ICE FORMATION AND SOLVENT NANOCONFINEMENT

IN PROTEIN CRYSTALLOGRAPHY

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfilment of the Requirements for the Degree of

Doctor of Philosophy

by

David Wayne Moreau

May 2020

ICE FORMATION AND SOLVENT NANOCONFINEMENT

IN PROTEIN CRYSTALLOGRAPHY

David Wayne Moreau, Ph. D.

Cornell University 2020

X-ray crystallography is the predominant method for macromolecular structure determination. In this method, crystals are illuminated with intense X-ray radiation that damages the crystal. If crystals are cooled to cryogenic temperatures, the effects of radiation damage are greatly reduced. Since protein crystals are aqueous, cryocooling creates the possibility of crystalline ice formation. X-ray diffraction from ice is difficult to disentangle from the protein diffraction and if the ice is internal to the protein crystal, the expansion of solvent into ice degrades the crystal's order and subsequent diffraction quality. While there has been significant research into how to prevent ice formation, there has been very little research into the basic science of ice formation in macromolecular crystallography. The solvent within protein crystals is nanoconfined so the ice formation process should differ significantly from bulk-like solvent. Many open questions remain, such as what is a protein crystal's freezing point? Once the crystal has been cooled below this temperature, how long does it take for ice to form? When ice does form, what is its structure and how much solvent at the protein's surface is restricted from crystallizing?

We performed experiments studying the ice formation process inside of protein crystals. The results from these experiments showed that freezing points are depressed, the structure of ice is stacking disordered and two monolayers of water remain uncrystallized at the protein's surface. Completely new results from our experiments show protein crystals can remain supercooled for 10's of seconds suggesting a dramatic suppression of ice nucleation rates unobserved in any other material. A novel observation was

made while performing these experiments, following an expected initial contraction on cooling to temperatures as low as 220 K, apoferritin crystals expand to volumes as large or larger than their volumes at room-temperature. This serves as further evidence that the solvent inside of the protein crystals is in liquid form at temperatures above the protein-solvent glass transition. Using data deposited to the Protein Data Bank and the Integrated Resource for Reproducibility in Macromolecular Crystallography, we show that crystals with larger solvent cavities and percent solvent content are more susceptible to ice formation and the prevalence of hexagonal like ice has dramatically increased in the last 10 years.

BIOGRAPHICAL SKETCH

David Moreau was born in Concord, North Carolina in 1989. He completed an Associates in Science degree at Central Piedmont Community College and transferred to North Carolina State University where he graduated with a B.S. in Physics in 2013. That fall, he started graduate studies in Applied Physics at Cornell University.

ACKNOWLEDGMENTS

Critical to the completion of this Ph. D. were all the people along the way, the friends I made, the people I have worked with and the people who have been there for me. I would like to thank the Thorne group for the support I have received over the last 6 years in the group. Hakan Atakisi and I worked over 40 beam-times together at CHESS, and I could not have made it through all those 24-hour work periods alone. Jesse Hopkins did a great deal to get me going in my first year in the group. Rob Thorne has always been a source of ideas, guidance, critical appraisals and always knew how to push these projects over the finish line.

All synchrotron data presented in this dissertation was collected at Cornell High Energy Synchrotron Source. The staff at CHESS and MacCHESS have been more than generous in their support for our research and their efforts were critical for the completion of the research making up this dissertation. CHESS has just completed a comprehensive upgrade and is now the Center for High Energy X-ray Sciences (CHEXS), which is supported by the National Science Foundation under award DMR-1829070, and the Macromolecular Diffraction at CHESS (MacCHESS) facility, which is supported by award 1-P30-GM124166-01A1 from the National Institute of General Medical Sciences, National Institutes of Health, and by New York State's Empire State Development Corporation (NYSTAR).

I would like to acknowledge financial support from the Applied Physics department and the Physics department, the National Institutes of Health Molecular Biophysics Training Grant at Cornell (T32GM0082567), the National Institute of Health (R01-GM127528) and the National Science Foundation (MCB-1330685).

TABLE OF CONTENTS

1 INTRODUCTION

A living organism is comprised of many different classes of molecules, of which proteins comprise about 30-40% of the dry mass (Forbes *et al.*, 1953). Proteins are integral to the structure of the organism, transport nutrients, transmit information and catalyze reactions, and are major target of medicine and structural biology. Once a protein has been isolated and purified, many experimental approaches can be applied to understand its structure, function and dynamics. Determining the atomic structure of a protein provides an enormous amount of insight into the protein's function. The protein's structure can reveal how a ligand binds and can guide improvements to the ligand. If a mutation modifies the protein's function, the protein's structure can show if the mutation alters the active site, disrupts an allosteric network, or affects the protein's stability.

As of April 2020, 90% of all Protein Data Bank (PDB) depositions were determined by macromolecular X-ray crystallography (MX), making it the most common technique for macromolecular structure determination (https://www.rcsb.org/stats/summary). The largest obstacle in the X-ray crystallography pipeline is obtaining crystals. Crystals need to be well ordered, and for most of the history of protein crystallography they needed to be large enough to yield adequate diffraction data before radiation damage becomes severe (Holton & Frankel, 2010).

Methods are now available that allow structural study using crystals too small to produce a complete data set. Diffraction data with adequate signal-to-noise can be collected from tens to tens of thousands of crystals and merged to form a single dataset, such as in the serial "diffract and destroy" approach used at X-ray free electron lasers (Chapman, 2019) and in serial microcrystallography performed at macromolecular crystallography beamlines at synchrotrons (Martin-Garcia *et al.*, 2017; Owen *et al.*, 2017; Weinert *et al.*, 2017). Recent advances in cryo-electron microscopy (Cryo-EM) have allowed near-atomic-structures to be determined of large proteins and protein complexes using

enormous numbers of individual molecules, bypassing the crystallization process entirely (Glaeser, 2019). Cryo-EM is allowing structures of many targets that have proved intractable to crystallographic methods to be determined. The importance of cryo-EM and serial crystallography to the future of structural biology cannot be understated. However, if adequate size and quality crystals can be grown, single-crystal X-ray crystallography is a relatively easy and expedient method to obtain atomic resolution crystallographic data and is especially well suited to ligand/fragment screening in drug discovery.

1.1. Cryocrystallography

The most significant approach to reducing the crystal volume needed to collect full datasets from a single crystal is cryocooling to 100 K. At cryogenic temperatures, the diffusion of free radicals produced by radiation damage is prevented by the vitrification of the solvent and the crystal lattice is immobilized, preventing degradation of its order (Garman, 2010; Holton, 2009; Warkentin et al., 2013). At 100K, the radiation sensitivity of a protein crystal is reduced by approximately 50 times relative to 300 K, dramatically increasing the amount of information that can be extracted from a single crystal. The utility of cryocooling is not limited to radiation sensitivity. Protein crystals are fragile and are highly susceptible to damage from mishandling or dehydration. Cryocooling prevents crystal dehydration and mechanical damage, and a large crystallography infrastructure has been developed around crystals stored, transported, and studied at 100 K. As a result, over 95% of all X-ray crystallographically determined structures in the Protein Data Bank (PDB) have been performed at 100 K (Deduced from PDB survey in Chapter 5.3.5).

Protein crystals are approximately 50% solvent by volume (Weichenberger et al., 2015), and this introduces a major drawback to cryocooling, ice formation. Typically, protein crystals are surrounded in aqueous liquid and flash frozen in liquid nitrogen. In the ideal case, the solvent turns to a vitreous form. In other cases, some amount of the solvent transitions into crystalline ice. The powder diffraction rings

from this ice overlap the protein's Bragg peaks. As discussed in Chapter 5 and illustrated by Parkhurst et al., (2017), the presence of ice rings in the X-ray diffraction images results in incorrect estimates of the protein Bragg peak intensities in the resolution regions near the ice peaks. Thorn et al. (2017) developed an algorithm to detect this biasing of the Bragg peaks from a dataset's deposited structure factors. They applied this algorithm to the broad set of structure factors deposited in the Protein Data Bank (PDB) and showed that roughly 20% of all depositions show a detectable amount of ice contamination. While this is a considerable fraction of depositions, it likely underestimates the extent of ice contamination in protein crystallography. The expansion of water as it turns to ice also cause a degradation of the order of the crystal's lattice which, in the best case, reduces the maximum resolution obtained from the crystal. More likely, the crystal will not produce diffraction suitable for further analysis. The data sets deposited to the PDB represent the best datasets collected for a given target, so a survey of the PDB will not capture all of the crystals shipped to a synchrotron only to find that ice formation has destroyed the crystal.

The primary method to prevent ice formation is to soak cryoprotectant into the crystals (Garman & Schneider, 1997). Cryoprotectants are cosolutes added to aqueous solutions that affect the ice formation process by reducing the freezing point (thermodynamics) and nucleation rates (kinetics). Crystals can also be cryoprotected by removing the external solvent either by wicking (Pflugrath, 2015), or by replacement with an oil (Kwong & Liu, 1999; Riboldi-Tunnicliffe & Hilgenfeld, 1999). This approach relies on the crystal itself as a "cryoprotectant". The solvent within the crystals is nanoconfined to dimensions on the order on nanometers, which also reduces the freezing point (Findenegg et al., 2008) and nucleation rates (Li et al., 2013) of the solvent internal to the protein crystals.

1.2. Relevance of protein crystals beyond structural biology

Protein crystals are almost entirely known for their use in structural biology, but they have broad reaching, novel uses in science and engineering. In the study of confined water, the most commonly

used materials are nanoporous silica, MCM-41 and SBA-15 (Beck et al., 1992). A major drawback to these materials is that the broader structure housing the water filled pores has dimensions of 10's of nanometers. When performing the experiments, it is critical to ensure that the pores are filled with water without excess solvent existing outside of the pores. The linear dimensions of the pockets, channels and pores within protein crystals are on the order of nanometers, making them also systems of confined water (Juers & Ruffin, 2014). An advantage of using protein crystals to study confined water is their size and perfection. Single crystals can be routinely grown to dimensions of > 100 µm. Their hydration is complete, and the removal of external solvent can be visually verified using a stereoscope. Their macroscopic size allows for macroscopic measurements, such as direct measurements of density, optical properties and single crystal diffraction, that would be considerably more difficult, if not impossible, on the sub-micrometer sized samples that can be produced of most other nanoporous materials.

