf{k}_\mathrm{s} \cdot \mathbf{x}) \exp(\mathrm{i}\mathbf{Q} \cdot \mathbf{r}). \tag{1.5}$$

The total scattering is offered by the sum of these two waves

$$\psi_1(\mathbf{x}) + \psi_2(\mathbf{x}) = \exp(\mathrm{i}\mathbf{k}_\mathrm{s} \cdot \mathbf{x}) \left[1 + \exp(\mathrm{i}\mathbf{Q} \cdot \mathbf{r})\right], \tag{1.6}$$

and thus the amplitude of the scattered wave is modified by a phase factor $F(\mathbf{Q})$,

$$F(\mathbf{Q}) = 1 + \exp(\mathrm{i}\mathbf{Q} \cdot \mathbf{r}). \tag{1.7}$$

This definition is notable, as the observable interference pattern is not sensitive to the wave vectors $\mathbf{k}_\mathrm{i}$ and $\mathbf{k}_\mathrm{s}$, but to their difference. Next we consider the scattering from a collection of particles, which can be generalized by a sum of phase shifts from

particles with position vectors $\mathbf{r}_j$:

$$F(\mathbf{Q}) = \sum_j \exp(\mathrm{i}\mathbf{Q} \cdot \mathbf{r}_j). \tag{1.8}$$

It is straightforward to then consider this *structure factor* $F(\mathbf{Q})$ from a continuous density distribution:

$$F(\mathbf{Q}) = \int \rho(\mathbf{r}) \exp(\mathrm{i}\mathbf{Q} \cdot \mathbf{r}) \, \mathrm{d}\mathbf{r}. \tag{1.9}$$

The scattering vector $\mathbf{Q}$ can be related back to the commonly known Bragg's law for scattering from lattice planes *via* its magnitude:

$$\lambda = 2d \sin\theta, \tag{1.10}$$

$$Q = \frac{4\pi}{\lambda} \sin\theta = \frac{2\pi}{d}, \tag{1.11}$$

where λ is the radiation wavelength, d is the separation distance between the planes, and θ is the scattering angle.

It is the experimental reality that we cannot directly measure the phase shift as a function of $\mathbf{Q}$, but the intensity of the shift, which is proportional to the square of the amplitude. This can be expressed for a discrete collection of particles

$$I(\mathbf{Q}) \propto |F(\mathbf{Q})|^2 = \sum_{j,k} \exp\left[\mathrm{i}\mathbf{Q} \cdot (\mathbf{r}_j - \mathbf{r}_k)\right], \tag{1.12}$$

or for a continuous scatterer density distribution

$$I(\mathbf{Q}) \propto \int\int \rho(\mathbf{r}_j - \mathbf{r}_k) \exp[\mathrm{i}\mathbf{Q}(\mathbf{r}_j - \mathbf{r}_k)] \, \mathrm{d}\mathbf{r}_j \mathrm{d}\mathbf{r}_k. \tag{1.13}$$

Herein lies the task of crystallographers. The theory we have developed provides a mapping between atomic position relative to a common origin and the phase shift of scattering, which contains both real and imaginary components; however, we are only capable of measuring the intensity $|F(\mathbf{Q})|^2$ and have lost information about

the phase of $F(\mathbf{Q})$. From the intensity, we only have information about the vectors between atoms and not their absolute positions. Deemed the *phase problem* in crystallography, this is at its essence a problem of information entropy in an inverse problem.

To elicit what sort of information content is available from diffraction — an indirect measurement of structure — we continue our discussion with crystalline materials that contain different types of atoms.

1.1.1 Diffraction in crystalline materials

Extending our analysis to crystalline materials, let us first formally define the ideal ordered state. A perfect crystal is composed of a motif and a lattice, an infinite periodic array of points. These points can be completely defined by three basis vectors: $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$. The vector between any two points on the lattice can be described by a linear combination of these vectors:

$$\mathbf{t}_{[UVW]} = U\mathbf{a} + V\mathbf{b} + W\mathbf{c}. \tag{1.14}$$

Using the basis vectors as the edges of the crystallographic unit cell with a single atom, the total lattice can be thought of mathematically as a summation of Dirac delta functions:

$$\mathcal{L}(x) = \sum_{U,V,W} \delta[\mathbf{r} - (U\mathbf{a} + V\mathbf{b} + W\mathbf{c})]. \tag{1.15}$$

Finally, the total crystal structure can be expressed as a convolution of this lattice with a motif (*i.e.*, the atomic scattering density of the repeat unit $\rho_0(\mathbf{r})$):

$$\rho(\mathbf{r}) = \sum_{U,V,W} \delta[\mathbf{r} - (U\mathbf{a} + V\mathbf{b} + W\mathbf{c})] * \rho_0(\mathbf{r}). \tag{1.16}$$

This expression of the crystal as a function of scatterer density is consistent with the primary experimental probe of average structure. In the case of X-ray diffraction, the function we eventually calculate is electron density.

It is useful to think of the crystal structure in terms of a convolution between the lattice, the unit cell, the electrons in atoms, and thermal motion. This allows us to make use of the convolution theorem and consider the Fourier transform of the crystal structure as the product of the Fourier transforms of those components. Firstly the Fourier transform of the lattice [eq. (1.15)] is the reciprocal lattice

$$\mathcal{R}(\mathbf{Q}) = \sum_{h,k,l} \delta[\mathbf{Q} - (h\mathbf{a}^* + k\mathbf{b}^* + l\mathbf{c}^*)], \tag{1.17}$$

which describes plane normal vectors within the real lattice. The Fourier transform of the unit cell is offered in equation (1.9), where we first derived the structure factor, considering the integral over a single unit cell. With just these two components, we have

$$\begin{aligned} F_{\text{crystal}}(\mathbf{Q}) &= F_{\text{cell}}(\mathbf{Q}) \times \mathcal{R}(\mathbf{Q}) & (1.18)\\ &= \int_{\text{cell}} \rho(\mathbf{r}) \exp(i\mathbf{Q}\cdot\mathbf{r})\,\mathrm{d}\mathbf{r} \times \sum_{h,k,l} \delta[\mathbf{Q} - (h\mathbf{a}^* + k\mathbf{b}^* + l\mathbf{c}^*)]. & (1.19) \end{aligned}$$

The above is only nonzero for $\mathbf{Q} = h\mathbf{a}^* + k\mathbf{b}^* + l\mathbf{c}^*$ (*i.e.* at the reciprocal lattice points). Thus, by considering the integral over the atoms in a unit cell we arrive at

$$F_{\text{crystal}}(hkl) = \int_{\text{cell}} \rho(\mathbf{r}) \exp[2\pi\mathrm{i}(hx + ky + lz)]\,\mathrm{d}\mathbf{r}, \tag{1.20}$$

where we have expanded the real space positions of the atoms $\mathbf{r}$ in terms of their components x, y, and z. In noting that $F(hkl) = 0$ for any non integer h, k, or l, we need only consider diffraction at these discrete locations in reciprocal space. These are the *Bragg peaks* (or Bragg reflections) of the diffraction pattern.

This formulation still retains the scatterer density, which depends on the probe used. X-rays are scattered by electrons, whereas neutrons are scattered by the nucleus. This results in different integrals to determine scattering factors, both of the same form of equation (1.9). In the case of electron density, each atom possesses

a Q-dependent atomic scattering factor (also known as the *atomic form factor*) that decreases monotonically with increasing Q. This is given by

$$f(\mathbf{Q}) = \int_{\text{atom}} \rho_{\text{el}}(\mathbf{r}) \exp(\mathrm{i}\mathbf{Q} \cdot \mathbf{r}) \, \mathrm{d}\mathbf{r}. \tag{1.21}$$

When inserted into (1.20), this builds our general X-ray scattering formalism to

$$F(hkl) = \sum_j f_j(Q_{hkl}) \exp[2\pi\mathrm{i}(hx + ky + lz)]. \tag{1.22}$$

Lastly, we will consider another common contribution to the structure factor. Thermal motion is an inescapable effect in real crystals, and effectively blurs the atomic positions. Thus, each atomic position must be convolved with the spread of displacements associated with thermal motion, giving rise to another product in our structure factor. There are multiple ways to account for thermal motion, one of which is the isotropic and harmonic assumption where the distribution of atomic displacements from the average position is Gaussian. There are also anisotropic options for a more sophisticated treatment of thermal motion. In the former case, the Fourier transform of the displacements is also Gaussian and offered by

$$T_j(Q) = \exp\left(\frac{-8\pi^2 \langle u_j^2 \rangle \sin^2\theta}{\lambda^2}\right), \tag{1.23}$$

where $\langle u_j^2 \rangle$ is the mean-squared atomic displacement of atom j. It is common to use the variable $B_j = -8\pi^2 \langle u_j^2 \rangle$ as shorthand. This *temperature factor* can finally be incorporated into our single crystal structure factor, giving

$$F(hkl) = \sum_j f_j(Q_{hkl}) \exp[2\pi\mathrm{i}(hx + ky + lz)] \exp\left(\frac{-8\pi^2 \langle u_j^2 \rangle \sin^2\theta_{hkl}}{\lambda^2}\right). \tag{1.24}$$

While the atomic scattering factor acts to adjust the magnitude of the Bragg reflections with respect to the Q-dependent scattering of unique atoms, this additional factor describes the attenuation of scattering caused by thermal motion. Moreover,

this temperature factor accounts for the effects of strictly independent atomic motion on the diffraction intensities. There are of course other corrections that can be applied to the structure factor, many of which are described elsewhere [13, 44].

What this formulation offers is not just a reciprocal space representation of single crystals, but a representation of the average structure of any crystalline material, including disordered crystals. Beyond the average structure, there is information contained in the *diffuse scattering* that is increasingly relevant in cases of disorder. This scattering appears underneath and between the Bragg peaks, and when taken in concert with the Bragg scattering is called *total scattering*. While there are multiple formalism to handling the total scattering in single crystals [47], we will focus on cases where the scattering information available is one-dimensional.

1.1.2 Scattering in non-crystalline materials

The above diffraction theory was developed primarily to deal with crystalline materials where the orientation of the crystal axes can be found (*i.e.*, in a single crystal). The crystallography community has developed numerous methods for understanding the resulting relative atomic positions with respect to a fixed orientation in a single crystal experiment [13]. The deviations from ideal crystallinity can be addressed by drawing from analysis techniques traditionally used to describe the structures of liquid and amorphous states, specifically the pair distribution function. This function can be conceptualized as either the inverse Fourier transform of the total scattering (Bragg plus diffuse) or as a measure of pairwise correlations in real space. It elicits information regarding the local structure in a system, which is particularly useful when this local structure deviates from the average structure.

Furthermore, for isotropic systems we can only produce one-dimensional scattering information to describe a three-dimensional structure. Samples can be considered isotropic when they possess no special orientations: occurring in liquid, amorphous, polycrystalline, and powder samples. Here we set out to define the PDF, discuss its real space significance, and develop its relation to the total scattering

pattern.

A linguistic labyrinth

Before we begin it is worth clarifying some notation. In the crystallographic community, $F(\mathbf{Q})$ is generally considered to be the complex function presented above and called the *structure factor* [13, 44]. Adding some ambiguity, the squared modulus of this function (a real function of intensity) is also often called the structure factor. On the contrary, in the total scattering convention, $F(\mathbf{Q})$ (again called the structure factor) refers to a real function related the squared modulus of the Fourier transform of a structure [45], and often appears with different notation [12]. This ambiguity of notation is further compounded by discrepancies in total scattering formalisms describing different normalizations of the PDF with the same language. This has not been neglected, with literature dedicated to casting a consistent formalism within the total scattering community [45].

The major discrepancy will be with the use of $F(\mathbf{Q})$. The function was presented above within the bounds of crystallographic conventions; however, Section 1.1.1 will be the only area of this thesis where $F(\mathbf{Q})$ and the *structure factor* are considered as complex functions. Hereafter, the conventions of the total scattering community, and more specifically Keen [45], will be followed. There are further discrepancies that will arise when considering the small angle scattering community. Nonetheless, these discrepancies do not contain redundancies in language, and will be described on a case-by-case basis.

Total scattering and the pair distribution function

To consider the total scattering from an isotropic system, let us start again with the basic diffraction equation for a beam of radiation scattered from a collection of N atoms with scattering factors b_j and positions $\mathbf{r}_j$:

$$S'(\mathbf{Q}) = \frac{1}{N} \sum_{j,k} b_j b_k \exp[i\mathbf{Q} \cdot (\mathbf{r}_j - \mathbf{r}_k)]. \tag{1.25}$$

Here we use the the term S' for uniqueness, and neutron scattering conventions for scattering factors for clarity; however, X-ray scattering factors could still be used with further complications introduced by the Q-dependence [48].

In an isotropic system it becomes necessary to average over all relative orientations of $\mathbf{Q}$ and $\mathbf{r}_j - \mathbf{r}_k$. Writing $|\mathbf{r}_j - \mathbf{r}_k|$ as r_{jk}, and the angle between the two vectors as θ, we have

$$\langle \exp[\mathrm{i}\mathbf{Q} \cdot (\mathbf{r}_j - \mathbf{r}_k)] \rangle = \frac{1}{4\pi} \int_0^{2\pi} \mathrm{d}\phi \int_0^{\pi} \exp(\mathrm{i}Qr_{jk}\cos\theta) \sin\theta \mathrm{d}\theta \tag{1.26}$$

$$= \frac{\sin(Qr_{jk})}{Qr_{jk}}. \tag{1.27}$$

Thus far we have only briefly mentioned the PDF as the distribution of distances between pairs of atoms in a material. Let us define the partial PDF $g_{jk}(r)$, with reference to an atom of type j, and the number of atoms n, of type k lying in a spherical shell of radius r and thickness $\mathrm{d}r$:

$$n_{jk}(r) = 4\pi r^2 \rho_k g_{jk}(r) \mathrm{d}r, \tag{1.28}$$

where the average number density of atoms of type k is given by ρ_k. If the total scattering structure factor $F(Q)$ is considered to include only distinct scattering [49] (*i.e.*, it does not include scattering contribution from atoms with zero separation), it can be written as

$$F(Q) = \rho_0 \sum_{j,k} c_j c_k b_j b_k 4\pi \int_0^{\infty} r^2 [g_{jk}(r) - 1] \frac{\sin(Qr)}{Qr} \, \mathrm{d}r. \tag{1.29}$$

From this we can define a total PDF

$$G(r) = \sum_{j,k} c_j c_k b_j b_k [g_{jk}(r) - 1] \tag{1.30}$$

and recast our total scattering structure factor as

$$F(Q) = \rho_0 \int_0^\infty 4\pi r^2 G(r) \frac{\sin(Qr)}{Qr}\, \mathrm{d}r. \tag{1.31}$$

It arises under these definitions that $F(Q)$ is the Fourier transform of $G(r)$, and $G(r)$ the inverse Fourier transform of $F(Q)$.

$$G(r) = \frac{1}{(2\pi)^3 \rho_0} \int_0^\infty 4\pi Q^2 F(Q) \frac{\sin(Qr)}{Qr}\, \mathrm{d}Q. \tag{1.32}$$

For completeness, our initial scattering law including the term for self scattering is simply

$$S'(Q) = F(Q) + \sum_i c_i b_i^2. \tag{1.33}$$

However, there is a more commonly used normalized total-scattering structure factor $S(Q)$, given by

$$S(Q) - 1 = F(Q) \Big/ \left(\sum c_j b_j\right)^2. \tag{1.34}$$

There are multiple approaches to normalizing the PDF, each with their own utility. I present a few normalizations here and their limiting behavior. For a more exhaustive discussion of total scattering nomenclature and normalizations I refer the reader to Keen [45].

A first alternative comes from normalizing to remove scaling from species concentrations and scattering lengths,

$$G'(r) - 1 = \frac{G(r)}{\left(\sum c_j b_j\right)^2}. \tag{1.35}$$

This gives a function with the same limiting values as the $g_{ij}(r)$ functions. Comparing (1.31) and (1.32), the transform is operating on the quantities $QF(Q)$ and $rG(r)$ [Table 1.1]. Therefore, an alternative normalization of the PDF which is commonly

Table 1.1: Limiting behavior of total-scattering functions.

Function	Equation	$\lim_{r\to 0} f(r)$	$\lim_{r\to\infty} f(r)$
$F(Q)$	(1.31)	$-\sum c_j b_b^2$	0
$S(Q)$	(1.34)	$\sum c_j b_j^2 / \left(\sum c_j b_j\right)^2$	1
$G(r)$	(1.32)	$-(\sum c_j b_j)^2$	0
$G'(r)$	(1.35)	0	1
$D(r)$	(1.36)	$-4\pi\rho(\sum c_j b_j)^2 r$	0
$T(r)$	(1.39)	0	$4\pi\rho(\sum c_j b_j)^2 r$

used throughout this thesis is $D(r)$, and its resepctive partial PDF $d_{jk}(r)$:

$$D(r) = 4\pi\rho r G(r), \tag{1.36}$$

$$d_{jk}(r) = r[g_{jk}(r) - 1], \tag{1.37}$$

$$D(r) = 4\pi\rho \sum_{j,k} c_j c_k b_j b_k d_{jk}(r). \tag{1.38}$$

A consequence of this renormalization is the linear behavior as r approaches 0. An alternative normalization y-shifts $D(r)$ by a linear term so that it tends to 0 at $r = 0$, but is linear at high r:

$$T(r) = D(r) + 4\pi\rho r \left(\sum c_j b_j\right)^2. \tag{1.39}$$

In Table 1.1 the limiting behavior of the functions utilized in this thesis are offered, as converting between two involves nothing more than an additive shift or multiplicative rescaling. With these definitions in hand, we can begin to explore the existing methods employed in structurally modeling disordered materials, and how the inverse problem of extracting real space configurations from information poor reciprocal space data is approached.

1.2 The nature of disorder

What is meant by 'disorder'? Canonical solid state physics depends significantly on periodicity in crystalline materials. From the foundation of a periodic crystal, the theories of electronic states, lattice vibrations, magnetization, and other phenomena can be built [2, 50]. The reciprocal space formalism used to describe the properties of crystalline systems is also necessary in determining the structure of crystalline materials *via* diffraction experiments. Despite the large exposition of periodicity in structural science, the assumption of a singular repeat unit does not hold universally. Many condensed matter systems do not possess this periodicity (*e.g.*, liquids, glasses, and polymers), and even crystalline systems boast behavior that disrupts order. It is this realm of materials that is considered to be 'disordered' — that is, any system of matter without the periodicity required of a crystal is considered a disordered material. The following will discuss how disorder manifests in materials, and its subsequent impact on our primary observable.

The range of disorder accessible in nature is vast, and it has historically been subjected to multiple overlapping classification schemes. It is possible to consider both the origin or the magnitude of the disorder in classifying the structural phenomenon; each yields different perspectives and structural analogues across disciplines, creating the ubiquity we described at the start of this thesis. I will begin by coarsely discretizing the degrees of disorder in condensed matter, paying close attention to the impact of the disorder on our main observable: the diffraction pattern [Fig. 1.2].

As described in Section 1.1, the ideal crystal is the epitome of order. Its diffraction pattern is composed of sharp Bragg peaks indicating long range periodicity. At the disordered end of the spectrum we will coarsely consider fluids. In the most disordered case of an ideal gas, there is random, uncorrelated disorder (no two-body correlations) and the scattering pattern yields isotropic radial diffuse scattering that follows the atomic form factors of the system.

The areas of complication — and thus excitement — lie in the region between

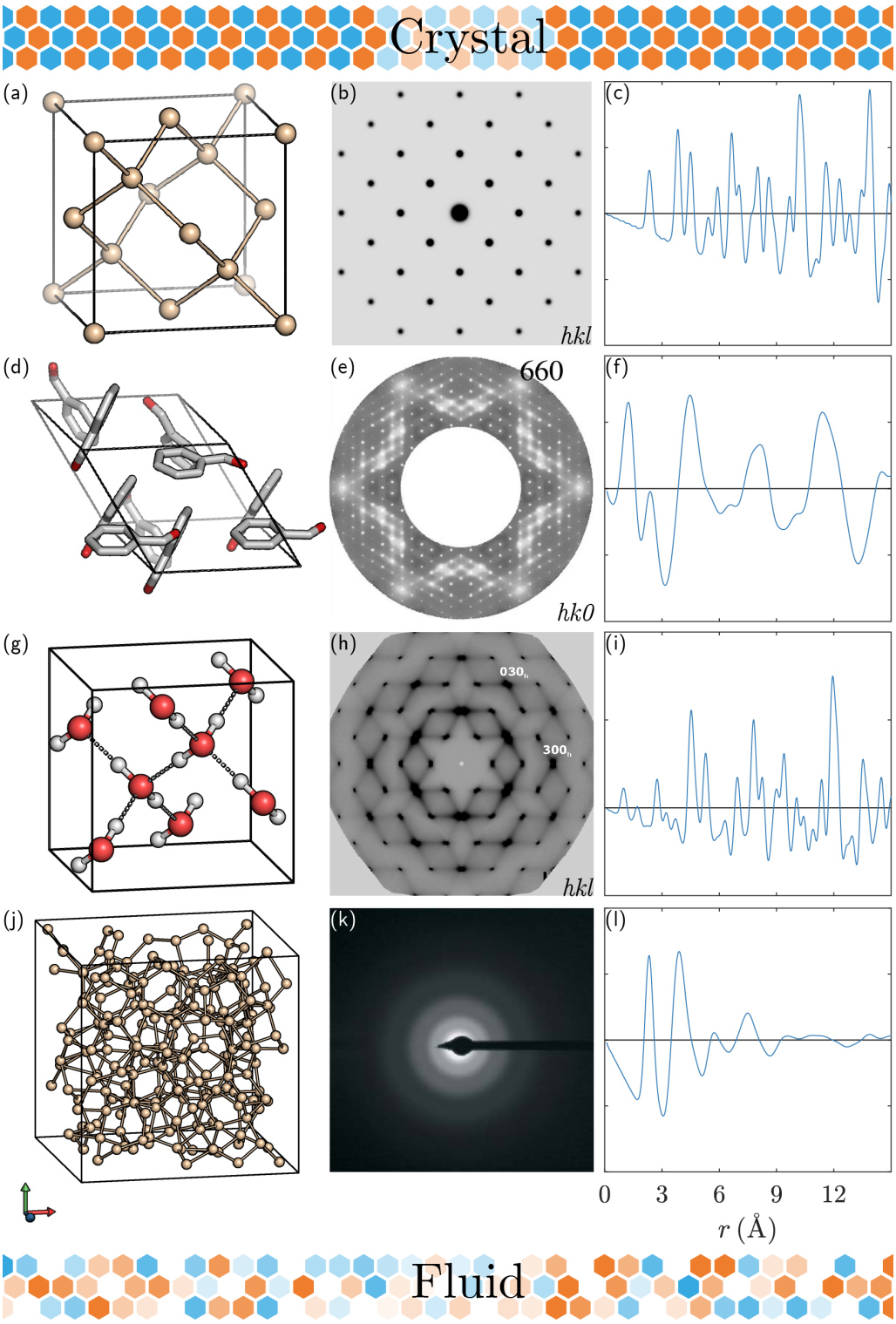


Figure 1.2: Increasing degrees of disorder. A discretized spectrum of disorder, ranging from an ideal crystal to a fluid. Each structure (left) is accompanied by its diffraction pattern (center) and PDF $D(r)$ (right) calculated with PDFGUI: (a) diamond type silicon and (b) its diffraction pattern as calculated by CRYSTALDIFFRACT; (d) benzil and (e) experimental neutron diffraction of the $(h\,k\,0)$ section (data from [51]); (g) cubic water ice (Ic) and (h) diffuse X-ray scattering data of the $(h\,k\,l)$ plane in Ih ice from Ref. [52]; and (j) amorphous silicon and its (k) electron diffraction pattern (data from [53]).

these two phases. Zero-point motion and uncorrelated atomic vibrations create an average atomic position, reducing the intensity of the Bragg diffraction and broadening the pair correlations seen in the PDF [Fig. 1.2(a)-(c)]. Realistic atomic motion involves correlated displacements of atoms [54]. This is the cause of the thermal diffuse scattering (TDS) centered about the Bragg peaks from crystalline benzil [Fig. 1.2(e)]. The considerable diffuse scattering (the less structured scattering between and around the Bragg peaks) prohibits the assumption of uniform periodicity; however, the diffuse scattering is highly structured and directional. Notably, the time-averaged positional disorder from thermal vibrations acts to significantly broaden the features of the PDF.

Next, I highlight the case of water ice, classical crystallography's earliest stumbling block and ardent antagonist [Fig. 1.2(g)-(i)]. It is here we begin to see states that clearly possess some intrinsic ordering (as indicated by the sharp local correlations in the PDF), contrasted with elements of disorder apparent from the diffuse scattering*. The disorder in ice is relatively static as compared to that in benzil, and this emerges in the short-range order seen in $D(r)$: because the oxygen atoms in cubic ice lie on a diamond lattice, the PDFs of silicon (c) and ice (i) display similar local correlations, with disorder introduced by the variation in hydrogen position.

Lastly, an amorphous material is instructive because of its isotropic scattering. Similar to powder diffraction, an amorphous material will have a scattering pattern independent of orientation. Here, the diffraction rings will be broadened by an underlying disorder in the crystal, qualitatively indicating a loose isotropic correlation. The PDF of amorphous silicon further demonstrates the presence of this local ordering. Beyond this we could consider condensed fluids and their two-body correlations, evident mainly in the PDF.

There are of course other examples of disorder that fall along this spectrum, and

*Available experimental scattering data for Ih ice is shown, whereas the structure and PDF of ice Ic is shown for comparison with silicon. The disorder in these two phases is qualitatively similar.

Figure 1.2 could be subdivided *ad infinitum*. The examples presented in the figure and below will nonetheless serve pertinent to the remainder of this thesis.

1.2.1 Dynamic positional disorder

Well known through the understanding of thermal motion of atoms, dynamic positional disorder occurs in every real material. That is, the energetically accessible lattice vibrations act to displace the atoms from their average sites. In the simplest case, this disorder can be approximated as uncorrelated harmonic motion about the average position. The crystallographic formalism of atomic displacement parameters (ADPs) simplify thermal motion into ellipsoids of vibration about an atom site [13]. The averaging associated with ADPs may often provide an incomplete picture of the structural behavior, particularly in the case of chemically bonded framework materials. It is these cases where thermal diffuse scattering need be considered as a result of correlated atomic motion. These correlated vibrations are described by the collective excitations of lattice dynamical states called *phonons*. In TDS, scattering intensity is inversely proportional to phonon frequency. Thus, TDS tends to be dominant near the zone center of reciprocal space (*i.e.* underneath the Bragg peaks), but can give a highly structured diffuse scattering in systems with low energy modes at the Brillouin zone boundary.

The TDS shown for benzil in Figure 1.2 is a result of correlated longitudinal displacements due to low energy phonon modes, but also contains information regarding the relative conformations of neighboring molecules [51]. Another fine example of TDS is encountered in the crystalline phases of silica. In the uncorrelated approximation the disorder is found to yield highly anisotropic (and unphysical) ADPs for the oxygens, which are much more accurately described by the accessible lattice vibrations [55].

Moreover, this raises the challenge associated with the timescale of disorder. Total scattering in theory integrates over all energy to give an instantaneous snapshot. Although a scattering event provides a snapshot of a point in a vibrating lattice,

the many scattering events needed — and the time it takes to experimentally measure them — to produce an effective signal relate to an average of the continuum of occupied positions. That is, our experimental probe operates on a longer timescale than our disorder. Nonetheless, inelastic scattering methods and spectroscopy [56] provide experimental access to these dynamics. Even in the case of proteins — which are dynamic in both solution and biological processes — methods are being developed to overcome the difference in timescale between the experiment and relevant phenomena [57]. Next we will examine cases of static disorder that lie near the center of the spectrum of disorder described in Figure 1.2, beginning with the renowned Ising antiferromagnet.

1.2.2 Frustrated overconstraint

Consider the Ising model on both a square and triangular lattice [58, 59]. In this model lattice points are decorated with spins that can point either up or down (or equivalently be positive or negative). In the antiferromagnetic case, spin exchange interactions are favorable when neighboring spins have opposite direction. The ordered ground state of the antiferromagnetic square lattice is facile to generate: fix any spin, set its 4 nearest neighbors to be antiparallel, and continue this process with each neighboring spin [Fig. 1.3]. If we attempted the same heuristic in generating the spin configuration of the triangular case, we may find ourselves frustrated by our inability to craft a ground state where all spin exchange interactions are satisfied. While any two spins in a triangle can be antiparallel, the third is forced to align parallel with one of the first two. There are six possible ways of achieving a degenerate ground state on a single triangle. This 'frustration' is also the technical term for the local energetics that form from a mismatch in the symmetry of the local geometry and the local states [23].

This Ising system actually provides a basis for interpreting any lattice where two states (up or down; positive or negative; occupied or unoccupied) have an incompatibility between their interactions and lattice geometry. Antiferromagnetic states

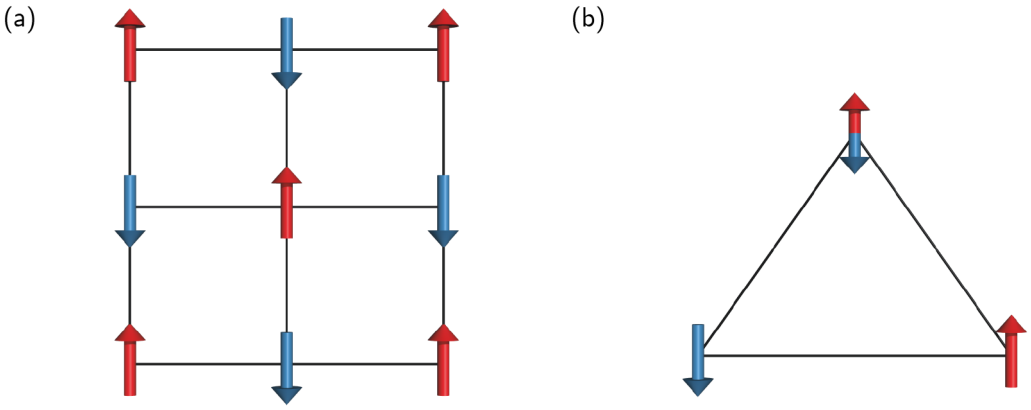


Figure 1.3: Ising model. The Ising antiferromagnet on a (a) square and (b) triangular lattice. The frustrated overconstraint of the triangular model becomes clear when following the heuristic for decoration.

can be extended to consider any system where interactions favor a state reversal between neighbors. Modified liquid crystals have been shown to map directly onto the xy-triangular Ising antiferromagnet due to geometric steric frustration [60]. The frustration characterized in Figure 1.4 emerges from the packing of subunits on a hexagonal lattice, where steric interactions provide competing interactions between nearest neighbors and across a hexagon. It is this local ordering that creates frustration on the sublattice, produces long range disorder, and manifests in characteristic diffuse scattering.

In the case of more continuous degrees of freedom, silver(I) and gold(I) cyanides form supramolecular assemblies with vertical displacements of chains as the key structural degree of freedom [61]. The structure of AuCN forms close packed chains of gold atoms. However, in the AgCN structure chains preferentially displace 1/3 of the chain repeat to bring the metal cations closer to the cyanide anions. If this vertical displacement is assigned a phase angle it directly maps onto the triangular antiferromagnet ground state. Other base lattices and interactions can established. For example, vertical displacements of tris-*tert*-butyl-1,3,5-benzene tricarboxamide packed in a hexagonal fashion — aligned with the c axis — are disordered, but strongly correlated; neighboring chains strongly prefer antiparallel arrangements [62].

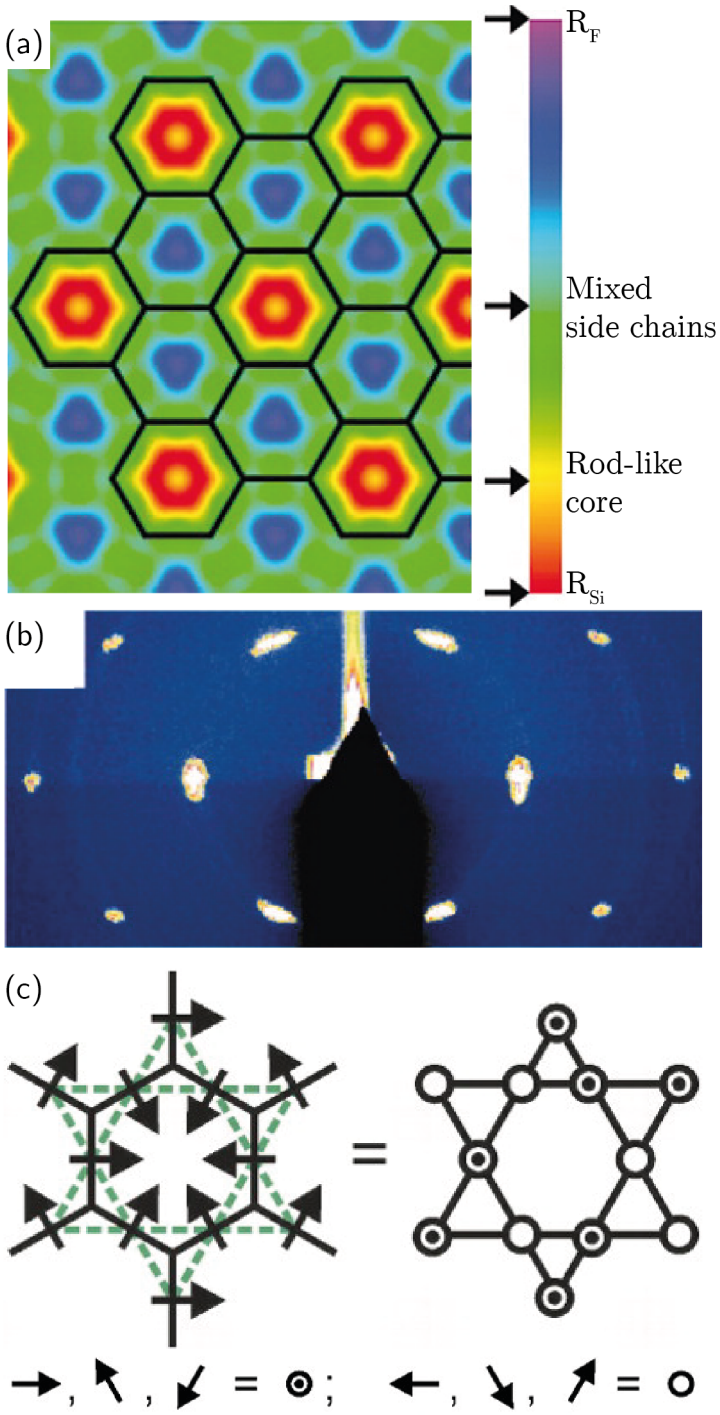


Figure 1.4: Complex tiling patterns in liquid crystal honeycombs. (a) Electron density map of a low-temperature two color liquid crystal phase, where R_X refers to a phase richness of chain components. (b) Small angle scattering patterns of thin films of the liquid crystal phase on Si substrates. (c) The hexagonal phase maps onto the spin-half Ising model on a kagome lattice, by equating the six orientations of the Si rich side chains with up or down spins. Figure adapted from [60].

From these examples, it is apparent that frustrated states can be generalized to other systems with different physical origins of interaction. Here we have introduced the concept of *local rules* that manifest themselves in real systems, and like the Ising antiferromagnet have no unique solution with long-range order. This mapping is useful because the statistical mechanics of these systems is well understood. When our local rules are relaxed to probabilistic enforcement, the disorder developed is still underpinned by local correlations.

1.2.3 Compositional disorder

While frustration can be caused by a mismatch in geometry, we can imagine other types of local interactions that will force disorder. Let us consider a specific example of frustrated overconstraint: the square lattice over which equal types of atoms A and B are distributed. This is the chemical realization of the square Ising model in Figure 1.3.

If A and B are ordered analogous to the antiferromagnetic square lattice, then each A atom will be surrounded by four B atoms, and *vice versa*, forming a checkerboard pattern. Thus, the local order at a given site propagates throughout the crystal lattice, and the average unit cell is identical to any local environment. Casting the antiferromagnetic ordering as a probabilistic rule, the probability that an site is occupied by A given its neighbor being B is 1 [*i.e.* $P(A_i|B_{i+1}) = 1.0$]. If we instead decorate a site environment such that the neighbors are chosen with a uniform distribution [*i.e.* $P(A_i|B_{i+1}) = 0.5$], the average structure will differ from the local environments. This approach with uniform random disorder will have a different average structure (and thus Bragg scattering) from the ordered case, with an obvious change in the diffuse scattering and pair correlations. Finally, if we use a nonuniform probability distribution and preferentially but not absolutely order like the antiferromagnetic analogue, with a site probability $P(A_i|B_{i+1}) = 0.8$, correlated disorder arises [23]. The local environment has a biased composition; however, the structure is disordered over the long range with the total composition fixed at

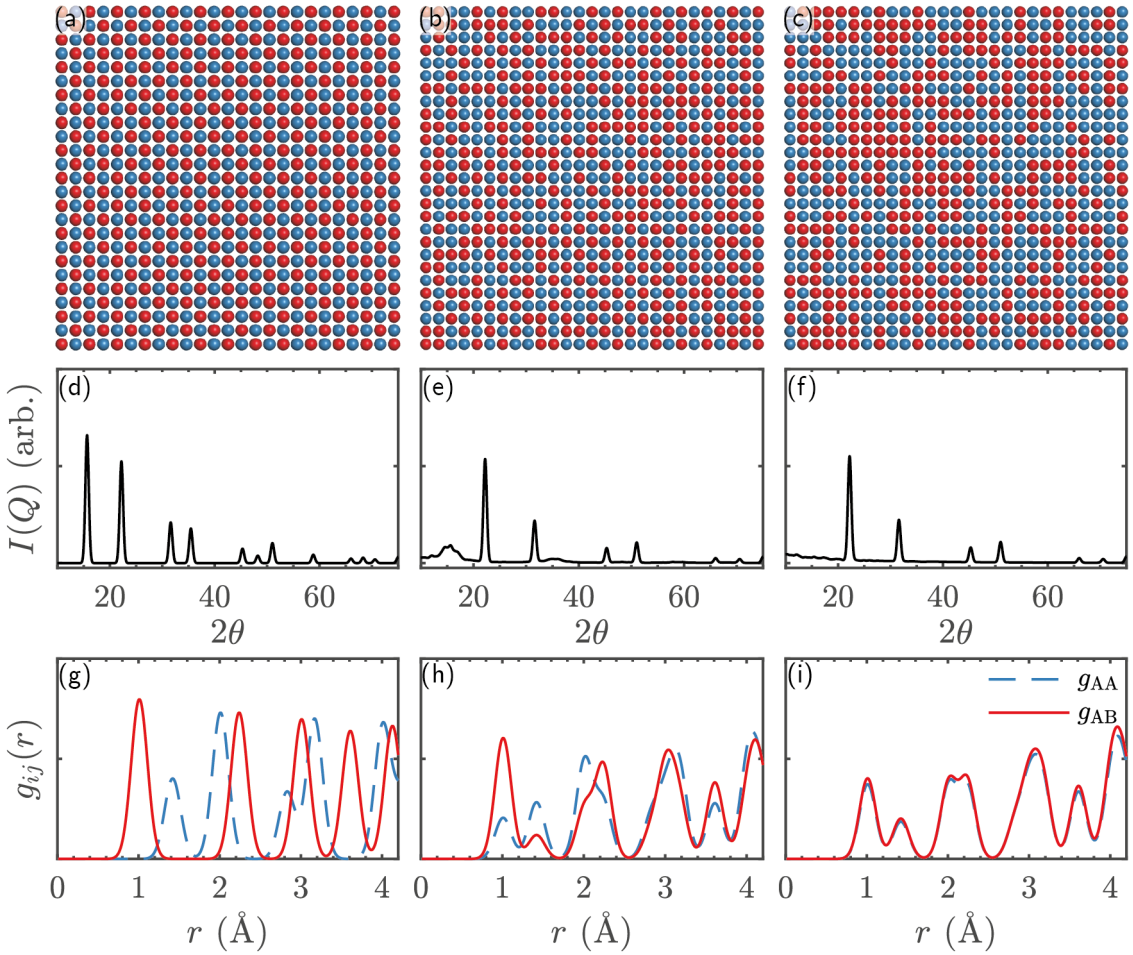


Figure 1.5: Square lattice compositional order, correlated and random disorder. A square lattice with varying degrees of compositional order. $P(A_i|B_{i+1}) =:$ (a) 1.0, (b) 0.8, (c) 0.5. The (d-f) total scattering and (g-i) pair distribution functions are shown below their respective structures. The partial PDF is capable of differentiating between the cases of correlated and random disorder, even though the Bragg diffraction is similar between the two. Structural coordinates courtesy of Matthew J. Cliffe.

$x_A = 0.5$. Examining the average structure of the correlated disorder case, it will appear the same as that of the random disorder. Naturally, the Bragg scattering will be similar between the two disordered cases. As an observer of the structure, how can we differentiate between the local environments? The key distinctions between these lie in the diffuse scattering and pair correlation functions [Fig. 1.5].

The capacity for pair correlations to differentiate between the local interactions governing a material's structure should come with a cautious optimism. Unfortunately, the diffuse scattering that could discriminate between the models of Fig-

ure 1.5 is incapable of recognizing any higher order correlations directly [11]. Another square lattice with disordered composition is offered in Figure 1.6. In this lattice, the completely random distribution is compared against three other examples that also possess zero two-body correlations, but have three-body and four-body correlations. The diffraction patterns for all four examples are indistinguishable, even though the physical nature of each is significantly different. These simple models serve to illustrate what kind of information is accessible in the diffraction pattern and what is not. Specifically, it refutes the notion that two matching diffraction patterns must have originated from the same structure, and demonstrates a particular challenge in understanding most disordered systems.

The practical system most similar to the square lattice case is that of binary cubic phase alloys. A prototypical example, Cu_3Au, demonstrates the full range of compositional disorder discussed above [64, 65]. At high temperature, a complete solid solution is formed on a face centered cubic (fcc) lattice. This uniform distribution yields an average structure unit cell with a single site, occupied by 0.75 Cu and 0.25 Au. Upon cooling, Cu_3Au undergoes a phase transition to an ordered structure, where the corners of the face centered unit cell are all occupied by Au and the faces all occupied by Cu [66]. Between these temperatures there is a regime of short-range order and long-range disorder characterized by various preference to local configurations [67–69]. This temperature-induced transition from random disorder to order through a state of correlated disorder is but one of many instances of compositional disorder.

Compositional disorder has profound influence across materials and technologies and occurs whenever a number of different species occupy the same average structural site. Nonstoichiometry can be exploited in even minuscule amounts, such as in the parts-per-million doping of semiconductors that is the foundation of modern technology [70, 71]. When the concentration of impurities increases, the interactions between species can no longer be neglected, and degrees of ordering arise. Local cor-

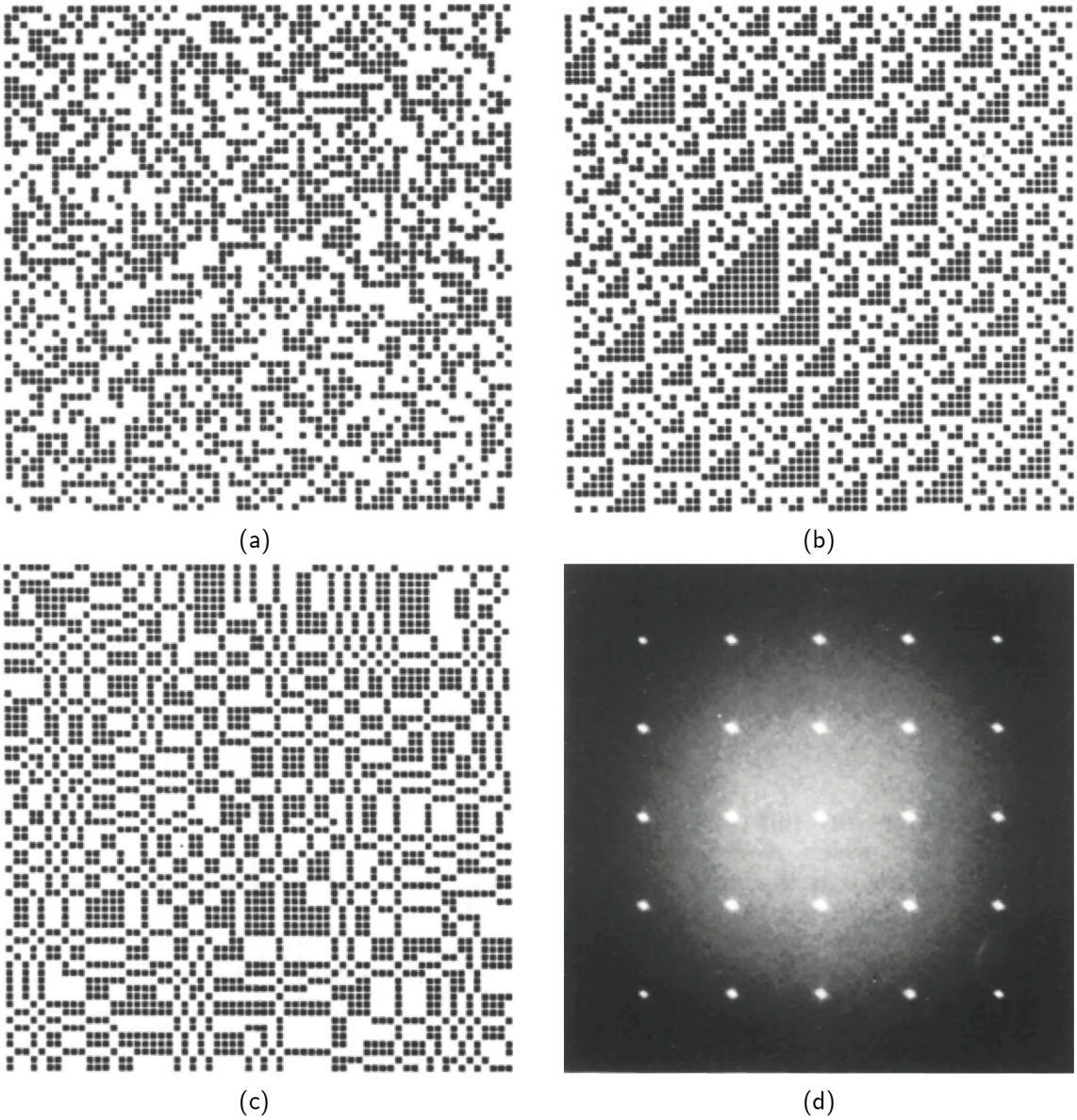


Figure 1.6: Diffuse scattering of higher order correlations on a square lattice. The representative scattering pattern (d) is indistinguishable between systems with (a) no, (b) three-body, and (c) four-body correlations. Figure adapted from Ref. [63].

relations impact the shape memory effect in $Ni_{50-x}Ti_{50+x}$ [72] and hydrogen storage in $LaNi_5$ [73]. Longer range correlations effect the microstructured organization of dopants in high temperature superconductors [74] and thermoelectrics [75]. Varied length scales of ordering explain the morphotropic phase boundary in lead zirconate titanate [76]. A lack of correlation is integral in the nitrogen integration in the $GaAs_{1-x}N_x$ structure [77]. Vacancy ordering in alloys is used as a different type

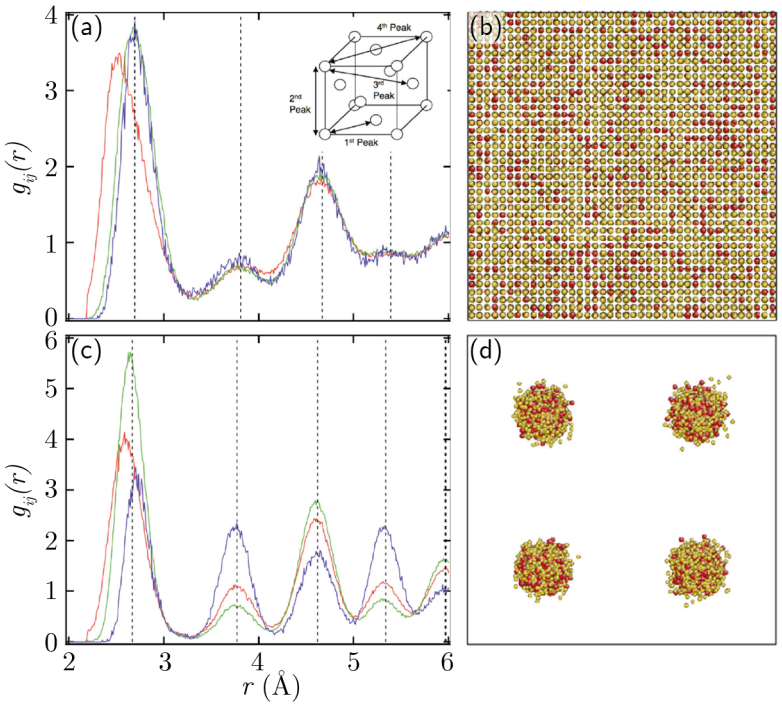


Figure 1.7: Short-range correlations in Cu_3Au. The copper gold alloy structure was refined using RMCPROFILE considering atom swaps and translations. Partial PDFs from these RMC fits for (a) 900 °C and (c) 400 °C samples are given. The variation in partials between the high temperature sample (complete solid solution) and low temperature sample (short-range order) highlight the preferential occupancy of nearest neighbors. The dotted lines indicate the expected distances from the fcc structure, with an inset demonstrating those distances. (b) The results of a 32000 atom simulation are (d) collapsed onto a single unit cell, showing the translational distribution of atom positions in the fcc lattice, as well as the compositional disorder (Cu atoms are represented with yellow spheres, and Au atoms are in red). Figure adapted from [69].

of defect in the control of volume, electrical conductivity [78], and magnetism [79]. Beyond alloyed materials, compositional disorder is employed in oxide frameworks for superconductivity [80] and colossal magnetoresistance [81, 82]. This displays the ubiquity of compositional disorder in materials; however, these ideas can effectively map onto other types of disorder.

1.2.4 Orientational disorder

It is worth presenting one more case study on the square lattice, to demonstrate a further category of correlated disorder that can appear in systems. Historically, the first geometrically underconstrained system to have been examined was crystalline

ice [83]. Following this seminal work, water has been subject to many 'toy model' studies, one of which we develop here.

Decorating the square lattice sites with the oxygen atoms of the water molecule, a choice is left for where to place the hydrogen atoms. The natural choice is the midpoints between neighboring oxygen atoms, such that no two hydrogens overlap. This satisfies both steric and hydrogen bonding considerations. Pauling cast these ice rules as follows: two hydrogens pointing into an O vertex, and two pointing out [83]. Following these rules, one suitable ground state is the perfectly ordered case; however, there are multitudes of degenerate ground states that do not possess translational order, but adhere to these rules and as a consequence possess correlated disorder [Fig. 1.8]. It should be clear that even though the correlated disordered state is not ordered in the traditional sense, it is certainly not random. In fact, if just one arbitrary row and column in the square lattice are decorated according to the 'two-in two-out' rule, then the orientation of every other water molecule becomes fixed, indicating the substantial degree of long range order even without translational order.

A 'two-in two-out' rule could well be different depending on the underlying lattice or chemical considerations. The critical idea is that these rules can govern physical behavior and manifest residual — or configurational — entropy in a crystalline state. Let us attempt to estimate this configurational entropy [84] in 3D ice I_h.

Firstly let's consider the case of tetrahedrally connected water ice with angles of $\sim$110°. With no other restrictions to a lattice of N water molecules, there are $2N$ lattice connections (hydrogen bonds) that can possess an 'on' or 'off' microstate, for 2^{2N} configurations. However, not all of these configurations are acceptable given the 'two-in two-out' rules. Of the 2^4 possible hydrogen microstates surrounding a given oxygen, 10 are inaccessible: 1 with all 4 on, $(H_4O)^{++}$; 4 with 3 on, $(H_3O)^{+}$; 4 with 1 on, $(OH)^{-}$; and 1 with 0 on, O^{--}. Thus the probability of a given microstate surrounding an oxygen being consistent with the rules is $\frac{6}{16}$. Thus the number of

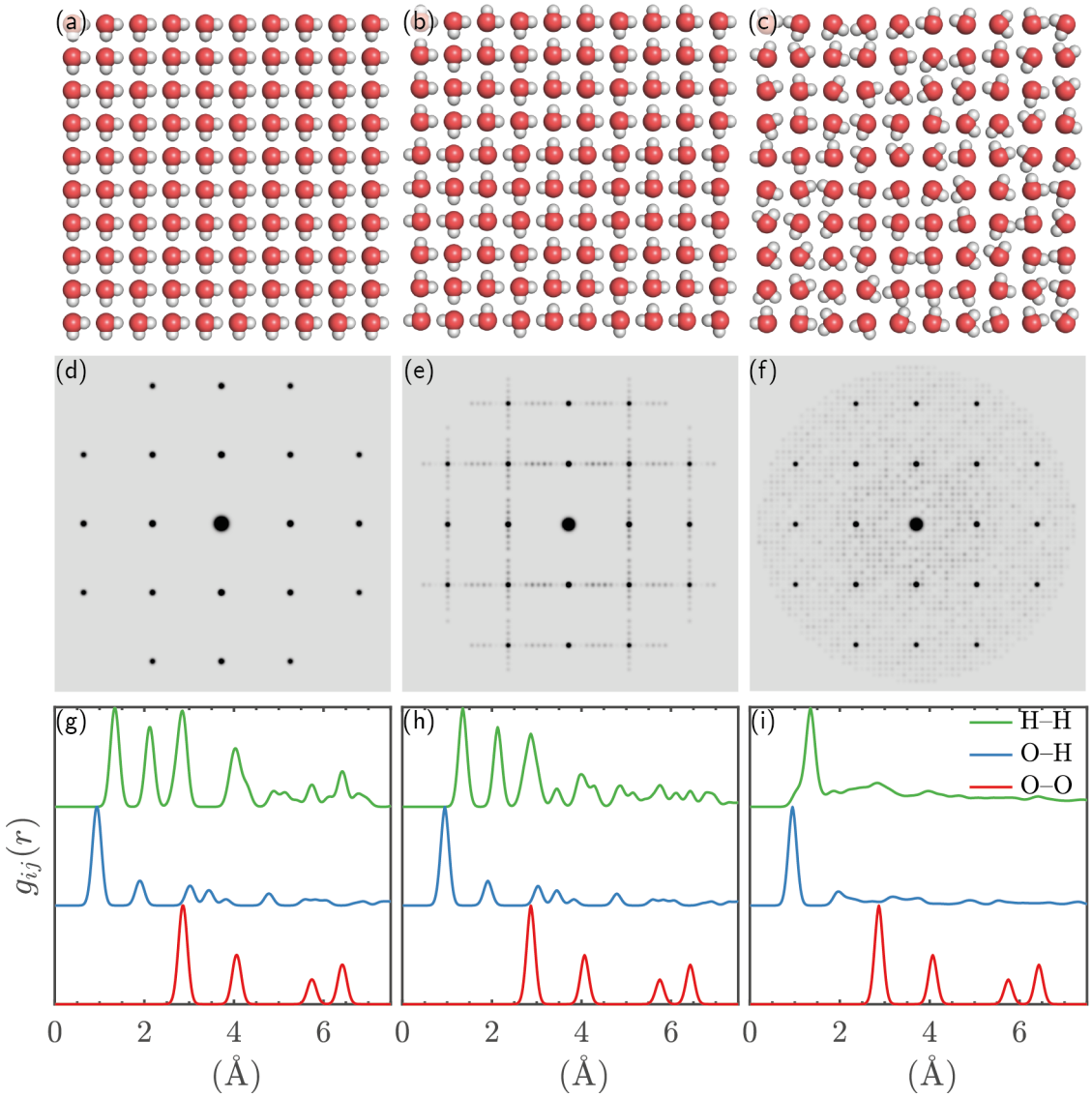


Figure 1.8: Square ice. Two energetically equivalent ground states of square ice that follow Pauling's ice rules: one is (a) perfectly ordered and the other has (b) correlated disorder. The (c) random configuration is shown for reference to demonstrate the differences in (d-f) diffuse scattering and (g-i) pair correlation function. The 1-dimensional ordering of the hydrogens along each row and column gives rise to the structured streaks of diffuse scattering seen in frame (e).

configurations is approximated by:

$$W_{\mathrm{P}} \simeq 2^{2N} \left(\frac{6}{16}\right)^{N} = \left(\frac{3}{2}\right)^{N}. \tag{1.40}$$

Equivalently, we can consider the 6 possible orientations of a water molecule with two hydrogens 'in'. Each of these orientations will have two neighbors adjacent to the 'in'

hydrogens that need to be configured such that their hydrogens are 'out' from the first water ($\frac{3}{6}$ orientations for each of two relevant neighbors). Thus the probability that a given orientation will be acceptable will be $\frac{1}{4}$, leading to $W_{\mathrm{P}} \simeq \left(\frac{6}{4}\right)^N = \left(\frac{3}{2}\right)^N$.

Generalizing Pauling's approximation, we have:

$$W_{\mathrm{P}} \simeq (\text{no. of possible local states}) \times (\text{proportion that obey the rules}). \tag{1.41}$$

Other approaches for characterizing configurational entropy have sought to compare directly the order of the point group of the local configuration and the order of the point group of the underlying lattice, casting the total entropy as

$$S_{\mathrm{config}} \simeq N k_{\mathrm{B}} \ln\left(\frac{n}{2^{d/2}}\right), \tag{1.42}$$

where n is given by the ratio of the order of the point group for the parent lattice and the order of the point group for the motif, and d is the underlying connectivity of the lattice. This can give an effective perspective on the degree of low dimensional order in a system of correlated disorder [85].

Although these three formulations are consistent, they become more interesting in the case of square ice where the connectivity is the same, but the local geometry of the network is different. Considering the square lattice, where the 'two-in two-out' rule is accompanied by a geometric restraint of a right angle H–O–H connection, with the first heuristic only four of 16 configurations are accessible (two of the six neutral H_2O configurations become linear). The same holds in the second heuristic, for a value of $W_{\mathrm{P}} \simeq 1$ and zero residual entropy according to (1.52). A similar result can be found using the symmetry arguments of (1.42). As designed, this model actually contains a subextensive entropy — that is, a configurational entropy that is nonzero for finite systems but approaches 0 as the size of the system approached infinity [86].

What makes the square ice system unique, is actually the information content.

A keen eye on Figure 1.8 may notice that each column or row progresses with a similar water orientation. That is, in a given column every water is oriented 'up' or 'down' when considering the hydrogen in the column with respect to their chemically bonded oxygen. In fact, if a self-consistent row and column were populated, the configuration of the entire box becomes determined. Therefore, in a 10×10 box of 100 water molecules, there are only 20 bits of information (19 relevant orientations and 2 bits to fix four degrees of freedom per orientation). But for a 50×2 box of 100 water molecules, there are 52 bits of information (51 independent orientations). What exists in this case is an N-independent function that adds to the residual entropy. When Pauling's approximation tends to zero and correlated disorder is present, there is subextensive entropy at 0 K and also a low dimensional order that is most apparent in the highly structured diffuse scattering.

The same 'two-in two-out' rules that govern the arrangement of water molecules in water ice I_h [83], can be mapped onto the anisotropic and correlated cation displacements in $Cd(CN)_2$ [40] and $Bi_2Ti_2O_7$ [87]. These also find analogues in the magnetic moment orientations in spin ices $Ho_2Ti_2O_7$ and $Dy_2Ti_2O_7$ [38], and in the structural disorder of transition metal oxynitrides [86]. The disparate examples of positional, orientational, and compositional disorder that find analogy in each other demonstrates the pervasive utility of disorder and the methods used to describe it.

1.2.5 Topological disorder

One final type of disorder warrants some exposition. The previous physical examples of disorder have all made reference to an underlying or average crystal lattice. In the instance of topological disorder, there is no long-range order, the topology is aperiodic, and thus the materials are considered amorphous [88]. That does not however mean that amorphous materials possess no elements of order or local correlations. In contrast, the complete loss of long-range order leads to a broad, diffuse, isotropic diffraction pattern, and a PDF with prominent low-r features [Fig. 1.9]. As we will come to see, this makes probing and describing the structure of amorphous

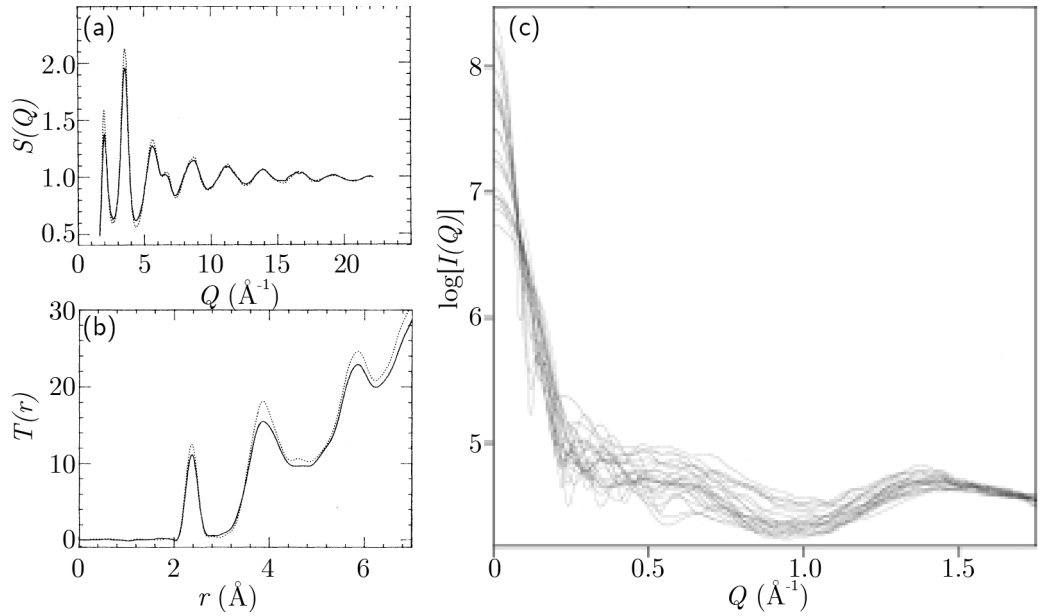


Figure 1.9: Examples of scattering from topologically disordered states. (a) The scattering of amorphous silicon does not contain Bragg peaks. However, the (b) PDF contains rich information regarding the local correlations in the disordered material. Shown are results from two separate synthesis methods in the dotted and solid line of the same material (adapted from [89]). Contrasting this, the small angle scattering for 25 different proteins of varying size is shown in (c). Although the pattern seems to possess far fewer features that that of even *a*-Si, the low Q region can meaningfully differentiate between proteins. The features of this region and the relative divergence as Q approaches 0, relate to the overall shape of a protein (adapted from [90]).

materials a unique challenge as compared to their crystalline analogues.

A most commonplace instance of disorder, the utility of everyday and laboratory glassware* depends on the amorphous phase of silica. Similar in their long-range translational or orientational disorder, amorphous materials depend on a large breadth of chemistries. A few of the many classes of amorphous materials are: organic pharmaceuticals [91] and semiconductors [92]; inorganic glasses [93], framework oxides [3], metallic glasses [94, 95], chalcogenide glasses [19], polymers [96], biomaterials [97], and thin films [21]. While there is some debate as to the thermodynamic nature of these materials, they are at the least metastable [98, 99], and often possess reversible ordering phase transitions [19]. The lack of long-range order has implications for the material properties — causing a direct band gap in

*Glasses being a subset of amorphous materials, which possess a glassy transition temperature region and a subsequent viscosity on human time scales [3].

a-Si [100], and low resistivity in metallic glasses [101] — but also for the diffraction pattern as described above.

The second example of isotropic diffraction offered in Figure 1.9 demonstrates the small angle scattering (SAS) of a protein in solution. Contrary to the characteristic scattering of a liquid — which has fewer features — the protein SAS pattern contains a rapidly decaying $I(Q)$ that is determined by particle shape. The protein SAS pattern is significantly less structured than the total scattering structure factor of amorphous silicon. A comparison of curves is shown in Figure 1.9 to demonstrate how this pattern varies across systems. At low values of Q, the curves are very different and determined by particle shape (*i.e.* large features in real space). However, at higher real space resolution, the features of these SAS curves are indistinguishable for practical purposes. Comparing this to the case of *a*-Si, the pattern at high-Q reveals significant detail of the local atomic structure of the material. Although it seems sparse, the SAS technique is a fundamental tool in the study of biological macromolecules [102], and will be implemented in this thesis as an upper bound of disordered scattering information.

To note, there remains a subset of topological disorder relevant to the topics explored in this thesis. If the connectivity of a framework can be cast as a 1-dimensional structure with no consistent topology and long range disorder, it is defined as a *spaghetti*. These spaghetti states are similar to the packing of Dirac strings in Coulomb phases such as pyrochlore spin ices [103]. Finding utility from geographic information systems [104] to astrophysics [105], this distinction can offer additional insight into frameworks similar to those developed herein [106].

It is clear that disorder is not limited by form or function. We have just seen the degeneracy of ground states from correlated disorder caused by both frustrated overconstraint and configurational underconstraint. While there a several approaches to understanding disorder — particularly non-random disorder — the ability for these simpler systems and ideas to map onto more complicated structures lies at the core

of this thesis. Throughout each of these examples of structural disorder, no matter how they were classified, the utility of the diffraction pattern remained fixed. Whether from a single crystal or after powder averaging, total scattering data is our best tool in discerning the structure of disordered materials. Next we will examine some of the historical attempts at inverting the diffraction pattern to ascertain a structure for these kinds of systems.

1.3 Structural modeling and disorder

Now that we have seen some of the ways disorder impacts structure and scattering, it is necessary to explore the reverse. Given these elements of Bragg and diffuse scattering, how have scientists approached the problem of deriving meaningful atomic structure? There is an abundance of real examples for each category of disorder discussed in the previous section. In what follows I will examine the historical attempts at producing relevant structural models of disorder with a focus on methodology, highlighting key systems and insights along the way.

A vast manifold of techniques is applied to modeling structure, some of which depend on fitting to experimental data, and others which focus strictly on *ab initio* modeling. As algorithms improve and computational costs decay, approaches are further combined to make use of compounding information. Before diving into these more contemporary mixed-modeling techniques, I will first describe how the problem of inverting the diffraction pattern is approached in the ordered case. In ordered systems, there exist approaches to solve the phase problem, and subsequently determine a structural solution that can be refined to reproduce the experimental scattering intensities as accurately as possible.

Following this, I will explore the many approaches to disordered structural solution and refinement. In the manifestation of increasing disorder, structure solution — in an absolute sense — becomes increasingly difficult; however, structure refinement is practically accessible. To begin, I will discuss approaches that directly analyze the diffuse scattering, both in the case of single crystals and powder

averaging. Secondly, I will examine *forward* modeling of diffuse scatting that is interaction-driven. This will be contrasted with *reverse* modeling of scattering that is primarily data-driven. There is significant overlap in the kinds of systems these distinct methods have been applied to. These methods can be be split by several overlapping dichotomies: big box and small box; model dependent and model independent; or forward and reverse. While it is helpful to use the latter descriptors in this section, I will make reference to the former while focusing on the relative benefits of each method. Finally, I will close the section with a discussion of the key challenges that exist in determining meaningful structural models from total scattering in disordered systems.

1.3.1 Inverting experimental data for ordered systems

Historically, the earliest crystal structures were solved by trial and error methods; however, this is not scalable and has been replaced by the utility of computational and algorithmic advances. In real space, the Patterson function

$$P(\mathbf{u}) = \rho(\mathbf{r}) * \rho(-\mathbf{r}) = \int \rho(\mathbf{r})\rho(\mathbf{r}-\mathbf{u})\mathrm{d}\mathbf{r} \tag{1.43}$$

will have maxima associated with all of the interatomic vectors in the unit cell, and can be calculated directly from the observed experimental diffraction pattern. This is starkly different from the inverse of the Fourier transform of the complex structure factor, which would in principle contain maxima associated with all the interatomic *positions* in the unit cell. Again, we encounter the phase problem. The Patterson function will have prominent maxima when the structure contains a limited number of strong scatterers (heavy elements for X-rays), or a molecular geometry with redundant interatomic vectors (*e.g.*, aromatic rings), and can also give valuable insight into the space group symmetry.

Single crystal diffraction

Creating a model from diffraction data is considered *structural solution*. Single crystal structural solution generally proceeds by first determining the unit cell parameters from the diffraction pattern. From the symmetry of the diffraction pattern, the Laue group can be determined [13]. The Laue classes are a set of eleven centrosymmetric point groups that describe the diffraction symmetry. Even though the diffraction pattern is always centrosymmetric, the intensity distribution across reciprocal space can indicate the likelihood of the crystal having a centrosymmetric space group. Systematic absences can then be used to narrow down the list of candidate space groups.

Following space group determination, single crystal structure solution can be split between real space and reciprocal space approaches. One direct space (real space) method makes use of strong scatterers to determine the positions of heavy atoms and takes the Fourier transform of their positions for an estimate of the phase, then subsequently calculates an electron density map using this phase as an estimate for the true phase [107, 108]. This process can be performed iteratively, as each new calculation of the electron density map reveals positions of lighter atoms until the structure is complete. This process of iteration is again employed in electron density modification methods, whereby an electron density map is calculated, the density is modified in a sensible way, the Fourier transform is calculated for new amplitudes and phases, these can then be modified, and the inverse Fourier transform is calculated to produce a new density map. These kinds of methods are accessible because of Fast Fourier Transform (FFT) algorithms and computational resources. Several modifications (in real and reciprocal space) are utilized, and have already been demonstrated in the heavy-atom approach. These include correcting the Fourier series through virtual reflections [109, 110], and charge flipping in real space [111–116]. In reciprocal space, the positivity and discrete nature of the electron density function are taken advantage of in determining the phases from the

observed amplitudes (a task not universally feasible given the square modulus of a Fourier transform) [117–119]. All of these approaches are different ways of dealing with the phase problem, that produce a set of atomic coordinates with limited precision.

Going from a structural solution to a precise set of coordinates that includes the amplitudes of thermal vibrations is accomplished through *structure refinement*. It is common for this refinement to utilize a least-squares method. The quality of a structure determination is given by the sum of residuals between the observed diffraction intensities and the calculated diffraction intensities. This avoids having to consider the phases as in structural solution. Computer algorithms are available that refine the values of the structural coordinates, atomic displacement parameters, and other effects that modify the structure factor (*e.g.*, inelastic scattering or absorption effects) to minimize the residuals. In practice this is all accomplished through well defined algorithms presented in user-friendly programs such as SHELXL [120], SUPERFLIP [121], BayMEM [122], and CRYSTALS [123].

Powder diffraction

In systems where suitable crystals cannot be grown for experiment, powder diffraction is the next alternative. The powder averaging used in this experiment is equivalent to the orientational averaging associated with amorphous and non-crystalline condensed materials [Section 1.1]. Although more difficult than the single crystal counterpart, structure solution and refinement of powders is still feasible.

The major challenge here is in using a 1D data set — which contains systemic overlap of non-symmetry-equivalent Bragg diffraction peaks [13] — to reconstruct a 3D structure [124]. Due to this peak overlap, Bragg intensities (and their integrals) cannot be refined independently, necessitating the refinement of the complete diffraction profile at once. Following the peak identification and indexing, the experimental pattern is decomposed in single diffraction intensities, generally by the Le Bail [125] or Pawley [126] methods. The Pawley method is a least-squares refine-

ment technique that presumes knowledge of the space group of the material — and thus the systematic absences — and refines the crystal cell parameters, background, peak shape and intensities of reflections. This technique is particularly useful for confirming the space group of a crystal and the lattice parameters in a primary analysis of a sample [127].

Once the pattern is decomposed, direct methods can be applied analogous to those for single crystals for structure solution. If the information about the molecular geometry is available, direct-space techniques can be used. These are global minimization schemes that act to fit the powder diffraction pattern to produce a structural model [124, 128]. A variety of schemes and approaches exist including grid search [129], Monte Carlo [130, 131], simulated annealing [132], and genetic algorithms [133]. The use of prior chemical information acts to drastically reduce the number of available degrees of freedom, and is particularly well suited for structures with defined building units such as zeolites and framework materials [124].

The Rietveld refinement is another least-squares minimization technique, but treats a greater number of parameters [134, 135] than the Pawley refinement. Here, atomic fractional coordinates, atomic displacement or thermal parameters, and fractional occupancy are capable of being refined to fit the experimental intensity. Other parameters can be included to account for peak structure, preferred orientation, absorption or scattering effects [13]. The available model space (see Section 1.4) is thus exceedingly complex in this approach, and like all non-linear least squares refinement the solution will be very dependent on the initialization [136]. Therefore, a judicious choice of initial coordinates and parameters must be made. Suites of powder diffraction refinement programs exist in the mainstream, including TOPAS [137, 138], and GSAS [139].

1.3.2 Direct analysis of diffuse scattering

The phase problem has presented a formidable challenge in determining structures of ordered materials, a process which is now routine. When structures no longer retain

the periodicity of an ideal crystal, how can we use diffraction data to back out a structural model? In the previous section we highlighted the impact of disorder on the diffraction pattern, and the origins of diffuse scattering. In a primary approach, information can be extracted directly from the diffuse scattering. This is most common for powdered systems, where the 1D total scattering pattern is transformed into the PDF.

Analysis of powder data

By considering the structural origins of different features, robust local insight can be gained from the PDF. The peak positions in the PDF give the atomic pair separations. The integrated peak intensities are related to the atomic coordination numbers. The peak shape offers an understanding of the atomic pair probability distribution function; peaks follow near Gaussian shapes in the case of harmonic thermal and uncorrelated static displacements, whereas non-Gaussian peak shapes in the PDF reflect some underlying divergence from uncorrelated disorder*.

The direct measurement of bond lengths was integral in understanding the structure of β-cristobalite, where the bond lengths in the real material deviate significantly from those indicated by the average structure [140]. This example material will be discussed further when considering how this information was combined with forward modeling techniques.

The number of neighbors at a specific distance can be elucidated from the integrated intensity of a PDF peak. This value is particularly useful in the study of glasses and amorphous materials when considering the ring statistics as a metric of topology. For instance, disordered carbons have been explored recently for their potential utility in battery technology. Petkov *et al.* examined a synthetic series of pyrolysis temperatures and extracted the coordination numbers from the PDF [141]. They were able to differentiate between regular three coordinate carbon in graphitic rings and and the lower coordination introduced by larger rings and defects

*Gaussian peak shapes do not necessarily guarantee a lack of correlation, as is the case near phase transitions [58].

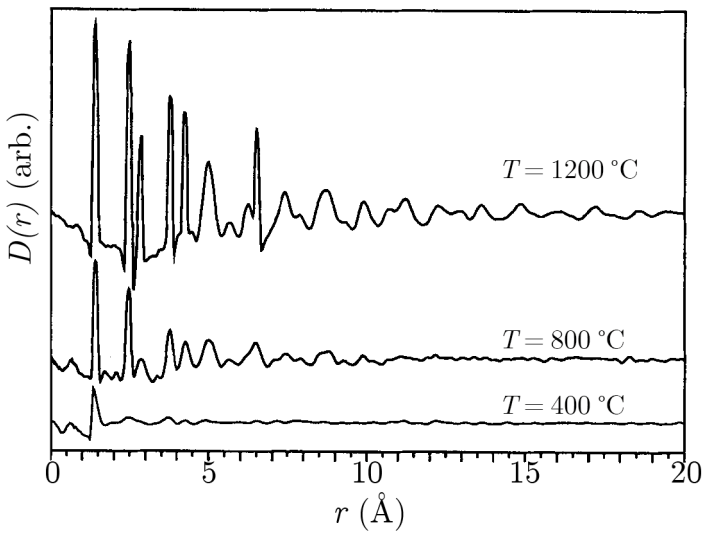


Figure 1.10: PDFs from disordered carbons. The series of PDFs from different pyrolysis conditions produce nanoporous products with varying degrees of disorder. The reaction at 1200 °C produces complete graphene sheets, whereas the cooler treatments introduce disorder in the form of greater sheet fragmentation and larger rings. Figure adapted from Ref. [141].

in the more disordered structures [Fig 1.10]. These results were confirmed by the topographic features in microscopy.

Lastly, peak width in the PDF reveals information about the static and dynamic disorder of atoms; however, this often warrants further modeling and consideration [12]. Across temperatures, the width gives information about the Debye temperature of a bond. The width as a function of interatomic separation offers information about the correlated atomic motions and underlying forces at play. Furthermore, the relative disorder across a phase diagram can be inferred from the peak widths across a range of compositions.

Analysis of single crystal data

There are also methods which attempt to make use of single crystal diffuse scattering directly [142]. Although these are less commonly used than their 1D counterparts due to experimental and computational considerations, they can elucidate robust information and serve as an important example of the discriminating power of diffuse scattering.

In work focused initially on binary alloys, Welberry and coworkers have expressed

the intensity of the diffuse scattering by first considering the perturbations to the average structure described in equation (1.12) on an atom by atom basis.

$$I(\mathbf{Q}) = \sum_n \sum_m f_m f_n \exp\left[\mathrm{i}\mathbf{Q}\cdot(\mathbf{r}_m + \mathbf{u}_m - \mathbf{r}_n - \mathbf{u}_n)\right] \tag{1.44}$$

Recall, $I(Q)$ is the scattered intensity and f_m is the scattering factor of atom m associated with lattice position $\mathbf{r}_m$. Now we have considered a small displacement from the average location $\mathbf{r}_m$ by the vector $\mathbf{u}_m$ at the root of the diffuse scattering. This equation can be expanded in a Taylor series to separate different contributions to the total scattering [11].

$$I(\mathbf{Q}) \simeq \sum_n \sum_m f_m f_n \exp\left[\mathrm{i}\mathbf{Q}\cdot(\mathbf{r}_m - \mathbf{r}_n)\right] \times \{1 + \mathrm{i}\mathbf{Q}\cdot(\mathbf{u}_m - \mathbf{u}_n) - \frac{1}{2}\left[\mathrm{i}\mathbf{Q}\cdot(\mathbf{u}_m - \mathbf{u}_n)\right]^2 + \frac{i}{6}\left[\mathrm{i}\mathbf{Q}\cdot(\mathbf{u}_m - \mathbf{u}_n)\right]^3 + \dots\} \tag{1.45}$$

This formulation immediately separates the Bragg component of scattering (the first term) from increasing moments of the displacements. In this form, we can break down the contributions to the diffuse scattering (the second term onward),

$$I_{\text{Diffuse}} = I_0 + I_1 + I_2 + I_3 + \dots. \tag{1.46}$$

Considering the total scattering expression in terms of correlations in displacements (see Appendix A of reference [11] for a complete derivation), we can begin to understand the nature of these discrete terms. The first term, I_0, is commonly known as the short-range order (SRO) diffuse intensity and is independent of displacements. It is a cosine Fourier sum with coefficients α^{ij}_{lmn}, where

$$\alpha^{ij}_{lmn} = 1 - \frac{P^{ij}_{lmn}}{c_j} \tag{1.47}$$

over interatomic vectors lmn for species i and j. These are the short-range order

parameters defined by Cowley [65]. A predictive metric appears here as the conditional probability of finding an atom of species j at the end of the interatomic vector lmn from the origin which contains an atom of species i, P^{ij}_{lmn}. The higher order terms all depend on displacement-parameter components X^{ij}_{lmn}, for each cartesian direction defined by

$$X^{ij}_{lmn} = u^{xj}_{lmn} - u^{xi}_{0}, \tag{1.48}$$

where u^{xj}_{lmn} is the displacement in the x direction of an atom with label j situated at the end of the vector $\mathbf{r}_{lmn}$ and $u^{x}_{0}i$ is the corresponding displacement in the x direction of an atom with label i at the origin of the vector. I_1, is known as the 'size-effect' where static atomic displacements cause an asymmetry in the diffuse scattering [143]. The next term involves displacements that include both static contributions and thermal diffuse scattering (TDS). It is generally expected that these terms sufficiently approximate the diffuse scattering due to the magnitudes of displacements encountered in real materials, and higher order terms generally lead to an impractical number of component intensities even for simple systems.

From this mathematical understanding of diffuse scattering, a primary step towards structural solution can be made. Firstly, an appropriate model for disorder needs to be formulated before any practical analysis can be accomplished. Fortunately, there are some intuitive rules for interpreting single crystal diffuse scattering in reciprocal space. For instance, small displacements from the average structure will only contribute strongly to the diffuse scattering at large values of $\mathbf{Q}$, and the width of diffuse features will be inversely related to the correlation length of local order [142]. This is apparent from even the most intuitive nature of the Fourier transform and reciprocal space [144]. In contrast, compositional disorder will have an impact at all scattering angles due to the difference in atomic form factor. We can also make perceptive interpretations of anisotropy in the diffuse scattering. We saw in the case of square ice (Fig. 1.8), 1D long range order can lead to diffuse planes with sharp profiles along the periodic directions. Anisotropy is also caused by the

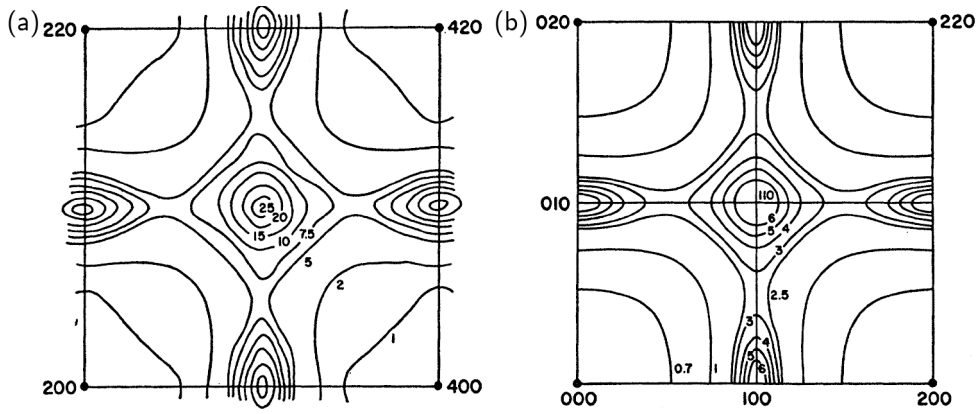


Figure 1.11: Single crystal diffuse scattering map for Cu_3Au. (a) The Cowley [147] diffuse scattering map for Cu_3Au at 405 °C in the h_1h_20 plane of reciprocal space. (b) A theoretical contour plot for the same system due to Moss and Clapp [150]. These approaches to understanding the diffuse scattering depend on direct calculation of short-range order parameters with probabilistic significance.

size-effect scattering term, I_1, as it is comprised of a sum of sine terms [142], with the directional difference in intensity related to the respective size and scattering factors of the species [145, 146].

However, more quantitative interpretations can be made from direct inversion of the diffuse scattering, particularly when the real structure is locally perturbed from the average structure. The earliest attempt at calculating the SRO coefficients [eq. (1.47)] came from Cowley in a study of Cu_3Au, in which the measured diffuse scattering was inverted and the α_{lmn}^{ij} normalized such that the probability that atoms of different type occupying the same site is zero [147]. Calculation of the SRO coefficients allowed for the testing of other theories of ordering in alloys [148–150]. This begins to elucidate some information, with the addition of higher order terms describing the static displacements caused by distinct species differing in atomic radius [143].

In fact, considering terms up to I_2 for a binary system, the diffuse scattering is represented by a total of 25 components (each component being the product of a Fourier summation and a unique coefficient that is dependent on the scattering factors and reciprocal space coordinates). Thus, given 25 linearly independent in-

tensity measurements a system of equations can be exactly solved for the quantities represented by the summations. Fourier inversion then yields the real space quantities: I_1 gives the average atomic displacement of atom pairs from their average structure sites; I_2 gives the degree of correlation between atomic displacements [47]. The procedure lends itself particularly well to alloys and minerals where the number of components can be limited, and the extracted parameters align intuitively with real space correlations or interaction potentials*. Short-range order and size effects have been separated for α-brass and used to determine effective pair potentials [151]. Beyond simple alloy systems, direct inversion has been used to examine non-stoichiometric oxides and halides such as TiO_x [152], VO_x [153], and calcia- and yttria-stabilized zirconia [154, 155]. Nonetheless, these systems become increasingly complicated and cannot be fully described by direct inversion alone.

An alternative function for examining local correlation and diffuse scattering is the 3D difference pair distribution function (3D-ΔPDF) obtained by the Fourier transform of the diffuse scattering [156, 157]. The function maps the difference between the total and average PDF, and thus indicates the pair correlations that are not described adequately by the average structure [Fig. 1.12]. Although this function in Patterson space contains the same information as the diffuse scattering, it is easier to understand intuitively. This real space approach aids in model building. Moreover, the 3D-ΔPDF has already been implemented and applied to refining the local structure of binary alloy systems [158], stacking faults in organic crystals [62], disordered defect layers in ternary alloys [159], and structural disorder in quasicrystals [160]. Finally, there have been recent efforts to use the 3D-ΔPDF in order to phase the diffraction pattern without the use of Bragg diffraction [161].

In practice, these approaches for directly interpreting the diffuse scattering — for powders or single crystals — deduce significant knowledge of the structure directly from the PDF or diffuse scattering. Nonetheless, the comprehensive analysis of

*These correlations were also considered in understanding the result generated by a more stochastic approach presented in Figure 1.7.