The phase transition of water to ice in confinement has been well studied, but there are still open questions. How ice forms in a biologically relevant form of confinement has received limited study. How the type of surface affects the ice formation process is typically studied by using surface modifications to mesoporous silica, and this can include modification using organic molecules. How ice forms in confinement by a soft material held together by non-covalent interactions is relatively unexplored. Protein crystals provide a highly ordered platform for experimenting on the ice formation process under nanoconfinement in soft materials.

1.3. Summary of research

A prerequisite for studying the process of ice formation within protein crystals is developing an understanding of the nature of the solvent within the protein crystals. A protein affects the composition (Timasheff, 2002), density (Svergun et al., 1998; Merzel & Smith, 2002) and dynamics (Laage et al., 2017)

of the solvent near its surface. The maximum length scale of these perturbations tends to be on the length scale of two diameters of a water molecule or 5.6 Å. The waters within this range of the protein surface are typically referred to as hydration layers. The solvent within the protein crystal is confined to nanoscale dimensions and the relative proportion of solvent in the protein's hydration layers to the bulk like solvent becomes significant. The implication is the solvent internal to the crystals is heavily influenced by the protein's surface. While this disordered solvent region comprises roughly 50% of a crystal's volume, very little attention is paid to it. Chapter 2 focuses on the properties of solvent within protein crystals, specifically on how its composition is altered and if there are observable changes to its density when it incorporates into the crystal from bulk solvent. The alteration of the solvent's composition near the protein's surface is known as the "preferential interaction" and has been extensively studied for proteins in solution (Timasheff, 2002). We demonstrate a new technique that measures this effect in protein crystals and make comparisons to the solution-based experiments. Our measurements are not the first experimental observation of compositional changes to a protein crystal's internal solvent. We aggregate these other results, interpret them in terms of the preferential interaction, and make comparisons to solution-based experiments. These results have practical implications for molecular dynamics (MD) simulations of protein crystals. These simulations are performed in environments of constant number of molecules, volume, and temperature, so the appropriate number of molecules in the crystal's unit cell must be specified before performing the simulations (Janowski et al., 2016). Our comparisons between solution and crystal-based measurements provide insights that could be used to constrain these simulations.

The science of ice formation within protein crystals is detailed in Chapter 3. We perform experiments on protein crystals cooled to temperatures between 180 and 260 K, and demonstrate that the ice formation process within protein crystals has striking similarities to that observed in other confining materials and geometries, while allowing new kinds of measurements and/or estimation of key

parameters not possible in these other systems. Nanoconfinement within the protein structure significantly depresses the freezing / melting point of water and reduces the ice nucleation rate. Ice that forms is heavily stacking disordered. Water within two monolayers (≈ 6 Å) of the protein surface or confined in pores with dimensions less than 20 Å does not freeze.

A novel observation was made while performing the experiments described in Chapter 3. Upon cooling to temperatures between 220 and 260 K, apoferritin crystals initially contract and become disordered. Then, on the timescale of seconds, the crystals expanded and regained their order. Chapter 4 documents this observation, and provides a detailed explanation for it, that has implications for applications of variable temperature crystallography.

Chapter 5 brings the dissertation back to ice in protein crystallography. Use a large amount of data, including our in-house data, Protein Data Bank depositions (PDB) and Integrated Resource for Reproducibility in Macromolecular Crystallography (IRRMC) depositions, we discuss the various types of ice observed in protein crystallography and why they are observed. We build off a recently developed algorithm that detects the influence of ice diffraction in structure factors deposited in the PDB (Thorn et al., 2017) and look at the relative prevalence of the different forms of ice in PDB depositions. We simulate the effects of ice diffraction in crystallographic data to examine how ice affects coordinate models and electron density maps.

1.4. References

Beck, J. S., Vartuli, J. C., Roth, W. J., Leonowicz, M. E., Kresge, C. T., Schmitt, K. D., Chu, C. T. W., Olson, D. H., Sheppard, E. W., McCullen, S. B., Higgins, J. B. & Schlenker, J. L. (1992). J. Am. Chem. Soc. **114**, 10834–10843.

Chapman, H. N. (2019). Annu. Rev. Biochem. **88**, 35–58.

Findenegg, G. H., Jähnert, S., Akcakayiran, D. & Schreiber, A. (2008). ChemPhysChem. **9**, 2651–2659.

Forbes, R. M., Cooper, A. R. & Mitchell, H. H. (1953). J. Biol. Chem. **203**, 359–366.

Garman, E. F. (2010). Acta Cryst. **D66**, 339–351.

Garman, E. F. & Schneider, T. R. (1997). J. Appl. Cryst. **30**, 211–237.

Glaeser, R. M. (2019). Annu. Rev. Biophys. **48**, 45–61.

Holton, J. M. (2009). J. Synchrotron Rad. **16**, 133–142.

Holton, J. M. & Frankel, K. A. (2010). Acta Cryst. **D66**, 393–408.

Janowski, P. A., Liu, C., Deckman, J. & Case, D. A. (2016). Protein Sci. **25**, 87–102.

Juers, D. H. & Ruffin, J. (2014). J. Appl. Cryst. **47**, 2105–2108.

Kwong, P. D. & Liu, Y. (1999). J. Appl. Cryst. **32**, 102–105.

Laage, D., Elsaesser, T. & Hynes, J. T. (2017). Struct. Dyn. **4**, 044018.

Li, T., Donadio, D. & Galli, G. (2013). Nat. Commun. **4**, 1–6.

Martin-Garcia, J. M., Conrad, C. E., Nelson, G., Stander, N., Zatsepin, N. A., Zook, J., Zhu, L., Geiger, J., Chun, E., Kissick, D., Hilgart, M. C., Ogata, C., Ishchenko, A., Nagaratnam, N., Roy-Chowdhury, S., Coe, J., Subramanian, G., Schaffer, A., James, D., Ketwala, G., Venugopalan, N., Xu, S., Corcoran, S., Ferguson, D., Weierstall, U., Spence, J. C. H., Cherezov, V., Fromme, P., Fischetti, R. F. & Liu, W. (2017). IUCrJ. **4**, 439–454.

Merzel, F. & Smith, J. C. (2002). Proc. Natl. Acad. Sci. USA. **99**, 5378–5383.

Owen, R. L., Axford, D., Sherrell, D. A., Kuo, A., Ernst, O. P., Schulz, E. C., Miller, R. J. D. & Mueller-Werkmeister, H. M. (2017). Acta Cryst. **D73**, 373–378.

Parkhurst, J. M., Thorn, A., Vollmar, M., Winter, G., Waterman, D. G., Fuentes-Montero, L., Gildea, R. J., Murshudov, G. N. & Evans, G. (2017). IUCrJ. **4**, 626–638.

Pflugrath, J. W. (2015). Acta Cryst. **F71**, 622–642.

Riboldi-Tunnicliffe, A. & Hilgenfeld, R. (1999). J. Appl. Cryst. **32**, 1003–1005.

Svergun, D. I., Richard, S., Koch, M. H., Sayers, Z., Kuprin, S. & Zaccai, G. (1998). Proc. Natl. Acad. Sci. USA. **95**, 2267–2272.

Thorn, A., Parkhurst, J., Emsley, P., Nicholls, R. A., Vollmar, M., Evans, G. & Murshudov, G. N. (2017). Acta Cryst. **D73**, 729–737.

Timasheff, S. N. (2002). Biochemistry. **41**, 13473–13482.

Warkentin, M., Hopkins, J. B., Badeau, R., Mulichak, A. M., Keefe, L. J. & Thorne, R. E. (2013). J. Synchrotron Rad. **20**, 7–13.

Weichenberger, C. X., Afonine, P. V., Kantardjieff, K. & Rupp, B. (2015). Acta Cryst. **D71**, 1023–1038.

Weinert, T., Olieric, N., Cheng, R., Brünle, S., James, D., Ozerov, D., Gashi, D., Vera, L., Marsh, M., Jaeger, K., Dworkowski, F., Panepucci, E., Basu, S., Skopintsev, P., Doré, A. S., Geng, T., Cooke, R. M., Liang,

M., Prota, A. E., Panneels, V., Nogly, P., Ermler, U., Schertler, G., Hennig, M., Steinmetz, M. O., Wang, M. & Standfuss, J. (2017). Nat. Commun. **8**, 542.

2 THE COMPOSITION AND NATURE OF SOLVENT INTERNAL TO PROTEIN CRYSTALS

Abstract Experimental and computation results have shown that proteins alter the composition, structure, and dynamics of the solvent adjacent to their surface, the protein's hydration layer. A large fraction of solvent found within protein crystals is in this hydration layer, magnifying the effects. Relatively little is known about the exact nature of the solvent within the protein crystals and basic insights to its composition or structure are not readily accessible from crystallographic data. This results in uncertainty over whether there may be differences in the perturbation to the protein's hydration layer within the protein crystals. The estimation of the electron density or composition of the solvent within protein crystals is quantitatively important for determining and accessing electron density maps and in simulations of protein and solvent dynamics in protein crystals. We demonstrate new approaches to protein crystal density measurements and chemical analysis. Our results, along with the results aggregated from past research, suggests that while the properties of the solvent within the protein crystals are perturbed away from the solvent the crystal is soaked or grown in, the effects in the crystals are quantitatively similar to effects observed in solution.

2.1 INTRODUCTION

The result of an X-ray protein crystallography experiment is an electron density map that is modelled and interpreted with an atomic structure of the protein. While this structure is rightfully the focus of the experimental results, the average protein crystal's volume is only 50% protein, the other half is solvent (Weichenberger & Rupp, 2014). Weichenberger *et al* (2015) recently provided a thorough review of the modelling of a protein crystal's solvent region. Very little is concluded about the solvent region of the protein crystal, but it is modelled in two ways. Near the protein's surface, usually within one or two hydration layers, peaks of electron density are observed and are modelled as ordered water molecules (Nittinger *et al.*, 2015) or other ions / molecules present in the "mother liquor", the solution the crystals are soaked or grown in. Further from the protein's surface, the rest of this solvent is disordered and is modelled as a region of constant electron density (Fokine & Urzhumtsev, 2002). The relative lack of information regarding this region prevents the determination of its composition or density.

This solvent could be assumed to be identical to the "mother liquor" but this fails to account for the fact that solvent is confined within and strongly interacts with a nanoporous protein network. The internal solvent is restricted to irregular pockets, channels and cavities with lateral dimensions of 10 to 30 Å that correlates with unit cell volume and solvent content (Juers & Ruffin, 2014; Moreau *et al.*, 2019). Protein surfaces, like most surfaces, interact strongly with solvent, modifying the solvent's composition, structure, and dynamics in their immediate vicinity. The alteration of the solvent's composition is an effect known as the preferential interaction (Timasheff, 2002). Experiments measuring the preferential interaction have shown that proteins either exclude or include cosolutes in their hydration layer due to the relative preference of interaction between water or the cosolutes with the surface of the protein. Many of the common crystallization salts and cryoprotectants used in crystallography are excluded from the protein's hydration layer. A few molecules, such as urea, are preferentially included in the hydration layer (Timasheff & Xie, 2003). Fig. 2.1 shows the preferential interaction coefficient for several

cryoprotectants measured using lysozyme (Timasheff & Xie, 2003; Arakawa & Timasheff, 1983, 1985; Bhat & Timasheff, 1992; Lee & Lee, 1981; Arakawa & Timasheff, 1982; Gekko, 1982; Arakawa & Timasheff, 1984; Arakawa *et al.*, 1990). Negative / positive values imply the cosolute is excluded / included from the protein's hydration shell, Section 2.7 gives further discussion of the preferential interaction coefficient. Data for bovine serum albumin and RNASE are qualitatively similar and are graphed in Fig. 6.1.1 with associated references in the supplementary materials.

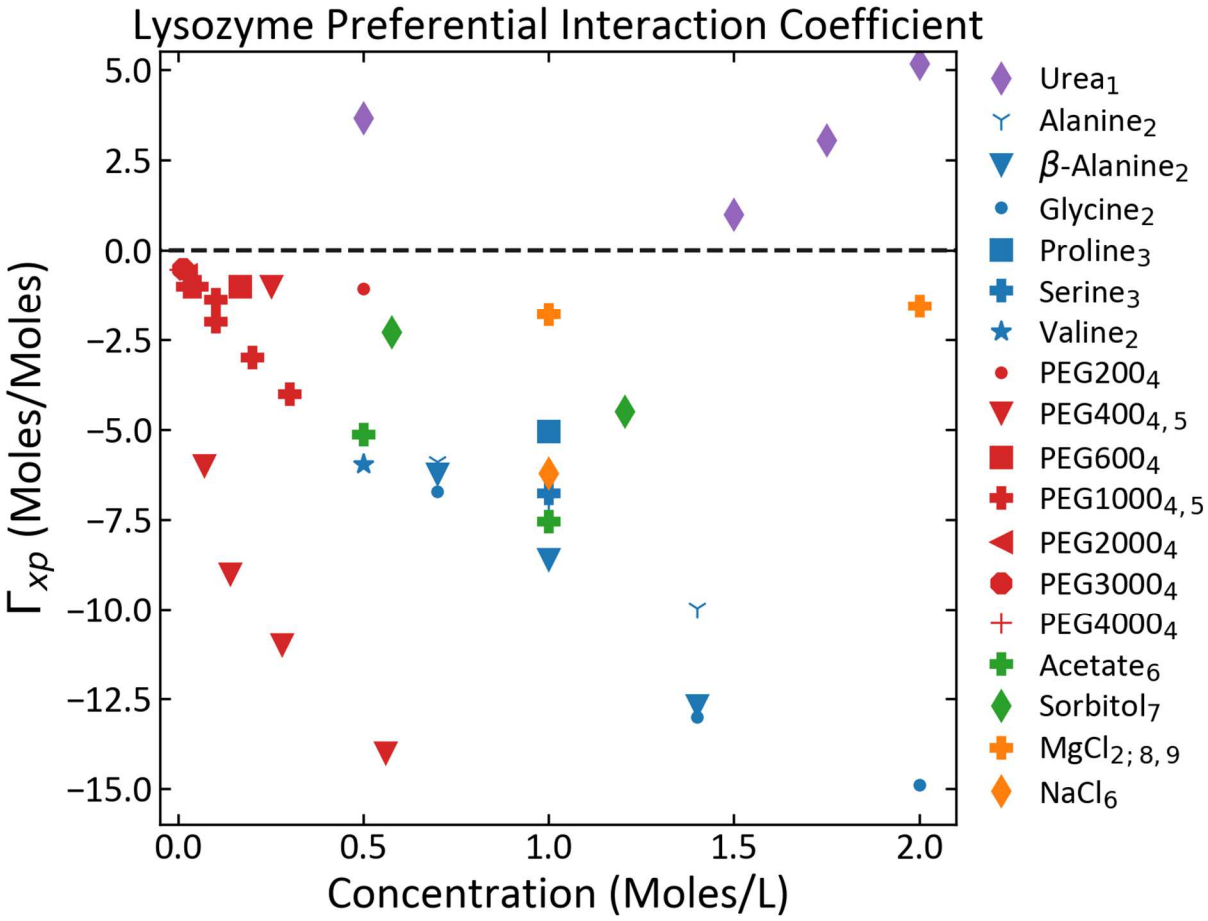


Figure 2.1 Preferential interaction coefficients measured on lysozyme in solution. Negative values indicate that the cosolute is excluded from the protein's surface, positive values indicate inclusion. Corresponding references are (1) Timasheff & Xie, 2003; (2) Arakawa & Timasheff, 1983; (3) Arakawa & Timasheff, 1983; (4) Bhat & Timasheff, 1992; (5) Lee & Lee, 1981; (6) Arakawa & Timasheff, 1982; (7) Gekko, 1982; (8) Arakawa & Timasheff, 1984; (9) Arakawa et al., 1990.

The preferential interaction must persist when proteins associate to form crystals, but does this effect differ in crystals from that of proteins in bulk solution? Cosolutes larger than the protein crystal's solvent channels, such as impurity proteins, are completely excluded from the crystal (McMeekin *et al.*, 1950; Low *et al.*, 1956; Vekilov *et al.*, 1996) and this is the basis for purification by recrystallization. Observations of the preferential interaction in crystals have been made. Schoenborn (1988) combined density measurements and neutron diffraction of myoglobin crystals to estimate the internal solvent to be 13% w/w ammonium sulfate compared to the 40% w/w solution that the crystals were grown in. Soares & Caspar (Soares & Caspar, 2017) took X-ray diffraction data of insulin crystals as grown and soaked in 41% w/v trehalose and used a novel method to determine the crystal's bulk-solvent electron density. The difference between the electron density of the two crystal preparations suggests a 10% w/v internal trehalose concentration. The cosolute concentration within protein crystals has also been deduced from measurements of crystal densities. Low *et al.* (1956) measured densities of several proteins in various polysaccharides and estimated the density of solvent within the crystal based on known unit cell properties. They then deduced the concentration of polysaccharide within the crystal's solvent channels to account for the crystal density. In all cases the concentration within the crystal was lower than the concentration that the crystals were soaked in. McMeekin *et al.* (1950) performed a similar experiments and showed that sucrose was found in lower concentrations within β-lactoglobulin crystals while the protein serum albumin was completely excluded from the crystals.

In solution-based measurements of the preferential interaction, NaCl ions is found to be excluded from a protein's hydration shell (Arakawa & Timasheff, 1982). Elgersma *et al.* (1992) and Vekilov *et al.* (1996) both observed an increase in the number of salt ions within the crystals. Elgersma performed batch crystallization experiments of lysozyme and monitored the Cl^- and protein concentration in the crystallization solution over time and explained the uptake of Cl^- ions in the crystal as a result of the lower dielectric constant of the solvent within the crystal. Vekilov performed similar experiments but also

measure Na^+ concentration. They varied the initial supersaturation of lysozyme and used protein stock with different amounts of macromolecular impurities. Their results suggested that crystals with larger amounts of defects incorporate more Na^+ and Cl^-, implicating defects in the crystal as the source of ion uptake. Palmer *et al.* (1948) air dried tetragonal lysozyme and measured mass lost during drying (assumed to be water) and the relative dried weight of Na^+, Cl^- and protein. From these measurements, they showed that there was an excess of Cl^- ions, relative to Na^+, left bound to dried proteins.

The water within a protein's hydration shell is thought to denser than the bulk-like water. This has been experimentally suggested through SAXS/SANS measurements (Svergun *et al.*, 1998; Ortore *et al.*, 2009; Kim *et al.*, 2016). For three separate proteins, Svergun used a three-term model for SAXS and SANS data including the protein's crystallographic model, excluded volume and a term representing a 3 Å thick hydration layer 2 Å from the protein's surface. The excluded volume per protein atom and the hydration layer's contrast with the bulk solvent were the only terms varied. The optimized contrast indicated a 10% denser hydration layer for lysozyme. The preferential interaction should affect their estimated hydration layer contrast, however with mM buffer concentrations, the effect would be insignificant. Merzel & Smith (2002) supported their results for lysozyme with MD simulations that demonstrated the perturbation of the hydration water by the protein's surface increased the density by 5% and attributed this to shortening of the average distance between waters and an "increase in coordination number". The capacity for MD simulations has dramatically improved since 2002, but recent MD simulations have suggested increased density of the first hydration layer to a similar degree (Persson *et al.*, 2018).

Molecular dynamic (MD) simulations provide atomic level details of a protein's structure and dynamics that can be inaccessible to experimental techniques and are frequently used to model and interpret experimental results. Within protein crystallography (see review Cerutti & Case 2019), MD simulations provide unprecedented insights into the hidden dynamics (Janowski *et al.*, 2013), solvation structure (Wall *et al.*, 2019) and diffuse scatter (Wall, 2018; Meisburger *et al.*, 2020). These simulations

are performed in a simulation box with fixed volume, temperature, and number of molecules (NVT). When setting up a MD simulation of a protein crystal, the number of cosolute and water molecules must be specified. Requiring the pressure of the NVT simulation to equal atmospheric pressure is used as a constraint to determine the appropriate number of water molecules to include (Wall *et al.*, 2014).