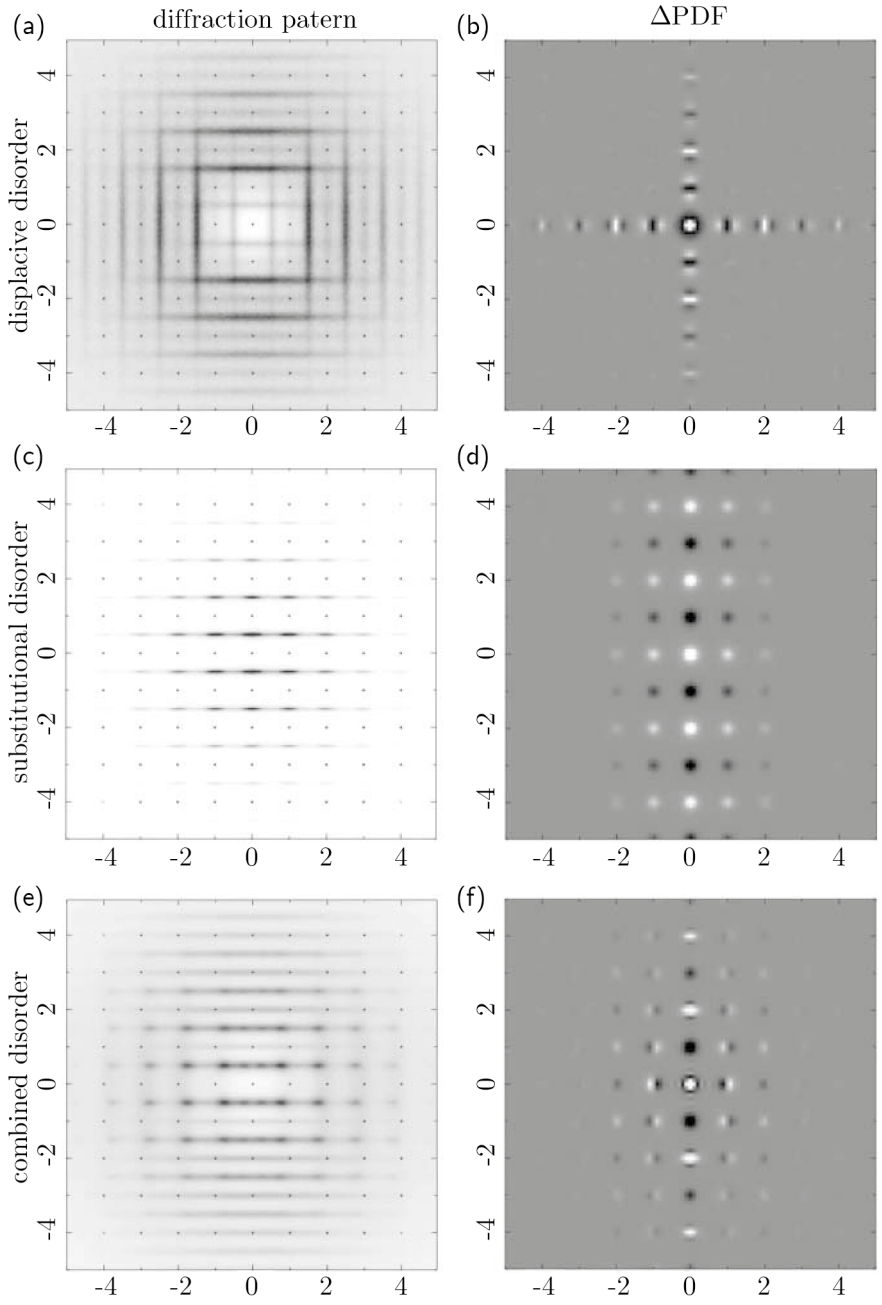


Figure 1.12: Model 3D-ΔPDF maps. Diffraction patterns and ΔPDF maps of synthetic Cu disorder models. The units are in fractional reciprocal ($\mathbf{a}^*$ = horizontal, $\mathbf{b}^*$ = vertical) and real space coordinates ($\mathbf{a}$ = horizontal, $\mathbf{b}$ = vertical), respectively. The grey-scale coding follows the convention that strong diffraction intensities are shown in dark and zero intensities are white, while positive PDF densities are bright, negative ones are dark and zero densities are shown in grey. The utility of the ΔPDF method is most apparent in considering the (e) diffuse scattering and (f) ΔPDF for combined substitutional and displacive disorder. The sombrero patterns of (b) indicate antiphase displacive disorder. The positive signals of (d) indicate a Cu-Cu (Void-Void) pairing, and negative signals indicating a preference for opposite pairs. In the combined case, these two phenomena are much easier seen in real space than in reciprocal space. Figure adapted from [142].

diffuse scattering requires some form of accompanying model. Although in the above examples some rules for reading the diffuse scattering were used to develop local order models, more complicated systems require further modeling and refinement that provides meaningful insight into the forces and mechanism behind the structural disorder.

1.3.3 Forward modeling of total scattering

The generation of appropriate models for disordered systems comes in two general forms: forward modeling, where interactions drive the structure prediction that is in turn validated by the total scattering pattern; and reverse modeling, where a fit to data drives the structure prediction that is in turn validated by physical sensibility. Each of these two fields of techniques can be further categorized by the size of the simulation environment.

Beginning with forward modeling of total scattering, the results of these structure predictions are model dependent. In general, some assumptions are made about the nature of the key energetic drivers in a material, and these are used to calculate an energy minimized system. A well established model dependent method for calculating a ground state structure is that of direct Monte Carlo [162].

Direct Monte Carlo

With the name originally coined in 1949 [163], and the first algorithm being published four years later by Metropolis *et al.* [164], its use has grown ubiquitous across fields of science and engineering today. A complete discussion of the method is offered in Section 2.4. Here it serves to understand that the total energy (Hamiltonian) of the crystal is cast as a high dimensional function of the relevant variables (positions, occupancy, orientation, *etc.*) and the Metropolis algorithm rapidly samples the Gibbs-Boltzmann distribution to find the energetic minimum. This minimization can be done *via* bespoke software of modern packages such as DISCUS [165]. From this model ground state, the diffraction pattern can be calculated and compared

against experimental data. Using this comparison, a new set of model parameters can be chosen and the process repeated iteratively to find sensible parameters. Although user dependent, this approach offers a direct description of the forces driving disorder. It is particularly useful in the case of analyzing three-dimensional diffraction data, because it lessens the computational load of regularly computing the complete diffraction data set. Nonetheless, it requires some assumptions to be made about the form of the Hamiltonian and the relevant variables. The direct Monte Carlo method has been used successfully in determining the structures of disordered molecular organics [166], relaxor ferroelectrics [167], piezoelectrics [168, 169], high temperature superconductors [170, 171], quasicrystals [172], and spin ices [173]. As an elaborate example, I present the detailed application of direct Monte Carlo to cubic stabilized zirconia.

The high-temperature cubic polymorph of ZrO_2 can be stabilized by Ca or Y ions, leading to anion-deficient materials with useful ceramic and superionic conduction properties [174, 175]. The cubic stabilized zirconia (CSZ) structure consists of a primitive cubic array of oxygen sites with cations in an fcc array filling alternate cubes of oxygens in each of the $\langle 1\,0\,0 \rangle$ directions [Fig. 1.13]. The cation ordering is however disordered, and the dopants create oxygen vacancies for charge balancing. Much of the early scientific debate rested on whether the dopant (Ca or Y) was eight coordinated [154, 155] or six coordinated [176, 177]. A series of thorough studies of Ca-CSZ revealed two types of microdomains: a single oxygen vacancy with relaxed neighboring ions and domains with a pair of oxygen vacancies separated by $\frac{1}{2}\langle 1\,1\,1 \rangle$. In both domains the calcium ion is associated with the vacancies [178–180].

Following the collection of comprehensive and detailed diffraction patterns for Y-CSZ [Fig. 1.13], a direct MC model was guided by the characteristic features of the single crystal data. The differences in the prominence of the dark lines when comparing Ca-CSZ and Y-CSZ stems from the scattering factors of the different dopants. Since $f_{\mathrm{Zr}} \simeq f_{\mathrm{Y}}$ the early terms of the diffuse scattering in equation (1.45)

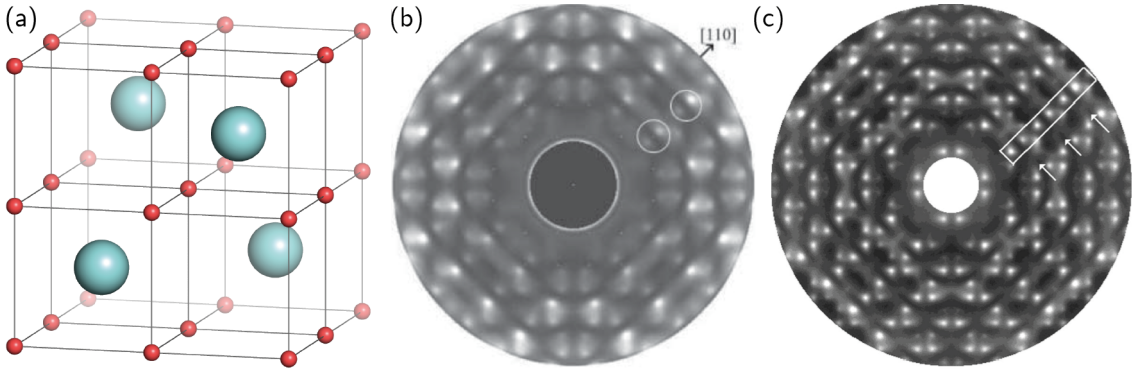


Figure 1.13: Structures and diffuse scattering of stabilized cubic zirconias. (a) Schematic diagram of the average unit cell of cubic stabilized zirconia. Small red spheres represent the primitive cubic oxygen array, which contains vacancies, and large teal spheres the cation array, which is complete but disordered. The differences in the diffuse scattering between form factors is demonstrated by the diffuse scattering in the $hk\frac{1}{2}$ section of (b) Y-CSZ [181] and (c) Ca-CSZ [182].

tend to zero, and higher order expansions are needed to account for the dark lines and asymmetry. However, in the case of Ca-CSZ where there is a significant difference in scattering factor, the lower terms are more relevant. This data set is especially useful in parsing the relative intensity of two peaks at low-Q and high-Q values where different terms from (1.45) dominate the diffuse scattering.

Combining this data and intuition with direct MC [183], a model for CSZs was established in which the O vacancies exist as mainly $\frac{1}{2}\langle 1\,1\,1\rangle$ pairs. Specifically, the MC considered modulations of these pairs across the average structure, and the probability of cation perturbations. The pairs account for the octahedral coordination expected for the small number of Zr cations, and distort the original cube of oxygen sites along body diagonal, propogating an isotropic strain field that repels similarly oriented 6-coordinated cation sites. This model has since been confirmed over a more extensive range of compositions [184]. It is pivotal to recognize that this structural model — however accurate — rests on the crystallographer's equivalent of a 'green thumb.' Without the reciprocal space intuition of the well practiced crystallographer, the potentials would not have been automatically formulated, making this procedure inaccessible to the non-expert. Moreover, in the case of TDS, these

static models will prove insufficient.

Empirical models

Physical models can also be cast by direct techniques that are not predicated on observational input. In the first approximation, empirical potentials can be refined within an MC scheme or a more deterministic minimization scheme. A popular implementation of a Newton-Raphson type method appears in the General Utility Lattice Program (GULP) [185]. Though seemingly simplistic, such an approximation allows for the rapid relaxation of structures and calculation of physical properties based on well established potentials, such as the Born-von Kármán model [186], the rigid-ion model [187], and the shell model [188], which also allow for phonon dispersion calculations that are consistent with experiment. These lattice dynamics were particularly relevant in diagnosing the source of thermal diffuse scattering in β-cristobalite.

Of the several polymorphs of silica, the β-cristobalite phase (average structure space group $Fd\bar{3}m$) has historically received considerable attention [189, 190]. Its structure defined by average atomic positions is analogous to that of elemental silicon; silicon atoms occupy the tetrahedral sites, where oxygen atoms occupy the *fcc* sites and positions half-way between nearest-neighbor silicon atoms. This creates an obvious problem of linear Si–O–Si bridges [140]. This model also casts the Si–O bond as shorter in the β-cristobalite phase than in the lower temperature α-cristobalite phase, and also predicts a displacive phase transition between the two. Using ADPs to describe the disorder observed in the thermal diffuse scattering yields oxygen ellipsoids that are highly anisotropic and orthogonal to the Si–Si vector [190]. As this is rather unphysical, an early approach considered multiple sites arranged about the average location with equal fractional occupation, casting the arrangement of oxygens as static positional disorder with sensible geometric environments. One step further would be to consider a continuum of oxygen locations [191]. This distribution of positions would be consistent with the average values extracted from

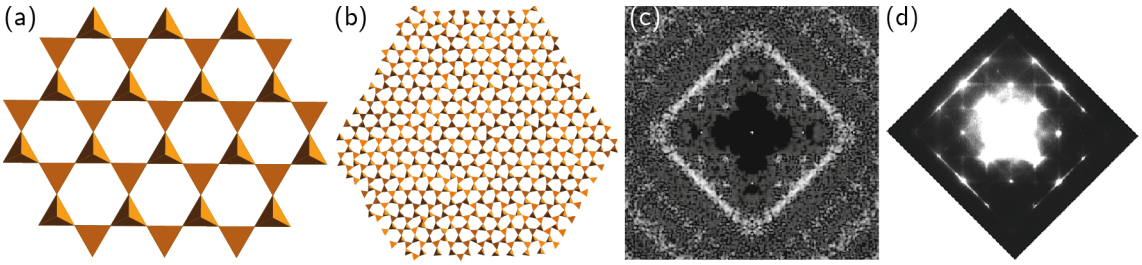


Figure 1.14: Thermal disorder in β-cristobalite is a dynamic phenomena that is best described by collective lattice vibrations of rigid polyhedral units. The apparent average structure in (a) has significantly higher symmetry than the model refined against neutron total scattering data at 900 K in (b). The deviations from the average structure correspond to occupancy of RUMs in the phonon dispersion. (c) The simulated total scattering pattern at 950 K reproduces the key features of the diffuse scattering observed in the (d) experimental electron diffraction data. The structures in (a) and (b) correspond to a $\langle 1\,1\,1 \rangle$ plane of polyhedra, and the $(h\,k\,0)$ diffraction is shown in (c) and (d). Figure adapted from Ref. [140].

refinements while leading to Si–O–Si bond angles nearer 145°.

Considering a distribution of positions lends itself to the discussion of dynamics, and the correlated atomic motions responsible for thermal diffuse scattering. The collective harmonic vibrations of a crystal are described by the phonons (discussed further in Section 3.1). At the heart of many physical phenomena, they also explain the positional disorder in β-cristobalite. Using force field models to describe the atomic equations of motion produced sensible matches to the diffraction pattern [189] with eventually sensible bond angles [191, 192]. With a polyhedral network it is possible — and convenient in the sense of a real space realization — to consider some phonon modes as rigid-unit modes (RUMs) [193]. These lattice vibrations create deformations at the linkages of SiO_4 polyhedra, while maintaining the geometry of the polyhedra.

This approach to structural solution is physically sensible and commensurate with real and reciprocal space approaches to scattering data. The utility of RUMs also allows the network structure to be described in terms of orientational disorder, rather than positional. Orientational disorder is more prevalent when dealing with fixed discrete units, such as rotating molecular guests in a porous framework material [91, 194]. While each of these models plays a role in eliciting particular understand-

ing, incorporating lattice dynamics into our understanding of positional disorder is the most precise and physical approach to analyzing thermal diffuse scattering. The concept of RUMs alone has been used to describe the thermal diffuse scattering observed in silica polymorphs [140, 195–199], zeolite frameworks [200–203], and negative thermal expansion materials [26, 27, 204–207].

Ab initio approaches

The utility of the approximations considered in empirical potentials and restricted model construction lies in their computational efficiency. In the examples considered above, these methods were used to describe large simulation boxes of many atoms. In the cases where these approximations will not suffice, or greater precision is needed to inform the application of these approximations, *ab initio* methods can be considered. Density functional theory (DFT) is used across the physical sciences to solve the time-independent ground state of the electronic structure in many-body systems [208]. A computationally efficient alternative to solving the Schrödinger equation, DFT has been applied to atoms, molecules, and condensed matter. In practice, this involves calculating the electronic structure of a unit cell without direct solution of the Schrödinger equation by optimizing the electron density as opposed to the electron wavefunction. The advantages of DFT come in the elegant formulation of N-particle quantum mechanics, offering conceptual simplicity combined with computational efficiency [209]. Nonetheless, its utility in describing disorder is limited by the number of unit cells required to encapsulate the mechanism of disorder. Implemented in a variety of packages, with a field dedicated to the approximations made in the calculations [210], DFT is capable of calculating ground state structures and probing the lattice dynamics. Relaxor ferroelectrics are just one example of functional materials that have been examined using DFT [211]. Often, the utility of DFT can be in parameterizing the force fields or potentials used in calculations without quantum mechanics [212, 213].

For information regarding the temporal nature of a system, molecular dynamics

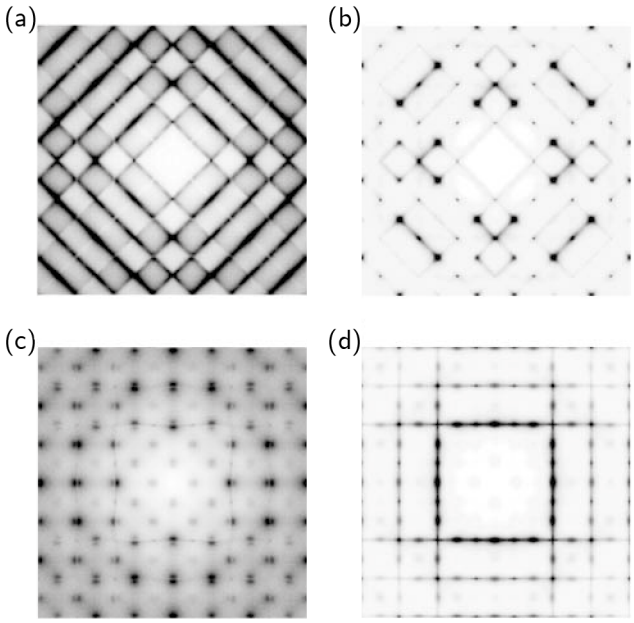


Figure 1.15: Diffuse scattering in MD simulated lead zirconate. $PbZrO_3$ diffuse scattering as calculated by molecular dynamics using an *ab initio* derived shell model. (a), (b) $hk0$ and (c), (d) $hk\frac{1}{2}$ planes calculated from all atoms (left) and from only O atoms (right). The intensities for the O atoms are scaled for clarity. By considering the correlations of different species in the MD result, the interactions between different phonon modes were observed. Figures taken from [214].

(MD) can be implemented to solve the classical equations of motion. The formalism used can be that of Lagrangian, Hamiltonian, or Newtonian mechanics, but in practice the simulation describes the positions and velocities of all the atoms in an ensemble [58]. The simulation progresses via a small time step, and the velocities and positions are updated on the basis of interatomic potentials and the laws of motion. The integration over the simulation can yield a trajectory that is in principle deterministic based on the starting coordinates. On the whole, this procedure is computationally expensive and doesn't regularly exceed simulation times on the order of nanoseconds. Several packages exist for implementing this procedure — each containing available sets of force fields — such as CHARM [215, 216] and GROMACS [217, 218], albeit they are generally specialized for unique purposes (*e.g.*, biomolecules). Though historically suffering from over-parameterization of force fields, modern developments in *ab initio* molecular dynamics are merging the quantum mechanical calculations of DFT with the Newtonian approach of MD [219].

For disordered crystals and diffuse scattering, the large (and slow) relevant atomic displacements require vast timescales of simulation. MD has been used to examine the phases of relaxor ferroelectrics [167, 212–214, 220], and to predict ground states of amorphous materials [221]. Even though MD offers a glimpse of the structural kinetics, for an understanding of crystalline ground states, it is often more effective to use lattice minimization procedures and lattice dynamical approaches to motion.

Model building for amorphous structures

A particularly unique subset of forward modeling has arisen from attempts at understanding topological disorder and amorphous materials. When the scattering data is isotropic and entirely diffuse, there is no Bragg scattering to describe a lattice foundation for the disorder. Looking back at our previous examples, the modeled disorder was approached as a deviation from an average crystalline structure. Without a periodic average structure, the task of characterizing and understanding disordered structure becomes even more challenging.

Unsurprisingly, initial speculation of amorphous structure was predicated on the relevant crystalline phases [222, 223]. The attempt to describe amorphous materials as crystalline structures of finite domain size often yields correlation lengths on the order of a single unit cell [224]. This would lead to defects throughout the entirety of the structure, an expectation inconsistent with the experimentally observed low concentrations of coordinative defects in amorphous frameworks such as Si and SiO_2 [225, 226]. The earliest reasonable approach to amorphous materials, involved casting the structure as a continuous random network (CRN).

A challenging task, building CRNs was initially done by hand in the well known work of Zachariasen in 1932, in a publication that included an illustration of a two dimensional random network with open boundary conditions [227]. Other hand built models have been made [Fig. 1.16]; however, without periodic boundaries conditions or a sensible box size, the ability for these models to predict bulk properties becomes limited [228]. As computational methods became accessible, a number or methods

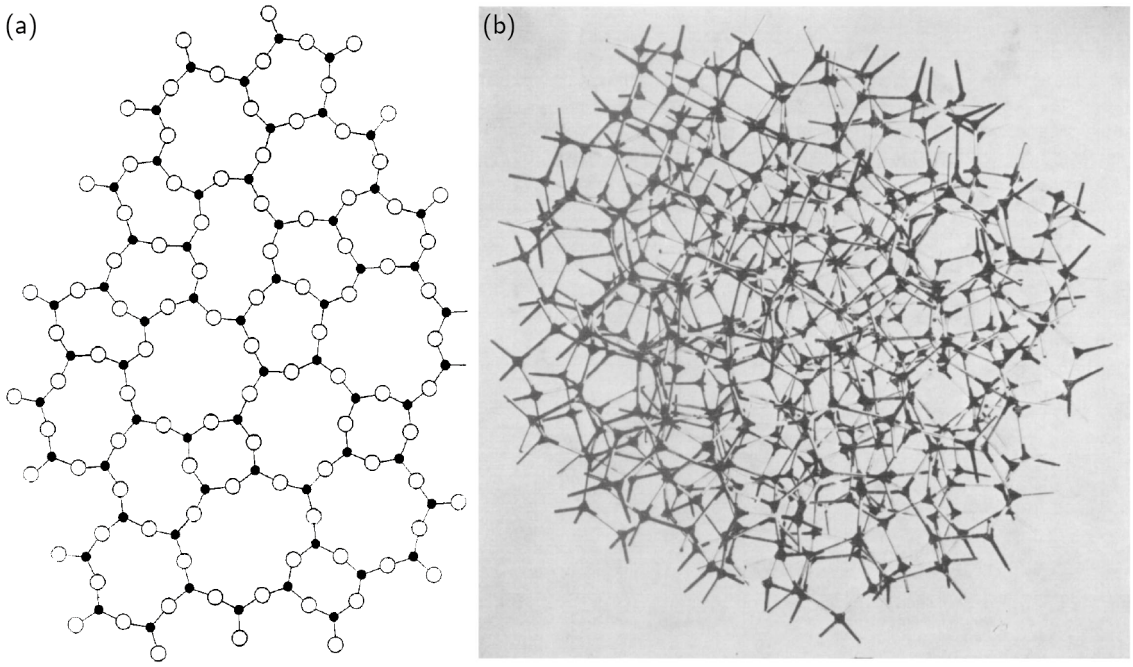


Figure 1.16: Early models of amorphous materials.(a) Zachariasen's original hand drawn model of a CRN under open boundary conditions. (b) a photograph of a physical model of *a*-Si. Figure adapted from [227] and [231], respectively.

were developed to model amorphous materials, specifically CRNs. The breadth of these methods is briefly described below, with a focus on the Wooten, Winer, and Weaire (WWW) approach [229, 230].

Many of the attempts at describing amorphous systems arose out of desire for a greater understanding of silicon (and germanium), due to its semiconducting amorphous phase. The first successful attempt was made by Guttman *et al.* with limited scaling ability [232]. This model was only later used to describe SiO_2 by inserting O atoms at the middle of the Si–Si bonds in the *a*-Si model, and rescaling the cell size [233]. Over the years this model has been reused, repurposed, and improved upon in other structural studies [26, 234–236].

Another early approach was in quench-from-the-melt molecular dynamics, with the Stillinger and Weber (SW) potential first generating an MD model of *a*-Si [221]. Other such models were developed, but all grappled with limited box size and highly strained configurations [237–240]. Another drawback of using dynamics, is the time constraints on the system force unphysically fast cooling rates [241].

The modern bond-switching algorithm was also developed around this time, and created models of *a*-Si and *a*-Ge that were in good agreement with diffraction data [229, 230]. The WWW method, and its derivatives have become the standard for a scalable means of constructing amorphous networks. A detailed description of the algorithm will be offered in Section 4.2. In brief, the WWW algorithm constructs a CRN from an ordered or random starting framework *via* a Markov chain Monte Carlo process, where the 'moves' in the Monte Carlo are treated as changes in the connectivity of the framework followed by a global energy minimization (structural relaxation). The energetics associated with the minimized states before and after the bond swap are used in determining the probability of acceptance of the move.

The original WWW simulations constructed a 216 atom cell of *a*-Si from a diamond cubic starting topology [230]. Useful measurements of the strain in these systems are the root-mean-squared (rms) bond length and angle deviations (from equilibrium tetrahedral values). In the original box in 1985, the rms bond length and bond angle deviation were 2.7 % and 10.9°, respectively. Historical values for these parameters are offered in Table 1.2. Larger models with 4096 atoms have been created using the standard WWW approach [242, 243]. The thermodynamics of the amorphous Si and Ge systems have been considered using statistical mechanics across the WWW simulations, treating bond swaps as excitation events [244].

The WWW algorithm has been improved upon by changing the potentials in the geometric relaxation and by considering different starting configurations. Barkema and Mousseau (BM) pioneered the idea of using an initially randomized network as a starting configuration to avoid any artificial artifacts of crystallinity from the starting structure [245]. The BM algorithm uses the same external bond switching procedure as WWW, but reduces runtime by aborting unfavorable configurations early in the geometric relaxation [228]. This method has been expanded with computational efficiency and by expanding the box size as a supercell of optimized WWW configurations [246–248]. The WWW method can also be combined with an inter-

Table 1.2: Simulation of amorphous silicon using bond-switching methods. $\Delta\theta$(rms) is the root mean squared deviation of the tetrahedral bond angle from its mean value, $\sigma_r(\%)$ the standard deviation in bond lengths. It is noted that these values are not offered by Ref. [248], but comparative distributions are given.

Year	Method	# atoms	$\Delta\theta$(rms)	$\sigma_r(\%)$	Ref.
1989	Exp.	-	9.3	-	[89]
1985	WWW	216	10.9	2.7	[230]
1995	WWW	4096	10.5	2.5	[242]
2000	BM	1000	9.2	3.7	[245]
2000	BM	4096	9.9	4.2	[245]
2001	BM	10k	10.5	4.2	[246]
2001	BM	20k	10.2	4.2	[246]
2013	NHN	20k	9.2	2.7	[249]

nal MD relaxation to model *a*-Si, and has recently been attempted by implementing the BM algorithm using Stillinger-Weber potentials to further refine each step of the external MC, creating a 'nearly hyperuniform network' [249].

Beyond generating models of *a*-Si, silica models have been of great interest. Because of the Si–O–Si geometry, these were first generated by decorating *a*-Si models [250, 251]. Bond switching algorithms can also be applied directly to SiO_2 [252, 253]. Further complexity can be introduced to allow for dangling bonds [254], or large configurations up to 300,000 atoms [255] [Fig. 1.17]. Importantly, other unrelated binary systems such as the amorphous semiconducting phase of GaAs have been cast using a modified WWW algorithm that incorporates dangling bonds and the exchange of chemical identity [256].

The CRN model has several limitations: (1) the models are unrealistically defect free [257], (2) have perfect coordination when real systems do not [258], (3) underestimate the density [259], and do not contain impurities such as hydrogen in *a*-Si [260]. In order to overcome these limitations, other avenues of simulation have been explored. Molecular dynamics has been used to understand any dynamics of the system; however, has significant drawbacks in scalability and unrealistic distortions. Contemporary MD has described amorphous systems using a machine

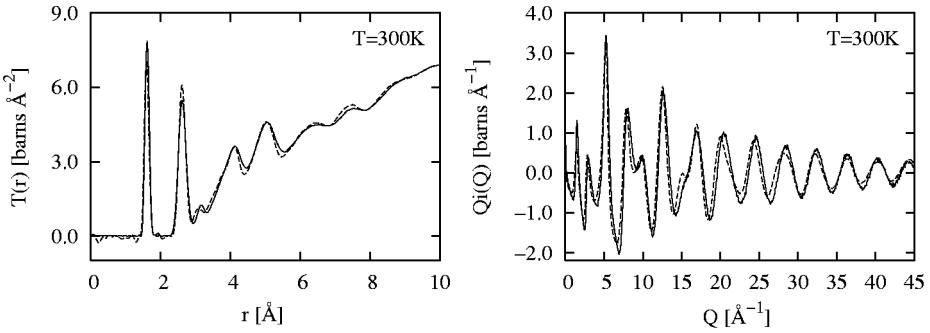


Figure 1.17: Amorphous silica CRN. Comparison of the PDF and neutron total scattering of *a*-SiO_2. The dotted line is from experimental data, and the solid line is from a WWW type simulation as produced by Barkema and coworkers [255]. The simulated data has been calculated to account for the 300 K data acquisition. These configurations have more narrow bond angle distributions and lower defect concentrations than networks produced by MD.

learning approach for incorporating DFT, reducing the cost of *ab initio* MD while maintaining a comparable accuracy [261]. A distinctly different approach — the activation–relaxation technique (ART) — considers paths in the energy landscape from local minima to saddle points [262, 263]. This method depends on SW potentials and has been applied to several amorphous materials, including GaAS [264], SiO_2 [265]. Both MD and ART give structures which are similar to successful WWW type simulations; however, have the freedom to adjust (sometimes too much) for the limitations of CRN models. Moreover, modeling amorphous structures based on potentials consistent with the liquid state leads to amorphous states which are much too liquid like [262], even though the liquid state and amorphous state have very different structures [266, 267]. Nonetheless, there is not an absolute consensus regarding the suitability of all of these methods to the study of archetypal amorphous materials, such as *a*-Si and SiO_2 [228].

A significant body of competing work has approached the drawbacks of CRNs by considering amorphous materials as paracrystalline [268]. A paracrystalline material in this case is defined by regions of crystalline medium-range order (crystallites) embedded in a truly amorphous material. The primary supporting data for the paracrystalline model comes from fluctuation electron microscopy (FEM), which is claimed to be incompatible with CRN models [269]. Some of this discrepancy comes

back to the non-uniqueness of the PDF that was explained in the prior section and the exploitation of experimental methods to go beyond atom–atom correlations to higher order pair–pair correlations [270]. With insight into the medium-range order of a material, recent work has suggested a volume fraction of around 50 % for crystallites in *a*-Si [271, 272]. The magnitude of this volume fraction has been contested by spectroscopic ellipsometry measurements [273].

Even when paracystalline silicon has been modeled using bond switching methods, the presence of a low volume fraction of crystallite grains does not significantly change agreement with the experimental PDFs as compared to CRN models [228, 274]. This fact is due to the information poor nature of the PDF [275]. As previously considered in kinetic studies, it would also be unsurprising if an amorphous structure had some memory of the previous crystalline phase [276]. Outside of the debate of how much paracrystallinity is present, the effects of paracrystallinity can equally be explained by voids, defects, and imperfections [276].

1.3.4 Reverse modeling of total scattering

The previous approaches in connectivity transformations or energetic relaxation relied on the assumptions of an explicit geometry or knowledge of the interactions within a material. When these forward models produce structures that are incommensurate with experimental data, it challenges their respective assumptions and becomes desirable to formulate data-driven approaches.

Real space Rietveld refinement

One such data-driven approach, is the real space parallel to Rietveld refinement. Using a least squares procedure, the total scattering data can be fitted against in real space by transforming it into the PDF. Introduced to satisfy the need in probing local instead of average structure, these procedures have similar operating principles as the Rietveld refinement, but in real space [277].

This utility comes from the impact of small modulations in the total scattering at

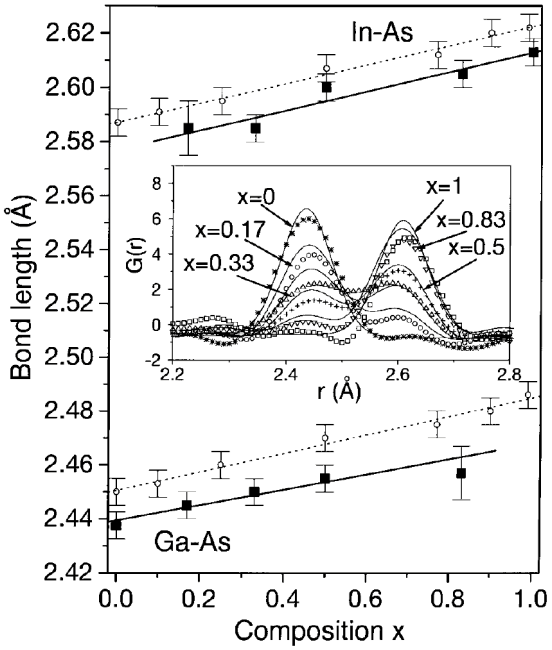


Figure 1.18: PDF refinement in $Ga_{1-x}In_xAs$. PDF refinement was used to analyze the peak splitting across compositions of $Ga_{1-x}In_xAs$. The bond distances as a function of composition are displayed as extracted from the PDF (solid squares) and compared with those from X-ray absorption fine structure (open circles). The inset shows the fits (solid lines) to the PDF data (symbols) for varying composition. The use of total scattering data with a PDF refinement is not only able to match the X-ray absorption fine structure data, but also make clearer statements about order beyond the first or second coordination shell. Figure from [278].

high-Q on observable low-r correlations in the PDF. The PDF resolution (and thus ability to separate peak overlap) is dependent of the maximum measured value of Q [12]. Although this experimental limitation is significant, increasing technological advances have brought about a surge in the utility of real space refinement. The program PDFGUI has been used in this thesis, and is capable of refining atomic positions, lattice parameters, atomic displacement parameters, and some factors for peak and profile shape [279]. Since its introduction, PDFGUI (and its predecessors) has been utilized in understanding the structures of disorder and strained materials, including: solid solutions that don't obey Vegard's law [280, 281] [Fig. 1.18], glasses [282, 283], spin glasses [284], nanostructured materials [285], intercalated layered materials [286, 287], and transition metal oxides with orbital disorder [288, 289].

A particularly interesting case study of real space refinement comes from that of $LaMnO_3$, in which local orbital states give rise to colossal magnetoresistance (CMR)

in select doped phases [290–293]. It has been identified that an orbital order/disorder transition in the material is discontinuous, creating a challenge in that the disordered state may differ dramatically from the corresponding and well-defined ordered states [289]. Although this thesis will not pursue CMR or materials with orbital disorder, this study highlights a difficulty in discriminating between local structural models when both static and dynamic disorder are present. Moreover, the non-uniqueness of the PDF readily manifests itself in a large simulation box, but here it is apparent in a small box PDF refinement. Such a refinement still depends on the initial structure prediction; however, PDF-driven modeling has demonstrable utility, even when considering the information entropy of the observation.

Reverse Monte Carlo

As an approach to large-box data-driven modeling, the reverse Monte Carlo (RMC) method applies the protocols of MC to minimize a residual function of the difference between experimental and simulated data [294]. Beginning from either a crystalline or random configuration, the procedure is ignorant of any symmetry or atomic interaction beyond the periodic boundary conditions. Thus, RMC has a strong tendency to favor the most random arrangement of atoms in a box that will provide a fit to the experimental data [295]. Initially developed for for the study on non-crystalline materials, RMC (and its most popular implementation RMCProfile [296, 297]) has had success in probing the structure of an array of materials, including: nanomaterials [298]; hydrated minerals [299]; liquid [300], amorphous [301], and crystalline metal-organic frameworks [302]; spin liquids [303]; and piezoelectric [76] and ferromagnetic solids [304].

The penchant for disorder is beneficial when RMC is applied to systems where the nature of disorder is initially unknown. As seen in the recent approaches to phase transitions in MOFs, an RMC approach was able to provide a more complete picture of the hydrogen bonding in dimethylammonium manganese(II) formate, without presupposing this behavior [305]. These forces were extracted from

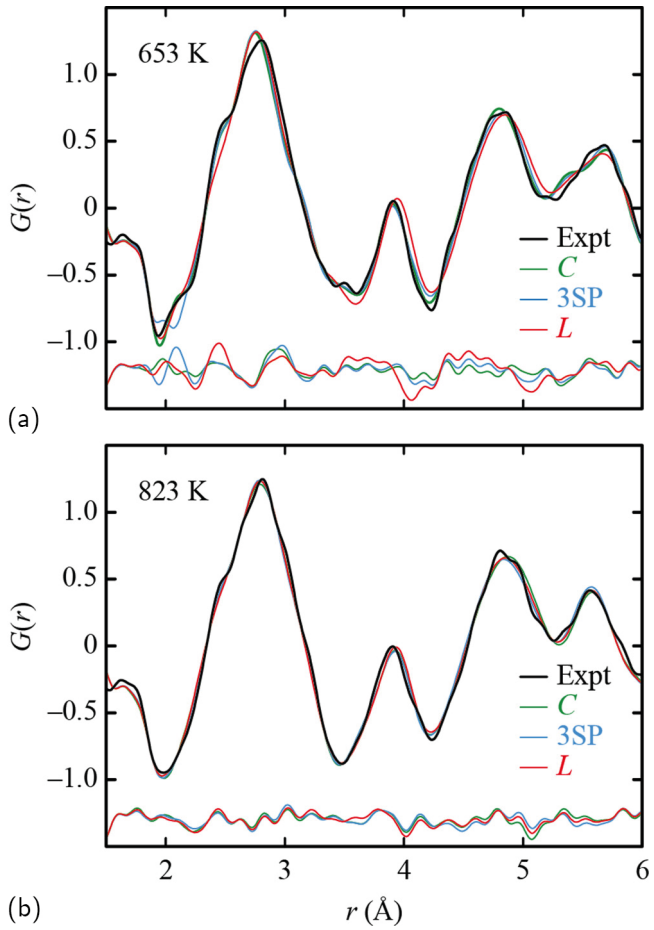


Figure 1.19: Competing models of $LaMnO_3$. Neutron PDF fits for data collected from (a) orthorhombic orbital ordered and (b) orthorhombic orbital disordered phases of $LaMnO_3$. The fitting was accomplished using PDFGUI with three different model types, with the difference functions (data−fit) shown in the respective color and vertically shifted for clarity. The ability for three unique models to produce remarkably similar fits to the PDF is indicative of the inherent limitations of this 1-dimensional function in isolation. Figure taken from [284].

the pairwise correlations that are inaccessible from the average structure, and developed a stronger understanding of the A-site cation orientation and local structure in the material. A similar approach was applied to Prussian blue analogs to develop a qualitative model of the competing interactions within the framework [306]. This model was able to explain the nature of the phase behavior (that underpins the unique electronic properties), revealing an inadequacy in the average structure description of disordered framework materials.

Continuing with our less ordered example of *a*-Si, the capabilities and limitations of RMC can be discussed. Naïve application of RMC to *a*-Si involves considering

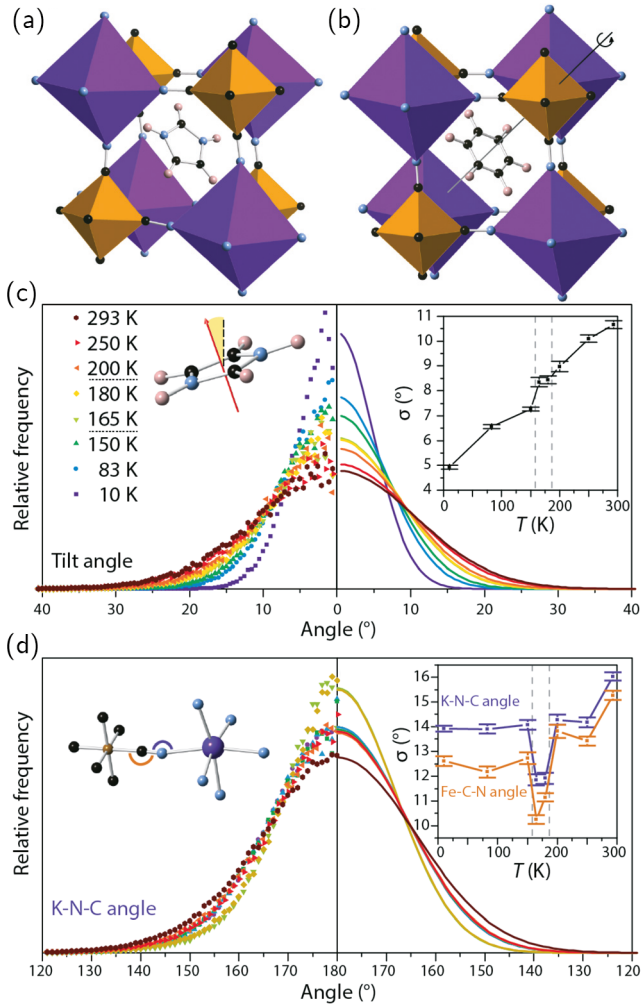


Figure 1.20: Cation disorder in Prussian blue analogs. The phase behavior of the intermediate- and high-temperature phases of potassium immidazolium hexacyanoferrate, $(C_3N_2H_5)_2K[Fe(CN)_6]$, was studied using RMC, with particular attention paid to the cation dynamics (the ordering of this cation in a low-temperature phase causes a stark change in the dielectric behavior of the material). The average structures of (a) the low-temperature ordered phase, and (b) high temperature phase where the imidazolium cation is disordered about the indicated threefold axis. Using RMCProfile to model temperature-dependent neutron total scattering data from a set of sensible starting configurations, the distributions (c) imidazolium tilts and (d) metal-cyanide angles were calculated to probe the dynamics. Gaussian fits to the distributions are given as solid lines, with the standard deviation σ presented as an inset. This indicated the difference between the high- and intermediate-temperature phases as related to cyanide tilting and a gradual decrease in bond length — an understanding that is only available in characterizing the local structure underpinning the disorder. Figure adapted from [306].

the simultaneous fit to the total scattering data and pair distribution function. It is expected that this will take advantage of the local correlations prominent in the PDF while maintaining consistency with the average structure by fitting in reciprocal space. This approach — not including any assumptions about the atomic interactions or topology — leads to structures that are inconsistent with other experimental measurements and contain high energy features [307, 308]. The failures of unrestrained RMC can be demonstrated in the variations in topology, local geometry, and coordination between methods. In Figure 1.21 it is shown that vastly unphysical coordination environments can emerge from RMC refinements: zero- and nine-coordinate silicon atoms exist in the result. All the while, the RMC and WWW results maintain comparable fits to the PDF. This is unsurprising given the previous discussion of the information content of two body correlations [11]. Moreover, the RMC result has a bond angle distribution that is much broader about the tetrahedral angle, and includes a population with 60° angles from three membered rings. These kinds of features are chemically nonsensical; however, this method is not predicated on chemical sensibilities.

With such an affinity toward disorder, the RMC method needs to be buttressed in order to be effective. There have been multiple approaches to keeping the efficacy of the data-driven algorithm, whilst still maintaining some levels of chemical sensibility. Each modification does however require some *a priori* knowledge or assumption about the nature of the material. Firstly, one can recognize the features of failure (*e.g.*, topological defects, unphysical coordination numbers, or geometric variance), and provide constraints in the minimization procedure that includes an energy term for these features [308, 310]. The use of restraints (particularly polyhedral restraints) has been implemented in modern editions of RMCProfile, to expand its utility beyond amorphous materials into disordered frameworks [297].

Another novel approach, and notably one that does not require any specific assumptions regarding local geometry, is the structural invariance approach im-

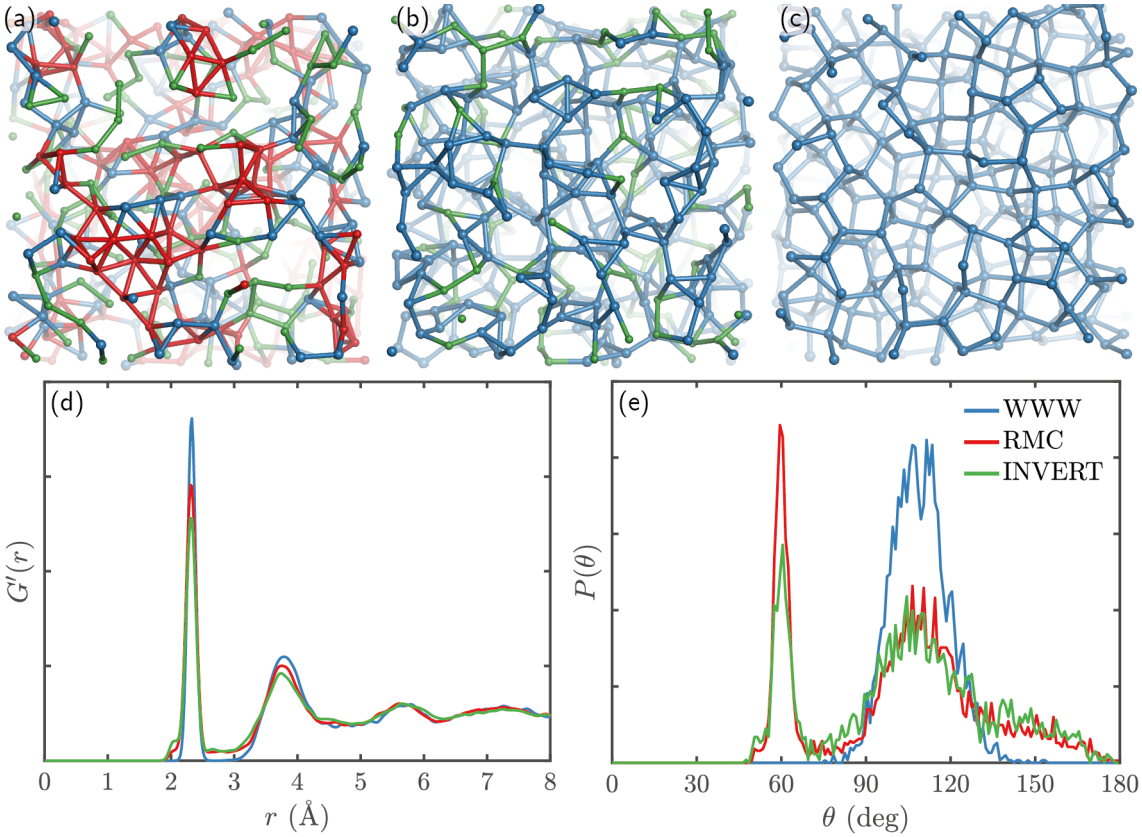


Figure 1.21: Comparison of different methods for modeling a-Si. The three representative slices are from structures that produce near equivalent PDFs: (a) a high quality model generated using the WWW method [230], (b) a model generated by naïve RMC refinement against PDF data, (c) a model generated RMC refinement against PDF data making use of in addition the INVERT PDF similarity constraint [309]. Four coordinate atoms are colored blue, atoms with higher coordination are shown in red, and three coordinate atoms are shown in green. (d) The PDFs from the WWW model (blue), the naïve RMC (red), and INVERT approach (green) all show near equivalent PDFs despite their divergent structures. (e) The bond-angle distributions for the three models highlight the structural differences. The naïve RMC has strongly distorted polyhedra, as shown by the broad peak centered at the tetrahedral angle ($109.5°$), and a large number of $109.5°$ bond angles corresponding to Si_3 triangles. The INVERT similarity constraint reduces the concentration of triangular defects (as seen in both the structure and angular distribution), and subsequently reduces the unphysical areas of high coordination. By design, the WWW model shows a single peak centered at the tetrahedral angle and no coordination defects. Structural coordinates courtesy of Cliffe *et. al* [309].

plemented in INVERT [309, 311]. Rather than assuming that silicon has a four-coordinate environment, the INVERT approach depends on the inference that similar atoms in a material will favor similar environments. This expectation retains the relative naïvity of RMC, whilst ameliorating the failures in physical sensibility. The methodology has since been expanded to successfully characterize disordered

spin systems [312–315].

One could also consider falling back on the assumption of interaction potentials to provide some physical influence to the data-driven approach. This still necessitates parametrization that makes assumptions about the chemical environments in the material, and in practice presupposes a local geometry [221, 238, 316]. To include energetics without the simplification of an empirical potential, first principles calculations would need to be implemented. This however imposes computational limitations, and confines the calculations to RMC relaxed structures in an iterative process [310].

Given that much of our chemical intuition informing any empirical potentials comes from spectroscopic data (*e.g.*, Raman and IR spectroscopy), it makes sense that we look to constrain the initial RMC approach through other experimental data. The challenge here lies in the computational expense. RMC was only made practical by the FFT, leading to a rapid calculation of total scattering from a structural model. In the case of prime candidates for local information, a quantum mechanical approach would be necessary. In order to calculate NMR or vibrational spectra, an expensive first-principles implementation is needed that makes the inclusion impractical. Extended X-ray absorption fine structure (EXAFS) has been successfully used in RMC to restrain the partial PDFs, offering insight into the local two-body correlations, but not completely overcoming the challenges of RMC [22]. Even though we are incapable of calculating some observables for direct comparison with experiment, the nature of additional experiments can inform the approach to parsing the scattering data.

Empirical Potential Structure Refinement

A conceptual extension of the RMC approach is that of Empirical Potential Structure Refinement (EPSR) [317]. This procedure refines pairwise interatomic potentials against the diffraction data (instead of fixed implementation in conjunction with structure refinement as in RMCProfile). Here the moves proceed similar

to any other MC, with molecular translations, rotations, and deformations being considered. This creates a crossover between data-driven and interaction-driven approaches. It tackles a similar problem as the INVERT technique (of ensuring chemical sensibility in a data-driven approach) by introducing explicit chemical information into the refinement. Notably, the refined potentials included in EPSR go beyond harmonic potentials of RMCProfile, and include Coulomb and Lennard-Jones potentials. This approach to modeling structure from total scattering has been particularly useful when considering liquid phases [318].

Although there are computational motivations for having differentiable functions driving a structural minimization procedure, the choice of potentials in EPSR will have a non-trivial impact of the outcome of the refinement, and there is some debate as to whether the potentials are useful outside of the structure modeling (*e.g.*, for thermodynamic calculations) [319]. This further complicates the interpretation of the structure prediction, as compared to RMC refinement.

1.3.5 Key challenges

In applying the routines from traditional crystallography, least-squares techniques can provide useful lattice information, refine existing models, and locate areas for more precise examination (we observed this in the large anisotropic ADPs in SiO_2 necessitating a lattice dynamical treatment). To create a model from scratch requires some ingenuity, and is often accomplished by beginning with an ordered analogue. From this, direct inversion of the diffuse scattering can extract the nature and magnitude of the disorder; however, this procedure requires some considerable intuition. These direct techniques are particularly useful when high quality single crystal diffraction data sets are available. As such, they are less accessible for any crystalline material that cannot be grown into a suitable crystal for the diffraction experiment, or any non-crystalline condensed materials.

When forced to work with one-dimensional diffraction data, additional efforts must be made to generate a sufficient model for refinement. For amorphous behav-

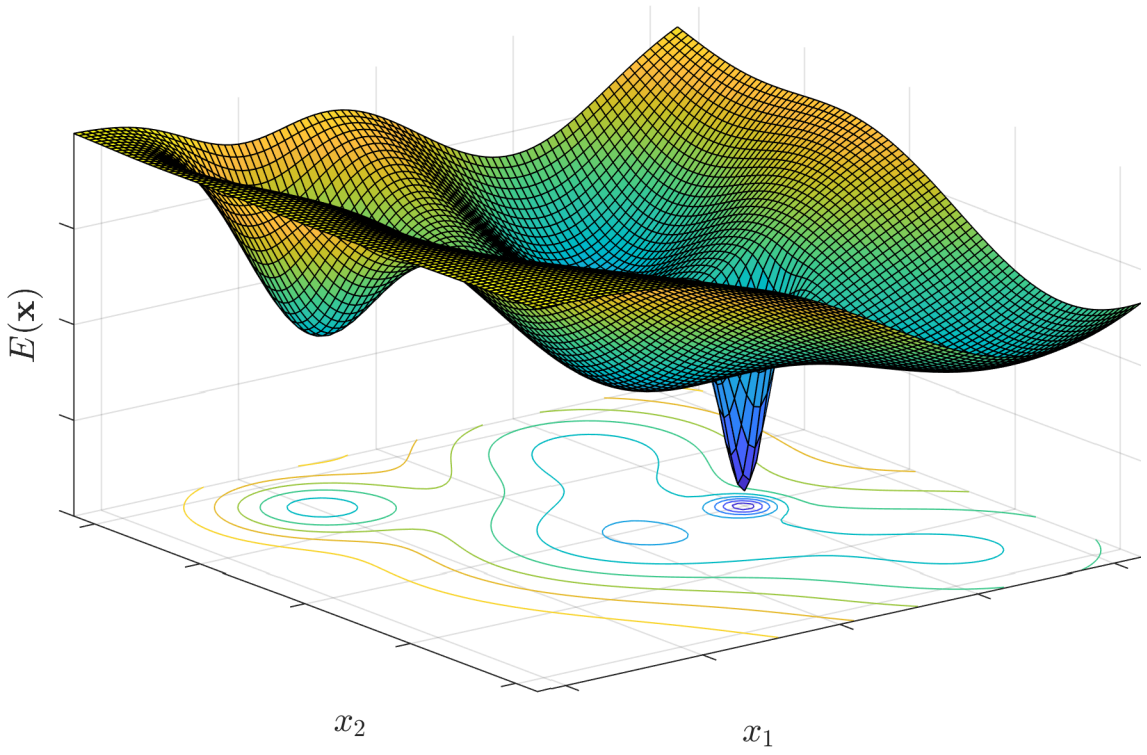


Figure 1.22: A demonstrative energy landscape. In just a two-dimensional slice of phase space, converging to a global minimum is not guaranteed using analytical or linear methods. When considering structural minimization these landscapes become exceedingly complex, posessing up to $3N$ dimensions.

ior, a purely topological approach was sufficient for early structures. In the case of physical modeling, there are varying degrees of precision available. In general, this precision has an inverse relation with computational expense. The quantum mechanical treatment of DFT limits the number of atoms to be included in the calculation, whereas empirical potentials are relatively facile to calculate for large boxes. The tradeoffs of this relationship are complicated by the requirements for a minimum number of atoms to encompass the origins of disorder.

One of the great challenges associated with disordered phases is finding the primitive unit to describe the structure. In the absence of the periodicity of crystalline materials, a unit cell is no longer a sufficient boundary for exploration. A potential approach would be to tile multiple unit cells, and apply the aforementioned small box techniques to a larger supercell. This however faces challenges of computational expense, and increased complexity of phase space. Many of the preceding minimiza-

tion procedures are extraordinarily sensitive to the initial conditions, and increasing the complexity of the available model space further obfuscates the route to a global minimum. This can be conceptualized by imagining a fictional energy landscape as shown in Figure 1.22. Increasing the size of model space effectively creates more ripples in the energy landscape, creating more opportunities for a minimization procedure to get trapped in a local minimum.

Minimization in a complex landscape is a task well suited for Monte Carlo methods. Naïve data driven RMC approaches this minimization with a prejudice for randomness. Nonetheless, when paired with effective restraints, this randomness can be curtailed by chemical knowledge. Even the WWW method depends on an internal MC minimization using empirical energetics. Although this still requires significant user input and iteration for solving disordered structures, the combination of methodology becomes crucial. The chemist using NMR to demonstrate a single chemical environment, then using that knowledge to program an RMC protocol that restrains local similarity (INVERT), is an inherently Bayesian approach to stochastic structural solution. It does not require recalculating the NMR spectra at each iteration, only taking prior probabilistic information regarding local environment from the initial experiment. Similar attempts at simplification by inference are made with other local probes such as EXAFS, XANES, and IR spectroscopy, or when using DFT to inform empirical potentials, lessening the computational load inside a minimization procedure.

With respect to the diffraction pattern alone, we encounter a growing set of degenerate structural solutions: multiple structures equivalently match the scattering. A continuing area of exploration, claims of uniqueness require evidence from multiple sources, especially in the case of disorder. Again we see the need for compounding techniques in probing structural information. With such requirements and dependence on experienced scientific input, can the structural solution of disordered materials ever become as routine as their ordered counterparts?

1.4 Structural modeling: an inverse problem

In an abstract sense, the task of the crystallographer amounts to the construction and solution of an inverse problem. Mathematically, the forward problem of predicting observables O given a set of model parameters, M, can be written:

$$\mathbf{M} \mapsto O = \mathbf{O}(M) \tag{1.49}$$

where M is a point in the space $\mathbf{M}$, and O is in $\mathbf{O}$. In most physical cases a well defined function $\mathbf{O}(M)$ exists, and the forward problem is resolved by an evaluation of this function. For instance, given the initial velocity of a particle subject to known forces, it is possible to solve the equations of motion for the particle's trajectory, or from a set of atomic coordinates and potentials, properties such as stress/strain parameters, thermal expansion, and dielectric susceptibility can be calculated. The inverse problem is given by the opposite: the objective is to generate a a model M from the observables O. Unless there is a well defined function $\mathbf{O}^{-1}(M)$, the solution of the inverse problem is significantly more challenging.

In some cases, this well defined inverse is attainable due to the linearity of the mapping

$$\mathbf{O}(\alpha M_1 + \beta M_2) = \alpha \mathbf{O}(M_1) + \beta \mathbf{O}(M_2) \tag{1.50}$$

where the problem can be broken down using the language of linear algebra as a forward operator or operations matrix, $\mathbf{O}$, and vectors of the model, $\mathbf{m}$, and observables $\mathbf{o}$.

$$\mathbf{o} = \mathbf{O}\mathbf{m} \tag{1.51}$$

Here, if the matrix $\mathbf{O}$ is invertible — it seldom is — the problem is solved. However, the problem is considered ill-posed in any case where the number of measurements or observations is lower than the dimensionality of the operations matrix [320]. Even in these circumstances, the linear inverse problem is still solvable by the optimization

of a penalty function that is generally convex (*e.g.*, linear least squares). These types of problems have been dealt with in the advancement of geophysics [321], signal processing [322], and machine learning [323].

Of further complexity is the frequent case of a non-linear mapping between the model and the observables, as is the case with scattering data [324]. In this instance one could consider exploring all of the space $\mathbf{M}$ and evaluating $\mathbf{O}(M)$ at each point, and stopping when a sufficiently correct match for O is found. This fails in the case of degeneracy, and becomes an intractable approach in systems of even moderate dimensionality [321]. Non-linear inverse problems appear across disciplines from heat transfer [325] to medicine [326], with applications of tomography being the canonical example. I have detailed some of the approaches for efficiently exploring phase space; however, many of these methods still struggle to handle degenerate solutions that produce the same observables [327].

The challenge of degeneracy in inverse problems was addressed by Patterson as early as 1939 [328]. As we have seen, X-ray scattering as an observable contains information about the pairwise correlations between atoms. Patterson keenly noted that X-ray analysis of crystal structures was inherently ambiguous, in that two non-equivalent structures (models which could not be superimposed on each other) could produce the same diffraction data [329]. By producing the same vector set of distances, he defined these sets of points as homometric [Fig. 1.23]. This is precisely the same challenge presented by Welberry's analysis of three-body correlations [Fig. 1.6]. It is this multiplicity of available solutions that create an additional challenge for inverse problems, as inaccurate models can produce seemingly accurate mappings of $\mathbf{O}(M)$.

Ambiguity of the observation is not the only source of degeneracy in the ground state. In fact, a multitude of accessible states is given in statistical mechanics by Boltzmann's equation:

$$S = k_{\mathrm{B}} \ln W \tag{1.52}$$

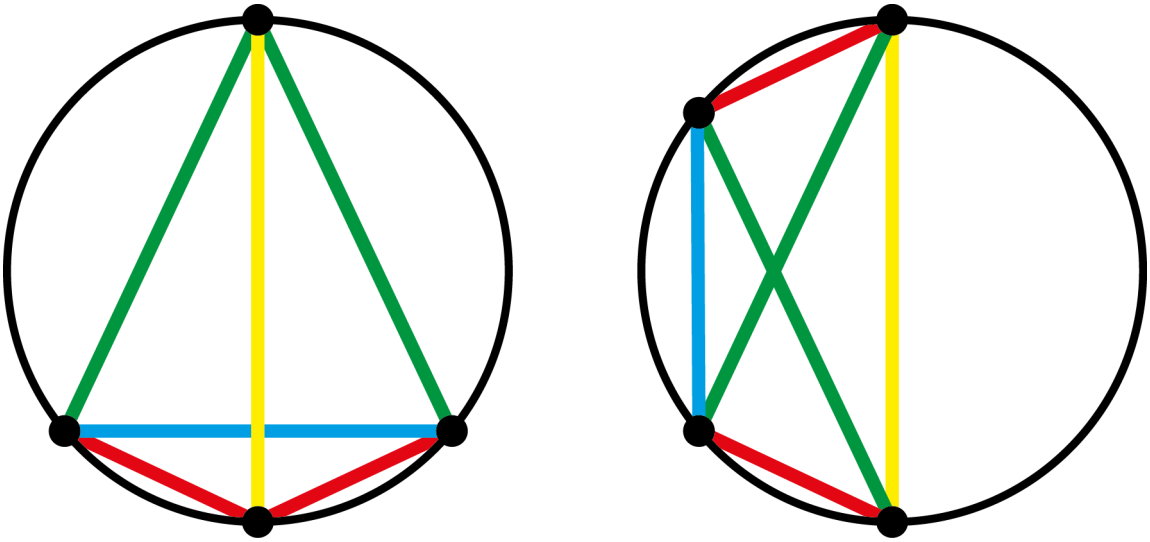


Figure 1.23: Homometric system of points. The homometric systems are not equivalent, in that the set of points on the left cannot be translated or rotated to match the set of points on the right. They are homometric because the pairwise distances between points in both the left and the right are the identical. Equivalent distances are drawn in the same color. This concept was adapted from Ref. [329], which contains a multitude of homometric examples.

describing the thermodynamic entropy as a function of the partition function. The partition function can be thought of as the number of accessible microstates to a system's given macroscopic state. This is directly related to our discussion of subextensive entropy in water ice [Section 1.2.4]. The original conception of Boltzmann's equation dealt with an equal probability of microstates, and is colloquially considered as the measure of disorder in a system. A similar concept exists in information theory called the Shannon entropy H, that measures the unpredictability of the macroscopic state or average information content [330]:

$$H = -\sum_i p_i \log p_i, \tag{1.53}$$

where p_i is the case probability. Originally conceptualized for communication systems, the Shannon entropy is related to and named in reference to the thermodynamic Gibbs entropy.

The Shannon entropy is maximized in the case where the probability distribution is uniform: if each possibility has an equal probability of existing, then it is most

difficult to predict the exact state. On the contrary, a sharp distribution such as a Dirac delta function would offer certainty and a minimized entropy. In modeling procedures, creating effective restraints can be viewed as sharpening the probability distribution and lowering the Shannon entropy. When we talk about experimenting with physical systems, there are two entropic forces at play. Firstly — and most well known to the chemist — is the entropy that comes from the energetics of ground states that are thermodynamically accessible. Secondly, is the kind of informational entropy associated with the experiment itself. In tandem these compound to increase the complexity of the inverse problem, with higher entropy offering an indication of a problem's insolubility.

The sources of this entropy are varied, and include both uncertainty in the model and the nature of the measurement. Naturally, experimental precision is the first and most usual source of uncertainty stemming from a measurement. The number of observables increases the certainty of the model — this can take the form of resolution in a multidimensional observable (*e.g.*, a spectrum), or multiple types of observables — as does the quality of their mappings to the models, $\mathbf{O}(M)$. With respect to model space $\mathbf{M}$, this is named phase space in statistical mechanics and contains all possible values of momentum and position for all objects in the system. This is a large multidimensional space that will grow with the increasing object degrees of freedom. A large model space will increase the complexity of the inverse problem and the information entropy; however, can be mitigated by a nonuniform probability of occupation. Disorder in a crystal has the impact of increasing the degeneracy of the energetic ground states, and removes the information content of the periodic lattice. Having seen the relation of disorder to our primary observable, the dearth of information content in this inverse problem is apparent. How we go about solving this inverse problem in a robust fashion, for a variety of disordered materials is the task of this thesis.

1.5 Aims and structure of this thesis

The primary objective of this thesis is to gain a structural understanding of disordered materials. The measured structure factor of disordered materials is inherently information poor, often only available as a one-dimensional function, and in many cases an insufficient characterization method in isolation. I present an approach to solving the inverse problem of transforming the one-dimensional total scattering pattern into a meaningful three-dimensional structure titled *informed modeling*. At its root, informed modeling builds on the historical approaches discussed in this chapter by supplementing the total scattering and PDF with some form of prior information or restraints. In practice, this is a data-driven approach that includes some aspects of model building, with constant reference to experimental and/or first principles information.

To examine the boundaries at which this approach is sensible, three structures of increasing degrees of disorder are presented in this thesis: $ZrW_2O_8 \cdot H_2O$, amorphous palladium imidazolate, and solution phase proteins. Although these subjects may seem disparate, they arise naturally as systems across our presented spectrum of disorder [Fig. 1.2]. Each study will be depend on a structure factor and PDF that is more information poor than the last; however, each subsequent system will also possess a more discriminating set of accessible prior information.

Chapter 2 will describe the methods central to the thesis and relevant to each study. Some of these methods will be expanded upon in the subsequent relevant chapters. As diffraction theory has already been discussed as a basis for the introduction, only experimental scattering methods will be presented in Chapter 2. This will be followed by a brief summary of the first principles approaches implemented in the study of crystalline materials.* Next, empirical approaches to calculating ground state structures will be described, with particular attention paid to the potentials

*Although all experimental scattering data and *ab initio* calculations utilized in this thesis were obtained through collaboration, it is important to understand the methods that support the developed models.

implemented in this thesis. Finally, a detailed discussion of Monte Carlo methods will be presented, as these types of methods are central to many of the informed modeling procedures developed herein.

Our first material of interest is presented in Chapter 3. The framework material ZrW_2O_8, well known for its anomalous thermal expansion behavior, also boasts an atypical contraction upon hydration. This volume reduction is accompanied by disorder observed in the powder diffraction data, which has to date made understanding the structure of this unique hydrate rather challenging. Here I use DFT calculations to aid in a model building procedure that propagates the structural changes observed in the small-box calculation throughout a larger supercell. The prior information is implemented as a set of local ice rules, and combined with total scattering data to validate the structural model, which is subsequently explained in the context of lattice dynamics.

Chapter 4 continues exploring correlated disorder by considering a more crystallographically challenging material. Palladium imidazolate, $Pd(im)_2$, is functionally interesting as an amorphous material with a square-planar metal center. Drawing on the approaches for modeling other amorphous frameworks by introducing topological disorder while fixing local geometry, this study creates a model using a sequential MC/RMC method. As amorphous materials contain fewer features in their diffraction data, more stringent prior assumptions are needed to develop a sensible model that is consistent with the data. From spectroscopic data, the local geometry of the metal center is fixed and the first known square-planar continuous random networks are generated from a modified WWW simulation. These provide a foundation for a $Pd(im)_2$ structure, which is refined against real and reciprocal space scattering data using RMC.

Chapter 5 develops the concept of informed modeling as an explicitly Bayesian technique. This chapter presents a case study for systems that possess robust and continuous prior statistics. Contrasted with the preceding examples of prior infor-

mation (local configurations as informed by DFT or spectroscopy), conformational statistics are available for protein structures in solution, and provide a means for converting the relatively sparse details contained in the protein PDF into a structure. An RMC type program for working with this data is developed and implemented in a proof of concept study, with a particular focus on the discriminating power of the method.

Finally, I conclude with a brief summary of the results presented in the thesis, followed by an outlook for the future. In the conclusion, I repeat the central question of this thesis: In what scenarios can we meaningfully invert the diffraction pattern into a structural model? I then discuss the boundaries of total scattering, and how and when informed modeling can be used to solve this problem though explicit or implicit use of prior information.

Chapter 2

Methods

2.1 Experimental X-ray scattering

The primary technique employed in solving atomic structure is based on diffraction. We have already seen the theoretical underpinnings of diffraction and total scattering [Section 1.1]. The results of this thesis depend on angular-dispersive X-ray total scattering measurements performed using in-house tube sources and synchrotron sources of radiation. Powder diffraction is the technique of choice when single crystals are inaccessible. In the experimental set up, the sample contains crystallites that are randomly oriented, resulting in a diffraction pattern that is an average of that from all orientations. For this reason, amorphous materials (which are isotropic in topology) behave as powder samples with respect to orientational averaging.

The geometries for the diffraction experiments used herein depend on the incident monochromatic beam of radiation, the sample stage, and the detector. For in-house experiments, the incident beam is generated in vacuum tubes in which electrons are fired with high energy into target materials to produce characteristic wavelengths of radiation. The wavelength of the incident beam is an important consideration, because it determines the accessible Q range, and some materials will fluoresce or scatter inelastically at certain wavelengths. Synchrotron radiation sources, on the other hand, generate X-rays from a particle accelerator that are focused, collimated, and monochromated along a beamline. The brightness (a measure of photon flux per area and solid angle of the beam) of radiation produced by synchrotron sources is significantly larger than that from a tube (up to 10^9 times), and wavelength selectivity allows for a broader range of reciprocal space vectors to be measured [331].

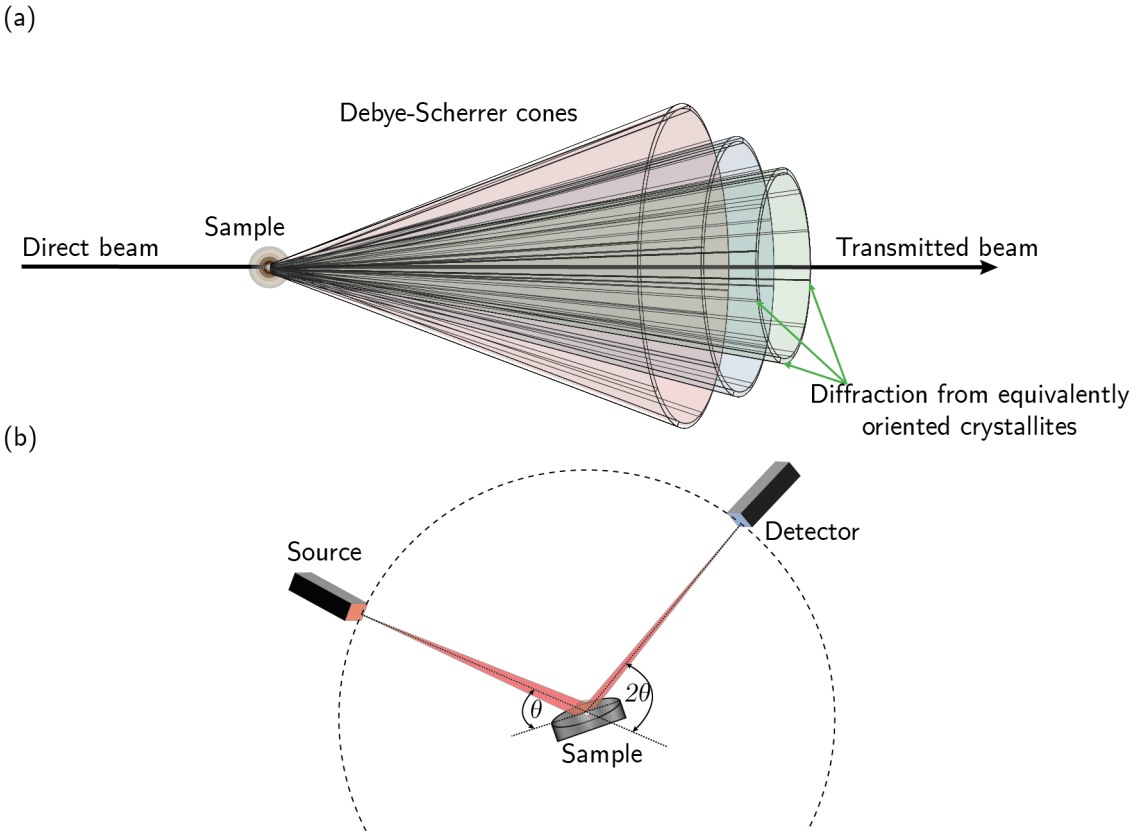


Figure 2.1: Powder averaging and experimental powder X-ray diffraction set up. (a) The orientational isotropy of powder diffraction samples create Debye-Scherrer rings of diffraction intensity, producing a 1D signal upon integration. (b) The 'Bragg-Brentano' geometry of an X-ray diffraction instrument, showing the position of the X-ray source, the sample, and the detector. Included in this set-up, but excluded from the diagram are associated slits and Sollers within the focusing ring, and any monochromator prior to the detector.

Orientational averaging in a powder X-ray diffraction experiment, results in cones of diffraction [Fig. 2.1] that can be detected using a point, strip, or 2D detector. A flat plate or 2D detector will detect the complete circumference of the diffracted cone as a ring (termed Debye-Scherrer rings). Integration of the rings renders the familiar 1D pattern. This is the simplest experimental set up, with the sample and detector placed in the direct beam, and is commonly used when carrying out total scattering studies at synchrotron sources. Strip or point detectors are generally used for in-house experiments. Here the detector sweeps an arc in 2θ around the sample. This requires precise control over the positions of the three main components of the experiment. While some experiments will keep the tube stationary and rotate the

sample and detector (Debye-Scherrer method), the instrument used in this thesis uses a Bragg-Brentano geometry in which all three components can be moved. The divergence of the incident and scattered beam can be reduced by slits or Soller slits (for sideways divergence). This reduction of instrument broadening comes at the expense of measured intensity, and thus can be adjusted for sample requirements.

Any experimental scattering data used in this work was produced through collaboration, and not by the author.

2.1.1 X-ray powder diffraction

Conventional X-ray powder diffraction experiments were performed in-house using a PANalytical Empyrean diffractometer equipped with a Ge monochromator, measuring 2θ=10–70° with a Bragg-Brentano geometry and incident Cu Kα_1 radiation (λ=1.540 56 Å). For reciprocal space refinements, TOPAS Academic v6 [137] was used to carry out non-linear least-squares Pawley and Rietveld refinements.

2.1.2 X-ray total scattering

For pair distribution function analysis, X-ray total scattering measurements were performed. We have already discussed the theory of total scattering and the pair distribution function. The experimental realization of this is complicated by the weak intensity of the diffuse scattering that is often discarded as background. In this case, accurate measurement of the background is necessary to subtract this from the measured data. Furthermore, equation (1.32) contains an integral over infinite Q. The experimental limitation of a finite $Q_{\max}$ introduces artifacts in the resolution and truncation effect of the Fourier transform. The resolution in real space is defined by a relation to $Q_{\max}$ [332]:

$$\Delta r = \frac{3.791}{Q_{\max}}. \tag{2.1}$$

This effectively relates the PDF precision — and ability to resolve overlapping peaks — to the wavelength of the incident beam by equation (1.10). Regular laboratory data (collected with Cu Kα radiation) only has a $Q_{\max} = 8.15\,\text{Å}^{-1}$, resulting in a real space resolution too large for practical purposes. Modern laboratory diffractometers can be equipped with Ag or Mo tubes with a $Q_{\max}$ of 22.48 and $17.72\,\text{Å}^{-1}$, respectively. Synchrotron radiation sources can offer significantly shorter wavelength X-rays; however, intensities fall off at high Q due to the size of the electron distribution and data above $25\,\text{Å}^{-1}$ are seldom used. Likewise, the Q-space resolution will impact the maximum real space values in the PDF. This is a lesser concern, as the PDF is primarily a tool for accessing local correlations.

For in-house PDFs, a PANalytical Empyrean diffractometer equipped with a Ag X-ray tube with Kα_1 radiation (λ=0.5594 Å) and a Galipix detector with a focusing mirror to remove the Kβ radiation was used. Synchrotron measurements were collected using the I12 instrument at Diamond Light Source, using high-energy (85 eV) radiation, Si (111) cryo-cooled double crystal bent Laue monochromator, and a large 2D area diffraction detector (Thales Pixium RF4343) [333].

GUDRUN was used to add background correction and data normalization to total scattering data and for PDF generation [334]. To refine against total scattering data, two other programs that have already been described in principle are utilized: RMCProfile for big box reverse Monte Carlo simulations that refine against both reciprocal and real space data [335]; and PDFGui for small box refinements against the PDF using a least-squares technique [279].

In this work, RMCProfile was used to fit against X-ray total scattering data. These refinements are fed a starting structure that contains the atomic coordinates and a set of geometric restraints. These restraints come in the form of two- and three-body harmonic potentials, specified by the bonding in the starting configuration. During the refinement, the bonding was not updated, and only the atomic positions were changed. Thus the available parameters in the refinement are the

force constants of the harmonic potentials, the temperature, and the weightings of the real and reciprocal space data.

The program PDFGUI was used to refine broadening factors and lattice parameters against the PDF for ZrW_2O_8. Although PDFGUI is capable of refining atomic positions and ADPs during a fit, this functionality was not used to limit the number of available degrees of freedom. In this context, the program is used for structure validation more than structure prediction. The program accepts `.cif` files for structural inputs. Refinements proceeded iteratively by adding more parameters to the refinement stepwise: first the scale factor is refined until convergence, followed by this and the δ_1 value, then the lattice parameters were added to consideration. These two programs are stand-alone, and the specific details of their functionalities will be left to their respective documentation.

2.2 DFT

Ab initio techniques are central to a realistic model-based structural solution. The objective of quantum mechanical methods is to solve the N-body Schrödinger equation:

$$\hat{H}\Psi(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N) = E\Psi(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N), \tag{2.2}$$

where $\hat{H}$ is the linear Hamiltonian operator, E is a constant energy, and Ψ is the wavefunction. Hartee-Fock theory neglects electron correlation in multi-electron systems due to a mean-field approximation, and depends on an ansatz for the structure of the wavefunction. In practice, calculations using this level of theory become intractable when considering extended solids.

The application of Density Functional Theory (DFT) circumvents the search for the exact N-particle wavefunction, and transforms the computationally challenging many-electron problem into a less computationally challenging problem of noninteracting electrons moving in a self consistent field. This is managed by instead casting

the Hamiltonian as a functional of the electron density:

$$\rho(\mathbf{r}) = \int \cdots \int \left|\Psi(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N)\right|^2 \mathrm{d}\mathbf{r}_2 \ldots \mathrm{d}\mathbf{r}_N. \tag{2.3}$$

The Hohenberg-Kohn theorems state that (1) the total energy is a unique functional of the density (degeneracy is not a problem) and (2) the ground state energy can be determined variationally, thus the exact electron density will have the lowest energy [336]. We again have a minimization problem.

Assuming a stationary nucleus (Born-Oppenheimer approximation), the electronic Hamiltonian can be separated into relevant kinetic and potential energy terms:

$$\hat{H}_e = \hat{T}_e + \hat{V}_{ee} + \hat{V}_{en}. \tag{2.4}$$

In the above, $\hat{T}_e$ is the electron kinetic energy operator, $\hat{V}_{en}$ is the potential energy due to electron-nucleus interactions plus any external potentials, and $\hat{V}_{ee}$ is the electron-electron repulsion and exchange potential. In this notation $\hat{T}_e$ is an operator, and T_e is the value of a functional. Considering a molecule with N electrons $(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N)$ and M nuclei at $\mathbf{R}_1, \mathbf{R}_2, \ldots, \mathbf{R}_M$ with nuclear charge Z, the external potential energy is trivial to calculate, as in the case of electron-nucleus interactions alone:

$$\hat{V}_{en} = -\sum_{i=1}^{N}\sum_{I=1}^{M} \frac{Z_I}{\mathbf{r}_i - \mathbf{R}_i} = \sum_{i=1}^{N} V_{\text{ext}}(\mathbf{r}_i, \mathbf{R}). \tag{2.5}$$

The kinetic and electron-electron functionals are more difficult to ascertain. The kinetic energy operator is given by the Laplacian,

$$\hat{T}_e = -\frac{1}{2}\sum_{i=1}^{N} \nabla_i^2. \tag{2.6}$$

The Kohn-Sham formulation describes a fictitious system of N non-interacting electrons in a self-consistent field, in a way that is exact for the ground-state energy and density [337]. This creates an approximation at the kinetic energy from a single

determinant wavefunction using N Kohn-Sham 'orbitals', ϕ:

$$T = \hat{T}[\rho] = -\frac{1}{2}\sum_{i=1}^{N} \langle \phi_i | \nabla^2 | \phi_i \rangle \,. \tag{2.7}$$

This circumvents the many-body problem by giving us a set of N independent particle equations that can be solved numerically. These orbitals are not to be taken as real orbitals, often given as plane waves or Gaussians; they are a means to calculating the total energy and charge density. The ground state density is then given by

$$\rho(\mathbf{r}) = \sum_{i=1}^{N} |\phi_i|^2 \,. \tag{2.8}$$

The electronic interaction is further split into the trivial Coulomb interaction $\hat{V}_C$, and the exchange-correlation functional $\hat{E}_{xc}$, where

$$\hat{V}_C = \frac{1}{2}\int\int \frac{\rho(\mathbf{r}_1)\rho(\mathbf{r}_2)}{|\mathbf{r}_1 - \mathbf{r}_2|}\, \mathrm{d}\mathbf{r}_1\, \mathrm{d}\mathbf{r}_2. \tag{2.9}$$

The exchange-correlation treatment is the error associated with using a non-interacting kinetic energy (Kohn-Sham) and in treating the electron-electron interaction classically (Coulomb). Quantum mechanical exchange stems from the electron's antisymmetric wavefunction leading to the Pauli exclusion principle. The correlations vary between material types, as electrons are not linearly spread over a material, but interact *via* electronic repulsions, dispersive forces, polar interactions, and hydrogen bonding.

The most challenging part of this approach is in choosing a form for the exchange-correlation functional. With that choice made, the variational principle can be applied with a minimization to calculate the ground state electron density. From this, we can calculate the ground state energy and meaningful ground state observables. These calculations were readily achieved using the packages CASTEP [338] and CRYSTAL14 [339].

There are many different flavors of functionals available for us in DFT simulations, resulting in a 'Jacobs ladder' of approximations [209, 210]. The choice of a functional depends on their results with respect to experimental data for the same or like systems. A number of functionals were used during the DFT work contained in this thesis. As such, it is worth detailing their operating principles.

The local density approximation (LDA) is one of the oldest and most simple functionals, and invaluable as a first approximation due to its low computational cost [336]. The LDA functional models use a homogeneous electron gas to calculate exchange-correlation energy. That is, they depend only on the local electron density in space. Although this seems like an egregious assumption, the LDA has been a fruitful approximation — calculating structure, vibrational frequencies, elastic properties, and phase stability of key systems — and underpins some of the more advanced functionals. Nevertheless, the LDA tends to lead to overbound structures and poorly estimated bond energies, due to an unrepresentative spherically symmetric exchange hole.

On the next rung of Jacobs ladder are generalized gradient approximations (GGAs). This approach uses both the density and the gradient of the density at each point in calculating exchange-correlation. This level of theory accommodates for some of the inadequacies of LDA. There is an abundance of unique forms for GGAs, including some advancements which also depend on the Kohn-Sham kinetic energy density (called *meta*-GGAs) [210].

For further accuracy, there exist hybrid functionals that involve some mixing of exact exchange as calculated from the Hartree-Fock method. This is naturally more computationally expensive, because exact exchange depends not just on the electron density but also on the density matrix. Hybrid functionals choose an empirical fraction of exact exchange to mix in because of fortuitous error cancellation [340]. These flavors (both GGA and hybrid functionals) have varying degrees of empirical components that can result in functionals over-fitting to the type of systems on which

they were parameterized. This section references the general principles associated with functional choice, whereas the exact functionals will be referenced in their respective sections.

2.3 Empirical simulations

When quantum mechanical treatment of the structure in question is either unnecessary or inaccessible, a fully empirical approach may be considered. This model based approach is achieved by considering the atomic interactions either classically, or through approximate potentials. Throughout this thesis, there will be empirical simulations that collapse the internal energy of a structure into functions that depend on point atoms with or without charge. The form of the explicit interaction between atoms is chosen based on physical insights as to the nature of the forces between the particles. The exact form of the potential will construct a unique Hamiltonian for each material treatment that is significantly easier to calculate — and thus minimize — than any *ab initio* Hamiltonian.

When deterministic minimization is appropriate, the General Utility Lattice Program (GULP) was used to develop structural solutions [185]. The program has effective implementation of all standard physical model force-fields, and is capable of minimizing a structure, calculating the internal energy, and calculating mechanical properties. The primary use of GULP in this thesis was for structural optimization using classical potentials. A two-body harmonic potential,

$$V_{\text{two}} = k_{ij} \sum_{ij} (r_{ij} - r_0)^2, \tag{2.10}$$

casts pairs of atoms i and j as harmonic oscillators with an equilibrium separation r_0. This is colloquially known as a bond length potential, as it can be parameterized for two atoms with only a length and oscillator strength (bond strength), k_{ij}. Another

potential that lends itself to molecular geometry is the three-body potential:

$$V_{\text{three}} = k_i \sum_i (\theta_i - \theta_0)^2. \tag{2.11}$$

This is a harmonic oscillator that depends on an equilibrium angle across three atoms, and is amenable to expected polyhedral geometry. The three-body harmonic potential can also be cast in terms of the cosine of the equilibrium angle:

$$V_{\cos} = k_i \sum_i (\cos\theta_i - \cos\theta_0)^2. \tag{2.12}$$

This formulation has its benefits in rapid calculation from the scalar product, but is inadequate when the sign of an angle is relevant (as is the case for torsion angles). With respect to non-bonded atoms, a diverse set of interatomic potentials are available. Pairwise interaction forces can be modeled by the Lennard-Jones (LJ) potential,

$$V_{\text{LJ}}(r_{ij}) = 4\epsilon \left[\left(\frac{\sigma}{r_{ij}} \right)^{12} - \left(\frac{\sigma}{r_{ij}} \right)^{6} \right], \tag{2.13}$$

which has an attractive potential that decays with r^{-6} to model dispersive forces, a term to approximate the repulsion of the core electrons that decays with r^{-12}, and only two parameters. An alternative to the Lennard-Jones potential — the Buckingham potential — is commonly used for ionic solids and replaces the repulsive term with an exponential:

$$V_{\text{Buck}}(r_{ij}) = A\mathrm{e}^{-Br_{ij}} - \frac{C}{r_{ij}^6}. \tag{2.14}$$

Here, A, B, and C are empirical constants. In ionic materials or molecules with charged species, a Coulomb potential can be used to approximate the electrostatic interaction by treating atoms as point charges:

$$V_{\text{C}}(r_{ij}) = \frac{1}{4\pi\epsilon_0} \frac{Z_1 Z_2 e^2}{r_{ij}}. \tag{2.15}$$

This classical potential is only parameterized by the nuclear charge, Z, where ϵ_0 is the dielectric permittivity of free space, and e is the electron charge. A final type of potential worth noting is the discontinuous hard spheres potential that is used to enforce distance restraints:

$$V_{\text{hs}}(\mathbf{r}_1, \mathbf{r}_2) = \begin{cases} 0, & \text{if } |\mathbf{r}_1 - \mathbf{r}_2| \geq \sigma \\ \infty, & \text{if } |\mathbf{r}_1 - \mathbf{r}_2| < \sigma. \end{cases} \tag{2.16}$$

This equation displays the potential for a single hard spheres interaction between atoms at points $\mathbf{r}_1$ and $\mathbf{r}_2$, where σ is twice the sphere radius. Collectively, these various classical approaches to calculating the energy of a system of atoms from their positions can be readily combined or augmented. They are representative of the potentials implemented with GULP and used in other procedures in this thesis.

The internal workings of GULP are well described in its program documentation. In brief, the optimizer proceeds via a Newton-Raphson method, using the second order Taylor expansion (*i.e.*, depending on second derivatives), and the Broyden-Fletcher-Goldfarb-Shannon (BFGS) algorithm for updating the matrix of second derivatives (the Hessian matrix). The algorithm converges to a local minimum, at which point it outputs a structure and requested physical data. The important consequence of this algorithm is that while it is rapid and effective, it is deterministic and does not search for a global minimum. Thus, much like the DFT approach, it will be dependent on the starting configuration.

When the energy surface produced by a set of potentials is too rough, or the starting configuration is too far from the global minimum, non-deterministic minimization techniques — specifically Monte Carlo type simulations — are employed.

2.4 Monte Carlo simulations

When reviewing the historical attempts at characterizing disorder, I described the Monte Carlo simulation in general conceptual terms. As it is a central method of this

thesis — used in some capacity for each problem addressed — it will be examined here.

Monte Carlo methods are a stochastic approach for evaluating high-dimensional integrals or minimizing high-dimensional functions. A stochastic process is the movement through a series of well-defined states in a way that involves some element of randomness. In the case of an integral, this amounts to the calculation of an expectation value of an observable over a many body system at thermal equilibrium. In the case of a minimization, this amounts to the finding the lowest expectation value (and accompanying microstate) at an effective zero temperature [341]. Often in the case of statistical physics, the objective is to sample a limiting distribution that is a Boltzmann distribution,

$$P_\mu = \frac{1}{Z} \mathrm{e}^{-\beta E_\mu}, \tag{2.17}$$

where the probability P_μ of microstate μ, is related to the partition function Z, the energy of the microstate E_μ, and the inverse temperature $\beta = \frac{1}{\mathrm{k}_B T}$.