What is lacking is experimental data to validate these initialization methods. The large data on the preferential interaction that could be used for this purpose has several deficiencies. These experiments were mostly performed by measuring the density or viscosity difference between a solution of cosolutes with and without protein present. These experiments cannot distinguish between the behaviors of the individual components of the solutions, so experiments that show that NaCl is excluded from the protein's hydration layer cannot distinguish separate the behaviors of the Na^+ and Cl^- ions or remove any contribution from the behavior of water molecules (Arakawa & Timasheff, 1982). Initializing these simulations need the information for each separate molecular species, which the density or viscosity-based measurements of the preferential interaction are unable to provide.

Direct measurements of the density and cosolute concentration of a crystal's internal solvent can be made with techniques accessible to a crystallography laboratory. Historically, protein crystal density measurements have been a routine technique used to estimate the number of molecules in the crystal's unit cell (Matthews, 1968). These measurements were performed by measuring the point of neutral buoyancy of the crystals in a gradient tube (Low & Richards, 1952) of various solution mixtures. Some common solutions include water saturated iodobenzene, carbon tetrachloride or xylene (Colman & Matthews, 1971). These chemicals are known to be dangerous (carbon tetrachloride is toxic to the liver and kidneys and xylene is highly flammable). Many of the solutions used in these crystal density measurements have some amount of water solubility that could be pulling water out of the crystals (Low *et al.*, 1956). Some of these solutions have also been observed to have deteriorating effects on the protein crystals (Westbrook, 1985). A new method for measuring the density of a protein crystal is demonstrated

using mixtures of non-hazardous, water insoluble oils: silicon fluid and 1-(perfluorohexyl)octane (1PO). 1PO is a fluorinated hydrocarbon oil commonly used in ophthalmology procedures and has a density much larger than typical oils (Zeana *et al.*, 1999).

Here, we demonstrate the power of direct chemical analysis and density measurements with protein crystals to determine the composition of the solvent within a protein crystal. Cosolute concentrations are estimated by dissolving crystals in water and measuring protein and cosolute the concentrations in this sample. The protein concentration of this sample can be estimated from the absorption of light at 280 nm using a spectrophotometer. The concentration of the cosolutes can be determined with multiple different techniques. Glycerol is measured with an enzyme-based assay, Na^+ with inductively coupled plasma mass spectrometry (ICP-MS) and Cl^- with ion chromatography. The relative concentrations of the cosolutes to the protein, along information about the crystal's unit cell, can be used to determine the concentration of the cosolutes within the crystal.

2.2 METHODS

2.2.1 Crystal growth

Purified lysozyme powder (Sigma L6876) was used as a starting point for lysozyme crystallizations. Tetragonal lysozyme crystals were grown at room temperature in hanging drops comprised of equal volumes of 80 mg/ml protein in 0.1 M sodium acetate buffer at pH 5.2 and a well solution of 2.0 – 2.5% w/v of NaCl in the same buffer. The pH was chosen to maximize crystal size (Judge *et al.*, 1999). Tetragonal crystals in the space group $P4_32_12_1$ grew to dimensions of 300-800 µm. Crystal appeared within one week and stopped growing within four weeks. The crystals have a Matthew's coefficient derived solvent content of 42% and a maximum size of the solvent cavity, determined using the program MAP-CHANNELS (Juers & Ruffin, 2014) is 13 Å.

Monoclinic lysozyme crystals were grown at room temperature in hanging drops comprised of equal volumes of 10 – 15 mg/ml protein in 0.1 M sodium acetate buffer at pH 5.2 and a well solution of 3.5% w/v of $NaNO_3$ in the same buffer. Monoclinic crystals in the space group $P12_11$ grew to dimensions of 100 - 300 µm. The crystals have a Matthew's coefficient derived solvent content of 34% and a maximum size of the solvent cavity is 10 Å.

Cubic crystals of equine spleen apoferritin (Sigma A-3641) were grown by the vapor diffusion method. Drops comprised of 4 µL of 10 mg /mL protein in 0.1M sodium acetate buffer at pH 6.5 and 4 µL of a well solution of 2% w/v $CdSO_4$ and 15% w/v (~1.1 M) $(NH_4)_2SO_4$ in the same buffer were equilibrated with well solution. Cubic crystals in space group F432 having room-temperature unit cell dimensions a = b = c = 183.5 Å and with sizes of 300-500 µm were obtained within a week. The crystals have a Matthews coefficient derived solvent content of 63% and a maximum size of the solvent cavity is 68 Å.

2.2.2 Soak solution preparation

Soak solutions along a range of cosolute concentrations were made for lysozyme and apoferritin. For lysozyme, glycerol and NaCl were varied and for apoferritin, glycerol and $(NH_4)_2SO_4$ were varied.

For tetragonal lysozyme, soak solutions with variable amounts of glycerol were made from 0.1M sodium acetate buffer at pH 5.2. The solutions were prepared by adding this buffer to a fixed amount of NaCl and glycerol so each solution will have the same amount of NaCl regardless of glycerol concentration, as opposed to mixing glycerol with a 3.5% NaCl buffer solution. The glycerol concentrations included were 0%, 10%, 20%, and 30% w/v. A second set of soak solutions were prepared with a variable amount of salt and a 0.01 M sodium acetate buffer at pH 5.2. The NaCl concentrations included 0.26M, 1M, 2M, 3M, 4M, and 5M.

For monoclinic lysozyme, a single soak solution was made comprised of 2% w/v $NaNO_3$ in 0.1 M sodium acetate buffer at pH 5.2.

For apoferritin, soak solutions with variable amounts of glycerol were made from 0.1M sodium acetate buffer at pH 6.8. The glycerol concentrations included were 0%, 20%, 40% and 60% w/v with 10% $(NH_4)_2SO_4$ and 2% $CdSO_4$.

The soak solutions were characterized by measuring their densities and glycerol or NaCl concentrations, with methods described in Sec. 2.2.4. The densities and cosolute concentrations are shown in Table S1.

Solution densities were estimated by measuring the buoyant force applied to a copper weight of volume V_{weight} suspended from a microbalance (Mettler: AE240) using a 25.4 µm thick Dyneema line (Berkley NanoFil: 1 lb, 0.001"). The weight's measured mass in air was set to zero and the weight was then placed in a solution, the change to the measured mass, Δm , is related to the density of the solution by

$$\rho_{solution} = \rho_{air} + \frac{\Delta m}{V_{weight}}$$

The density of air at normal temperature and pressure $\rho_{air} = 0.0012 \pm 0.0001\,\mathrm{g/cm^3}$ (Davis, 1992). The volume of the mass was calibrated by measuring the change in mass resulting from placing the mass in methanol (0.792 g/cm^3), propylene glycol (1.036 g/cm^3), ethylene glycol (1.1132 g/cm^3), and methylene chloride (1.3266 g/cm^3). The volume of the mass was then optimized to minimize the difference between the known and estimated densities and found to be $0.3096 \pm 0.0001\ cm^3$ (Fig. 6.1.2). The volume of the line in the solution is estimated to be $5 \cdot 10^{-6}$ cm^3 and is neglected.

2.2.3 Crystal density measurements

The densities of crystals were measured by finding their point of neutral buoyancy similar to traditional crystal density measurements in a gradient tube (Low & Richards, 1952). The method was to take two oils with different densities and make a series of mixtures with different concentrations, or equivalently, densities. Upper and lower bounds on their densities were set by visual observation of

whether they sunk or floated. The two oils used were silicon fluid (Thomas Scientific: SF96/50), a polydimethylsiloxane oil, and 1-(perfluorohexyl)octane (1PO) (Fluorxy Labs: FC12-T6Octane). 1PO is a fluorinated hydrocarbon oil commonly used in ophthalmology procedures and has a density much larger than typical oils (Zeana *et al.*, 1999). The density of the silicon fluid and 1PO were measured to be 0.9630 ± 0.0004 g/cm^3 and 1.3376 ± 0.0005 g/cm^3 respectively. Fig. 6.1.3 shows the densities of 0, 20.25, 40.78, 59.13, 64.93, 74.78, 79.58, and 100 %w/w 1PO in silicon fluid and a calibration curve fit to the densities.

Before measuring the crystal's density, crystals were soaked in a specific soak solution. For solutions with similar compositions to the mother liquor, the crystals could be transferred directly to the solution, otherwise, careful soaking was necessary to prevent the crystals from cracking. For these solutions, the hanging drop slide was place in a gas stream with the drop facing upwards. The humidity of the gas was measured to be approximately 95% and was used to slow any dehydration. Solution was gradually added to the hanging drop and left to equilibrate for a few minutes, the crystals were soaked in a series of solutions with increasing concentration to reach the target concentration. To compensate for any change in the soak solution when equilibrating with the crystal, crystals were transferred between several drops of the target concentration and left in each for several minutes. The oil that the crystal densities were to be measured in was pipetted onto a 12 mm circular coverslip and the crystal was stirred in the oil to clear the external solvent (Warkentin & Thorne, 2009). The crystal was then placed in 40 µL oil in a 96 well plate (Mitegen: In Situ-1 Crystallization Plate). Crystals that floated could be floating because of external solvent present on their surface. Floating crystals were rotated to view all sides under high magnification. If external solvent was observed, the crystal was removed from the well and further cleaned, otherwise it was recorded as floating. At each concentration of the oil, 2 to 5 crystals were tested. Fig. 6.1.4 demonstrates theses sink-float measurements.

2.2.4 Cosolute concentration measurements

Crystals cleared of their external solvent were dissolved into deionized water in centrifuge tubes and the relative concentrations of protein and cosolutes was used to estimate the concentration of cosolutes inside of the crystal. Typically, 10 crystals with linear dimensions of 500 µm were needed per trial to achieve concentrations high enough for accurate measurements. Crystals were either taken from the density measurements or crystals were identically prepared, except their external solvent was cleaned in NVH oil (Warkentin & Thorne, 2009).