A Markov chain Monte Carlo (MCMC) is one of the most commonly used approaches for this objective. A Markov chain is generated by a sequence with a set of transition probabilities, π_{ij}, that are independent of the history of the chain (*i.e.*, the next state depends only on the current state and no previous states). The transition probability, π_{ij}, is the probability of moving from state i to state j. The transition probability matrix, $\mathbf{\Pi}$ does not vary over time, such that

$$\sum_j \pi_{ij} = 1. \tag{2.18}$$

A further condition of *ergodicity* applies to MCMC, requiring that it should be possible for the Markov process to reach any state of the system from any other state, if it is run long enough. This does not require every π_{ij} to be nonzero, just that there exists a nonzero path of π_{ij}'s between every state.

Let us recast the limiting distribution of states, such that state i exists with

a probability, π_i. A sufficient — but not necessary — condition for ensuring the limiting distribution of the n-step Markov chain, with transition probabilities given by $\mathbf{\Pi}^n$, is that of *detailed balance* (or *microscopic reversibility*):

$$\pi_i \pi_{ij} = \pi_j \pi_{ji}, \tag{2.19}$$

which is more conveniently written as

$$\frac{\pi_{ij}}{\pi_{ji}} = \frac{\pi_j}{\pi_i}. \tag{2.20}$$

The Metropolis algorithm [163], later generalized by Hastings [342], is by far the most successful approach to satisfying detailed balance. In this algorithm, the transition probability is split into a proposal probability, τ_{ij}, and an acceptance probability, A_{ij}:

$$\pi_{ij} = \tau_{ij} A_{ij}. \tag{2.21}$$

To be consistent with equation (2.19), the acceptance probabilities must satisfy

$$\frac{A_{ij}}{A_{ji}} = \frac{\pi_j \tau_{ji}}{\pi_i \tau_{ij}}. \tag{2.22}$$

The Metropolis-Hastings acceptance probability that fulfills this condition is

$$A_{ij} = \text{MIN} \left\{ 1, \frac{\pi_j \tau_{ji}}{\pi_i \tau_{ij}} \right\}. \tag{2.23}$$

Therefore, the algorithm proceeds as follows. After generating an initial state i, propose a candidate state j according to the probability, τ_{ij}. Calculate the acceptance probability according to equation (2.23). Next we generate a uniform random number using a pseudo random number generator, $u \in [0, 1]$. If $u < A_{ij}$ we accept the new state j, otherwise we reject the candidate state and remain in the initial state. This process is repeated with a new proposal for an appropriate number of

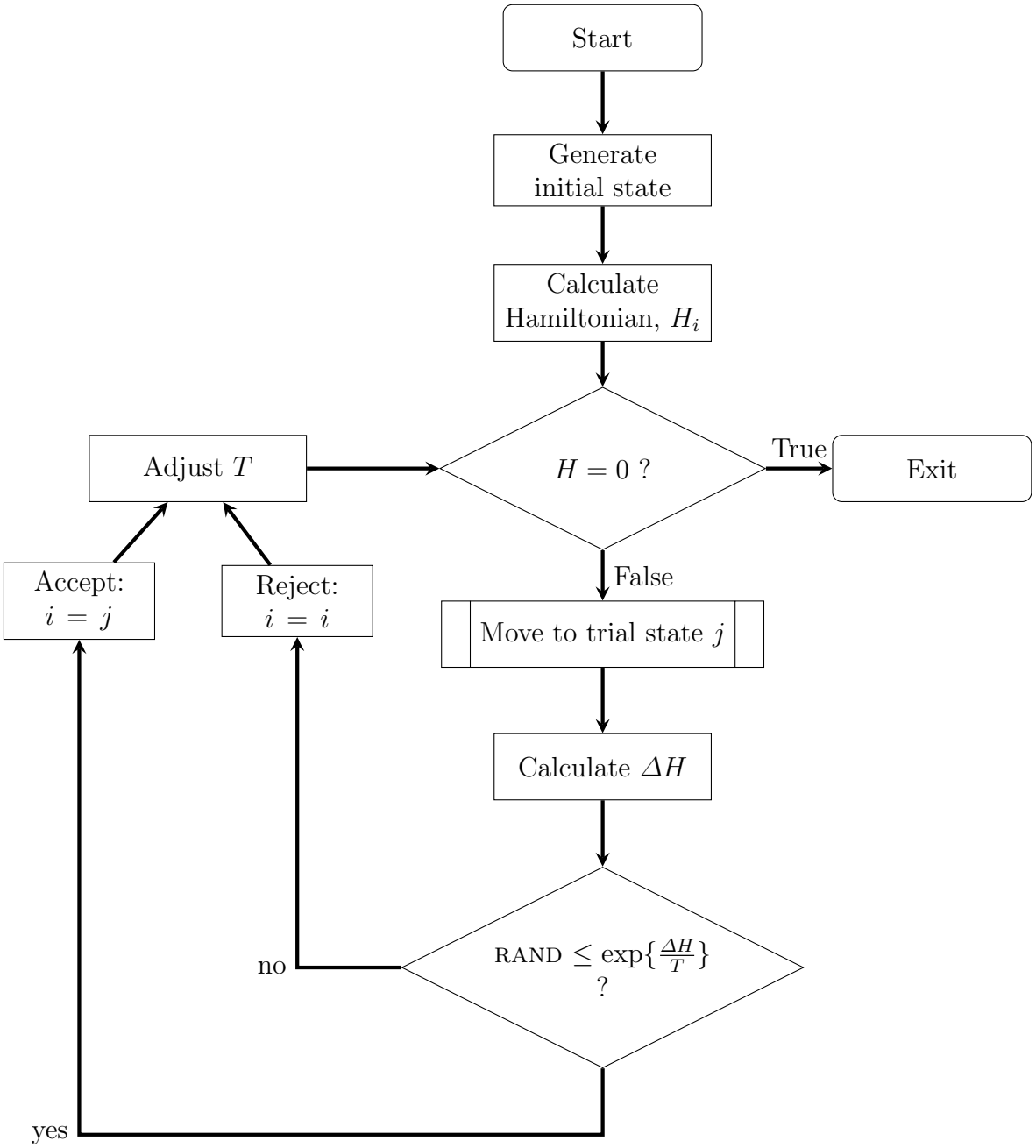


Figure 2.2: Simulated annealing flowchart. The simulated annealing algorithm is shown above using a uniform proposal probability and Hamiltonian, H. The regular MCMC algorithm is trivially similar: instead of the the exit condition searching for a global minimum $H = 0$, it would depend on a certain number of MC moves. During simulated annealing — if a zero minimum is not expected — an additional exit condition could be included to exit after a number of MC moves.

times. The entire algorithm is displayed graphically in Figure 2.2, with an additional step to adjust the acceptance rate over time.

In the practice of atomistic simulations, it is often convenient to choose an atom

at random and adjust its position in a random direction for a random bounded distance. This amounts to a uniform proposal probability, such that $\tau_{ij} = \tau_{ji}$. Often, Boltzmann statistics are chosen for the limiting distribution [Eq. (2.17)]. Thus given any arbitrary Hamiltonian H, to describe the relevant energetics of a many-body simulation, the probability ratio in (2.23) becomes a function of the difference in energy between two states:

$$\frac{\pi_j}{\pi_i} = \exp[-\beta(H_j - H_i)]. \tag{2.24}$$

This gives a computationally facile acceptance probability of

$$A_{ij} = \text{MIN}\left\{1, \mathrm{e}^{-\beta\Delta H}\right\}. \tag{2.25}$$

With this understanding of the MC algorithm, it is worth examining some terminology. The acceptance condition will always accept a proposed MC move to a more likely state. This is considered a 'good' move. In the context of statistical physics, it is a move that reduces the energy cast by the Hamiltonian. When thinking of the Hamiltonian as any function, it results in a downhill step in a search. If the proposed state is less likely than the initial, it is termed a 'bad' move. These moves are accepted with a finite probability — so long as the temperature is positive and finite — enabling the simulation to increase the energy or Hamiltonian, or make uphill steps in a search. A MC *sweep* is defined by the the number of moves it takes on average to adjust each degree of freedom, and is generally given by the number of atoms or spins in a system. This assortment of terms is used interchangeably throughout the thesis and literature depending on the system in question.

In progressing along a Markov chain of microstates, the point of equilibration becomes important. Depending on the initial state generated at the start of the algorithm, this could take some time that also depends on the temperature of the system. There are a number of methods for ensuring equilibration that depend on

tracking the value of an observable quantity, Q, over time. This quantity could be the energy or the magnetization of a sample. Once a system has reached equilibrium, it is necessary to generate independent measurements of Q for robust statistical analysis. Because the Monte Carlo method has no kinetic energy terms built into it, there is no access to the real dynamics of the configurational states. There is however a measure of Metropolis 'dynamics'. That can be derived from the *time autocorrelation function*

$$c_{QQ}(k) = \langle Q_n Q_{n+k} \rangle - \langle Q_n \rangle^2, \tag{2.26}$$

where n labels the Monte Carlo time step, and k refers to a number of intermediate steps. Thus, if MC steps separated by k intermediate steps are uncorrelated, $c_{QQ}(k) = 0$. A more commonly used measure of correlation that is more easily calculated is the *integrated correlation time*:

$$\tau_{\text{int}} = \sum_k \frac{c_{QQ}(k)}{c_{QQ}(0)}. \tag{2.27}$$

At a given temperature, the MC can be run for an initial number of moves to calculate τ_{int}, then continued for a multiple of this time to extend past equilibrium. This is done by calculating the expectation value of the observable throughout the τ_{int} calculation phase, the expectation value of the square of the observable, as well as a value similar to the square — separated by a time delay. It is most efficiently accomplished by utilizing a bidirectional linked list which can update with constant complexity instead of $\mathcal{O}(N)$ [58]. At this point, the calculation of the quantity Q should be separated by an interval of at least $2\tau_{\text{int}}$. The autocorrelation time is also useful in approximating a minimum time for equilibration during a simulation in which temperature varies.

I mentioned that MC could be used not only for generating a thermodynamic ensemble and calculating observables, but also as a minimization technique. This

can be achieved by setting the temperature to zero and only accepting good moves that lower the energy of the system. A zero-temperature MC would result in a stochastic downhill search, and is susceptible to getting trapped in a local minimum. In other words, the exploration of phase space is limited to negative changes in the Hamiltonian. Simulated annealing circumvents this problem by introducing a nonzero MC temperature and cooling schedule [343–345]. The algorithm works by emulating the physical process whereby a solid is slowly cooled to a structure with a minimum energy configuration. At high temperature, there is a broad exploration of phase space, and many bad moves are accepted. As the system cools, the simulation progresses downhill to an eventual global minimum.

In the circumstances of modeling disorder, we are often searching for a minimum energy configuration. Given a Hamiltonian as in direct MC [295], this process fits directly into the above and is amenable to calculating physical quantities for comparison with experiment. However, in the case of RMC, the Hamiltonian is replaced by a goodness of fit function.The fitting probability is given as an exponential of the cost function of each state, χ_i, that is the sum over the squared residuals

$$\chi_i^2 = \sum_n \left[\frac{O_{\text{sim}}(n) - O(n)}{\sigma(n)} \right]^2 , \tag{2.28}$$

where σ is the experimental uncertainty and O_{sim} is the predicted observable of the model state i. Thereby, the acceptance probability is implemented with a proxy temperature as

$$A_{ij} = \min \left\{ 1, \exp \left[\frac{-(\chi_j^2 - \chi_i^2)}{T} \right] \right\} . \tag{2.29}$$

Additional restraints can be added to a joint Hamiltonian that includes χ^2 and energetic terms. The MC and RMC methods are modified and used throughout much of this thesis with a modern Fortran or Python implementation. The self contained package RMCProfile [335] is also used where appropriate.

Chapter 3

Negative hydration expansion in ZrW_2O_8

At the close of Chapter 1, I discussed the implications of disordered states in scientific problems, and some of the available modeling techniques for understanding this disorder. I introduced the concept of a spaghetti phase, and how it can arise from correlated disorder. Herein, I will describe an instance of these spaghetti phases in a crystal structure through some of the mixed modeling techniques previously presented.

The most information rich problem contained in this thesis, the hydrated phase of ZrW_2O_8 and its subsequent anomalous behavior, has remained a conundrum in the scientific community. That is, ZrW_2O_8 contracts upon adsorption of water [346], but not as significantly with other adsorbates, such as ammonia [347]. A relatively unexplored problem, many questions remain: what is the mechanism of hydration in ZrW_2O_8, why is this mechanism different from other ZrW_2O_8 adsorption phases, how does the mechanism relate to the negative thermal expansion in ZrW_2O_8, and is this mechanism related to other transitions in ZrW_2O_8?

In this Chapter, I explore how compounding techniques to provide information can elicit unique structural solutions in a disordered crystal. I begin by presenting the anomalous hydration behavior of ZrW_2O_8 against the backdrop of the known crystalline and amorphous phases of the material. I elaborate on negative thermal expansion (the most historically notable property of ZrW_2O_8) and its relationship to lattice dynamics and structure. To understand the mechanism for this hydration, I turn to scattering techniques — in both reciprocal and real space — to inform and validate any modeling procedures. Density functional theory calculations are used

in a single unit cell to develop local rules, while supercell lattice energy minimization procedures are used to propagate these local rules through a larger box. I consider the dynamics of the phase transition in the context of the phonon dispersion of ZrW_2O_8. Together, these analyses present an understanding of correlated structural disorder in $ZrW_2O_8{\cdot}H_2O$. Finally, I conclude with a discussion of both the particular material relationships and a broader discussion on how these modeling procedures can be applied elsewhere.

I would like to acknowledge Dr. Fabien Trousselet, Dr. F-X Coudert, and Mia Baise for their DFT calculations, Mia Baise who completed the powder diffraction experiments, and Dr. Nicholas Funnell for assisting with the total scattering and data normalization.

3.1 Background

Zirconium tungstate, ZrW_2O_8, is perhaps best known as a canonical negative thermal expansion (NTE) material. That is, the volume of its unit cell *decreases* by 2 % on heating from 0.3 K up to its decomposition temperature of 1050 K [348]. First synthesized in 1959 *via* solid state reactions [349], and later characterized by X-ray crystallography in 1967 [350], α-ZrW_2O_8 was initially examined for sorption characteristics for selective ion exchange and column chromatography applications [351, 352]. This early utility alludes to the flexibility of the framework material. More recently, ZrW_2O_8 has been found to contract upon adsorption of water [346], a phenomenon serving as the impetuous for this work.

3.1.1 Negative hydration expansion

The hydration of ZrW_2O_8 with stoichiometrically one water molecule per formula unit occurs with a decrease in volume of ∼10 % [Fig. 3.1]. This 'inverse sponge' behavior is counterintuitive, in that an adsorption process creates an internal stress equivalent to a negative thermodynamic pressure, and would be expected to expand the volume. This negative hydration expansion (NHE) effect has been observed

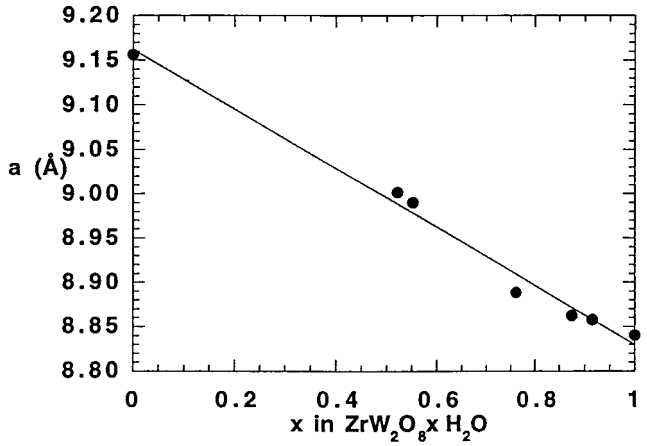


Figure 3.1: Plot of lattice parameter as a function of water content in $ZrW_2O_8{\cdot}H_2O$. Figure obtained from Ref. [346].

in anisotropic 'breathing' framweworks (namely MIL-52 [353, 354]), but is rare in cubic systems [355, 356]. Moreover, ZrW_2O_8 has no pore network in the traditional sense, whereas the pore sizes in these breathing frameworks ranges between 5 and 20 Å. This unique behavior has obvious implications for the engineering the swelling response in composite materials (*e.g.*, cements or dental implants), for which isotropy is essential in minimizing internal strain. Given the importance of shrinkage cracking — a $48 billion market in American cement alone [357] — this NHE property of ZrW_2O_8 may actually prove more practically relevant than its longer studied NTE counterpart.

Despite the potential importance of NHE, little is understood of its microscopic mechanism in ZrW_2O_8. Previous X-ray and neutron powder diffraction studies of $ZrW_2O_8{\cdot}H_2O$ demonstrated a preservation of the basic framework structure, an apparent increase in crystal symmetry from $P2_13$ to $Pa\bar{3}$, and a quenching of the NTE response [346, 358, 359]. A decrease in the $Zr\cdots W$ separation was observed and interpreted in terms of a correlated tilts of ZrO_6 and WO_4 polyhedra [359]. Nonetheless, a meaningful structural model of $ZrW_2O_8{\cdot}H_2O$ has yet to be developed. During a high-resolution neutron powder diffraction study, Duan and coworkers were fully capable of characterizing the disordered β-phase of ZrW_2O_8; however, they claim that severe structural disorder due to the partial water occupancy and

orientational disorder of water and WO_4 tetrahedra precluded the construction of a model for $ZrW_2O_8{\cdot}H_2O$. The absence of structural understanding — of the position and orientation of the water, and any changes in the framework — has left the community to infer that NHE is driven by strong hydrogen bonding that 'pulls' the framework together while increasing W coordination from four to five as suggested by the change in space group. The limitations of these assumptions point to a need for a more accurate structural model that encapsulates the role of disorder in the hydration mechanism.

The NHE phenomena and subsequent quenching of NTE are made more interesting by recent work in the adsorption of ammonia. This adsorption process does not produce a significant volume change, nor quench the negative thermal expansion [347]. Ammoniated ZrW_2O_8 has a slight increase in lattice parameter, and weakens the thermal expansion coefficient from $-7.8 \times 10^{-6}\,K^{-1}$ to $-2.1 \times 10^{-6}\,K^{-1}$. It is clear that NHE and NTE are not exclusive phenomena, and suggests that the unique topology of ZrW_2O_8 plays a role in the hydration driven volume collapse. As described previously, there is significant debate regarding the precise real-space mechanism of NTE. A mechanistic understanding of NHE may elicit a greater understanding of the mechanism of NTE.

3.1.2 ZrW_2O_8 structure and phases

The mechanism of negative hydration expansion is best explained with an understanding of the accessible ZrW_2O_8 phases, and the constitution of the ZrW_2O_8 unit cell. ZrW_2O_8 exhibits a number of anomalous topological transitions on increasing temperature [360], pressure [361], and hydration [346]. The structure of α-ZrW_2O_8 [Fig. 3.2] can be conceptualized with reference to the rock salt structure, where a set of ZrO_6 octahedra sit on the Na sites. These octahedra share all six corners with WO_4 tetrahedra, with a set of two tetrahedra sitting on the Cl sites as a W_2O_8 unit. Each tetrahedron shares three of its four corners with a ZrO_6 octahedron, with the fourth corner being a terminal W–O bond. This ambient phase of ZrW_2O_8

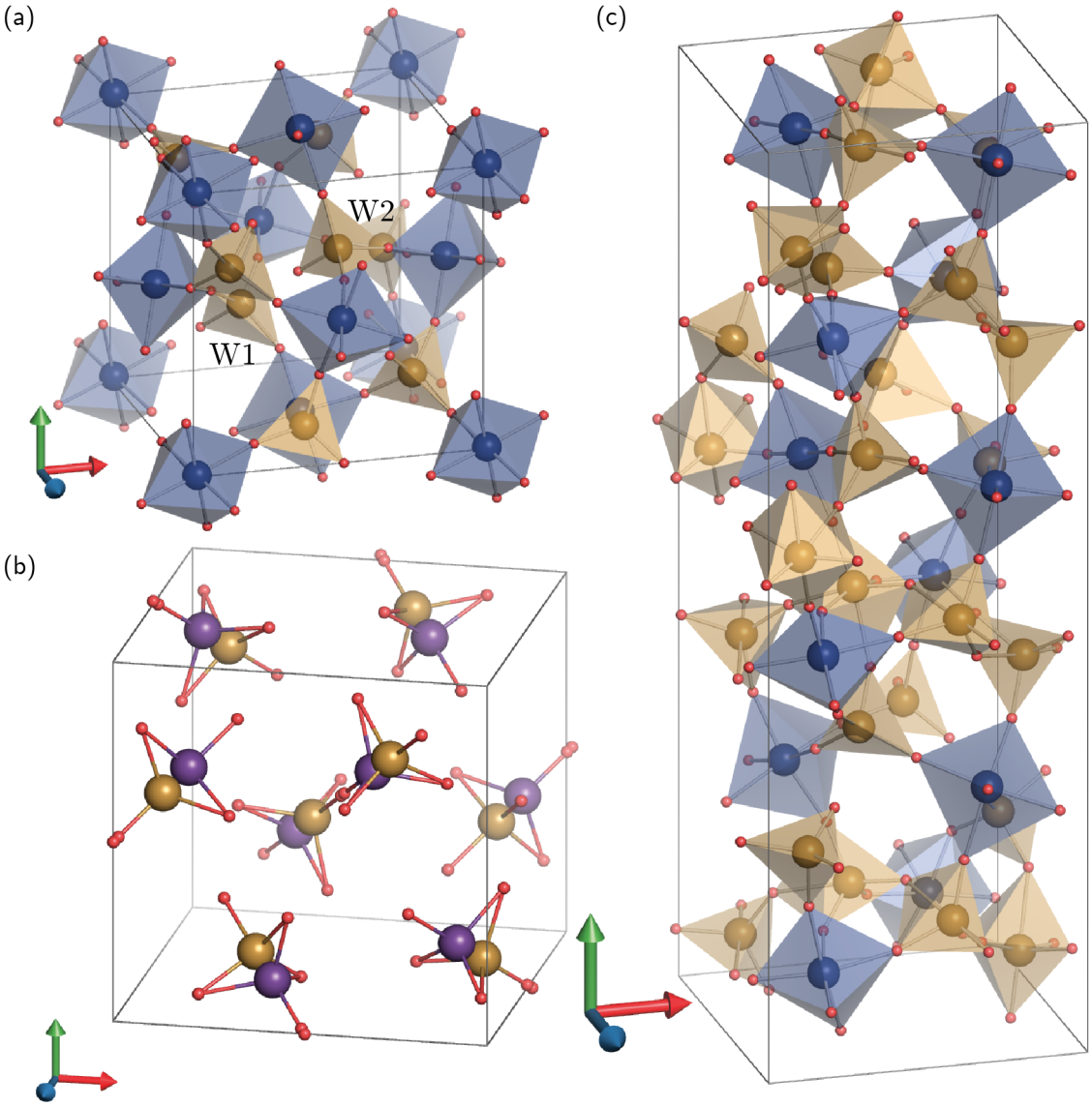


Figure 3.2: Select phases of ZrW_2O_8. (a) α-ZrW_2O_8 is shown with tungsten polyhedra in yellow, and zirconium polyhedra in blue. The naming convention of W2 and W1 along the threefold axis is shown. (b) The high temperature disordered β-phase is a result of the unconconstrained oxygen migration and indicative of the material's inherent flexibility. In the tungsten sublattice, each threefold axis is decorated either all yellow or all purple tungstens, resulting in 50% occupation of the distinct W sites and an inversion center in the average structure. (c) A similar kind of distortion as seen in the β-phase underpins the oxygen migration in the high pressure transition to the γ-phase. In both phases, the oxygen migration leads to the formation of W_2O_8 clusters.

has $P2_13$ symmetry with a lattice parameter $a = 9.15\,\text{Å}$, and the zirconium and tungsten polyhedra are only slightly distorted from regular geometries [362].

The flexibility in the tungsten unit arises from the single unconstrained oxygen atom in each tetrahedron. The W_2O_8 unit contains two crystallographically distinct

WO_4 tetrahedra that are not formally bonded to each other. The unconstrained oxygen in both of the tetrahedra lies along the same 3-fold axis. Another perception of this framework is as sets of two WO_3 units, in which one unit (W2) is bonded to an oxygen (O3) which is weakly interacting (at 2.4 Å) with the second tungsten (W1), and the final oxygen (O4) is truly non-interacting and unconstrained [Fig. 3.2].

Above 430 K, α-ZrW_2O_8 undergoes a second-order order-disorder phase transition to β-ZrW_2O_8 as a direct consequence of the unconstrained oxygen in the structure [348, 360]. During this transition, the terminal oxygen, O4, is mobile and migrates from its original site 1.7 Å from W1 to a vacant site 1.7 Å from W2. The displacement of the bridging oxygen (O3) is randomly disordered and completes the coordination of a W1 that has lost its terminal O4. These changes explain the emergence of a center of inversion, and the apparent $Pa\bar{3}$ average structure can be considered as containing two WO_4 tetrahedra that are dynamically disordered about the 3-fold axis [363]. Outside of the β-phase, other high temperature phases have been shown to exist experimentally [362].

Changes in pressure can also induce flexible phase transitions in ZrW_2O_8. At 0.2 GPa there is a transition from the α-phase to the γ-phase, accompanied by a 4.95 % volume contraction [364]. In this phase transition, there is a similar oxygen migration mechanism to the transition to the β-phase with an anisotropic tripling of the unit cell, yielding an orthorhombic $P2_12_12_1$ unit cell. The O3 atom along the 3-fold axis, deviates from the axis and breaks the symmetry while drifting towards the neighboring W1, forming W_2O_8 groups. Through a series of concerted tilts, the tungsten polyhedra and unconstrained terminal oxygens act to decrease the nonbonding W$\cdots$O distances [365].

At even higher pressures over 1.5 GPa, a topologically similar amorphous phase develops. Rotations of the W1 polyhedra away from the the threefold axis result in a O4 drift towards one of three neighboring W2 polyhedra. The weakly interacting O3 bridges the W2 and single neighboring W1. This new connectivity of –[O3–W1–

O4$\cdots$W2]– linkages produces a disordered spaghetti phase [366]. As described in Section 1.2, these spaghetti phases arise because there is no unique trajectory for a component of the connectivity. That is, the O4$\cdots$W2 connection has a choice of three equivalent W2 atoms, thus following a 'one in, two out' ice rule. This degeneracy produces an overall amorphous structure that is well demonstrated by the diffuse scattering pattern [367]. Other models have considered the PIA of ZrW_2O_8 to be a kinetically hindered decomposition into constituent metal oxides [368]; albeit, this model has been shown to be inconsistent with other experimental data [367, 369, 370].

3.1.3 Negative thermal expansion

The observation of linear NTE in ZrW_2O_8 — most notably, isotropic NTE that retains the cubic unit cell — has served as the impetus for a large body of research [371]. As the behavior is not only relevant to our interest in ZrW_2O_8, but our understanding of NHE, a brief background on NTE and its physical origins is offered here.

As expected, most common materials expand upon heating due to the anharmonicity of vibrations between chemically bonded atoms. This is illustrated in Figure 3.3. In the case of a harmonic potential, the energy increases quadratically with internuclear separation away from an equilibrium distance; however, anharmonicity produces an asymmetry in the potential well, such that the increase in energy when internuclear separation decreases (bond length decreases) is larger than the increase in energy for an equivalent increase in internuclear separation (bond length increase). Thus, as a material is heated and more vibrational energy states are populated, bonds within the material tend to increase in length. Overall, this results in a bulk positive thermal expansion (PTE) from the accumulation of many bonds increasing in length.

This first approximation yields important consequences of thermal expansion. Firstly, as bonds in a lattice have orientation, there should be a way to compare

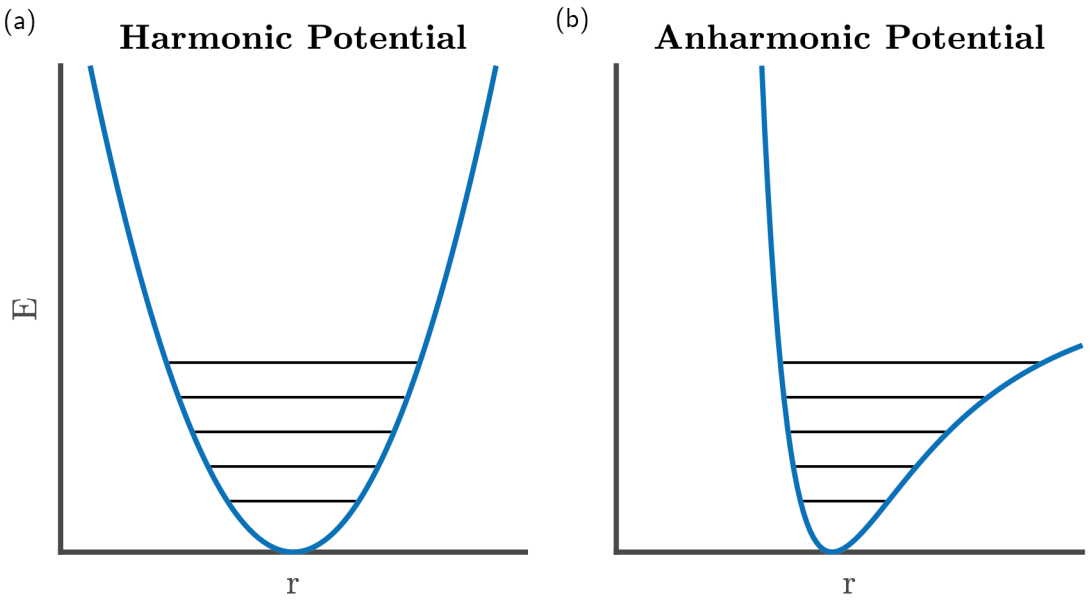


Figure 3.3: Harmonic and anharmonic potential well. Comparison of a typical (a) harmonic and (b) anharmonic potential well, with energy levels indicated by horizontal lines. The anharmonic terms of the latter invoke an asymmetry in the potential well.

directional thermal expansion and bulk thermal expansion between materials. Secondly — and most importantly — the thermal expansion behavior is dependent on the vibrational states of the system and their respective energetics. It follows that strongly bonded materials would be expected to be more rigid than weakly bonded materials due to a deeper potential well associated with the chemical bond. Later, this intuitive argument will be expanded to include lattice dynamics and phonon dispersion.

From a bulk perspective, thermal expansion behavior can be formally defined as a rank 2 tensor containing the coefficients of thermal expansion α:

$$\alpha_{ij} = \left(\frac{\partial x_{ij}}{\partial T} \right)_P , \tag{3.1}$$

where x_{ij} is the directional strain. In cases where eigenvalues of the tensor are equivalent, the behavior is isotropic and identical in each direction. Given the dependence on direction, it is convenient to define a volumetric coefficient of thermal

expansion α_V:

$$\alpha_V = \frac{1}{V}\left(\frac{\partial V}{\partial T}\right)_P. \tag{3.2}$$

For crystalline materials where the principle axes of the tensor are aligned with the crystallographic axes, α_V is given by the trace of the thermal expansion tensor:

$$\alpha_V = \alpha_{11} + \alpha_{22} + \alpha_{33}. \tag{3.3}$$

Thus for cubic materials (*i.e.* α-ZrW_2O_8), the volumetric coefficient of thermal expansion is three times as large as the thermal expansion coefficient along any crystallographic axis.

In order to determine thermal expansion coefficients, diffraction experiments can be used to refine lattice parameters at a range of temperatures. Thermal expansion parameters are fairly constant over large temperature ranges, although they can vary significantly in the region of phase transitions.

In order to quantify the complete thermal expansion behavior of a lattice, it is necessary to consider not just individual bond vibrations, but the collective lattice vibrational modes, or phonon modes. Solving for phonon displacement vectors and frequencies, is actually an eigenproblem of the form

$$\mathbf{e}(\mathbf{k}) \cdot \mathbf{\Omega}(\mathbf{k}) = \mathbf{D}(\mathbf{k}) \cdot \mathbf{e}(\mathbf{k}), \tag{3.4}$$

where $\mathbf{e}$ is the matrix of orthonormal mode eigenvectors, $\mathbf{\Omega}$ is a diagonal matrix of squared vibrational frequencies, and $\mathbf{D}$ is the Hermitian dynamical matrix containing the relevant forcefield information from the equations of motion, all at a given point in $\mathbf{k}$-space [56]. The dynamical matrix $\mathbf{D}(\mathbf{k})$ is defined at wave vector $\mathbf{k}$ in terms of the elements

$$D_{\alpha j,\beta j'}(\mathbf{k}) = \frac{1}{(m_j m_{j'})^{1/2}} \sum_{\ell'} \left[\Phi_{\alpha\beta} \begin{pmatrix} j & j' \\ 0 & \ell' \end{pmatrix} \times \exp\{\mathrm{i}\mathbf{k} \cdot [\mathbf{r}(j'\ell') - \mathbf{r}(j0)]\} \right]. \tag{3.5}$$

Here, $\binom{j}{\ell}$ denotes the atom j in unit cell ℓ with position $\mathbf{r}(j\ell)$, $\alpha, \beta \in \{x, y, z\}$, and Φ is the force constant matrix

$$\Phi_{\alpha\beta}\begin{pmatrix} j & j' \\ 0 & \ell' \end{pmatrix} = \frac{\partial^2 W}{\partial u^{\alpha}_{j\ell} \partial u^{\beta}_{j'\ell'}} \tag{3.6}$$

that relates variations in the lattice energy W to displacements u. The dynamical matrix is thus the mass-reduced Fourier transform of the force constant matrix (a matrix of second derivatives of the energy with respect to atomic displacements).

The eigenvector is composed of mass weighted displacements, such that at a single mode, ν, the eigenvector can be broken down as

$$\mathbf{e}(\mathbf{k}, \nu) = \begin{pmatrix} \sqrt{m_1} U_x(1, \mathbf{k}, \nu) \\ \sqrt{m_1} U_y(1, \mathbf{k}, \nu) \\ \sqrt{m_1} U_z(1, \mathbf{k}, \nu) \\ \sqrt{m_2} U_x(2, \mathbf{k}, \nu) \\ . \\ . \\ . \\ \sqrt{m_n} U_z(n, \mathbf{k}, \nu) \end{pmatrix}, \tag{3.7}$$

where $U_z(n, \mathbf{k}, \nu)$ is the z component of displacement for the n^{th} atom, and m_n is the atom's mass. Furthermore, the time independent displacement vector for the j^{th} atom is then given by

$$\mathbf{u}(j) = \sum_{\mathbf{k},\nu} \mathbf{U}(j, \mathbf{k}, \nu) \exp\{2\pi \mathrm{i}[\mathbf{k} \cdot \mathbf{r}(j)]\}. \tag{3.8}$$

Comparing the displacements between the harmonic force field model and the mode eigenvectors makes regular use of equations (3.7), and (3.8).

In order to identify which phonon modes are characteristic to NTE, we must calculate the Grüneisen parameters

$$\gamma_{\mathbf{k},\nu} = -\frac{V}{\omega(\mathbf{k}, \nu)} \frac{\partial \omega(\mathbf{k}, \nu)}{\partial V} = -\frac{\partial[\ln \omega(\mathbf{k}, \nu)]}{\partial(\ln V)} \tag{3.9}$$

for individual phonon modes using frequencies calculated for the same system at different volumes. The Grüneisen parameter establishes a relationship between a material's volumetric coefficient of thermal expansion α_V, heat capacity at constant volume C_V, volume V, and its isothermal compressibility K:

$$\alpha_V = \frac{\gamma C_V K}{V}. \tag{3.10}$$

The Grüneisen parameter of equation (3.10) is the overall Grüneisen parameter, and is related to the contribution of each mode Grüneisen parameter by the equation

$$\gamma = \frac{\sum_{\mathbf{k},\nu} \hat{C}_V(\mathbf{k},\nu)\gamma(\mathbf{k},\nu)}{\sum_{\mathbf{k},\nu} \hat{C}_V(\mathbf{k},\nu)}, \tag{3.11}$$

where $\hat{C}_V$ weights each mode by their contribution to the overall heat capacity C_V.

From the relation of equation (3.10), it is clear that phonon modes with positive Grüneisen parameters will contribute to PTE, whereas modes with negative Grüneisen parameters will contribute to NTE. The most obvious case of this would be to consider a metal–oxygen–metal linkage [Fig. 3.4]: a transverse vibration of the oxygen will keep the metal–oxygen distances fixed, while decreasing the metal–metal distance [372]. The volumetric thermal expansion coefficient will in turn depend on the relative weightings of the different phonon modes, which in turn depend on their energies. Modes with low energy have a larger probability of being occupied given the thermal energy of a system. Therefore, in the case where low energy phonons are dominated by modes with negative Grüneisen parameter, the overall γ may be negative and the material may experience NTE. An excellent visualization of this relationship is presented in Ref. [207].

Historically, the NTE behavior in ZrW_2O_8 has been explained by a set of low energy vibrational states involving rigid polyhedra, that is, RUMs. The assumptions of the RUM model are that the key vibrations do not involve the distortion of

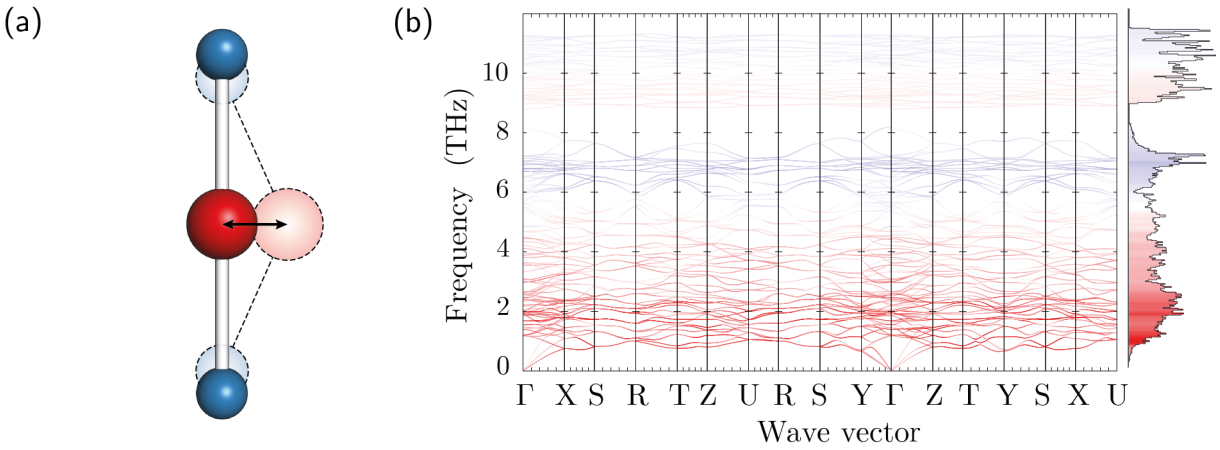


Figure 3.4: Example NTE phonon modes. (a) The transverse 'guitar string' vibrations of a metal–oxygen–metal linkage decrease the overall metal–metal separation and tend to lead to NTE (*i.e* have negative mode Grüneisen parameters). (b) Low energy phonon dispersion curves and densities of states for NTE material $Y_2W_3O_{12}$. The data is shaded according to the value of $\gamma_{\mathbf{k},\nu}$ of each mode at each wave vector: red corresponding to large negative Grüneisen parameters and blue corresponding to large positive Grüneisen parameters. Adapted from Ref. [207].

metal–ligand bond lengths or metal center coordination geometries, and that these vibrations tend to exist at lower energy than more distortive phonon modes [373]. Consequently, these RUMs can in principle contribute substantially in the summation presented in equation (3.11), and if their respective Grüneisen parameters are large and negative it indicates a potential for NTE behavior in the bulk material [374]. These energetically accessible modes are also prime candidates for soft modes that also lead to displacive phase transitions [200] and the the formation of local or extensive defects [373].

Surprisingly, the phase transition at 430 K doesn't have a drastic effect on the thermal expansion, and early calculations were able to reconcile this behavior in ZrW_2O_8 by RUMs with wave vectors across complex three-dimensional surfaces in reciprocal space [204, 205]. Evidence suggesting no significant changes to the metal–oxygen bond distances as a function of temperature affirm the RUM model, and indicate the importance of framework topology in dictating these phenomena [360]. The high density of states of low energy phonons consistent with RUMs, is further confirmed by experimental heat capacity and inelastic neutron scattering

experiments, with only slight deviations due to model assumptions [375–377].

Although the RUM model provides a strong conceptual tool for understanding NTE in ZrW_2O_8, the precise local mechanism is still contested [374]. The transverse vibrations in the Zr–O–W linkages have been attributed to concerted rotations of ZrO_6 octahedra and translations (not rotations as in the previous RUM models) of WO_4 tetrahedra [26, 297, 378–380]. Even amongst this body of work there is debate on the rigidity of certain units as obtained from local measurements such as X-ray absorption fine structure [378]. First principles calculations have described a scenario where one of the WO_4 tetrahedra is involved in concerted rotation with ZrO_6 octahedra; however, the other tetrahedron is capable of translation along the three-fold axis [381]. Moreover, there is increasing evidence that the overall negative coefficient of thermal expansion is contributed to by higher energy phonon modes [382, 383] across the Brillouin zone [384].

Finally, recent *ab initio* lattice dynamical calculations show the driving force of NTE to be a composite of phonon modes that resemble vibrations of many discrete units: polyhedra, rods, and other polyhedral subunits [27]. Each of the many modes responsible for NTE appears to involve a remarkably different combination of polyhedral translations, rotation, and distortions, such that previously proposed flexibility models map onto many phonon eigenvectors instead of neatly onto a few. This work marks the furthest step from the RUM models: considering polyhedral subunits effectively allows for polyhedral distortions, disregarding the primary assumption of the RUM model. This epitomizes how complicated the mechanism of behavior is for the unique topology of ZrW_2O_8.

In what follows, we relate this NTE back to the mechanism of hydration. We show that the hydration of ZrW_2O_8 proceeds *via* a concerted formation of one-dimensional strings of new O–W bonds. Through the combined use of X-ray total scattering experiments and quantum mechanical calculations, we develop a set of local rules and harmonic potentials that can be used in a lattice energy minimiza-

tion procedure. We relate the simulated contraction of this minimization back to the scattering information and to the calculated phonons across the BZ in order to validate the final structure and describe the mechanism in the context of NTE. This mixed modeling procedure elucidates a mechanism that is consistent with the unique topology of ZrW_2O_8 and the observed disorder in the hydrated phase, and furthermore has potential to be applied elsewhere in topologically intriguing materials.

3.2 Developing local constraints

Our starting point for developing an atomic-scale model of the hydrated phase of ZrW_2O_8 was to use quantum mechanical calculations to determine an energetically sensible $ZrW_2O_8{\cdot}H_2O$ structure. Although ZrW_2O_8 is not traditionally considered to be a porous material, within the unit cell there is only one set of possible positions with void space for adsorbate occupation. The parent structure contains a set of cavities of diameter 2.2 Å about the three-fold $\langle 111 \rangle$ axes [Fig. 3.5]. Even though it is the largest cavity, it is not large enough to fit a water molecule (whose kinetic diameter is 2.6 Å) without distorting the framework. An important inconsistency also exists in symmetry considerations: the C_{2v} point group of a water molecule is incompatible with the point group symmetry at the void site in the anhydrous crystal. Thus, there are three possible orientations for each of the four water molecules to be placed in the unit cell, with each set of particular water orientations breaking the cubic crystal symmetry. Several starting configurations were considered and used; however, the following developments focus on the unique arrangement with the least severe symmetry lowering — namely to orthorhombic $P2_12_12_1$ [Fig. 3.5].

The DFT minimization (expanded on in detail in Appendix A) was remarkably sensitive to the precise initial geometry of the water molecules. Two levels of theory were used in the DFT protocol (LDA and PBESOL0 [385, 386]) on a series of starting configurations, which led to the identification of a series of five closely-related low-energy structures, with volume smaller by 6 to 8 % than that of the original ZrW_2O_8

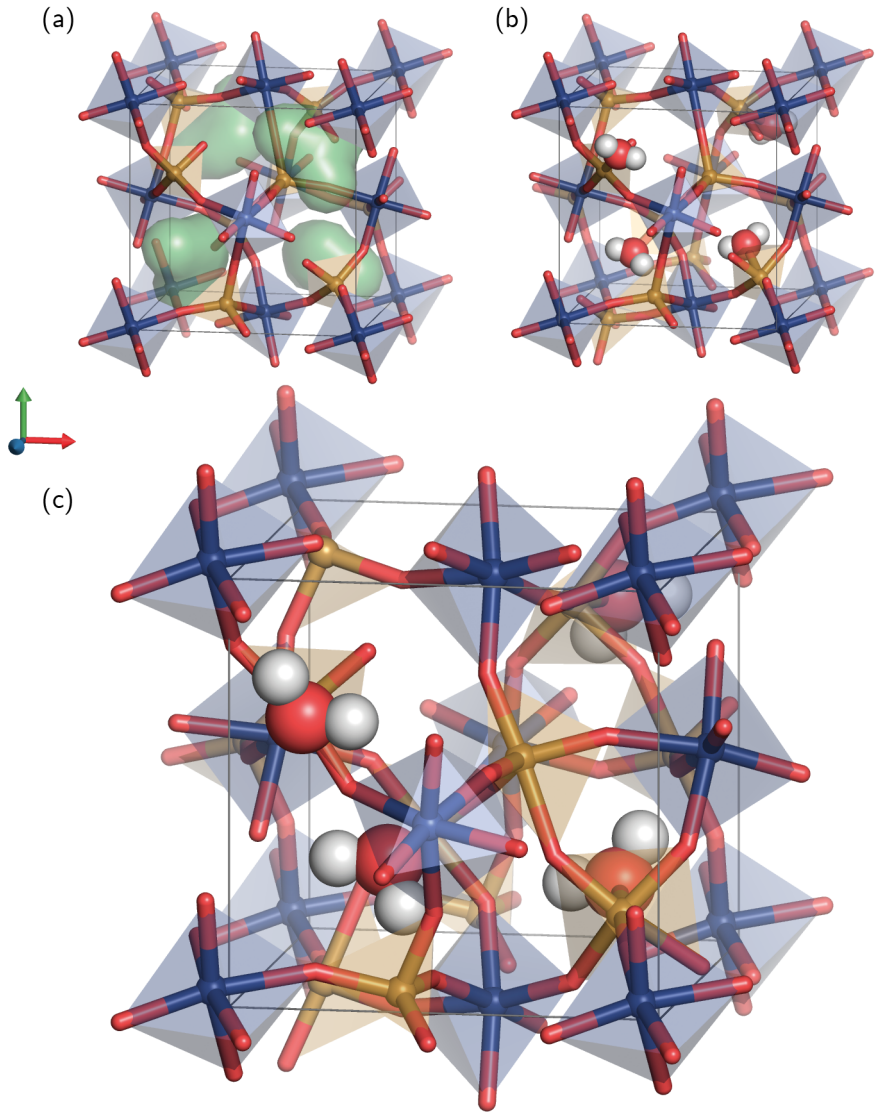


Figure 3.5: $ZrW_2O_8{\cdot}H_2O$ structures. (a) $P2_13$ structural model showing void locations (green) suitible for water placement. $P2_12_12_1$ structural model showing (b) the starting configuration and (c) the DFT relaxed configuration. The framework connections and distortions are shown with a stick representation accompanied by shaded polyhedra: Zr (blue), W (yellow), O (red). Waters are shown as space-filling.

unit cell. The enthalpy of water adsorption for these structures was found to be in the 60–90 kJ mol^{-1} range for the PBESOL0 functional, indicating a strong physisorption of the water. This is consistent with the reversible nature of hydration, demonstrated in experiment. To inform the local behavior of the empirical simulations, we focused the analysis on one of these DFT-derived structures [Fig. 3.5].

The structure of this initial model for $ZrW_2O_8{\cdot}H_2O$ reveals how hydration affects

Table 3.1: $ZrW_2O_8{\cdot}H_2O$ DFT coordinates. Fractional atomic coordinates for $ZrW_2O_8{\cdot}H_2O$ ($P2_12_12_1$) from DFT calculations. Lattice parameters $a = 8.7797$ Å and $b = 9.3019$ Å.

Atom	x	y	z
Zr	0.5310	0.9474	0.5109
W1	0.6470	0.8137	0.1298
W2	0.1140	0.9622	0.3919
O11	0.5025	0.6594	0.0876
O12	0.1005	0.6392	0.6867
O13	0.2691	0.5159	0.9126
O21	0.9689	0.2724	0.0871
O22	0.0755	0.9663	0.1960
O23	0.1950	0.0876	0.9348
O3	0.5222	0.4477	0.4563
O4	0.8008	0.6791	0.1541
O	0.3028	0.2878	0.6035
H1	0.3454	0.1888	0.6433
H2	0.1886	0.2925	0.6171

a reduction in cell volume [Fig. 3.5]. Each water molecule is chemically bound to one of the two crystallographically distinct W atoms (W1 in Figure 3.6) along a direction that is no longer parallel to a triad axis of the original unit cell.

This new coordination couples to a set of rotations and translations of both types of W-centred polyhedra that results in a new set of W–O–W connections that draws the framework in on itself: the terminal W1–O and W2–O bonds of the original structure are forced within bonding distance of nearby W2 and W1 atoms, respectively [Fig 3.6]. Although all of the W atoms are four-coordinate in ZrW_2O_8, the DFT results suggest that half the W atoms (W1) are six-coordinate and half (W2) are five-coordinate in $ZrW_2O_8{\cdot}H_2O$. We will come to show that this increase in coordination number is consistent with scattering experiments probing local structure.

The resulting connectivity from the DFT minimization of $ZrW_2O_8{\cdot}H_2O$ mimics that implicated in the mechanism for PIA described above [366]. That is, a set of ice rules can be developed to dictate the connectivity of self-avoiding pathways in a larger box of the structure. While there is only one W1 atom accessible to each

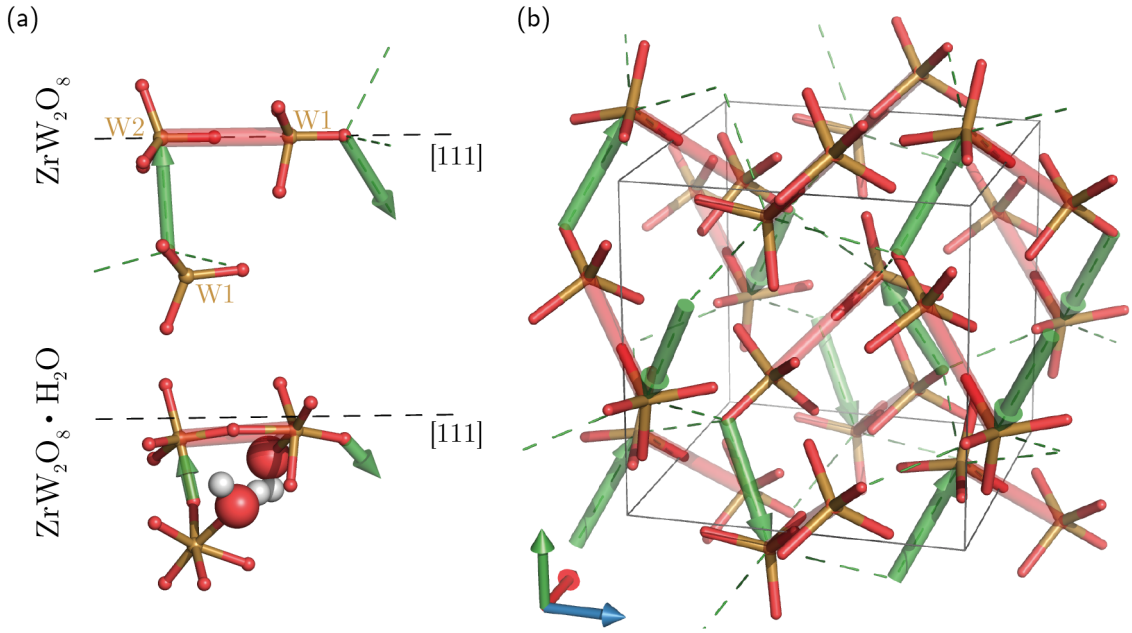


Figure 3.6: Trajectory of ZrW_2O_8 hydration mechanism. (a) The trajectory of the W–O–W chains shows the unique W1 atom for the W2–O3$\cdots$ bond to access, the anticipation of this connection shown as a red cylinder. There remains a choice of three W2 atoms for the W1–O4$\cdots$ bond to approach (the three directions indicated by a a dashed green line), which is made by the water occupancy. The eventual directionality of the W1–O4$\cdots$ bond is shown by a green arrow. (b) The α-ZrW_2O_8 structure is shown, omitting Zr polyhedra, highlighting the trajectories formed in the $P2_12_12_1$ DFT simulation (results presented in Figure 3.5). The calculation suggests a five-coordinate W2 environment and a six-coordinate W1 environment, as shown for example tungsten centers along the $[1\,1\,1]$ axis.

terminal W2–O3$\cdots$ bond in the parent ZrW_2O_8 structure, there are three equally-spaced W2 atoms which each terminal W1–O4$\cdots$ bond might approach. These three choices of W2 atom are related by the same three-fold axis that is broken by the symmetry of the H_2O molecule [Fig 3.6]. In effect, the orientation of each water molecule dictates the choice in a set of 'one-in two-out' rules, and selects one of three possible W1–O4$\rightarrow$W2 linkages. Since the H_2O orientation throughout the structure determines the overall framework connectivity, we sought to establish how different combinations of orientations affect the nature of the relaxed structure.

3.3 Modeling a supercell

The DFT solution represented an over-constrained case of a single unit cell, which possessed anisotropic behavior. In order to resolve some of the isotropic changes and

experimentally observed disorder, the set of ice rules was propagated throughout a larger supercell lattice. We created a subset of configurations using a Monte Carlo simulation to dictate the connectivity, and a simplified harmonic lattice dynamical model to facilitate the structural relaxation of these representative configurations with larger unit cells. The number of these configurations grows exponentially with system size to give a macroscopic number of viable hydration states; as introduced in Section 1.2, the configurational entropy of this system is at least $S_{\text{config}} \simeq R\ln(9/8)$. Thus, attempting to relax a large system is not computationally feasible at the quantum mechanical level.

To generate a supercell with an appropriate set of self-avoiding W–O–W pathways, a Hamiltonian was formulated that satisfied the ice rules, and a MC simulation used to solve for a ground state. For this purpose, a set of adjacency lists $\mathbf{A}_i$ was developed to track the connection of each W1–O4$\cdots$ terminus to the three potential W2 atoms, where an item in the adjacency list $A_i(j)$ has a value of 1 if the corresponding bond is 'on' and a value of 0 if it is 'off'. As such the Hamiltonian for the system is

$$H = \sum_{i}^{n} \left[\sum_{j}^{m} (1 - A_i(j))^2 \right], \tag{3.12}$$

where the outer sum is over n terminal W1–O4$\cdots$ bonds, and the inner sum is over m accessible W2 atoms. It follows that the ground state energy is 0 when every O4 is singly connected, and every W2 satisfies the '1 in 2 out' ice rule.

With the construction of an appropriate Hamiltonian, a Monte Carlo simulation could be used to navigate to a ground state solution by randomly switching bonds and keeping all energetically favorable changes and some energetically poor changes with a given probability (in this case following Boltzmann statistics). Simulated annealing was used to prevent the simulation getting trapped in a local minimum with any defects, as only a ground state with zero energy would be acceptable. A flow chart describing a general SA algorithm is shown in Figure 2.2, with the Fortran implementation offered in Appendix F.

Table 3.2: ZrW_2O_8 empirical potentials. The set of characteristic parameters used in GULP relaxation of the spaghetti topologies.

Parameter	Value	Units
k_{W-O}	1.0	eV Å^{-2}
k_{Zr-O}	2.5	eV Å^{-2}
r_{W-O}	[1.807, 2.086]	Å
r_{Zr-O}	[1.986, 2.129]	Å
k_{O-Zr-O}	1.0	eV
k_{O-W-O}	1.0	eV
θ_{O-Zr-O}	90.0	deg
θ_{O-W-O}	[80.93, 173.94]	deg

Once the connectivity of the network was fixed, the structure was allowed to relax and the distortions tracked. A set of harmonic potentials between bonded atoms — including the connectivity developed by the MC — was developed using the equilibrium bond lengths and bond angles from the DFT calculations. These two- and three-body harmonic potentials described the potential energy surface for a GULP simulation (described in Section 2.3), which sought to minimize the energy through structural relaxation. The force constants for the harmonic potentials were initially pulled from standard bond potential values and adjusted over trials to aid in the deterministic GULP approach and enforce sensible structural changes (such as a drift of the W1–O4$\cdots$). The simulations were completed with $P1$ symmetry, and would avoid any shearing instability. A sample GULP input file is given in Appendix F.

The lattice energy could be minimized at constant pressure (allowing for a flexibility in the lattice parameters) or at constant volume (fixing the unit cell throughout the simulation). In the constant volume case the experimentally obtained lattice parameters from a Rietveld refinement of powder diffraction data of $ZrW_2O_8{\cdot}H_2O$ were used. The procedure was first checked for consistency, by assuring that the single unit cell relaxation reproduced the DFT calculated structure. The GULP program was incapable of successfully minimizing supercells larger than 5×5×5 dimensions

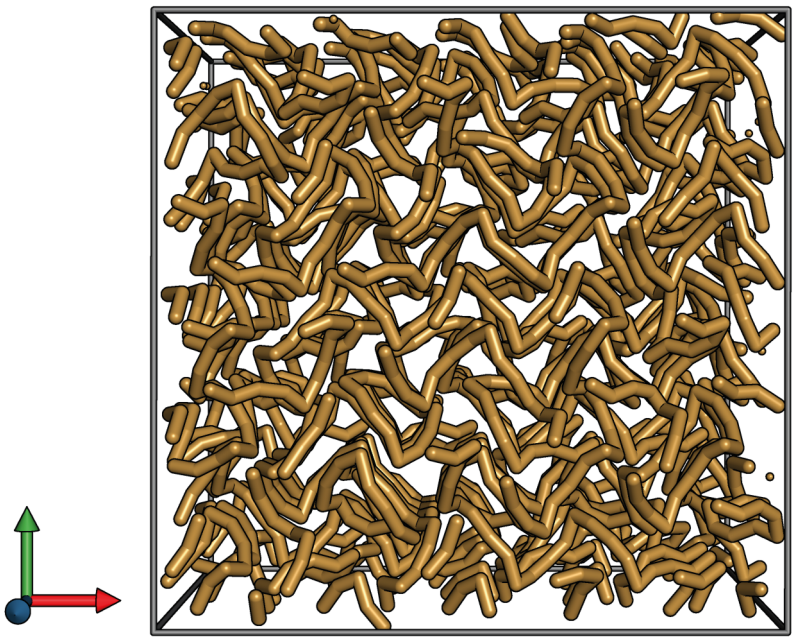


Figure 3.7: Relaxed 5×5×5 supercell The W–O–W connectivity following the GULP relaxation. Both the oxygen and tungsten atoms are displayed as connected, tan cylinders.

(5,500 atoms), with a typical simulation time taking 150 minutes on a single logical core of a 2.7 GHz Intel Core i5 processor.

These relaxed structures are then decorated with water molecules according to the connectivity. It is found that the average equilibrium structure of the constant pressure GULP simulations overestimates the volume reduction compared to the DFT calculation, though the isotropy of the contraction increases with supercell size [Fig 3.8]. The overestimation is likely an artifact of the under-constrained nature of the simple harmonic potentials. A relaxed 5×5×5 supercell configuration is shown in Figure 3.7, highlighting the topology of the spaghetti structured W–O–W chains.

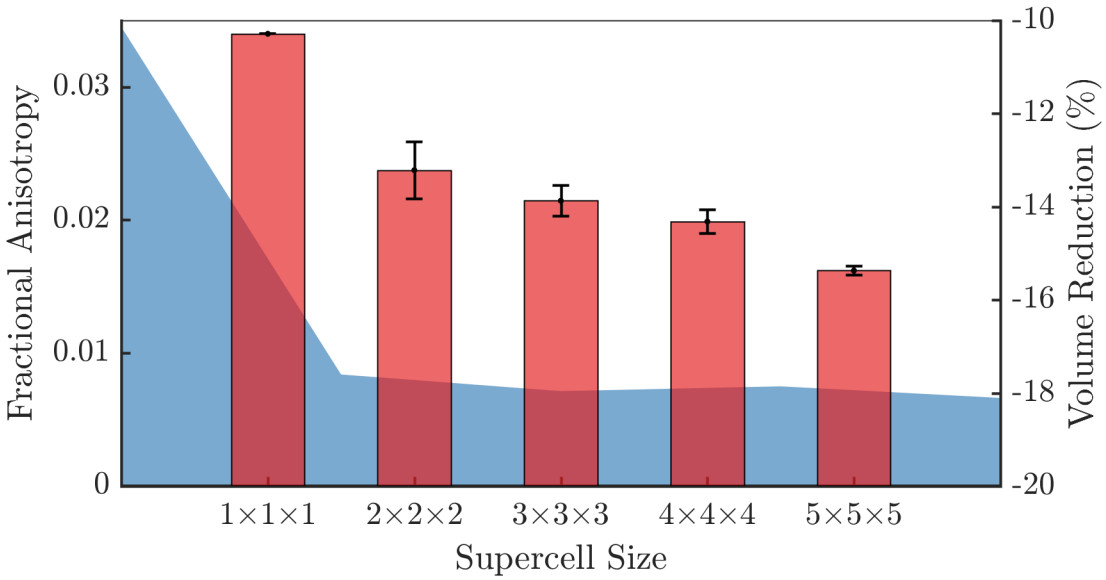


Figure 3.8: Simulation behavior across supercell size. The fractional anisotropy (red bars) is monotonically decreasing with the simulation size considered. The predicted volume reduction (blue shading) at constant pressure is larger than the experimental observation. This is likely due to the relatively weak constraints of the harmonic potentials. Nonetheless, the propagation of spaghetti topology creates a structure which would contract isotropically in the limit of large box size.

3.4 Results

Following the production of a set of spaghetti models, their structural relevance was evaluated by considering the X-ray diffraction pattern and PDF. Figure 3.9 shows the diffraction pattern for a disordered $Pa\bar{3}$ unit cell derived from the atomic coordinates determined in the supercell calculations.

All key features of the experimental diffraction pattern are reproduced, including the absences that had previously been interpreted as indicating $Pa\bar{3}$ crystal symmetry [346, 358]. The inversion symmetry is expected to arise from defects in the hydrated spaghetti network (accounted for by the likely proton migration in a hydrated oxide and topological defects in the network), even though the local symmetry does not contain an inversion center. This phenomena has been observed in other topologically disordered states [387]. The structural model itself is shown as set of probability isosurfaces for the metal atoms in Figure 3.10; the

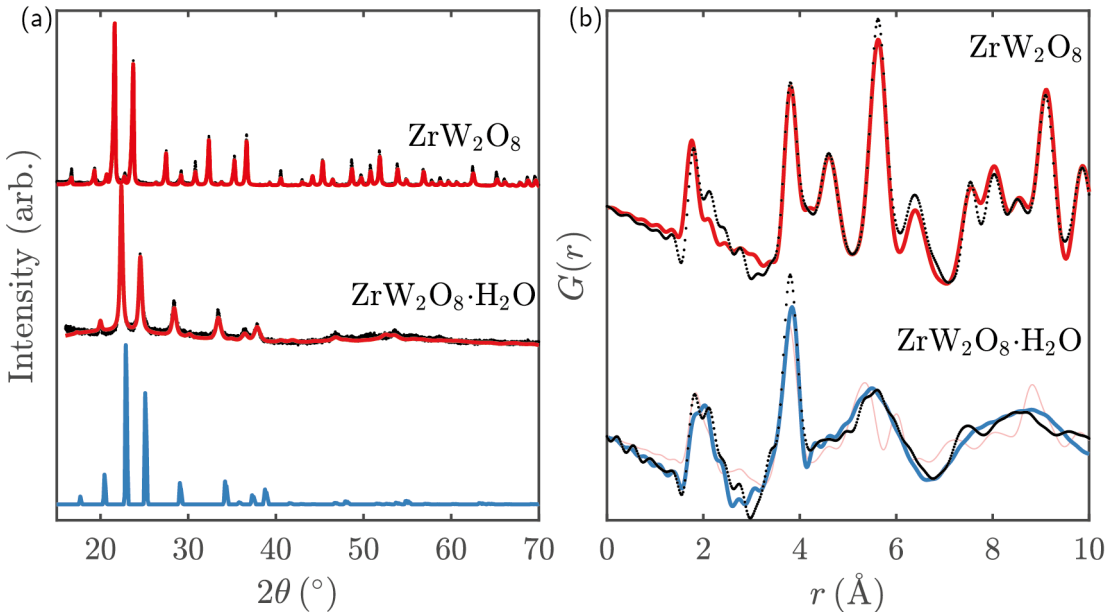


Figure 3.9: ZrW_2O_8 diffraction and PDF. (a) Diffraction data (black) and Rietveld refinement fits (red) for anhydrous and hydrated ZrW_2O_8. Included is the simulated diffraction pattern from a collapsed 5×5×5 supercell model (blue). As discussed regarding Figure 3.8, the limited constraints of the GULP model allow for a larger than experimental volume decrease, and thus a positive shift in 2θ. Nonetheless, from these simple potentials and the spaghetti model, all of the features of the diffraction pattern are reproduced. Detailed refinements and residuals can be found in Figure A.2. (b) X-ray PDFs of anhydrous and hydrated ZrW_2O_8 (black). Included are the fits from PDFGUI using the DFT calculated ZrW_2O_8 (red), $ZrW_2O_8{\cdot}H_2O$ (translucent red), and the 3×3×3 (blue) supercell model. Similar results are produced for a 5×5×5; however, the PDFGUI program was not stable with such a large box size. The DFT calculated $P2_12_12_1$ structure is capable of reproducing the features at low r, but over-predicts ordering at correlations where $r > 4$ Å. The spaghetti model as relaxed in GULP reproduces all of the key features of low- and medium-range correlations. Detailed refinements and residuals can be found in Figures A.3–A.5.

substantial disorder present explains the previously-noted difficulty of interpreting even high-quality neutron diffraction data [346]. Consequently we can be confident that hydration-driven volume collapse in $ZrW_2O_8{\cdot}H_2O$ is driven by a degenerate family of spaghetti-like lattice distortions, modulated with periodicities distributed continuously throughout the BZ.

Using PDFGUI, in-house X-ray PDF data were analyzed for both ZrW_2O_8 and $ZrW_2O_8{\cdot}H_2O$ with the aim of establishing the consistency between the local structure of different models and the data. Starting from the well known crystallographic structure of ZrW_2O_8, and the literature values for the isotropic atomic displacement

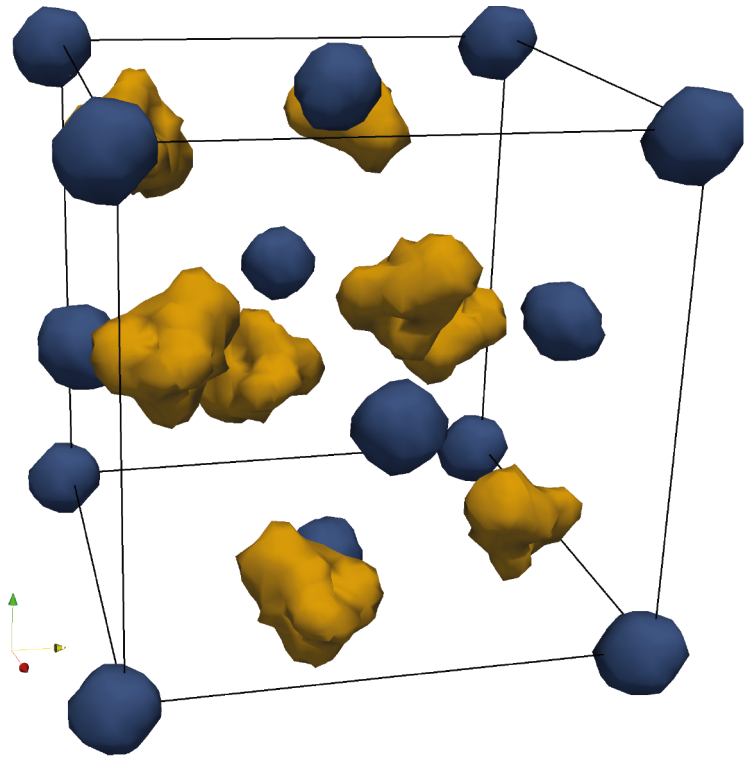


Figure 3.10: Probability isosurface of disordered $Pa\bar{3}$ $ZrW_2O_8{\cdot}H_2O$. A relaxed supercell configuration from GULP is collapsed onto a single unit cell and $Pa\bar{3}$ symmetry is applied. The unit cell is then divided into voxels and statistics for each site are calculated. The isosurfaces show a probability of occupation greater than 0.85 for the tungsten (yellow) and zirconium (blue). Oxygen atoms are omitted.

parameters [360], the PDF could be refined using only the scale and S_{ratio} (a useful parameter when there is a known first metal–metal peak). From this refinement, a starting set of parameters for the modeling of the hydrated phase was gained. The atomic displacement parameters in the hydrated phase were fixed to these values; if they were allowed to vary freely, the mobile tungsten (W1) would (as expected) develop a non-realistic and large ADP. The refinement from the DFT and GULP simulations used only the scale factor and S_{ratio}. Comparison of the DFT- and GULP-produced PDFs show a stark difference in the respective local structure predicted by each level of modeling. The quantum mechanical calculations are capable of reproducing the short range order in the system; however, they over-predict the long range order beyond the first unit cell. This artificial periodicity is resolved in the supercell model.

Given that the correlated disorder in the system helps explain the diffraction behavior observed, it is worth considering how the distortions relate to NTE. The lattice distortions associated with this hydration model are dense throughout the Brillouin zone, due to the configurational degeneracy associated with the many accessible spaghetti states. Are the same atomic displacements of the NHE mechanism somehow related to the phonons responsible for NTE in ZrW_2O_8?

Published data [27] allowed for access to the frequencies $\omega(\mathbf{k}, \nu)$, eigenvectors $\mathbf{e}(\mathbf{k}, \nu)$, and Grüneisen parameters for each phonon branch, ν, and wave-vector, $\mathbf{k}$. The NTE modes are those for which ω is small and γ is large and negative. By comparing the atomic coordinates of each ZrW_2O_8 framework atom in the anhydrous and hydrated DFT structures, we formed a normalized, mass-weighted displacement vector $\mathbf{u}$, akin to that of equation (3.7), which describes the volume-reducing framework deformations induced by hydration. The projections

$$c(\mathbf{k}, \nu) = \mathbf{u} \cdot \mathbf{e}(\mathbf{k}, \nu) \tag{3.13}$$

partition these deformations amongst the normal modes; for the single unit cell DFT structures, the $c(\mathbf{k}, \nu)$ necessarily vanish for $\mathbf{k} \neq 0$. Remarkably, 88% of the displacements described by $\mathbf{u}$ can be accounted for in terms of two of the lowest-energy $\mathbf{k} = 0$ NTE modes [Fig. 3.11]. The same result is found for zone center phonons generated by the quantum mechanical calculations used in this study.

Extending this calculation to the $3 \times 3 \times 3$ supercell, the relevant NTE modes are associated with a more diverse set of wave-vectors $\mathbf{k} \in \langle \frac{n_1}{3}, \frac{n_2}{3}, \frac{n_3}{3} \rangle$, $n_i \in (0, \pm 1)$. Considering the full ensemble of hydration states with complex periodicities, we demonstrated projections onto NTE modes distributed throughout the BZ, such that large-amplitude activation of the most important NTE phonons can be said to account for nearly all of the ZrW_2O_8 framework response to hydration [Fig. 3.12]. The projections onto lowest energy modes (mostly acoustic) account for $\sim$45% of the displacement and a shearing of the lattice, whereas inclusion of all modes with neg-

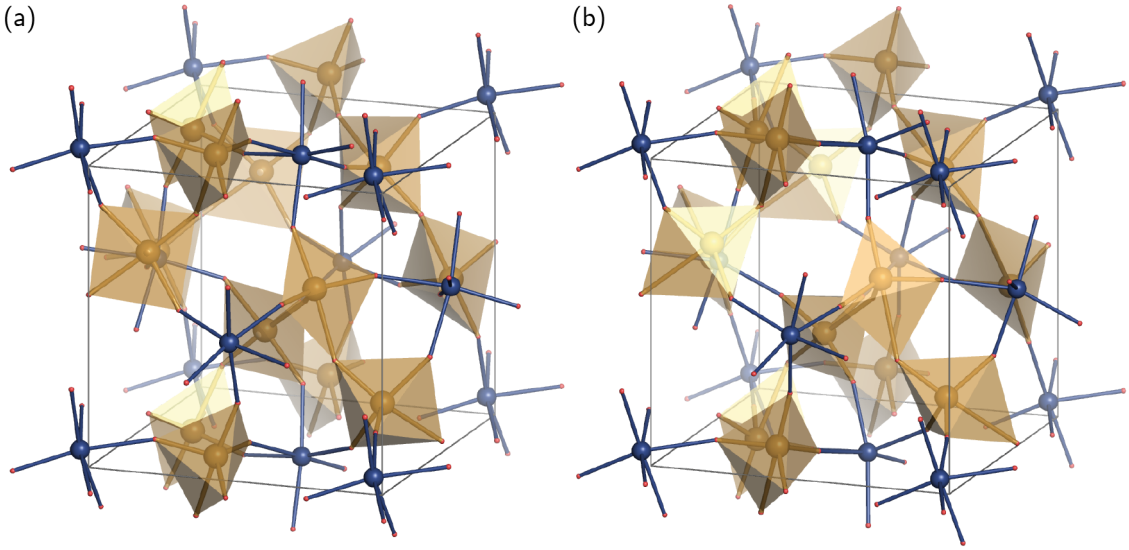


Figure 3.11: Linear combination of displacements from phonon projections. (a) $P2_12_12_1$ structural model DFT relaxation is shown as a reference for the hydration mechanism, omitting water molecules as these cannot be included in the phonon projections for phonons calculated from anhydrous ZrW_2O_8. (b) The spaghetti dynamics are reconstructed from a linear combination of only two modes at zone center to recover 88% of the hydration induced displacement. These structures are nearly indistinguishable, demonstrating the extent to which low energy phonon modes can recreate the spaghetti dynamics and unique W–O–W connectivity. Beginning from the same alpha phase, the changes from the (b) two-mode linear combination and the (a) complete DFT relaxation produce a similar resulting structure.

ative Grüneisen parameters reveals ∼88% of the framework response. Animations of these deformations are presented in Appendix F.

From this result, the key NTE phonon branches are implicated in the NHE response of ZrW_2O_8, and the quenching of NTE is explained. The spaghetti displacements involved in the NHE are shown to be intimately involved in the NTE dynamics by the use of the projection values, $c(\mathbf{k}, \nu)$. Whereas previous studies have focused either on individual collective modes (such as RUMs [384]) or on the correlated motion of isolated structural fragments [207, 380], the set of $c(\mathbf{k}, \nu)$ values allows for reconstruction of the NTE modes with an intuitive real-space realization. For example, the $P2_12_12_1$ DFT configuration of Fig. 3.5 gives a framework displacement vector $\mathbf{u}$ that projects almost entirely onto just two zone-centre modes ν_1, ν_2 (note $\omega_1 = \omega_2 = 44.1$ cm^{-1} and $\gamma_1, \gamma_2 = -4.14$). The combination mode $c(0, \nu_1)\nu_1 + c(0, \nu_2)\nu_2$ then describes one specific collective vibra-

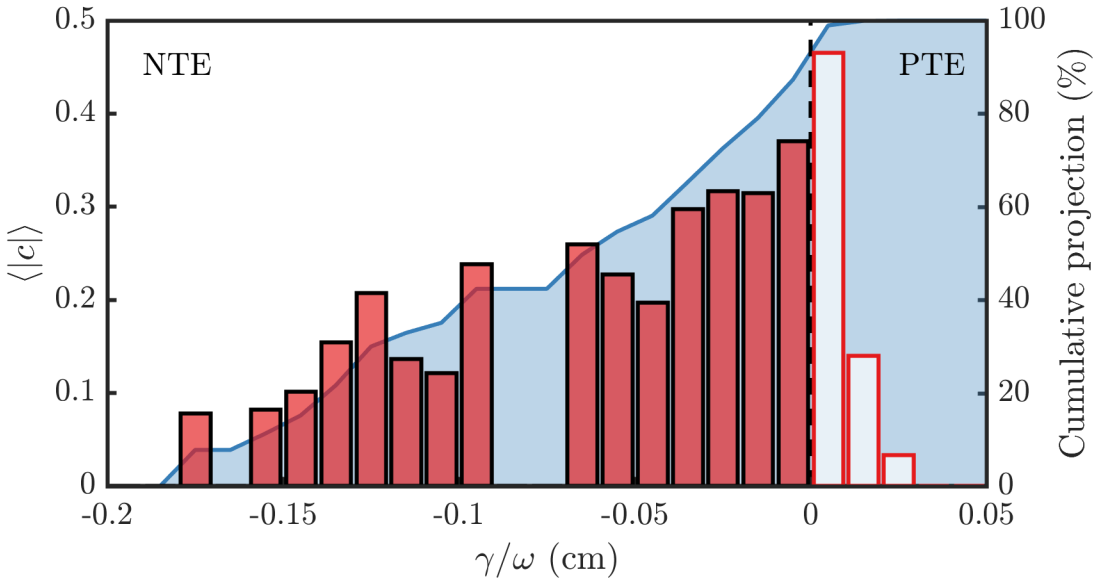


Figure 3.12: Projection v Grüneisen parameters. Average magnitude of projection c (bars) as a function of γ/ω for 10 different $3 \times 3 \times 3$ supercell relaxations. Nearly 90% of the cumulative projection (blue shaded region) is accounted for by NTE modes.

tion of ZrW_2O_8 contributing to NTE. Crucially, the associated displacement vector $\mathbf{e}' = c(0,\nu_1)\mathbf{e}(0,\nu_1) + c(0,\nu_2)\mathbf{e}(0,\nu_2)$ does indeed involve fluctuations towards the very same spaghetti states assumed by $ZrW_2O_8{\cdot}H_2O$ [Fig. 3.11]. The same holds for the more complex family of NTE modes associated with the $3 \times 3 \times 3$ spaghetti configuration [App. F].

3.5 Conclusions

From these results, we have demonstrated the utility of mixed modeling techniques in eliciting understanding of complicated systems, such as ZrW_2O_8, by establishing a key real-space mechanism for NTE in ZrW_2O_8 and its relation to NHE. Through a correlated rotation and translation of ZrO_6 and WO_4 polyhedra toward the spaghetti states adopted in $ZrW_2O_8{\cdot}H_2O$ and the PIA phase of ZrW_2O_8, the anhydrous framework structure is capable of a thermally invoked volume collapse. The configurational entropy associated with this concerted motion — a function of the unique ZrW_2O_8 topology — is so large that these fluctuations are described by a family of phonons spread throughout the BZ: individual spaghetti states are thus described

by the population of many NTE phonons. Consequently a 'spaghetti dynamics' mechanism for NTE in ZrW_2O_8 finally explains (1) why NTE modes are distributed throughout the BZ, (2) why it is so difficult to interpret individual NTE phonons in isolation, and (3) how the local polyhedral rotations/translations identified in Refs. 26, 27, 380 actually propagate throughout the crystal lattice. Moreover, in building a relationship between NHE and NTE, this explains the recent demonstration that not all adsorbates produce the same volume collapse [347]; the point group of NH_3 is compatible with the space group of the anhydrous framework, therefore, there is no need to break the threefold axis aligned with the W1–O4 bond and no emergent configurational entropy.

In the context of this thesis, this breakthrough is indicative of the power of informed modeling techniques. Experimental scattering information offered not only a validation of the final model, but advised the changes in coordination seen in the first principles calculations. In a way that a standard Rietveld refinement could not, a mixed modeling approach balances the information poor scattering from disordered phases, limits in the computational reach of quantum mechanical calculations, and accessible local and bulk experimental probes. Conceptually, we created a physical model that reproduces the features of the scattering pattern with relatively few degrees of freedom in the procedure. This concept of prior information to dictate local rules or simplified potentials, has far reaching implications for the study of disordered materials, and the problems addressed in this thesis.

Chapter 4

Square-planar amorphous zeolitic imidazolate frameworks

By constructing the spaghetti topology from a set of local rules, we took an inferential approach to understanding the structure of hydrated ZrW_2O_8. In what follows, I will use the same concepts of informed modeling — using the inference from local probes to guide a model with few degrees of freedom — to approach an amorphous system. We examined amorphous materials and the historical attempts at understanding their structure in Chapter 1 through the lens of locally tetrahedral *a*-Si. In this chapter, I demonstrate a procedure to generate amorphous structures with a square-planar local geometry. I then use these isotropic structures to model an amorphous metal-organic framework.

In particular, the focus will be on platinum and palladium imidazolate, $M(im)_2$, an analogue to traditional zeolitic imidazolate frameworks (ZIFs) with a square-planar d^8 metal center (Pd^{2+} and Pt^{2+}). Placed against the backdrop of ZIF developments and previous work in modeling continuous random networks (CRNs), the key results of this chapter are the generation of square-planar continuous random networks for the first time, and the local structural understanding of amorphous $Pd/Pt(im)_2$ produced by informed modeling procedures. As in the previous study, the PDF is used in conjunction with local constraints to produce a viable structural model through stochastic sampling of topology. I would like to acknowledge Dr. Andrew Cairns, Dr. David Keen, Dr. Matthew Tucker, and Dr. Leigh Connor for their experimental contributions to the scattering and normalization used in this work; Jerome Luchesa-Smith for his work in prototyping this model; and Prof. Mark Wilson for insightful discussions.

4.1 Amorphous metal-organic frameworks

A contemporary class of materials, metal-organic frameworks have been the subject of a growing body of research over the past 30 years since conception [388]. Crystalline MOFs are network solids in which inorganic nodes (clusters or metals) are bound by organic linkers. The ability to substitute the nodes and linkers has created rapid progress in material and crystal engineering: the number of novel MOF structures reported each year is greater than 6000 [389]. This customizability in conjunction with the porous architecture has given rise to significant research interest, mainly because of the vast prospective applications in gas storage and separation [390, 391], catalysis [392], sensing [393], and drug delivery [394]. As more creative node and linker combinations are developed, the opportunity for tuning materials to specific applications grows. Moreover, there has been a surge in work on functionalizing and altering existing MOF structures to retain the framework topology, but alter chemical and physical properties. A number of recent reviews have been produced to describe this bourgeoning field of synthesis, structure, and applications [392, 395].

Far less researched, amorphous MOFs (*a*-MOFs) maintain the main components of their crystalline analogues; however, they possess disordered topologies. That is, they are well differentiated from their crystalline counterparts by their lack of long-range order, resulting in diffraction patterns without Bragg peaks.

From an industrial viewpoint, they may lead to more efficacious applications than their crystalline counterparts, especially in areas that would benefit from a reversible collapse of a porous MOF structure. The first applications to come to mind would be reversible gas storage or effluent scrubbing. Using PIA for reversible gas storage or to change the equilibrium thermodynamics of adsorption for long-term storage has already been presented [30, 396]. Notably, *a*-MOFs are often more mechanically robust than crystalline materials, which lends to their utility in industrial applications [389].

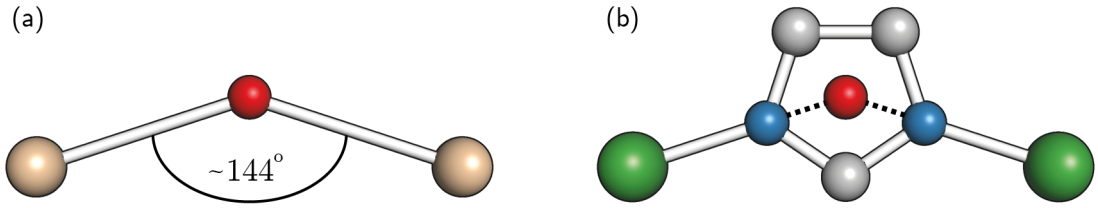


Figure 4.1: Comparative representation of the Zn(im)$_2$ and silicate local geometry. The node–linker–node relationship between the two tetrahedrally coordinated structures is made apparent in the highlighted angle. Atom coloring: Si (tan), O (red), Zn (green), N (blue), C (grey).

A relatively untapped field, there are only ∼30 unique *a*-MOFs that have been reported. In conjunction, liquid MOF phases have been the subject of a growing body of research [300, 397]. Most *a*-MOFs have been prepared through stress to a crystalline MOF *via* hydrostatic PIA or mechanical ball milling [30, 398]. Thermally induced amorphization [31] and *in situ* formation [399] have been reported as well.

4.1.1 Zeolitic imidazolate frameworks

A particularly interesting subclass of MOFs, zeolitic imidazolate frameworks (ZIFs) are defined in analogy to the family of aluminosilicate materials. Instead of an oxygen, metal centers of a ZIF are bridged by an imidazolate ion. These materials are of great interest for amorphization because their structures are closely related to zeolitic silica polymorphs by virtue of their local geometry and connectivity [Fig. 4.1]. In other words, amorphous ZIFs are the MOF analogs of *a*-SiO_2.

ZIFs share in the useful properties of both zeolites and MOFs. Not only are they thermally and chemically stable with high surface areas, they are diverse in topology and easily functionalized chemically [392]. The isoreticular functionalization of MOFs can modify the organic linkers (imidazolate) without disrupting the connectivity, using traditional synthetic approaches [400]. Subsequently, the majority of research has been conducted on ZIFs which share the aluminosilicate topologies, namely those with tetrahedrally coordinated metal centers (Zn^{2+} or Co^{2+}) [401]. This potential for diversity is limited by experimental accessibility both thermody-

namically and kinetically [402]. To date only around 200 aluminosilicate topologies are known [403], with over 100 — and rapidly growing — ZIF topologies experimentally observed [400]. Besides synthetic accessibility and utility, having a prominent and well researched structural analogue facilitates the development of structure–property relationships in the novel materials.

A material possessing an interesting amorphous phase and no exact zeolitic sister topology, ZIF-4, is shown in Figure 4.2 [401]. Its ambient phase adopts the **cag** topology and space group *Pbca*. This material — as well as a set of similar ZIFs [31] — undergoes an irreversible temperature-driven phase transition to the amorphous phase, *a*-ZIF, and then to a high density phase, ZIF-zni [28]. Amorphous ZIFs can potentially combine the utility of silica glass with the functionalization associated with MOFs. After heating to 300 °C, the framework becomes amorphous and can be cooled back to ambient temperature without undergoing an ordering phase transition. The irreversibility and subsequent transition to the high density phase at 400 °C, has been investigated in depth to show an intermediate order-disorder liquid phase transition and kinetic trapping [397]. Other means of producing the amorphous phase are possible *via* pressure-induced amorphization [29, 404], mechanical shearing [405], or starting from other ZIF topologies [301, 398]. The accessibility of a reversible pressure-induced amorphization in a pore occupied ZIF-4 further demonstrates the diversity of this material class [30].