Glycerol concentrations in the centrifuge tube were measured using the Megazyme glycerol assay. This assay measures the concentration of glycerol by a series of two enzymatic reactions that results in the conversion of NADH to NAD^+ that produces a measurable reduction in the absorption of light at 340 nm. These measurements were performed in 1 cm pathlength cuvettes (Brand UV cuvette: Sigma z637157) with a Spectronic Genesys 5 spectrophotometer. Samples were submitted to the Analytical and Technical Services at SUNY College of Environmental Science and Forestry for Na^+ measurements by ICP-MS. Samples were submitted to the Center for Environmental Systems Engineering at Syracuse University for Cl^- measurements by ion chromatography.

Protein concentrations were measured by uv absorption in 1 cm pathlength cuvettes (Brand UV cuvette: Sigma z637157) with a Spectronic Genesys 5 spectrophotometer. Extinction coefficient of 2.64 and 0.982 ml⁄(mg·cm) were used for lysozyme (Aune 1969) and apoferritin (Bryce & Crichton 1973) respectively.

2.2.5 Crystallographic data collection and processing

X-ray data was collected on station F1 at the Cornell High-Energy Synchrotron Source (CHESS) using a Pilatus 6M detector. Each crystal was placed in the X-ray beam at room temperature in a Mitegen RT tube in equilibrium with their soak solution. For tetragonal lysozyme, three datasets were collected at

different locations on the crystal. For apoferritin, one data set was recorded for each crystal. The diffraction images were indexed and integrated using DIALS (Winter *et al.*, 2018). Unit cell volumes for the crystals soaked in the different soak solutions are shown in Fig. 6.1.5.

2.2.6 Internal solvent density

The density of the solvent within a crystal, ρ_{is}, can be estimated from the crystal's measured density, $\rho_{crystal}$, by subtracting the known density of the protein in the crystal and correcting for the solvent fraction:

$$\rho_{is} = \frac{\rho_{crystal} - z \cdot MW \big/ Na \cdot V_{uc}}{f_s} \tag{1}$$

Here, z is the number of protein monomers of molecular weight MW in the crystal's unit cell of volume V_{uc} with a solvent fraction f_s; Na is Avogadro's number. The solvent fraction of a protein crystal determined by the Matthews coefficient is

$$f_s^m = 1 - \frac{1.23}{V_m}$$

Where the Matthews coefficient is defined as

$$V_m = \frac{V_{uc}}{z \cdot MW}.$$

The prefactor 1.23 is a result of dividing the average partial specific volume of a protein, $\bar{v} \approx 0.74\,\mathrm{cm^3/g}$ by Avogadro's number (Na) (Matthews, 1968). Proteins smaller than 20 kDa, such as lysozyme, tend to be more compact than this average (Fischer *et al.*, 2004) due to less internal voids (Liang & Dill, 2001). A simple modification to the solvent fraction determined through this method is to use an experimentally determined partial specific volume of the protein, v

$$f_s^I = 1 - \frac{v \cdot z \cdot MW}{Na \cdot V_{uc}}$$

Experimentally measured values of lysozyme's partial specific volume reported in the literature were aggregated from 16 sources and averaged to give an estimate of lysozyme's specific volume of 0.72 ± 0.02 cm^3/g, the individual values and references are in Table 6.1.2. Apoferritin's partial specific volume has been experimentally determined to be 0.732 ± 0.016 (Ghirlando *et al.*, 2015). Most of these measurements were made using density measurements where the partial specific volume is deduced from the difference in density of a solution with and without added protein. Physically, the partial specific volume of a protein is the change in a solutions volume when the protein is added. This includes the primary contribution from the protein's excluded volume and additional contributions if the protein's hydration water's density is modified. Therefore, results obtained using this partial specific volume should show the internal and external solvent's density are the same. The protein's volume could be deduced from crystal structures, but the available methods are not accurate enough for these calculations (Section S1)

Additional corrections need to be made to protein volume due to the changes in the solvent excluded volume (SEV) due to crystal contacts. The SEV in the unit cell, calculated from atomic coordinates, differs if it is calculated from a monomer and multiplied by the number of monomers in the unit cell or it is calculated using the entire unit cell. Using atomic coordinates, the SEV volume of a monomer, $V_{monomer}^{SEV}$, and within the unit cell, V_{uc}^{SEV}, was calculated by a custom ball rolling program described in Supplementary Section 5.1. The ratio of the SEV within the unit cell calculated by these two methods,

$$r = \frac{V_{uc}^{SEV}}{z \cdot V_{monomer}^{SEV}}.$$

This ratio was calculated to be 1.018 and 1.023 for tetragonal lysozyme and monoclinic lysozyme, respectively. Since the partial specific volume of apoferritin was determined from the 24mer, the relevant ratio is the ratio of the SEV of the spherical shell to the unit cell volume. This was determined to be 1.0. The protein's partial specific volume was scaled by this ratio to correct for the effective increase in the protein's SEV due to crystal packing

$$f_s^I = 1 - \frac{r \cdot v \cdot z \cdot MW}{Na \cdot V_{uc}}. \quad (2)$$

2.2.7 Relation between protein and cosolute concentrations

The effect that the protein's surface has on the local concentration of a cosolute can be quantified by the Kirkwood-Buff formulation (Pierce *et al.*, 2008). The relevant quantity here is the excess solvation number, N_{XP}, where the subscript X and P denote the specific cosolute and the protein respectively. The excess solvation number represents the number of cosolutes excluded or included from the region surrounding the protein's surface. It is positive or negative depending on if there is a net inclusion or exclusion of the cosolute. This number comes about by integrating the difference between the cosolute concentration a distance r from the protein's surface, $c_x(r)$ with the concentration infinitely far away from the protein's surface, c_x^o:

$$N_{XP} = \int c_x(r) - c_x^o dV. \quad (3)$$

The excess solvation number can also be arrived at through thermodynamic treatments (Timasheff, 2002). The preferential interaction parameter, Γ_{XP}, estimates these effects, but does not disentangle them from the exclusion or inclusion of water from the hydration layer

$$\Gamma_{XP} = N_{XP} - \frac{c_X^o}{c_W^o} N_{WP}.$$

Here, c_W^o is the concentration of water infinitely far from the protein and N_{WP} is the excess solvation number for water. The excess solvation number and preferential interaction coefficient reported with units either as the moles of cosolute per mole of protein, or as grams of cosolute per gram of protein.

In the sample that the crystals were dissolved in, we are directly measuring the protein, c_P, and cosolute, c_X, concentrations and can use these estimates to calculate the excess solvation numbers. The protein concentration is used to estimate the total volume of solvent in the sample that originated from within the protein crystal:

$$c_{is} = \frac{c_P \cdot N_a \cdot V_{uc} \cdot f_s^I}{MW_P \cdot z} \quad \left(\frac{ml \text{ of internal solvent}}{ml} \right).$$

The concentration of the cosolute within the crystal's solvent channels can be calculated using the concentration of cosolute found within the sample:

$$c_X^{\text{crystal}} = \frac{c_X}{c_{is}} = \frac{c_X \cdot MW_P \cdot z}{c_P \cdot N_a \cdot V_{uc} \cdot f_s^I} \quad \left(\frac{mg \text{ of cosolute } X}{ml \text{ of internal solvent}} \right).$$

With the measured concentration of the cosolute in the soak solutions, c_X^{soak}, the excess solvation number is then

$$N_{XP} = \left(C_X^{\text{crystal}} - C_X^{\text{soak}} \right) \frac{V_{uc} \cdot f_s^I \cdot N_a}{n \cdot MW_X} \quad \left(\frac{\text{moles cosolute } X}{\text{moles protein}} \right). \qquad (4)$$

2.2.8 Number of water molecules in the unit cell

The number of water molecules in the unit cell are estimated by taking the total mass of the unit cell, determined from the crystal density, and subtracting the total mass of the protein and the cosolutes. The number of acetate molecules within the crystal's solvent cavities was assumed to be equal to the concentration of acetate in the soak solution. It was present in the soak solution at a concentration of

0.01 M, one to two orders of magnitude lower than any other cosolute so any the effects of any error should be negligible.

2.3 RESULTS

2.3.1 Glycerol and NaCl exclusion

Fig. 2.2 shows the excess solvation numbers for glycerol, Cl^- and Na^+ in tetragonal lysozyme crystals and glycerol in apoferritin crystals. The excess solvation number represents the number of cosolute molecules within the unit cell found in excess (positive) or deficient (negative) compared to an equivalent volume of the soak solution, divided by the number of protein monomers within the unit cell. These values were calculated using Eq. 4 and are proportional to the difference in concentrations of the cosolutes within the crystals and in the soak solutions, shown in Fig. 6.1.6. In lysozyme, glycerol was found to be excluded from the unit cell at all concentrations. In apoferritin, the relatively large error bars obscure the behaviour at 20 and 40% w/v glycerol, but glycerol was observed to be excluded from the crystals at 60% glycerol. Different behaviour was observed for Na^+ and Cl^- in tetragonal lysozyme. At all concentrations, Na^+ was found to be excluded from the crystal and the exclusion gradually increased with increasing concentration. Cl^- was found to be preferentially included into the unit cell and the inclusion was found to be the highest at the two lowest concentrations.

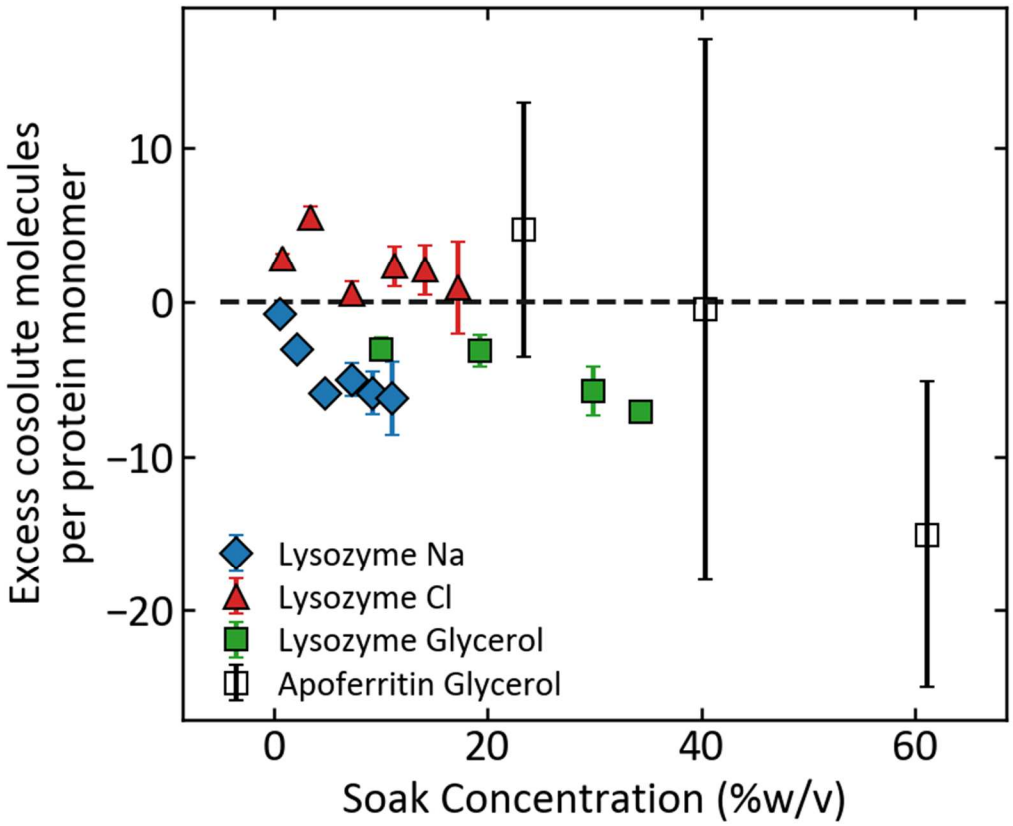


Figure 2.2 Excess solvation numbers for Na^+, Cl^- and glycerol within tetragonal lysozyme crystals and glycerol within apoferritin crystals. The excess solvation number represent the total number of cosolute molecules excluded from the protein crystal per protein monomer.

2.3.2 Crystal and internal solvent densities

Fig. 2.3 shows the density of the solvent within the crystal plotted against the density of the soak solutions, corrected for the measured exclusion or inclusion of cosolutes. The measured crystal densities are plotted against the soak solution densities in Fig. 6.1.7. For lysozyme, the calculated density of the solvent within the unit cell is highly dependent on the assumed value of the solvent fraction. No objective method is available to calculate the solvent fraction directly from the crystal structures to the precision required for this calculation (Section 6.1.1). The solvent fraction was calculated using the modified version of the Matthew's coefficient where the experimentally determined partial specific volume of the protein is used instead of the average (Section 2.6). The experimentally determined partial specific volume of a protein incorporates any modifications that the protein makes to the hydration shell. Given that the internal solvent density of lysozyme calculated with the experimental partial specific volume is equivalent to the density of the soak solutions, our observations suggests that no modifications are occurring the solvent within the unit cell that are not already occurring when the protein is in solution form.

Shown in the open square symbols, the solvent within the unit cell of apoferritin was estimated to be much larger than the soak solutions. This could be interpreted by an accumulation of Cd^{2+} ions into the apoferritin crystal. Cd^{2+} ions attach to apoferritin's surface at a rate highly dependent on pH. Pead (Pead *et al.*, 1995) measured 3 and 54 Cd^{2+} ions bound per apoferritin 24mer at pH 5.5 and 7.5 respectively. At pH 5.5, six Cd^{2+} binding sites were identified in anomalous difference maps of crystal structures (Granier *et al.*, 1998). This additional Cd^{2+} could account for a small increase in measured density, however, approximately 10 times more Cd^{2+} would need to be incorporated within the unit cell to account for the measured discrepancy. The density of apoferritin crystals in 1% $CdSO_4$ has previously been measured to be 1.134 g/cm^3 (Harrison, 1963), as opposed to our measured crystal density of 1.215 g/cm^3 (Fig. 6.1.7).

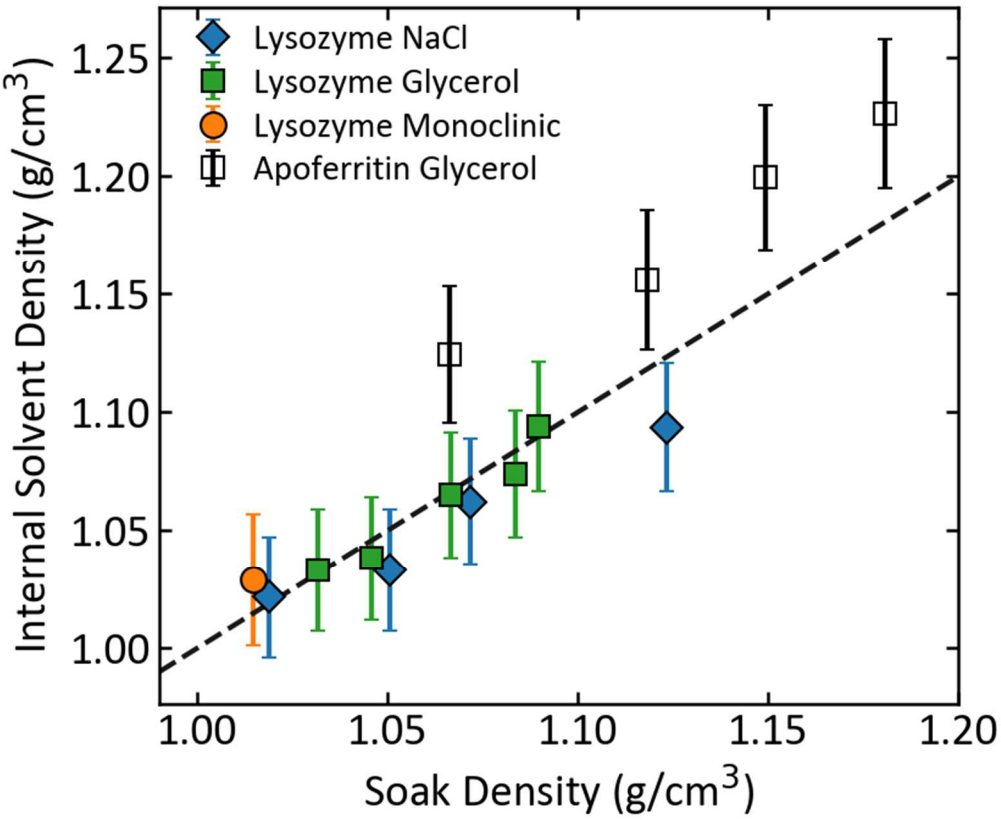


Figure 2.3 Estimated density of the solvent within the protein crystals.

2.3.3 Crystal composition

For lysozyme soaked in various concentrations of NaCl, the concentration of all the cosolutes within the soak solution present within the crystal were determined, with the exceptions of water and acetate. From these numbers, we can determine the total number of Na^+, Cl^-, and water molecules within the unit cell. The acetate was present in the soak solution at a concentration of 0.01 M and for the purpose of determining the number of water molecules within the unit cell, was assumed to be the same within the crystal. Since this is at least an order of magnitude lower than the concentration of the other cosolutes, error from this assumption is negligible. Fig. 2.4 shows the number of Na^+, Cl^-, and water atoms within the unit cell of lysozyme. The isoelectric point of lysozyme is 11.3 (Wetter & Deutsch, 1951; Kuehner *et al.*, 1999) and represents the pH at which lysozyme is electrically neutral. At pH values less than 11.3, lysozyme is found to have a net positive surface charge. Given a soak solution pH of 5.2, lysozyme has a net positive charge of approximately 10 electron units (Kuehner *et al.*, 1999), which should translate to an excess of 80 electron units per unit cell. Averaged over the different soak concentrations, there were 55 more Cl^- ions in the unit cell than Na^+ ions. The excess of the negatively charged Cl^- ions within the unit cell could be a result of neutralizing the ionic charges within the protein crystal.

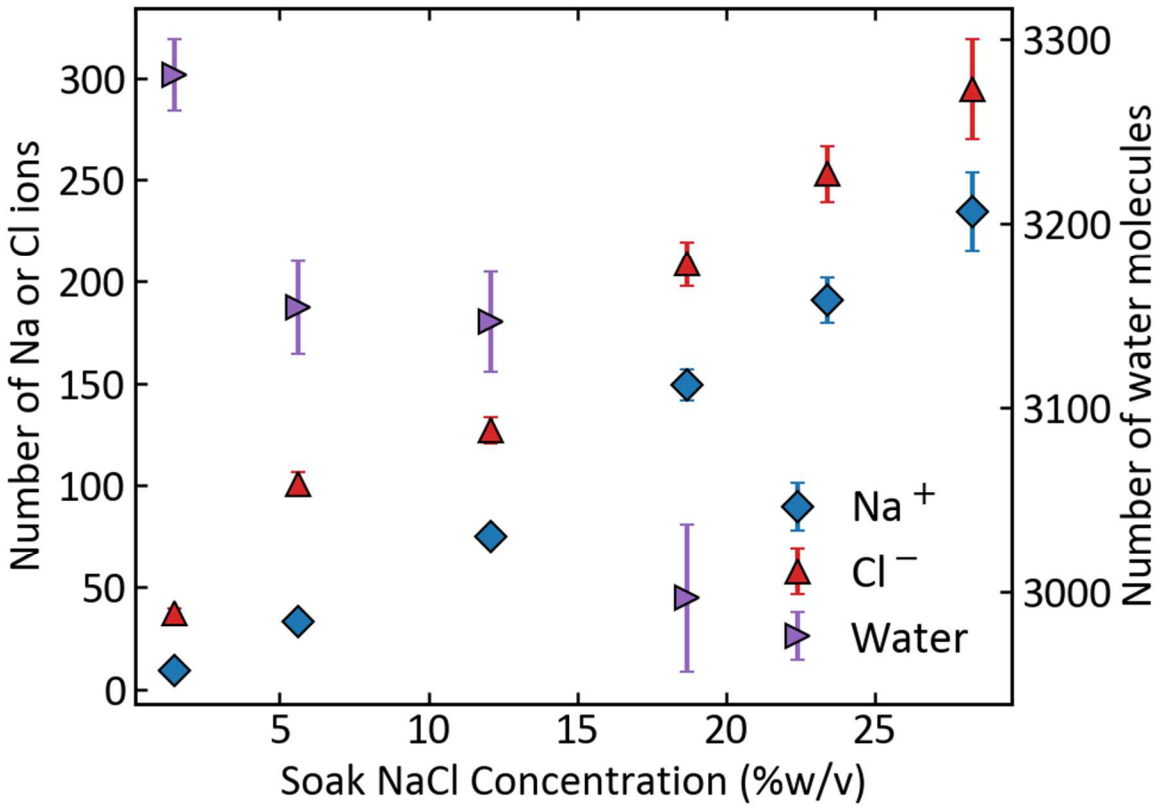


Figure 2.4 Total number of atoms within the tetragonal lysozyme crystals soaked in NaCl solutions.