Using neutron and X-ray total scattering data, the structure of *a*-ZIF has been previously investigated by reverse Monte Carlo (RMC) refinement [28]. It was noted that the PDF of *a*-ZIF and two crystalline structures were nearly identical in the low-r region, confirming the consistent tetrahedral coordination environment of Zn and structure of the imidazolate bridges between the three phases. An attempt at refining a crystalline starting configuration (both ZIF-4 and ZIF-zni) against the *a*-ZIF total scattering data and PDF was unsuccessful at producing a sensible model commensurate with the data [28]. In the same study, a continuous random network

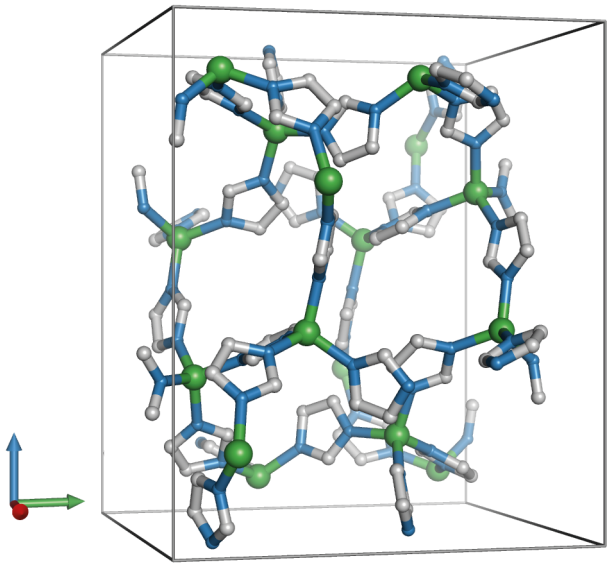


Figure 4.2: ZIF-4 structure. ZIF-4 possesses the space group *Pbca* (known as the **cag** topology in the MOF community). The overall composition is $Zn(im)_2$ with Zn atoms shown in green, N atoms in blue, and C atoms in grey.

(CRN) topology based on amorphous silica was decorated with $Zn(im)_2$ to initiate a refinement which closely fit the *a*-ZIF total scattering data. The motivation for this decoration is apparent when qualitatively comparing the PDFs of *a*-SiO_2 and *a*-ZIF [Fig. 4.3]. In conjunction with microscopy data across phases, this demonstrated that the *a*-ZIF is truly amorphous, maintains the local order of the crystalline counterparts (the tetrahedral coordination and imidazole molecular structure), and is not a highly disordered crystal-like topology.

4.1.2 Square-planar imidazolate frameworks

In contrast with the previously successful work, $Pt(im)_2$ and $Pd(im)_2$, do not possess a structural analogue for their amorphous phases. They are synthesized through an orthorhombic intermediate, $M(Him)_2(im)_2$ *via* heating in vacuum or a water bath [Fig. 4.4]. Both of these compounds have similar IR spectra, and are diamagnetic, colorless solids, indicating a square-planar metal center in both cases [406]. This is to be expected for Pd/Pt(II); however, due to the insufficiency of crystalline structures as starting points for modeling *a*-ZIF, quite the quandary remains in un-

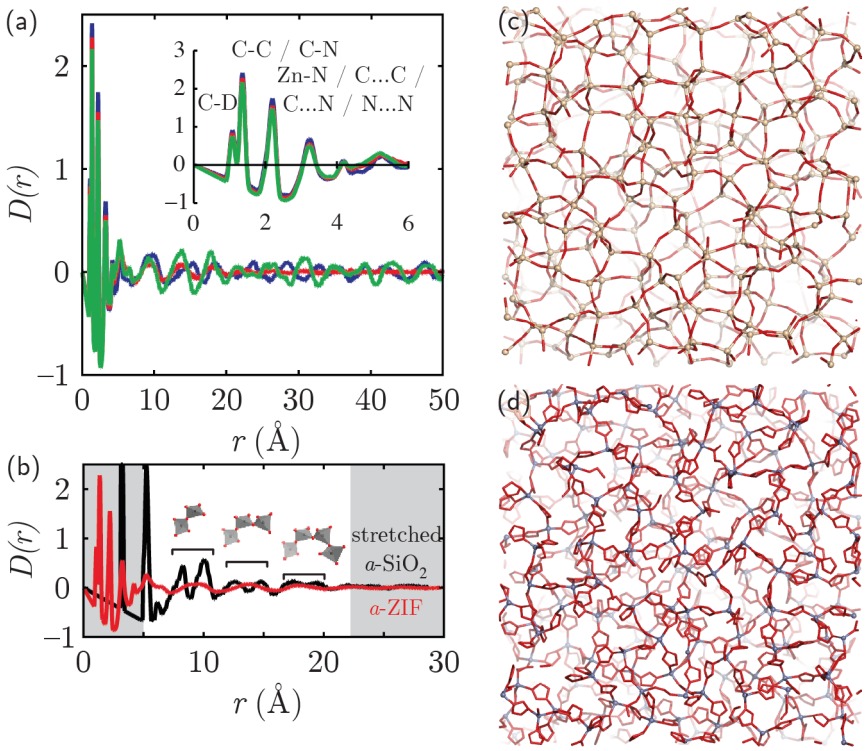


Figure 4.3: Scattering behavior of an amorphous tetrahedral ZIF. (a) Neutron differential correlation functions $D(r)$; the inset illustrates the practically identical local structure in all phases. (b) $D(r)$ functions compared for *a*-ZIF and *a*-SiO_2, where the latter data have been stretched in r by a factor of two to reflect approximately the difference in Zn$\cdots$Zn and Si$\cdots$Si distances in the two materials. Figure adapted from Ref. [28]. On the right, a ball and stick representation of (c) *a*-SiO_2 and (d) *a*-ZIF. The imidazolate linkers are colored (red) the same as the oxygen linkers to accentuate the similarities between the two structures, while the Si atoms are shown in tan, and the Zn atoms in grey.

derstanding the structure of an amorphous square-planar ZIF. The crystal structure of $M(Him)_2(im)_2$ is held together by hydrogen bonding and is not a covalent network structure (like the CRNs we have seen). Thus, it is not a suitable analog for modeling $Pd/Pt(im)_2$. Like *a*-ZIF, it can not be considered a glass since it it is not formed through a melt, does not posses a glassy transition, and involved the making a breaking of bonds at the phase transition [28].

Historically, well characterized materials amorphous materials are predominantly composed of tetrahedral subunits. Although previous work has examined triangular (in the case of B_2O_3 [407–409]) and octahedral (in the case of FeF_3 [240, 410]) subunits in amorphous materials, to the best of our knowledge, square-planar amor-

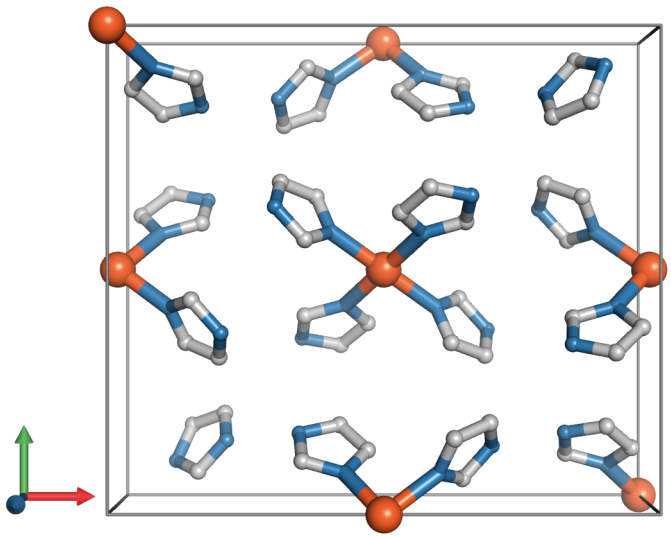


Figure 4.4: The crystal structure of Pd(Him)$_2$(im)$_2$ as reported in Ref. [406]. This crystal structure, space group $Pnca$, was the precursor polymeric network for *a*-Pd(im)$_2$. The Pd atoms are shown in orange, N atoms in blue, and C atoms in grey, and hydrogens omitted for clarity.

phous materials are entirely unexplored. Nonetheless, the square-planar subunit is distinctive in its quadrapolar nature and opens new avenues for this class of engineered materials. The propagation of anisotropy within a locally D_{4h} amorphous network will be fundamentally different than for the locally T_d CRNs used to describe SiO_2 and *a*-ZIF. Multipolar order could have ramifications for the optical properties of a material if this order is on the length scale of light. Moreover, the anisotropy may impact the mechanical properties (for casting, processing, and material stability) and pore behavior (for gas separation). Thus developing an understanding of an amorphous network of quadrapoles would mark a significant advancement in the engineering and design of functional materials.

In order to develop a model, we considered how the original silica models were made, and pursued an avenue for casting such a model with a square-planar local geometry.

4.2 Methods

In Section 1.3 we explored the canonical and modern approaches for generating models of amorphous materials. Particularly, the case of modeling *a*-Si — and subsequently *a*-SiO_2 — has played an important role in differentiating between approaches, and benchmarking their key results. Of these, the Wooten, Winer, and Weaire (WWW) method has grown since its conception to include various modifications and approaches to structural relaxation. This method has been used to produce models consistent with the structure factor, pair correlation functions, elastic constants, and regions of the vibrational density of states [247]. Recent work has expanded the model to work across temperature-driven phase transitions (the original algorithm being simply an optimization tool for structurally relaxed, topologically disordered network), overcoming a major challenge that previously necessitated molecular dynamics [411]. Moreover, research using the WWW method has determined that the total scattering structure factor of an amorphous material cannot necessarily be considered a universal quantity for the disordered network, and amorphous silicon spans a vast diversity of ground states [249].

Given the considerable success of the continuous random network model, it is sensible in a first pass to treat the amorphous square-planar ZIF phases as CRNs. Following the methods utilized in generating tetrahedral CRNs, we sought to apply these techniques in the context of a square-planar continuous random network as an attempt to analyze the structure of platinum and palladium imidazolate*. To accomplish this, we constructed a WWW simulation to create starting structures that could be refined against total scattering data.

4.2.1 WWW method

At its core, the WWW method is a structural minimization technique based on the Monte Carlo methods presented in Section 2.4. The algorithm proposes a topological

*This work will continually make reference to $Pd(im)_2$ instead of explicitly stating both metal centers, as the structure of the two materials is similar.

rearrangement *via* a bond-swap move, and accepts that move with the Metropolis probability (2.25). That is, if the change in connectivity acts to reduce the energy, the move is accepted; however, if the change increases the overall energy, the move is rejected with a finite probability related to the Boltzmann distribution. In order to determine the change in energy after a bond-swap, a structural minimization is used to calculate the total energy with all of the atoms at their new equilibrium positions. This minimization will also often take the form of a MC simulation.

Importantly, the WWW method depends on topologically dependent potentials. If there is no variability associated with the bond swap, the outer MC becomes meaningless. The types of two- and three- body potentials explored in 2.3, as well as others such as the Keating potential are amenable to this requirement.

The significant benefit of this approach is that it allows for topological rearrangement in a computationally facile manner, in a way that is considerably challenging for *ab initio* methods alone. Because the topological rearrangements are pair-wise and equivalent bond swaps, this method also ensures a local coordination that is propagated throughout the bulk structure. Although a defect-free structure is not entirely physical, it is highly amenable to probing a system where the limited available information indicates a well defined local coordination.

4.2.2 Starting configuration

A primary consideration in the construction of any CRN is the starting configuration. A completely randomized network, or a known crystalline structure can be used initially. Although a random network will have no memory of any particular crystal state [245], it will start in a highly strained configuration with unphysical bond separations. Particularly in the case where the amorphous phase is related to a known crystalline phase, beginning the simulation at the crystalline state and using temperature to prompt disorder in the system is appropriate. This has been done successfully with using cubic β-cristobalite in constructing vitreous silica by a WWW simulation [252]. The similarity between the phases is apparent in comparing the

local vand medium-range order and overall atomic density: $0.0657\,\text{Å}^{-3}$ for *a*-Si versus $0.0661\,\text{Å}^{-3}$ for β-cristobalite [191].

When considering the starting configuration of $Pd(im)_2$, it is not feasible to start from a randomized state due to the lower symmetry of the subunit. Contrasted with *a*-Si, the local coordination has two unique environments that define the 90° and 180° angles. There are also no potential three-dimensional square-planar CRNs to decorate or use as a starting configuration, even though there is a two-dimensional square-planar CRN in the literature [412]. As such, we used a crystal structure of compatible local geometry and connectivity as a starting configuration.

Similar to the utility of β-cristobalite in casting a starting topology for the construction of *a*-SiO_2, NbO* was used as an initial square-planar network to feed the WWW simulation. The **nbo** net was chosen as the highest symmetry net that satisfied the potentials of the internal Monte Carlo, and in aid of the search for an isotropic result. Only a small set of nets with square-planar subunits is known, with an even fewer number of nets with D_{4h} local symmetry. The choice of different starting nets is expected to have less impact than the box density in the stochastic procedure [252]. Thus, exploring the available starting nets — including nets with greater than single penetration — and their impact on resulting CRNs will be left as a subject for further study.

Both the oxygen and niobium coordinates were replaced with metal centers (Pd). A chalcogenide (S or O) was placed at the midpoint of each Pd–Pd bridge in the crystallographically distinct positions [Fig 4.5]. This decoration allowed for the adjacent chalcogenide bridges coordinating the metal center to be distinct — a necessary condition when considering the angles in the square geometry: X–Pd–X bonds always have 180° angles and X–Pd–Y bonds always have 90°. Next, the unit cell was scaled so that knowledge of *a*-SiO_2 geometry could be exploited ($a = 6.1125\,\text{Å}$, $\rho = 0.0788\,\text{Å}^{-3}$). Extensive historical efforts have developed well parameterized

*The topology of the NbO structure is conveniently given the topological notation **nbo**.

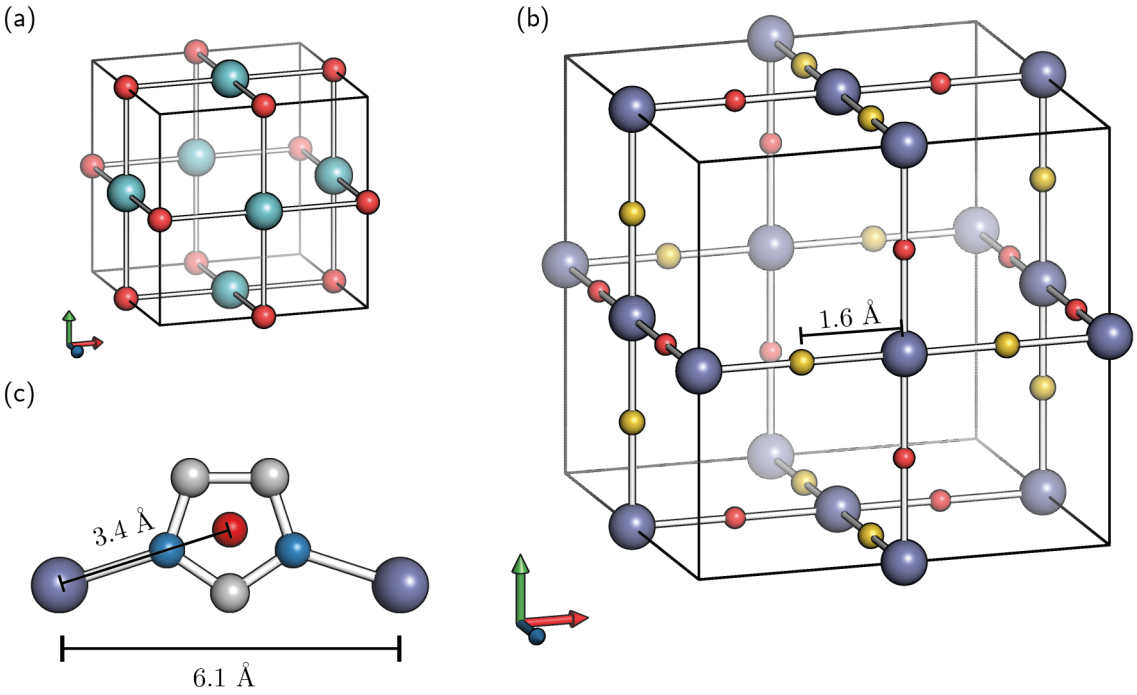


Figure 4.5: WWW starting configuration for a-Pd(im)$_2$ generation. (a) Niobium oxide crystal structure used as basis for generating a starting configuration for the *a*-Pd(im)$_2$ WWW simulation. This structure was (b) scaled and decorated with chalcogenides at the midpoints. This network is scaled such that the initial Pd–X distance is $\approx 1.6\,\text{Å}$. The fictitious PdSO is then used as a starting square-planar network for the WWW runs. Following completion of the WWW simulation, structure is again scaled to be consistent with the Pd–Pd distance observed in the PDF ($6.1\,\text{Å}$): (c) from the law of cosines, the expected Pd–X distance is calculated as $\approx 3.4\,\text{Å}$, where X serves as the origin for the imidazolate decoration.

potentials for characterizing silicon, including potentials based on the bond lengths and angles. In this scaled NbO configuration, the expected Pd–X bonding distance is similar to that of Si–O, and the angle at the chalcogenide linker will be similar to that of oxygen in SiO_2. Although there is no extensive IR data to use in estimating the relative harmonic strengths, we can use the *a*-SiO_2 modeling of the literature as a starting point. A good discussion of these amorphous silica potentials is offered by von Alfthan *et al.* [252].

From this newly decorated fictitious structure, a supercell was made by tiling the unit cell in space. To mitigate edge effects, the cell length of the cubic supercell should be at least twice that of the largest pair correlation length of interest. On completion of the WWW simulation, the cell was expanded once more to

match the expected value distance between the metal and linker in $Pd(im)_2$, and the chalcogenides were replaced by imidazolate ions to form a completed $Pd(im)_2$ model ($a = 12.295\,\text{Å}$, $\rho = 0.0097\,\text{Å}^{-3}$) [Fig. 4.5c].

4.2.3 Square-planar WWW algorithm

The WWW algorithm proceeds as a matryoshka doll of Monte Carlo simulations [Fig 4.6]. The external MC has moves which are considered bond swaps that adjust the topology of the network. The acceptance of the bond swap is dependent on a Boltzmann relation where the change in energetics across the sweep is in reference to the relaxed structure with each topology. It follows that the internal MC minimization is the structural relaxation that feeds the external energetics.

The external bond swapping is equipped with two distinct thermostats. The first is the simulated annealing (SA) process already described [Section 2.4]. Secondly, an adaptive temperature scheme was implemented to ensure disorder in the system. An adaptive thermostat adjusts the temperature based on an acceptance criterion: if too many moves are being accepted, the temperature is reduced by a fraction, $\frac{1}{\xi}$; if too many moves are being rejected, the temperature is increased by a fraction, ξ. This adjustment is made after a set number of moves throughout the simulation in which the acceptance ratio, η, measured:

$$T_{i+1} = \begin{cases} T_i, & \text{if } \eta_{\min} < \eta < \eta_{\max} \\ \xi T_i, & \text{if } \eta < \eta_{\min} \\ \frac{1}{\xi} T_i, & \text{if } \eta > \eta_{\max} \end{cases} . \tag{4.1}$$

The topology of the network is dictated and tracked by an adjacency matrix. An adjacency matrix is an $N \times N$ symmetrical matrix, where N is the number of atoms in the system, which keeps track of any adjacency (or bonding) in a binary fashion. This matrix accelerates the energetic calculations in the Fortran standard by not considering non-bonded atoms, and by efficiently searching memory addresses for

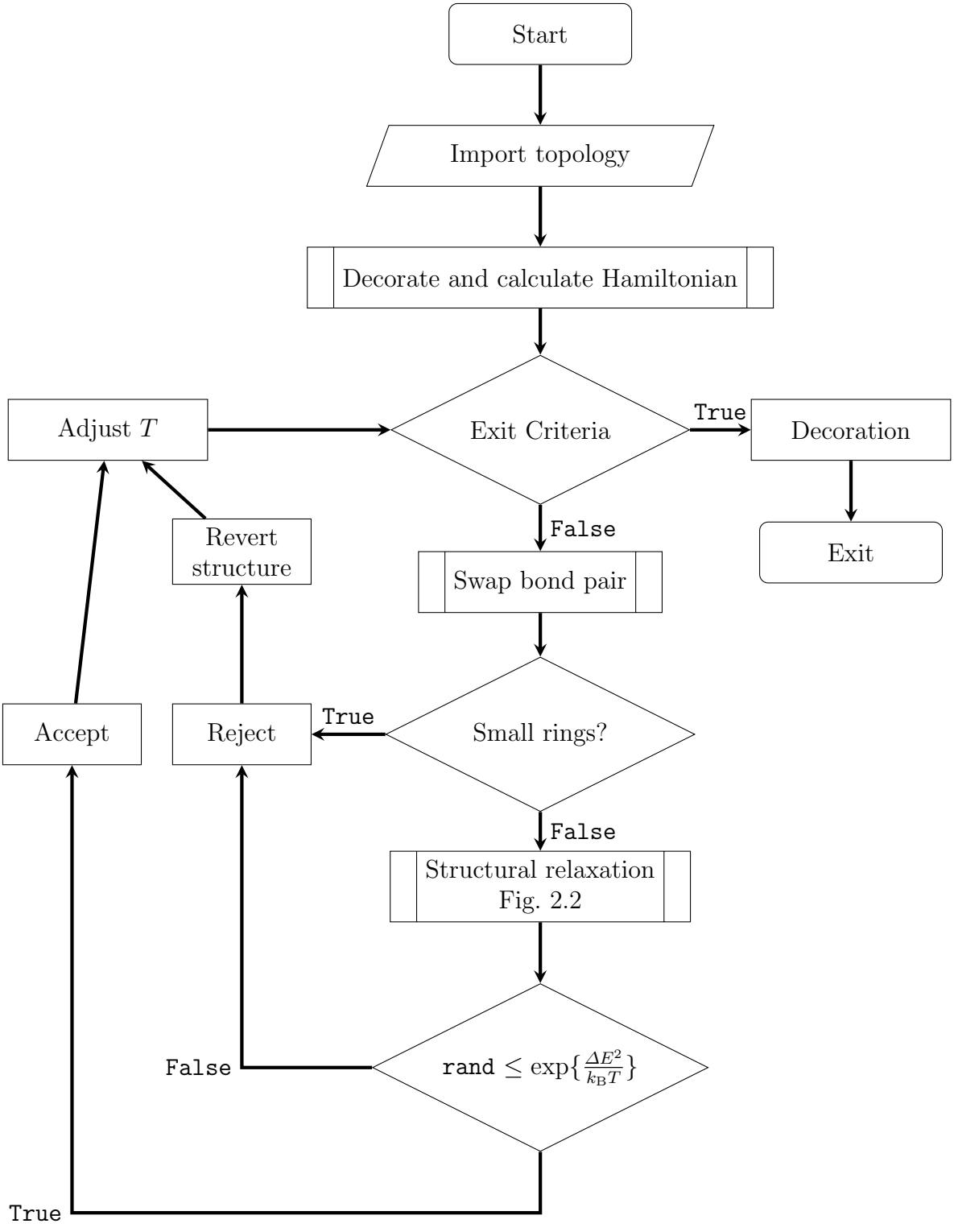


Figure 4.6: WWW simulation flow chart. The WWW bond swapping algorithm that builds an amorphous structure. Details of predefined processes (initialization, bond swapping, and structural relaxation) are omitted for clarity and described in the main text.

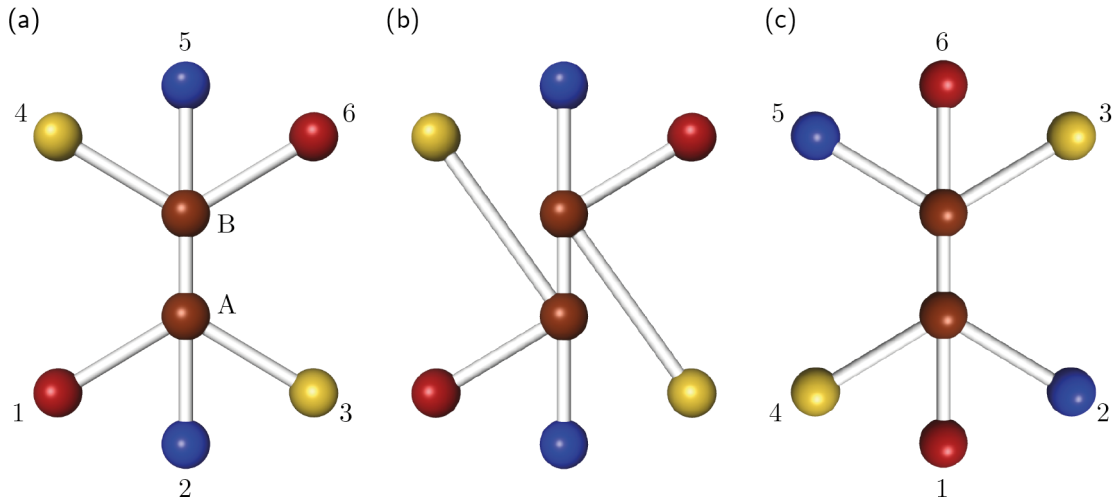


Figure 4.7: Illustrative bond swap. (a) A starting eight atom cluster at equilibrium in open boundary conditions. (b) The same cluster after a bond swap and (c) subsequent structural relaxation. The atoms are labeled with reference to the explanatory adjacency lists.

adjacent members.

In a single species system, the bond swap is achieved by first selecting two adjacent nodes at random, A and B. By indexing the adjacent nodes to the selected nodes (excluding the selected nodes), a random node can be selected from each list. In this case, there is node A, node B, list $\mathbf{L}_A$ of nodes adjacent to node A, and list $\mathbf{L}_B$. Node A' can be selected from $\mathbf{L}_A$ at random, and likewise for node B'. The adjacency matrix for the new structure is adjusted such that A' is now adjacent to B and not adjacent to A, likewise B' becomes adjacent to A and not B. A simple illustration of this process in a freely mobile system of eight atoms is shown in Figure 4.7 (A' is atom 3, and B' is atom 4), and includes the subsequent structural relaxation.

In Figure 4.7 the atoms are labeled A and B and 1 through 6. A given adjacency list (that is, a single column of the adjacency matrix) would be given by,

$$\mathbf{L}_i = \begin{bmatrix} u_{iA} & u_{iB} & u_{i1} & u_{i2} & u_{i3} & u_{i4} & u_{i5} & u_{i6} \end{bmatrix}, \tag{4.2}$$

where u_{ij} is binary, and the values of the initial adjacency lists [Fig. 4.7(a)] are

$$\mathbf{L}_A^0 = \begin{bmatrix} 0 & 1 & 1 & 1 & \mathbf{1} & 0 & 0 & 0 \end{bmatrix}, \tag{4.3}$$

$$\mathbf{L}_B^0 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & \mathbf{1} & 1 & 1 \end{bmatrix}. \tag{4.4}$$

The values associated with the change in Figure 4.7(b) are bold. Following the swap, the adjacency lists are updated such that

$$\mathbf{L}_A^f = \begin{bmatrix} 0 & 1 & 1 & 1 & 0 & \mathbf{1} & 0 & 0 \end{bmatrix}, \tag{4.5}$$

$$\mathbf{L}_B^f = \begin{bmatrix} 1 & 0 & 0 & 0 & \mathbf{1} & 0 & 1 & 1 \end{bmatrix}. \tag{4.6}$$

For a binary system, the process of a bond swap is the same, except in this case each adjacent metal node has a coordinate and adjacency associated with a linker atom. For the **nbo** net decorated as PdOS, a single adjacency matrix is used in order to simplify the internal structural relaxation. With a single metal node A selected, an adjacent metal node can be found by first searching for an adjacent linker (chalcogenide) node X. As the linker node is only adjacent to two metals, there is only one option for an adjacent metal node B. With the A–X–B bridge selected, the bond swap proceeds by choosing other linkers, A' and B', that are adjacent to A and B. This is where the PdSO nomenclature is implemented: if the first selected node for the swap, A', is a sulfur atom, the second selected node for the swap, B', must be a sulfur atom; the same condition holds for oxygen selection. This restriction is unique to the anisotropic WWW, and enforces the geometry by ensuring there are two straight angles and four right angles to relax after each swap. This proceeds exactly as shown in Figure 4.7, except along each white connection there is a subliminal node X.

In profiling the overall program, it is apparent that the majority of the time is spent in the structural relaxation. It raises the question, can we recognize bad moves early without a complete MC simulation? An effort at this is attempted in

the BM method [228, 245]. Here, a set of topological constraints is implemented to prohibit unrealistic ring statistics: any move which creates a topological ring that only contains three unique metal centers is rejected without consideration. The formation of these rings would introduce significant strain in the system, and would likely be rejected after the structural relaxation. The tracking of individual rings is far simpler than overall ring statistic calculations. The metal nodes of interest, node A and B above, can be used to compile lists of next neighbors. That list can then be iterated over to compile a new list of second order neighbors, careful to keep record of the directionality of connectivity (as the adjacency matrix creates a non-directional graph, A would always be considered a second order neighbor to itself if the search was not directional). Finally, if node A or B are in this list of second order neighbors, a three-member ring exits.

Prior to a bond swap, a copy of the structure is made. This way, rejecting the move and reverting the structure does not require the costly structural relaxation. This optimization is carried out as a standard MC simulation where the moves are atomic displacements of the linker (O or S) and node (Pd) atoms. Simulated annealing is used with a Boltzmann like acceptance criterion to allow for some structural rearrangement and avoid local minima. It isn't expected that a ground state energy would always be reached for each new topology; however, a decent estimate can be approached and a fixed number of sweeps is used for the exit condition.

The energetics of the structural relaxation are dictated by a set of harmonic potentials. A metal–chalcogenide bond length potential could take one of two forms: one that fixes the equilibrium bond lengths by geometric arguments [eq. (2.10)] and allows for flexibility by increasing temperature, or one that allows flexibility by treating the equilibrium bond length as the expectation value of the bond length.

$$r_0 = \langle r \rangle \tag{4.7}$$

The latter potential has a similar operating principle as INVERT [309]. Considering that the box could be scaled to an arbitrary size the important result of the simulation would be a structure with a narrow distribution of bond lengths. On the contrary, the recalculation of the $\langle r \rangle$ is $\mathcal{O}(N)$ at each Monte Carlo move, whereas the energetic changes for the standard two body potential scale as $\mathcal{O}(1)$. Therefore, the first metal–metal peak in the PDF was used to calculate the metal–linker distance assuming the equilibrium angle, and the former two-body potential is used. The starting topology is scaled to match an Si–O distance from a-SiO_2, and subsequently rescaled at the close of the simulation to match this PDF derived distance restraint.

The local geometry of the nodes is fixed by a set of harmonic bond angle potentials with

$$\theta_0 = \begin{cases} 180^\circ, & \text{if X–Pd–X} \\ 90^\circ, & \text{if X–Pd–Y} \end{cases}. \tag{4.8}$$

As described above, the utility of two chalcogenide linkers allows for making the distinction between the straight and right angles by atomic symbol. Likewise, the local geometry of the linkers has an associated harmonic bond angle potential, with $\alpha_0 = 144^\circ$. In the implementation, the three-body harmonic potentials is cast in terms of the cosines, thus our overall potential resembles a simplified Keating potential [252],

$$V = \sum_{ij \in \text{bonds}} k_{\text{two}} \left(r_{ij} - r_0\right)^2 + \sum_{i \in \text{nodes}} k_\theta (\cos\theta_i - \cos\theta_0)^2 + \sum_{i \in \text{linkers}} k_\alpha (\cos\alpha_i - \cos\alpha_0)^2. \tag{4.9}$$

With just these three harmonic potentials, the structural relaxation will accommodate unrealistic atomic distances when treating atoms that are not adjacent. Therefore, a repulsive potential is implemented to prevent atom overlap during the internal MC. We took two approaches to this problem. First, we considered hard

spheres potential that treats atoms like 'billiard balls' with a fixed radius, and will not allow MC moves that create overlapping spheres [eq. (2.16)]. Although it contains a delta function, it can be implemented with a finite force constant to produce 'soft' spheres that can overlap with an energy penalty. As flexible repulsion was considerably more effective, a Lennard-Jones type repulsion (2.13) was eventually used instead of hard spheres, and is implemented in the results. The potential only includes the repulsive term of the Lennard-Jones potential, acts between atoms within a cut-off distance, and ignores topological neighbors to the second degree,

$$V_{\mathrm{r}}(r) = \begin{cases} k_{\mathrm{r}} \sum\limits_{i,j \in \mathrm{atoms}} \left(\frac{1}{r_{ij}}\right)^{12}, & r_{ij} < r_{\mathrm{r}} \\ 0, & r_{ij} \geq r_{\mathrm{r}} \end{cases} . \tag{4.10}$$

This give a rapidly decaying repulsive potential for atoms within a distance of closest approach, but offers some flexibility near the cut-off.

As per the program design, there are a set of parameters that are tuned. The supercell dimensions are bounded by the relvant range of the PDF. For the internal MC, adjustable parameters include (1) the maximum move size, (2) the number of MC sweeps, (3) the starting temperature, (4) the final temperature, (5) the harmonic force constants k_{two}, k_θ, and k_α, and (6) the repulsive force constants and (7) cut-offs. The external bond swapping MC has fewer parameters: (1) the number of moves, (2) the starting temperature, and (3) the thermostat type. In simulated annealing the temperature will generally be allowed to approach an effective zero. In the adaptive thermostat, the (4) fractional change ξ is self correcting over most of parameter space, and (5) η is an adjustable parameter.

This process — using the program provided in Appendix F — completes 1000 bond swaps in a 1152 atom system, with 200 internal MC sweeps, on a single logical core of a 3.06 GHz Intel Xeon Hex processor, in approximately 30 minutes.

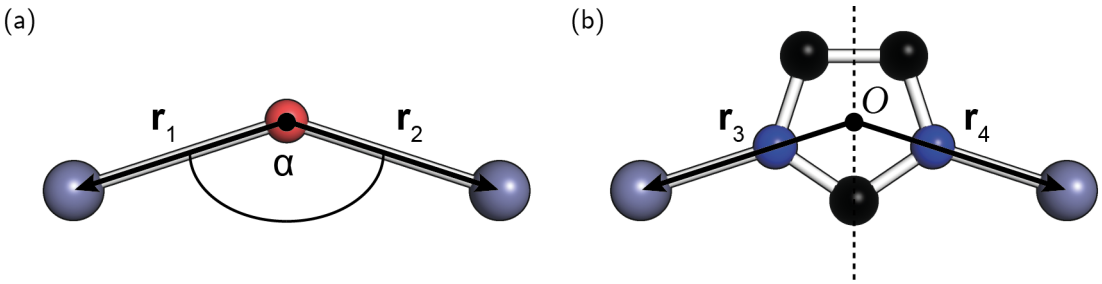


Figure 4.8: Representation of decorating PdSO structure with imidazolate ions. The vectors from a linker chalcogenide to the metal center, $\mathbf{r}_1$ and $\mathbf{r}_2$, are used to define a plane in which the imidazolate will lie.

4.2.4 Decoration and refinement

Once a sensible topology of PdSO is constructed, the linker chalcogenide atoms need to be replaced with imidazolate ions for a complete structure. Although the PDF will be dominated by the metal cations, a complete $Pd(im)_2$ structure is needed for general relaxation in RMCProfile. The process of converting the structure proceeds analogous to the geometric relation between *a*-Si and *a*-ZIF [Fig 4.1]. A linker plane is defined by the unit normal, $\hat{\mathbf{n}}_l$, from the normalized cross product of two vectors $\mathbf{r}_1$ and $\mathbf{r}_2$ [Fig. 4.8]. A second plane is constructed from a reference Pd–(im)–Pd, which can be thought of as a regular pentagon with an origin, O, such that $\angle O = 144°$. Likewise this plane can be defined by a unit normal, $\hat{\mathbf{n}}_p = \mathbf{r}_3 \times \mathbf{r}_4$.

In the WWW result, there is a distribution of bond angles centered about ∼144°. As a first approximation, this distribution is considered in the flexibility of the C–N–Pd angle, and the imidazolate is left rigid across linker sites. The decoration is then completed with relatively facile linear algebra. Firstly, the plane defined by $\hat{\mathbf{n}}_p$ is rotated to align with the plane defined by $\hat{\mathbf{n}}_l$ elsewhere in coordinate space, thus having a torsional angle of 0°. The axis of rotation $\hat{\mathbf{n}}_a$, and angle of rotation θ are

defined as follows:

$$\hat{\mathbf{n}}_a = \hat{\mathbf{n}}_p \times \hat{\mathbf{n}}_l, \tag{4.11}$$

$$\det = \hat{\mathbf{n}}_a \cdot (\hat{\mathbf{n}}_p \times \hat{\mathbf{n}}_l), \tag{4.12}$$

$$\theta = \texttt{atan2}(\det, \hat{\mathbf{n}}_p \cdot \hat{\mathbf{n}}_l). \tag{4.13}$$

The `atan2` function is a standard library arctangent function that returns a value for θ that is the clockwise angle. With a value for $\hat{\mathbf{n}}_a$ and θ, a rotation matrix can be constructed that will rotate the vector from the origin to any atomic coordinates to the appropriate place in space.

The Rodrigues rotation formula (4.14), is thus used to transform all the atomic coordinates in the reference imidazolate to a new imidazolate that lies in the same plane as the linker:

$$\mathbf{v}_{\text{rot}} = \mathbf{v}\cos\theta + (\hat{\mathbf{n}}_a \times \mathbf{v})\sin\theta + \hat{\mathbf{n}}_a(\hat{\mathbf{n}}_a \cdot \mathbf{v})(1 - \cos\theta), \tag{4.14}$$

where $\mathbf{v}_{\text{rot}}$ is the vector $\mathbf{v}$ following rotation by the matrix. The same formula is applied a second time; however, in this case it is rotating the two planes so their orientation in two-dimensions is aligned. In the second iteration, the vector $\hat{\mathbf{n}}_l$ is directed from the linker coordinates to the midpoint between the two Pd atoms, and $\hat{\mathbf{n}}_p$ is directed from the pentagon origin to the carbon bonded to two nitrogens. Finally, a translation moves the origin of the pentagon to coincide with the linker coordinates, and the adjacency matrix is updated to remove the chalcogenide and complete the Pd–(im)–Pd subunit. This process is repeated for every chalcogenide linker in the structure in a Python script that converts the output `.pdb` files of the WWW simulation (PdSO) to $Pd(im)_2$ [Appendix F].

The Python script that completes the decoration also outputs a set of files for use in the program RMCProfile. Refinement against real and reciprocal space X-ray data and was completed in RMCProfile version 6.5.2 using two- and three-body

harmonic restraints to maintain the topology invoked in the WWW model.

4.2.5 Total scattering

The X-ray total scattering data was collected by Andrew Cairns using the I12 instrument at Diamond Light Source, using high-enery radiation, Si (1 1 1) cryo-cooled double crystal bent Laue monochromator, and a large 2D area diffraction detector (Thales Pixium RF4343). A beam of wavelength $\lambda = 0.145\,77\,\text{Å}$, was used with a cylindrical sample geometry in a 1 mm SiO_2 capillary. The experimental Q was in the range 0.4 to 22.5 Å^{-1}. Data from an empty silica capillary were also collected using the same diffractometer settings for background subtraction. Corrections for background, multiple scattering, container scattering, Compton scattering, absorption, and Bremsstrahlung effects were applied using the GUDRUN program. The corrected data were placed on an absolute scale using a procedure derived from the Krogh-Moe method which uses a single-atom scattering sharpening factor to emphasise scattering amplitudes for increasing Q. The Fourier transform yielded the PDF. This conversion was accomplished using the program GUDRUN. It should be noted that the low-r conversion produces negative peaks in this data set due to assumptions regarding form factor, and the practical challenge of knowing the true number density of an amorphous sample. This will not impact the understanding from the final total scattering refinement, as the low-r data are the least relevant.

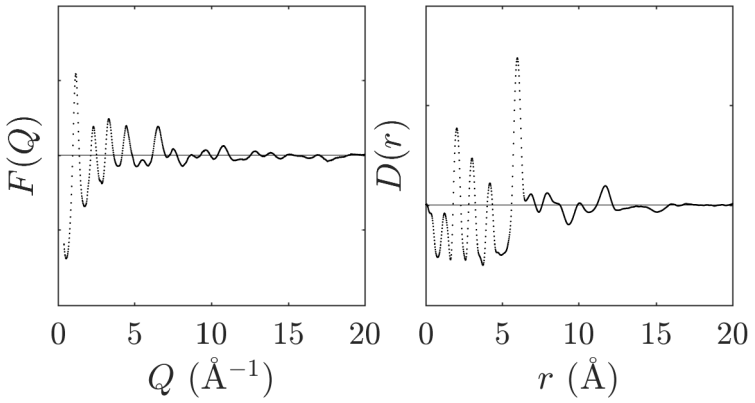


Figure 4.9: Total scattering data for Pd(im)$_2$. (a) The normalized total scattering data $F(Q)$, and (b) the processed PDF, $D(r)$, from GUDRUN.

4.3 Results and discussion

In our quest to understand the structure of amorphous Pd/Pt(im)$_2$, we come across three novel results. Firstly, we have developed and parameterized a method to generate continuous random networks of a D_{4h} subunit. Secondly, we have validated the consistency of that network with the standards of statistical variation expected from known amorphous materials. Finally, we develop the first ever structural model of Pd(im)$_2$ (and thus Pt(im)$_2$) that is consistent with total scattering data.

4.3.1 Parameterization

Our first task was in parameterizing the simulation to produce CRNs of a uniform density, with low variation in local geometry. This was accomplished by addressing the internal structural relaxation, followed by the bond switching algorithm.

In generating the square-planar CRN, we are not modeling an existing material. As such, the values of the parameters don't have the same physical significance as with modeling silica. Nonetheless, the insight from silica comes in the comparative rigidity of the bond, node geometry, and linker geometry. Parameters similar to those in [252] provided a starting point, with an increase in the three-body potentials to force more structure about the metal center.

The internal thermostat and simulation length was set to balance computational

expense and the need for near — not complete — structural relaxation. Because the initial topology satisfies the harmonic potentials as a minimum, an effective internal MC would be able to relax any structure with the same topology as a modulated initial structure. During testing of the internal MC, we set the linker bond angle $\alpha = 180°$, such that the input structure was a global minimum. The parameters found in Table 4.1 were generated for simulations without a bond swap, that partially melted the structure and were reproducibly relaxed back to the initial minimum. Thus, at any given bond swap, significant mobility is introduced into the system to potentially overcome any internal stress caused by the bond swap. In order to allow for scaling between ensembles of different size, the MC time parameter is measured in sweeps*. A relaxation that is consistent for 100 atoms will be consistent for 1000 atoms, but will take approximately 10 times as long.

With a functional internal structural relaxation, the WWW simulation was explored over logarithmic steps in parameter space using simulated annealing. An effective zero temperature was initially found that prohibited any 'bad' bond swaps. The length of the simulation time and initial randomization needed to be well balanced. The initial randomization came in two forms: (1) The WWW simulation was run for $N/2$ moves to create a 'melted' topology, or (2) the initial temperature was set high enough that the early period of SA explored a large region of phase space. A consequence of using simulated annealing from an initially crystalline structure is a potential memory of the initial state. A network that does not become sufficiently random will return to the crystalline state upon annealing; however, a network that is initially too strained will require long periods of time to explore phase space and decrease in energy [244]. Three orders of magnitude of initial temperatures, $k_{\mathrm{B}}T_0$, were explored, while monitoring the energetics over time, and the final structure statistics, to arrive at the parameter set used. This is a representative parameter

*Contrary to the internal MC relaxaxation, the WWW moves were left programmed as moves, due of the expensive nature of each bond swap. Consistency between ensembles of different sizes was manually implemented.

Table 4.1: WWW simulations parameters. A set of characteristic parameters used in the bond swapping algorithm for a 3×3×3 supercell cell of the NbO structure (486 atoms): The relevance of each symbol is described in 4.2. Values for both WWW thermostats are included, where the adaptive temperature scheme was updated on a 500 move basis. The maximum move distance for each internal MC step is listed as $d_{\max}$.

Parameter	Value	Units
Structural Relaxation		
k_{b}	30.0	eV Å^{-2}
r_0	1.6	Å
k_θ	10.0	eV
k_α	10.0	eV
α	144	deg
k_{r}	200.0	eV Å^{-12}
r_{r}	1.8	Å
$k_{\mathrm{B}}T_0$	$10^{-1.0}$	eV
$k_{\mathrm{B}}T_f$	$10^{-5.0}$	eV
$d_{\max}$	1.5	Å
Sweeps	$10^{2.3}$	
Bond Swapping		
$k_{\mathrm{B}}T_0$	$10^{0.8}$	eV
$k_{\mathrm{B}}T_f$	$10^{-1.0}$	eV
η	$[1, 2]$	%
ξ	$\frac{4}{3}$	
Moves	$10^{5.0}$	

set, with reasonable flexibility.

The common approach to exploring phase space during minimization is to increase $k_{\mathrm{B}}T_0$ and run the simulation for a longer period. In creating a square-planar CRN, we were not looking for a true minimum (as these are not always meaningful [249]), but a network with sufficient topological disorder to reproduce the total scattering. To gain efficiency in this period of encouraging entropy, we also implemented an adaptive thermostat.

These three approaches to the WWW simulation are shown in Figure 4.10. The initial randomization statistically alters only half of the structure, yet creates significantly more strain in this short time than the SA alone. The consequence is that any early changes to the randomized network will act to reduce the energy, but the

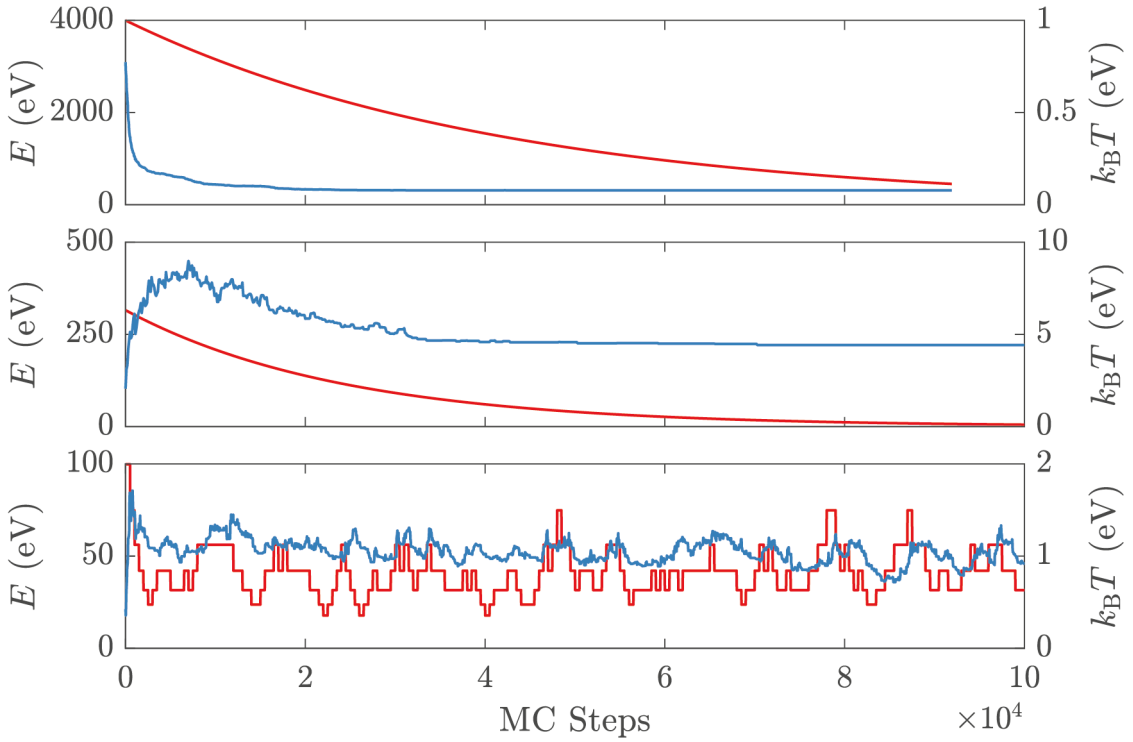


Figure 4.10: WWW results from three different thermostats. In the graph, the overall energetics are shown (blue) with the temperature (red) over 90,000 bond swaps for the (top) simulated annealing of a melt, (middle) simulated annealing, and (bottom) adaptive thermostats.

network will freeze into the earliest minimum it finds. To overcome this, an infinite time would need to be spent annealing. Whereas in the case of simulated annealing with a sufficient initial temperature, the early steps of the simulation will act to increase the overall energy, and allow for an exploration of phase space nearer to the starting crystalline structure. This can be judged by the variation in the energy profiles. A rough profile indicates moves are being accepted that change the energy, that is, the simulation is moving throughout phase space.

This roughness is most apparent when using an adaptive temperature scheme, which effectively forces the simulation to explore new configurations. As is clear in Figure 4.10, once a topological minimum is found in the simulation, the structure will inevitably change as the temperature grows. A sensible value of η will limit how fast this occurs, as has been found for other applications [413].

We also fed results from the adaptive thermostat simulation back through the

Table 4.2: Square-planar CRN statistics. Shown are expectation value and standard deviation of r_b (Pd–X bond length), θ_{90} (X–Pd–Y bond angle), θ_{180} (X–Pd–X bond angle), and α (Pd–X–Pd bond angle) in the scaled CRNs produced by the three different thermostats: Simulated annealing with a melt, simulated annealing without a melt, and the adaptive thermostat. The angular distributions are weighted by $1/\sin\theta$. These statistics correspond to the structures of Figure 4.11 and simulations of Figure 4.10.

	$\langle r_\mathrm{b}\rangle$	σ_r	$\langle\theta_{90}\rangle$	σ_{90}	$\langle\theta_{180}\rangle$	σ_{180}	$\langle\alpha\rangle$	σ_α
Melt+SA	3.25	0.04	94.7	21.2	140.2	23.6	138.2	23.8
SA	3.18	0.02	92.3	10.43	155.8	14.4	149.5	14.2
Adaptive thermostat	3.23	0.03	92.6	14.9	150.6	15.4	136.6	19.6

WWW under mild simulated annealing conditions. From entropic considerations, it was not likely that the global minimum will be found, neither was it desirable, as that would remove the topological disorder from our system. Combined, the two steps act similarly to an initial randomization (as considered in other CRN models [252]) and subsequent minimization. Nonetheless, the second SA did not produce significant changes in the overall topology of the result.

4.3.2 CRN model validation

A first step in analyzing the sensibility of a CRN is to calculate the bond-angle distributions. Because this model is the first of its kind, we are only able to offer a qualitative comparison to the statistics of other network structures. A set of WWW results from the different thermostats is shown in Figure 4.11.

Considering the structures produced by the three methods, the distribution about the 90° angle is more narrow than that of the 180° angle in the Pd coordination [Table 4.2]. This is sensible given that distortions out of the plane can maintain the 90° angle, but inevitably reduces the 180° counterpart. The breadth of the bond-angle distribution at the linker atom is characteristic of the flexibility found in framework materials, and similar to that of SiO_2. The mean value of 144° is somewhat larger than equilibrium value set in the potential; however, this would be a sensible accommodation to the fixed density of the simulation.

From the three networks and their respective distributions, the differences in

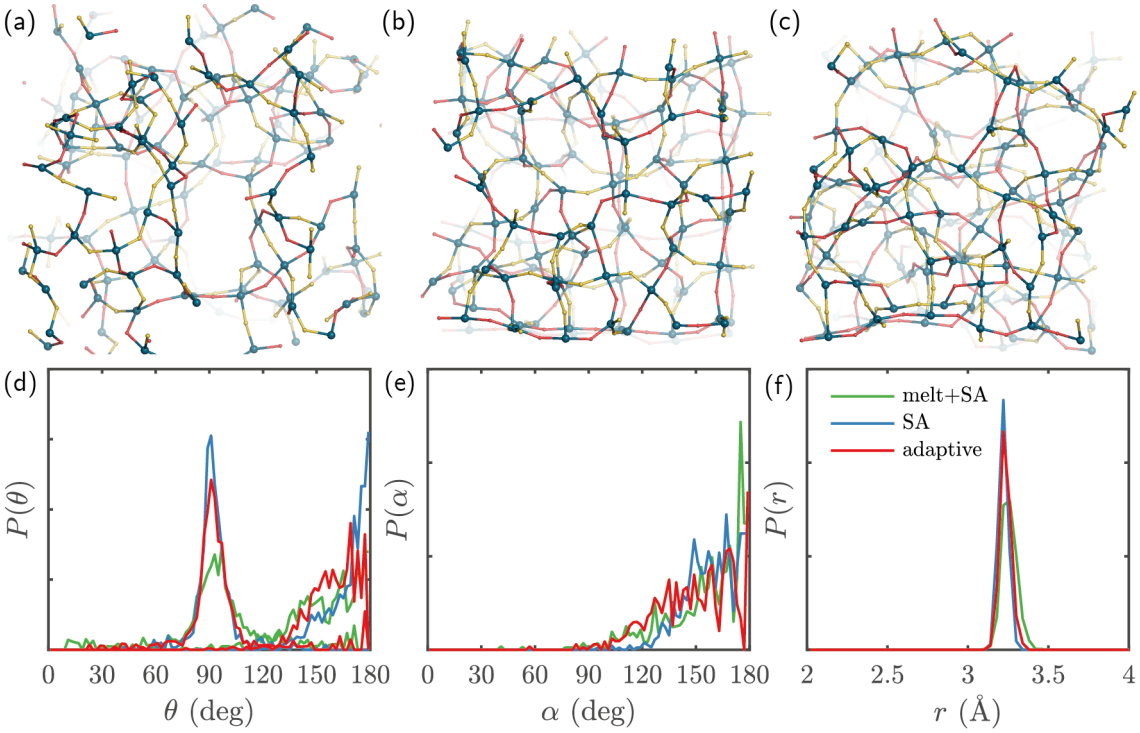


Figure 4.11: Effect of thermostat choice on WWW. A 3×3×3 **nbo** supercell (486 atoms, 162 metal centers) was used as input for the three available thermostat approaches. The structures from the (a) melt and simulated annealing, (b) only simulated annealing, and (c) adaptive thermostat are shown along the top row. Qualitatively, (b) shows significantly less disorder than the other two, with (a) containing highly strained clusters. Although the example from the (c) adaptive thermostat is not uniform, it shows similar geometric distributions to (b). Histograms of the bond angles at (d) the Pd and (e) the linker atom, and Pd–X bond lengths are shown in corresponding colors: melt and simulated annealing, red; simulated annealing, green; and adaptive thermostat, red.

topological and structural disorder are apparent. Without a well defined set of target distributions — as exists for silica — we searched for a range of topological entropy to be refined against the total scattering. Considering the highly unphysical structures resulting from the initial melt (and the computational requirements to mitigate them), we focused our attention on applying simulated annealing and an adaptive thermostat to an initially ordered network.

Significant variation can be introduced across structures from small changes in the initial temperature. Naturally, a simulation that spends longer time at higher temperatures will produce more disordered results. This is highlighted by a set of experiments in Figure. 4.12. When considering the geometric statistics of the

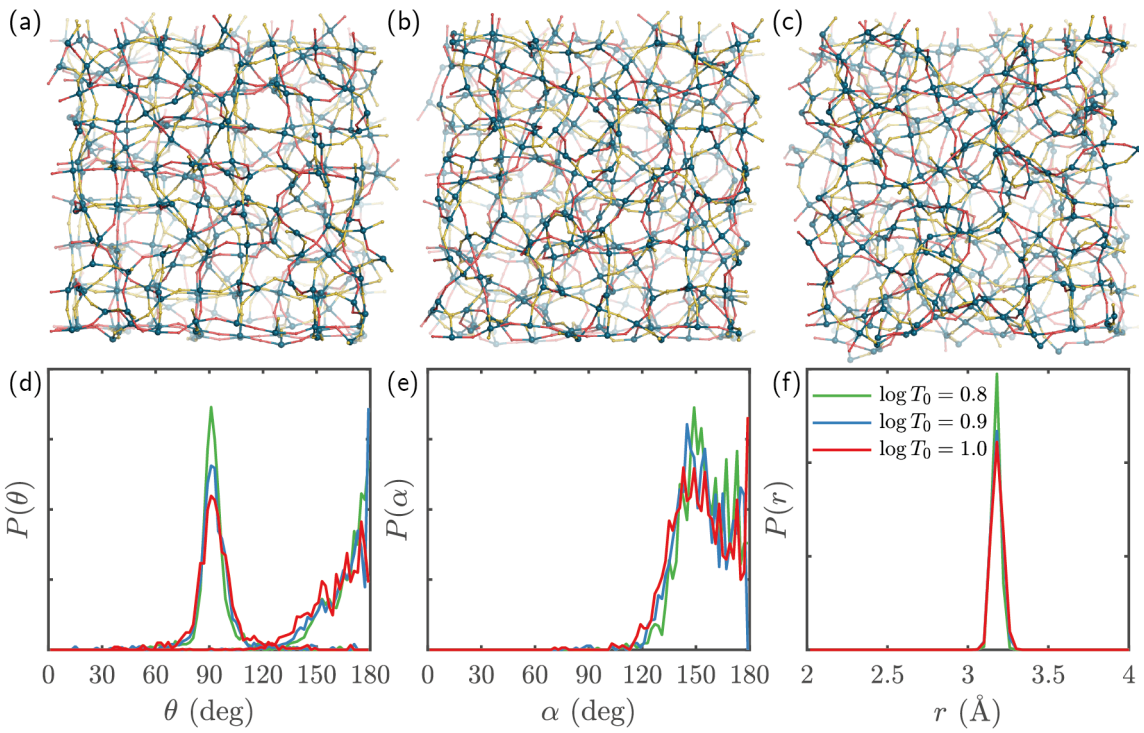


Figure 4.12: Comparing WWW simulations at different temperatures. A set of 4×4×4 unit cells at increasing initial temperature $k_{\mathrm{B}}T_0 =:$ (a) $10^{0.8}$ (b) $10^{0.9}$ (c) $10^{1.0}$ eV. The angular and bond length statistics (d)-(f) from the ensemble show limited variability between experiments. The distributions become broader, but there is much more variability in the topology between the three simulations.

the set, some features remain consistent across the distributions: (1) there are no secondary peaks indicating unfavorable or non-physical ring structures, (2) the bond-length distribution is sharp and centered about the equilibrium value, r_0, and (3) the angles associated with the squares maintain the 90° angle and reduce the 180° angle indicating an out of plane distortion, and (4) the bond-angle at the linker is shifted slightly higher than the set equilibrium value.

Even though SA is capable of introducing topological disorder into the network, after an initial exploration of configurational space, the algorithm inevitably remains trapped in a local minimum. As seen elsewhere [268], multiple WWW simulations at the same conditions can produce starkly different results. Furthermore, our objective was to produce a starting topology to decorate and refine against total scattering data. Considering the potential inconsistencies of SA, we chose to move forward

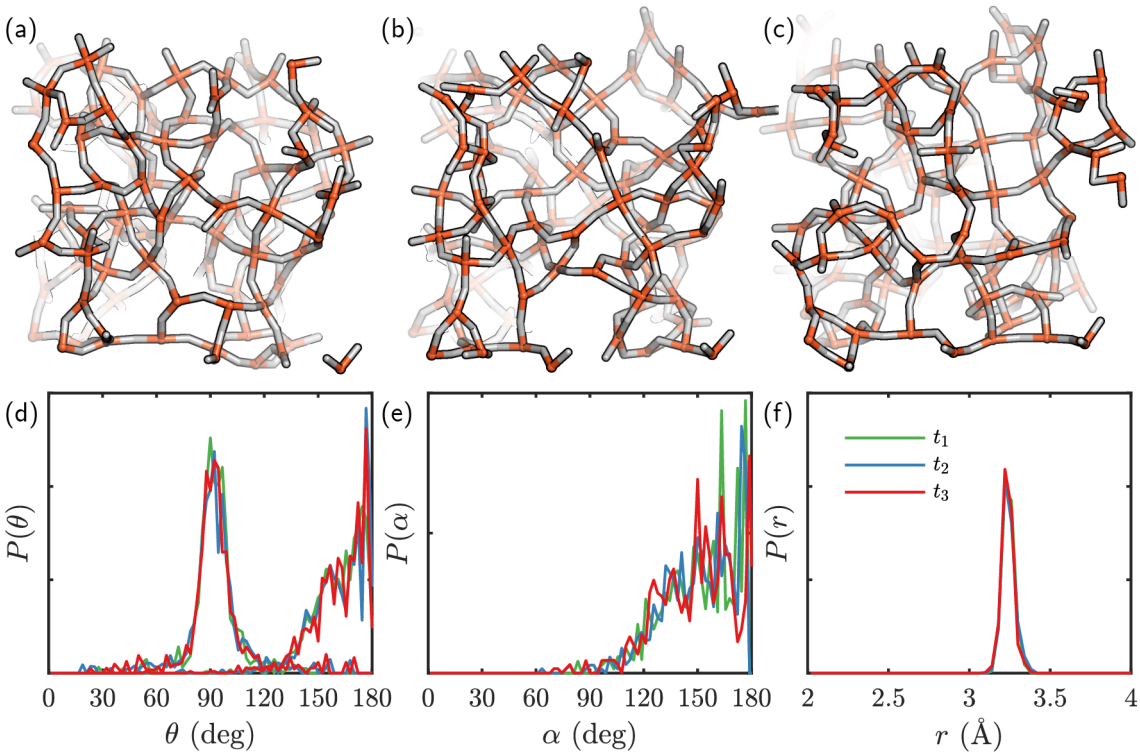


Figure 4.13: Sampling from an adaptive thermostat WWW. (a)-(c) A set of structures produced by a single WWW run with an adaptive thermostat. The topology becomes increasingly disordered with time. The angular and bond length statistics (d)-(f) from the ensemble show the variability between samples. From this sampling, we could select sensible structures for decoration and refinement.

using the adaptive displacement approach with the temperature regime found successful for SA. This approach created similar structures to the SA, but instead of freezing into a single topology, many could be sampled [Fig. 4.13].

Although the distributions of Figure 4.13 have well-distributed local geometry, a CRN should also be uniform over the long range. In order to accommodate distortions in the network topology, unphysical voids can emerge in the structure. This produces a nonuniform density, that would be observed in the simulated total scattering pattern as a divergence or multiple peaks at low-Q [App. B]. The resulting nonuniform density can also be viewed as a 'memory' of the initial configuration: the considerably open **nbo** net.

We accounted for these voids by revisiting one of the initial assumptions in our model. The density of the CRN [and subsequently the model $Pd(im)_2$], was

assumed during the scaling of the **nbo** network: by analogy to the relationship between β-cristobalite and silica glass, we set the initial bond length of the crystalline starting configuration to be similar to the expected equilibrium bond length for the amorphous structure. While this was suitable for silica, because of the physical relationship between the two phases, it proved unsuitable here.

We examined a set of initial densities ranging from those above, to the number density of crystalline $Pd(Him)_2(im)_2$. Using the same WWW parameters we created dense CRNs that were uniform, and did not manifest unphysical voids or crystallinity in the structure. This discriminating characterization utilized RMC refinement techniques, which we we will come to demonstrate in the next section, as a means of rapid screening. Through the use of a modified WWW simulation — equipped with an adaptive thermostat, and initialized from an **nbo** net with a number density related to the crystal phase of $Pd(Him)_2(im)_2$ — we have generated the first ever square-planar continuous random network [Fig. 4.14].

4.3.3 Total scattering refinement

Given the sensible structure and geometric statistics produced by the square-planar CRN model, we sought to describe the structure of amorphous $Pd(im)_2$ using this topology. Proceeding to decoration and refinement, the RMC box measured $49.18\,\text{Å}\times49.18\,\text{Å}\times49.18\,\text{Å}$ and had a density of $0.082\,50\,\text{Å}^{-3}$. During the refinement, no changes to the topology were made, and harmonic potentials were applied to restrain the local geometry. A square-well restraint was also implemented to fix the maximum allowed Pd–N distance. We also decorated a previous model of an amorphous tetrahedral ZIF (512 metal centers for a total of 8,704 atoms) with Pd ions, and performed a refinement under the same conditions. These square-planar CRN models of $Pd(im)_2$ reproduced all key aspects of the experimental X-ray total scattering data [Fig. 4.15].

Comparing the two models, the square-planar and tetrahedral generated CRNs have similar profiles in the low-r region of the PDF. This is unsurprising, as these

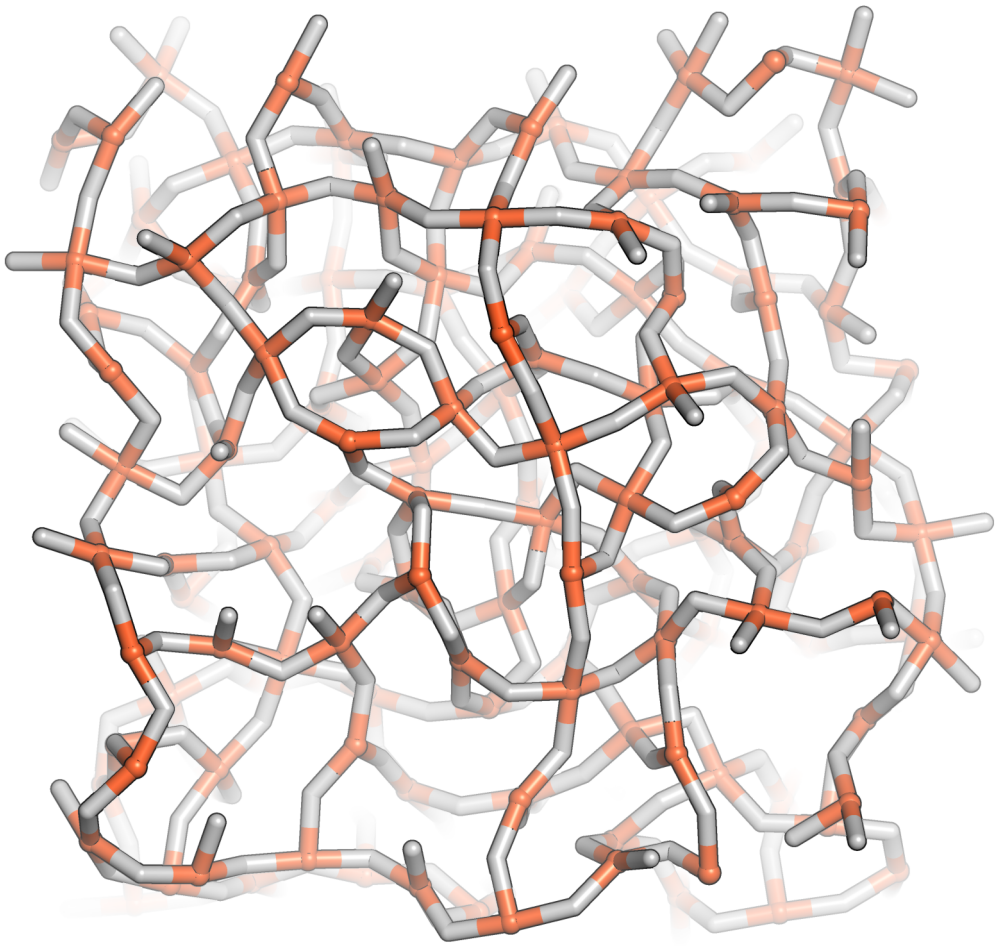


Figure 4.14: Slice of a square-planar CRN. A slice of the CRN as produced from the WWW simulation at $\rho = 0.0998\,\text{Å}^{-3}$. This network shows a sensible square geometry at the metal center, and uniform density. It was employed in the modeling of amorphous $Pd(im)_2$.

peaks are dependent on the imidazolate, which was decorated onto both structures by the same procedure. Beyond this molecular geometry, only the square-planar CRN is capable of reproducing the peak $r \approx 12\,\text{Å}$, and other mid-range structural modulations. It is important to note a key assumption contained within RMCProfile: We calculating the PDF of the model to compare to the experimental data, the program assumes that the atomic form factor of every atom in the box varies the same way with Q. This is a fair assumption when the atoms considered in a box are similar (*e.g.*, an organic structure composed of all period 2 elements, or a pure phase like *a*-Si). Because of the significant difference in scattering behavior between Pd/Pt atoms and the atoms of imidazolate, this assumption necessarily fails, and a

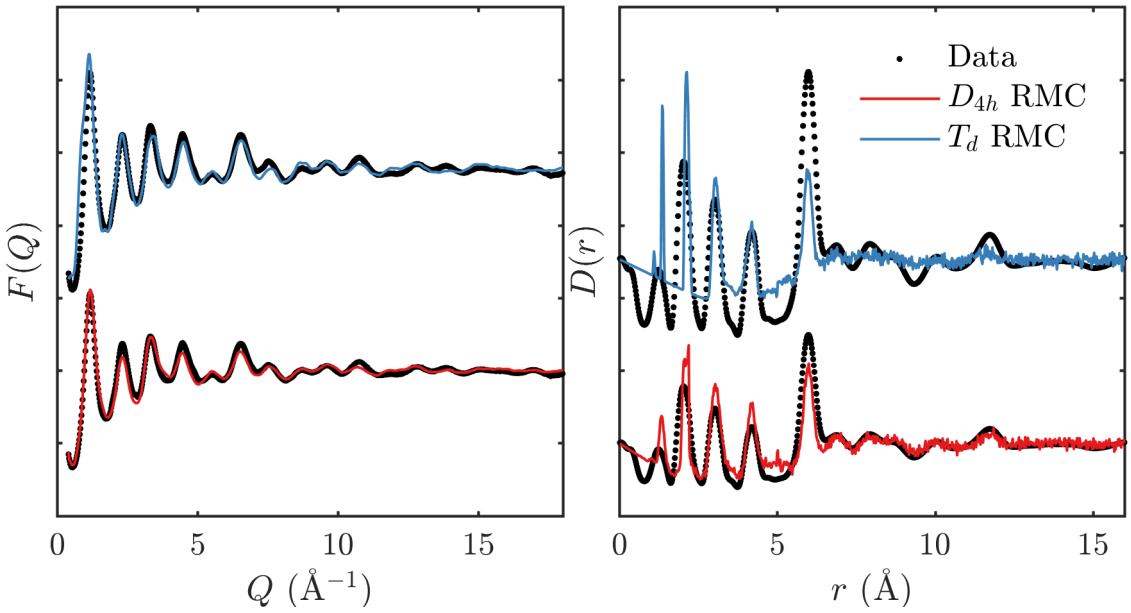


Figure 4.15: Comparison of Pd(im)$_2$ scattering data and fit. The $F(Q)$ (left) and $D(r)$ (right) measurements are shown in black circles, overlaid by the RMCProfile refinement of a square-planar CRN. The CRN was generated using an adaptive displacement WWW simulation. Offset is the RMCProfile refinement of a model based on the topology of *a*-ZIF (blue). The minor difference in density between the two models is accounted for in the PDF rescaling allowed during the refinement procedure, considering the general uncertainty associated with an amorphous material's density. Detailed refinements including residuals are offered in Appendix B.

precise fit the the PDF would not vindicate the structural model — in fact, it would likely cause some concern with the model.

The reciprocal space fits are even more distinctive, as this assumption regarding form factor Q-dependence is not made. Thus, the fit in Q-space is a more significant metric of model accuracy. Comparing the two decorations, the tetrahedral CRN model is incapable of fitting most features of the scattering profile, whereas the square-planar CRN model fits the entirety of the $F(Q)$ pattern.

The structural solution is presented in Figure 4.16, comparing the as-produced, decorated, and RMC refined square-planar CRN model. The geometric distributions at the metal validate our constrained RMC approach. In this we also highlight the impact of a range of model densities. These span from that of *a*-ZIF [28] to that of Pd(Him)$_2$(im)$_2$ corrected for hydrogen [406], with the structural model [Fig. 4.16(a)-(c)] density near the average of the two.

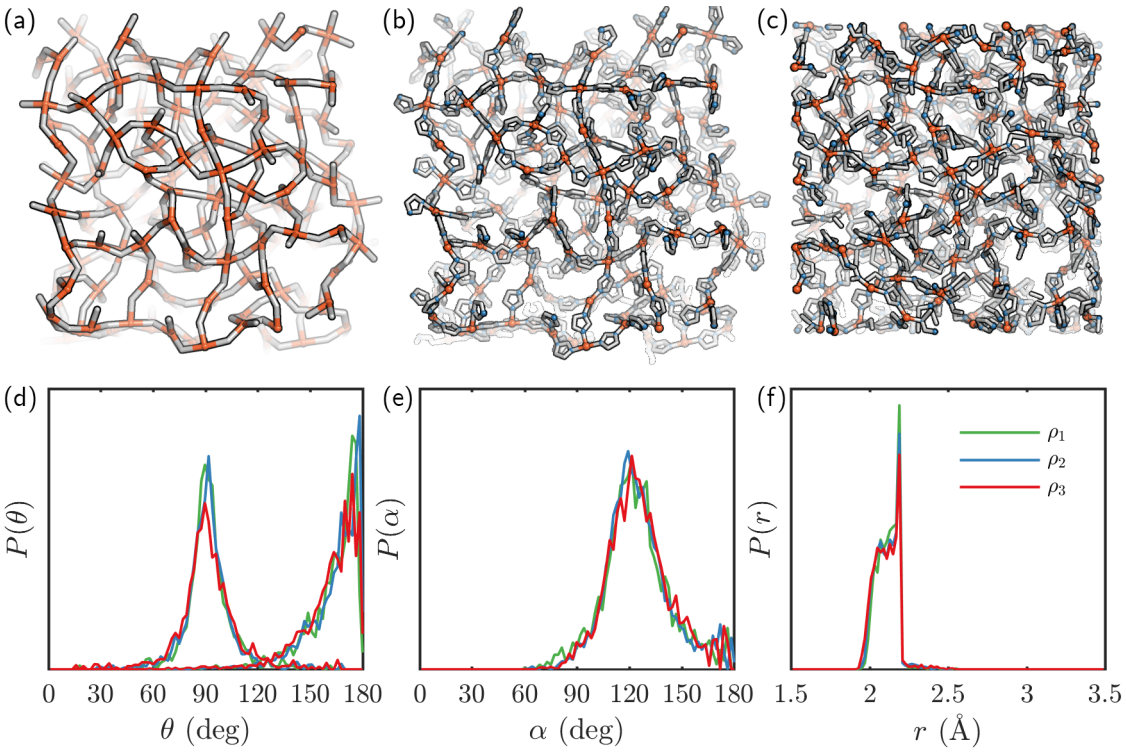


Figure 4.16: Structural solution of Pd(im)$_2$. (a) The CRN as produced from the WWW simulation at $\rho_3 = 0.0998\,\text{Å}^{-3}$; Pd atoms are orange, N atoms blue, and C atoms and connectivity grey. (b) After decoration, the Pd(im)$_2$ topology is refined against the total scattering data using RMCProfile. (c) The refined structure has narrow angular statistics, consistent with CRN models of tetrahedral materials. Histograms of the (d) the N–Pd–N angle, (e) Pd–N–C angle, and Pd–N bond lengths from the RMC refined configurations. The coloring corresponds to three densities: green, $\rho_1 = 0.0722\,\text{Å}^{-3}$ (*a*-ZIF density); blue, $\rho_2 = 0.0825\,\text{Å}^{-3}$ (mid-range density); and red, $\rho_3 = 0.0998\,\text{Å}^{-3}$ (corrected Pd(Him)$_2$(im)$_2$ density). The truncated shape of the (f) Pd–Pd bond distance peaks is a consequence of the square-well restraints placed on the RMC to prohibit unphysical changes in metal coordination.

Notably, the fit to total scattering data was comparable for structures of different density — insofar as that density was large enough to produce a uniform network from the open **nbo** structure — and the model itself had to be examined to ensure physical sensibility. The total scattering fits for the three densities are available in Appendix B. Models of the three structures and histograms of all bond lengths and angles are also presented there for reference.

The square-planar CRN approach to modeling amorphous Pd(im)$_2$ is robust across a range of densities and unique decorated networks. This demonstrates the importance of the novel square-planar CRN topology in understanding the material

structure. Moreover, these models create a snapshot approximating the average of a large disordered network, and as such are subject to moderate variations that account for the limited box size. The geometry of the metal center and imidazolate persist throughout the refinement with some expected and acceptable error related to the finite size of the simulation. Therefore, we have demonstrated strong evidence that $Pd(im)_2$ is related to a square-planar CRN described by Pd centers and imidazolate linkers.

The question remains whether or not we can be sure of the absolute topology of the material. While the total scattering remains sensitive to differences between D_{4h} and T_d CRNs, it not likely to differentiate between networks of equivalent local units with unique ring statistics. Moreover, a challenge of RMC modeling and using total scattering as structural validation was demonstrated in Figure 4.16: refined models producing different degrees of structural distortions had similar fits to the total scattering data.

This influence of box density on bond stretching is apparent in the structural models and geometric statistics [App. B]. A decrease in cell density will allow for a higher concentration of larger voids in the structure. To compensate for this, structural units will stretch toward the voids during the RMC refinement, acting to make the structure more uniform: a hallmark characteristic of amorphous materials. This behavior is not obvious from the approach, and further imparts the need for informed approaches to data-driven procedures.

This structural solution is particularly meaningful, in that it provides the foundations for studies of structure-property relationships in a CRN composed of anisotropic subunits. How local orientations of quadrapoles interact and order in MOF materials is an under-investigated subject [414]. The development of accurate structural models for real amorphous materials with quadrapolar building units opens up new areas for studying multipolar order in frameworks.

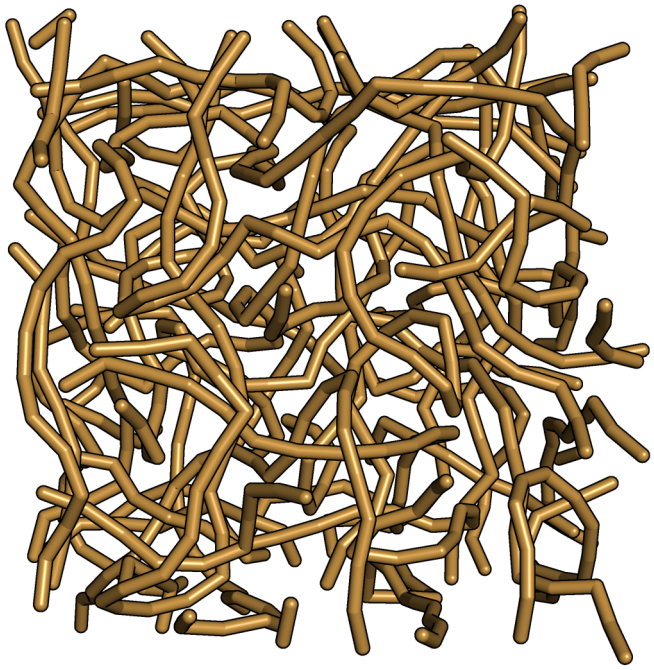


Figure 4.17: One-dimensional spaghetti like substructure in square-planar CRNs. An CRN produced from a 4×4×4 tiling of the **nbo** unit. The fictitious PdSO stoichiometry is left in place for the sake of visualizing the trajectory of like members in the final network. Similar to the topological disorder in ZrW_2O_8, this network can be described as following a set of rules. The two-in two-out rules are translated as oxygen linkers (left) being 'in' and sulfur linkers (right) being 'out'.

4.4 Conclusions and future work

In the above we were able to demonstrate a new type of continuous random network through the use of local rules and topological augmentations of a know square-planar net. Not only is this network the first of its kind, but the local anisotropy of the model allows for a visual analogy to the topologically disordered states in ZrW_2O_8. This interesting relation arises from different treatments of linkers coordinating the square-planar cation. Unlike tetrahedral networks, where all coordination angles are equivalent, the square-planar network construction forces a distinction between two types of linkers to enforce the right and straight angles (the fictitious PdSO structure). What results is a one-dimensional disordered trajectory similar to the spaghettis seen in $ZrW_2O_8{\cdot}H_2O$ [Fig. 4.17].

The metal coordination can be considered to follow 'two-in two-out' rules over the

underlying net. These are not conventional ice rules, since ice rules are applied to a fixed topology decorated with a subunit with a symmetry mismatch. The spaghetti in $ZrW_2O_8{\cdot}H_2O$ was an emergent topological feature. Here the local rules are always satisfied (by the WWW conditions), but the topology of the underlying network is altered. This observation raises the potential for defect states in square-planar amorphous materials to be studied in analogy to other one dimensional topological features, such as spin ices or Coulomb phases [103].

Beyond representing an interesting and unique topology, the square-planar CRN model was implicated in describing the amorphous structures of $Pd/Pt(im)_2$. Through the use of RMC refinement, we captured the local structure and topology of the amorphous network, with an approximation for the density. The three network densities also represent three unique CRN topologies (as they are from different WWW simulations) and demonstrate the robustness of the procedure.

The development of this distinctive topology creates new avenues for research in an already rich field. Future experimental work to extract the true density and angular statistics could be paired with more precise modeling of the structure, including a cell volume variation in the WWW algorithm. As RMC is technically a model independent refinement technique, further approaches could include model dependence with more physical significance. Molecular dynamics could also be implemented to relax the imidazolate ion (assumed to be a regular pentagon in this study), and provide more precise interaction potentials between non-bonded atoms. Nuclear magnetic resonance or IR studies to extract bond information would further inform the structural relaxation. As with *a*-ZIF, an accurate model would allow for calculation of electronic and vibrational properties [415]. Overall, we have opened up avenues for pursuing more diverse amorphous networks as they arise experimentally.

Framing this study in the scope of this thesis, we have further pressed the power of the one dimensional total scattering pattern. Despite the limited information content of the pattern, we developed a meaningful structural solution from a set of

local rules and informed principles. This approach is capable of discerning between the amorphous metal-organic networks formed by different symmetry subunits with the same coordination number. Armed with the prior information of the coordination geometry from crystal field theory, we will come to argue that we made an effectively Bayesian inference regarding the topology of a material. It is this inference that facilitates the extraction of accurate 3D structural information from the 1D pattern.

Chapter 5

Bayesian approach for information poor structure determination

Prior to this point, when faced with an information poor inverse problem, we compounded insight from multiple sources to gain confidence in our structural prediction. In contrast to this, we have seen the consequence of naïve RMC in fitting the most random (often nonsensical) structure to data. Throughout these informed modeling procedures, an implicit Bayesian approach was taken.

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} \tag{5.1}$$

Bayes' theorem relates the posterior probability of one condition given a second [$P(A \mid B)$], to the likelihood of the second given the first [$P(B \mid A)$], and the prior probabilities of each condition [$P(A)$ and $P(B)$]. According to Bayesian theory, a maximal posterior probability (or best estimate) comes from incorporating priors with the evidence [416]. As the evidence becomes more uncertain — defined peaks becoming diffuse — the optimal estimate would rely more heavily on the distribution of the priors than the evidence itself. This is precisely how our brains process information [417].

When we considered ZrW_2O_8 [Ch. 3], instead of presupposing a structure based solely on the scattering data, we implemented a set of priors in the form of a distribution of topological spaghetti states, themselves derived from informed symmetry and physical arguments. In the context of *a*-MOFs, we relied more heavily on the priors [Ch. 4]. Chemical knowledge and crystal field theory expected a square-planar metal center. This casts a distribution of acceptable underlying frameworks: the

priors indicated isotropic disorder and D_{4h} metal centers. In both of these cases, the structural predictions didn't just fit to the data, they were informed by prior probability distributions of relaxed topological states, under the guise of local rules.

These kinds of approaches are not unheard of in the field of structure determination. Contemporary approaches to RMC implement harmonic potentials within bonded systems or fixed polyhedral subunits in frameworks [335]. More elaborate empirical potentials have been used to refine structures with thermodynamic sensibility and extract relevant data (*e.g.*, internal energy, pressure, or phase behavior) [316, 418]. The INVERT approach depends on the prior expectation of similar local environments for similar atoms in a system [311]. Each of these modeling techniques exercise varied dependence on priors — often in the form of a Boltzmann distribution — when fitting to scattering data.

In the context of informed modeling, the incorporation of priors has acted to reduce the size of explorable phase space or configurational space. What can be considered 'implicitly Bayesian' is bounded by the influence of a probability distribution that limits the size of phase space in the modeling procedure. In the preceding chapters, we have considered use of local rules to cast a discrete distribution of topological space that was combined with a continuous energetic distribution, which limits the overall explored region of phase space in the inverse problem. Initially this was the most extreme case of not sampling topological states incommensurate with the spaghetti model, and only sampling energetic microstates from this subset. When combining RMC with added potentials for $Pd(im)_2$, the microstates sampled fall along a combination of the Boltzmann distribution and χ^2 distribution. Simply reducing the size of phase space is not necessarily Bayesian, as this would include any assumptions about the nature of a system in the definition; however, the implementation of a prior probability distribution — implicitly or explicitly — is.

This brings us back to the earliest question posed in this thesis: when are the

scattering data sufficiently discriminating to dictate a structural model? Even in ordered materials, homometric structures introduce information entropy [Fig. 1.6]. The degeneracy of unique structures with similar fits to the pair correlation offered by total scattering complicates the search for the ground state [419]. The roughness of the hypersurface associated with a fit-dependent penalty function only increases with increasing degeneracy [420]. Moreover, the Shannon entropy will only grow as scattering data becomes more diffuse. In the cases of such daunting information poverty, seeking out priors can elicit meaningful predictions. We have seen this at large in the RMCProfile [335] and EPSR [418] approaches, and in our study of ZrW_2O_8 and *a*-MOFs. However, we have yet to implement any continuous priors.

Considering this set of fundamental ideas, let us turn our attention to a case study that is unparalleled in information poverty, but possesses a set of robust prior statistics: the prediction of aqueous protein structure from the pair distribution function.

Here, I expand these ideas and the knowledge from ZrW_2O_8 and $Pd/Pt(im)_2$ to cast an explicitly Bayesian route to structure determination for information poor systems. In this chapter I will address a proof of concept of a program that produces coarse-grained models of protein structures using the pair distribution function and sparse nuclear magnetic resonance (NMR) data as input. The aim is to investigate whether the one-dimensional pair distribution function contains sufficient enough information to extract a three-dimensional macromolecular structure, given the available statistics. There is a parallel objective in producing a method for creating structural models of proteins that are difficult to crystallize.

To set out, I offer a review of the contemporary literature and the relevant general science. The chapter begins by explaining the terminology associated with proteins — insofar as they are relevant to structural studies — followed by examining the many techniques used to model protein structure with a particular focus on the nuclear magnetic resonance (NMR) based approaches implicated in this study.

From this base, the algorithm and associated program are presented along with the primary parameterization and results for a simple protein. The chapter concludes with a discussion of the approach and its impact on our parallel aims.

I would like to acknowledge William Fletcher for his insights into this project, as well as Prof. Lorna Smith for her assistance with protein NMR understanding. Though no molecular dynamics are included in this chapter, I would like to thank Prof. Jane Allison for her helpful discussions and assistance with MD. Lastly, I would like to thank Dr. Nicola Steinke and Dr. Sylvia McLain for the insight into the experimental outlook for this project.

5.1 Background

Immense effort in the crystallographic community alone goes into solving the structure of proteins, with X-ray crystallography being the primary method for atomic resolution structure determination of biological macromolecules [421] since its first application 50 years ago [422]. Structural biologists have a variety of methods at their disposal but the impetus for the field is clear: the three-dimensional conformation of a protein is directly related to its function, thus understanding a protein's structure is a prerequisite for understanding its role in biology, and how to harness, manipulate, or remove that role entirely [423]. Nonetheless, it is exceedingly difficult to determine some of the most desirable protein structures [4]. These techniques and their challenges will be discussed in more detail below, but it is important to note their failure outside the realm of water-soluble globular proteins.