2.4 Discussion

There have been several experimental measurements of the preferential interaction within protein crystals. The results from Schoenborn (1988) showed that myoglobin crystals with an external concentration of 32% w/w ammonium sulfate had an internal concentration of just 13% w/w. Soares & Caspar (2017) showed a reduction in trehalose concentration from 41.1% w/v external to 9.9% w/v internal. Several early experiments observed the exclusion of sucrose and other polysaccharides from β-lactoglobulin crystals (McMeekin *et al.*, 1950; Low *et al.*, 1956), however very little was understood about protein solvent interactions at the time. McMeekin *et al.* suggested that their observations were the result of a unit cell contraction as sucrose was added to the crystals. Low *et al.* followed up their results, they point out that β-lactoglobulin crystal's unit cell expands as sucrose is added, invalidating McMeekin's interpretation. Low *et al.* performed similar experiments and estimate the polysaccharide concentrations within β-lactoglobulin crystals by the crystal's optical density. They used the polysaccharides comprised of 2 (sucrose), 3 (raffinose & melezitose), 6 (α-dextrin), 8 (γ-dextrin) and 30 (inulin) repeating hexose rings. In all their measurements, they observe the polysaccharides excluded from the crystals and the degree of exclusion increased with increasing size, with inulin completely excluded from the crystals. They suggest their observations were due to the steric exclusion of the polysaccharides from the protein's surface, as the larger polysaccharides were less able to fit into the cavities in the protein's surface.

Elgersma *et al.* (1992) and Vekilov *et al.* (1996) observed, respectively, an excess of Cl^- and Na^+ ions within lysozyme crystals, in contrast to an exclusionary effect observed in solution (Arakawa & Timasheff, 1982). Elgersma and Vekilov's explanations for their results contradicted each other, Elgersma stated that Cl^- incorporation was likely due to changes in the dielectric constant within the crystal. Vekilov's experimental data showed that Na^+ incorporation was due to accumulation in defects within the crystals. The remaining questions from these results are, why are excess Na^+ and Cl^- ions found within protein crystals and where are these ions located, in defects or within the solvent cavities? The

discrepancy between the solution and crystal-based measurements occurs because the solution-based measurements do not distinguish between the different species of ions. Our data shows distinctly different behavior between Na^+ and Cl^- similar to the observations of Palmer (1948). Na^+ is clearly excluded from the crystal while Cl^- is either excluded or found at similar concentrations internally and externally to the crystal. If we add the separate effects of Na^+ and Cl^-, we see that the net effect is NaCl is excluded from the crystal.

Since we were using large crystals, grown slowly over a month of time from high purity protein stock, they should have a much lower number of defects than the crystals used in Vekilov's experiments, where they intentionally used protein stock with impurities to introduce defects. In our crystals we observe that Na^+ is excluded from the crystals and Vekilov's observation of excess Na^+ must have been due to crystal defects. Our observations of Cl^- incorporation were also in line with Elgersma's measurements and we interpret the results as neutralizing the large amounts of positive charge found on lysozyme.

This charge neutralization is due to the Gibbs-Donnan effect. In this effect, charged particles are found unevenly distributed across a semi-permeable membrane because a different charged particle that is too large to pass through the membrane exists only on one side. In a way, the protein crystal is a charged membrane, water and salt ions freely exchange between it at the external solvent, but the protein molecules are bound to the crystal. The Gibbs-Donnan effect is known to occur with charged proteins (Marrack & Hewitt, 1927, 1929) and has been previously discussed in the context of protein crystals (Cvetkovic *et al.*, 2004; López-Jaramillo *et al.*, 2002, 2003; Kundrot & Richards, 1988).

Fig. 2.5 shows the excess solvation number that we calculate using the experimental observations of cosolute exclusion from protein crystals and compare them to the preferential interaction coefficients determined from solution-based measurements (opaque). The units of the preferential interaction or

excess solvation number are converted to grams of cosolute excluded per gram of protein to facilitate comparison between different proteins and cosolutes. These observed exclusion of cosolutes from the crystals made by Schoenborn and Soares & Caspar are interpreted in terms of the preferential interaction and compared to the exclusion measured for similar cosolutes with different proteins. Our observations of the preferential interaction with glycerol, Na^+ and Cl^- are included in this plot along with our interpretations of the observations of polysaccharide exclusion from β-lactoglobulin crystals (McMeekin *et al.*, 1950; Low *et al.*, 1956) and Na^+ and Cl^- exclusion from tetragonal lysozyme crystals (Palmer *et al.*, 1948). All these observations of cosolutes being excluded from protein crystals are comparable to solution-based measurements.

Crystal density measurements could be a useful means to initialize MD simulations of protein crystals. In our new method for density measurements, we set upper and lower bounds on the crystal's density based on sink / float observations in mixtures of 1PO and silicone oil. In all measurements, we were able to reduce the sink / float transition region to a width of 0.5% w/w 1PO in silicone oil, which translates to 2 mg /cm^3. Systematic uncertainty in the conversion between mixture ratio and density as well as uncertainty in the precision to which we could mix the 1PO and silicone oil increases our uncertainty on the crystal densities to 5 mg/cm^3, which translates to 0.4% uncertainty for a 1.25 g/cm^3 crystal.

Using our density measurements, we were able to estimate the density of the solvent within the unit cell. The estimated internal density is very sensitive to the assumptions that go into the estimation of the solvent fraction. We chose to use the protein's partial specific volume from previously reported, solution-based experiments. Any modification that the protein makes to its hydration layer is accounted for in these values and will not be captured in our density measurements. For lysozyme, our results show that the density of the solvent within the unit cell is indistinguishable from the external soak solution,

suggesting that no modifications to the water are occurring within the crystal that are not already happening in solution.

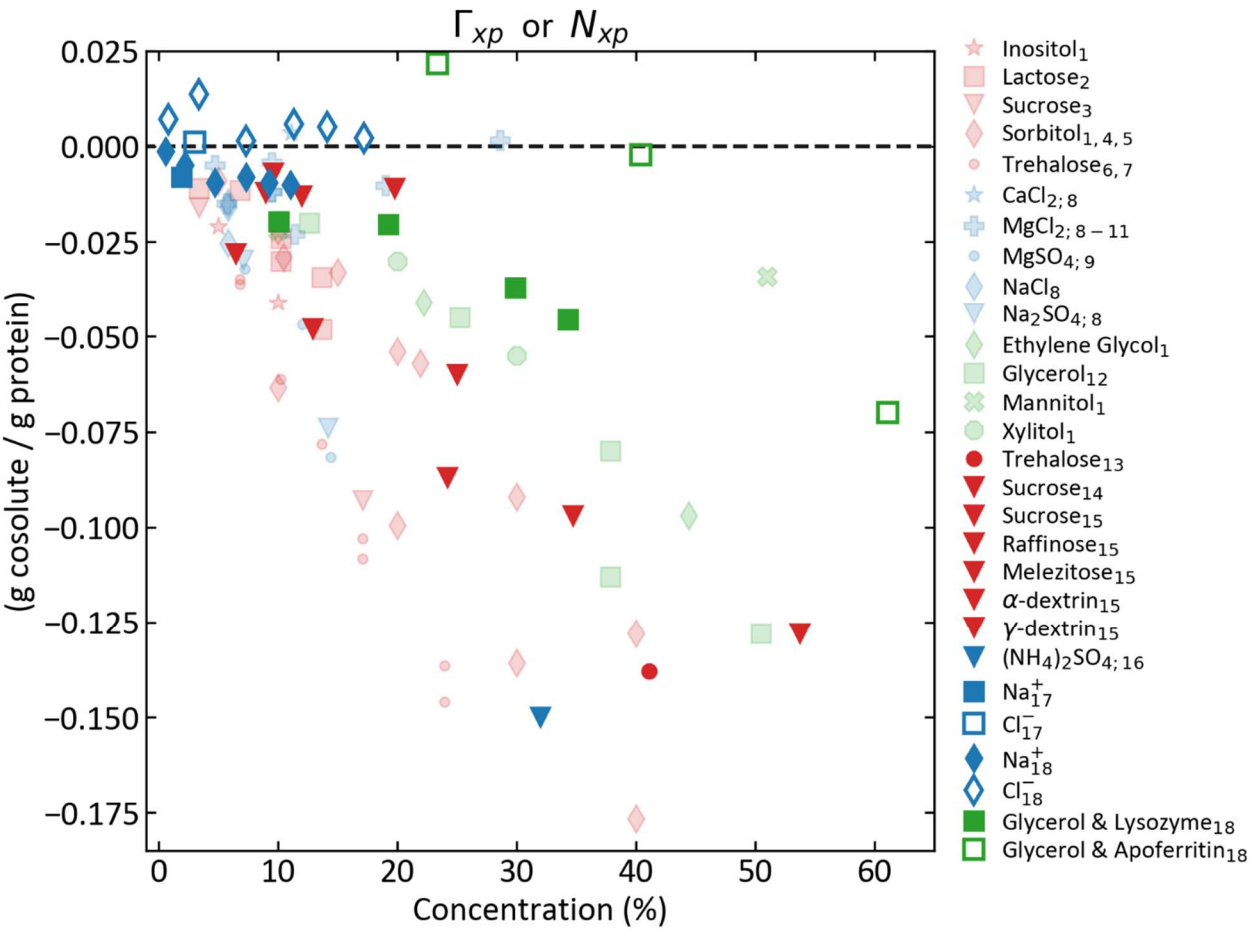


Figure 2.5 Preferential interaction coefficients or excess solvation numbers from data measured in solution (partially transparent) and in crystals (bold). Units are converted to grams of cosolute per gram of protein to enable comparison between different proteins. Corresponding references are (1) Gekko & Morikawa, 1981; (2) Arakawa & Timasheff, 1982; (3) Lee & Timasheff, 1981; (4) Gekko, 1922; (5) Xie & Timasheff 1997a; (6) Xie & Timasheff, 1997; (7) Lin & Timasheff, 1996; (8) Arakawa & Timasheff, 1982a; (9) Arakawa et al., 1990a; (10) Arakawaw & Timasheff 1984; (11) Arakawa et al. 1990b; (12) Gekko & Timasheff 1981; (13) Soares & Caspar 2017; (14) McMeekin et al. 1950; (15) Low et al. 1956; (16) Schoenborn 1988; (17) Palmer et al. 1948; (18) This paper.