Despite representing between 20% and 30% of all proteins [424], and 60% of all drug targets [425], membrane proteins account for less than 0.1% of the entries in the the Protein Data Bank (PDB). This is in part due to their incongruence with the methods of canonical crystallography. They cannot be expressed in high concentration, do not purify well, and can unfold outside of the membrane environment [426]. This particular environmental behavior accounts for much of the underrepresentation in the PDB. In order to address this deficiency, non-crystalline (thereby

non-periodic) routes to structural solutions need to be found.

Small-angle scattering (SAS) is an attractive alternative to traditional crystallography for dealing with solution-phase measurements [427]. Due to atomic motion and random orientations in a fluid, these measurements are subject to the same orientational averaging associated with powder diffraction. Effectively probing the low-Q region of reciprocal space, SAS is capable of resolving information between 10 and 250 Å [102]. Assuming no protein-protein interactions (the dilute monodisperse case), this experimental data is related to the Fourier transform of the intra-protein pair correlations at nanometer scale resolution [90]. These measurements have traditionally been considered insufficient for atomic resolution models, only for probing the overall shape of the protein.

In alignment with the previous work addressing condensed matter structural solution, the PDF — as obtained from SAS — is sought as a source of local, non-periodic, structural information. The content of a protein PDF appears far less instructive than in the cases of crystalline or glassy functions. In Figure 5.1 a clear comparison of PDFs is offered. In the case of a crystalline sample, the Bragg diffraction dominates the scattering pattern and produces a rich PDF at high r values, indicative of long range order. The scattering pattern of the glassy sample is mostly diffuse scattering, and produces a PDF that only shows local, short-range order. The protein PDF — at first glance — doesn't have the same wealth of information (it would be impossible to calculate a particular coordination number from a finite integral). I will show that that although this PDF isn't as information dense, it offers considerable insight into the overall shape of the macromolecule.

5.1.1 Protein formalism

Prior to further exploration of how protein structures are obtained, it is first necessary to introduce some terminology and functional details. There are four key levels to describing protein structure, each of which is intimately related with the other.

Firstly, there is the *primary structure*, or the sequence. The protein itself is a

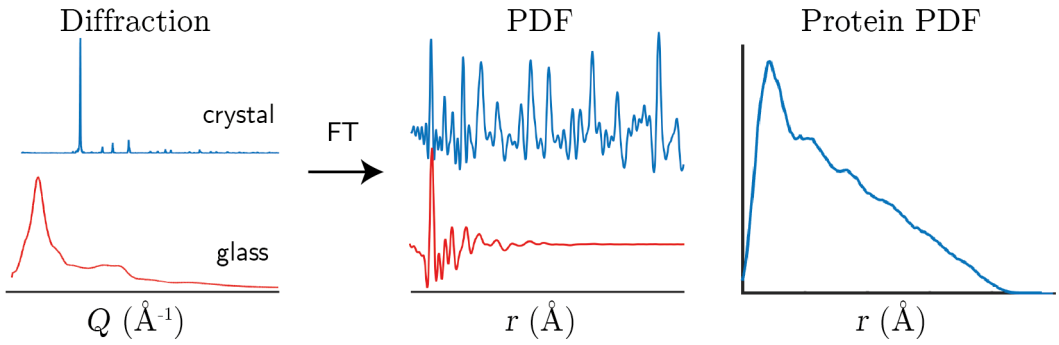


Figure 5.1: Comparing 1-dimensional PDFs across systems. The pair distribution function is a normalized Fourier transform of the total scattering pattern. Shown are simulated $S(Q)$ and $D(r)$ for β-crystabolite and amorphous SiO_2, and the simulated $G(r)$ for melittin.

polypeptide chain made up of 20 naturally occurring amino acids. The amino acid is assembled by the ribosome during *in vivo* synthesis by the creation of peptide bonds between two amino acid *residues**. The intricate processes of transcription and translation depend on the genetic code [428], although it is possible to post-modify a protein *via* hydrolysis [429], or chemical alteration of the sidechains. Primary structure is readily determined by chemical [430] or nanopore [431] sequencing of the gene or protein itself, or by statistically-aided mass spectrometry [432]. The *sidechain* of the protein is the defining component of the amino acids, all of which possess the same backbone ($-C_{i-1}-N_i-C_{\alpha i}-C_i-N_{i+1}-$) [433]. It is here where the chemical uniqueness is found: varying polarity and charge distribution in different sidechains separate the residues of a protein into hydrophobic and hydrophilic subgroups, and impacting the local conformation and secondary structure of a protein.

The *secondary structure* of a protein is defined by the backbone dihedral, or torsion angles [Fig. 5.2]. Thanks to chemical principles behind the bonding of the repeat units, the majority of the variability comes down to these torsion angles. Given reference names, the ω angle is that about the peptide bond, the ϕ angle is that about the $N-C_{\alpha i}$, and the ψ angle is that about the $C_{\alpha i}-C$ bond. Due to the sp^2 hybridization associated with the peptide bond, the ω angle is always close to

*A residue is used to describe a unique instance of a given amino acid type in a protein.

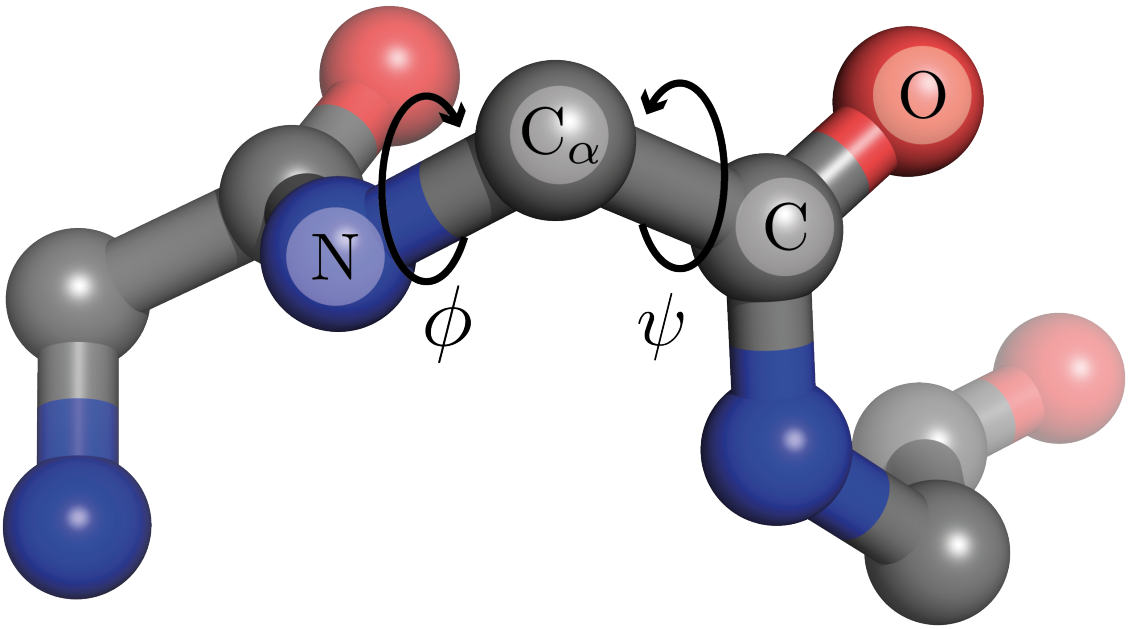


Figure 5.2: Protein backbone structure. A representative protein backbone segment is shown with the $-N_i-C_{\alpha i}-C_i-$ chain of a single amino acid labeled. The axes associated with the (ϕ, ψ) torsion angles are labeled.

planar and only varies between cis and trans bonding. In practice the majority of the peptide bonds are cis, and those proteins with all cis peptide bonds will be the focus of this study. As such, the majority of the secondary structure can be refined with only two degrees of freedom for each residue, (ϕ, ψ). Special recurrences of secondary structure have assigned names (*e.g.*, α-helix, π-helix, and β-sheet) characterized by their hydrogen bonding structure [434, 435]. This list is not exhaustive and each of these have subclasses of secondary structures [433]. These secondary structures can assemble into larger motifs, and in their entirety form the *tertiary structure*: the complete 3-dimensional descriptions of a single protein backbone strand. These motifs become an effective means of both visualizing a protein and characterizing its functionality. Solving the tertiary structure is the real goal of the protein problem [436]. At the highest level there is the *quaternary structure*, which is the arrangement of separate single polypeptide chain subunits that form the full functional protein. Examining quaternary structure is not the focus of this work; however, it would be feasible to examine by similar means. Within these four layers of classification,

there is a plethora of restraints to describe structure. There is covalent bonding at disulfide bridges [437], symmetry related oligomerization of subunits, flexibility constraints [438, 439], and the previously discussed chemical restraints associated with the peptide bonding [440, 441]. Most integral to this work are the statistical constraints and understanding developed by considering where amino acid dihedral angles exist in (ϕ, ψ) space.

Secondary structure is highly amenable to statistical analyses such as Ramachandran plots [442, 443], which consider where each residue lies in (ϕ, ψ) space. Population statistics have been developed and examined from the Protein Data Bank (PDB) to divide this space into occupied regions [444]. The Ramachandran space has several disallowed regions due to steric hindrance, and these plots have long been used as an assessment for the sensibility of a proposed protein structure [445, 446]. There are some exceptions to these rules which preclude any rigidity in assessment [447]. A broad-strokes characterization of dihedral angles is given in Table 5.1 with an example Ramachandran plot presented in Figure 5.3. Most work is focused around the backbone dihedrals for tertiary structure determination; however, a significant body of research has been dedicated to the sidechain torsion angle distribution as well [448]. This (ϕ, ψ) conformational space is integral to the Bayesian algorithm presented in this chapter.

With this understanding of protein structures as fixed chains determined by their respective sequence of torsion angles, it is worth examining the experimental and computational techniques currently used in solving tertiary structure. Each of these may approach the problem of structural determination with reference to different degrees of freedom, and each of them offer particular utility and insights.

5.1.2 Existing experimental techniques

X-ray crystallography has historically been the primary means of protein structure determination and proceeds *via* a refinement of a real space structural model against reciprocal space scattering data collected from a crystalline protein sample.

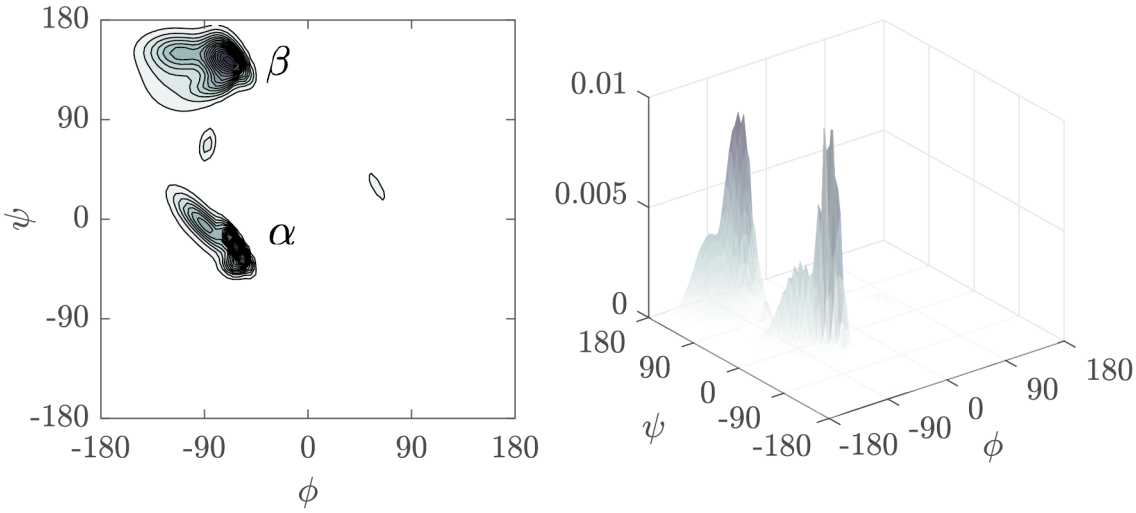


Figure 5.3: A typical Ramachandran plot. Shown is a surface plot and contour plot of the prior probability of a residue's dihedral angles with regions of α-helices, and β-sheets noted.

Naturally, these techniques depend on coercing a flexible organic polymer into a crystalline phase. Not only is this process exceedingly intensive for those subjects that do crystallize, but it is not necessarily indicative of their biologically native state. The experimental protocols for crystallization include genetic modification of an organism; gene expression that produces a correctly folded protein; purification of the protein following expression, and finally temperature and chemical potential control to allow for crystallization [449]. These samples also tend to have large unit cells and limited symmetry since a protein solution is a nonracemic mixture of chiral molecules.

Crystallization is an even more daunting challenge when considering membrane proteins. These proteins are extracted from the cell membrane using detergents, and are relatively flexible and unstable [450]. This leads to a host of challenges across the purification and structure solution pipeline. Even with high-throughput screens, this has served as a limiting step in the process of membrane protein structure determination. If and when the proteins are crystallized, there is a further consideration that the crystal is actually comprised of a detergent-protein complex, and not the protein in isolation [451].

Table 5.1: Ranges of (ϕ, ψ) values associated with particular secondary structures [444].

	ϕ	ψ
α-helix	$(-180°, 0°)$	$(0°, 45°)$
β-sheet	$(-180°, -45°)$	$(45°, 255°)$
turn	$(0°, 180°)$	$(-90°, 90°)$

Beyond the challenge of crystallization is the problem of beam damage. During a measurement, the 'soft' crystals associated with proteins are exposed to high energy X-ray radiation, which causes absorptive destruction of the sample [452]. Occurring by disrupting electronic structure or heat dissipation, this degradation of the sample precludes the use of long exposure times to enhance resolution. Therefore, high intensity beams — necessitating a synchrotron — must be used to increase resolution and the crystal must be large enough for the equipment available. This worsens the practical concerns associated with traditional crystallographic methods. Flash-freezing has been proposed as an effective means of minimizing beam damage; however, this pushes the crystal further away from the biological conditions desired in the probe [453].

To circumvent the need for large crystals, X-ray free-electron laser (XFEL) crystallography has been used recently on small mircrocrystals, less than 1 micrometer by 1 micrometer by 3 micrometers in size [454]. A novel means of producing traditional crystallographic profiles in reciprocal space includes rapid crystallization of 'soft crystals' of membrane bound proteins, and femtosecond laser exposure [57]. Because the laser can be pulsed at extreme rates, the data resolution problem can be solved not by a large crystal under high intensity or a cold crystal with long exposure, but by the rapid succession of many small crystals which are destroyed upon exposure [57, 455]. Even though the samples are disintegrated, the experiment only collects time resolved scattering data prior to destruction for damage-free observation [456]. These collective profiles reduce the need for long exposure times and

subsequent sample damage, yet are expensive to produce and require large quantities of sample. All in all there are only a few XFELs in existence.

A third — more recently well-established — method is the use of electron microscopy (EM) to access an electron density mapping of a protein sample [457, 458]. This method is especially useful for understanding quaternary structure. Initially facing challenges in resolving tertiary structure [459], more recent advancements in cryo-EM and image processing have allowed for 3–5 Å resolution. With a crystalline packing, an electron diffraction experiment can yield protein structures. Here again beam damage is encountered and the same signal-to-noise ratio concerns arise. This can be mitigated by freezing the target proteins in vitreous ice [460]. This freezing process needs to be accomplished rapidly in liquid ethane in order to maintain the amorphous water environment and native structure [461]. Aided by advanced computation methods for data processing (ensemble single-particle analysis), cryo-EM has determined atomic resolution structures [462, 463] [Fig. 5.4]. Notably, a mixture of cryo-elctron tomography, mass spectrometry, and biochemical and evolutionary *de novo* methods were used to understand the architecture of the macromolecular nuclear pore [464, 465]. With the development of better microscopes and detectors this procedure shows promise; however, it is still experimentally taxing, and a stronger probe of high symmetry and quaternary structure [466, 467].

Accessing the native biological conformation will ideally be done in the solution-phase. Biochemical NMR crystallography attempts to accomplish just this. Over the past 30 years, significant development in understanding the structure and dynamics of proteins through the use of NMR has been made [468, 469]. These structures are generally size limited ($<\sim$15 kDa) and require rich information from chemical shifts, nuclear Overhauser effects (NOE), and residual dipolar coupling (RDC) data [470]. Alternative uses of NMR to create distance or shape restraints for refinements against diffraction data or molecular force fields have also been applied with much success [471, 472]. It is this latter portion of study that is most intriguing with

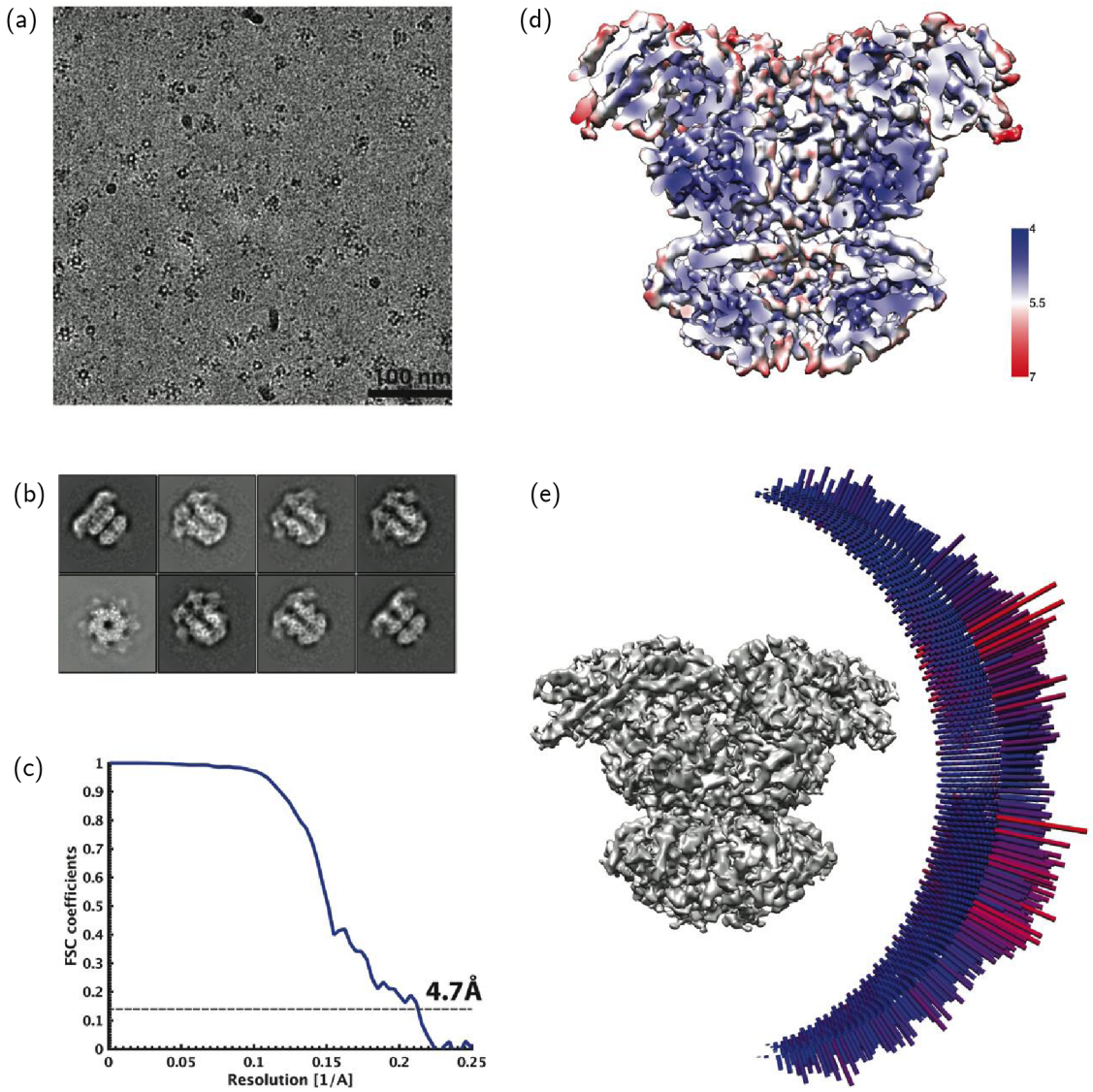


Figure 5.4: CryoEM single particle analysis. CryoEM single particle analysis of a protein complex (KaiCBA). (a) Part of a cryoEM micrograph of the protein. (b) Reference-free 2D class averages from the dataset. (c) Fourier shell correlation (FSC) for two independently refined halves of the data. Shown is the curve for reconstructions with C_6 symmetry imposed during refinement and the resolution at which FSC = 0.143. (d) Cut-open density of the atomic density colored according to its local resolution. (e) Orientation distribution plot of all particles that contribute to the reconstruction. Figure from Ref. [463].

regards to Bayesian influence of structural relaxation. Here, the goal is to develop methods to probe the native structure in the solution-phase from the accessible, though information poor, experimental techniques.

5.1.3 Existing modeling techniques

Homology modeling

Producing a method for creating structural models of proteins that are difficult to crystallize from the PDF would further develop the field of *de novo* protein structure solution, which is presently dense with programs and methods, using the precepts of modeling disorder. A predominant approach in *de novo* protein modeling has been through comparative or homology modeling [473]. The major assumption of this field is that proteins with similar primary structure (amino acid sequence) or evolutionary history will have similar secondary (conformational) and tertiary (geometric) structure. Thus, structure prediction usually proceeds along four major steps: fold identification and template selection, template-target alignment, model building, and model evaluation. The primary step can be accomplished by evolutionary knowledge, and/or database programs which use scoring functions to find potential homologues such as BLAST [474], PSI-Blast [475], and FASTA [476]. Generally, alignment between sequences can be improved after template selection by similar programs, and sequence identity of over 40% is considered to be correctly aligned. The protein model can then be built from rigid bodies, segments, or spatial restraints. Model building expresses particular difficulties in the loop regions between areas of canonical secondary structure and in placing the sidechains and hydrogen atoms. Finally, in a purely comparative model, the proposed protein structure is evaluated via any combination of similarity scoring, stereochemistry, traditional spatial features, and physical force fields such as CHARMM [215].

Particular attention should be given to the most well received and robust program, Rosetta [477], which has been continually developed by the Baker group

including the addition of NMR capabilities [478]. ROSETTA is unique in that it uses fragments of unrelated protein structures to assemble the query structure by fragment sequence similarity. The modeling proceeds by simulated annealing with a Bayesian scoring function (a type of data driven Hamiltonian) that accounts for the radius of gyration, relative position and orientation of structural elements, residue pair interactions, and solvent accessible surface area. These knowledge-based methods rely heavily on the breadth, diversity, and accuracy of the databases of known structures [479]. The scoring function was continually developed to separate sequence dependent features from sequence independent features (such as the assembly of super-secondary structures from secondary structures). ROSETTA was soon combined with a high resolution full atom potential for further refinement — initially for short structurally variable regions like loops [480] and later larger structures [481] — as well as including distance restraints for better sampling of conformational space [482]. The most recent and comprehensive comparative modeling procedure is ROSETTACM [481] and is most notable for its ability to accurately bridge ordered regions of structure with a loop closure method, and resolve high accuracy sidechain and backbone conformations with limited sequence alignment. Particularly, this speaks to the relative utility of Bayesian inference in low resolution modeling and the need for mixed methods in high resolution or large scale protein problems.

Relating structure and chemical shift

In producing a Bayesian mixed method, we turned to a solution-phase accessible measurement, NMR, to provide priors, and thus further insight into our information poor problem. There are several families of algorithms that calculate chemical shift from structural coordinates, and their complements that approach the inverse problem. The approaches use what is known as the secondary chemical shift ($\Delta\delta$), defined as the difference between the absolute chemical shift (δ_{abs}) and the corre-

sponding random coil (δ_{rc}) shift [483]:

$$\Delta\delta = \delta_{\mathrm{abs}} - \delta_{\mathrm{rc}}. \tag{5.2}$$

A detailed review of these programs and their historical development is given in Appendix C, with particular attention paid to the methodology and influence of the programs invoked in this study.

The approaches for calculating chemical shift range from using empirically derived surfaces that capture physically relevant parameters [484], to homology modeling that depends on structural likeness in a database [485]. The most modern algorithms utilize artificial neural networks (ANNs). These machine learning approaches, train a high dimensional empirical equation by adjusting weights during a learning period [486]. Consistent between all of these methods is the dependence on a structural database. ANNs are a blind Bayesian approach and depend on a robust set of priors. Similarly, any empirical equation will be biased to the data set used in parameterization. Of these algorithms, SPARTA+ [487] generally outperforms its competitors when considering the backbone structure [488].

Most relevant to our present study are the programs which provide predictions of structure based on chemical shifts. These will be able to provide prior probabilities that will complement the information scarcity in the scattering data. Again there is a large scope of programs, combinations, and specializations in this research area. Nonetheless, two programs stand out, because of their ability to produce probability distributions: DANGLE [489] and TALOS-N [490, 491].

Most other programs will make a prediction regarding categories of secondary structure or a dihedral angle set with associated errors. DANGLE searches a fragment database and utilizes Bayesian inference to produce a posterior probability distribution from the database likelihoods. This distribution, titled the global likelihood estimate, is discretized over Ramachandran space. Similar plots are created during the TALOS-N execution. Depending on a two-layer neural network, the

TALOS-N program makes predictions from a smooth (ϕ, ψ) probability density map. This map is not the standard output of TALOS-N, but an exploitable component for this body of work. Examples of these distributions are offered in Appendix D. Having already encountered the Ramachandran statistics as a viable source of priors for a Bayesian approach, it should be clear that these two programs can serve a similar purpose.

Mixed methods

After considering the breadth of programs which attempt to solve structure from shift or the inverse problem, it is worth examining the work which incorporates these algorithms into other existing procedures. Traditional NMR restraints (shifts, NOEs, J-couplings, residual dipolar couplings) are still in use for mixed Monte Carlo and molecular dynamics packages [470, 492, 493] and for stronger sampling of conformational space in refining oligomers against small angle X-ray scattering data [471, 472]. Other programs such as CHESHIRE use chemical shifts as restraints in molecular mechanics and in defining secondary structure, within a typical ROSETTA style homology model [494]. Chemical shifts are also used to create structural restraints as tunable soft-square harmonic wells calculated from the differences in experimental shifts and those predicted by SHIFTX [495]. This biases the conformational search to regions of phase space consistent with the chemical shifts.

An explicit combination of methods is demonstrated in CS-ROSETTA [496] [Fig. 5.5]. Using the ROSETTA protocol modified by SPARTA for selecting fragments that are similar in sequence and in shift, and using shift to determine cis-proline peptide bonds [487], the model was capable of converging for proteins smaller than 130 residues. Improvements of the CS-ROSETTA algorithm to include evolutionary information resulted in a program capable of modeling protein structures up to 225 residues in size. Similar in principle to CS-ROSETTA, the CS23D web server contains a back end collaboration of many aforementioned programs [497]. The program generates multiple conformations and geometries which are shift threaded by THRIFTY

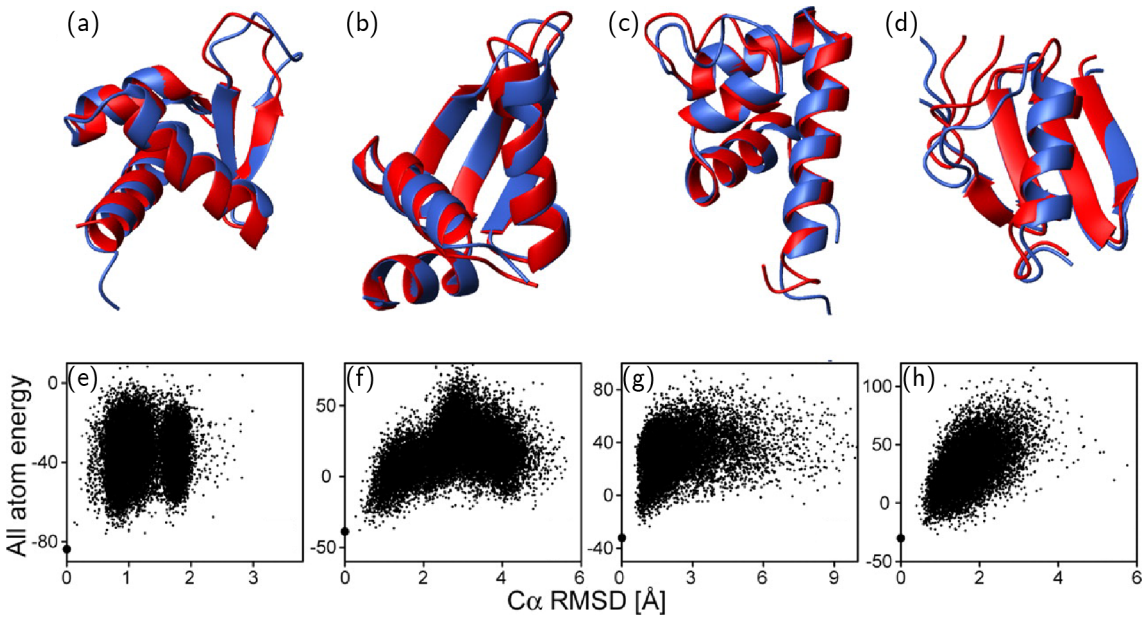


Figure 5.5: Blind structure prediction with CS-Rosetta. Results from blind CS-Rosetta structure generation for four structural genomics targets. (a)-(d) Superposition of lowest-energy CS-Rosetta models (red) with experimental NMR structures (blue), with superposition optimized for ordered residues. (e)-(h) Plots of the rescored all-atom energy versus C_α RMSD relative to the lowest-energy model (bold dot on vertical axis). (a) and (e): StR82. (b) and (f): RpT7. (c) and (g): VfR117. (d) and (h): NeT4. Taken from Ref. [496].

[498] then assembled and refined using Rosetta and a knowledge based potential function, GAFolder.

POMONA is a development to using sparse NOE data, and (ϕ, ψ) probability maps as extracted from TALOS-N, to search the PDB for structures which coincide with the maps [491]. Templates are constructed and then submitted to CS-Rosetta as input. Another alteration to molecular fragment replacement techniques includes biased sampling based on a continuous probabilistic model which meets the requirements of detailed balance and is suitable for Markov chain Monte Carlo simulations such as those in the early stages of Rosetta [499, 500]. These probability distributions can also introduce bias through force field perturbations.

As stated, these shift programs have also made an impact in molecular dynamics simulations, particularly when an appropriate starting configuration is available. One such program utilizes the continuous and differentiable shift predictions of CamShift [501] to augment the molecular mechanics potentials [502]. The method

can also be combined with other structural restraints, and works effectively if used following molecular fragment replacement techniques to generate a starting configuration. Shift-driven molecular dynamics in a genetic algorithm have also been proven effective for refining a near correct structure from sparse NMR data [503]. Recently, TALOS-N data has been used in tandem with traditional NMR structural restraints and the Xplor-NIH molecular dynamics package in AUDANA [504] to determine structure directly from NOESY data.

Although this description of methods has been brief and general with respect to their operating principles, it should be clear that there is no single perfect way to resolve a protein structure from experimental data. Such a complicated problem requires information from experiment, historical data, intuition, and physical mechanics. As such, an open avenue for novel research is the bridging of some of the aforementioned approaches with each other and/or disparate concepts to solve protein structures rapidly and effectively. Also yet to be mentioned, and of particular interest, is the goal of approaching refinement of structures which are not canonical in the databases and historical information on which many of these approaches rely.

5.2 Algorithm

To achieve the goals set out at the start of this study, we developed an RMC-type algorithm for fitting protein structure to an experimental PDF. Within the RMC procedure, the likelihood of a given protein fold determines the probability of its proposal, and the quality of that proposal's fit to the experimental data in turn determines the probability of accepting the proposed move. This approach emerges naturally from Bayesian statistics, and as we have seen, there are robust prior statistics available for proteins. Our Bayesian reverse Monte Carlo approach to protein structure prediction — the Prior Information Protein folding Algorithm (PIPA) — is a new approach to handling the degeneracy of solutions that fit information poor data, which accounts for likelihood in a rigorous fashion.

PIPA is implemented in the Fortran 2003 standard in a modular approach similar

to other programs in this thesis. This was chosen with the speed of calculation in mind. A given instantiation of a protein class is assembled of a coordinates list and set of logical masks for atom types. This is feasible — and works rapidly — because there are only a few types of element relevant to a protein and exploits the memory indexing important to Fortran. The protein class also contains a few other relevant attributes, an instantiation of a residue class, and methods associated with bending the protein (changing a given torsion angle) and calculating the PDF. The residue class is implemented as a list of residues — or amino acids — that track relevant information and break the residue into its backbone and sidechain constituents. In conjunction with a library detailing the amino acids, this effectively tracks the correct number of χ dihedrals, and allows for backbone only simulations. The smallest subunit of the class hierarchy is the atom class, which tracks individual coordinates, labels, symbols, and an integer index for book-keeping on each atom. Significantly more detail can be found in the program source in Appendix F.

5.2.1 Initial structure

The program is initiated by calling the executable followed by a command argument that is the location of an input file. The input file dictates the experimental parameters and has some keywords for testing and debugging [App E]. PIPA begins with a configuration file familiar amongst the protein crystallography community. The RCSB protein data bank file type (`.pdb`) is used as input to generate a starting configuration. If a starting structure is not available, PIPA will read in the sequence information of the protein without any coordinates. From a sequence, a planar secondary structure is generated using the standard geometries of an amino acid backbone [505]. The bond lengths and angles used are readily extracted from the all atom potential AMBER [506] (MD packages such as GROMACS [218] come equipped with pre-processing capable of converting between the equilibrium geometry of two potentials). Once a planar secondary structure is created, the torsion

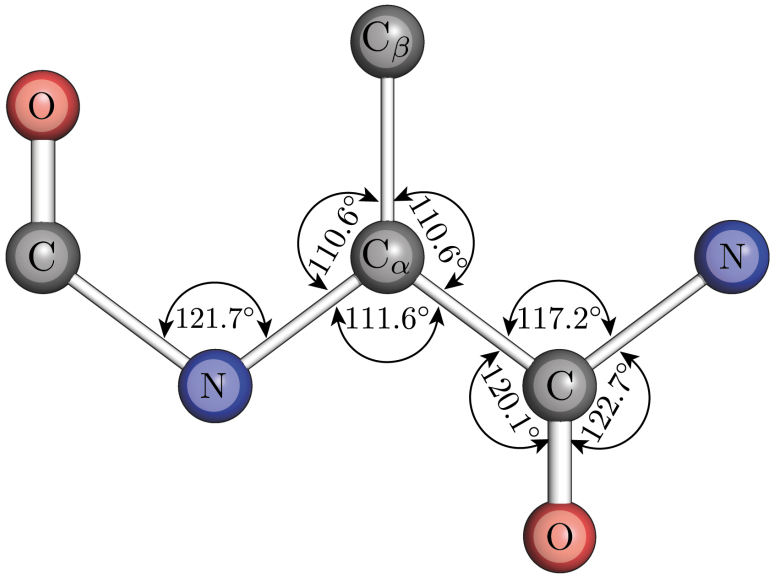


Figure 5.6: Standard amino acid backbone geometry. The reference backbone geometry used to assemble a protein in two dimensions from only its sequence.

angles are randomized and the protein bent to create a randomly configured polypeptide chain in space.

The initial randomization allows for a greater search of configurational space and avoids the memory of a starting configuration. This functionality can be removed for continuing simulations or when starting from a structural configuration with complete atomic coordinates. If the randomization is not used, the initial torsional angles are then calculated. A torsion angle is given by the clockwise angle between two planes, and thus is defined by four atoms. Considering Figure 5.7, the first plane can be defined by the unit normal given by the cross product of the unit vectors, $\hat{\mathbf{n}}_1 = \hat{\mathbf{r}}_{12} \times \hat{\mathbf{r}}_{23}$. Likewise, the second plane can be defined by $\hat{\mathbf{n}}_2 = \hat{\mathbf{r}}_{23} \times \hat{\mathbf{r}}_{34}$. Finally, the below equation gives the signed clockwise torsion angle from the `atan2` function:

$$\phi = \texttt{atan2}\left[\hat{\mathbf{r}}_{23} \cdot (\hat{\mathbf{n}}_1 \times \hat{\mathbf{n}}_2), (\hat{\mathbf{n}}_1 \cdot \hat{\mathbf{n}}_2)\right]. \tag{5.3}$$

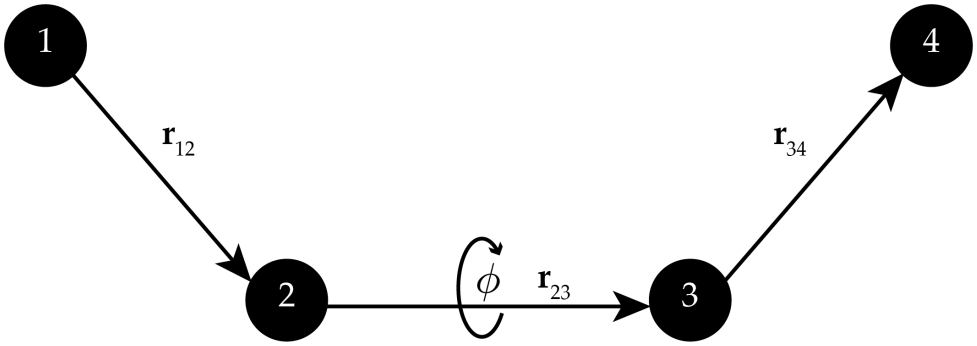


Figure 5.7: Defining a torsion angle. Four atoms define a torsion angle about an axis.

5.2.2 Bayesian approach

The core of the algorithm is a reverse Monte Carlo that depends on a Bayesian approach [eq. (5.1)]. Bayesian probability is a particularly useful tool in information poor problems, and has been used in inverse problems in biology [507], geophysics [321], meteorology [508], medical imaging [326], and commonly political science [509]. In this case, I recast the traditional phrasing of Bayes' theorem in equation (5.4): the posterior probability of a model state i given an observable O is equal to the likelihood of the observable given the state, times the prior probability of the state, divided by the prior probability of the observable.

$$P(i \mid O) = \frac{P(O \mid i)P(i)}{P(O)} \tag{5.4}$$

This is important in the context of a RMC refinement. Unlike traditional RMC, the algorithm does not search for the model that best fits the data, the refinement searches for the *most probable* model — as given by some prior distribution — that best fits the data. This distinction has already been employed conceptually in generating sensible starting configurations that represent likely local coordination in $ZrW_2O_8{\cdot}H_2O$ and $Pd/Pt(im)_2$, and more explicitly in considering local environments to likely be similar across a structure in programs such as INVERT [309]. Here the

explicit implementation will come in the transition probabilities of the Markov chain, generated from the Ramachandran statistics or NMR based predictions.

Within a Markov process, the probability of accepting a new state, π_{ij}, is given as a function of proposal probabilities τ and state probabilities π [eq. (5.5)]. The general process is as follows.

$$\pi_{ij} = \frac{\pi_j}{\pi_i} \times \frac{\tau_{ij}}{\tau_{ji}} \tag{5.5}$$

With a proposal probability, τ_{ij}, move from state i to a trial state j. That state is the accepted with probability, π_{ij}, obeying detailed balance. The probability of a given state is then considered as proportional to the posterior probability of the state given the observable:

$$\pi_j = P(j \mid O) \propto \frac{P(O \mid j)P(j)}{P(O)}. \tag{5.6}$$

Substituting equation (5.6) into (5.5), we have

$$\pi_{ij} = \frac{P(O \mid j)}{P(O \mid i)} \frac{P(i)}{P(j)} \frac{\tau_{ij}}{\tau_{ji}}, \tag{5.7}$$

and using the prior probabilities of the states as the proposal probability matrix

$$\tau_{ji} = P(i), \tag{5.8}$$

we are left with an acceptance probability that is a ratio of the likelihood of the observable given the two states. This is best cast as a Boltzmann type function for the later incorporation of a pseudo temperature':

$$\pi_{ij} = \frac{P(O \mid j)}{P(O \mid i)} = \frac{\exp\{-\chi_j^2\}}{\exp\{-\chi_i^2\}}. \tag{5.9}$$

This results in a complete separation of the proposal and acceptance parts of the algorithm, and only requires data fitting in the acceptance step. From an implemen-

tation standpoint this is not dissimilar from a standard RMC acceptance step. The principle distinction comes in the more efficient exploration of configuration space that maintains chemical and biological sensibility.

The fitting probability is given as an exponential of the cost function of each state, χ_i, that is the sum over the squared residuals [eq. (2.28)]. Thereby, the acceptance probability is implemented with a pseudo temperature as

$$\pi_{ij} = \min \left\{ 1, \exp \left[\frac{-(\chi_j^2 - \chi_i^2)}{T} \right] \right\}. \tag{5.10}$$

With this set of definitions, the algorithm proceeds as described in 2.4. Each move is proposed based on the prior probability distributions (5.8). The new value of χ^2 is calculated and the move is accepted with a probability from (5.10). This probability is dependent on a temperature variable, which allows simulated annealing to be implemented in the algorithm.

5.2.3 Observable

The acceptance probability is a function of the experimental observable, in this case the PDF, and the observable as calculated from the model. In order to account for the effects of thermal motion within the protein, the model PDF would need to be altered by some means. One possible approach is to convolve the simulated PDF with a broadening function that approximates the effects of thermal motion. This is however just an approximation and the convolution operation is rather computationally expensive. Overcoming the approximation would require significantly more vibrational detail, and overcomplicate the process. With both of these ideas in mind, using a convolution to approximate thermal motion is an unattractive option.

Alternatively, the PDF can be reformulated in terms of a distance list (as accomplished in [311]). First a cumulative PDF is formed from $G(r)$ or $G'(r)$:

$$C_{\mathrm{PDF}}(r) = \sum_{i}^{r} G(r). \tag{5.11}$$

$C_{\mathrm{PDF}}(r)$ is then normalized by its maximum, to run over the interval [0,1]. The number of distances expected is known, so the distance list can be populated with values of r by searching for fractional reductions of values of the normalized $C_{\mathrm{PDF}}(r)$. These floored bins will generate multiple identical distances where there will be unique distances in the model. Importantly, this distance list is automatically sorted as it is created. Thus, a corresponding χ^2 can be calculated by a vector subtraction of the simulated and experimental distance lists.

It is important to note that while PIPA has been developed to handle real experimental data, the study contained in this chapter is a proof-of-concept that operates on data generated from an exact model. In this first pass there would be no need to concern ourselves with thermal motion in the PDF; however, the distance list formulation will continue to be used to demonstrate its utility and eventually allow for incorporation of inexact data.

5.2.4 Monte Carlo moves

Following the creation of a distance list and initialization of a structural model, the set of prior probabilities are read in. These priors are in Ramachandran space and functions $P(\phi, \psi)$ for each residue in the protein chain. They can in turn be constructed from chemical shift-to-structure programs (TALOS-N or DANGLE), Ramachandran statistics ignorant of chemical shifts [510], or the renormalized product of the two. For work without chemical shifts information, the Dunbrack group has generated a set of neighbor-dependent Ramachandran distributions, $P_{\ldots XR}(\phi_X, \psi_X)$ and $P_{LX\ldots}(\phi_X, \psi_X)$ for all residues X with all possible neighboring residues L or R [510]. At each residue, a probabilistically combined triplet Ramachandran distribution at each medial residue, dependent on *both* its neighbors:

$$P_{LXR}\left(\phi_X, \psi_X\right) = \frac{1}{N}[P_{\ldots XR}\left(\phi_X, \psi_X\right) \cdot P_{LX\ldots}\left(\phi_X, \psi_X\right)]. \tag{5.12}$$

Here, N is a normalization constant. The resulting distribution is already smoothed and discretized in (ϕ, ψ) bins of $5° \times 5°$. Regardless of the source of probability distribution, the discretization at this point is the same, so the sampling process will be the same.

PIPA creates a cumulative probability distribution function $C(x)$ from the joint or triplet probability distributions.

$$C(x) = \sum_{n}^{x} P\left(\phi_n, \psi_n\right) \tag{5.13}$$

These can be thought of — although the implementation is different — as a set of lists of $(\phi, \psi, C(x))$ combinations, with one list for each residue. This is the proposal probability matrix. Thus, given three pseudorandom numbers (x,y, and z), the distributions can be sampled.

$$\phi = 72 \left\lfloor \frac{C^{-1}(x)}{72} \right\rfloor + 5y \tag{5.14}$$

$$\psi = 72 \left[C^{-1}(x) \bmod 72\right] + 5z \tag{5.15}$$

The value of $C^{-1}(x)$ can be determined by the smallest value of the cumulative distribution that a random number $x \in (0, 1]$ is less than. This gives the bin for the (ϕ, ψ) combinations, which is then offset by a random fraction (y or z respectively) of the bin size for the proposal.

Once the distributions are sampled for a randomly chosen residue, the protein needs to be bent about the torsional axes. For the change of a single torsion angle, four atoms define the angle [Fig. 5.7]. As this is already defined with respect to an axis $\hat{\mathbf{r}}_{23}$, what becomes relevant is the clockwise change in the angle, θ. In this case, a vector from the atom at point 2 in Figure 5.7 to the atom at point 4 can be created and rotated about the axis $\hat{\mathbf{r}}_{23}$. This is accomplished using the Rodrigues rotation formula (4.14), and is then completed for every atom to the right of the atom at point 3.

With the new conformation in place, it is necessary to calculate a distance list. The uniqueness of this approach allows for a rapid calculation of only the C_α distances for a coarse refinement. The new χ_j is calculated by a vector subtraction of the simulated and experimental distance lists (assuming an uncertainty of $\sigma(n) = 1$), followed by the dot product of this difference with itself to give the sum of the squares of the components [eq. (2.28)]. It was recognized during development that the variation in long-range distances is likely to be larger than the variation in short-range distances. We rectified this by increasing the nominal experimental uncertainty σ in proportion to the distance (*i.e.* setting $\sigma(n) = r_n$). This effectively weights the short-range correlations within the simulation. Alternative weightings can be used (such as r_n^{-2}), but I found the former to be more successful. Finally, the move is accepted with the probability as defined in equation (5.10), with rejected moves reverting to the prior structure as described above with the negative angular change.

5.2.5 Temperature control

Yet to be discussed is that the temperature variable in the simulation also impacts the acceptance probability. Approaches to temperature and an annealing schedule were offered in Section 2.4; however, many of these were designed with real thermodynamic systems in mind. In such a case, it is prudent to allow an ensemble to equilibrate prior to taking measurements of any observables. In this case, I will allow a protein to relax well past equilibrium before changing the temperature by calculating the integrated auto correlation time, τ_{int}. At a given temperature, the RMC refinement can be run for an initial number of moves to calculate τ_{int}, then continued for a multiple of this time to extend past equilibrium. In the PIPA implementation, the user can optionally dictate a number of MC sweeps to use to calculate τ_{int}. The simulation will run at the given temperature to calculate τ_{int} and then continue at that temperature for a fixed multiple of τ_{int} (if it has not already exceeded that time). After this point, the simulation is cooled according to an annealing schedule. Multiple annealing steps are available, but the default implementation uses an

Table 5.2: Available annealing schedules in PIPA. To note, $0 < c < 1$ for exponential cooling.

Constant	$T(t)$	$= T_0$
Linear	$T(t)$	$= T_0 - ct$
Exponential	$T(t)$	$= T_0 c^t$
Logarithmic	$T(t)$	$= \frac{c}{\log(t+1)}$

exponential annealing schedule [Table 5.2].

The benefit of the exponential annealing approach is threefold. Firstly, it reduces the danger of continually boosting the temperature in a method such as adaptive displacement. Secondly, it automates the annealing schedule and reduces the number of adjustable simulation parameters. Thirdly, it opens the opportunity for calculating thermodynamic properties, such as the second moment of χ^2, allowing for observation of any obvious 'phase transitions' in the simulation.

5.2.6 Output

Within a given run of PIPA, several useful outputs can be obtained in a folder corresponding to the run title in the input `.pin` file. A verbose `.pout` file can be written dynamically to track temperature, acceptance rate, χ^2, and the likelihood given the prior probability distributions. A more succinct file can produce the data over time, which is significantly more useful for plotting. The program will also output files to track the structure with less frequency (defaulting to 100 outputs per simulation). These include `.pdb` files for progressive visualizations, data files for plotting the PDF, and the distance list. Within the output folder, subfolders will be created for each of these file types. There is also an option to produce a cleaner version of the prior probability distributions in Ramachandran space for easier visualization. Lastly, there is an option to include the thermodynamic data previously discussed: temperature, χ^2, and $\langle(\chi^2)^2\rangle - \langle\chi^2\rangle^2$. The extent of these files and the inputs parameters available within PIPA (detailed in Appendix E) allow for a significant exploration of the parameter space of the program.

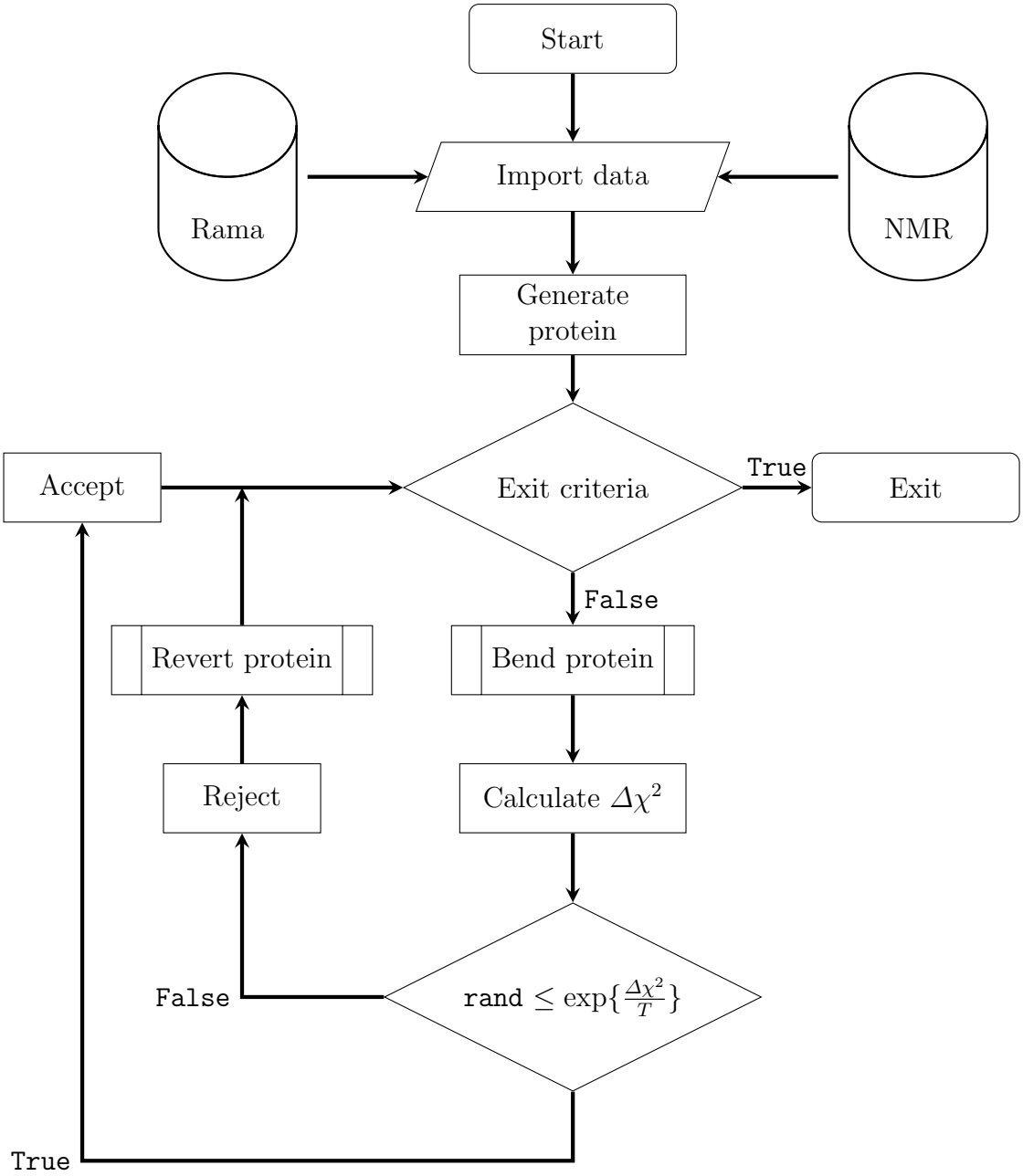


Figure 5.8: PIPA flowchart describing the overall algorithm for the PIPA program.

5.3 Results and discussion

The demonstration of utility and robustness were the first priority in developing PIPA. Can we resolve an accurate structural prediction from the PDF using the priors available? In order to answer this, we worked with a relatively simple protein: melittin. Composed of just 27 residues, melittin forms an alpha helix with an off

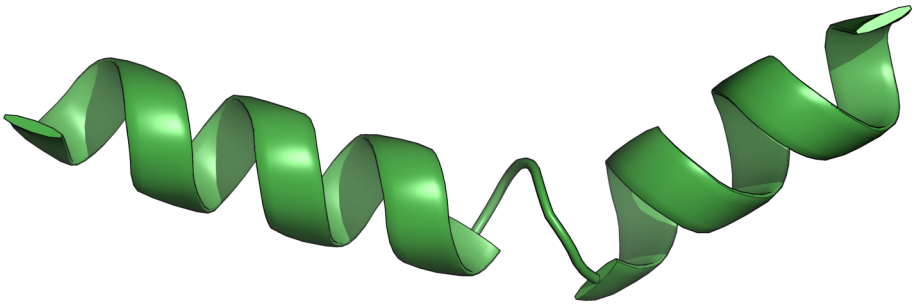


Figure 5.9: Melittin structure.

center bend [Fig 5.9] [511]. Primarily a proof of concept, such a simple protein also allowed for a complete and efficient exploration of parameter space of the program. Below, we explore the nature of the prior probability distributions that are sampled within the algorithm, followed by an analysis of the robustness of these probability distributions and the program overall.

5.3.1 Proposal distributions

A qualitative understanding of the types of distributions can be established by examining a subset of residues in melittin. Any subset would do; however, it is illustrative to chose the residues across the bend in melittin as this is a non-canonical secondary structure. The proposal distributions are given by the *a priori* probabilities of a residue occupying a particular conformation, and subsequently forming a relevant secondary structure. Often, Ramachandran space is cast as a 2D contour plot highlighting particular regions as belonging to secondary structures, as opposed to refined (ϕ, ψ) combinations. With this in mind, the significance of regions of high probability is clear — these are the most likely secondary structures to be formed.

There were three sensible choices for the prior probability distributions: the neighbor dependent Ramachandran statistics produced by Dunbrack and coworkers [510], the distributions produced by NMR based predictors (in this case TALOS-N was strongest option [490]), or the normalized product of the two distributions. Each of these methods produces a prior probability distribution $P_X(\phi, \psi)$ for each

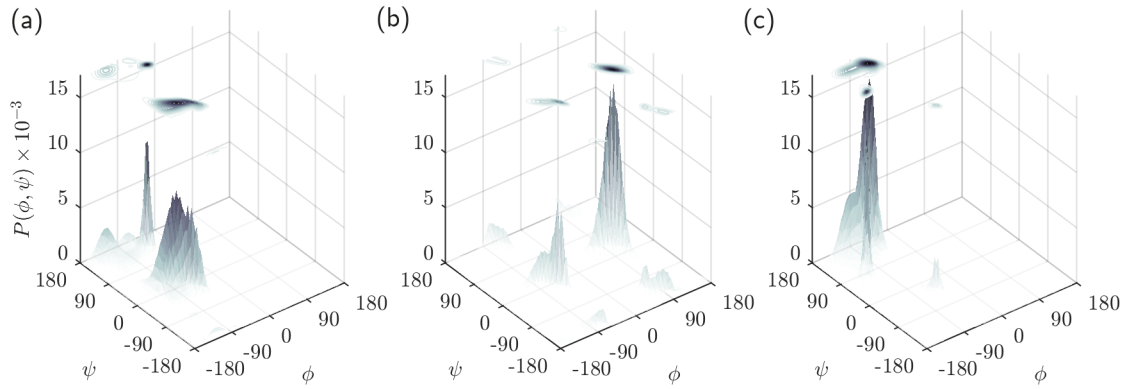


Figure 5.10: Prior probability Ramachandran distributions. Distributions produced from Ramachandran statistics for three residues in melittin: (a) THR-11, (b) GLY-12, (c) LEU-13. These encompass the more complex bent region of melittin, and respectively have reference values of (ϕ, ψ): $(-130.9°, -41.8°)$, $(-84.3°, -32.8°)$, and $(-61.8°, -46.0°)$.

residue, X. As this work set out to explore the limits of information poor problems, the primary approach was to operate PIPA without any input from chemical shifts.

Figure 5.10 shows a characteristic set of Ramachandran distributions. At a glance, it appears that the 11th residue would likely be an alpha-helical conformation of threonine with some likelihood of a beta sheet conformer. The middle frame highlights the flexible nature of glycine, an amino acid commonly present in flexible and loop regions of a protein structure. Despite being part of an alpha helix, the distribution in the third frame of Figure 5.10 indicates a high likelihood of a beta sheet conformation. In the first two residues the correct regions of Ramachandran space will be frequently sampled, but the third will suffer from poor sampling.

It is important to highlight how narrow these distributions are, and where the reference angles lie. The actual (ϕ, ψ) torsion angles of THR-11 (threonine at the 11th amino acid), GLY-12, LEU-13 are $(-130.9°, -41.8°)$, $(-84.3°, -32.8°)$, and $(-61.8°, -46.0°)$ respectively. To efficiently measure both the accuracy and width of the probability distributions, we use the likelihood quotient

$$Z = \frac{P_X(\varphi_0, \psi_0)}{P_{max}}, \tag{5.16}$$

which compares the likelihood of the reference angle with the maximum value of

Table 5.3: Likelihood quotients Z, for 3 residues across the distribution types.

	THR-11	GLY-12	LEU-13
Ramachandran	6.6×10^{-3}	2.5×10^{-3}	1.8×10^{-2}
Product	5.5×10^{-4}	1.8×10^{-3}	3.8×10^{-2}
TALOS-N	7.6×10^{-2}	6.8×10^{-1}	9.9×10^{-1}

the probability distribution. An ideal case of $Z = 1$ would have the maximum of the distribution centered at the reference (ϕ, ψ) values, and Z would approach 0 if the correct torsion angles were never sampled in the distribution. A table of Z values is offered below [Table 5.3], showing that the value is very indicative of the width of the distribution. LEU-13 has the highest Z value of the three for the Ramachandran statistics, even though it appears as though the distribution would lead to over-sampling of the wrong region in configuration space.

Expanding this discussion to the product distribution, it was expected that the chemical shift information would help isolate a large region of Ramachandran space for each residue with positive probability. Effectively, this would take the multiple potential peaks offered by the Ramachandran statistics and isolate one or two of them as the more likely regions. The distribution is calculated by the product of the Ramachandran probability distributions, P_X^R, and the TALOS-N* produced distributions, P_X^T, then normalized such that the integral over the new distribution is unity.

$$P_X(\phi, \psi) = \frac{P_X^R(\phi, \psi) P_X^T(\phi, \psi)}{\sum_{\phi\psi} P_X^R(\phi, \psi) P_X^T(\phi, \psi)} \tag{5.17}$$

In the case of THR-11, the beta sheet likelihood is all but annihilated and in the GLY-12 case, 6 potential regions of Ramachandran space are isolated down to two [Fig. 5.11]. The consequence of this is apparent on renormalization of the probability distributions. The peaks become amplified but more narrow. This is encompassed in

*For the TALOS-N distributions, first chemical shifts were predicted using SPARTA, a program trained on the same database as TALOS-N. These chemical shifts were then formatted as TALOS-N input, and dihedral angles were then predicted. The program offers varying levels of certainty on its predictions, but most importantly outputs probability distributions for the torsion angles of each residue.

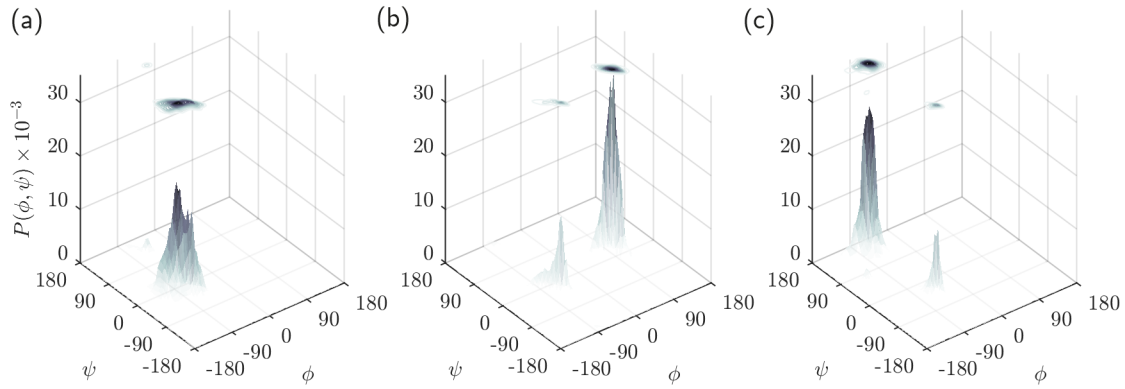


Figure 5.11: Prior probability product distributions. Distributions for three residues in melittin: (a) THR-11, (b) GLY-12, (c) LEU-13. The distributions are the renormalized products of the Ramachandran and TALOS-N distributions. These encompass the more complex bent region of melittin, and respectively have reference values of (ϕ, ψ): $(-130.9°, -41.8°)$, $(-84.3°, -32.8°)$, and $(-61.8°, -46.0°)$.

the Z quotient that is significantly smaller for the product distributions even though the reference angles are no farther from the peak maxima [Table 5.3]. Oversampling of incorrect angles due to a narrow probability distribution would be detrimental if there is much variability in the angles associated with a secondary structure. Examining some of the more general contour plots [App. D] demonstrates that this is in fact the case.

Because the desired utility of the prior probability distributions is better sampling, the narrow nature of the product distributions raised concerns and directed attention toward the TALOS-N produced distributions. Although these are very wide distributions, they tend to have their maxima in the correct regions of Ramachandran space. The tradeoff with a wide distribution is that the sampling of configurational space becomes less efficient as a larger region is sampled. However in the case where the maxima of the distributions deviate significantly from the equilibrium values, a broader distribution becomes necessary. A sensible answer slowly is far more useful than a nonsense answer quickly.

On further examination of the TALOS-N produced probability distributions, it becomes clear that the program does not predict (ϕ, ψ) angles in zero probability regions of the Ramachandran statistics. In fact, the peak location from TALOS-

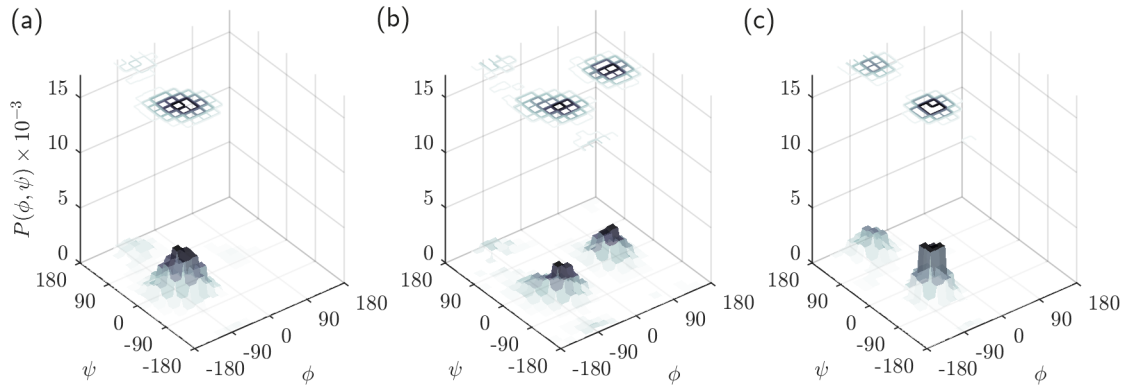


Figure 5.12: Prior probability TALOS-N distributions. Distributions from TALOS-N output for three residues in melittin: (a) THR-11, (b) GLY-12, (c) LEU-13. These encompass the more complex bent region of melittin, and respectively have reference values of (ϕ, ψ): $(-130.9°, -41.8°)$, $(-84.3°, -32.8°)$, and $(-61.8°, -46.0°)$.

N is similar to that from the purely Ramachandran with adjusted magnitude and spread of the data. This is a consequence of the artificial neural networks used in the prediction method. The refinement of the hidden layers within a neural network depends on the learning set and validation set used. Because these data sets used to refine the ANN model are a representative sampling of known proteins, and those known proteins already follow the Ramachandran statistics, it wouldn't be sensible for the ML algorithm to predict outside of the positive nonzero Ramachandran regions. Thus, the Ramachandran statistics are already wrapped inside the TALOS-N algorithm — although hidden — and the multiplying them as in the above is very much 'a hat on a hat': unnecessary and cumbersome. This is not necessarily true for other means of conformation prediction, from chemical shifts or other information, and is strictly a consequence of the Bayesian nature of an ANN.

Having scrutinized the three prior probability distributions utilized in this chapter, it is worth noting that PIPA is capable of handling output from other NMR predictors and can be easily updated with new modules to handle new sources of priors. With that in mind, the tradeoffs associated with these distributions are important. A narrow distribution is only effective if correct, and detrimental if the error is underestimated. In the limiting case this would be an absolute prediction

of the dihedral angles — only useful if accurate. In contrast, a wide distribution will take significantly more simulation time to explore, but still acts to reduce the overall size of configuration space. This was also the impetus for continuing forward with the first neural network output of TALOS-N instead of DANGLE (for a full comparison of the distributions, see Appendix D). Any modular developments to PIPA should apply these insights and preliminary use of the likelihood quotient.

5.3.2 Predictive accuracy

With priors in hand, the program needed to be proven effective and parameterized. This process is routine for any MC simulation. A core component of the simulated annealing algorithm is temperature control, namely the initial temperature and rate of cooling. Firstly, an effective zero temperature must also be found so as not to waste time in the asymptotic decay of the annealing schedule. As the conformation of the structure is randomized at the start of the simulation, there will be an immediate set of 'good' moves which lower the χ^2 value across conformational adjustments. As χ^2 represents a fictitious energetic term, the temperature cannot be directly compared to a Kelvin temperature. Too low of an initial temperature would lead to a down-hill search that would get caught in a local minima. Too high of an initial temperature would lead to a random exploration of conformational space with little penalty for drifting from the correct structure.

The rate of cooling must also be scrutinized. This rate is governed by the variable c in Table 5.2 and cast as $\alpha = 1 - c$ in the figures below. Cooling rate is intimately related to the initial temperature, T_0; ideally one would start with a very high temperature and cool slowly, but this is time consuming and expensive. A balance between a high initial temperature and slow cooling rate needs to be struck in order to effectively approach a global minima.

In order to find fruitful experimental parameters, a large set of simulations was automated and run using PIPA to solve the structure of melittin with idealized simulated data. This was first done with only the Ramachandran distribution and

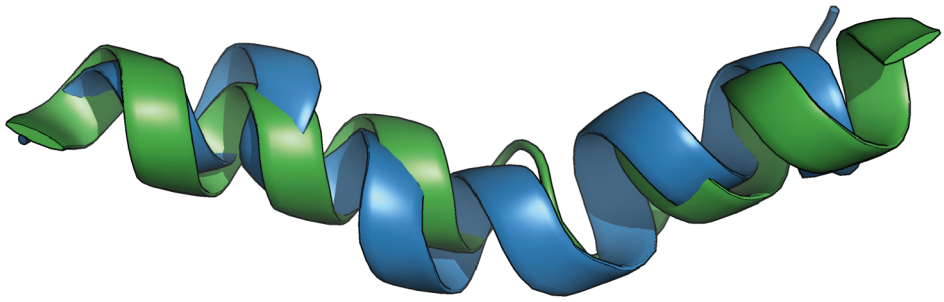


Figure 5.13: Threshold RMSD. Comparison of the crystallographic structure (green) and a structure with an RMSD value of 1.97 Å (blue).

later with the other two probability distributions from above. At each member of the set of parameters (T_0, α), five experiments were run for $10^{6.5}$ MC sweeps with an effective zero temperature of $T_f = 10^{-5}$. Due to the structure of the algorithm in calculating autocorrelation times, a given run could spend longer at one temperature than a different run, and was not guaranteed to reach T_f in the maximum number of sweeps. In total, this amounts to 100 million Monte Carlo moves, which can be achieved in under an hour on a single core of a 2.7 GHz Intel processor.

On reviewing the quality of the structures with low values of χ^2, a benchmark value of 2.0 was chosen for the root-mean-square deviation of atomic positions (RMSD) of an accurate simulation. The RMSD can be defined by

$$\text{RMSD} = \sqrt{\frac{1}{N}\sum_{i}^{N}\delta_i^2}, \tag{5.18}$$

where δ_i is the distance between atom i and the reference structure, and the sum goes over all atoms in the backbone. This metric of success judged the forming of the overall backbone structure of melittin — an alpha helix with an appropriate deviation in the correct residues. A more rigid benchmark could be considered; however, the objective of the study is not necessarily to solve for a perfect structure, but to refine a coarse model that can be used as input for further refinement and to understand the value of experimental information content. In this case a positive

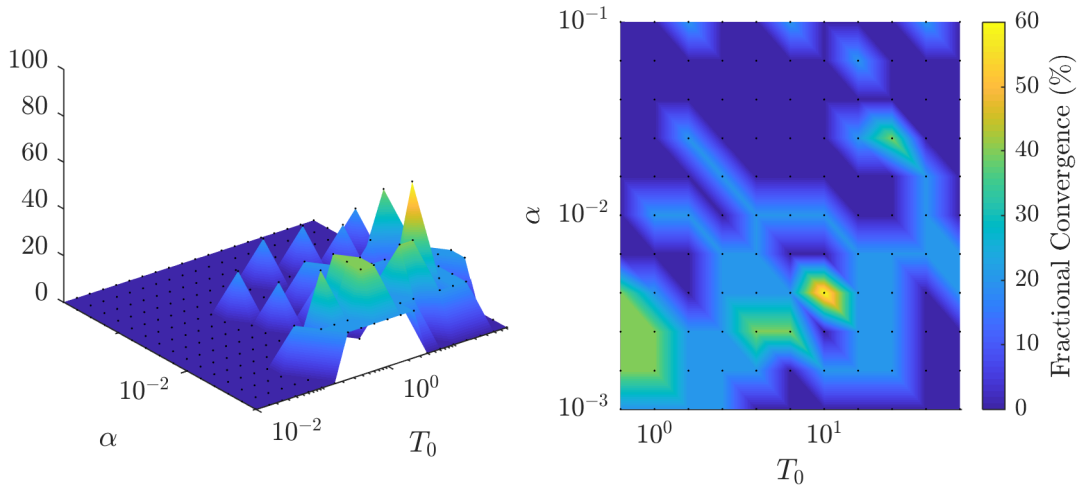


Figure 5.14: Success rate for Ramachandran sampling. Success (RMSD ≤ 2.0) rate for Ramachandran distribution sampling across a set of experimental parameters. These simulations were run for $10^{6.5}$ MC sweeps to the noted starting temperature, T, and exponential cooling rate, α, such that $T_i = (1 - \alpha)T_{i-1}$. A subset of the parameters is shown in the two-dimensional contour plot for clarity.

result would be an accurate if imprecise result. For reference, the crystallographic structure of melittin is overlaid with a simulated structure with an RMSD of 1.97 Å in Figure 5.13. At these lower bounds of accuracy, the local structure is correct across the protein, with minor deviations compensating with each other for the long range correlations.

With this metric, a fractional convergence at each point in parameter space could be calculated. It is expected that there will be some regions that start too cold for the simulation to converge near a global minimum, and some regions that cool too fast to approach a that minimum. Thus a flat region of potential parameters is anticipated where there is a sufficient balance between starting temperature and cooling rate. In the following [Fig. 5.14–5.16] that region is shown in as a two-dimensional contour plot with each black point corresponding to a set of parameters used in PIPA.

When using only the Ramachandran statistics, some simulations can produce accurate models and a flat region of sensible parameters arises in the plots of Fig-

ure 5.14. At best, a given set of parameters only leads to an accurate model 60 % of the time. Although this is promising — using prior probabilities from the gross statistics available, a Bayesian RMC can produce an accurate structural model using only the PDF as experimental information — it leaves something to be desired. An effective proof of concept, this demonstrates the potential for the PDF (currently underutilized in the protein crystallography community) to have disruptive impact in the field of protein structural science. The next logical step was to improve the rate of success and robustness of algorithm by adding additional priors.

Without repeatable success in producing accurate structural models, the approach is not of sufficient use to the community. It could be modified by increasing the timescale of the simulation and allowing for slower cooling, but this process would be divergent in time. We looked to better sampling and improving the prior probability distributions. As discussed above, we expected that any information that isolated a particular region of Ramachandran space would have a substantial impact of the accuracy of the program. In fact, if each residue based probability distribution only possessed two peaks and additional information could effectively isolate one of those peaks, then the effectiveness of sampling — and subsequently rate of convergence — could double. Over 27 residues, this amounts to an $\mathcal{O}(10^8)$ times increase in rate.

In this vein, we considered many of the NMR approaches for estimating conformation, and chose to utilize the TALOS-N program to provide a complementary probability distribution. The new prior probability distribution [eq. (5.17)] had a dramatic impact on the fractional convergence of the PIPA runs [Fig. 5.15]. Overall, the same region of parameter space was successful in producing an accurate model as in the Ramachandran only case, yet the rate of success was significantly higher with some parameters producing accurate models on 100 % of the simulations. The 3D surface plot, clearly shows a cutoff where the initial temperature is too low. The region of very high initial temperature and slow cooling ($T_0 \in [10^{0.4}, 10^2]$ and

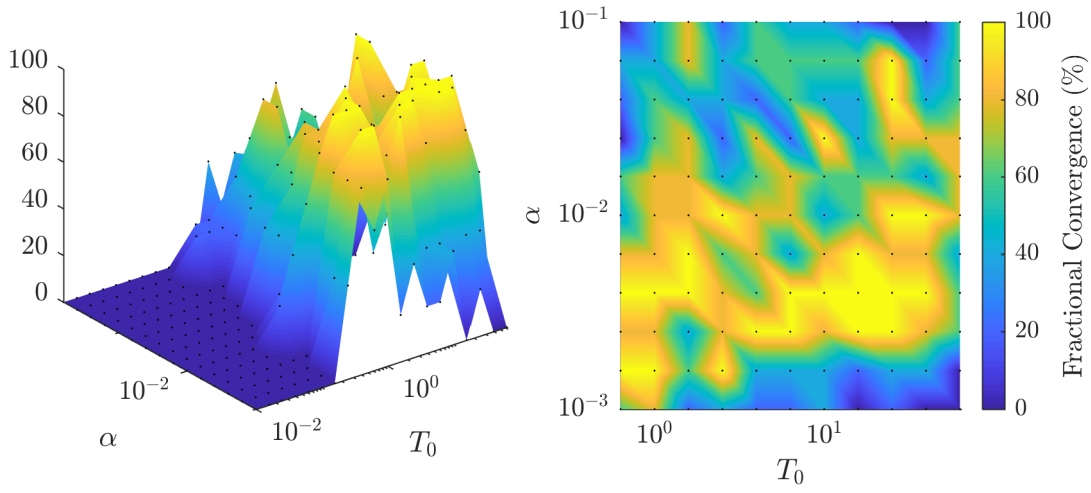


Figure 5.15: Success rate for product sampling. Success (RMSD ≤ 2.0) rate for product distribution sampling across a set of experimental parameters. These simulations were run for $10^{6.5}$ MC sweeps to the noted starting temperature, T, and exponential cooling rate, α, such that $T_i = (1 - \alpha)T_{i-1}$. A subset of the parameters is shown in the two-dimensional contour plot for clarity.

$\alpha = 10^{-3}$) is discontinuous from the expected because these simulations did not have enough time to cool completely with only $10^{6.5}$ sweeps. Nonetheless, the variance across the successful region does create some cause for concern.

Concerned with the variance in fractional convergence, we looked back to sampled probability distributions and considered the possibility that the TALOS-N produced distributions would themselves, though broad, be sufficient to provide informed sampling. This analysis is presented above. Although the product distributions highlight the correct areas to be sampled, the renormalization makes the distributions so narrow that they ineffectively sample the correct angles and oversample some deviation from those angles. The likelihood quotients in Table 5.3 show this trend directly across three different types of amino acids. When the reference torsion angles are sampled up to 1000 times less frequently than the most likely angles, and the maxima are approximately Gaussian, the simulation does not spend a significant enough time exploring near the reference angles. Therefore, the use of an accurate prior probability distribution with a high value of Z is key to a robust

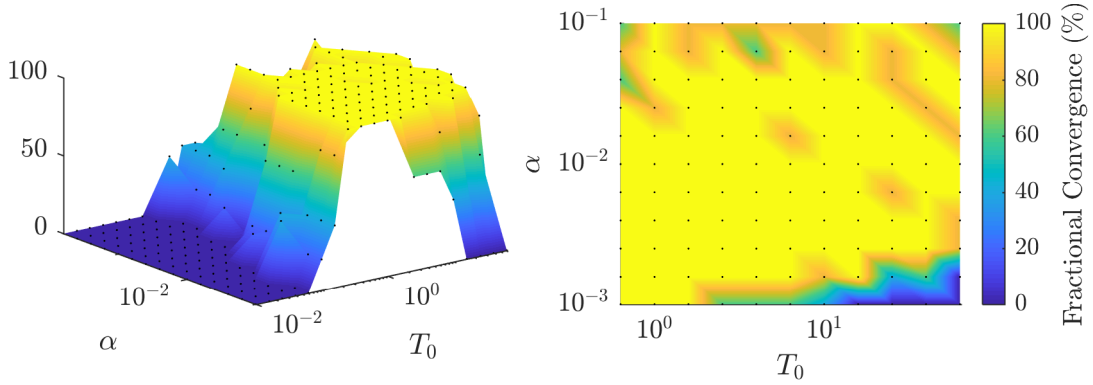


Figure 5.16: Success rate for TALOS-N sampling. Success (RMSD ≤ 2.0) rate for TALOS-N distribution sampling across a set of experimental parameters. These simulations were run for $10^{6.5}$ MC sweeps to the noted starting temperature, T, and exponential cooling rate, α, such that $T_i = (1 - \alpha)T_{i-1}$. A subset of the parameters is shown in the two-dimensional contour plot for clarity.

procedure.

This becomes fully apparent when examining the reproducibility associated with using the TALOS-N produced distributions. Not only is there a clear region of parameter space that produces an accurate model, but this accuracy is continuous across different experimental trails.

Up to this point, idealized simulated data have been used to examine the feasibility of producing an accurate protein model from facile solution phase accessible measurements, and the information content of individual priors. In practice, however, getting complete chemical shift referencing for a protein is challenging and PDF data are not perfect and have a degree of error associated with them. To explore the robustness of the algorithm we convolved the simulated PDF with a Gaussian to mimic thermal fluctuations and then extracted a new distance list. We also used only the ^{15}N and proton NMR shifts as input for TALOS-N. Amino acid doping and developing a ^{13}C loaded protein, and complete shift assignment would be more akin to full NMR crystallography, thus undermining the facile intent of this work. The resulting exploration of parameter space with this incomplete set of input data is shown in Figure 5.17. The program is still more successful in this case than in the use of Ramachandran data. This promising result could further be improved

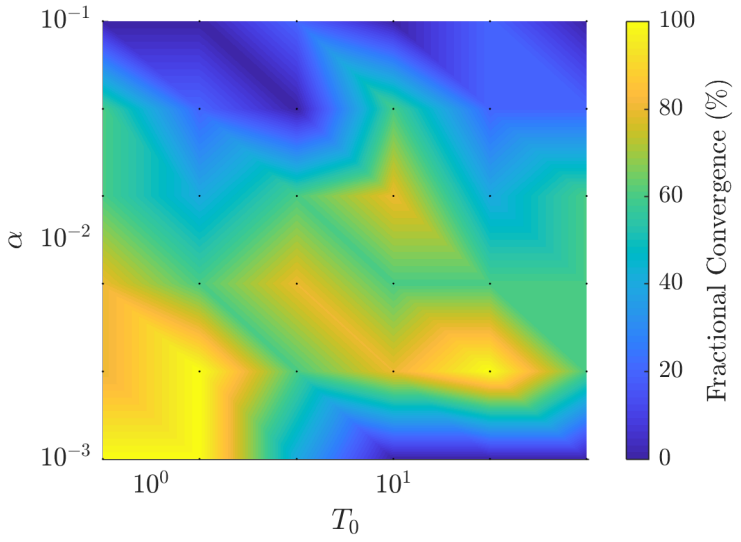


Figure 5.17: Success rate for incomplete data. Success (RMSD ≤ 2.0) rate for TALOS-N distribution sampling across a set of experimental parameters. The TALOS-N distributions were compiled using ^{15}N and proton shifts only, and the distance lists assembled from an ideal PDF convolved with a gaussian. These simulations were run for $10^{6.5}$ MC sweeps to the noted starting temperature, T, and exponential cooling rate, α, such that $T_i = (1 - \alpha)T_{i-1}$.

by allowing for a longer simulation time and slower cooling rates; even if the useful parameter space is less vast, the ability to repeatably produce an accurate model is the priority.

5.3.3 Importance of prior information

Given the promising outcomes associated with melittin structure determination using PIPA, it is fascinating how much information can be extracted from the PDF. We thus turn this investigation to discuss the nature of each type of prior information employed in PIPA. Using the successful parameter space determined above, a set of predicted structures was generated by modifications on the PIPA program. These modifications allowed for input of the chemical shift and/or Ramachandran based probability distributions (for sampling), and/or the PDF (for move acceptance) in any combination. The resulting structures give an understanding of what each piece of prior or experimental information offers the solution of the inverse problem, and the benefits of the combinations of inputs.

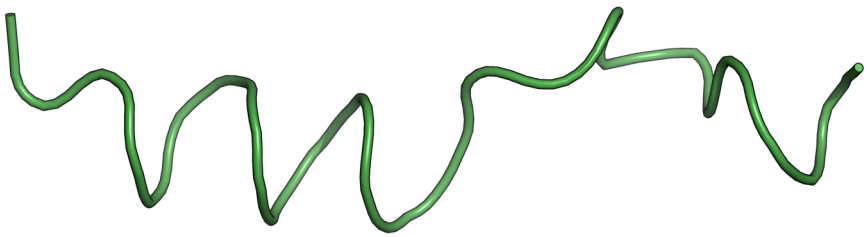


Figure 5.18: Scattering only result. Characteristic RMC solution using only the scattering information, ignorant of Ramachandran statistics and chemical shifts.

In the case of using only the scattering information, random moves are attempted uniformly across Ramachandran space and accepted with the criterion of traditional RMC. This results in a nonphysical local structure with torsion angles inaccessible to proteins, but an overall shape that resembles that of melittin [Fig. 5.18]. This makes sense in the context of the nature of a protein PDF [Fig. 5.1]. Very limited local structural peaks are present in the PDF, yet the overall shape can change and shift with the collective long range correlations of the protein.

In contrast, the Ramachandran statistics offer very little information in isolation [Fig. 5.19]. Sampling regions of disparate secondary structure across residues, no cohesive secondary structure forms. Without any structural probes beyond statistics, the model melts and creates an effectively denatured protein with randomized — yet physical — dihedral angles.

The TALOS-N produced distributions offer the same type of information as the Ramachandran statistics, albeit with more content. The more predictive nature of the chemical shift based probabilities allows for secondary structures to form in the model, akin to the expected [Fig. 5.20]. Nonetheless, the chemical shift information is highly flexible and ignorant beyond small clusters of neighbors, thus cannot reproduce the overall shape of melittin.

To compile these details in a sensible fashion, the graphic below was designed [Fig. 5.21]. The prior information triangle summarizes the type of structural information present in the set of data, and the types of models produced at the intersec-

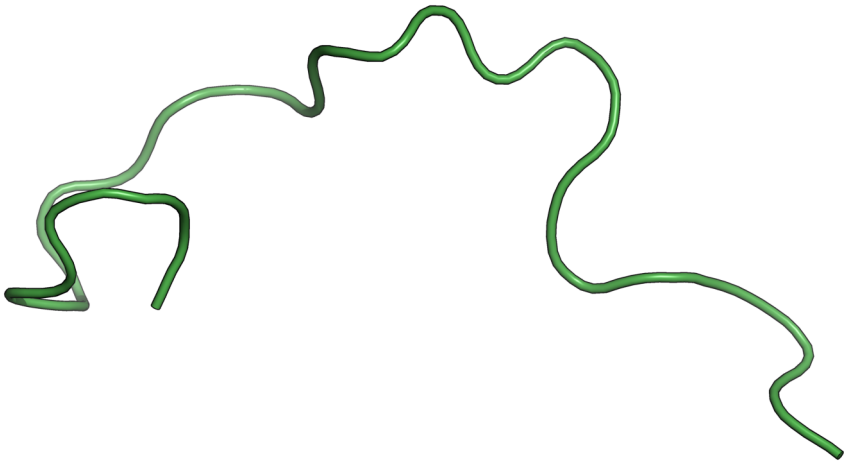


Figure 5.19: Ramachandran only result. Characteristic RMC solution using only the Ramachandran distribution, ignorant of scattering information and chemical shifts.

tions of the information. There are hundreds of examples for each edge and vertex, with those shown in the figure being characteristic in nature. The key conclusions from the graphic are (i) the PDF supplies the simulation with long range information describing the overall protein shape, (ii) the Ramachandran statistics provide better than random sampling, (iii) the chemical shift based priors can estimate local structure completely ignorant of overall shape, and (iv) in concert these information sources offer a robust means of structure solution.

5.3.4 Solution quality

With this remarkable proof of concept and development of the value of prior information sources, a defining question remains between concept and applicability. Using PIPA, how can one discriminate blindly between a good model and a bad model with low χ^2. The program was built explicitly to find not just the model that best fits the data (a problem with PDF based methods already discussed elsewhere in this thesis), but the most likely model that best fits the data. In Figure 5.22, we compare the simulation energetics with the resulting RMSD values (it should be noted that RMSD scales as $N^{\frac{1}{3}}$ and as such, RMSD values should not be compared directly between dissimilar proteins as the program is utilized elsewhere [512]). This shows

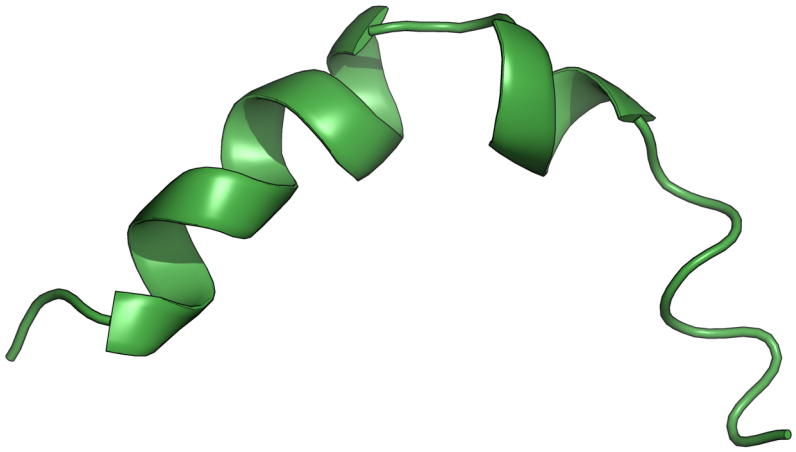


Figure 5.20: TALOS-N only result. Characteristic RMC using only the TALOS-N produced distribution, ignorant of scattering information and Ramachandran statistics.

that there is not a precise monotonic relationship between χ^2 and RMSD. Instead there is a tail of the relationship, where the lowest energy solutions certainly have a low RMSD. Furthermore, the ability to distinguish goodness of fit from goodness of model is developed with the quality of prior probability. Not only is the RMSD trend with respect to χ^2 more distinct for the TALOS-N based simulation, but the trailed likelihood also forms a trend with RMSD. In the case of the product distribution, there is little correlation between RMSD and likelihood.

This is not dissimilar from the discriminatory process of leading programs such as CS-Rosetta [496]. While there may be many low energy solutions, there is a clear threshold in energy where the accuracy of the model is highly correlated. Elsewhere, there is a significant spread in the data, *i.e.* structures with the same χ^2 values can have drastically different RMSD values. When considering the relationship of model strength with likelihood (P_X), this is more dependent on the strength of the prior probability distributions. Again, this is akin to the trends seen when implementing chemical shift information with homology modeling and all-atom potentials [491]. Overall, the χ^2 parameter is effective at distinguishing the highest quality structural predictions; however, it is not sufficient in detecting structures which may be suitable candidates for further refinement. One could place a χ^2 cutoff at 10^{-1} for high quality

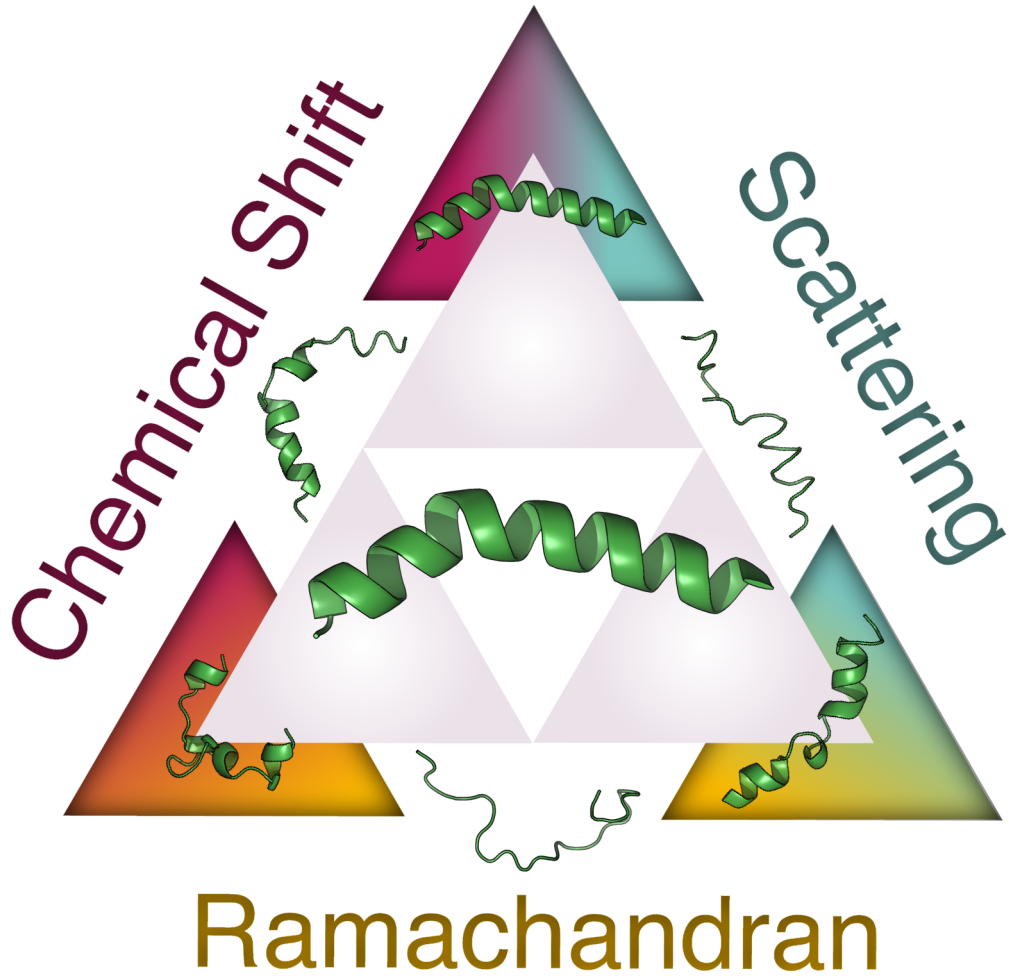


Figure 5.21: Prior information triangle. Graphic representation of the value of specific types of prior information. The sides of the triangle show the characteristic results using only the labeled priors in a Monte Carlo simulation. The vertices include the shared types of priors, and the center demonstrates the characteristics result using the combination of the three types of prior information.

structural models and a cutoff at $10^{-0.5}$ for reasonable structural models, but the former threshold is more rigorous.

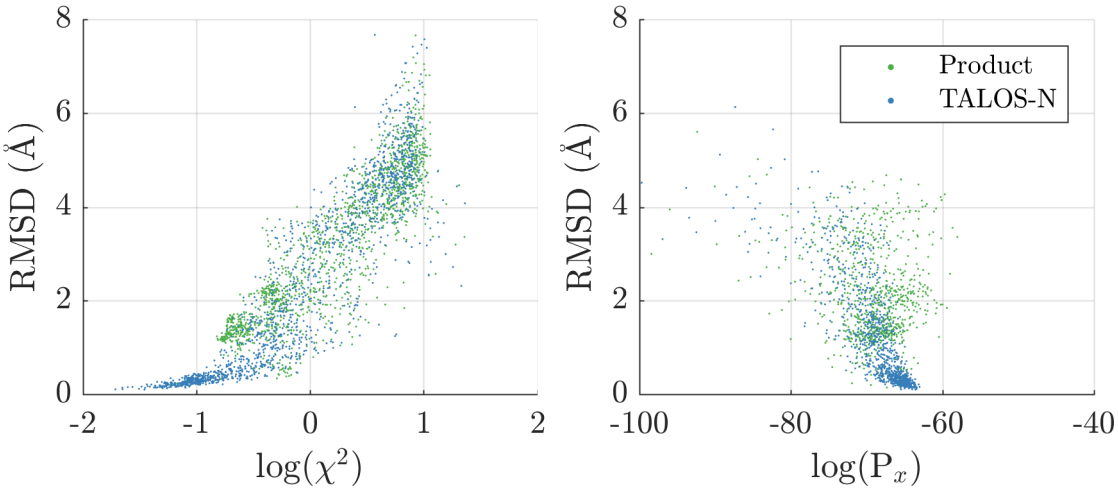


Figure 5.22: RMSD dependence. Plot of RMSD vs χ^2 (left) and RMSD vs $\log P_X$ (right) for the sets of simulations from Figures 5.16 and 5.15.

5.4 Further developments

Following the introduction of the Prior Information Protein folding Algorithm (PIPA), some practical considerations are necessary. The field of protein structural science is diverse and rich with opportunity, therefore such a proof of concept has numerous avenues for development. Firstly, a blind metric needs to be developed for quantifying the strength of a model beyond goodness of fit. This would involve a full atomic potential and structural relaxation. In order to accomplish this, the protein first needs to be decorated with appropriate sidechains. Furthermore, work with real experimental data would serve as an important test of feasibility with increasing information entropy. Beyond these developments, many areas of improvement are available for such a program: hybrid methods are of growing necessity in protein structure determination, and algorithmic improvements are also available.

A program was built to decorate the PIPA output with appropriately oriented side chains for further refinement. With the side chains added, it becomes possible to consider hydrogen bonding, which can dominate the energetics of a protein fold. Specifically, many proteins can be thought of as having a hydrophobic core and hydrophilic exterior. As PIPA is ignorant to these chemical interactions — insofar

as they are not contained in chemical shift information — other means are necessary for further refinement and ensuring chemical sensibility. Thus, sidechain decoration opens a channel for feeding PIPA output structures into structural validation or as starting configurations in molecular dynamics simulations. The decoration is programmed (full detail in Appendix E) to work intimately with the PIPA pipeline, and treated as a packing problem to prioritize simple physical sensibility prior to chemical potential. The program also exists as a stand-alone platform titled SCD (SideChainDecoration).

Furthermore, larger proteins must be explored using the same procedures used for melittin. Accomplished in Figure 5.23, this test set confirms the robustness of PIPA across larger structures; however, corroborates the need for a stronger metric for blind structure determination. The PDF is already degenerate with respect to models, the protein PDF is particularly information sparse, and larger models only add more degrees of freedom in the fitting procedure. We would expect a combined approach — such as that seen in CS-Rosetta [491] — would mitigate this problem by adding more sources of information (such as atomistic potentials). In Figure 5.23 we show two examples of larger proteins aligned to their crystallographic structures, that can be discerned from the test set via RMSD, but not via χ^2.

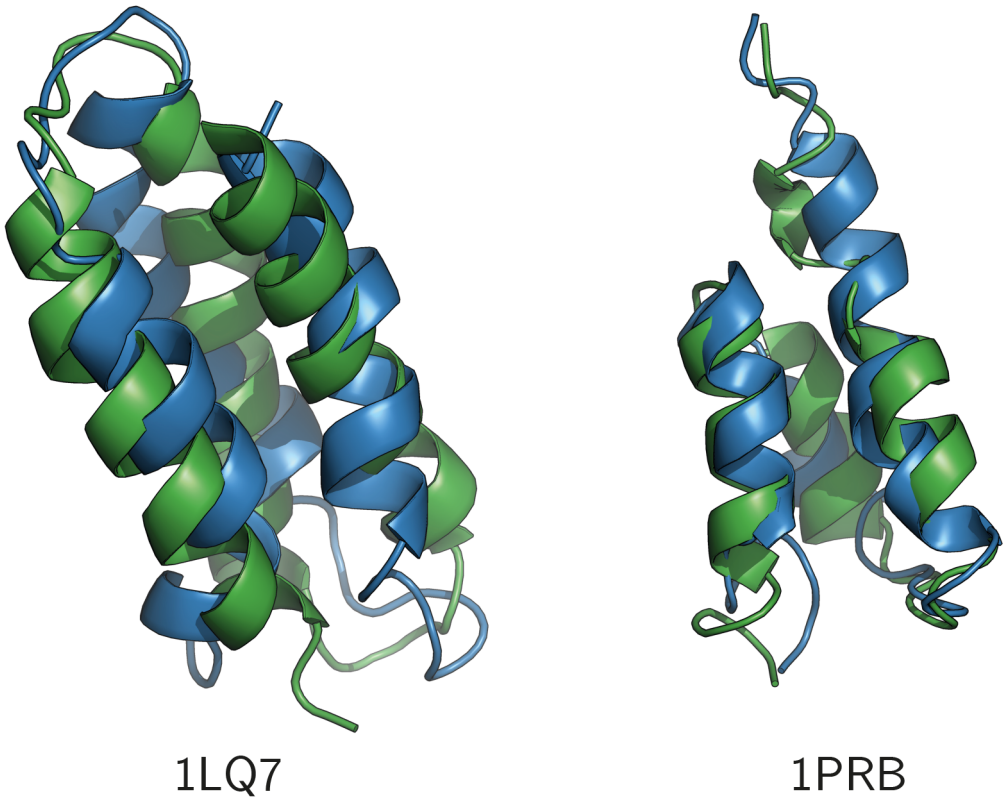


Figure 5.23: Larger protein structures generated by PIPA. The structure of two proteins generated by PIPA (orange) and overlaid on their crystallographic structures (green). The protein 1LQ7 (left) is refined to $\chi^2 = 0.245$ and RMSD $= 3.678\,\text{Å}$. The protein 1PRB (right) is refined to $\chi^2 = 0.462$ and RMSD $= 2.300\,\text{Å}$

5.5 Conclusions

In this study, I have directly implemented Bayesian inference in the modeling of protein structure. Under the umbrella of informed modeling, the approach developed depends on robust prior information. Combining this principle with the vast databases of proteins, and Monte Carlo methods, I have developed a Bayesian reverse Monte Carlo algorithm that creates the most likely structural model of a protein that best fits the pair distribution function. Not only is this a development in protein structural science, but an advancement of the power of the PDF in the crystallography community.

The program, PIPA, has been developed in a proof-of-concept study to demonstrate its potential with the protein melittin. Using this protein with simulated

data, I parameterize the program for use and discuss the utility of particular types of prior information. When considering an ensemble of structure refinements, the algorithm effectively discriminates sensible structure predictions from other ensemble members by using the χ^2 parameter. In considering the models in isolation, the approach suffers from a lack of precision, but could potentially be used as a starting configuration in molecular dynamics. This opens the opportunity for inclusion of other information sources, while still demonstrating the pair distribution function as having sufficient information content to determine the overall tertiary structure of a protein. Combined with sparse chemical shift data, this work demonstrates a facile solution phase probe into protein structural solution, bringing the community one step closer to effective understanding of challenging proteins, especially membrane proteins.

With such a demonstration there are many avenues for future developments in the program and in protein structural science. Mixed methods have begun to dominate the literature, and PIPA fits well into that niche. Future work could very easily combine PIPA with genetic algorithms, homology modeling, or molecular dynamics, without creating the need for crystallization in the experimental pipeline. From a Bayesian perspective, each advancement in mixed modeling is an additional prior to increase confidence in the *a priori* probability distribution of states. Furthermore, the use of PIPA here assumed a cis peptide bond to fix ω, but the program is built to accommodate moves that flip the peptide bond if desired. Finally, the program should be applied to increasingly larger proteins and experimental PDF data to understand any limits to the method. With these ideas for development in mind, the opportunity presented within the PIPA program is transformative in the approach to understanding structure in difficult to crystallize proteins, and furthers our comprehension as to how far we can stretch the diffraction pattern and extract meaningful structural models.

Chapter 6

Conclusions and outlook

At the outset of this thesis, I proclaimed that we would press the limits of using reciprocal space information to derive meaningful real space structures. To this end, we kept our focus on only the most information poor systems, with access to just one-dimensional scattering functions to extract three-dimensional structure. Each subsequent effort, turned to some form of inference that was inherently — if not explicitly — Bayesian. The informed modeling procedures developed often included mixed methodologies to approach a structural interpretation that was not necessarily the structure that best fit the data, but the *most probable* structure that best fit the data. This concept is reiterated word for word in the development of PIPA; however, it can be seen in each of the three systems explored in this thesis.

In the first instance, we used powder X-ray diffraction and the pair distribution function to validate a structural model. The critical hypothesis of local rules driving the hydration mechanism in ZrW_2O_8 was informed from first principles calculations. The local rules are akin to a prior probability given the information from DFT. The propagation of these local rules and subsequent atomistic relaxation yielded a model that could be validated by the total scattering data and offered a meaningful and consistent explanation of the anomalous properties and transitions in the material.

Our second study depended on total scattering data to refine the atomic degrees of freedom of a model. Here, the concepts of local rules manifested themselves in constraining the topology of an amorphous network. With a hypothesis of regularity in subunit geometry, a defect-free amorphous structure was cast as a continuous random network. Not only did this approach deliver insight into a particular material, it demonstrated the ability for the 1D total scattering pattern to differentiate between amorphous topologies and local geometries. Hypothesis driven modeling

extracted information from scattering that isn't natively available.

In our final system of interest, we built protein models from an ostensibly featureless pair distribution function. Depending on the solution phase PDF for complete model generation, Bayesian inference was explicitly implemented into an RMC refinement. What biological macromolecules lack in terms of the periodic information (the choice driver for the crystallographer) they make up for in robust statistics. The Ramachandran distributions created local neighbor-dependent rules for the torsion angles in a protein. These rules — unlike those previously enforced — were empirically probabilistic and not absolute or based on energetic assumptions. The ability to enforce a geometry from chemical constraints, and then generate a model according to the PDF from statistical constraints, is a remarkable testament to the power of informed modeling.

Alongside each other, these three topics couldn't seem more dissimilar. The research composed in this thesis has required collaboration across disciplines, and continual support from expert advice. Nonetheless, each system contained a puzzle, its nature rooted in disorder, that required the application of total scattering in conjunction with some degree of model dependence. At its core, the methodology applied to understanding ZrW_2O_8, is the same as that used to generate tertiary protein structures. That methodology is informed modeling of disorder in the context of total scattering.

The success with Bayesian methods — whether implicit or explicit — helps to identify which inverse problems may be tractable. In the case of highly discriminating data (*e.g.*, single crystal diffraction data for crystals) structural solution proceeds routinely without any priors (specifically, without any prior probability distributions though restrictions are made — such as the positivity of the inverse Fourier transform — in this process). However, upon the introduction of disorder, some form of prior information needs to be introduced. We first saw this in Chapter 1 in the form of seasoned interpretation of single crystal diffuse scattering informing model-driven

structural solution. As our data becomes less discriminating (moving from 3D to 1D, and representing more disordered systems), more robust priors are needed in order to extract meaningful information from the total scattering and solve for structure. These priors could be empirical statistics, relying on abundant data sets, as in the case of our approach toward protein structural solution. They could be experimental, using spectroscopic information to dictate local structure, as was necessary in developing our understanding of amorphous imidazolates with unique coordination geometry. Or the priors could be developed on a case by case basis from first principles (*ab initio* calculations initiated our model building in $ZrW_2O_8{\cdot}H_2O$).

Moving forward, it is expected that the concepts of informed modeling will continue to prosper in the field of crystallography and elsewhere. In any system where local rules, symmetry requirements, or the trends intuitive to chemistry can be enforced, the modeling methods developed here can be adapted. It serves to note, that chemistry as a science often depends on trends, approximations, and historical information; this makes it inherently Bayesian. When these trends and approximations can be defined in terms of a probability distribution in phase space, the analogy becomes clear. It is in this vein that I have continually used the language of Bayesian probability and information theory throughout this thesis. The methods presented here tackled problems with higher dimensionality, under the same principles. Even MCMC can be conceptualized as a way to solve for the stationary distribution of posterior probabilities of parameters given a set of data, without having to calculate the marginal probabilities of the data. It is my prediction that in the not too distant future, the most advanced algorithms with inherently Bayesian approaches will be applied as informed modeling tools for systems similar to those tackled in this thesis.

Future work invoking the principles of informed modeling should work to make explicit use of the stationary distribution. This has been explored in the magnetic analog of INVERT, SPINVERT [313], and can expand upon the Bayesian inference. In the present case — specifically when dealing with PIPA — the Bayesian imple-

mentation assumed the prior probability distributions to be the true probability of torsional microstates. If, as suggested in the close of Chapter 5, further energetic (Boltzmann) distributions of microstates were included in the stationary distribution, the question of how to weight the distributions becomes meaningful. The novelty of the Bayesian implementation is removing this parameterization from the approach, effectively casting the MCMC stationary distribution as a simple product of the empirical and χ^2 distributions; however, as multiple distributions or means of reducing the volume of phase space are added, care must be given to how — if at all — these distributions are weighted relatively. This would allow for a more general analysis of phase space to improve applicability to further materials systems.

Such an explanation is not directly amenable to the topological approaches present in the study of ZrW_2O_8 or $Pd/Pt(im)_2$. An interesting line of inquiry however, is defining and characterizing the topological states in these and other materials. Though beyond the scope of this thesis, the theory of the topology of defects [513] and disorder [514] has been explored in depth. A pressing computational problem is calculating the equivalence between complicated topologies such as those seen in ZrW_2O_8 [515] or in rod packings [516]. This is as important an opportunity for materials engineering as it is from a theoretical perspective. With regards to amorphous square-planar networks, the validation of their existence comes with new opportunities in materials design. A potential next step comes in rigorously describing the structure-property relationships of these networks, and exploring the consequences of quadrapolar order in their magnetic, electronic, and optical properties. While only **nbo** was implemented as an initial topology for generating CRNs, it is important to explore the impact of starting configuration on the final CRN topology, especially in considering the degree of initial interpenetration of the network. I expect that this new class of amorphous materials will be characterized and predicted prior to being synthesized.

Of the advanced statistical approaches available for material characterization

and discovery, machine learning algorithms are becoming more routine for non-experts. Database dependent machine learning is already implemented in the crystallography of mixtures [517] and in crystallization prediction [518, 519]. Recently, breakthroughs in artificial intelligence for synthetic chemistry have also been made [520–522]. Against the backdrop of recent developments in the community — *e.g.* ML-based *ab initio* molecular dynamics for amorphous structure prediction [261], and automated chemical design *via* natural language processing representations [523] — there are a number of compelling research directions for materials discovery. In particular, the energy–structure–function framework for materials discovery [524] has areas that could benefit from ML techniques: (i) a first principles calculated molecular structure could be fed to an end-to-end neural network to produce a set of crystal structure predictions, (ii) in the absence of end-to-end procedures, the present combination of Monte Carlo and Newtonian structural relaxation could implement ML-based *ab initio* potentials, and (iii) the structure function prediction could be accomplished with a combination of neural networks and Gaussian predictors. Each of these research paths would require *de novo* synthesis of rich data sets for algorithm training, but would readily benefit from transfer learning (*i.e.*, borrowing early layers from previously trained neural networks). Outside of materials engineering, machine learning approaches can be further implemented in the characterization of complicated or crystallographically challenging materials.

The underlying necessity for these potential research directions, or any modern techniques that depend on statistical approaches, is a robust prior data set. This creates challenges in developing useful catalogs (attempts seen in the RCSB and bioinformatics), as well as managing the 'bias' and 'variance' of these techniques [524]. Conceptually, the means of preventing the over-fitting of data sets are no different than managing the relative weights between multiple distributions during a search in phase space.

The more scientists implement these algorithms, the more 'black-box' methods

will develop. While these kinds of methods need to be taken with a grain of salt [525], there are many self contained programs that make the crystallography of ordered materials — a once challenging task for experts — routine. It is my expectation that even the structural characterization of disordered materials will eventually become a routine analytical tool, and the contents of this thesis will be reduced to a series of undergraduate practicals.

Bibliography

[1] P. Delhaes, *Carbon Science and Technology: From Energy to Materials* (John Wiley & Sons, Inc., 2012), chap. 2.

[2] C. Kittel, *Introduction to Solid State Physics* (John Wiley & Sons, Ltd., 2004), eighth edn.

[3] J. E. Shelby, *Introduction to Glass Science and Technology* (The Royal Society of Chemistry, 2005), second edn.

[4] A. Bhattacharya, *Nature* **459**, 24 (2009).

[5] W. L. Bragg, *Proc. R. Soc. Lond. A* **89**, 248 (1913).

[6] W. H. Bragg, W. L. Bragg, *Proc. R. Soc. Lond. A* **88**, 428 (1913).

[7] W. H. Bragg, W. L. Bragg, *Proc. R. Soc. Lond. A* **89**, 277 (1913).

[8] J. H. Warner, E. R. Margine, M. Mukai, A. W. Robertson, F. Giustino, A. I. Kirkland, *Science* **337**, 209 (2012).

[9] U. Diebold, *Surf. Sci. Rep.* **48**, 53 (2003).

[10] R. D. Piner, J. Zhu, F. Xu, S. Hong, C. A. Mirkin, *Science* **283**, 661 (1999).

[11] T. R. Welberry, B. D. Butler, *J. Appl. Cryst.* **27**, 205 (1994).

[12] T. Egami, S. Billinge, *Underneath the Bragg Peaks: Structural Analysis of Complex Materials* (Pergamon, 2003).

[13] C. Giacovazzo, ed., *Fundamentals of Crystallography* (Oxford University Press, 2011), third edn.

[14] D. V. Denisov, K. A. Lőrincz, W. J. Wright, T. C. Hufnagel, A. Nawano, X. Gu, J. T. Uhl, K. A. Dahmen, P. Schall, *Sci. Rep.* **7**, 43376 EP (2017).

[15] C. J. Byrne, M. Eldrup, *Science* **321**, 502 (2008).

[16] S. Weiner, I. Sagi, L. Addadi, *Science* **309**, 1027 (2005).

[17] O. A. Tertuliano, J. R. Greer, *Nat. Mater.* **15**, 1195 EP (2016).

[18] N. Ciocchini, M. Laudato, M. Boniardi, E. Varesi, P. Fantini, A. L. Lacaita, D. Ielmini, *Sci. Rep.* **6**, 29162 EP (2016).

[19] A. L. Greer, N. Mathur, *Nature* **437**, 1246 EP (2005).

[20] J. Kido, M. Kimura, K. Nagai, *Science* **267**, 1332 (1995).

[21] K. Nomura, H. Ohta, A. Takagi, T. Kamiya, M. Hirano, H. Hosono, *Nature* **432**, 488 EP (2004).

[22] I. Levin, V. Krayzman, J. C. Woicik, *Phys. Rev. B* **89**, 024106 (2014).

[23] D. A. Keen, A. L. Goodwin, *Nature* **521**, 303 (2015).

[24] C. A. Young, A. L. Goodwin, *J. Mater. Chem.* **21**, 6464 (2011).

[25] S. J. L. Billinge, M. G. Kanatzidis, *Chem. Commun.* pp. 749–760 (2004).

[26] M. G. Tucker, A. L. Goodwin, M. T. Dove, D. A. Keen, S. A. Wells, J. S. O. Evans, *Phys. Rev. Lett.* **95**, 255501 (2005).

[27] L. H. N. Rimmer, M. T. Dove, K. Refson, The mechanism for negative thermal expansion in zirconium tungstate, ZrW2O8 (2015). ArXiv: 1510.00361.

[28] T. D. Bennett, A. L. Goodwin, M. T. Dove, D. A. Keen, M. G. Tucker, E. R. Barney, A. K. Soper, E. G. Bithell, J.-C. Tan, A. K. Cheetham, *Phys. Rev. Lett.* **104**, 115503 (2010).

[29] K. W. Chapman, G. J. Halder, P. J. Chupas, *J. Am. Chem. Soc.* **131**, 17546 (2009). PMID: 19916507.

[30] T. D. Bennett, P. Simoncic, S. A. Moggach, F. Gozzo, P. Macchi, D. A. Keen, J.-C. Tan, A. K. Cheetham, *Chem. Commun.* **47**, 7983 (2011).

[31] T. D. Bennett, D. A. Keen, J.-C. Tan, E. R. Barney, A. L. Goodwin, A. K. Cheetham, *Angew. Chem. Int. Ed.* **50**, 3067 (2011).

[32] S. Moggach, T. Bennett, A. Cheetham, *Angew. Chem. Int. Ed.* **48**, 7087 (2009).

[33] J. W. Lynn, D. N. Argyriou, Y. Ren, Y. Chen, Y. M. Mukovskii, D. A. Shulyatev, *Phys. Rev. B* **76**, 014437 (2007).

[34] S. J. L. Billinge, T. Proffen, V. Petkov, J. L. Sarrao, S. Kycia, *Phys. Rev. B* **62**, 1203 (2000).

[35] H. Shintani, H. Tanaka, *Nat. Phys.* **2**, 200 EP (2006).

[36] B. C. Wood, N. Marzari, *Phys. Rev. Lett.* **97**, 166401 (2006).

[37] T. Ogitsu, E. Schwegler, G. Galli, *Chem. Rev.* **113**, 3425 (2013). PMID: 23472640.

[38] S. T. Bramwell, M. J. P. Gingras, *Science* **294**, 1495 (2001).

[39] S. T. Bramwell, M. J. Harris, B. C. den Hertog, M. J. P. Gingras, J. S. Gardner, D. F. McMorrow, A. R. Wildes, A. L. Cornelius, J. D. M. Champion, R. G. Melko, T. Fennell, *Phys. Rev. Lett.* **87**, 047205 (2001).

[40] V. E. Fairbank, A. L. Thompson, R. I. Cooper, A. L. Goodwin, *Phys. Rev. B* **86**, 104113 (2012).

[41] M. Christensen, A. B. Abrahamsen, N. B. Christensen, F. Juranyi, N. H. Andersen, K. Lefmann, J. Andreasson, C. R. H. Bahl, B. B. Iversen, *Nat. Mater.* **7**, 811 EP (2008).

[42] G. Xu, J. Wen, C. Stock, P. M. Gehring, *Nat. Mater.* **7**, 562 EP (2008).

[43] L. E. Cross, *Ferroelectrics* **151**, 305 (1994).

[44] M. T. Dove, *Structure and Dynamics: An Atomic View of Materials* (Oxford University Press, 2002).

[45] D. A. Keen, *J. Appl. Cryst.* **34**, 172 (2001).

[46] P. Juhás, D. M. Cherba, P. M. Duxbury, W. F. Punch, S. J. L. Billinge, *Nature* **440**, 655 (2006).

[47] T. R. Welberry, B. D. Butler, *Chem. Rev.* **95**, 2369 (1995).

[48] C. Wagner, *J. Non-Cryst. Solids* **76**, 29 (1985). Proceedings of the Workshop on Research Opportunities in Solids with Pulsed Neutrons Sources.

[49] G. Placzek, *Phys. Rev.* **86**, 377 (1952).

[50] N. W. Ashcroft, N. D. Mermin, *Solid State Physics* (Cengage Learning, 1975).

[51] D. J. Goossens, T. R. Welberry, A. P. Heerdegen, M. J. Gutmann, *Acta Cryst. A* **63**, 30 (2007).

[52] B. Wehinger, D. Chernyshov, M. Krisch, S. Bulat, V. Ezhov, A. Bosak, *J. Phys. Condens. Matter* **26**, 265401 (2014).

[53] K. Evans, R. Beanland, *Microscopy and Microanalysis* **20**, 1601 (2014).

[54] C. Hammond, *The Basics of Crystallograhy and Diffraction* (Oxford University Press, 2009).

[55] J. D. Dunitz, V. Schomaker, K. N. Trueblood, *J. Phys. Chem.* **92**, 856 (1988).

[56] M. T. Dove, *Introduction to Lattice Dynamics* (Cambridge University Press, 1993).

[57] S. Boutet, *et al.*, *Science* **337**, 362 (2012).

[58] D. Chandler, *Introduction to Modern Statistical Mechanics* (Oxford University Press, 1987).

[59] R. J. Baxter, *Exactly Solved Models in Statistical Mechanics* (Academic Press, 1982).

[60] X. Zeng, R. Kieffer, B. Glettner, C. Nürnberger, F. Liu, K. Pelz, M. Prehm, U. Baumeister, H. Hahn, H. Lang, G. A. Gehring, C. H. M. Weber, J. K. Hobbs, C. Tschierske, G. Ungar, *Science* **331**, 1302 (2011).

[61] A. B. Cairns, M. J. Cliffe, J. A. M. Paddison, D. Daisenberger, M. G. Tucker, F.-X. Coudert, A. L. Goodwin, *Nat. Chem.* **8**, 442 EP (2016).

[62] A. Simonov, T. Weber, W. Steurer, *J. Appl. Cryst.* **47**, 2011 (2014).

[63] T. R. Welberry, R. L. Withers, *J. Appl. Cryst.* **24**, 18 (1991).

[64] W. L. Bragg, E. J. Williams, *Proc. R. Soc. Lond. A* **145**, 699 (1934).

[65] J. M. Cowley, *Phys. Rev.* **77**, 669 (1950).

[66] H. Okamoto, D. J. Chakrabarti, D. E. Laughlin, T. B. Massalski, *J. Phase Equilib.* **8**, 454 (1987).

[67] P. C. Clapp, *Phys. Rev. B* **4**, 255 (1971).

[68] L. Owen, H. Playford, H. Stone, M. Tucker, *Acta Mater.* **115**, 155 (2016).

[69] L. Owen, H. Playford, H. Stone, M. Tucker, *Acta Mater.* **125**, 15 (2017).

[70] X. Chen, S. Shen, L. Guo, S. S. Mao, *Chem. Rev.* **110**, 6503 (2010). PMID: 21062099.

[71] S. M. Sze, K. K. Ng, *Physics and Properties of Semiconductors—A Review* (John Wiley & Sons, Inc., 2006), pp. 5–75.

[72] S. Sarkar, X. Ren, K. Otsuka, *Phys. Rev. Lett.* **95**, 205702 (2005).

[73] J.-M. Joubert, M. Latroche, R. Černý, A. Percheron-Guégan, K. Yvon, *J. Alloy. Compd.* **330-332**, 208 (2002). Proceedings of the International Symposium on Metal-Hydrogen Systems, Fundamentals and Applications (MH2000).

[74] M. Fratini, N. Poccia, A. Ricci, G. Campi, M. Burghammer, G. Aeppli, A. Bianconi, *Nature* **466**, 841 EP (2010).

[75] L. Brockway, V. Vasiraju, S. Vaddiraju, *Nanotechnology* **25**, 125402 (2014).

[76] N. Zhang, H. Yokota, A. M. Glazer, Z. Ren, D. A. Keen, D. S. Keeble, P. A. Thomas, Z. G. Ye, *Nat. Commun.* **5**, 5231 EP (2014).

[77] M. Felici, R. Trotta, F. Masia, A. Polimeni, A. Miriametro, M. Capizzi, P. J. Klar, W. Stolz, *Phys. Rev. B* **74**, 085203 (2006).

[78] J. B. Goodenough, *Phys. Rev. B* **5**, 2764 (1972).

[79] P. J. Saines, M. G. Tucker, D. A. Keen, A. K. Cheetham, A. L. Goodwin, *Phys. Rev. B* **88**, 134418 (2013).

[80] J. Bobroff, H. Alloul, S. Ouazi, P. Mendels, A. Mahajan, N. Blanchard, G. Collin, V. Guillen, J.-F. Marucco, *Phys. Rev. Lett.* **89**, 157002 (2002).

[81] C. P. Adams, J. W. Lynn, Y. M. Mukovskii, A. A. Arsenov, D. A. Shulyatev, *Phys. Rev. Lett.* **85**, 3954 (2000).

[82] E. S. Božin, S. J. L. Billinge, G. H. Kwei, H. Takagi, *Phys. Rev. B* **59**, 4445 (1999).

[83] L. Pauling, *J. Am. Chem. Soc.* **57**, 2680 (1935).

[84] E. A. DiMarzio, F. H. S. Jr., *J. Chem. Phys.* **40**, 1577 (1964).

[85] A. R. Overy, A. B. Cairns, M. J. Cliffe, A. Simonov, M. G. Tucker, A. L. Goodwin, *Nat. Commun.* **7**, 10445 (2016).

[86] P. J. Camp, A. Fuertes, J. P. Attfield, *J. Am. Chem. Soc.* **134**, 6762 (2012).

[87] D. P. Shoemaker, R. Seshadri, A. L. Hector, A. Llobet, T. Proffen, C. J. Fennie, *Phys. Rev. B* **81**, 144113 (2010).

[88] Z. H. Stachurski, *Fundamentals of Amorphous Solids* (Wiley-VCH, 2015).

[89] J. Fortner, J. S. Lannin, *Phys. Rev. B* **39**, 5527 (1989).

[90] D. I. Svergun, M. H. J. Koch, *Curr. Opin. Struct. Biol.* **12**, 654 (2002).

[91] L. Yu, *Adv. Drug Deliv. Rev.* **48**, 27 (2001). Characterization of the Solid State.

[92] C. D. Dimitrakopoulos, D. J. Mascaro, *IBM Journal of Research and Development* **45**, 11 (2001).

[93] B. T. Holland, C. F. Blanford, T. Do, A. Stein, *Chem. Mater.* **11**, 795 (1999).

[94] L. Zhong, J. Wang, H. Sheng, Z. Zhang, S. X. Mao, *Nature* **512**, 177 EP (2014).

[95] Y. H. Liu, G. Wang, R. J. Wang, D. Q. Zhao, M. X. Pan, W. H. Wang, *Science* **315**, 1385 (2007).

[96] D. I. Bower, *An Introduction to Polymer Physics* (Cambridge University Press, 2002).

[97] A. Navrotsky, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 12096 (2004).

[98] P. F. Damasceno, M. Engel, S. C. Glotzer, *Science* **337**, 453 (2012).

[99] L. Rovigatti, F. Smallenburg, F. Romano, F. Sciortino, *ACS Nano* **8**, 3567 (2014).

[100] D. E. Carlson, C. R. Wronski, *Appl. Phys. Lett.* **28**, 671 (1976).

[101] R. Hasegawa, *Mater. Sci. Eng. A* **375-377**, 90 (2004).

[102] D. I. Svergun, M. H. J. Koch, *Rep. Prog. Phys.* **66**, 1735 (2003).

[103] C. L. Henley, *Annu. Rev. Condens. Matter Phys.* **1**, 179 (2010).

[104] H. Z. Selvi, I. O. Bildirici, M. Yerci, *The Cartographic Journal* **47**, 157 (2010).

[105] A. Sarkar, A. Bhattacharjee, F. Ebrahimi, *The Astrophysical Journal* **783**, 65 (2014).

[106] D. W. Schaefer, B. J. Olivier, C. Ashley, G. Beaucage, D. Richter, B. Farago, B. Frick, D. A. Fischer, *J. Non-Cryst. Solids* **172-174**, 647 (1994). Relaxations in Complex Systems.

[107] F. Pavelčík, O. Pivovarcikova, *J. Appl. Cryst.* **35**, 526 (2002).

[108] G. M. Sheldrick, *Acta Cryst. A* **64**, 112 (2008).

[109] I. Usón, C. E. M. Stevenson, D. M. Lawson, G. M. Sheldrick, *Acta Cryst. D* **63**, 1069 (2007).

[110] R. Caliandro, B. Carrozzini, G. L. Cascarano, L. De Caro, C. Giacovazzo, D. Siliqi, *Acta Cryst. D* **61**, 556 (2005).

[111] L. Palatinus, *Acta Cryst. B* **69**, 1 (2013).

[112] G. Oszlányi, A. Süto, *Acta Cryst. A* **60**, 134 (2004).

[113] J. Wu, K. Leinenweber, J. C. H. Spence, M. O'Keeffe, *Nat. Mater.* **5**, 647 EP (2006).

[114] G. Oszlányi, A. Süto, *Acta Cryst. A* **61**, 147 (2005).

[115] A. A. Coelho, *Acta Cryst. A* **63**, 400 (2007).

[116] G. Oszlányi, A. Süto, *Acta Cryst. A* **67**, 284 (2011).

[117] D. Sayre, *Acta Cryst.* **5**, 60 (1952).

[118] C. Giacovazzo, *Direct phasing in crystallography : fundamentals and applications* (Oxford University Press, 1998).

[119] H. Hauptman, J. Karle, *Acta Cryst.* **9**, 45 (1956).

[120] G. M. Sheldrick, *Acta Cryst. C* **71**, 3 (2015).

[121] L. Palatinus, G. Chapuis, *J. Appl. Cryst.* **40**, 786 (2007).

[122] C. Hübschle, S. van Smaalen, *Acta Cryst. A* **70**, C103 (2014).

[123] P. W. Betteridge, J. R. Carruthers, R. I. Cooper, K. Prout, D. J. Watkin, *J. Appl. Cryst.* **36**, 1487 (2003).

[124] R. E. Dinnebier, S. J. L. Billinge, eds., *Powder Diffraction Theory and Practice* (RSC Publishing, 2008).

[125] A. L. Bail, H. Duroy, J. Fourquet, *Mater. Res. Bull.* **23**, 447 (1988).

[126] G. S. Pawley, *J. Appl. Cryst.* **14**, 357 (1981).

[127] A. Le Bail, *Powder Diffraction* **20**, 316 (2005).

[128] K. D. M. Harris, M. Tremayne, B. M. Kariuki, *Angew. Chem. Int. Ed.* **40**, 1626 (2001).

[129] F. L. Hirshfeld, *Acta Cryst. A* **24**, 301 (1968).

[130] K. D. M. Harris, M. Tremayne, P. Lightfoot, P. G. Bruce, *J. Am. Chem. Soc.* **116**, 3543 (1994).

[131] Y. G. Andreev, P. Lightfoot, P. G. Bruce, *J. Appl. Cryst.* **30**, 294 (1997).

[132] S. Brenner, L. B. McCusker, C. Baerlocher, *J. Appl. Cryst.* **30**, 1167 (1997).

[133] B. M. Kariuki, H. Serrano-González, R. L. Johnston, K. D. Harris, *Chem. Phys. Lett.* **280**, 189 (1997).

[134] H. M. Rietveld, *J. Appl. Cryst.* **2**, 65 (1969).

[135] H. M. Rietveld, *Acta Cryst.* **22**, 151 (1967).

[136] A. Constantinides, N. Mostoufi, *Numerical Methods for Chemical Engineers with MATLAB Applications* (Prentice Hall, 1999).

[137] A. A. Coelho, Topas-academic v6, `http://www.topas-academic.net/`.

[138] A. A. Coelho, *J. Appl. Cryst.* **36**, 86 (2003).

[139] B. H. Toby, R. B. Von Dreele, *J. Appl. Cryst.* **46**, 544 (2013).

[140] M. G. Tucker, M. P. Squires, M. T. Dove, D. A. Keen, *J. Phys. Condens. Matter* **13**, 403 (2001).

[141] V. Petkov, R. G. Difrancesco, S. J. L. Billinge, M. Acharya, H. C. Foley, *Philos. Mag. B* **79**, 1519 (1999).

[142] T. Welberry, T. Weber, *Crystallography Reviews* **22**, 2 (2016).

[143] B. E. Warren, B. L. Averbach, B. W. Roberts, *J. Appl. Phys.* **22**, 1493 (1951).

[144] J. F. James, *A Student's Guide to Fourier Transforms* (Cambridge University Press, 2011), third edn.

[145] T. R. Welberry, *Diffuse X-ray scattering and models of disorder* (Oxford University Press, 2004).

[146] M. Kobas, T. Weber, W. Steurer, *Phys. Rev. B* **71**, 224206 (2005).

[147] J. M. Cowley, *J. Appl. Phys.* **21**, 24 (1950).

[148] P. C. Clapp, S. C. Moss, *Phys. Rev.* **142**, 418 (1966).

[149] P. C. Clapp, S. C. Moss, *Phys. Rev.* **171**, 754 (1968).

[150] S. C. Moss, P. C. Clapp, *Phys. Rev.* **171**, 764 (1968).

[151] L. Reinhard, B. Schönfeld, G. Kostorz, W. Bührer, *Phys. Rev. B* **41**, 1727 (1990).

[152] H. Terauchi, J. B. Cohen, *Acta Cryst. A* **35**, 646 (1979).

[153] M. Morinaga, J. B. Cohen, *Acta Cryst. A* **35**, 975 (1979).

[154] M. Morinaga, J. B. Cohen, J. Faber, Jnr, *Acta Cryst. A* **35**, 789 (1979).

[155] M. Morinaga, J. B. Cohen, J. Faber, Jnr, *Acta Cryst. A* **36**, 520 (1980).

[156] T. Weber, A. Simonov, *Z. Kristallogr.* **227**, 238 (2012).

[157] A. Simonov, T. Weber, W. Steurer, *J. Appl. Cryst.* **47**, 1146 (2014).

[158] T. Weber, A. Simonov, C. D. Malliakas, M. G. Kanatzidis, *Acta Cryst. A* **69**, s572 (2013).

[159] P. Urban, A. Simonov, T. Weber, O. Oeckler, *J. Appl. Cryst.* **48**, 200 (2015).

[160] P. Schaub, T. Weber, W. Steurer, *J. Appl. Cryst.* **44**, 134 (2011).

[161] A. Simonov, T. Weber, A. Goodwin, *Acta Cryst. A* **73**, C1045 (2017).

[162] T. Welberry, E. Chan, D. Goossens, A. Heerdegen, *Phase Transitions* **83**, 80 (2010).

[163] N. Metropolis, S. Ulam, *J. Am. Stat. Assoc.* **44**, 335 (1949).

[164] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, E. Teller, *J. Chem. Phys.* **21**, 1087 (1953).

[165] R. B. Neder, T. Proffen, *Diffuse scattering and defect structure simulations: a cook book using the program DISCUS* (Oxford University Press, 2008).

[166] T. R. Welberry, D. J. Goossens, A. J. Edwards, W. I. F. David, *Acta Cryst. A* **57**, 101 (2001).

[167] M. Paściak, T. R. Welberry, J. Kulda, M. Kempa, J. Hlinka, *Phys. Rev. B* **85**, 224109 (2012).

[168] K. Z. Baba-Kishi, T. R. Welberry, R. L. Withers, *J. Appl. Cryst.* **41**, 930 (2008).

[169] R. L. Withers, Y. Liu, T. Welberry, *J. Solid State Chem.* **182**, 348 (2009).

[170] M. Izquierdo, S. Megtert, D. Colson, V. Honkimäki, A. Forget, H. Raffy, R. Comès, *J. Phys. Chem. Solids* **72**, 545 (2011). Spectroscopies in Novel Superconductors 2010.

[171] Z. Islam, S. K. Sinha, D. Haskel, J. C. Lang, G. Srajer, B. W. Veal, D. R. Haeffner, H. A. Mook, *Phys. Rev. B* **66**, 092501 (2002).

[172] T. R. Welberry, *J. Appl. Cryst.* **24**, 203 (1991).

[173] G. Sala, M. J. Gutmann, D. Prabhakaran, D. Pomaranski, C. Mitchelitis, J. B. Kycia, D. G. Porter, C. Castelnovo, J. P. Goff, *Nat. Mater.* **13**, 488 EP (2014).

[174] D. A. Keen, *J. Phys. Condens. Matter* **14**, R819 (2002).

[175] S. Hull, *Rep. Prog. Phys.* **67**, 1233 (2004).

[176] J. G. Allpress, H. J. Rossell, H. G. Scott, *J. Solid State Chem.* **14**, 264 (1975).

[177] H. J. Rossell, H. G. Scott, *J. Phys. Colloques* **38**, 28 (1977).

[178] R. B. Neder, F. Frey, H. Schulz, *Acta Cryst. A* **46**, 792 (1990).

[179] T. Proffen, R. B. Neder, F. Frey, W. Assmus, *Acta Cryst. B* **49**, 599 (1993).

[180] T. Proffen, R. B. Neder, F. Frey, D. A. Keen, C. M. E. Zeyen, *Acta Cryst. B* **49**, 605 (1993).

[181] T. Welberry, R. Withers, J. Thompson, B. Butler, *J. Solid State Chem.* **100**, 71 (1992).

[182] T. R. Welberry, D. J. Goossens, D. R. Haeffner, P. L. Lee, J. Almer, *J. Synchrotron Radiat.* **10**, 284 (2003).

[183] T. Welberry, R. Withers, S. Mayo, *J. Solid State Chem.* **115**, 43 (1995).

[184] J. P. Goff, W. Hayes, S. Hull, M. T. Hutchings, K. N. Clausen, *Phys. Rev. B* **59**, 14202 (1999).

[185] J. D. Gale, *J. Chem. Soc., Faraday Trans.* **93**, 629 (1997).

[186] R. Xu, T. Chiang, *Z. Kristallogr.* **220**, 1009 (2005).

[187] M. Born, K. Huang, *Dynamical theory of crystal lattices* (Oxford: Clarendon Press, 1968).

[188] R. A. Cowley, W. Cochran, B. N. Brockhouse, A. D. B. Woods, *Phys. Rev.* **131**, 1030 (1963).

[189] G. L. Hua, T. R. Welberry, R. L. Withers, J. G. Thompson, *J. Appl. Cryst.* **21**, 458 (1988).

[190] W. W. Schmahl, I. P. Swainson, M. T. Dove, A. Graeme-Barber, *Z. Kristall.* **201**, 125 (1992).

[191] D. A. Keen, M. T. Dove, *J. Phys. Condens. Matter* **11**, 9263 (1999).

[192] T. R. Welberry, G. L. Hua, R. L. Withers, *J. Appl. Cryst.* **22**, 87 (1989).

[193] A. L. Goodwin, *Phys. Rev. B* **74**, 134302 (2006).

[194] J. C. W. Folmer, R. L. Withers, T. R. Welberry, J. D. Martin, *Phys. Rev. B* **77**, 144205 (2008).

[195] F. S. Tautz, V. Heine, M. T. Dove, X. Chen, *Phys. Chem. Miner.* **18**, 326 (1991).

[196] I. P. Swainson, M. T. Dove, *J. Phys. Condens. Matter* **7**, 1771 (1995).

[197] S. A. Wells, M. T. Dove, M. G. Tucker, K. Trachenko, *J. Phys. Condens. Matter* **14**, 4645 (2002).

[198] M. Gambhir, M. T. Dove, V. Heine, *Phys. Chem. Miner.* **26**, 484 (1999).

[199] M. T. Dove, K. D. Hammonds, V. Heine, R. L. Withers, Y. Xiao, R. J. Kirkpatrick, *Phys. Chem. Miner.* **23**, 56 (1996).

[200] K. D. Hammonds, V. Heine, M. T. Dove, *J. Phys. Chem. B* **102**, 1759 (1998).

[201] A. P. Giddy, M. T. Dove, G. S. Pawley, V. Heine, *Acta Cryst. A* **49**, 697 (1993).

[202] K. D. Hammonds, V. Heine, M. T. Dove, *Phase Transitions* **61**, 155 (1997).

[203] M. T. Dove, A. K. A. Pryde, V. Heine, K. D. Hammonds, *J. Phys. Condens. Matter* **19**, 275209 (2007).

[204] A. K. Pryde, K. D. Hammonds, M. T. Dove, V. Heine, J. D. Gale, M. C. Warren, *Phase Transitions* **61**, 141 (1997).

[205] A. K. A. Pryde, K. D. Hammonds, M. T. Dove, V. Heine, J. D. Gale, M. C. Warren, *J. Phys. Condens. Matter* **8**, 10973 (1996).

[206] M. J. Conterio, A. L. Goodwin, M. Tucker, D. A. Keen, M. Dove, L. Peters, J. S. O. Evans, *J. Phys. Condens. Matter* **20**, 255225 (2008).

[207] L. H. N. Rimmer, M. T. Dove, *J. Phys. Condens. Matter* **27**, 185401 (2015).

[208] J. Seminario, P. Politzer, eds., *Modern density functional theory: a tool for chemistry* (Elsevier, 1995).

[209] J. P. Perdew, A. Ruzsinszky, L. A. Constantin, J. Sun, G. I. Csonka, *J. Chem. Theory Comput.* **5**, 902 (2009). PMID: 26609599.

[210] D. Rappoport, N. R. M. Crawford, F. Furche, K. Burke, *Approximate Density Functionals: Which Should I Choose?* (John Wiley & Sons, Ltd, 2006).

[211] P. Ganesh, E. Cockayne, M. Ahart, R. E. Cohen, B. Burton, R. J. Hemley, Y. Ren, W. Yang, Z.-G. Ye, *Phys. Rev. B* **81**, 144102 (2010).

[212] M. Sepliarsky, Z. Wu, A. Asthagiri, R. E. Cohen, *Ferroelectrics* **301**, 55 (2004).

[213] M. Sepliarsky, A. Asthagiri, S. Phillpot, M. Stachiotti, R. Migoni, *Curr. Opin. Solid State Mater.* **9**, 107 (2005).

[214] M. Paściak, T. Welberry, A. Heerdegen, V. Laguta, T. Ostapchuk, S. Leoni, J. Hlinka, *Phase Transitions* **88**, 273 (2015).

[215] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, M. Karplus, *J. Comput. Chem.* **4**, 187 (1983).

[216] B. R. Brooks, *et al.*, *J. Comput. Chem.* **30**, 1545 (2009).

[217] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, P. A. Kollman, *J. Am. Chem. Soc.* **117**, 5179 (1995).

[218] B. Hess, C. Kutzner, D. van der Spoel, E. Lindahl, *J. Chem. Theory Comput.* **4**, 435 (2008). PMID: 26620784.

[219] T. D. Kühne, *WIREs Comput. Mol. Sci.* **4**, 391 (2014).

[220] M. Sepliarsky, R. E. Cohen, *J. Phys. Condens. Matter* **23**, 435902 (2011).

[221] F. H. Stillinger, T. A. Weber, *Phys. Rev. B* **31**, 5262 (1985).

[222] J. T. Randall, H. P. Rooksby, B. S. Cooper, *Z. Krist.* **75**, 196 (1930).

[223] B. E. Warren, *J. Chem. Phys.* **2**, 551 (1934).

[224] F. M. Michel, L. Ehm, S. M. Antao, P. L. Lee, P. J. Chupas, G. Liu, D. R. Strongin, M. A. A. Schoonen, B. L. Phillips, J. B. Parise, *Science* **316**, 1726 (2007).

[225] S. Roorda, W. C. Sinke, J. M. Poate, D. C. Jacobson, S. Dierker, B. S. Dennis, D. J. Eaglesham, F. Spaepen, P. Fuoss, *Phys. Rev. B* **44**, 3702 (1991).

[226] D. L. Griscom, *J. Non-Cryst. Solids* **73**, 51 (1985). Glass Science and Technology Problems and Prospects for 2004.

[227] W. H. Zachariasen, *J. Am. Chem. Soc.* **54**, 3841 (1932).

[228] N. Mousseau, G. T. Barkema, S. M. Nakhmanson, *Philos. Mag. B* **82**, 171 (2002).

[229] F. Wooten, D. Weaire, *J. Non-Cryst. Solids* **64**, 325 (1984).

[230] F. Wooten, K. Winer, D. Weaire, *Phys. Rev. Lett.* **54**, 1392 (1985).

[231] D. Polk, *J. Non-Cryst. Solids* **5**, 365 (1971).

[232] L. Guttman, W. Y. Ching, J. Rath, *Phys. Rev. Lett.* **44**, 1513 (1980).

[233] W. Y. Ching, *Phys. Rev. Lett.* **46**, 607 (1981).

[234] M.-Z. Huang, W. Y. Ching, *Phys. Rev. B* **54**, 5299 (1996).

[235] N. Li, W.-Y. Ching, *J. Non-Cryst. Solids* **383**, 28 (2014). The XIII International Conference on the Physics of Non-Crystalline Solids.

[236] N. Li, R. Sakidja, S. Aryal, W.-Y. Ching, *Phys. Chem. Chem. Phys.* **16**, 1500 (2014).

[237] R. Biswas, D. R. Hamann, *Phys. Rev. Lett.* **55**, 2001 (1985).

[238] J. Tersoff, *Phys. Rev. Lett.* **61**, 2879 (1988).

[239] R. Car, M. Parrinello, *Phys. Rev. Lett.* **60**, 204 (1988).

[240] J. Greneche, J. Teillet, J. Coey, *J. Phys. France* **48**, 1709 (1987).

[241] A. Tilocca, *J. Chem. Phys.* **139**, 114501 (2013).

[242] B. R. Djordjević, M. F. Thorpe, F. Wooten, *Phys. Rev. B* **52**, 5685 (1995).

[243] J. M. Holender, G. J. Morgan, *J. Phys. Condens. Matter* **3**, 7241 (1991).

[244] J. Wejchert, D. Weaire, F. Wooten, *J. Non-Cryst. Solids* **122**, 241 (1990).

[245] G. T. Barkema, N. Mousseau, *Phys. Rev. B* **62**, 4985 (2000).

[246] R. Vink, G. Barkema, W. van der Weg, N. Mousseau, *J. Non-Cryst. Solids* **282**, 248 (2001).

[247] R. L. C. Vink, G. T. Barkema, M. A. Stijnman, R. H. Bisseling, *Phys. Rev. B* **64**, 245214 (2001).

[248] J. Samela, S. A. Norris, K. Nordlund, M. J. Aziz, *Nucl. Instr. Meth. Phys. Res. B* **269**, 1568 (2011). Computer Simulations of Radiation Effects in Solids.

[249] M. Hejna, P. J. Steinhardt, S. Torquato, *Phys. Rev. B* **87**, 245204 (2013).

[250] W. Y. Ching, *Phys. Rev. B* **26**, 6610 (1982).

[251] H. HongXing, *J. Non-Cryst. Solids* **89**, 402 (1987).

[252] S. von Alfthan, A. Kuronen, K. Kaski, *Phys. Rev. B* **68**, 073203 (2003).

[253] Y. Tu, J. Tersoff, *Phys. Rev. Lett.* **84**, 4393 (2000).

[254] V. M. Burlakov, G. A. D. Briggs, A. P. Sutton, Y. Tsukahara, *Phys. Rev. Lett.* **86**, 3052 (2001).

[255] R. L. C. Vink, G. T. Barkema, *Phys. Rev. B* **67**, 245201 (2003).

[256] N. Mousseau, G. T. Barkema, *J. Phys. Condens. Matter* **16**, S5183 (2004).

[257] P. V. Sushko, S. Mukhopadhyay, A. S. Mysovsky, V. B. Sulimov, A. Taga, A. L. Shluger, *J. Phys. Condens. Matter* **17**, S2115 (2005).

[258] K. Laaziri, S. Kycia, S. Roorda, M. Chicoine, J. L. Robertson, J. Wang, S. C. Moss, *Phys. Rev. Lett.* **82**, 3460 (1999).

[259] J. S. Custer, M. O. Thompson, D. C. Jacobson, J. M. Poate, S. Roorda, W. C. Sinke, F. Spaepen, *Appl. Phys. Lett.* **64**, 437 (1994).

[260] P. Biswas, R. Atta-Fynn, S. R. Elliott, *Phys. Rev. B* **93**, 184202 (2016).

[261] V. L. Deringer, N. Bernstein, A. P. Bartók, M. J. Cliffe, R. N. Kerber, L. E. Marbella, C. P. Grey, S. R. Elliott, G. Csányi, *J. Phys. Chem. Lett.* pp. 2879–2885 (2018).

[262] G. T. Barkema, N. Mousseau, *Phys. Rev. Lett.* **77**, 4358 (1996).

[263] G. Barkema, N. Mousseau, *Comput. Mater. Sci.* **20**, 285 (2001). 9th Int. Workshop on Computational Materials Science.

[264] N. Mousseau, L. J. Lewis, *Phys. Rev. Lett.* **78**, 1484 (1997).

[265] N. Mousseau, G. T. Barkema, S. W. de Leeuw, *J. Chem. Phys.* **112**, 960 (2000).

[266] S. Sastry, C. Austen Angell, *Nat. Mater.* **2**, 739 EP (2003).

[267] C. Desgranges, J. Delhommelle, *J. Am. Chem. Soc.* **133**, 2872 (2011).

[268] A. Wright, M. F. Thorpe, *Phys. Status Solidi B* **250**, 931 (2013).

[269] M. M. J. Treacy, J. M. Gibson, *Acta Cryst. A* **52**, 212 (1996).

[270] M. M. J. Treacy, K. B. Borisenko, *Science* **335**, 950 (2012).

[271] J. M. Gibson, M. M. J. Treacy, T. Sun, N. J. Zaluzec, *Phys. Rev. Lett.* **105**, 125504 (2010).

[272] J. M. Gibson, *Science* **335**, 929 (2012).

[273] X. Wang, B. Chen, H. Wang, X. Zheng, S.-J. Liu, J. Wang, B. Li, S. Yu, Z. Cui, *J. Appl. Cryst.* **49**, 528 (2016).

[274] S. Nakhmanson, N. Mousseau, G. T. Barkema, P. M. Voyles, D. A. Drabold, *Int. J. Mod. Phys. B* **15**, 3253 (2001).

[275] J. M. Gibson, *J. Phys. Condens. Matter* **19**, 455217 (2007).

[276] S. Roorda, L. J. Lewis, *Science* **338**, 1539 (2012).

[277] T. Proffen, S. J. L. Billinge, *J. Appl. Cryst.* **32**, 572 (1999).

[278] V. Petkov, I.-K. Jeong, J. S. Chung, M. F. Thorpe, S. Kycia, S. J. L. Billinge, *Phys. Rev. Lett.* **83**, 4089 (1999).

[279] C. L. Farrow, P. Juhas, J. W. Liu, D. Bryndin, E. S. Božin, J. Bloch, T. Proffen, S. J. L. Billinge, *J. Phys. Condens. Matter* **19**, 335219 (2007).

[280] W. Zhao, Z. Wang, C. D. Malonzo, T. E. Webber, A. E. Platero-Prats, F. Sotomayor, N. A. Vermeulen, T. C. Wang, J. T. Hupp, O. K. Farha, R. L. Penn, K. W. Chapman, M. Thommes, A. Stein, *Chem. Mater.* (2018).

[281] F. X. Zhang, S. Zhao, K. Jin, H. Xue, G. Velisa, H. Bei, R. Huang, J. Y. P. Ko, D. C. Pagan, J. C. Neuefeind, W. J. Weber, Y. Zhang, *Phys. Rev. Lett.* **118**, 205501 (2017).

[282] Š. Michalik, J. Ďurišin, D. Balga, K. Saksl, M. Ďurišin, M. Drakopoulos, *J. Alloy. Compd.* **687**, 188 (2016).

[283] S. Tominaka, K. Kawakami, M. Fukushima, A. Miyazaki, *Mol. Pharmaceutics* **14**, 264 (2017).

[284] P. M. M. Thygesen, J. A. M. Paddison, R. Zhang, K. A. Beyer, K. W. Chapman, H. Y. Playford, M. G. Tucker, D. A. Keen, M. A. Hayward, A. L. Goodwin, *Phys. Rev. Lett.* **118**, 067201 (2017).

[285] S. S. M. Bhat, B. Babu, M. Feygenson, J. C. Neuefeind, M. M. Shaijumon, *ACS Appl. Mater. Interfaces* **10**, 437 (2018).

[286] B. Li, H. Wang, Y. Kawakita, Q. Zhang, M. Feygenson, H. L. Yu, D. Wu, K. Ohara, T. Kikuchi, K. Shibata, T. Yamada, X. K. Ning, Y. Chen, J. Q. He, D. Vaknin, R. Q. Wu, K. Nakajima, M. G. Kanatzidis, *Nat. Mater.* (2018).

[287] H. D. Yoo, Y. Liang, H. Dong, J. Lin, H. Wang, Y. Liu, L. Ma, T. Wu, Y. Li, Q. Ru, Y. Jing, Q. An, W. Zhou, J. Guo, J. Lu, S. T. Pantelides, X. Qian, Y. Yao, *Nat. Commun.* **8**, 339 (2017).

[288] A. E. Maughan, A. M. Ganose, A. M. Candia, J. T. Granger, D. O. Scanlon, J. R. Neilson, *Chem. Mater.* **30**, 472 (2018).

[289] P. M. M. Thygesen, C. A. Young, E. O. R. Beake, F. D. Romero, L. D. Connor, T. E. Proffen, A. E. Phillips, M. G. Tucker, M. A. Hayward, D. A. Keen, A. L. Goodwin, *Phys. Rev. B* **95**, 174107 (2017).

[290] G. Jonker, J. V. Santen, *Physica* **16**, 337 (1950).

[291] J. B. Goodenough, *Phys. Rev.* **100**, 564 (1955).

[292] N. Mathur, *Nature* **390**, 229 EP (1997).

[293] A. P. Ramirez, *J. Phys. Condens. Matter* **9**, 8171 (1997).

[294] R. L. McGreevy, L. Pusztai, *Mol. Simul.* **1**, 359 (1988).

[295] T. R. Welberry, T. Proffen, *J. Appl. Cryst.* **31**, 309 (1998).

[296] D. A. Keen, M. G. Tucker, M. T. Dove, *J. Phys. Condens. Matter* **17**, S15 (2005).

[297] M. G. Tucker, D. A. Keen, J. S. O. Evans, M. T. Dove, *J. Phys. Condens. Matter* **19**, 335215 (2007).

[298] M. Eremenko, V. Krayzman, A. Gagin, I. Levin, *J. Appl. Cryst.* **50**, 1561 (2017).

[299] D. A. Keen, D. S. Keeble, T. D. Bennett, *Phys. Chem. Miner.* (2017).

[300] R. Gaillac, P. Pullumbi, K. A. Beyer, K. W. Chapman, D. A. Keen, T. D. Bennett, F.-X. Coudert, *Nat. Mater.* **16**, 1149 EP (2017).

[301] S. Cao, T. D. Bennett, D. A. Keen, A. L. Goodwin, A. K. Cheetham, *Chem. Commun.* **48**, 7805 (2012).

[302] E. O. R. Beake, M. T. Dove, A. E. Phillips, D. A. Keen, M. G. Tucker, A. L. Goodwin, T. D. Bennett, A. K. Cheetham, *J. Phys. Condens. Matter* **25**, 395403 (2013).

[303] J. A. M. Paddison, H. Jacobsen, O. A. Petrenko, M. T. Fernández-Díaz, P. P. Deen, A. L. Goodwin, *Science* **350**, 179 (2015).

[304] A. L. Goodwin, M. G. Tucker, M. T. Dove, D. A. Keen, *Phys. Rev. Lett.* **96**, 047209 (2006).

[305] H. D. Duncan, M. T. Dove, D. A. Keen, A. E. Phillips, *Dalton Trans.* **45**, 4380 (2016).

[306] H. D. Duncan, E. O. R. Beake, H. Y. Playford, M. T. Dove, A. E. Phillips, *CrystEngComm* **19**, 7316 (2017).

[307] S. Kugler, L. Pusztai, L. Rosta, P. Chieux, R. Bellissent, *Phys. Rev. B* **48**, 7685 (1993).

[308] O. Gereben, L. Pusztai, *Phys. Rev. B* **50**, 14136 (1994).

[309] M. J. Cliffe, A. L. Goodwin, *Phys. Status Solidi* **250**, 949 (2013).

[310] P. Biswas, D. Drabold, *J. Non-Cryst. Solids* **354**, 2697 (2008). Amorphous and Nanocrystalline Semiconductors.

[311] M. J. Cliffe, M. T. Dove, D. A. Drabold, A. L. Goodwin, *Phys. Rev. Lett.* **104**, 125501 (2010).

[312] J. A. M. Paddison, A. L. Goodwin, *Phys. Rev. Lett.* **108**, 017204 (2012).

[313] J. A. M. Paddison, J. R. Stewart, A. L. Goodwin, *J. Phys. Condens. Matter* **25**, 454220 (2013).

[314] J. A. M. Paddison, M. J. Gutmann, J. R. Stewart, M. G. Tucker, M. T. Dove, D. A. Keen, A. L. Goodwin, *Phys. Rev. B* **97**, 014429 (2018).

[315] D. R. Harcombe, P. G. Welch, P. Manuel, P. J. Saines, A. L. Goodwin, *Phys. Rev. B* **94**, 174429 (2016).

[316] G. Opletal, T. Petersen, I. Snook, S. Russo, *Comput. Phys. Comm.* **184**, 1946 (2013).

[317] A. K. Soper, *Phys. Rev. B* **72**, 104204 (2005).

[318] A. K. Soper, K. Weckström, *Biophys. Chem.* **124**, 180 (2006). Ion Hydration Special Issue.

[319] R. L. McGreevy, *J. Phys. Condens. Matter* **13**, R877 (2001).

[320] V. Chandrasekaran, B. Recht, P. A. Parrilo, A. S. Willsky, *Found. Comput. Math.* **12**, 805 (2012).

[321] K. Mosegaard, A. Tarantola, *J. Geophys. Res.* **100**, 12431 (1995).

[322] S. Mallat, G. Peyre, *A Wavelet Tour of Signal Processing : The Sparse Way* (Elsevier Science, 2008), third edn.

[323] J. Suykens, J. Vandewalle, *J. Neural Processing Letters* **9**, 293 (1999).

[324] M. J. Ablowitz, D. J. Kaup, A. C. Newell, H. Segur, *Stud. Appl. Math.* **53**, 249 (1974).

[325] J. Muller, S. Siltanen, *Linear and Nonlinear Inverse Problems with Practical Applications* (Society for Industrial and Applied Mathematics, 2012).

[326] V. Kolehmainen, A. Vanne, S. Siltanen, S. Järvenpää, J. P. Kaipio, M. Lassas, M. Kalke, *e & i Elektrotechnik und Informationstechnik* **124**, 248 (2007).

[327] K. Mosegaard, M. Sambridge, *Inv. Prob.* **18**, R29 (2002).

[328] A. L. Patterson, *Nature* **143**, 939 EP (1939).

[329] A. L. Patterson, *Phys. Rev.* **65**, 195 (1944).

[330] C. Shannon, *Bell Syst. Tech. J.* **27**, 379 (1948).

[331] B. Fultz, J. Howe, *Transmission Electron Microscopy and Diffractometry of Materials* (Springer Berlin Heidelberg, 2013), fourth edn.

[332] A. C. Hannon, *Nucl. Instr. Meth. Phys. Res. A* **551**, 88 (2005). Proceedings of the E-MRS Fall Meeting 2004, Symposium D.

[333] M. Drakopoulos, T. Connolley, C. Reinhard, R. Atwood, O. Magdysyuk, N. Vo, M. Hart, L. Connor, B. Humphreys, G. Howell, S. Davies, T. Hill, G. Wilkin, U. Pedersen, A. Foster, N. De Maio, M. Basham, F. Yuan, K. Wanelik, *J. Synchrotron Radiat.* **22**, 828 (2015).

[334] A. K. Soper, E. R. Barney, *J. Appl. Cryst.* **44**, 714 (2011).

[335] M. G. Tucker, D. A. Keen, M. T. Dove, A. L. Goodwin, Q. Hui, *J. Phys. Condens. Matter* **19**, 335218 (2007).

[336] P. Hohenberg, W. Kohn, *Phys. Rev.* **136**, B864 (1964).

[337] W. Kohn, L. J. Sham, *Phys. Rev.* **140**, A1133 (1965).

[338] S. J. Clark, M. D. Segall, P. J. Hasnip, M. I. J. Probert, K. Refson, M. C. Payne, *Z. Kristallogr.* **220**, 567 (2004).

[339] R. Dovesi, R. Orlando, A. Erba, C. M. Zicovich-Wilson, B. Civalleri, S. Casassa, L. Maschio, M. Ferrabone, M. De La Pierre, P. D'Arco, Y. Noël, M. Causà, M. Rérat, B. Kirtman, *Int. J. Quant. Chem.* **114**, 1287 (2014).

[340] J. P. Perdew, V. N. Staroverov, J. Tao, G. E. Scuseria, *Phys. Rev. A* **78**, 052513 (2008).

[341] M. E. J. Newman, G. T. Barkema, *Monte Carlo Methods in Statistical Physics* (Oxford University Press, 1999).

[342] W. K. Hastings, *Biometrika* **57**, 97 (1970).

[343] S. Kirkpatrick, C. D. Gelatt, M. P. Vecchi, *Science* **220**, 671 (1983).

[344] D. Bertsimas, J. Tsitsiklis, *Statist. Sci.* **8**, 10 (1993).

[345] Y. Nourani, B. Andresen, *J. Phys. A* **31**, 8373 (1998).

[346] N. Duan, U. Kameswari, A. Sleight, *J. Am. Chem. Soc.* **121**, 10432 (1999).

[347] W. Cao, Q. Huang, Y. Rong, Y. Wang, J. Deng, J. Chen, X. Xing, *Inorg. Chem. Front.* **3**, 856 (2016).

[348] T. A. Mary, J. S. O. Evans, T. Vogt, A. W. Sleight, *Science* **272**, 90 (1996).

[349] J. Graham, A. D. Wadsley, J. H. Weymouth, L. S. Williams, *J. Am. Ceram. Soc.* **42**, 570 (1959).

[350] V. Trunov, L. Kovba, *Zh. Neorg. Khim.* **12** (1967).

[351] S. Ahrland, B. Nihlgard, L. Johansson, J. Albertsson, L. Nilsson, *Acta Chem. Scand.* **18** (1964).

[352] K. A. Kraus, T. A. Carlson, J. S. Johnson, *Nature* **177**, 1128 (1956).

[353] C. Serre, F. Millange, C. Thouvenot, M. Noguès, G. Marsolier, D. Louër, G. Férey, *J. Am. Chem. Soc.* **124**, 13519 (2002). PMID: 12418906.

[354] A. V. Neimark, F.-X. Coudert, A. Boutin, A. H. Fuchs, *J. Phys. Chem. Lett.* **1**, 445 (2010). PMID: 26700628.

[355] D. Bousquet, F.-X. Coudert, A. G. J. Fossati, A. V. Neimark, A. H. Fuchs, A. Boutin, *J. Chem. Phys.* **138**, 174706 (2013).

[356] S. Krause, V. Bon, I. Senkovska, U. Stoeck, D. Wallacher, D. M. Többens, S. Zander, R. S. Pillai, G. Maurin, F.-X. Coudert, S. Kaskel, *Nature* **532**, 348 EP (2016).

[357] U. G. Survey, Cement (2015).

[358] N. A. Banek, H. I. Baiz, A. Latigo, C. Lind, *J. Am. Chem. Soc.* **132**, 8278 (2010).

[359] M. I. Ahmad, M. Akinc, *Mater. Chem. Phys.* **145**, 403 (2014).

[360] J. S. O. Evans, T. A. Mary, T. Vogt, M. A. Subramanian, A. W. Sleight, *Chem. Mater.* **8**, 2809 (1996).

[361] C. A. Perottoni, J. A. H. d. Jornada, *Science* **280**, 886 (1998).

[362] T. Hashimoto, T. Katsube, Y. Morito, *Solid State Commun.* **116**, 129 (2000).

[363] J. S. O. Evans, W. I. F. David, A. W. Sleight, *Acta Cryst.* **B55**, 333 (1999).

[364] J. S. O. Evans, Z. Hu, J. D. Jorgensen, D. N. Argyriou, S. Short, A. W. Sleight, *Science* **275**, 61 (1997).

[365] J. D. Jorgensen, Z. Hu, S. Teslic, D. N. Argyriou, S. Short, J. S. O. Evans, A. W. Sleight, *Phys. Rev. B* **59**, 215 (1999).

[366] D. A. Keen, A. L. Goodwin, M. G. Tucker, M. T. Dove, J. S. O. Evans, W. A. Crichton, M. Brunelli, *Phys. Rev. Lett.* **98**, 225501 (2007).

[367] T. Varga, A. P. Wilkinson, A. C. Jupe, C. Lind, W. A. Bassett, C.-S. Zha, *Phys. Rev. B* **72**, 024117 (2005).

[368] A. K. Arora, V. S. Sastry, P. C. Sahu, T. A. Mary, *J. Phys. Condens. Matter* **16**, 1025 (2004).

[369] X. Yan, X. Ren, G. Sun, D. Li, X. Li, D. He, W. Yang, *Phys. Rev. B* **95**, 224205 (2017).

[370] C. Perottoni, J. Zorzi, J. da Jornada, *Solid State Commun.* **134**, 319 (2005).

[371] C. Martinek, F. A. Hummel, *J. Am. Ceram. Soc.* **51**, 227 (1968).

[372] J. Evans, *J. Chem. Soc., Dalton Trans.* pp. 3317–3326 (1999).

[373] M. T. Dove, M. Gambhir, K. D. Hammonds, V. Heine, A. K. Pryde, *Phase Transitions* **58**, 121 (1996).

[374] J. Tao, A. Sleight, *J. Solid State Chem.* **173**, 442 (2003).

[375] G. Ernst, C. Broholm, G. R. Kowach, A. P. Ramirez, *Nature* **396**, 147 (1998).

[376] A. P. Ramirez, G. R. Kowach, *Phys. Rev. Lett.* **80**, 4903 (1998).

[377] K. Wang, R. R. Reeber, *Appl. Phys. Lett.* **76**, 2203 (2000).

[378] D. Cao, F. Bridges, G. R. Kowach, A. P. Ramirez, *Phys. Rev. Lett.* **89**, 215902 (2002).

[379] D. Cao, F. Bridges, G. R. Kowach, A. P. Ramirez, *Phys. Rev. B* **68**, 014303 (2003).

[380] F. Bridges, T. Keiber, P. Juhas, S. J. L. Billinge, L. Sutton, J. Wilde, G. R. Kowach, *Phys. Rev. Lett.* **112**, 045505 (2014).

[381] C. A. Figueirêdo, C. Perottoni, *Phys. Rev. B* **75**, 184110 (2007).

[382] T. R. Ravindran, A. K. Arora, T. A. Mary, *J. Phys. Condens. Matter* **13**, 11573 (2001).

[383] V. Gava, A. L. Martinotto, C. A. Perottoni, *Phys. Rev. Lett.* **109**, 195503 (2012).

[384] M. K. Gupta, R. Mittal, S. L. Chaplot, *Phys. Rev. B* **88**, 014303 (2013).

[385] C. Adamo, V. Barone, *J. Chem. Phys.* **110**, 6158 (1999).

[386] J. P. Perdew, A. Ruzsinszky, G. I. Csonka, O. A. Vydrov, G. E. Scuseria, L. A. Constantin, X. Zhou, K. Burke, *Phys. Rev. Lett.* **102**, 039902 (2009).

[387] Y. Wang, G. Y. Cho, T. L. Hughes, E. Fradkin, *Phys. Rev. B* **93**, 134512 (2016).

[388] B. F. Hoskins, R. Robson, *J. Am. Chem. Soc.* **111**, 5962 (1989).

[389] T. D. Bennett, A. K. Cheetham, *Acc. Chem. Res.* **47**, 1555 (2014). PMID: 24707980.

[390] O. M. Yaghi, G. Li, H. Li, *Nature* **378**, 703 (1995).

[391] V. ctor Manuel Aceituno Melgar, J. Kim, M. R. Othman, *Ind. Eng. Chem. Res.* **28**, 1 (2015).

[392] B. Chen, Z. Yang, Y. Zhu, Y. Xia, *J. Mater. Chem. A* **2**, 16811 (2014).

[393] G. Lu, J. T. Hupp, *J. Am. Chem. Soc.* **132**, 7832 (2010). PMID: 20486704.

[394] P. Horcajada, C. Serre, M. Vallet-Regí, M. Sebban, F. Taulelle, G. Férey, *Angew. Chem. Int. Ed.* **45**, 5974 (2006).

[395] B. R. Pimentel, A. Parulkar, E. Zhou, N. A. Brunelli, R. P. Lively, *ChemSusChem* **7**, 3202 (2014).

[396] T. D. Bennett, P. J. Saines, D. A. Keen, J.-C. Tan, A. K. Cheetham, *Chem. Eur. J.* **19**, 7049 (2013).

[397] T. D. Bennett, J.-C. Tan, Y. Yue, E. Baxter, C. Ducati, N. J. Terrill, H. H. M. Yeung, Z. Zhou, W. Chen, S. Henke, A. K. Cheetham, G. N. Greaves, *Nat. Commun.* **6** (2015).

[398] T. D. Bennett, S. Cao, J. C. Tan, D. A. Keen, E. G. Bithell, P. J. Beldon, T. Friscic, A. K. Cheetham, *J. Am. Chem. Soc.* **133**, 14546 (2011). PMID: 21848328.

[399] M. G. Goesten, E. Stavitski, J. Juan-AlcaÃ±iz, A. MartiÃ±ez-Joaristi, A. V. Petukhov, F. Kapteijn, J. Gascon, *Catalysis Today* **205**, 120 (2013). Operando IV: The 4th International Congress on Operando Spectroscopy.

[400] A. Phan, C. J. Doonan, F. J. Uribe-Romo, C. B. Knobler, M. O'Keeffe, O. M. Yaghi, *Acc. Chem. Res.* **43**, 58 (2010).

[401] K. S. Park, Z. Ni, A. P. Côté, J. Y. Choi, R. Huang, F. J. Uribe-Romo, H. K. Chae, M. O'Keeffe, O. M. Yaghi, *Proc. Natl. Acad. Sci. U.S.A.* **103**, 10186 (2006).

[402] V. A. Blatov, G. D. Ilyushin, D. M. Proserpio, *Chem. Mater.* **25**, 412 (2013).

[403] B. M. Weckhuysen, J. Yu, *Chem. Soc. Rev.* **44**, 7022 (2015).

[404] A. U. Ortiz, A. Boutin, A. H. Fuchs, F. X. Coudert, *J. Phys. Chem. Lett.* **4**, 1861 (2013). PMID: 26283122.

[405] E. F. Baxter, T. D. Bennett, A. B. Cairns, N. J. Brownbill, A. L. Goodwin, D. A. Keen, P. A. Chater, F. Blanc, A. K. Cheetham, *Dalton Trans.* **45**, 4258 (2016).

[406] N. Masciocchi, G. A. Ardizzoia, G. Lamonica, A. Maspero, S. Galli, A. Sironi, *Inorg. Chem.* **40**, 6893 (2001).

[407] B. E. Warren, H. Krutter, O. Morningstar, *J. Am. Ceram. Soc.* **19**, 202 (1936).

[408] T. F. Soules, *J. Chem. Phys.* **73**, 4032 (1980).

[409] W. Soppe, C. van der Marel, W. van Gunsteren, H. den Hartog, *J. Non-Cryst. Solids* **103**, 201 (1988).

[410] J. Coey, P. Murphy, *J. Non-Cryst. Solids* **50**, 125 (1982).

[411] R. L. C. Vink, *J. Chem. Phys.* **140**, 104509 (2014).

[412] J. Greneche, J. Coey, *J. Phys. France* **51**, 231 (1990).

[413] V. Favre-Nicolin, R. Černý, *J. Appl. Cryst.* **35**, 734 (2002).

[414] N. L. Evans, P. M. M. Thygesen, H. L. B. BostrÃ¶m, E. M. Reynolds, I. E. Collings, A. E. Phillips, A. L. Goodwin, *J. Am. Chem. Soc.* **138**, 9393 (2016). PMID: 27414161.

[415] P. Adhikari, M. Xiong, N. Li, X. Zhao, P. Rulis, W.-Y. Ching, *J. Phys. Chem. C* (2016).

[416] T. Bayes, *Phil. Trans. R. Soc. Lond.* **53**, 370 (1763).

[417] K. P. Körding, D. M. Wolpert, *Nature* **427**, 244 EP (2004).

[418] A. Soper, *Chem. Phys.* **202**, 295 (1996).

[419] Y. Jiao, F. H. Stillinger, S. Torquato, *Phys. Rev. E* **81**, 011105 (2010).

[420] C. J. Gommes, Y. Jiao, S. Torquato, *Phys. Rev. Lett.* **108**, 080601 (2012).

[421] M. Frank, D. B. Carlson, M. S. Hunter, G. J. Williams, M. Messerschmidt, N. A. Zatsepin, A. Barty, W. H. Benner, K. Chu, A. T. Graf, S. P. Hau-Riege, R. A. Kirian, C. Padeste, T. Pardini, B. Pedrini, B. Segelke, M. M. Seibert, J. C. H. Spence, C.-J. Tsai, S. M. Lane, X.-D. Li, G. Schertler, S. Boutet, M. Coleman, J. E. Evans, *IUCrJ* **1**, 95 (2014).

[422] J. C. Kendrew, G. Bodo, H. M. Dintzis, R. G. Parrish, H. Wyckoff, D. C. Phillips, *Nature* **181**, 662 (1958).

[423] M. Sarikaya, C. Tamerler, A. K. Y. Jen, K. Schulten, F. Baneyx, *Nat. Mater.* **2**, 577 (2003).

[424] A. Krogh, B. Larsson, G. von Heijne, E. L. Sonnhammer, *J. Mol. Biol.* **305**, 567 (2001).

[425] J. P. Overington, B. Al-Lazikani, A. L. Hopkins, *Nat. Rev. Drug Discov.* **5**, 993 (2006).

[426] M. Baker, *Nat. Methods* **7**, 429 (2010).

[427] S. Skou, R. E. Gillilan, N. Ando, *Nat. Protocols* **9**, 1727 EP (2014).

[428] F. H. C. Crick, L. Barnett, S. Brenner, R. J. Watts-Tobin, *Nature* **192**, 1227 (1961).

[429] G. A. Khoury, R. C. Baliban, C. A. Floudas, *Sci. Rep.* **1**, 90 (2011).

[430] P. Edman, *Acta Chem. Scand.* **4**, 283 (1950).

[431] J. Clarke, H.-C. Wu, L. Jayasinghe, A. Patel, S. Reid, H. Bayley, *Nat. Nanotech.* **4**, 265 (2009).

[432] D. F. Hunt, J. R. Yates, J. Shabanowitz, S. Winston, C. R. Hauer, *Proc. Natl. Acad. Sci. U.S.A.* **83**, 6233 (1986).

[433] A. M. Lesk, *Introduction to Protein Science: Architecture, Function and Genomics* (Oxford University Press, 2016), third edn.

[434] L. Pauling, R. B. Corey, *Proc. Natl. Acad. Sci. U.S.A.* **37**, 251 (1951).

[435] L. Pauling, R. B. Corey, *Proc. Natl. Acad. Sci. U.S.A.* **37**, 235 (1951).

[436] K. A. Dill, S. B. Ozkan, M. S. Shell, T. R. Weikl, *Annu. Rev. Biophys.* **37**, 289 (2008). PMID: 18573083.

[437] A. Gronenborn, D. Filpula, N. Essig, A. Achari, M. Whitlow, P. Wingfield, G. Clore, *Science* **253**, 657 (1991).

[438] M. V. Berjanskii, D. S. Wishart, *J. Am. Chem. Soc.* **127**, 14970 (2005).

[439] M. V. Berjanskii, D. S. Wishart, *J. Biomol. NMR* **40**, 31 (2008).

[440] W. Kabsch, C. Sander, *Biopolymers* **22**, 2577 (1983).

[441] R. A. Laskowski, D. S. Moss, J. M. Thornton, *J. Mol. Biol.* **231**, 1049 (1993).

[442] G. Ramachandran, C. Ramakrishnan, V. Sasisekharan, *J. Mol. Biol.* **7**, 95 (1963).

[443] G. Ramachandran, V. Sasisekharan, *Advances in Protein Chemistry* **23**, 283 (1968).

[444] S. Hovmöller, T. Zhou, T. Ohlson, *Acta Cryst. D* **58**, 768 (2002).

[445] R. A. Laskowski, M. W. MacArthur, D. S. Moss, J. M. Thornton, *J. Appl. Cryst.* **26**, 283 (1993).

[446] A. L. Morris, M. W. MacArthur, E. G. Hutchinson, J. M. Thornton, *Proteins* **12**, 345 (1992).

[447] D. Pal, P. Chakrabarti, *Biopolymers* **63**, 195 (2002).

[448] M. V. Shapovalov, J. Dunbrack, Roland L., *Structure* **19**, 844 (2011).

[449] J. Drenth, *Principles of protein X-ray crystallography* (Springer-Verlag, New York, 2007).

[450] E. P. Carpenter, K. Beis, A. D. Cameron, S. Iwata, *Curr. Opin. Struct. Biol.* **18**, 581 (2008).

[451] J. L. Parker, S. Newstead, *Membrane Protein Crystallisation: Current Trends and Future Perspectives* (Springer International Publishing, Cham, 2016), pp. 61–72.

[452] Y. Teng, K. Moffat, *J. Synchrotron Rad.* **7**, 313 (2000).

[453] E. F. Garman, R. L. Owen, *Acta Cryst. D* **62**, 32 (2006).

[454] L. Redecke, *et al.*, *Science* **339**, 227 (2013).

[455] H. N. Chapman, *et al.*, *Nature* **470**, 73 (2011).

[456] K. Hirata, K. Shinzawa-Itoh, N. Yano, S. Takemura, K. Kato, M. Hatanaka, K. Muramoto, T. Kawahara, T. Tsukihara, E. Yamashita, K. Tono, G. Ueno, T. Hikima, H. Murakami, Y. Inubushi, M. Yabashi, T. Ishikawa, M. Yamamoto, T. Ogura, H. Sugimoto, J.-R. Shen, S. Yoshikawa, H. Ago, *Nat. Methods* **11**, 734 (2014).

[457] M. Beer, J. Frank, K.-J. Hanszen, E. Kellenberger, R. C. Williams, *Quart. Rev. Biophys.* **7**, 211 (1974).

[458] S. Zhang, D. Vasishtan, M. Xu, M. Topf, F. Alber, *Bioinformatics* **26**, i261 (2010).

[459] R. Henderson, P. N. T. Unwin, *Nature* **257**, 28 (1975).

[460] E. Knapek, J. Dubochet, *J. Mol. Biol.* **141**, 147 (1980).

[461] R. Henderson, *Proc. R. Soc. Lond. B* **241**, 6 (1990).

[462] Z. H. Zhou, *Adv. Protein Chem. Struct. Biol.* **82**, 1 (2013).

[463] J. Snijder, J. M. Schuller, A. Wiegard, P. Lössl, N. Schmelling, I. M. Axmann, J. M. Plitzko, F. Förster, A. J. R. Heck, *Science* **355**, 1181 (2017).

[464] J. Kosinski, S. Mosalaganti, A. von Appen, R. Teimer, A. L. DiGuilio, W. Wan, K. H. Bui, W. J. Hagen, J. A. G. Briggs, J. S. Glavy, E. Hurt, M. Beck, *Science* **352**, 363 (2016).

[465] D. H. Lin, T. Stuwe, S. Schilbach, E. J. Rundlet, T. Perriches, G. Mobbs, Y. Fan, K. Thierbach, F. M. Huber, L. N. Collins, A. M. Davenport, Y. E. Jeon, A. Hoelz, *Science* **352** (2016).

[466] I. Nederlof, *Acta Cryst. A* **73**, C66 (2017).

[467] H. Matsunami, C. S. Barker, Y.-H. Yoon, M. Wolf, F. A. Samatey, *Nat. Commun.* **7**, 13425 EP (2016).

[468] M. Kovermann, P. Rogne, M. Wolf-Watz, *Quart. Rev. Biophys.* **49** (2016).

[469] C. Goebl, T. Madl, B. Simon, M. Sattler, *Prog. Nucl. Magn. Reson. Spectrosc.* **80**, 26 (2014).

[470] Y. Tang, Y. J. Huang, T. A. Hopf, C. Sander, D. S. Marks, G. T. Montelione, *Nat. Methods* **12**, 751 (2015).

[471] J. Wang, X. Zuo, P. Yu, I.-J. L. Byeon, J. Jung, X. Wang, M. Dyba, S. Seifert, C. D. Schwieters, J. Qin, A. M. Gronenborn, Y.-X. Wang, *J. Am. Chem. Soc.* **131**, 10507 (2009).

[472] A. Grishaev, J. Wu, J. Trewhella, A. Bax, *J. Am. Chem. Soc.* **127**, 16621 (2005).

[473] M. A. Martí-Renom, A. C. Stuart, A. Fiser, R. Sánchez, F. Melo, A. Šali, *Annu. Rev. Biophys.* **29**, 291 (2000).

[474] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, *J. Mol. Biol.* **215**, 403 (1990).

[475] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, D. J. Lipman, *Nucleic Acids Res.* **25**, 3389 (1997).

[476] W. R. a. Pearson, [*5*] *Rapid and sensitive sequence comparison with FASTP and FASTA* (Academic Press, 1990), vol. Volume 183, pp. 63–98.

[477] K. T. Simons, C. Kooperberg, E. Huang, D. Baker, *J. Mol. Biol.* **268**, 209 (1997).

[478] J. M. Thompson, N. G. Sgourakis, G. Liu, P. Rossi, Y. Tang, J. L. Mills, T. Szyperski, G. T. Montelione, D. Baker, *Proc. Natl. Acad. Sci. U.S.A.* **109**, 9875 (2012).

[479] K. T. Simons, I. Ruczinski, C. Kooperberg, B. A. Fox, C. Bystroff, D. Baker, *Proteins* **34**, 82 (1999).

[480] C. A. Rohl, C. E. Strauss, D. Chivian, D. Baker, *Proteins* **55**, 656 (2004).

[481] Y. Song, F. DiMaio, R. Y.-R. Wang, D. Kim, C. Miles, T. Brunette, J. Thompson, D. Baker, *Structure* **21**, 1735 (2013).

[482] K. M. S. Misura, D. Chivian, C. A. Rohl, D. E. Kim, D. Baker, *Proc. Natl. Acad. Sci. U.S.A.* **103**, 5361 (2006).

[483] N. E. Hafsa, D. S. Wishart, *J. Biomol. NMR* **60**, 131 (2014).

[484] S. Neal, A. M. Nip, H. Zhang, D. S. Wishart, *J. Biomol. NMR* **26**, 215 (2003).

[485] Y. Shen, A. Bax, *J. Biomol. NMR* **38**, 289 (2007).

[486] B. Han, Y. Liu, S. W. Ginzinger, D. S. Wishart, *J. Biomol. NMR* **50**, 43 (2011).

[487] Y. Shen, A. Bax, *J. Biomol. NMR* **48**, 13 (2010).

[488] D. Li, R. Brueschweiler, *J. Biomol. NMR* **62**, 403 (2015).

[489] M.-S. Cheung, M. L. Maguire, T. J. Stevens, R. W. Broadhurst, *J. Magn. Reson.* **202**, 223 (2010).

[490] Y. Shen, A. Bax, *J. Biomol. NMR* **56**, 227 (2013).

[491] Y. Shen, A. Bax, *Nat. Methods* **12**, 747 (2015).

[492] C. D. Schwieters, J. J. Kuszewski, N. Tjandra, G. Marius Clore, *J. Magn. Reson.* **160**, 65 (2003).

[493] J. R. Allison, P. Varnai, C. M. Dobson, M. Vendruscolo, *J. Am. Chem. Soc.* **131**, 18314 (2009). PMID: 20028147.

[494] A. Cavalli, X. Salvatella, C. M. Dobson, M. Vendruscolo, *Proc. Natl. Acad. Sci. U.S.A.* **104**, 9615 (2007).

[495] P. Robustelli, A. Cavalli, C. M. Dobson, M. Vendruscolo, X. Salvatella, *J. Phys. Chem. B* **113**, 7890 (2009).

[496] Y. Shen, O. Lange, F. Delaglio, P. Rossi, J. M. Aramini, G. Liu, A. Eletsky, Y. Wu, K. K. Singarapu, A. Lemak, A. Ignatchenko, C. H. Arrowsmith, T. Szyperski, G. T. Montelione, D. Baker, A. Bax, *Proc. Natl. Acad. Sci. U.S.A.* **105**, 4685 (2008).

[497] D. S. Wishart, D. Arndt, M. Berjanskii, P. Tang, J. Zhou, G. Lin, *Nucleic Acids Res.* **36**, W496 (2008).

[498] D. S. Wishart, D. A. Case, V. D. Thomas L. James, U. Schmitz, [*1*] - *Use of Chemical Shifts in Macromolecular Structure Determination* (Academic Press, 2002), vol. Volume 338, pp. 3–34.

[499] W. Boomsma, K. V. Mardia, C. C. Taylor, J. Ferkinghoff-Borg, A. Krogh, T. Hamelryck, *Proc. Natl. Acad. Sci. U.S.A.* **105**, 8932 (2008).

[500] W. Boomsma, P. Tian, J. Frellsen, J. Ferkinghoff-Borg, T. Hamelryck, K. Lindorff-Larsen, M. Vendruscolo, *Proc. Natl. Acad. Sci. U.S.A.* **111**, 13852 (2014).

[501] K. J. Kohlhoff, P. Robustelli, A. Cavalli, X. Salvatella, M. Vendruscolo, *J. Am. Chem. Soc.* **131**, 13894 (2009).

[502] P. Robustelli, K. Kohlhoff, A. Cavalli, M. Vendruscolo, *Structure* **18**, 923 (2010).

[503] M. Berjanskii, D. Arndt, Y. Liang, D. S. Wishart, *J. Biomol. NMR* **63**, 255 (2015).

[504] W. Lee, C. M. Petit, G. Cornilescu, J. L. Stark, J. L. Markley, *J. Biomol. NMR* **65**, 51 (2016).

[505] D. S. Berkholz, M. V. Shapovalov, J. Dunbrack, Roland L., P. A. Karplus, *Structure* **17**, 1316 (2009).

[506] Y. Duan, C. Wu, S. Chowdhury, M. C. Lee, G. Xiong, W. Zhang, R. Yang, P. Cieplak, R. Luo, T. Lee, J. Caldwell, J. Wang, P. Kollman, *J. Comput. Chem.* **24**, 1999 (2003).

[507] N. S. Burkoff, C. Várnai, S. A. Wells, D. L. Wild, *Biophys. J.* **102**, 878 (2012).

[508] R. Cano, C. Sordo, J. M. Gutiérrez, *Applications of Bayesian Networks in Meteorology* (Springer Berlin Heidelberg, Berlin, Heidelberg, 2004), pp. 309–328.

[509] G. Kernell, *Political Analysis* **17**, 215 (2009).

[510] D. Ting, G. Wang, M. Shapovalov, R. Mitra, M. I. Jordan, R. L. Dunbrack, Jr, *PLoS Comput. Biol.* **6**, 1 (2010).

[511] T. C. Terwilliger, D. Eisenberg, *J. Biol. Chem.* **257**, 6016 (1982).

[512] B. A. Reva, A. V. Finkelstein, J. Skolnick, *Fold. Des.* **3**, 141 (1998).

[513] N. D. Mermin, *Rev. Mod. Phys.* **51**, 591 (1979).

[514] J. M. Ziman, *Models of disorder* (Cambridge University Press, 1979).

[515] M. Baise, P. M. Maffettone, F. Trousselet, N. P. Funnell, F. m. c.-X. Coudert, A. L. Goodwin, *Phys. Rev. Lett.* **120**, 265501 (2018).

[516] A. R. Overy, R. Pandya, P. M. Maffettone, P. A. Chater, A. Simonov, A. L. Goodwin, Extreme cooperative swelling in topologically disordered fibre entanglements (2018). ArXiv: 1802.07629.

[517] P. Bosetti, M. Ragni, M. Leoni, *Acta Cryst. A* **73**, C562 (2017).

[518] J. G. P. Wicker, L. M. Crowley, O. Robshaw, E. J. Little, S. P. Stokes, R. I. Cooper, S. E. Lawrence, *CrystEngComm* **19**, 5336 (2017).

[519] M. Pillong, C. Marx, P. Piechon, J. G. P. Wicker, R. I. Cooper, T. Wagner, *CrystEngComm* **19**, 3737 (2017).

[520] M. H. S. Segler, M. Preuss, M. P. Waller, *Nature* **555**, 604 EP (2018).

[521] T. Klucznik, B. Mikulak-Klucznik, M. P. McCormack, H. Lima, S. Szymkuć, M. Bhowmick, K. Molga, Y. Zhou, L. Rickershauser, E. P. Gajewska, A. Toutchkine, P. Dittwald, M. P. Startek, G. J. Kirkovits, R. Roszak, A. Adamski, B. Sieredzińska, M. Mrksich, S. L. Trice, B. A. Grzybowski, *Chem* **4**, 522 (2018).

[522] B. Sanchez-Lengeling, A. Aspuru-Guzik, *Science* **361**, 360 (2018).

[523] R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams, A. Aspuru-Guzik, *ACS Cent. Sci.* **4**, 268 (2018).

[524] I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning* (MIT Press, 2016). `http://www.deeplearningbook.org`.

[525] A. Wlodawer, W. Minor, Z. Dauter, M. Jaskolski, *FEBS J.* **280**, 5705 (2013).

[526] R. Dovesi, R. Orlando, B. Civalleri, C. Roetti, V. R. Saunders, C. M. Zicovich-Wilson, *Z. Krist.* **220**, 571 (2005).

[527] C. Gatti, V. R. Saunders, C. Roetti, *J. Chem. Phys.* **101**, 10686 (1994).

[528] G. Sophia, P. Baranek, C. Sarrazin, M. Rérat, R. Dovesi, *Phase Transitions* **86**, 1069 (2013).

[529] D. Andrae, U. H außermann, M. Dolg, H. Stoll, H. Preuß, *Theor. Chim. Acta* **77**, 123 (1990).

[530] H. J. Monkhorst, J. D. Pack, *Phys. Rev. B* **13**, 5188 (1976).

[531] J. P. Perdew, A. Zunger, *Phys. Rev. B* **23**, 5048 (1981).

[532] D. S. Wishart, B. D. Sykes, F. M. Richards, *J. Mol. Biol.* **222**, 311 (1991).

[533] A. Pastore, V. Saudek, *J. Magn. Reson.* **90**, 165 (1990).

[534] S. Spera, A. Bax, *J. Am. Chem. Soc.* **113**, 5490 (1991).

[535] X.-P. Xu, D. A. Case, *J. Biomol. NMR* **21**, 321 (2001).

[536] A. T. Frank, S.-H. Bae, A. C. Stelzer, *J. Phys. Chem. B* **117**, 13497 (2013).

[537] D. S. Wishart, M. S. Watson, R. F. Boyko, B. D. Sykes, *J. Biomol. NMR* **10**, 329 (1997).

[538] A. S. Christensen, T. E. Linnet, M. Borg, W. Boomsma, K. Lindorff-Larsen, T. Hamelryck, J. H. Jensen, *PLoS ONE* **8**, e84123 (2014).

[539] S. W. Ginzinger, M. Coles, *J. Biomol. NMR* **43**, 179 (2009).

[540] S. Schwarzinger, G. J. A. Kroon, T. R. Foss, P. E. Wright, H. J. Dyson, *J. Biomol. NMR* **18**, 43 (2000).

[541] S. Schwarzinger, G. J. A. Kroon, T. R. Foss, J. Chung, P. E. Wright, , H. J. Dyson, *J. Am. Chem. Soc.* **123**, 2970 (2001).

[542] Y. Wang, D. S. Wishart, *J. Biomol. NMR* **31**, 143 (2005).

[543] H. Dashti, M. Tonelli, W. Lee, W. M. Westler, G. Cornilescu, E. L. Ulrich, J. L. Markley, *J. Biomol. NMR* **64**, 17 (2016).

[544] H. Zhang, S. Neal, D. S. Wishart, *J. Biomol. NMR* **25**, 173 (2003).

[545] B. R. Seavey, E. A. Farr, W. M. Westler, J. L. Markley, *J. Biomol. NMR* **1**, 217 (1991).

[546] J. F. Doreleijers, S. Mading, D. Maziuk, K. Sojourner, L. Yin, J. Zhu, J. L. Markley, E. L. Ulrich, *J. Biomol. NMR* **26**, 139 (2003).

[547] G. Cornilescu, F. Delaglio, A. Bax, *J. Biomol. NMR* **13**, 289 (1999).

[548] X.-P. Xu, D. A. Case, *Biopolymers* **65**, 408 (2002).

[549] A. B. Sahakyan, W. F. Vranken, A. Cavalli, M. Vendruscolo, *Angew. Chem. Int. Ed.* **50**, 9620 (2011).

[550] A. B. Sahakyan, W. F. Vranken, A. Cavalli, M. Vendruscolo, *J. Biomol. NMR* **50**, 331 (2011).

[551] J. Meiler, *J. Biomol. NMR* **26**, 25 (2003).

[552] D.-W. Li, R. Brüschweiler, *J. Biomol. NMR* **54**, 257 (2012).

[553] A. Kumari, K. Dorai, *J. Mol. Struct.* **1041**, 200 (2013).

[554] C.-C. Wang, W.-C. Lai, W.-J. Chuang, *J. Biomol. NMR* **59**, 175 (2014).

[555] D. S. Wishart, B. D. Sykes, F. M. Richards, *Biochemistry* **31**, 1647 (1992).

[556] S. Neal, M. Berjanskii, H. Zhang, D. S. Wishart, *Magn. Reson. Chem.* **44**, S158 (2006).

[557] J. Song, H. Tan, M. Wang, G. I. Webb, T. Akutsu, *PLOS ONE* **7** (2012).

[558] S. Wu, Y. Zhang, *PLoS ONE* **3** (2008).

[559] P. Kountouris, J. D. Hirst, *BMC Bioinformatics* **10** (2009).

[560] Y. Shen, F. Delaglio, G. Cornilescu, A. Bax, *J. Biomol. NMR* **44**, 213 (2009).

[561] M. V. Berjanskii, S. Neal, D. S. Wishart, *Nucleic Acids Res.* **34**, W63 (2006).

[562] L. Willard, A. Ranjan, H. Zhang, H. Monzavi, R. F. Boyko, B. D. Sykes, D. S. Wishart, *Nucleic Acids Res.* **31**, 3316 (2003).

[563] A. A. Canutescu, A. A. Shelenkov, R. L. Dunbrack, *Protein Science* **12**, 2001 (2003).

[564] D. S. Wishart, B. D. Sykes, *J. Biomol. NMR* **4**, 171 (1994).

[565] N. E. Hafsa, D. Arndt, D. S. Wishart, *Nucleic Acids Res.* **43**, W370 (2015).

[566] M. V. Berjanskii, D. S. Wishart, *J. Am. Chem. Soc.* **135**, 14536 (2013).

[567] R. Kaźmierkiewicz, A. Liwo, H. A. Scheraga, *Biophys. Chem.* **100**, 261 (2002).

[568] J. Ryu, M. Lee, J. Cha, R. A. Laskowski, S. E. Ryu, D.-S. Kim, *Nucleic Acids Res.* **44**, W416 (2016).

[569] J. K. Strayer, *Geometric Linear Programming* (Springer New York, New York, NY, 1989), pp. 5–26.

Publications

M. Baise, P. M. Maffettone, F. Trousselet, N. P. Funnell, F.-X. Coudert, and A. L. Goodwin, *Phys. Rev. Lett.* **120**, 265501 (2018).

Editors' Suggestion | Featured in Physics

Negative Hydration Expansion in ZrW_2O_8: Microscopic Mechanism, Spaghetti Dynamics, and Negative Thermal Expansion

Mia Baise,[1,2] Phillip M. Maffettone,[1] Fabien Trousselet,[3] Nicholas P. Funnell,[1,4] François-Xavier Coudert,[3] and Andrew L. Goodwin[1,*]

[1]*Department of Chemistry, University of Oxford, Inorganic Chemistry Laboratory, South Parks Road, Oxford OX1 3QR, United Kingdom*

[2]*Department of Chemistry, University College London, Gower Street, London WC1E 6BT, United Kingdom*

[3]*Chimie ParisTech, PSL Research University, CNRS, Institut de Recherche de Chimie Paris, 75005 Paris, France*

[4]*ISIS Neutron and Muon Facility, Rutherford Appleton Laboratory, Didcot OX11 0QX, United Kingdom*

(Received 13 April 2018; published 29 June 2018)

We use a combination of x-ray diffraction, total scattering, and quantum mechanical calculations to determine the mechanism responsible for hydration-driven contraction in ZrW_2O_8. The inclusion of H_2O molecules within the ZrW_2O_8 network drives the concerted formation of new W—O bonds to give one-dimensional $(—W—O—)_n$ strings. The topology of the ZrW_2O_8 network is such that there is no unique choice for the string trajectories: the same local changes in coordination can propagate with a large number of different periodicities. Consequently, $ZrW_2O_8 \cdot H_2O$ is heavily disordered, with each configuration of strings forming a dense aperiodic "spaghetti." This new connectivity contracts the unit cell via large shifts in the Zr and W atom positions. Fluctuations of the undistorted parent structure towards this spaghetti phase emerge as the key negative thermal expansion (NTE) phonon modes in ZrW_2O_8 itself. The large relative density of NTE phonon modes in ZrW_2O_8 actually reflects the degeneracy of volume-contracting spaghetti excitations, itself a function of the particular topology of this remarkable material.

DOI: 10.1103/PhysRevLett.120.265501

Zirconium tungstate, ZrW_2O_8, is an important material because it harbors a number of useful and counterintuitive mechanical responses to external stimuli. First, on heating, its molar volume *decreases* such that the material is 2% denser at 1050 K than at 0.3 K [1–4]. This negative thermal expansion (NTE) effect is about as strong as the usual positive thermal expansion (PTE) of conventional materials, and so it can be exploited in athermal composites [5–7]. Second, under hydrostatic pressure, ZrW_2O_8 amorphizes unusually easily, transforming into a disordered phase that is denser by more than 20% [8–10]. This switching between crystalline and amorphous states is related to the mechanism exploited for data storage in phase-change chalcogenides [11]. And, third, hydration of ZrW_2O_8 gives the monohydrate $ZrW_2O_8 \cdot H_2O$, which has a molar volume ∼10% smaller than ZrW_2O_8 itself [12,13]. In this sense, the material—which has no appreciable pore network—is nonetheless a kind of "inverse sponge" that shrinks instead of swells when it takes up water. We call this effect "negative hydration expansion" (NHE). NHE is known to occur in anisotropic "breathing" frameworks such as MIL-53 [14,15], but it is rare in cubic systems [16,17]; pore sizes are typically 5–20 Å in all cases. The phenomenon is largely unexplored—and particularly so in ZrW_2O_8—yet NHE may eventually have applications in counteracting hydration-driven swelling of construction materials (e.g., concretes and clays).

While the mechanism responsible for NHE in ZrW_2O_8 remains unclear [12,13], both NTE and pressure-induced amorphization (PIA) have been studied extensively and are understood to arise from a peculiar phonon spectrum. The material supports an extensive family of low-energy modes with large and negative Grüneisen parameters [18–23]; i.e., their frequencies soften with pressure—the thermodynamic requirement for NTE in cubic systems [6]. Under hydrostatic pressure, the energies of these modes fall to zero [8,24]. Because the modes are distributed throughout the Brillouin zone (BZ), the resulting structural transition is associated with a family of modulation periodicities and, hence, amorphization [10].

Despite the many experimental and computational studies of ZrW_2O_8, the microscopic origin of its phonon anomalies remains contentious. It is the sheer number of NTE vibrational modes that is difficult to explain: each one appears to involve a remarkably different combination of polyhedral translations, rotations, and distortions [22]. Early indications of the importance of "rigid unit modes" (RUMs) [25–27] have increasingly lost experimental [28,29] and computational [22,23] support, with consensus only emerging regarding the role of polyhedral distortions and the "skipping rope" or "tension" mechanism proposed in Ref. [3].

Intriguingly, hydration switches off NTE in ZrW_2O_8: $ZrW_2O_8 \cdot H_2O$ is a conventional PTE material [13,30].

 © 2018 American Physical Society

One implication is that the mechanism responsible for NHE saturates the same set of displacements from which NTE arises. Consequently, we sought to determine the structure of $ZrW_2O_8 \cdot H_2O$, establish a mechanism for NHE, and explore the relationship between NHE and NTE in ZrW_2O_8 itself.

In this Letter, we show that hydration of ZrW_2O_8 proceeds via the concerted formation of one-dimensional strings of new W—O bonds. The topology of the ZrW_2O_8 network is such that there is no unique choice for the trajectory of these strings: the same local changes in coordination can propagate with a macroscopically large number of different periodicities. Each configuration of strings forms a dense "spaghetti" that describes an additional connectivity responsible for contracting the unit cell via surprisingly large shifts in the Zr and W atom positions. So, $ZrW_2O_8 \cdot H_2O$ is heavily disordered, albeit in a highly correlated fashion. Remarkably, fluctuations of the anhydrous parent towards this spaghetti phase emerge as the key NTE phonon modes in ZrW_2O_8. Consequently, the density and distribution of NTE modes throughout reciprocal space actually reflect the degeneracy of volume-contracting spaghetti excitations, itself a function of network topology.

The starting point for our study was to prepare and characterize polycrystalline samples of both ZrW_2O_8 and its monohydrate, following the hydrothermal methodologies of Ref. [31]. The corresponding x-ray powder diffraction patterns [Fig. 1] reflect all hydration-driven structural responses noted in earlier studies [12,13]: the crystal symmetry increases on hydration from $P2_13$ to $Pa\bar{3}$, the lattice parameter a decreases from 9.153(3) to 8.876(4) Å, and there is a marked increase in strain-driven peak broadening. We found hydration and dehydration cycles (~36 h) to be reversible in all respects, including the appearance and disappearance of strain effects. Rietveld refinement of the parent ZrW_2O_8 diffraction pattern using the published structure [3] gave an excellent fit, but we encountered the same difficulty noted in [12] of obtaining a structural model for the monohydrate: the scattering density at the Zr and W sites indicated severe positional disorder, and the O atom positions were impossible to determine.

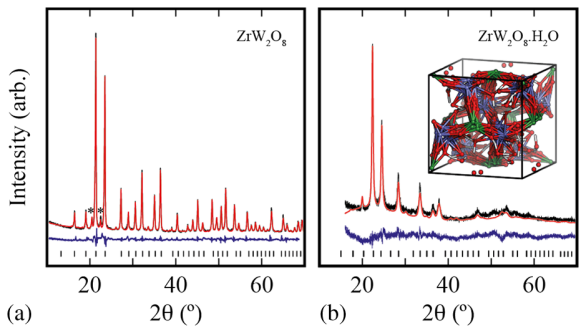


FIG. 1. X-ray diffraction patterns ($\lambda = 1.54056$ Å) of (a) anhydrous and (b) hydrated ZrW_2O_8: data in black, fits in red, difference (data − fit) in blue, and reflection positions as tick marks. The fit in (a) is from a Rietveld refinement using the coordinates of Ref. [3]; $*$ = minor impurity. The fit in (b) used our DFT coordinates, with $Pa\bar{3}$ symmetry applied (inset: the corresponding structural model represents the configurational average of a disordered spaghetti).

Given the difficulty of directly interpreting the x-ray diffraction pattern of $ZrW_2O_8 \cdot H_2O$, we used quantum mechanical calculations to determine an energetically-sensible candidate structure for this phase. The structure of ZrW_2O_8 contains a single symmetry-related set of cavities, with each cavity of the correct size to host one H_2O molecule [Fig. 2(a)]. There is one cavity per formula unit, and four formula units per unit cell. The cavities are positioned on the $4a$ Wyckoff sites and coincide with threefold rotation axes of the $P2_13$ space group. Clearly, this point symmetry is incompatible with that of the H_2O molecule, so there is a choice of three possible orientations for each of the four water molecules to be placed in a single unit cell. This situation contrasts that of NH_3 binding in $ZrW_2O_8 \cdot NH_3$, for which NH_3 orientations are ordered and

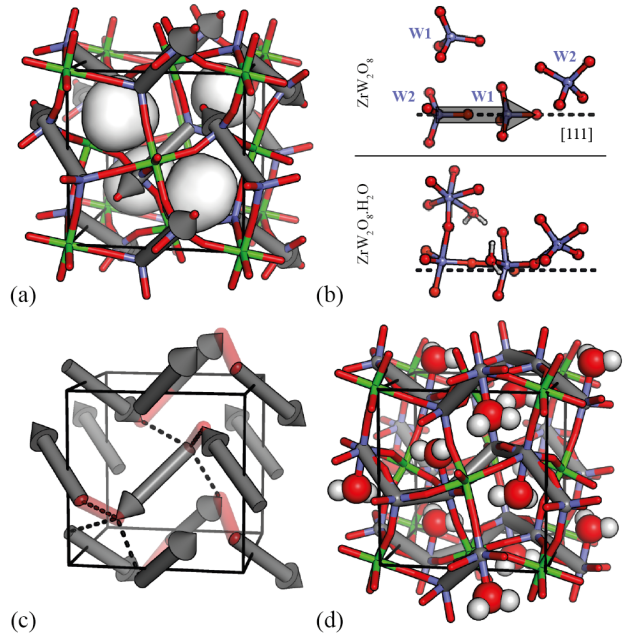


FIG. 2. (a) The crystal structure of ZrW_2O_8: Zr in green, W in blue, O in red, and the four cavities of each unit cell in surface representation. (b) DFT geometries of the key structural fragment before (top) and after (bottom) hydration. Note the formation of additional W—O—W bonds. W2…W1 pairs are indicated by a grey arrow; the location of these pairs within the structure is shown in (a). (c) On hydration, neighbouring W2…W1 pairs connect to form W—O—W— chains. Note the symmetry-equivalent choices for these new connections (dashed lines). The red cylinders denote the connections made in the DFT ($P2_12_12_1$) structure. (d) The final $P2_12_12_1$ structure, with W—O—W strings shown in grey and H_2O molecules shown in space-filling representation.

$P2_13$ symmetry is maintained [32]—the molecular point symmetry of NH_3 ($3m$) being a supergroup of the $4a$ site symmetry (3) [33]. Instead, the arrangement of H_2O molecules that results in the least severe symmetry lowering gives $P2_12_12_1$ symmetry. We will come to show that $ZrW_2O_8 \cdot H_2O$ is, in fact, heavily disordered, with $Pa\bar{3}$ symmetry emerging as a configurational average over states related to this $P2_12_12_1$ arrangement we now consider in detail.

Starting with this highest-symmetry H_2O arrangement, we carried out density functional theory (DFT) relaxations at two different levels of theory (the local density approximation (LDA) and the PBESOL0 functional; see [34]). Energy minimization was sensitive to the precise starting positions of the additional H_2O molecules within the framework cavities, yet we were able to identify a series of closely-related low-energy structures with molar volumes smaller by 6–8% than that of ZrW_2O_8. The corresponding enthalpies of water adsorption were in the range 60–90 kJ mol^{-1} (PBESOL0 functional), indicating strong physisorption [48]. Hence, our calculations are entirely consistent with a NHE effect of the same magnitude as that observed experimentally.

The structure of our lowest-energy $P2_12_12_1$ model for $ZrW_2O_8 \cdot H_2O$ reveals how hydration brings about a contraction of the unit cell. In this model, each H_2O molecule is chemically bonded to one of the two crystallographically-distinct W atoms [W1 in Fig. 2(b)] along a direction that is no longer parallel to a triad axis of the original unit cell. This new coordination induces a series of coupled rotations and translations of both types of W-centered polyhedra that collectively results in a set of new W—O—W connections. These connections draw the framework in on itself: the terminal W1—O and W2—O bonds of the original structure are forced within bonding distance of nearby W2 and W1 atoms, respectively. Not only does this connectivity reduce the average W…W separation—and in turn the unit-cell volume—but it results in substantial displacement of both Zr and W atoms away from their original sites. Indeed, provided we can account for the $Pa\bar{3}$ crystal symmetry, these displacements are fully consistent with the experiment: Figure 1(b) shows the excellent match between the experimental diffraction pattern and that derived from our DFT coordinates. Moreover, our model accounts for the NTE/PTE switch on hydration: zone-center phonon calculations show a dramatic reduction in the number of NTE modes, with the corresponding coefficient of thermal expansion now positive [34].

Given the existence of this low-energy $P2_12_12_1$ structure, why would the experimental system retain cubic symmetry? We suggest that the key is the existence of a diverging number of different ways of propagating the local changes in coordination identified by our DFT calculations. This degeneracy, we claim, leads to a strongly disordered state. We proceed to describe this state and in turn test our claims using experimental x-ray pair distribution function (PDF) measurements.

Consider the new W—O—W bonds formed during hydration [Fig. 2(b)]. While there is only one W1 atom accessible to each terminal W2—O bond, there are three equally-spaced W2 atoms that any terminal W1—O bond might approach. These three choices of W2 atoms are related by the same triad axis that is broken by the symmetry of an H_2O molecule. So, in selecting one of three possible states, the particular orientation of each H_2O molecule effectively determines the trajectory of a corresponding W1—O → W2 linkage [Fig. 2(c)]. These new W—O—W bonds form one-dimensional strings that in our DFT structure reflect the same $P2_12_12_1$ crystal symmetry of the H_2O molecule arrangements [Fig. 2(d)], but more disordered H_2O arrangements are associated with equally disordered arrangements of W—O—W strings (we will call collections of strings a "spaghetti," noting the relationship to the "loop" models [49] of, e.g., Coulomb phases [50,51]). Such states are not random. Note that in the DFT structure each W1—O bond approaches only one W2 atom, and each W2 atom is approached by only one W1—O bond. Preserving these local constraints enforces correlations between the new W—O—W bonds formed, and in turn between the orientations of successive H_2O molecules. It can be shown [34] that the number of configurations satisfying these local rules grows exponentially with system size to give a macroscopic number of viable hydration states; the corresponding configurational entropy is $S_{\mathrm{config}} \gtrsim R\ln(9/8)$. Hence, there is an entropic driving force for $ZrW_2O_8 \cdot H_2O$ to adopt spaghetti states, with the bulk crystal symmetry reflecting the (cubic) configurational average. There is a robust analogy between this model and icelike states on the pyrochlore lattice, which also preserve cubic symmetry [50,52,53].

In order to test the relevance of this spaghetti model to the NHE mechanism of $ZrW_2O_8 \cdot H_2O$, we proceeded to measure x-ray total scattering data for both ZrW_2O_8 and its hydrate. Such data are directly sensitive to the local structure changes induced by hydration. We then developed a supercell model of a disordered spaghetti state, informed by the DFT structure we have already described, from which an x-ray-weighted PDF might be calculated and compared against experiment.

Our total scattering measurements were performed using a PANalytical Empyrean x-ray diffractometer fitted with an Ag anode ($Q_{\mathrm{max}} = 20$ Å^{-1}) and a GaliPIX3D detector; the corresponding PDFs [54], normalized using GUDRUNX [55], are shown in Fig. 3(a). We were able to model the PDF of ZrW_2O_8 in PDFGUI [56] using the crystal structure of Ref. [3]. By contrast, the $P2_12_12_1$ model for $ZrW_2O_8 \cdot H_2O$ is capable of fitting convincingly only the very lowest-r region of the corresponding PDF, and it cannot account for the positional disorder evident at distances $r \gtrsim 4$ Å.

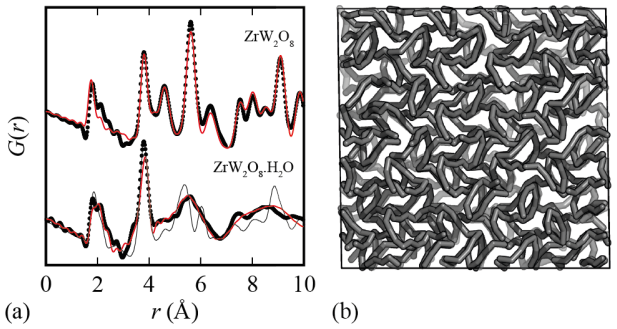


FIG. 3. (a) X-ray PDFs $G(r)$ of anhydrous (top) and hydrated (bottom) ZrW_2O_8: data as points and fits as lines. For ZrW_2O_8 a satisfactory fit is obtained using the model of Ref. [3] (red lines). For the hydrate, the DFT $P2_12_12_1$ model fits at low r but cannot reproduce the broad oscillations at $r > 4$ Å (thin black line). Instead, the PDF calculated from our disordered spaghetti model (red line) captures all features. (b) The relaxed GULP model for a $ZrW_2O_8 \cdot H_2O$ spaghetti configuration as described in the text: only W—O—W linkages are shown, using grey cylinders [*cf.* Fig. 2(d)].

Because it was not computationally feasible to explore a disordered spaghetti configuration at the DFT level, we used harmonic lattice-dynamical methods, informed by the bonding geometries observed in our $P2_12_12_1$ DFT structure, to generate a $5 \times 5 \times 5$ approximant [57] [Fig. 3(b)]. First, we determined an appropriate disordered spaghetti of self-avoiding W—O—W pathways using a custom Monte Carlo algorithm, guided by the connectivity rules of Fig. 2(c). This spaghetti was used to assign both connectivity and H_2O orientations within a supercell of the $P2_13$ ZrW_2O_8 structure. Atomic positions and unit cell dimensions were then relaxed using the GULP code [34,58], with harmonic potentials constructed from the equilibrium bond distances and angles of our earlier DFT calculation. The cell volume decreased during relaxation, reflecting activation of the NHE mechanism as described above. We consider this GULP-relaxed configuration to be representative of the degree of structural disorder implied by the spaghetti model. The connectivity of the relaxed structure is shown in Fig. 3(b), and the PDF calculated from the full model compared against experiment in Fig. 3(a). The quality of this fit-to-data shows that the displacive disorder introduced by spaghetti W—O—W connectivity can indeed account for the broad oscillations in the experimental PDF.

A necessary consequence of the macroscopic configurational entropy of spaghetti formation on the ZrW_2O_8 lattice is that the corresponding hydration-driven distortions are dense throughout the BZ. Hence, the structure of $ZrW_2O_8 \cdot H_2O$ reflects a superposition of spaghetti states with arbitrarily large periodicities. Anisotropy vanishes in this limit such that the corresponding global symmetry is cubic, with an average structure given by the projection of all possible distortions onto a single unit cell [Fig. 1(b)]. The large displacements involved account for the substantial crystal strain observed experimentally, and they also account for the difficulty of interpreting even high-quality neutron diffraction data [12]. In turn, proton migration and topological defects in spaghetti state formation account for the introduction of inversion symmetry. Consequently, we conclude that NHE is driven by a degenerate family of "spaghetti"-like lattice distortions modulated with periodicities distributed continuously throughout the BZ.

Having established this mechanism, we sought to determine whether the same distortions that drive NHE are related to the phonons responsible for NTE in ZrW_2O_8. Published data [23] allow access to the frequencies $\omega(\mathbf{k}, \nu)$, eigenvectors $\mathbf{e}(\mathbf{k}, \nu)$, and Grüneisen parameters,

$$\gamma(\mathbf{k}, \nu) = -\frac{V}{\omega(\mathbf{k}, \nu)} \frac{\partial}{\partial V} \omega(\mathbf{k}, \nu), \tag{1}$$

for each phonon branch ν and wave-vector $\mathbf{k}$. NTE modes are those for which $\gamma < 0$. By comparing the atomic coordinates of each ZrW_2O_8 framework atom in the anhydrous and hydrated DFT structures, we formed a normalized, mass-weighted displacement vector $\mathbf{u}$ that describes the volume-reducing framework deformations induced by hydration. The projections $c(\mathbf{k}, \nu) = \mathbf{u} \cdot \mathbf{e}(\mathbf{k}, \nu)$ partition these deformations amongst the normal modes; for the ($\mathbf{k} = \mathbf{0}$) DFT structure the $c(\mathbf{k}, \nu)$ necessarily vanish for $\mathbf{k} \neq \mathbf{0}$. Remarkably, we find that 88% of the displacements described by $\mathbf{u}$ can be accounted for in terms of two of the lowest-energy $\mathbf{k} = \mathbf{0}$ NTE modes. The same result is found for zone-center phonons generated by our quantum mechanical calculations [34]. Extending our analysis to wave-vectors $\mathbf{k} \neq \mathbf{0}$ requires access to supercell configurations. So, using the same GULP approach described above, we generated a family of ten different $3 \times 3 \times 3$ spaghetti configurations. For each $ZrW_2O_8 \cdot H_2O$ configuration, we extracted the corresponding framework displacement vectors $\mathbf{u}$ and projected these onto the phonon eigenvectors $\mathbf{e}(\mathbf{k}, \nu)$ of ZrW_2O_8 for $\mathbf{k} \in \langle (n_1/3), (n_2/3), (n_3/3) \rangle^*$, $n_i \in (0, \pm 1)$. We find that the projections onto the very lowest energy modes (mostly acoustic) account for ~45% of the hydration-driven displacement, and on including all modes with negative Grüneisen parameters, this total projection rises to 87% of the total framework response [Fig. 4] [34].

So, the most important NTE phonon branches are strongly implicated in the NHE response of ZrW_2O_8, but the question remains of whether the spaghetti displacements invoked for NHE are necessarily involved in the NTE dynamics. The $c(\mathbf{k}, \nu)$ play a crucial role in answering this question: whereas previous studies have focused either on individual collective modes [22] or on the correlated motion of isolated structural fragments [29], the $c(\mathbf{k}, \nu)$ allow us to reconstruct combination NTE modes with intuitive real-space realizations. For example, the $P2_12_12_1$

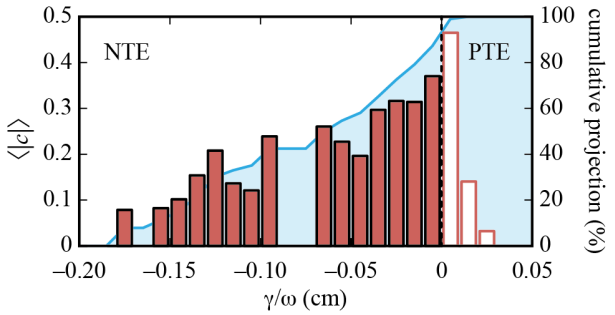


FIG. 4. Hydration-driven framework displacements in $ZrW_2O_8 \cdot H_2O$ project strongly onto NTE phonon eigenvectors in ZrW_2O_8. The average magnitude of projection coefficients c (vertical bars) is shown as a histogram of γ/ω; nearly 90% of the cumulative projection (blue shaded region) is accounted for by NTE modes.

DFT configuration of Fig. 2(b) gives a framework displacement vector $\mathbf{u}$ that—as we have said—projects almost entirely onto just two zone-center modes ν_1, ν_2 (with $\omega_1 = \omega_2 = 44.1$ cm^{-1} and $\gamma_1 = \gamma_2 = -4.14$). The combination mode $c(0,\nu_1)\nu_1 + c(0,\nu_2)\nu_2$ then describes one specific collective vibration of ZrW_2O_8 contributing to NTE. Crucially, the associated displacement vector $\mathbf{e}' = c(0,\nu_1)\mathbf{e}(0,\nu_1) + c(0,\nu_2)\mathbf{e}(0,\nu_2)$ does indeed involve fluctuations towards the very same spaghetti states assumed by $ZrW_2O_8 \cdot H_2O$; an animation of this combination mode is provided as Supplemental Material [34]. The same result also holds for the more complex family of NTE modes associated with the $3 \times 3 \times 3$ spaghetti configurations described above [34].

The significance of this result is that the key real-space mechanism for NTE in ZrW_2O_8 is the correlated translation, rotation, and distortion of ZrO_6/WO_4 polyhedra towards the volume-reducing spaghetti states adopted in $ZrW_2O_8 \cdot H_2O$. The configurational degeneracy of these states—a property of the peculiar ZrW_2O_8 topology—is such that these fluctuations are described by a family of phonons necessarily spread throughout the BZ, with individual spaghetti distortions described by the population of a family of NTE phonons. Consequently, a "spaghetti dynamics" mechanism for NTE in ZrW_2O_8 finally explains (i) why NTE modes are distributed throughout the BZ, (ii) why it is so difficult to interpret individual NTE phonons in isolation, and (iii) how the local polyhedral rotations or translations identified in previous studies [27,29] actually propagate throughout the crystal lattice. Moreover, in determining this mechanism, we have established an unexpected link between the correlated structural disorder associated with NHE, on the one hand, and the property of NTE, on the other hand, that is of significant conceptual importance in the broader context of understanding disorder–property relationships in functional materials [53].

The authors gratefully acknowledge financial support from the ERC (Grant No. 279705), the EPSRC (U.K.), the Leverhulme Trust (Grant No. RPG-2015-292), a CNRS/Oxford collaboration grant, and from the Marshall Foundation to P. M. M. This work benefitted from access to HPC platforms provided by a GENCI grant (A0030807069). A. L. G. thanks J. S. O. Evans (Durham), J. P. Attfield (Edinburgh), B. Slater (UCL), and T. Ledsam (Oxford) for valuable discussions.

*andrew.goodwin@chem.ox.ac.uk

[1] C. Martinek and F. A. Hummel, J. Am. Ceram. Soc. **51**, 227 (1968).
[2] J. S. O. Evans, T. A. Mary, T. Vogt, M. A. Subramanian, and A. W. Sleight, Chem. Mater. **8**, 2809 (1996).
[3] T. A. Mary, J. S. O. Evans, T. Vogt, and A. W. Sleight, Science **272**, 90 (1996).
[4] J. S. O. Evans, W. I. F. David, and A. W. Sleight, Acta Crystallogr. Sect. B **55**, 333 (1999).
[5] J. S. O. Evans, J. Chem. Soc. Dalton Trans. 3317 (1999).
[6] G. D. Barerra, J. A. O. Bruna, T. H. K. Barron, and N. L. Allan, J. Phys. Condens. Matter **17**, R217 (2005).
[7] C. Lind, Materials **5**, 1125 (2012).
[8] C. A. Perottoni and J. A. H. da Jornada, Science **280**, 886 (1998).
[9] T. Varga, A. P. Wilkinson, A. C. Jupe, C. Lind, W. A. Bassett, and C.-S. Zha, Phys. Rev. B **72**, 024117 (2005).
[10] D. A. Keen, A. L. Goodwin, M. G. Tucker, M. T. Dove, J. S. O. Evans, W. A. Crichton, and M. Brunelli, Phys. Rev. Lett. **98**, 225501 (2007).
[11] J. Hegedüs and S. R. Elliott, Nat. Mater. **7**, 399 (2008).
[12] N. Duan, U. Kameswari, and A. W. Sleight, J. Am. Chem. Soc. **121**, 10432 (1999).
[13] N. A. Banek, H. I. Baiz, A. Latigo, and C. Lind, J. Am. Chem. Soc. **132**, 8278 (2010).
[14] C. Serre, F. Millange, C. Thouvenot, M. Noguès, G. Marsolier, D. Louër, and G. Férey, J. Am. Chem. Soc. **124**, 13519 (2002).
[15] A. V. Neimark, F.-X. Coudert, A. Boutin, and A. H. Fuchs, J. Phys. Chem. Lett. **1**, 445 (2010).
[16] D. Bousquet, F.-X. Coudert, A. G. J. Fossati, A. V. Neimark, A. H. Fuchs, and A. Boutin, J. Chem. Phys. **138**, 174706 (2013).
[17] S. Krause, V. Bon, I. Senkovska, U. Stoeck, D. Wallacher, D. M. Többens, S. Zander, R. S. Pillai, G. Maurin, F.-X. Coudert, and S. Kaskel, Nature (London) **532**, 348 (2016).
[18] A. P. Ramirez and G. R. Kowach, Phys. Rev. Lett. **80**, 4903 (1998).
[19] R. Mittal and S. L. Chaplot, Phys. Rev. B **60**, 7234 (1999).
[20] K. Wang and R. R. Reeber, Appl. Phys. Lett. **76**, 2203 (2000).
[21] V. Gava, A. L. Martinotto, and C. A. Perottoni, Phys. Rev. Lett. **109**, 195503 (2012).
[22] M. K. Gupta, R. Mittal, and S. L. Chaplot, Phys. Rev. B **88**, 014303 (2013).
[23] L. H. N. Rimmer, M. T. Dove, and K. Refson, arXiv: 1510.00361.

[24] R. Mittal, S. L. Chaplot, H. Schober, and T. A. Mary, Phys. Rev. Lett. **86**, 4692 (2001).

[25] A. K. A. Pryde, K. D. Hammonds, M. T. Dove, V. Heine, J. D. Gale, and M. C. Warren, J. Phys. Condens. Matter **8**, 10973 (1996).

[26] J. Z. Tao and A. W. Sleight, J. Solid State Chem. **173**, 442 (2003).

[27] M. G. Tucker, A. L. Goodwin, M. T. Dove, D. A. Keen, S. A. Wells, and J. S. O. Evans, Phys. Rev. Lett. **95**, 255501 (2005).

[28] D. Cao, F. Bridges, G. R. Kowach, and A. P. Ramirez, Phys. Rev. Lett. **89**, 215902 (2002).

[29] F. Bridges, T. Keiber, P. Juhas, S. J. L. Billinge, L. Sutton, J. Wilde, and G. R. Kowach, Phys. Rev. Lett. **112**, 045505 (2014).

[30] Md. Imteyaz Ahmad and M. Akinc, Mater. Chem. Phys. **145**, 403 (2014).

[31] N. A. Banek, Autohydration of Nanosized Cubic Zirconium Tungstate, Ph.D. Thesis, The University of Toledo, 2011.

[32] W. Cao, Q. Huang, Y. Rong, Y. Wang, J. Deng, J. Chen, and X. Xing, Inorg. Chem. Front. **3**, 856 (2016).

[33] Note that the molar volume of $ZrW_2O_8 \cdot NH_3$ is only marginally smaller than that of ZrW_2O_8, and both systems show NTE [32].

[34] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.120.265501 for details concerning methods, the $P2_12_12_1$ DFT structural model of $ZrW_2O_8 \cdot H_2O$, and configurational entropy calculations, which includes Refs. [35–47].

[35] M. S. Dadachov and R. M. Lambrecht, J. Mater. Chem. **7**, 1867 (1997).

[36] L. C. Kozy, M. N. Tahir, C. Lind, and W. Tremel, J. Mater. Chem. **19**, 2760 (2009).

[37] A. A. Coelho, TOPAS-ACADEMIC, version 4.1, 2007.

[38] R. Dovesi, R. Orlando, B. Civalleri, C. Roetti, V. R. Saunders, and C. M. Zicovich-Wilson, Z. Kristallogr. **220**, 571 (2005).

[39] C. Gatti, V. R. Saunders, and C. Roetti, J. Chem. Phys. **101**, 10686 (1994).

[40] G. Sophia, P. Baranek, C. Sarrazin, M. Rérat, and R. Dovesi, Phase Transitions **86**, 1069 (2013).

[41] C. A. Figueirêdo and C. A. Perottoni, Phys. Rev. B **75**, 184110 (2007).

[42] D. Andrae, U. Häußermann, M. Dolg, H. Stoll, and H. Preuß, Theor. Chim. Acta **77**, 123 (1990).

[43] H. J. Monkhorst and J. D. Pack, Phys. Rev. B **13**, 5188 (1976).

[44] C. Adamo and V. Barone, J. Chem. Phys. **110**, 6158 (1999).

[45] J. P. Perdew, A. Ruzsinszky, G. I. Csonka, O. A. Vydrov, G. E. Scuseria, L. A. Constantin, X. Zhou, and K. Burke, Phys. Rev. Lett. **102**, 039902 (2009).

[46] J. P. Perdew and A. Zunger, Phys. Rev. B **23**, 5048 (1981).

[47] L. Pauling, J. Am. Chem. Soc. **57**, 2680 (1935).

[48] V. Bolis, in *Calorimetry and Thermal Methods in Catalysis*, edited by A. Auroux (Springer, Berlin, 2013).

[49] L. D. C. Jaubert, M. Haque, and R. Moessner, Phys. Rev. Lett. **107**, 177202 (2011).

[50] C. L. Henley, Annu. Rev. Condens. Matter Phys. **1**, 179 (2010).

[51] L. D. C. Jaubert, S. Piatecki, M. Haque, and R. Moessner, Phys. Rev. B **85**, 054425 (2012).

[52] R. Moessner and A. P. Ramirez, Phys. Today **59**, 24 (2006).

[53] A. R. Overy, A. B. Cairns, M. J. Cliffe, A. Simonov, M. G. Tucker, and A. L. Goodwin, Nat. Commun. **7**, 10445 (2016).

[54] D. A. Keen, J. Appl. Crystallogr. **34**, 172 (2001).

[55] A. K. Soper and E. R. Barney, J. Appl. Crystallogr. **44**, 714 (2011).

[56] C. L. Farrow, P. Juhas, J. W. Liu, D. Bryndin, E. S. Bozin, J. Block, T. Proffen, and S. J. L. Billinge, J. Phys. Condens. Matter **19**, 335219 (2007).

[57] P. M. M. Thygesen, J. A. M. Paddison, R. Zhang, K. A. Beyer, K. W. Chapman, H. Y. Playford, M. G. Tucker, D. A. Keen, M. A. Hayward, and A. L. Goodwin, Phys. Rev. Lett. **118**, 067201 (2017).

[58] J. D. Gale, J. Chem. Soc., Faraday Trans. **93**, 629 (1997).

Appendix A

Supplementary information: ZrW_2O_8

The following appendix details the DFT calculations, refinements, and entropy estimation included in Chapter 3. I would like to gratefully acknowledge Mia Baise, Dr. Fabien Trousselet, and Dr. F-X Coudert, who completed the DFT calculations contained herein. This appendix consists of work published in reference [515].

A.1 Density functional theory calculations

All quantum chemistry calculations were performed using the CRYSTAL14 software [526]. This approach is based on density functional theory (DFT) and adapted to ordered solids as it uses localised basis sets built on atom-centred (Gaussian-type) functions. The following sets were retained (acronyms in italics; for H, O, and Zr, the corresponding coefficients can be found at `www.crystal.unito.it/basis-sets.php`):

- For H: *H 3-1p1G gatti 1994* (Ref. 527)
- For O: *O 8-411(1d)G baranek 2013 BaZrO3* (Ref. 528)
- For Zr: *Zr SC HAYWSC-3111(32111df)G baranek 2013 BaZrO3* (Ref. 528)
- For W: basis set for 14 valence electrons [381] and pseudopotential [529]

The **k**-point mesh was generated using the Monkhorst-Pack method [530], with a $3 \times 3 \times 3$ supercell.

To take into account exchange and correlation (XC), the PBESOL0 functional was used: it is a hybrid exchange–correlation functional combining a generalised gradient approximation and a certain amount of exact Hartree-Fock exchange.[385] We checked that, of the functionals tested, PBESOL0 gives the best agreement with

available experimental data on ZrW_2O_8 (structural and mechanical properties). In addition, we also used an LDA-based functional with the Perdew-Zunger parameterisation (LDA-PZ) [386, 531] to evaluate the influence of the ZC functional on our results. Unless specified, the results reported are those obtained with PBESOL0.

Angular frequencies ω_ν of the $3N$ harmonic vibration modes ($N = 44$ for ZrW_2O_8 and $N = 56$ for $ZrW_2O_8{\cdot}H_2O$) were computed using the method described in the CRYSTAL14 manual. For a given optimised structure (local minimum in the energy landscape), this approach consists of performing small displacements of individual atoms and estimating from these the Hessian matrix. These calculations were performed at the harmonic level (keyword `FREQCALC`).

To estimate Grüneisen parameters γ_ν, the structures were re-optimised at pressures $k\delta p$, with $\delta p = 200\,\mathrm{MPa}$ and $-3 \leq k \leq 3$; for each pressure (corresponding to a volume $V(k\delta p)$), angular frequencies ω_ν were computed as described above and assigned to the correct mode label ν with help of their symmetry properties. The corresponding Grüneisen parameters

$$\gamma_\nu = -\frac{V}{\omega_\nu}\frac{\mathrm{d}\omega_\nu}{\mathrm{d}V} \tag{A.1}$$

was then obtained *via* a linear fit to the $\omega_\nu(V)$ curve.

A.1.1 DFT-relaxed structure of ZrW_2O_8

The unit cell parameters and atom coordinates for the DFT-calculated structural model for $ZrW_2O_8{\cdot}H_2O$ discussed in the main text are given in Table A.1.

Table A.1: Crystallographic details for the $ZrW_2O_8 \cdot H_2O$ structure model discussed in the main text.

Crystal system	Orthorhombic
Space group	$P2_12_12_1$
Z	4
a(Å)	8.780
b (Å)	8.757
c (Å)	9.302

Atom	**Coordinates**
Zr	$(-0.031, 0.053, 0.011)$
W1	$(0.353, 0.314, 0.370)$
W2	$(0.614, 0.538, 0.608)$
O11	$(0.199, 0.179, 0.346)$
O12a	$(0.231, 0.484, 0.413)$
O12b	$(0.498, 0.159, 0.412)$
O12c	$(0.399, 0.361, 0.187)$
O21	$(0.522, 0.448, 0.456)$
O22a	$(0.805, 0.588, 0.565)$
O22b	$(0.531, 0.728, 0.587)$
O22c	$(0.575, 0.534, 0.804)$
Ow	$(0.303, 0.288, 0.608)$
Hwa	$(0.189, 0.293, 0.617)$
Hwb	$(0.345, 0.189, 0.643)$

A.1.2 Criteria for modeling

This section describes the DFT-optimised, hydrated structure discussed in the paper and possesses the key properties of $ZrW_2O_8{\cdot}H_2O$. It has a smaller unit cell than the anhydrous structure:

$$V(ZrW_2O_8{\cdot}H_2O) = V_H < V_0 = V(ZrW_2O_8),$$

which reflects the hydration-driven volume reduction observed experimentally. For both versions of this structure (each corresponding to a XC functional, either LDA or PBESOL0) we list the shortest O...W distances in Table A.2; the corresponding O...H distances corresponding to hydrogen bonds can also be found in Table A.3. The various structures (A–E) we report correspond to the output of relaxations using different relative H_2O positions and orientations. All have $P2_12_12_1$ symmetry. The structure considered in the main text is the one we label 'structure E', but the conclusions are independent of which solution is considered.

Table A.2: Distances (in Å) between each of the two inequivalent W atoms and the oxygen atoms of its coordination sphere, given for structure E and for both XC functionals considered. The label Ow refers to the oxygen atom of a bound H_2O molecule.

Functional	**LDA**	**PBESOL0**
d(W1–O)	1.80	1.77
”	1.81	1.79
”	1.88	1.87
”	1.89	1.88
”	2.06	2.11
d(W1-Ow)	2.23	2.23
d(W2–O)	1.78	1.77
”	1.81	1.78
”	1.83	1.80
”	1.85	1.83
”	2.09	2.18

Table A.3: Structural and energetic comparison between the various hydrated structures relaxed from different initial configurations. Hydration enthalpies ΔE_{hyd} are in kJ/mol and distances are in Å. The parameter p(O...H) denotes the fraction of hydrogen atoms involved in hydrogen bonding, using the geometric criterion of d(O...H) < 2.5 Å.

Structure label	**XC functional**	$(V_0 - V_{\mathrm{H}})/V_0$ (%)	$-\Delta E_{\mathrm{hyd}}$	p(O...H)	d_{min}(O...H)	d_{max}(O...H)
A	LDA	6.59	148.6	1.0	1.59	1.71
A	PBESOL0	5.54	61.2	1.0	1.71	1.86
B	LDA	6.97	152.7	0.5	1.57	1.57
B	PBESOL0	5.68	74.2	0.5	1.67	1.67
C	LDA	6.13	165.8	1.0	1.64	2.27
C	PBESOL0	5.72	84.5	1.0	1.72	2.41
D	LDA	6.96	148.2	0.75	1.56	1.71
D	PBESOL0	5.66	64.9	0.75	1.68	1.89
E	LDA	7.52	175.5	1.0	1.60	1.75
E	PBESOL0	6.24	88.11	1.0	1.66	1.82

Independent of the particular model and the particular XC functional used, all structures are characterised by three remarkable features:

(i) The H_2O unit is not separated into H and O–H, but is inserted as a whole.

(ii) Hydrogen bonds were formed between both H atoms and the surrounding O atoms of the zirconium tungstate framework.

(iii) Even more importantly, one of the two inequivalent W atoms acquires 6-fold coordination, when including weak W...O bonds of length 2.3(1) Å (the other four oxygens are at much smaller distances $\simeq$ 1.8 Å); and the other becomes five-coordinated (again by formation of a weaker/longer W–O bond), allowing for the spaghetti state described in the main text.

We note that when re-optimising any hydrated structure model with other XC functionals, the same features remained unchanged.

A.1.3 Analysis of vibrational and thermal expansion properties

The Grüneisen parameters for zone-center phonon modes in both ZrW_2O_8 and our $P2_12_12_1$ structure of $ZrW_2O_8{\cdot}H_2O$ were calculated. The corresponding values and their frequency dependence are shown in Fig. A.1.

From Grüneisen parameters of individual vibration modes (in either the anhydrous or hydrated structure), we extracted the thermal expansion coefficient $\alpha_V = \mathrm{d}V/V\mathrm{d}T$ as follows:

$$\alpha_V = \frac{1}{BV}\sum_\nu \gamma_\nu c_V^{(\nu)}(T), \tag{A.2}$$

$$c_V^{(\nu)}(T) = k_\mathrm{B}\exp(-\xi_\nu)\left\{\frac{\xi_\nu}{[1-\exp(-\xi_\nu)]^2}\right\}^2, \tag{A.3}$$

$$\xi_\nu = \frac{\hbar\omega_\nu}{k_\mathrm{B}T}. \tag{A.4}$$

The resulting values are listed in Table A.4. We find NTE for the anhydrous structure and PTE for the hydrated structure. Both observations are consistent with the experimental behavior.

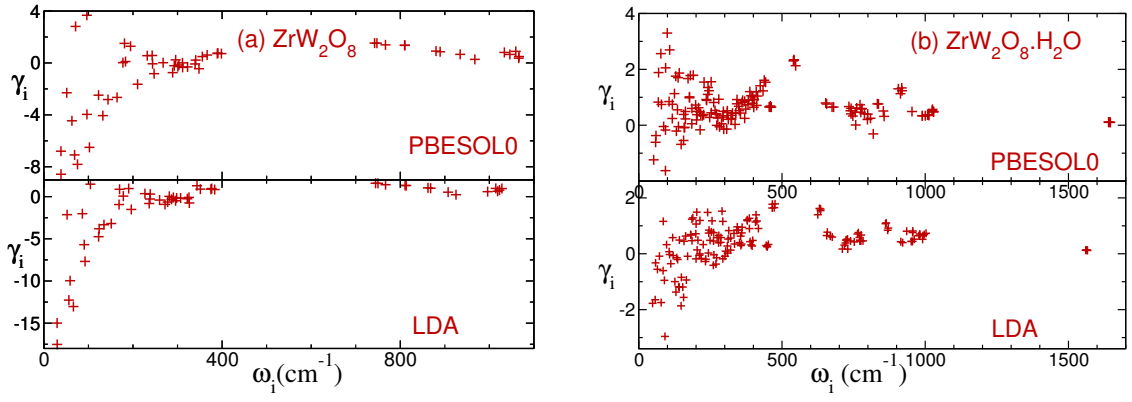


Figure A.1: DFT mode Grüneisen parameters. Grüneisen parameters γ_ν of the individual harmonic modes as a function of their angular frequencies ω_ν (obtained at $P = 0, V = V_0$ from DFT calculations), for both XC functionals considered. Left: anhydrous structure' right: hydrated structure.

Table A.4: Thermal expansion coefficients (in MK^{-1}) for both anhydrous and hydrated structures. Unless specified, calculated values of α_V correspond to $T = 300\,\mathrm{K}$, obtained via linear fits of $\omega_\nu(V)$ plots.

Method / Functional	**LDA**	**PBESOL0**	**PBESOL0** ($T \to \infty$)	**Experiment**
ZrW_2O_8	-42.91	-20.50	-16.84	-27 (Ref. 360)
$ZrW_2O_8{\cdot}H_2O$	7.90	18.36	24.61	> 6 (Ref. 346)

A.2 Detailed refinements

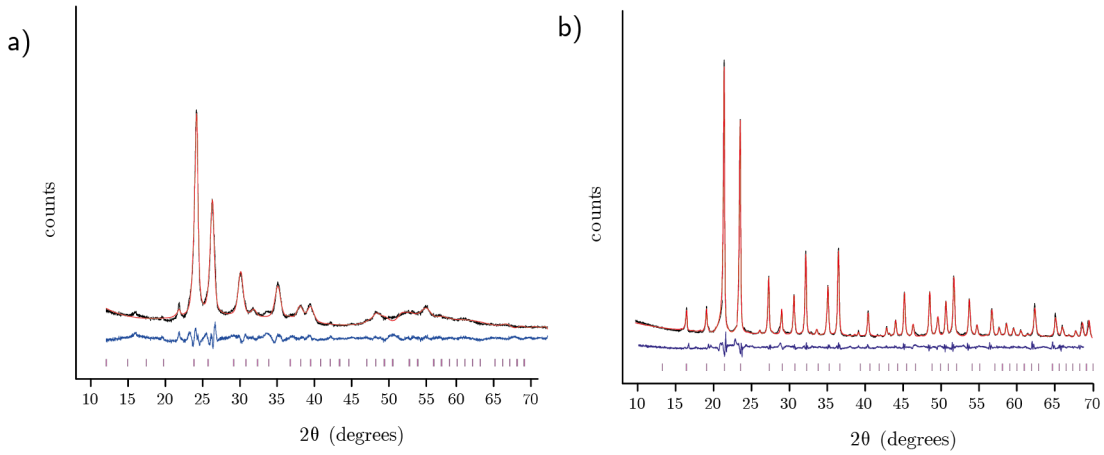


Figure A.2: ZrW_2O_8 Rietveld refinements. Included are the refinements for the (a) anhydrous (b) hydrated phases detailed in Chapter 3.

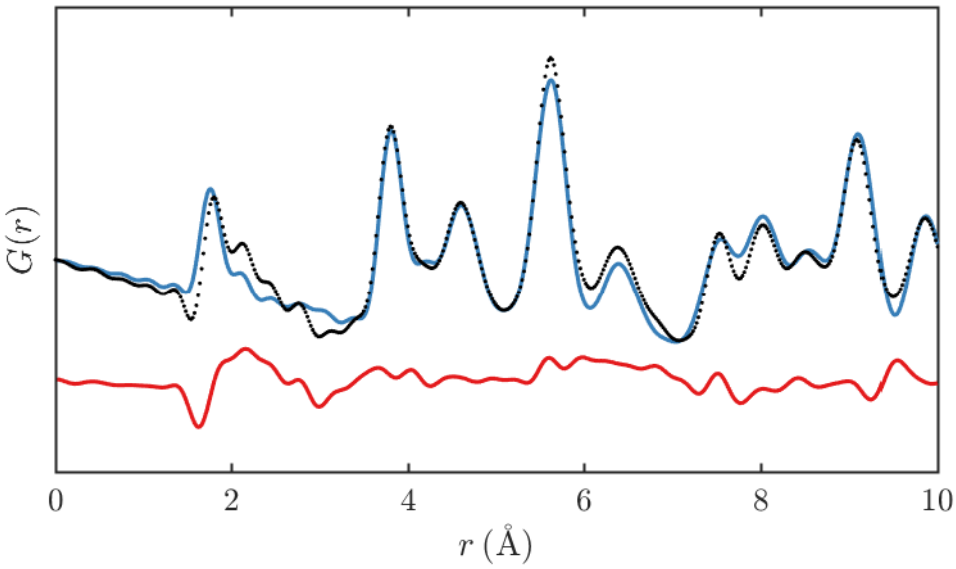


Figure A.3: ZrW_2O_8 PDF refinement. Refinement based on a single unit cell of the Rietveld result.

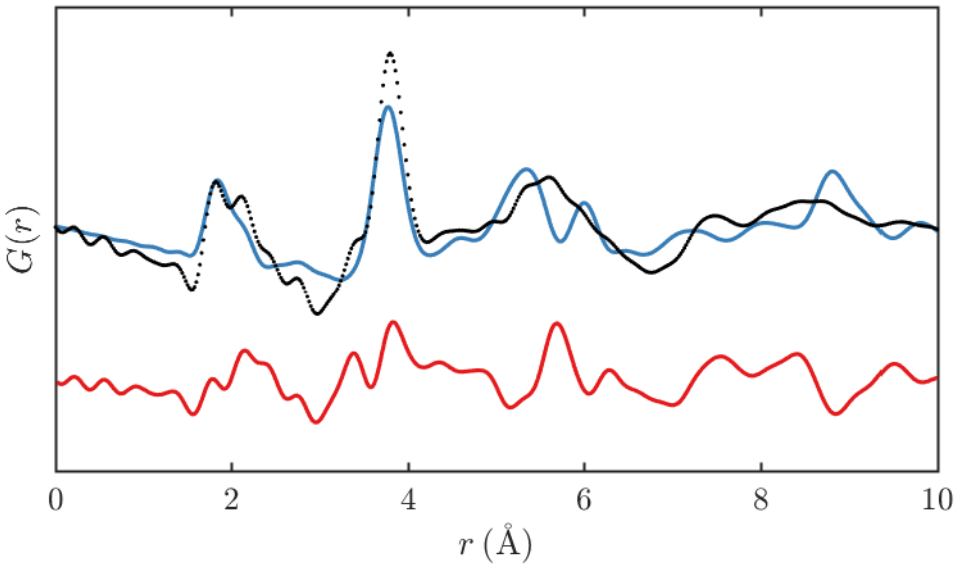


Figure A.4: $ZrW_2O_8 \cdot H_2O$ PDF refinement. Refinement based on a single unit cell of the DFT based model.

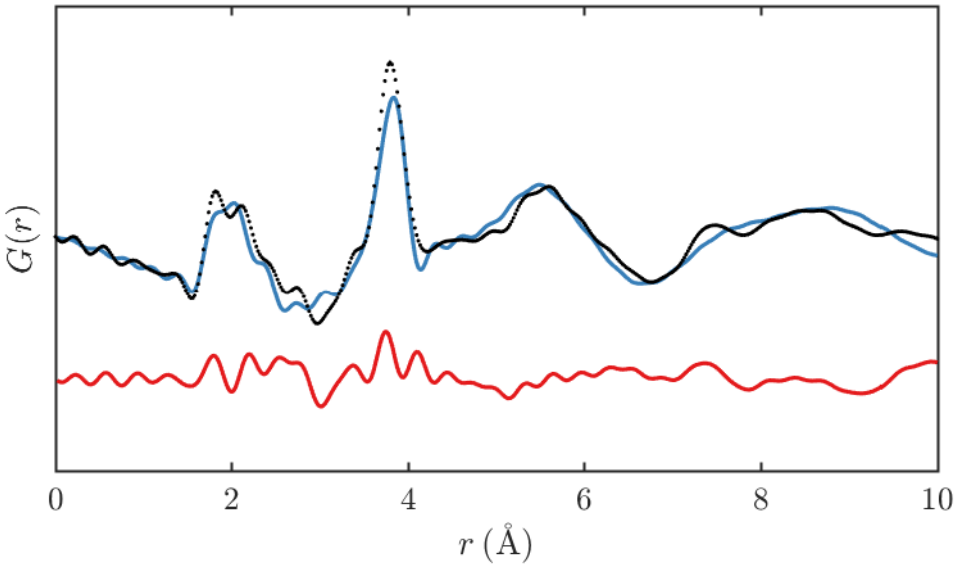


Figure A.5: $ZrW_2O_8{\cdot}H_2O$ PDF refinement. Refinement based on a supercell GULP based model.

A.3 Configurational entropy of the spaghetti phase

In order to estimate the configurational entropy associated with spaghetti formation we use an approach based on Pauling's approximation to the configurational entropy of ice, and as extended to the class of procrystals in Ref. 85. Our basis is to consider the number of ways in which N W_2O_8 units might connect subject to the icelike rules described in the main text. Note that there will be $3N$ possible inter-unit connection pathways. We assign to each pathway an Ising variable $e \in \{0, 1)$ that denotes whether it is to be connected (1) or not (0). Clearly there are 2^{3N} possible pathway states if rules are ignored. Pauling tells us now to determine the number of microstates by multiplying this number by the fraction that satisfy local rules for a given W_2O_8 unit. Hence we have

$$\Omega = 2^{3N} \times \left(\frac{9}{2^6}\right)^N, \tag{A.5}$$

$$= \left(\frac{9}{8}\right)^N, \tag{A.6}$$

since there are 2^6 possible states for each W_2O_8 unit, and only nine of these obey the local rules established in the text. Consequently the Pauling estimate for the configurational entropy of the spaghetti phase is $S_{\text{config}} \simeq R \ln(9/8) > 0$.

Appendix B

Supplementary data: square-planar CRNs

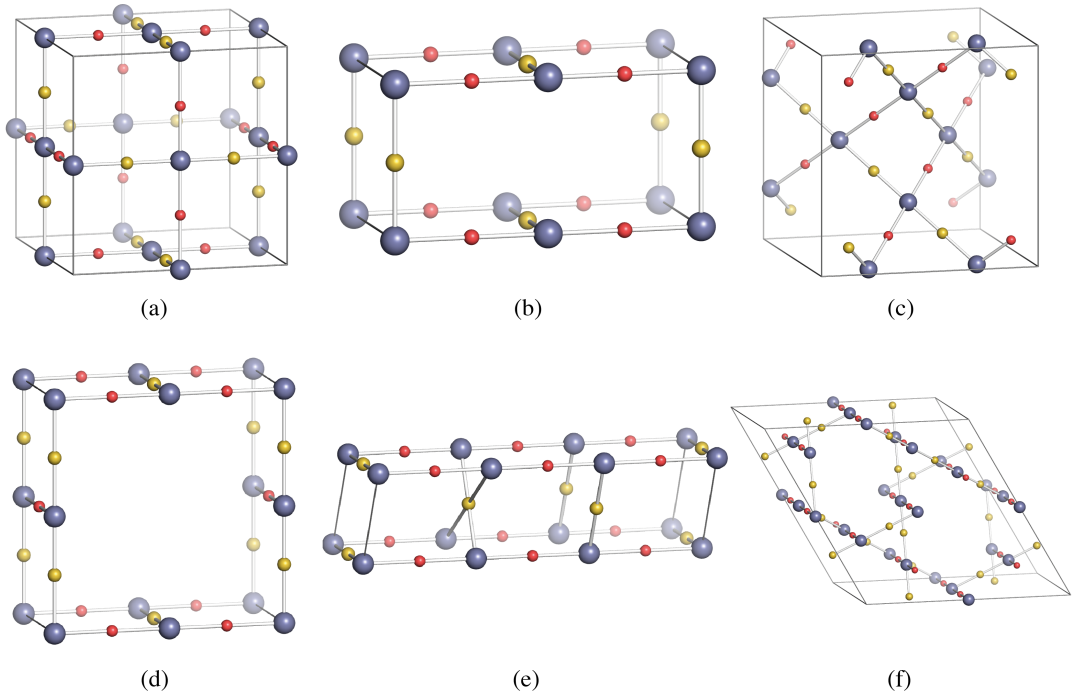


Figure B.1: Available four-coordinate planar nets. There is only a small set of known nets with four-coordinate planar subunits. This includes: (a) **nbo**, (b) **cds**, (c) **lvt**, (d) **mot**, (e) **qzd**, and (f) **usf**.

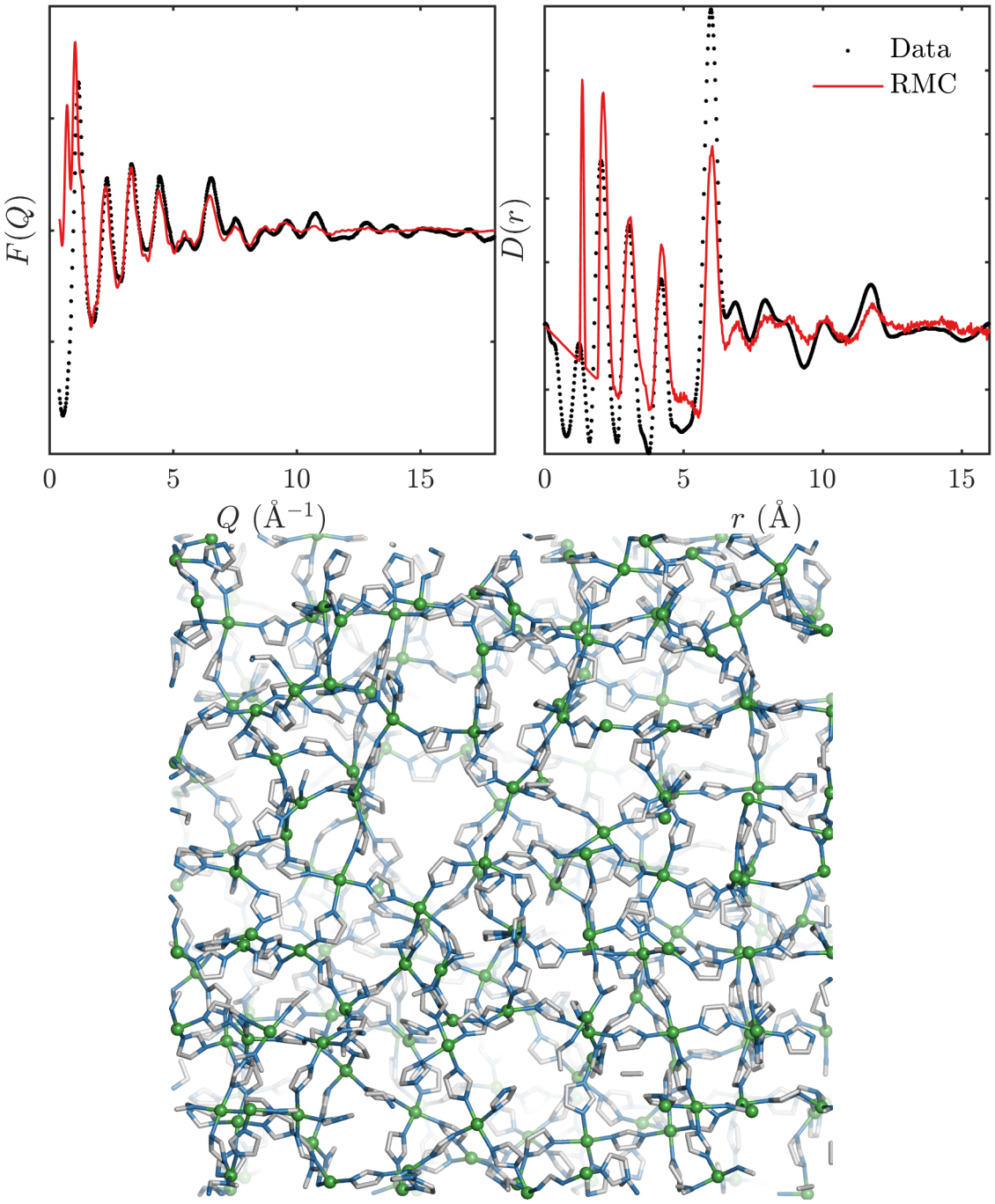


Figure B.2: Unphysical voids in a CRN. The voids — and subsequent memory of the open **nbo** starting structure — are evident from the low-Q behavior of the scattering fit.

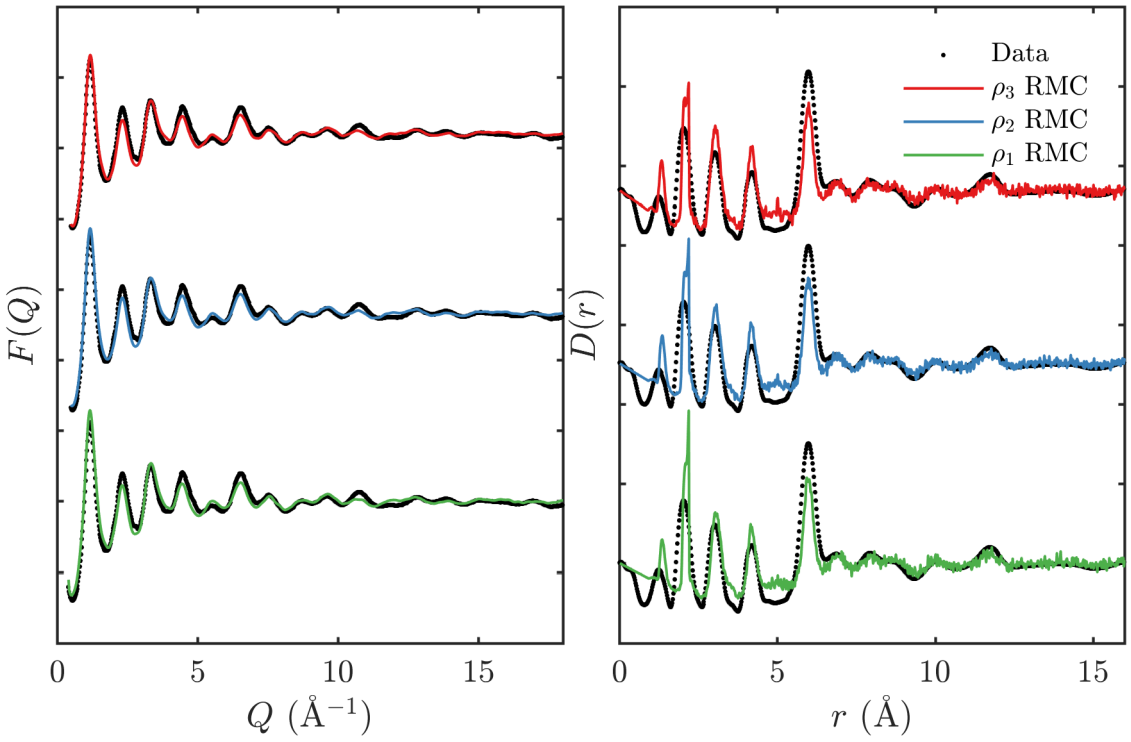


Figure B.3: RMC fit for different densities. The $F(Q)$ (left) and $D(r)$ (right) measurements are shown in black circles, overlaid by the RMCProfile refinement of a square-planar CRN. The CRN was generated using an adaptive displacement WWW simulation. The coloring corresponds to three densities: green, $\rho_1 = 0.0722\,\text{Å}^{-3}$ (*a*-ZIF density); blue, $\rho_2 = 0.0825\,\text{Å}^{-3}$ (mid-range density); and red, $\rho_3 = 0.0998\,\text{Å}^{-3}$ (corrected Pd(Him)$_2$(im)$_2$ density).

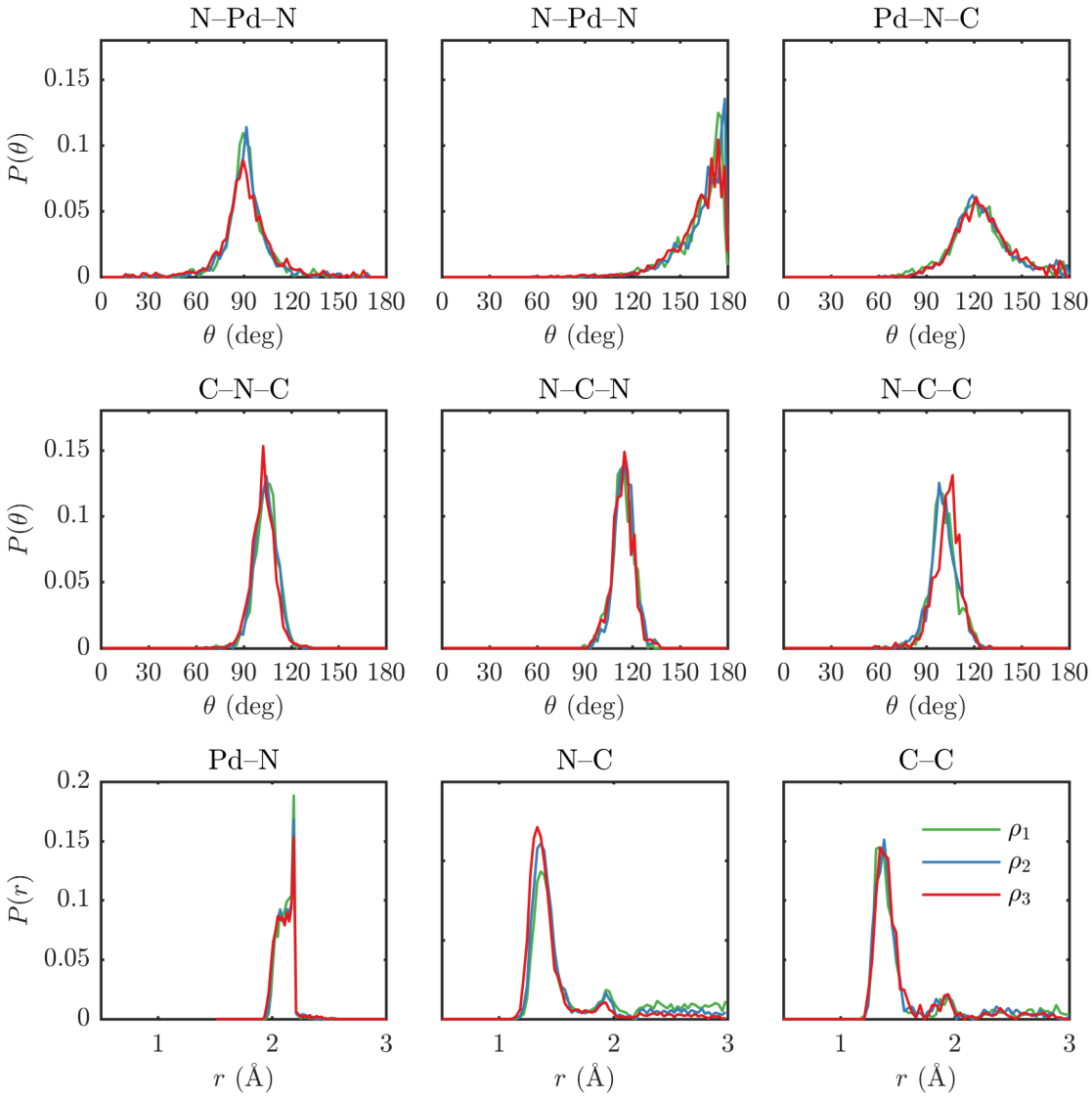


Figure B.4: Pd(im)$_2$ geometry histograms for different densities. The distribution of bond lengths and angles corresponding to the RMC fits herein. The angular distributions are weighted by $1/\sin\theta$. The coloring corresponds to three densities: green, $\rho_1 = 0.0722\,\text{Å}^{-3}$ (*a*-ZIF density); blue, $\rho_2 = 0.0825\,\text{Å}^{-3}$ (mid-range density); and red, $\rho_3 = 0.0998\,\text{Å}^{-3}$ (corrected Pd(Him)$_2$(im)$_2$ density).

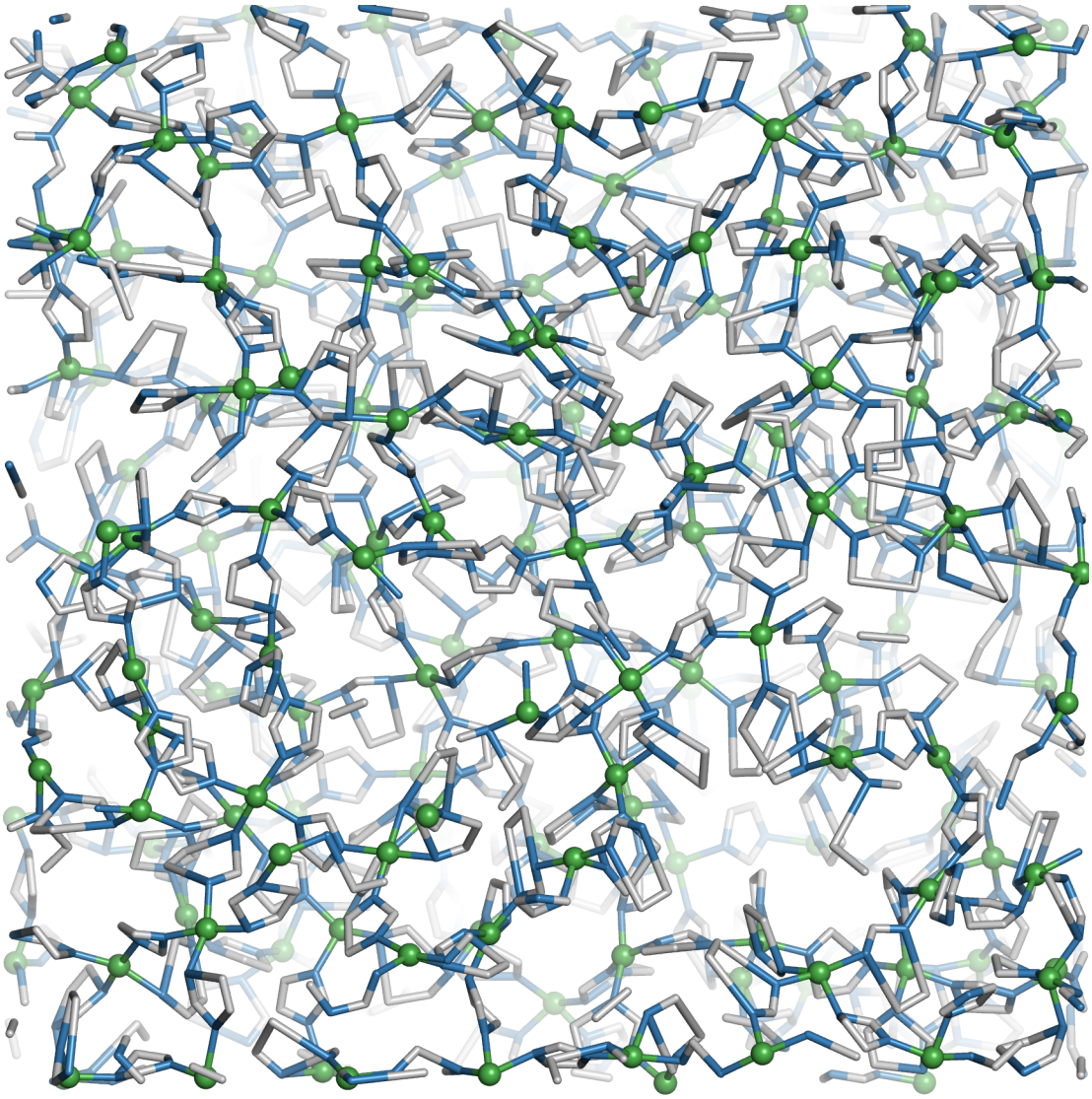


Figure B.5: Refined Pd(im)$_2$ at a-ZIF density. $\rho_1 = 0.0722\,\text{Å}^{-3}$.

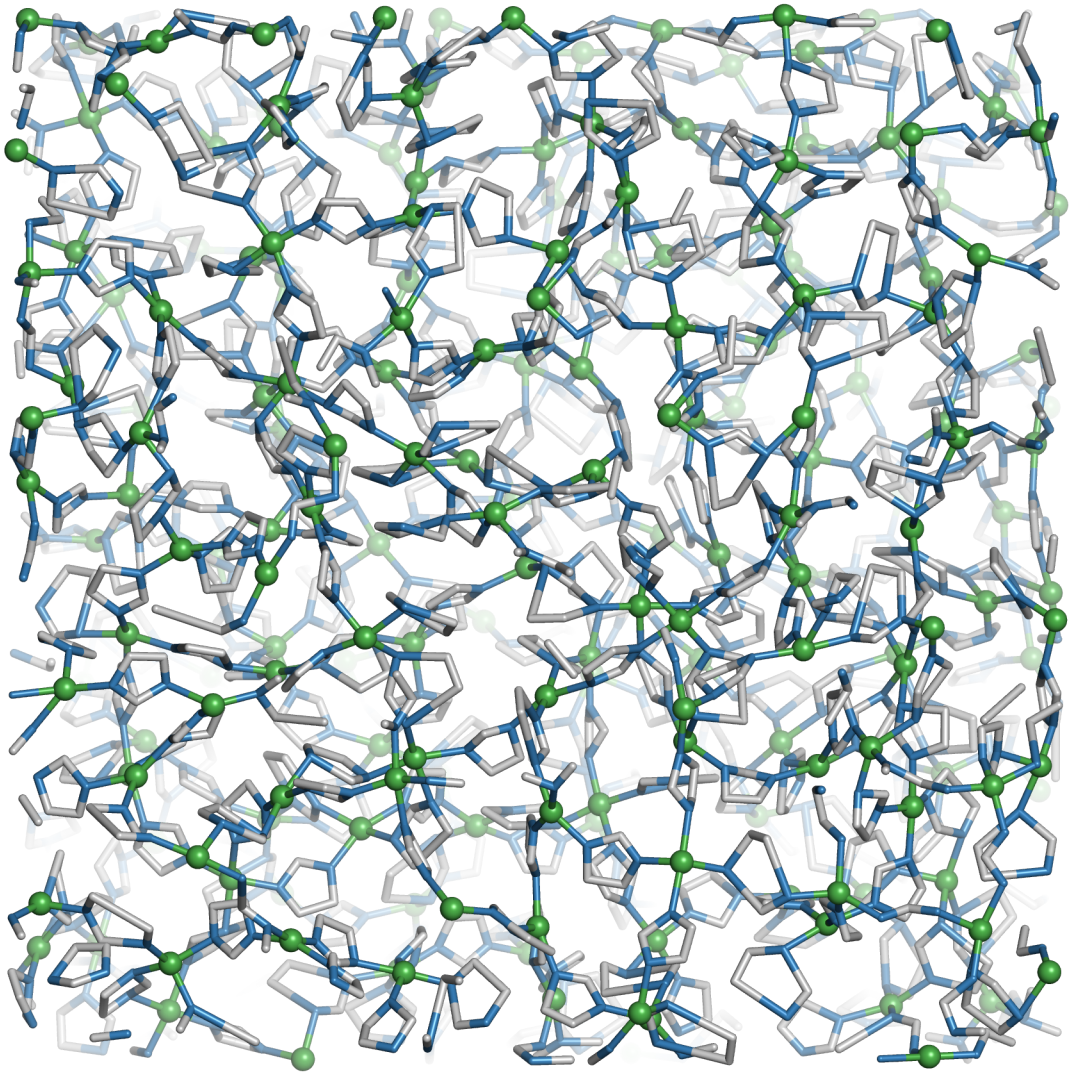


Figure B.6: Refined Pd(im)$_2$ at mid-range density. $\rho_2 = 0.0825\,\text{Å}^{-3}$

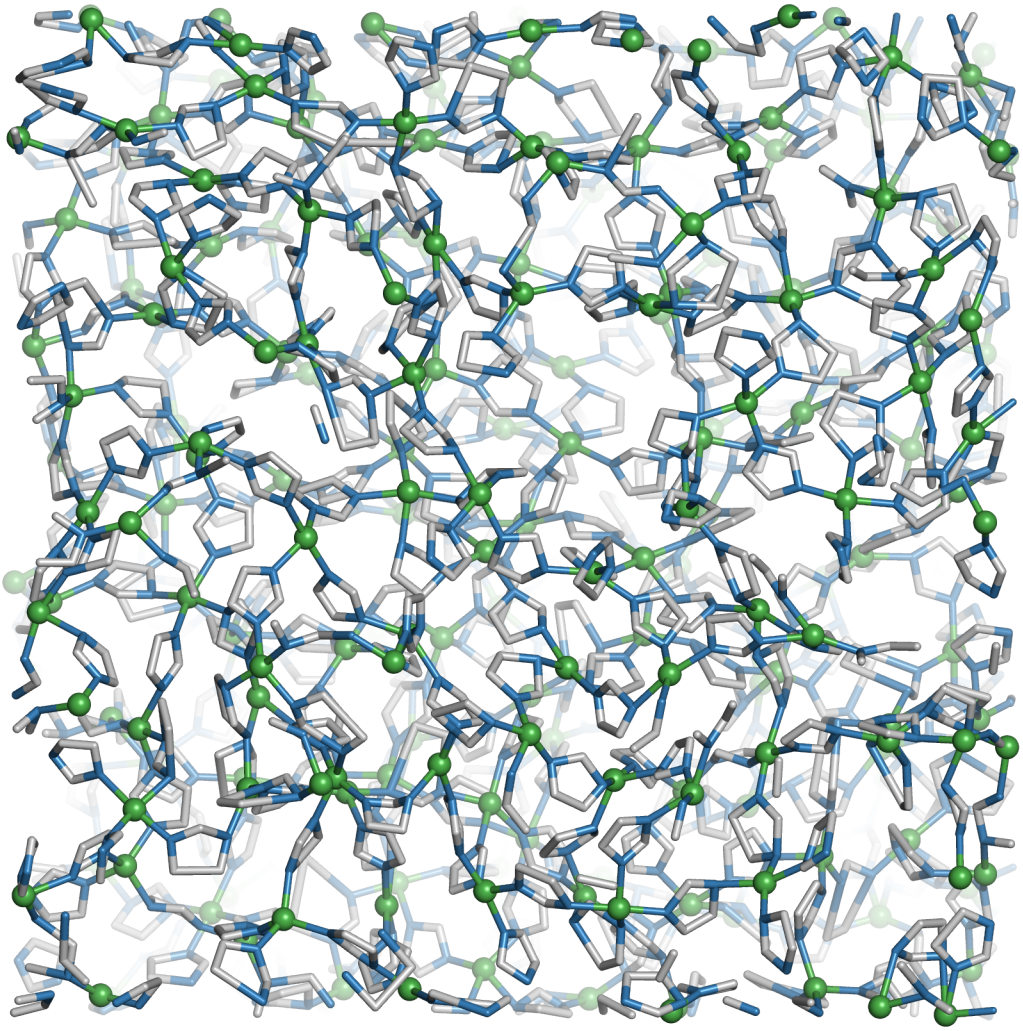


Figure B.7: Refined Pd(im)$_2$ at corrected Pd(Him)$_2$(im)$_2$ density density. $\rho_3 = 0.0998\,\text{Å}^{-3}$.

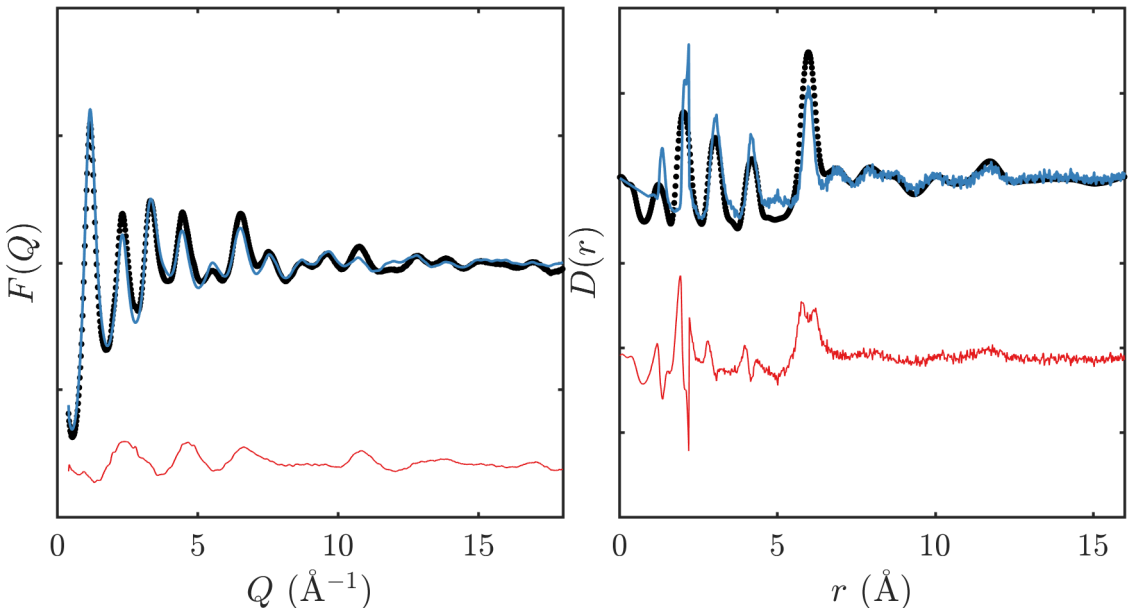


Figure B.8: Detailed total scattering refinements for a D_{4h} model. Total scattering (left) and PDF refinements (right) with residuals offset (red).

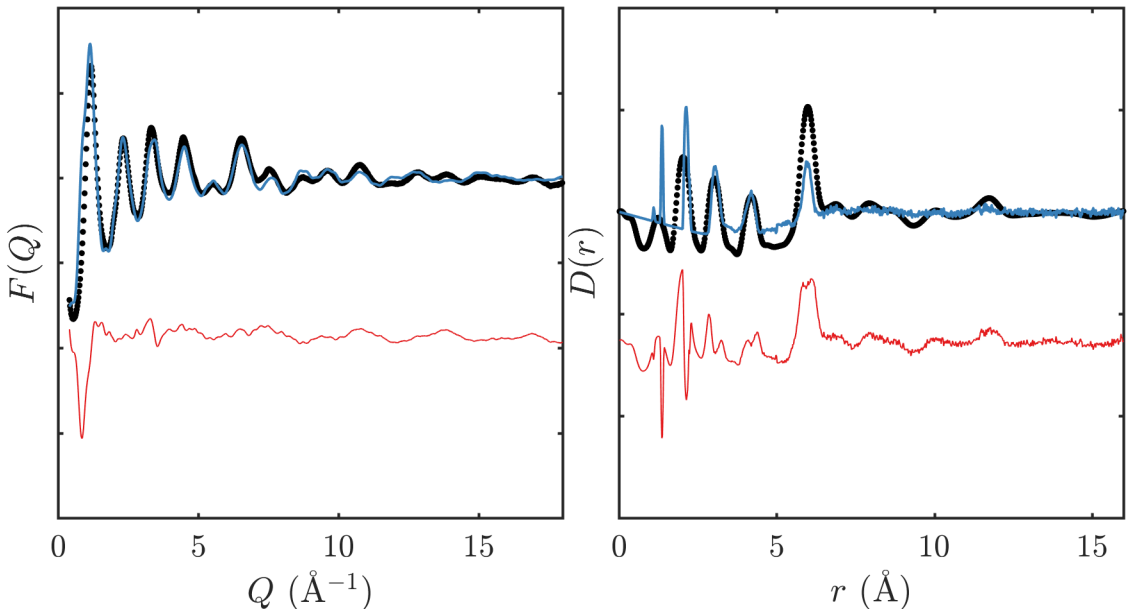


Figure B.9: Detailed total scattering refinements for a T_d model. Total scattering (left) and PDF refinements (right) with residuals offset (red).

Appendix C

Relating chemical shift and structure

The relationship between structure and chemical shift has long been recognized as an important avenue for research and understanding [532]. Initially recognized as a qualitative relationship [533], interest in relating the two characteristics has yielded developments in empirical [484, 534], *ab initio* [535], and machine learning [485, 536] approaches for estimating chemical shifts from structural data. The diversity of approaches is unsurprising given the breadth of structural features available to influence shifts.

Contained in this section is a historical review of the methods relating chemical shift and protein structure. Throughout, particular note is taken as to how these approaches are invoked to inform other modeling techniques. We begin with approaches in the forward problem of calculating chemical shift from a known structure, and then move to the inverse problem of constraining the structure based on the shifts.

C.1 Structure to chemical shift

There are a multitude of opportunities for employing calculated chemical shifts. One such example is in the area of chemical shift assignment. In biomolecular NMR, resonance assignment schemes often depend on prior knowledge of approximate chemical shifts for the species under investigation [537]. It is expected that with accurate prediction of ^{1}H, ^{15}N, and ^{13}C chemical shifts prior to protein assignment, it would make the sequential referencing more efficient and less dependent on NOE or scalar coupling information. Another avenue of utility involves using

calculated chemical shifts as a measure of validity or scoring for structures refined by other techniques, either dynamically [503] or following refinement [538]. Shift predictors have also been utilized in creating large databases of protein structures with assosicated shifts for mining or machine learning [539]. Finally, in order to solve the inverse problem of going directly from chemical shifts to complete tertiary structures, it is considered helpful to first solve the forward problem of predicting chemical shifts from a given structure [501].

Many of the programs that follow rely substantially on databases, and thus need to be cautious of not only the breadth of the structures available, but also the accuracy of the information contained within. The relevant observable for these algorithms is the secondary chemical shift ($\Delta\delta$), defined as the difference between the absolute chemical shift (δ_{abs}) and the corresponding random coil (δ_{rc}) shift [483].

$$\Delta\delta = \delta_{\mathrm{abs}} - \delta_{\mathrm{rc}} \tag{C.1}$$

The random coil chemical shifts are extracted from databases or model peptides under experimental conditions which induce a denatured state. These random coil values however, are starkly dependent on the experimental conditions used to produce the denatured states and the environment of the unfolded proteins [540]. Further, shifts are strongly sequence dependent, particularly in the case of $^{15}\mathrm{N}^{\mathrm{H}}$, $^{1}\mathrm{H}^{\mathrm{N}}$, and $^{13}\mathrm{CO}$, and additive correction factors can be constructed [541]. These needs for correction are compounded with potential mis-assignments, typographical errors, and chemical referencing errors in the BMRB (a user-policed database). Programs such as REFCOR [542] and ARECA [543] attempt to use available information to determine referencing corrections. RefDB is a database created and commonly implemented by the Wishart group that utilizes SHIFTX to create a reference corrected database from the BMRB [544]. Many of the programs which depend on large sets of prior information will carry out some means of reference correction on their data set.

There are several families of algorithms which have been developed for predicting chemical shifts. Of these, one of the earliest methods depended on sequence homology to make predictions of chemical shifts [537]. Operating on the assumption of structural — and thus shift — similarity amongst homologous proteins, the SHIFTY algorithm mined the BioMagResBank (BMRB) [545, 546] for similar proteins and developed a procedure for transferring the shift assignment from the homologues to the query protein. Bax and coworkers took a different approach to homology, focusing on triplets of adjacent residues. Searching a database developed for TALOS [547], SPARTA [485] searches for triplets which provide the best match in torsion angles and sequence similarity to the query triplet, and creates a weighted average to estimate the secondary shifts of the central residue in the query triplet. Another early advance in accurate prediction was the program SHIFTS [535], which depended on a database of chemical shifts, computed *ab initio* with density functional theory, and an additive model of contributions by corresponding conformational effects found in the database. Programs of this nature are characteristically dependent on the quality of the database, and even upon expansion of the database and number of fitted terms included, they can often fail to accurately handle non-canonical structures [548].

In keeping with diversity of nomenclature, the early benchmark program SHIFTX was developed by the Wishart group for empirical shift calculations [484]. This program uses empirically derived hypersurfaces (capturing dihedral angle, sidechain orientation, secondary structure and nearest neighbor effects) in combination with classical or semi-classical equations (for ring current, electric field, hydrogen bond, and solvent effects) to calculate shifts from atomic coordinates. This is an effective compromise between the speed of the empirical approaches and accuracy of the quantum mechanical approaches. To generate the hypersurfaces, data mining techniques were used to extract the largest pairwise contributions from the RefDB [544] and a set of high resolution X-ray structures solved from traditional crystallographic

techniques. The nature of the semi-empirical method is a resulting set of discontinuous functions for shift which keeps the program from being amenable to refinement techniques which require first derivatives; however, methods that implement a simplex minimization, or Monte Carlo or Genetic algorithm could readily exploit the SHIFTX program. The necessity for differentiable functions lead to a set of prediction methods which expresses shifts as a function of interatomic distances. These include CamShift [501] for backbone shifts that shows a modest improvement over SHIFTX, ArShift [549] for aromatic proton shifts, and CH3Shift [550] for methyl shifts. The utility of continuous, differentiable functions for predicting shifts is most useful in developing better molecular mechanics force fields for proteins.

Other efforts to calculate chemical shifts have utilized increasingly accessible and improved machine learning techniques. The first of these models was an artificial neural network (ANN) implemented by PROSHIFT [551] that successfully mapped well known dependencies between the secondary structure and the backbone chemical shifts. PROSHIFT had limited success with predicting sidechain shifts, but was most notable for its gains in speed over DFT models with comparable accuracy. As databases such as the BRMB became richer with submissions, the training for machine learning algorithms became more efficacious. An ANN algorithm which outperformed PROSHIFT and the aforementioned empirical methods, was SPARTA+, which used input parameters similar to SPARTA to train the network [487]. The artificial neural networks train an empirical equation, by adjusting the weights and biases in the learning phase. This gives a rapid calculation of chemical shifts, as well as potential for on the fly utility. Using the RefDB, SHIFTX2 was developed to implement the structure-based concepts included in SHIFTX [484] and sequence based concepts from SHIFTY [537] with advanced machine learning techniques [486]. The significant gains in accuracy were a result of training on a high quality database, and including more features than the two programs used in isolation. A final implementation of machine learning involves the mixed methodology of PPM_One [488]. The

shift predictor makes utility of an ANN and most notably emulates the effect of local protein dynamics via interatomic steric contacts. Previous predictors produced by the same group demonstrated the improvement gained in understanding side chain chemical shifts by calculating shifts from MD ensembles instead of isolated structures [552]. PMM_One is comparable or underperforms SPARTA+ for all backbone atoms, but supersedes all available programs when addressing sidechain atoms.

C.2 Chemical shift to structure

With respect to our objective of structural refinement, a similar pace of advancements have been made that elicit structural information or restraints from chemical shift data alone. These restraints can come as secondary structure assignments [553–555], torsion angle prediction [547, 556], or statistical distributions [489]. There are also several sequence based techniques that predict torsion angle [557, 558] and/or secondary structure [559] with expectedly less accuracy. Because these programs are implicated in the *de novo* modeling pipeline that follows, they will be described in significant detail.

The first family of programs to make significant progress in predicting structure from chemical shift was produced by the Bax group. TALOS [547] was the predecessor to SPARTA, and originally depended on a database of 20 proteins, which it searched for similarity in sequence and chemical shift of three residue fragments. The average of the dihedral angles (ϕ, ψ) from the center residue of the 10 best triplets was then predicted for the query residue. These predictions had confidence values of 'good/strong', 'ambiguous', or 'bad'. The similarity score depended on an empirically adjusted weighting between sequence homology and shift matching. Even though this first iteration of the program showed promise, the small database tended to sample the most populated regions of the Ramachandran map. A later development of the program included a larger database (200 proteins from the BMRB) and a two layer neural network filter which drastically increased the prediction rate in TALOS+ [560]. The network took 78 inputs from each tripeptide and output

a 3-state backbone ϕ/ψ distribution, the DSSP secondary structure [440], and the predicted dihedral angles. The program also instituted the random coil index (RCI) [438, 439] as a measure of protein flexibility, and found that the prediction failed in loop regions. The TALOS programs all utilize chemical shift data from ^{15}N, $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$, and $^{1}H^{\alpha}$; however, can attempt and succeed in structure prediction with incomplete shift data.

The competitive program put forth by the Wishart group to calculate protein torsion angles is a web server, PREDITOR [561] (A similar algorithm is presented as source, titled SHIFTOR [556]). Actually a combination of many subprograms, PREDITOR accepts sequence and/or chemical shift data as input and generates torsion angle predictions and predicted errors for ϕ, ψ, ω, and χ_1. Operating under a a homology principle, chemical shifts are extracted from the RefDB and homology from the PDB for tripeptide segments. VADAR [562] is used to generate torsion angles from the PDB segments and sequence alignment is accomplished by BLAST [474, 475]. Other additional information is available in the server, including secondary structure prediction from CSI [555], RCI flexibility [439], and chemical shift reference corrections from REFCOR [542]. Furthermore, the probabilistic hypersurfaces calculated by Dunbrack [563] are used to predict χ_1 angles based on ϕ/ψ angles. Finally, the confidence scores are based on the similarity scores, $S(i,j)$ used in searching the database and given in Equation C.2.

$$C = 2^{(3*S(i,j)/239)^2} \tag{C.2}$$

The Wishart group also developed a series of programs designed to accurately predict secondary structure. Usually, secondary and super-secondary structure are predicted using NOE data. The Chemical Shift Index (CSI) is a method for predicting secondary structure from only shift information [555, 564], which applied a three part digital filter to separate shifts into clustering of secondary structure type and was more accurate for helix structures than β-sheets or coils. Improvements

to this early method to include a support vector machine trained from a data set of 240 proteins, produced accuracy which rivaled secondary structure classification from atomic coordinates [483]. In this method, CSI 2.0, missing info was filled in with predictions from programs similar to those in described in the previous section, and many feature vectors were used including information from the RCI. The culmination of this series is in a combination of several programs to create a rich set of outputs to predict secondary structure [565]. CSI 3.0 combines 4 programs in order (CSI 2.0 [483], RCI [439], TALOS-N [490], and Sidechain RCI [566]) to identify 11 types of secondary structure: helicies, β-strands, coil regions, five common β-turns (type I, II, I′,II′, and VIII), β hairpins, and interior and edge β-strands. The CSI 3.0 predictions show greater than 90% agreement with high resolution structures, and has potential to produce rich topological pictures or structural restraints for refinement.

Recently, a program titled DANGLE was developed that employs Bayesian inference to estimate the likelihood of all possible values of the dihedral angles (ϕ,ψ) for each residue in the query protein [489]. The major distinction associated with this program is degeneracy available in the output. As chemical shifts may be consistent with multiple conformations, other programs such as TALOS and PREDITOR predict angles from the major cluster (most probability dense region) in Ramachandran space, whereas DANGLE produces a probability distribution for the end user to work with. This avoids the failure many programs face when handling non-canonical structures, such as 3_{10} helices, positive values of ϕ, the flexibility of glycine, and the possible cis peptide bond in residues preceding proline. Contrary to the low prediction rate and accuracy in comparison with TALOS, when the full GLE diagrams are included — not just the principle cluster — $>$ 90% of all experimental ϕ and ψ would appear within the boundary ranges of the estimate. The quality of considerations in conjunction with the non-specificity of the output, makes DANGLE a strong program for implementation in an already Bayesian structure refinement pro-

cedure. Therefore, significant detail will be presented below outlining its operating principles and procedure.

DANGLE begins its algorithm by dividing its residues into four classes: glycines, prolines, residues preceding proline, and 'generic'. Adapting some homology and the 186 protein database associated with TALOS [547], Dangle generates pentapetide fragments and stores the primary structure, experimental ϕ and ψ angles, and a three state classification of secondary structure for the central residue from the DSSP program [440]. Next the four classes from above were specified. Then population frequency maps spanning Ramachandran space, discretized in 10° intervals, were constructed with the reference minimum population being small and non-zero.

The input sequence and shift data are used to compute secondary shifts (^{15}N, $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$, and $^{1}H^{\alpha}$), and then search the fragment database for close matches, saving the 10 lowest scoring matches to build the query scatter pattern (ϕ, ψ) (QSP) that is transformed into a smoothed frequency matrix. The matches are scored according to a function similar to that employed by TALOS,

$$S(i,j) = \sum_{n=-2}^{+2} \left[\begin{array}{lllll} & k_n^0 \Delta^2_{\mathrm{ResType}} & & + & k_n^1(\Delta\delta \mathrm{C}^{\alpha}_{i+n} - \Delta\delta \mathrm{C}^{\alpha}_{j+n})^2 \\ + & k_n^2(\Delta\delta \mathrm{N}_{i+n} - \Delta\delta \mathrm{N}_{j+n})^2 & & + & k_n^3(\Delta\delta \mathrm{C}^{\beta}_{i+n} - \Delta\delta \mathrm{C}^{\beta}_{j+n})^2 \\ + & k_n^4(\Delta\delta \mathrm{C}^{'}_{i+n} - \Delta\delta \mathrm{C}^{'}_{j+n})^2 & & + & k_n^5(\Delta\delta \mathrm{H}^{\alpha}_{i+n} - \Delta\delta \mathrm{H}^{\alpha}_{j+n})^2 \end{array} \right] \quad \text{(C.3)}$$

where i is the central residue in the query window and j the central residue in the database window; $\Delta^2_{\mathrm{ResType}}$ is the square of a residue-type homology score, and the k_n^x are empirically adjusted weighting factors; n represents the position of the current residue with respect to the central residue; and $\Delta\delta \mathrm{X}_{i+n}$ is the secondary chemical shift of nucleus X fro the current residue.

Central to the method is the use of Bayesian inference to produce the posterior probability of a given backbone conformation (ϕ, ψ) given a QSP, $P(\phi, \psi|\mathrm{QSP})$. From Bayes' theorem,

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} \quad \text{(C.4)}$$

it can be seen that the posterior probability is proportional to the likelihood and prior probability. Thus for DANGLE we have,

$$P(\phi, \psi|\text{QSP}) \propto P(\text{QSP}|\phi, \psi) \times P(\phi, \psi) \tag{C.5}$$

Likelihood is given by the population frequency maps described above, assembled by a database of 500 high resolution X-ray structures. The prior probability $P(\text{QSP}|\phi, \psi)$ represents the conditional probability that query scatter pattern could be produced by a conformation in a given 10° bin. For each occurrence of experimental (ϕ, ψ) in the bin, a new similarity search identified the 10 lowest scoring pentapeptides to construct a QSP 'scattergram', thus defining the shape of QSP's commensurate with the binned conformation. A QSP will be representative a conformation if it resembles the scattergram associated with that bin. This similarity is quantified using the Kullback-Leibler divergence, $D_{\text{KL}}(\phi, \psi)$:

$$D_{\text{KL}}(\phi, \psi) = \sum_{\phi'} \sum_{\psi'} Q(\phi'\,\psi') \times \ln\left(\frac{Q(\phi'\,\psi')}{S(\phi'\,\psi'|\phi, \psi)}\right), \tag{C.6}$$

where $Q(\phi'\,\psi')$ is the value of the cell $(\phi'\,\psi')$ in the QSP and $S(\phi'\,\psi'|\phi, \psi)$ is the value of the cell $(\phi'\,\psi')$ in the scattergram matrix associated with the bin (ϕ, ψ). The behavior of $D_{\text{KL}}(\phi, \psi)$ tends to zero if the query frequency matrix and the current scattergram are a perfect match. To equate this value of zero to a probability of unity in a normalized distribution, $P(\text{QSP}|\phi, \psi)$ is calculated as:

$$P(\text{QSP}|\phi, \psi) = \frac{\exp(-D_{\text{KL}}(\phi, \psi))}{\sum_{\phi'} \sum_{\psi'} \exp(-D_{\text{KL}}(\phi', \psi'))} \tag{C.7}$$

Finally, combining equations (C.7) and (C.5), the posterior probability score, B, for a conformational bin (ϕ, ψ) is given by the equation:

$$B(\phi, \psi) = \frac{\exp(-D_{\text{KL}}(\phi, \psi))}{\sum_{\phi'} \sum_{\psi'} \exp(-D_{\text{KL}}(\phi', \psi'))} \times P(\phi, \psi) \tag{C.8}$$

The values of $B(\phi, \psi)$ are assembled into a global likelihood estimate (GLE) diagram which resembles a lower resolution Ramachandran plot. The program goes further to designate islands in the GLE diagram and predict values of ϕ/ψ. The program was competitive when considering accuracy of the predicted angle (being within a 30° margin of error), but superseded other programs when considering the full range of predicted possible values. This is due to the GLE capturing degeneracy that is not available with a single prediction and confidence interval.

The most recent iteration of the TALSO suite, has a component similar to a GLE diagram buried between two neural networks. TALOS-N [490, 491], relies significantly more on the use of ANNs than its predecessors, and is trained with two databases: one of 580 proteins with experimental shift data, and one with 9,523 proteins and calculated shift data using the prediction program SPARTA+ [487]. Within the two layer network, the first layer produces a Ramachandran like probability density map of likelihood — instead of the 3-state ϕ/ψ distribution in TALOS — which serves as input to the second layer, which outputs a similar smooth probability density map. Cluster analysis is then used to predict ϕ/ψ conformations from the voxels in the the map. Other changes include the use of 5 residue pentapetides instead of tripeptides, and the ability to separate flexible regions into multiple high density pockets in the distribution, encapsulating some of the ensemble averaging of the NMR. A comparison of three recent programs in presented in Table C.1.

	PREDITOR	TALOS-N	DANGLE[c]
Method	Homology	ANN	Bayesian
Database Size	141	580	189
Prediction Rate (%)	97	90	83
Prediction Accuracy (%)	73	87[a]	82
χ' (%)	84	50	0
Q3[b] (%)	—	88	85

Table C.1: Comparison of PREDITOR, TALOS-N, and DANGLE. DANGLE prediction rate is for the accuracy statistic and removal of probability distributions with multiple islands. The accuracy metrics for PREDITOR and DANGLE were calculated on the same dataset, whereas the metrics for TALOS-N are those from the program authors. (a) Accuracy for TALOS is judged by the function $[(\phi_{\mathrm{pred}} - \phi_{\mathrm{Xray}})^2 + (\psi_{\mathrm{pred}} - \psi_{\mathrm{Xray}})^2] < 60\%$, others are for the percentage of angles within 30° of the actual structure. (b) Q3 is a measure of the fraction of secondary structure accurately predicted (alpha helix, beta sheet, loops). (c) Values are considered for principle island and single conformation predictions.

Appendix D

Melittin prior probability distributions

The following contains the prior probability distributions for melittin, considered in the development of PIPA. Empty plots are left intentionally where residue dihedral angle distributions are not included in the priors.

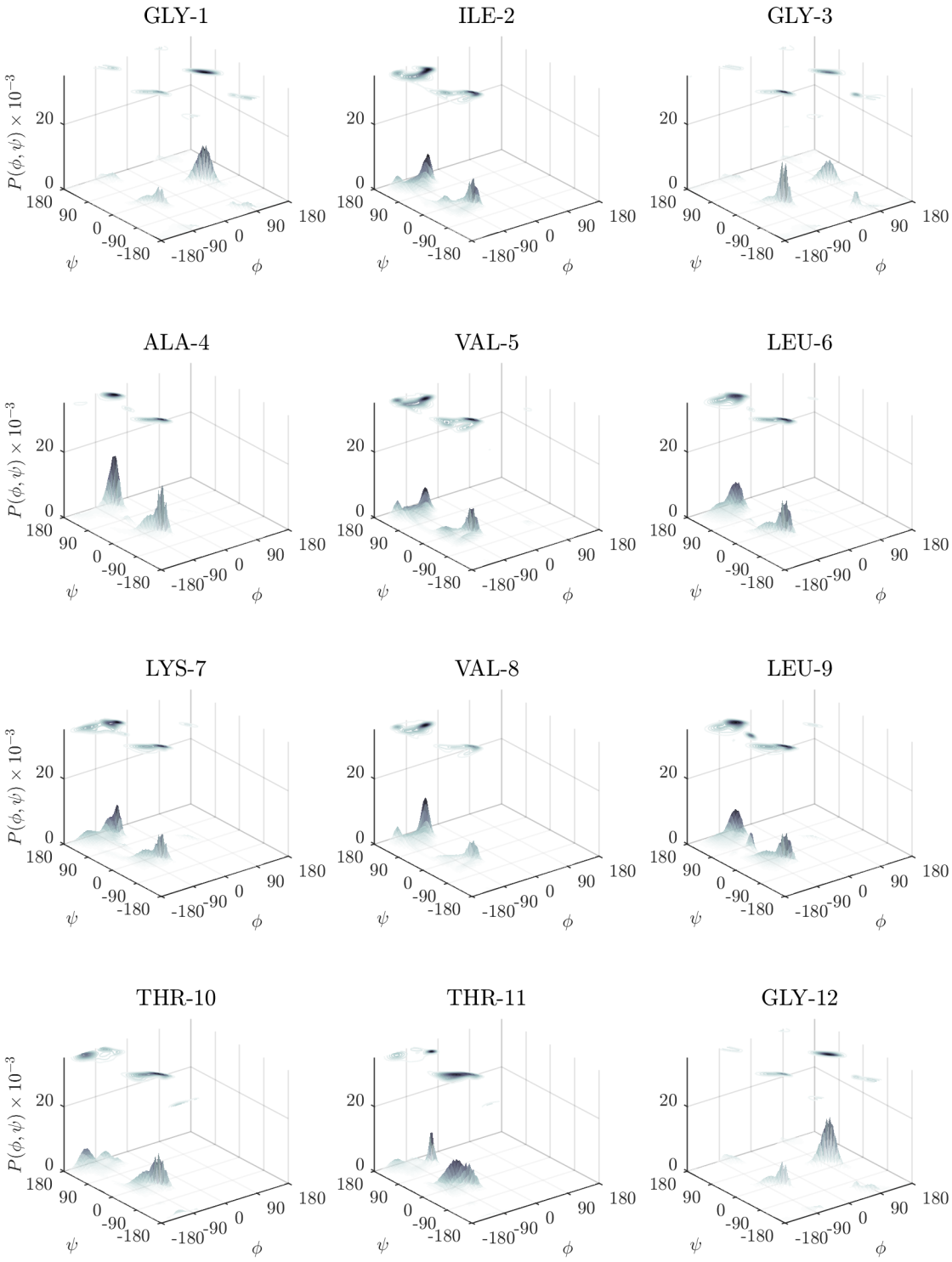


Figure D.1: Neighbor dependent Ramachandran distributions for melittin residues.

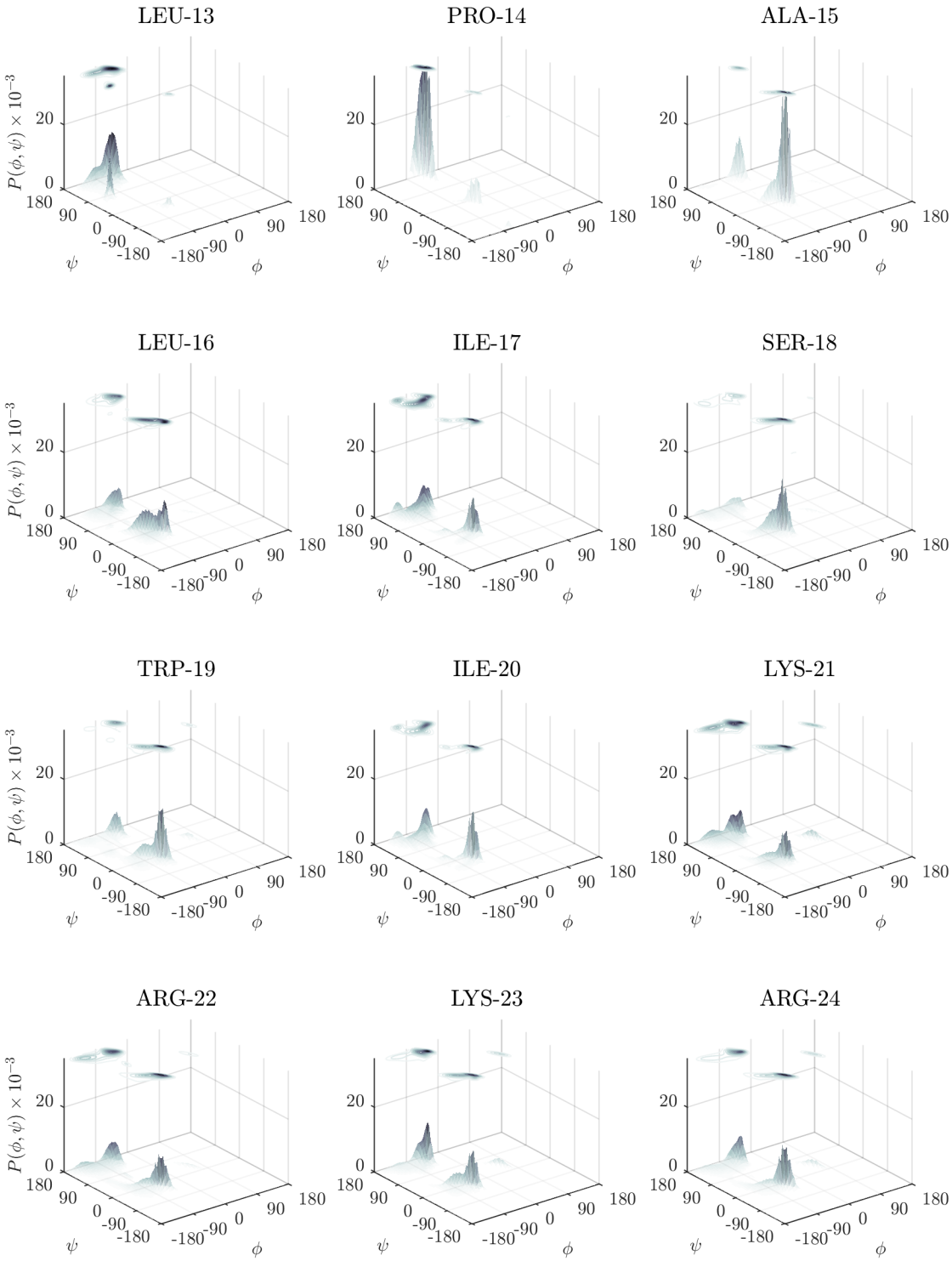


Figure D.2: Neighbor dependent Ramachandran distributions for melittin residues.

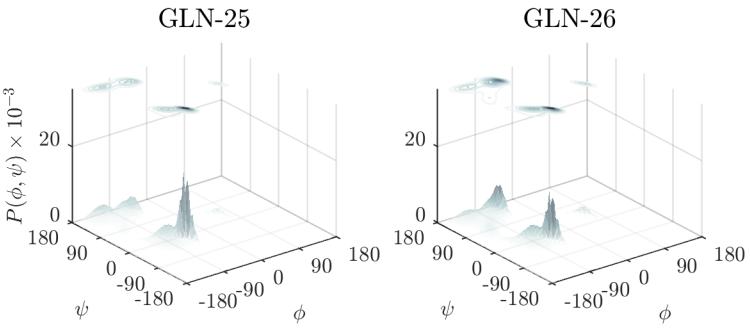


Figure D.3: Neighbor dependent Ramachandran distributions for melittin residues.

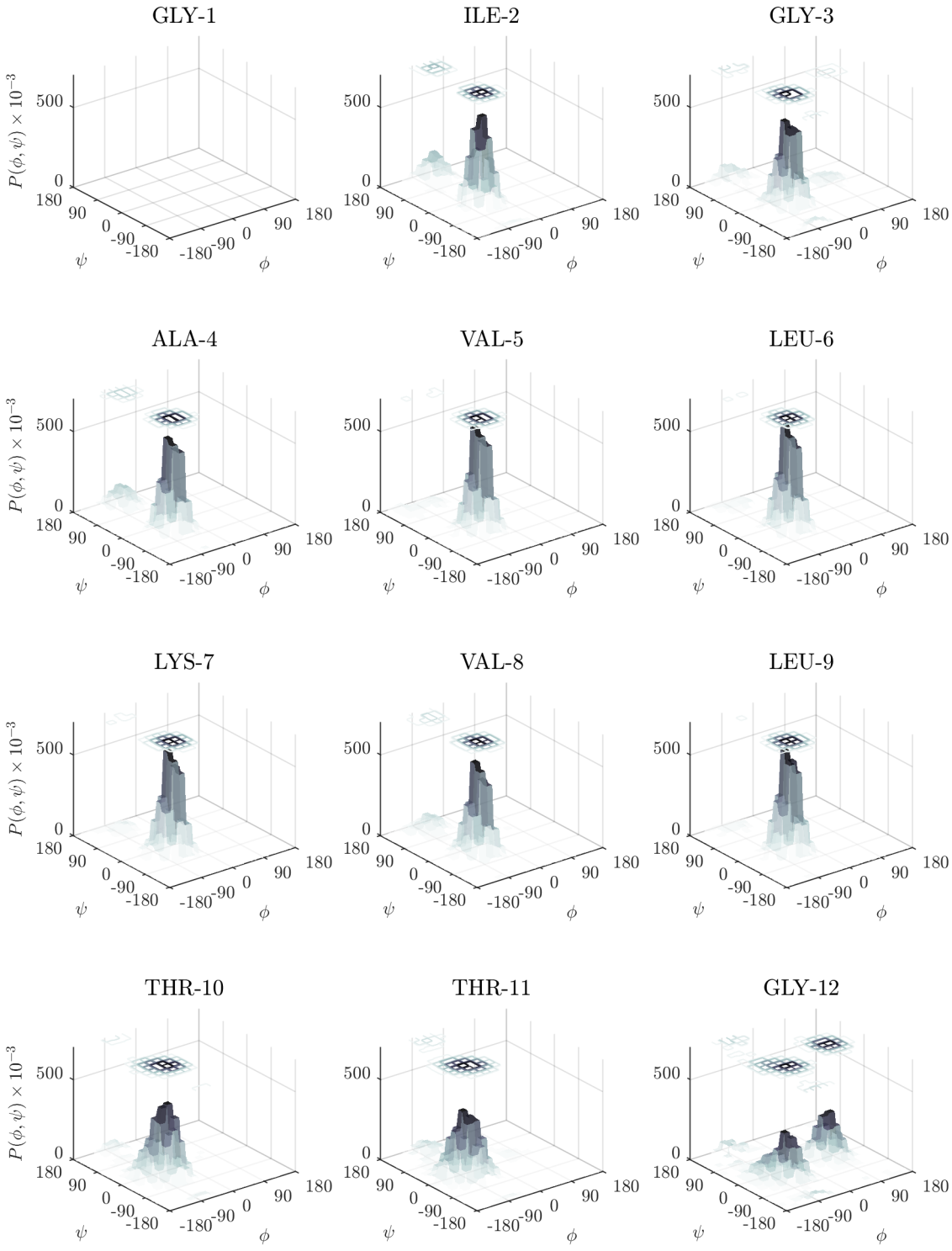


Figure D.4: As produced distributions from TALOS-N for melittin residues.

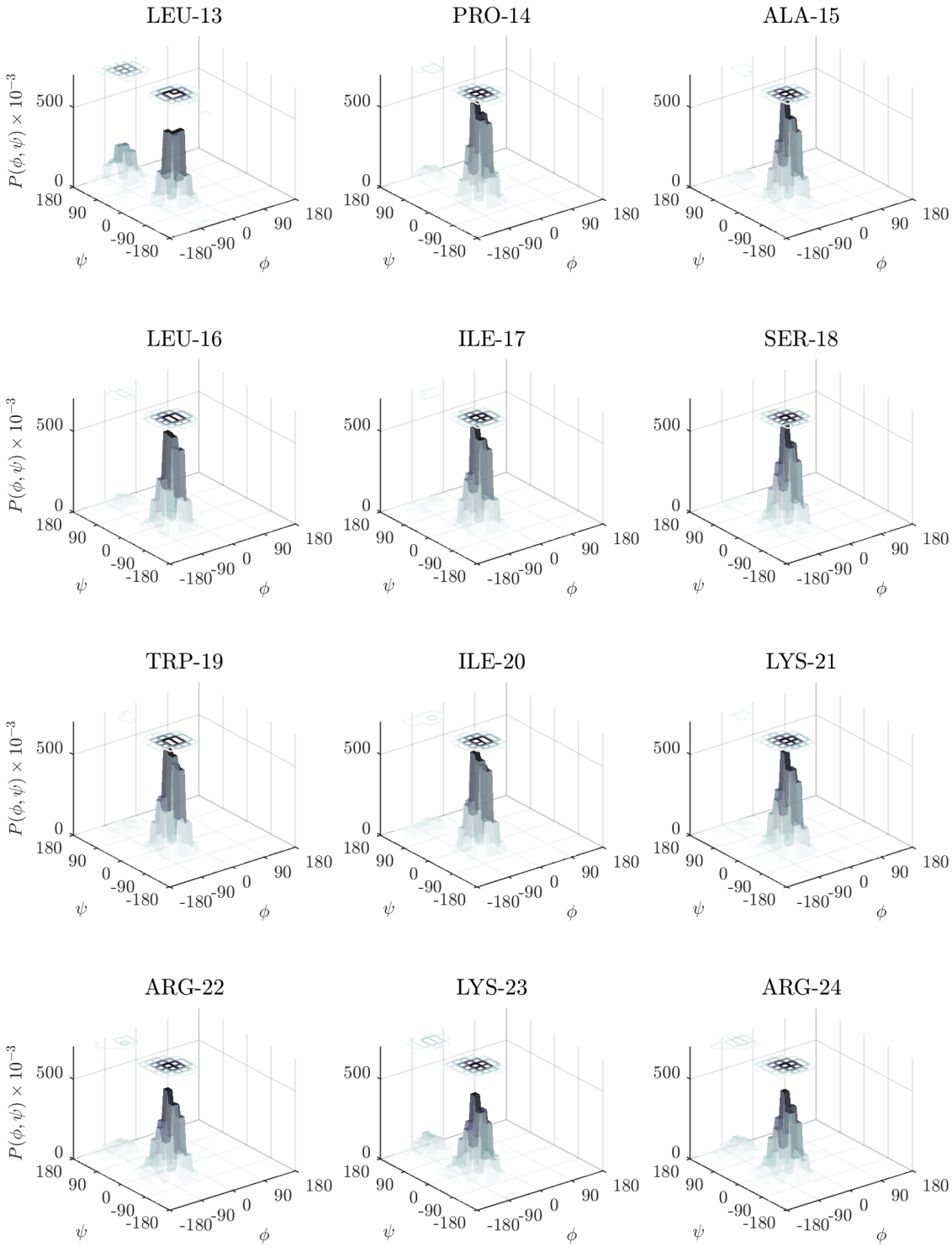


Figure D.5: As produced distributions from TALOS-N for melittin residues.

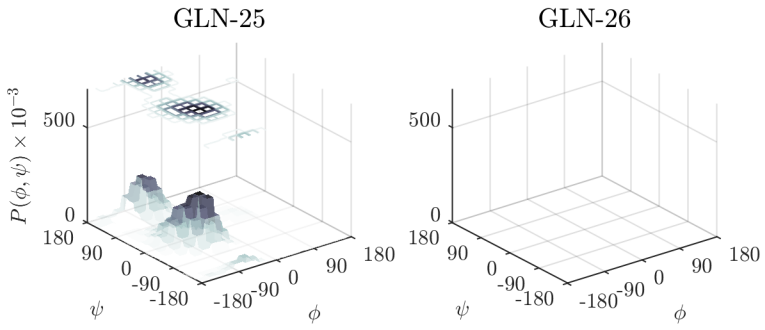


Figure D.6: As produced distributions from TALOS-N for melittin residues.

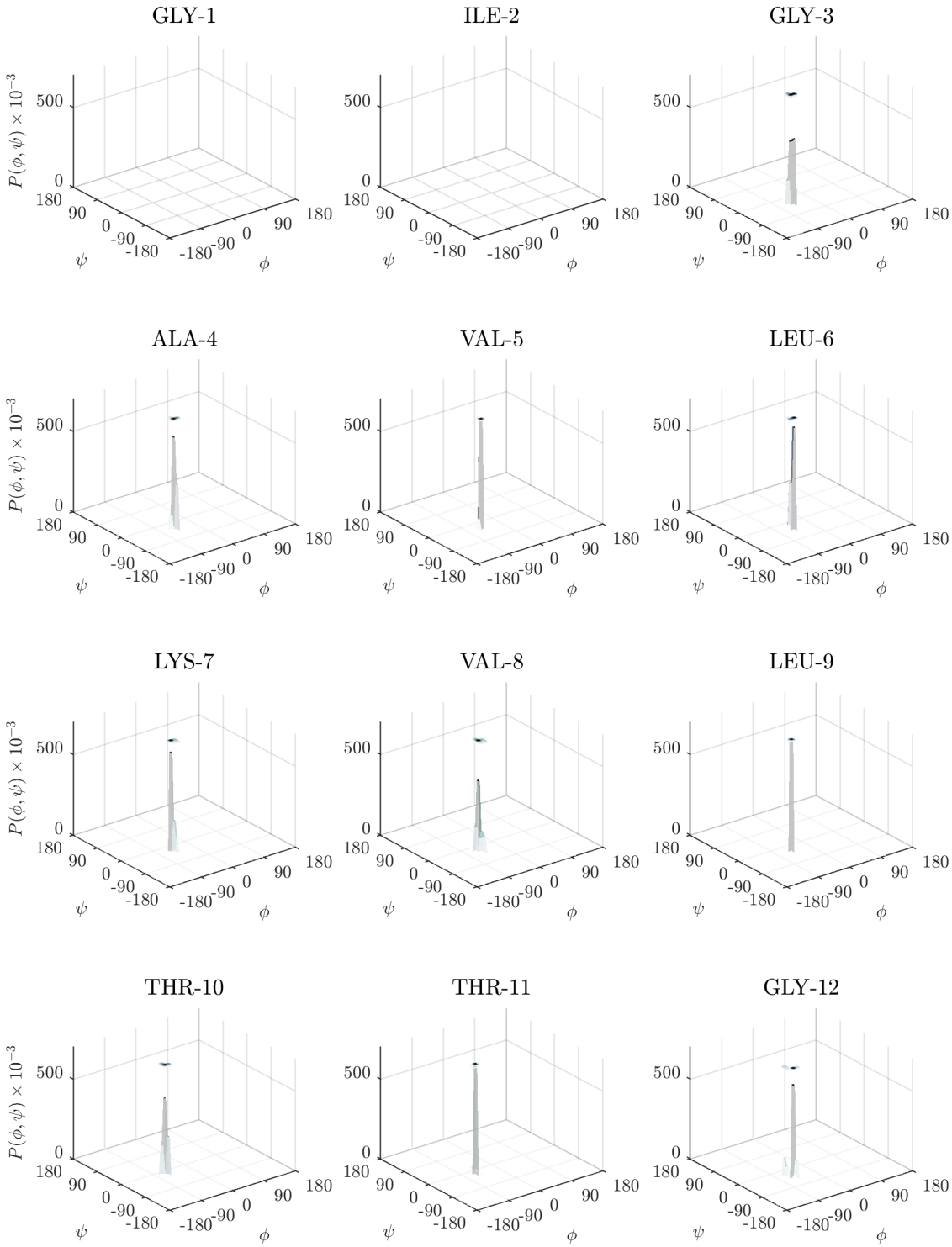


Figure D.7: As produced distributions from DANGLE for melittin residues.

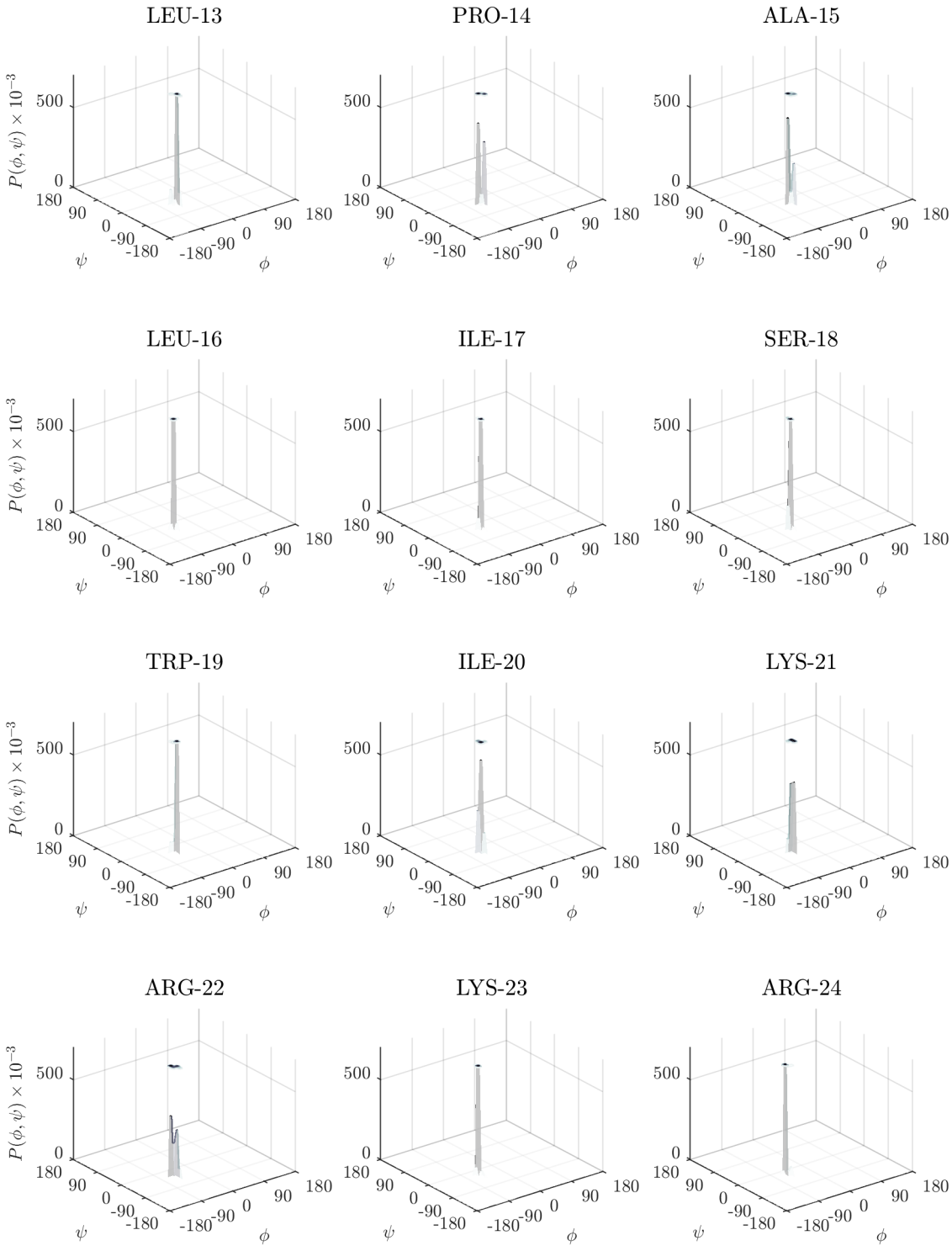


Figure D.8: As produced distributions from DANGLE for melittin residues.

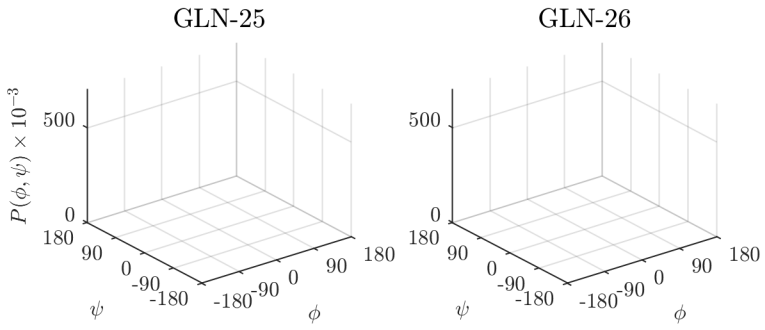


Figure D.9: As produced distributions from DANGLE for melittin residues.

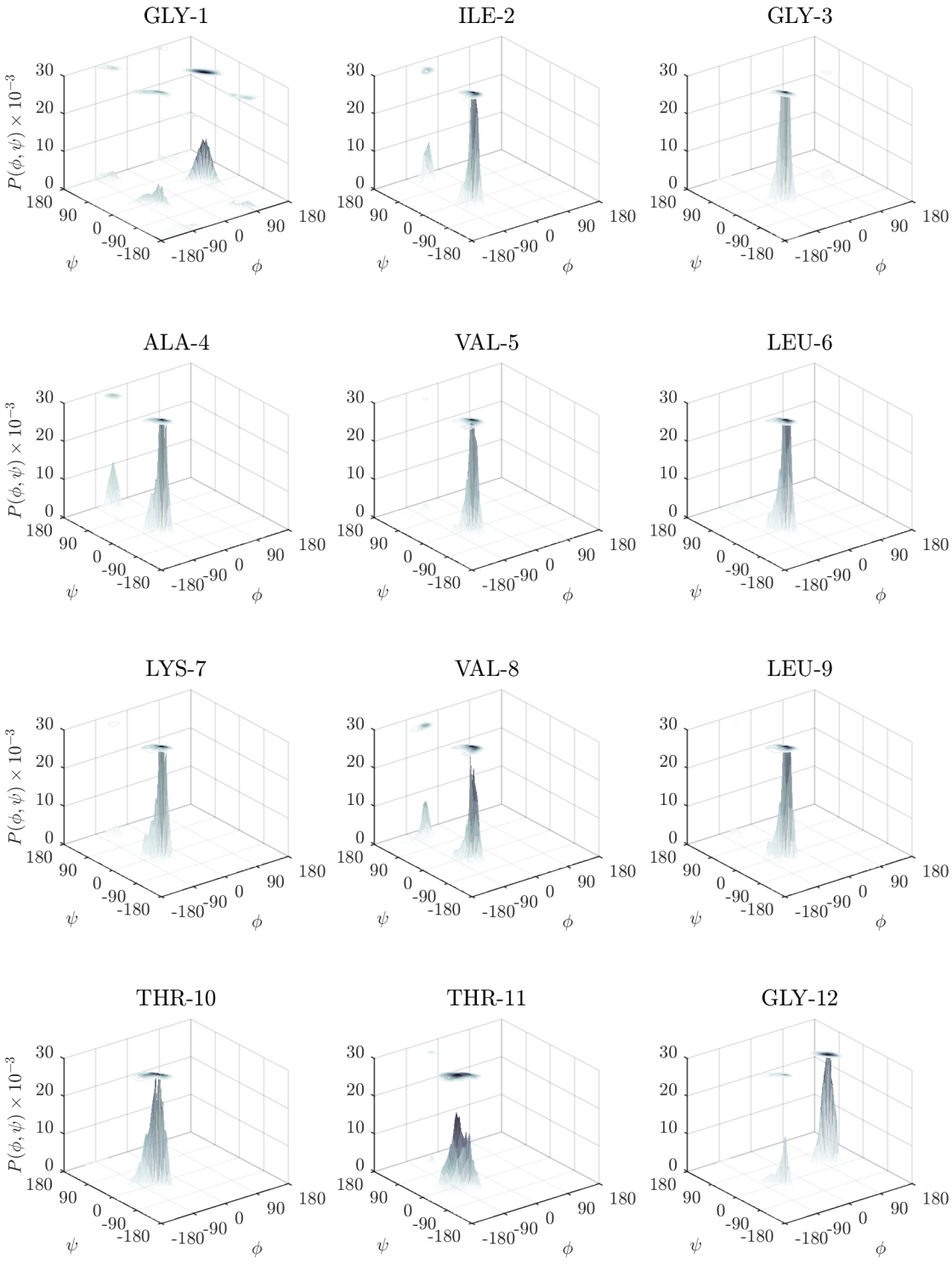


Figure D.10: Distributions from the normalized product of the TALOS-N output and neighbor dependent Ramachandran data for melittin residues.

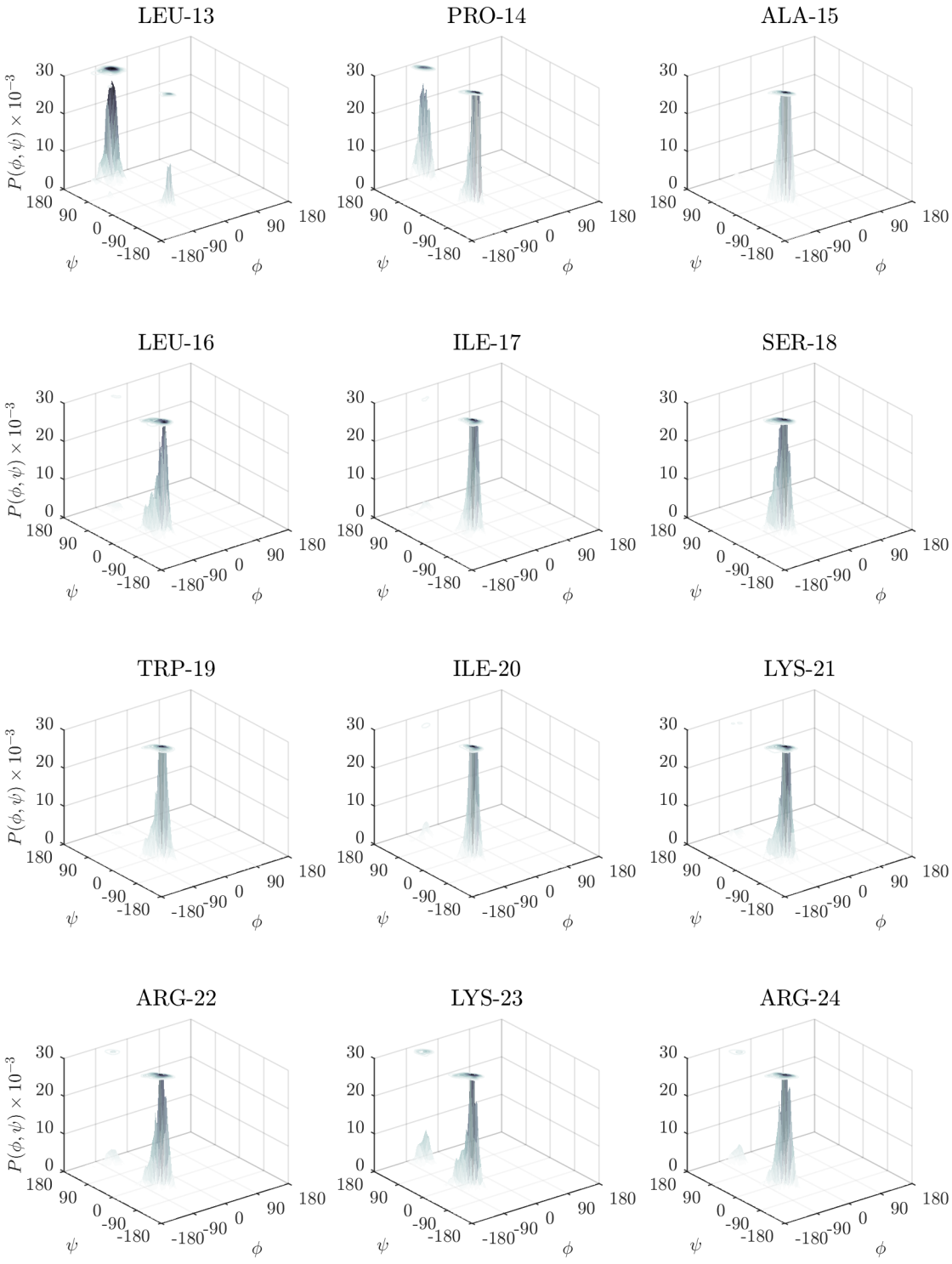


Figure D.11: Distributions from the normalized product of the TALOS-N output and neighbor dependent Ramachandran data for melittin residues.

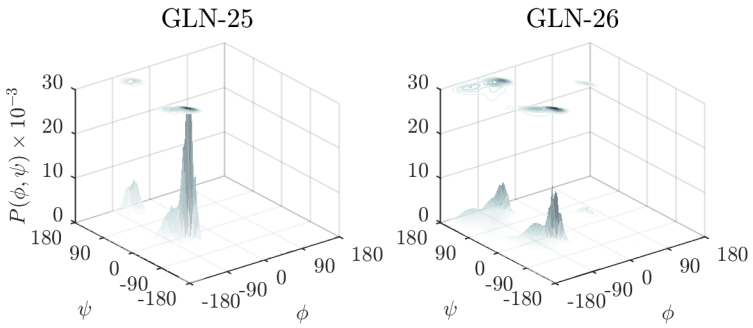


Figure D.12: Distributions from the normalized product of the TALOS-N output and neighbor dependent Ramachandran data for melittin residues.

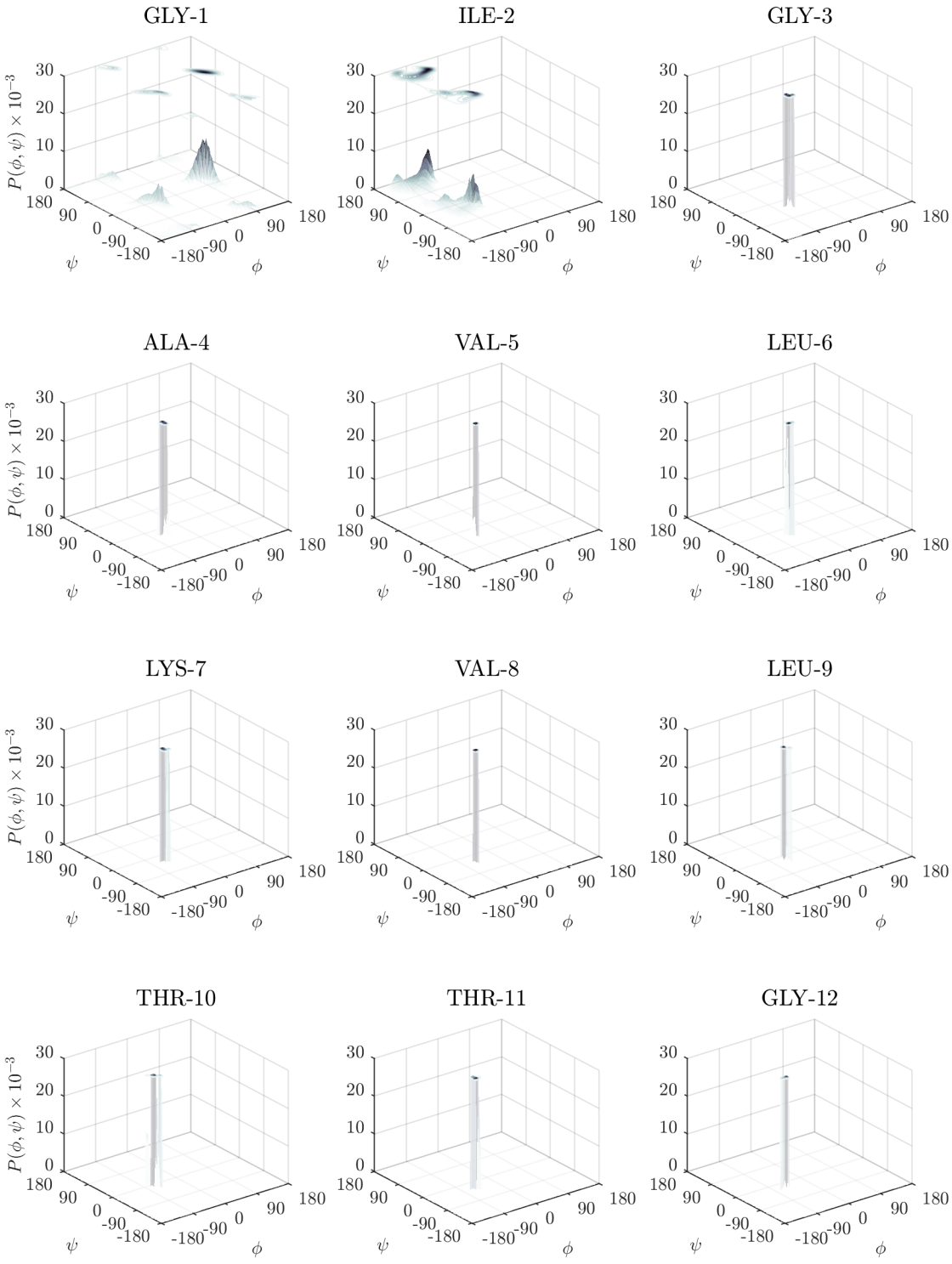


Figure D.13: Distributions from the normalized product of the DANGLE output and neighbor dependent Ramachandran data for melittin residues.

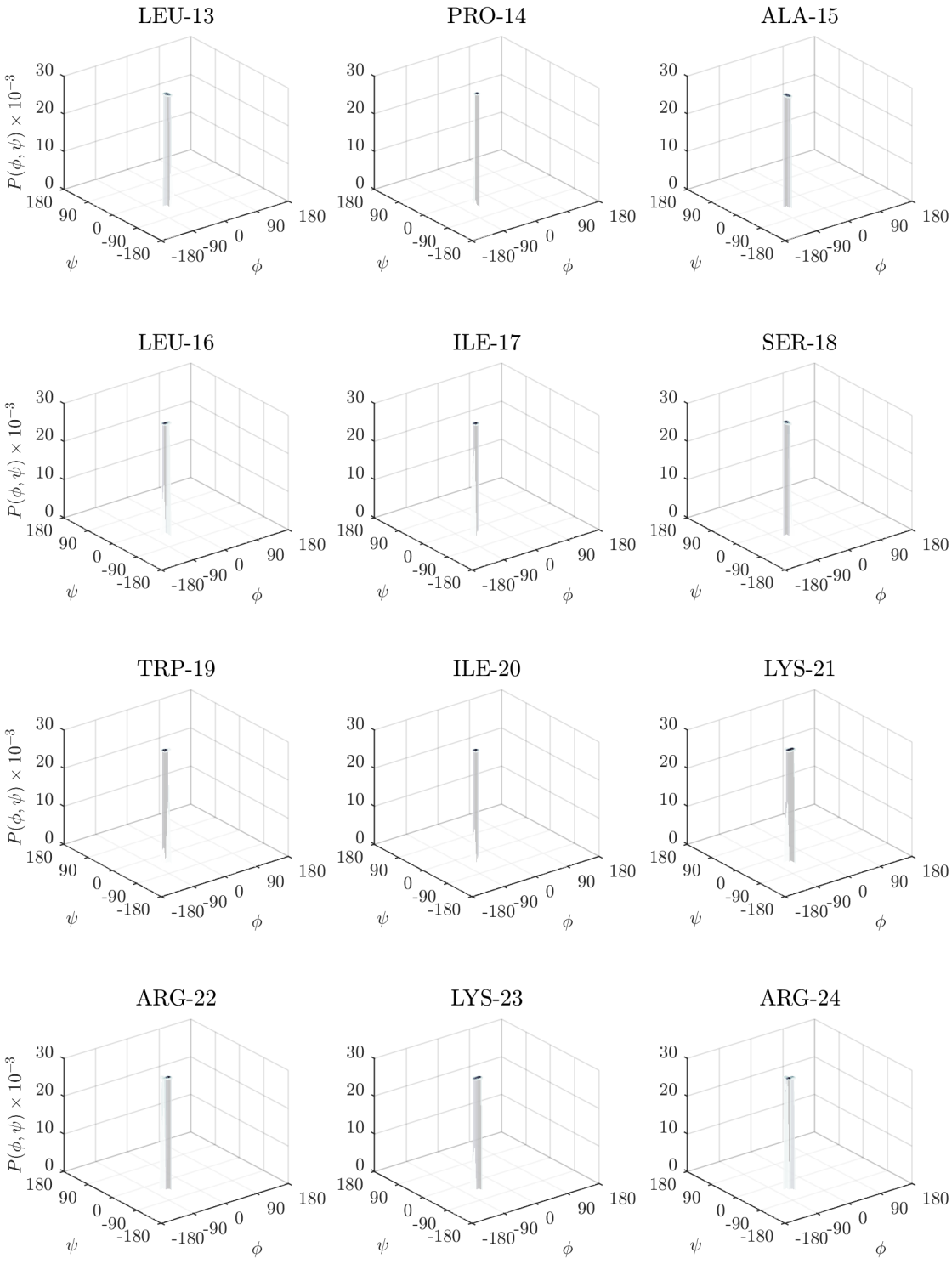


Figure D.14: Distributions from the normalized product of the DANGLE output and neighbor dependent Ramachandran data for melittin residues.

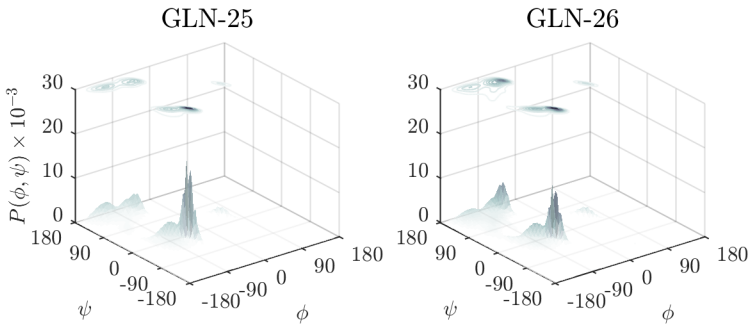


Figure D.15: Distributions from the normalized product of the DANGLE output and neighbor dependent Ramachandran data for melittin residues.

Appendix E

PIPA side chain decoration

Previous work in adding side chains to a protein backbone has had mixed success [567, 568].

In this case the goal in designing a side chain decoration algorithm was not to access the correct orientations of the side chains, but to create a sensible starting structure for further refinement. As clashing between chains and atomic overlap produces divergent potentials in MD, avoiding this common pitfall was a priority.

The side chain decoration program (SCD) was built using the modern Fortran standard and the same objects and methods utilized in PIPA. A modular design allows SCD to act as a standalone program that can be implemented in the workflow of other methods, as well as integrate directly into the PIPA output. As shown in the flowchart in Figure E.1, the major steps are (1) input of data and generating coordinates; (2) appending side chains from the AMBER94 library; (3) generating k-dimensional oriented polytopes (k-DOPs) which contain all rotamers; (4) generating an adjacency matrix of feasible interactions; (5) running Monte Carlo minimization using hard spheres potentials; and (6) outputting the results. Each of these steps will be described in detail below in context of the standalone SCD program.

The input file for SCD is a standard RSCB structural file with a `.pdb` extension. This structure is assembled into a PIPA Protein object, with a chain of residues assembled in a Peptide object array. Only the backbone atoms of the protein are retained. In order to append side chains onto the backbone, the locations of the β-carbon and α-hydrogen need to be chosen. They are placed in a peptide in tetrahedral positions with respect to the α-carbon, backbone nitrogen and carbon. The chirality is accounted for, and each residue is in the levorotary configuration. This is a standard in biological systems, and equivalent to the side chain being external

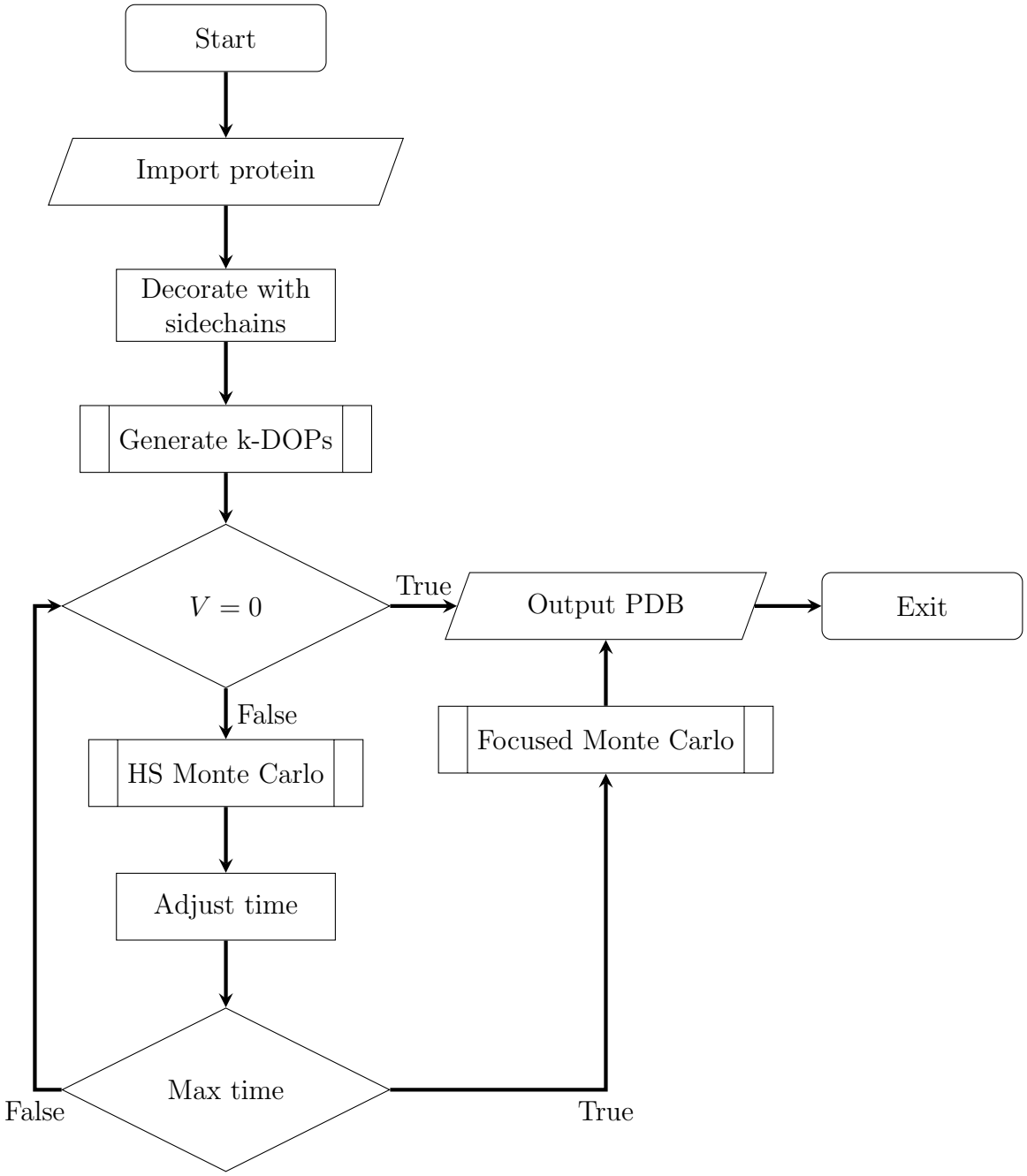


Figure E.1: Flowchart for the side chain decoration program.

to an α-helix. Once the location of the β-carbon has been specified, the rest of the sidechain can be populated from the geometry of fragments in a library. The AMBER94 library was chosen for convenience and formatting; however, other libraries could easily be implemented [217].

Once the geometry is loaded, the orientational degrees of freedom associated

with each side chains are cast in terms of the dihedral angles. Each residue has an available set of dihedral angles $(\chi_1 ... \chi_n)$ where n is dependent on the amino acid type, and each χ_i is defined by a unique set of four atoms. All of this amino acid specific information is contained in a separate library accessed by the program. A first approximation of the side-chain dihedral angles is given by a discrete set of available rotamers, $\chi_i \in (-60°, 180°, 60°)$. Some amino acids have more continuous rotameric degrees of freedom, and there are compiled statistical distributions that could be implemented to sample these states [448]. The program will sample different microstates by changing individual rotamers. To calculate hard spheres potentials between every sidechain in a new microstate would be an $\mathcal{O}(N^2)$ process, so a neighbor sensitive algorithm was developed using k-DOPs.

A polytope is a geometric object with flat sides, commonly used is graphics rendering and video game applications [569]. The prime utility of a k-dimensional oriented polytope is in testing for overlap between two shapes, and best demonstrated in the context of rectangles. A rectangle in $\mathbb{R}^2$ is type of 4-dimensional polytope, defined by the vectors to the edges of the shape. If the a rectangle is oriented along the cartesian axes, it can similarly be defined by the maximum and minimum values of two of its basis vectors (Fig. E.2). This latter definition is key for the simple condition of overlap. Two oriented rectangles will not overlap if for each basis vector, the minimum value of the first rectangle is greater than the maximum of the second or the maximum of the first is less than the minimum of the second (Fig. E.2). This rule works for any dimension of oriented polytope.

$$\text{No overlap iff: } (min_i^A > max_i^B \vee max_i^A < min_i^B) \tag{E.1}$$

Finally, these k-DOPs can be cast as bounding volumes for a quick estimation of overlap of complicated objects. It is computationally intensive to reorient the the bounding volume; however, it is a cheap to combine bounding volumes and compare bounding volumes as k-DOPs. Figure E.3 demonstrates how a k-DOP can be built

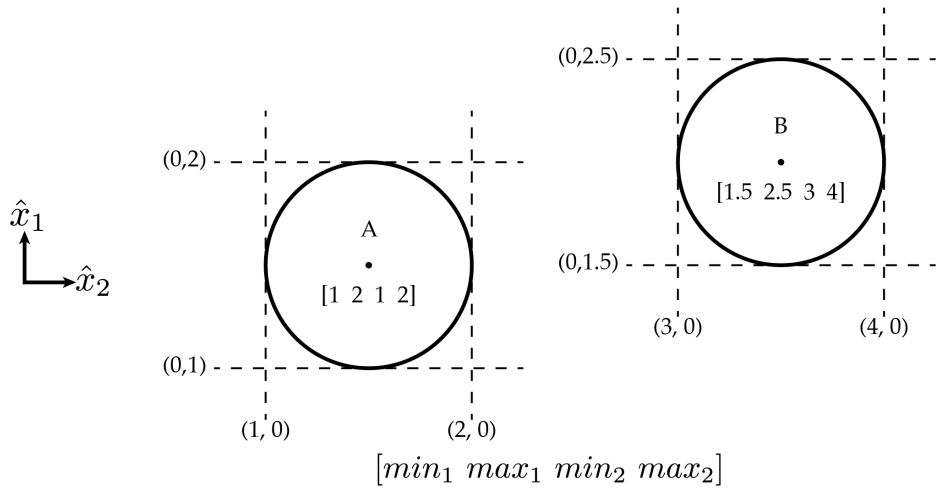


Figure E.2: A rectangle can be thought of as a 4-dimensional oriented polytope described by 2 basis vectors and respective minima and maxima.

around a side chain as a generous bounding volume. Each atom is bound by a k-DOP according to the van der Waals atomic radius, and these are combined to make one large k-DOP that bounds the residue. Higher dimensional k-DOPs will bound a volume more tightly (*i.e.* with less of the bounding volume unoccupied by the actual volume) [569]. We expand upon this by casting 14-DOPs around each side-chain rotamer, and combining the volumes to create a bounded region of rotameric freedom. From the 14-DOPs that describe the available space to be occupied by an amino acid residue, rapid comparison can be used to check for any potential overlap and build and adjacency matrix. This changes the best case complexity for calculation of energetics between side chains to $\mathcal{O}(N)$ — a significant improvement in the case of larger proteins.

With the record of potential conflicts stored in an adjacency matrix, a standard MCMC minimization algorithm can be used to find a ground state with 0 energy. The energetics are defined by hard spheres (eq. E.2), with radii given by Br_{cov}, where r_{cov} is the covalent radius of an atom and B is factor (default $B = 1.4$) to ensure the simulation does not underestimate steric clash.

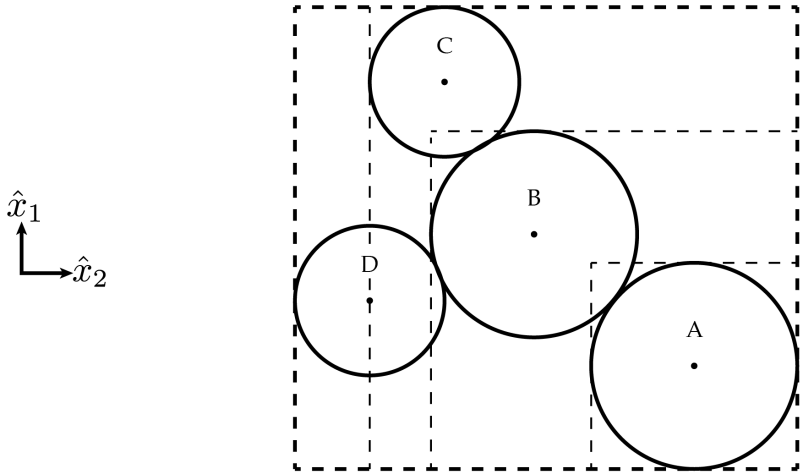


Figure E.3: Iterative building process of bounding volumes around a side chain. Shown with rectangular 4-DOPs for clarity.

$$V(\mathbf{r_1}, \mathbf{r_2}) = \begin{cases} 0, & \text{if } |\mathbf{r_1} - \mathbf{r_2}| \geq \sigma \\ 1, & \text{if } |\mathbf{r_1} - \mathbf{r_2}| < \sigma \end{cases} \tag{E.2}$$

The algorithm proceeds as a standard MC process, with simulated annealing. The starting temperature is automatically set such that $T_0 = 2V_0$ with an exponential temperature decay such that the 0 temperature reached in a set number of moves. If no minimum is reached, the set number of moves (starting at 100 MC sweeps) is increased by a factor of two (at maximum 10 times). Finally, in a last ditch effort, the program indexes any clashing residues and completes a short MC only sampling this subset. The program has been tested successfully on proteins with backbones of up to 1000 residues and is presented in its entirety in Appendix F. The results from SCD can be considered as accurate structures with low precision or high thermal parameters on the side chains and then fed to structural validation or further refinement in the structure prediction pipeline.

Appendix F

Mixed media

The following contents are provided on an attached compact disk:

- Source code
  - ZrW_2O_8 spaghetti generation
  - SqWWW
  - CRN_decoration
  - PIPA
  - SCD
- Sample input files
  - GULP
  - SqWWW
  - RMCProfile
  - PIPA
- Mixed media
  - mode_projection_animation.gif
  - PIPA_refinement_1PRB.avi
  - PIPA_refinement_1LQ7.avi