2.5 REFERENCES

Arakawa, T., Bhat, R. & Timasheff, S. N. (1990). *Biochemistry*. **29**, 1914–1923.

Arakawa, T. & Timasheff, S. N. (1982). *Biochemistry*. **21**, 6545–6552.

Arakawa, T. & Timasheff, S. N. (1983). *Arch. Biochem. Biophys.* **224**, 169–177.

Arakawa, T. & Timasheff, S. N. (1984). *Biochemistry*. **23**, 5912–5923.

Arakawa, T. & Timasheff, S. N. (1985). *Biophys. J.* **47**, 411–414.

Bhat, R. & Timasheff, S. N. (1992). *Protein Sci.* **1**, 1133–1143.

Cazals, F. (2006). *Protein Sci.* **15**, 2082–2092.

Colman, P. M. & Matthews, B. W. (1971). *J. Mol. Biol.* **60**, 163–168.

Connolly, M. L. (1985). *J. Am. Chem. Soc.* **107**, 1118–1124.

Cousido-Siah, A., Petrova, T., Hazemann, I., Mitschler, A., Ruiz, F. X., Howard, E., Ginell, S., Atmanene, C., Van Dorsselaer, A., Sanglier-Cianférani, S., Joachimiak, A. & Podjarny, A. (2012). *Proteins Struct. Funct. Bioinforma.* **80**, 2552–2561.

Cvetkovic, A., Zomerdijk, M., Straathof, A. J. J., Krishna, R. & Van Der Wielen, L. A. M. (2004). *Biotechnol. Bioeng.* **87**, 658–668.

Davis, R. S. (1992). *Metrologia*. **29**, 67–70.

Elgersma, A. V., Ataka, M. & Katsura, T. (1992). *J. Cryst. Growth*. **122**, 31–40.

Fischer, H., Polikarpov, I. & Craievich, A. F. (2004). *Protein Sci.* **13**, 2825–2828.

Fokine, A. & Urzhumtsev, A. (2002). *Acta Cryst. D*. **58**, 1387–1392.

Gekko, K. (1982). *J. Biochem.* **91**, 1197–1204.

Ghirlando, R., Mutskova, R. & Schwartz, C. (2015). *Nanotechnology*. **27**, 45102.

Granier, T., Comberton, G., Gallois, B., Langlois D'Estaintot, B., Dautant, A., Crichton, R. R. & Précigoux, G. (1998). *Proteins Struct. Funct. Genet.* **31**, 477–485.

Harrison, P. M. (1963). *J. Mol. Biol.* **6**, 404–422.

Janowski, P. A., Cerutti, D. S., Holton, J. & Case, D. A. (2013). *J. Am. Chem. Soc.* **135**, 7938–7948.

Judge, R. A., Jacobs, R. S., Frazier, T., Snell, E. H. & Pusey, M. L. (1999). *Biophys. J.* **77**, 1585–1593.

Juers, D. H. & Ruffin, J. (2014). *J. Appl. Cryst.* **47**, 2105–2108.

Kim, H. S., Martel, A., Girard, E., Moulin, M., Härtlein, M., Madern, D., Blackledge, M., Franzetti, B. & Gabel, F. (2016). *Biophys. J.* **110**, 2185–2194.

Kuehner, D. E., Engmann, J., Fergg, F., Wernick, M., Blanch, H. W. & Prausnitz, J. M. (1999). *J. Phys. Chem. B*. **103**, 1368–1374.

Kundrot, C. E. & Richards, F. M. (1988). *J. Mol. Biol.* **299**, 401–410.

Lee, J. C. & Lee, L. L. (1981). *J. Biol. Chem.* **256**, 625–631.

Li, A. J. & Nussinov, R. (1998). *Proteins Struct. Funct. Genet.* **32**, 111–127.

Liang, J. & Dill, K. A. (2001). *Biophys. J.* **81**, 751–766.

López-Jaramillo, F. J., Moraleda, A. B., González-Ramirez, L. A., Carazo, A. & García-Ruiz, J. M. (2002). *Acta Cryst. D*. **58**, 209–214.

López-Jaramillo, F. J., Otálora, F. & Gavira, J. A. (2003). *J. Cryst. Growth*. **247**, 177–184.

Low, B. W. & Richards, F. M. (1952). *J. Am. Chem. Soc.* **74**, 1660–1666.

Low, B. W., Richards, F. M. & Berger, J. E. (1956). *J. Am. Chem. Soc.* **78**, 1107–1113.

Marrack, J. & Hewitt, L. F. (1927). *Biochem. J.* **21**, 1129–1140.

Marrack, J. & Hewitt, L. F. (1929). *Biochem. J.* **23**, 1079–1089.

Matthews, B. W. W. (1968). *J. Mol. Biol.* **33**, 491–497.

McMeekin, T. L., Groves, M. L. & Hipp, N. J. (1950). *J. Am. Chem. Soc.* **72**, 3662–3666.

Meisburger, S. P., Case, D. A. & Ando, N. (2020). *Nat. Commun.* **11**, 1–13.

Merzel, F. & Smith, J. C. (2002). *Proc. Natl. Acad. Sci. USA*. **99**, 5378–5383.

Moreau, D. W., Atakisi, H. & Thorne, R. E. (2019). *IUCrJ*. **6**, 346–356.

Nittinger, E., Schneider, N., Lange, G. & Rarey, M. (2015). *J. Chem. Inf. Model.* **55**, 771–783.

Ortore, M. G., Spinozzi, F., Mariani, P., Paciaroni, A., Barbosa, L. R. S., Amenitsch, H., Steinhart, M., Ollivier, J. & Russo, D. (2009). *J. R. Soc. Interface*. **6**,.

Palmer, K. J., Ballantyne, M. & Galvin, J. A. (1948). *J. Am. Chem. Soc.* **70**, 906–908.

Pead, S., Durrant, E., Webb, B., Larsen, C., Heaton, D., Johnson, J. & Watt, G. D. (1995). *J. Inorg. Biochem.* **59**, 15–27.

Persson, F., Söderhjelm, P. & Halle, B. (2018). *J. Chem. Phys.* **148**, 215101.

Pierce, V., Kang, M., Aburi, M., Weerasinghe, S. & Smith, P. E. (2008). *Cell Biochem. Biophys.* **50**, 1–22.

Richards, F. M. (1977).

Schoenborn, B. P. (1988). *J. Mol. Biol.* **201**, 741–749.

Soares, A. S. & Caspar, D. L. D. (2017). *J. Struct. Biol.* **200**, 213–218.

Svergun, D. I., Richard, S., Koch, M. H., Sayers, Z., Kuprin, S. & Zaccai, G. (1998). *Proc. Natl. Acad. Sci. USA*. **95**, 2267–2272.

Timasheff, S. N. (2002). *Biochemistry*. **41**, 13473–13482.

Timasheff, S. N. & Xie, G. (2003). *Biophys. Chem.* **105**, 421–448.

Vekilov, P. G., Monaco, L. A., Thomas, B. R., Stojanoff, V. & Rosenberger, F. (1996). *Acta Cryst. D*. **52**,

785–798.

Voss, N. R. & Gerstein, M. (2010). *Nucleic Acids Res.* **38**, 555–562.

Wall, M. E. (2018). *IUCrJ*. **5**, 172–181.

Wall, M. E., Van Benschoten, A. H., Sauter, N. K., Adams, P. D., Fraser, J. S. & Terwilliger, T. C. (2014). *Proc. Natl. Acad. Sci. USA*. **111**, 17887–17892.

Wall, M. E., Calabró, G., Bayly, C. I., Mobley, D. L. & Warren, G. L. (2019). *J. Am. Chem. Soc.*

Warkentin, M. & Thorne, R. E. (2009). *J. Appl. Cryst.* **42**, 944–952.

Weichenberger, C. X., Afonine, P. V., Kantardjieff, K. & Rupp, B. (2015). *Acta Cryst. D*. **71**, 1023–1038.

Weichenberger, C. X. & Rupp, B. (2014). *Acta Cryst. D*. **70**, 1579–1588.

Westbrook, E. M. (1985). *Methods Enzymol.* **114**, 187–196.

Wetter, L. R. & Deutsch, H. F. (1951). *J. Biol. Chem.* **192**, 237–242.

Winter, G., Waterman, D. G., Parkhurst, J. M., Brewster, A. S., Gildea, R. J., Gerstel, M., Fuentes-Montero, L., Vollmar, M., Michels-Clark, T., Young, I. D., Sauter, N. K. & Evans, G. (2018). *Acta Cryst. D*. **74**, 85–97.

Zeana, D., Becker, J., Kuckelkorn, R. & Kirchhof, B. (1999). *Int. Ophthalmol.* **23**, 17–24.

3 ICE FORMATION AND SOLVENT NANOCONFINEMENT IN PROTEIN CRYSTALS

Synopsis Nanoconfinement dramatically modifies the behavior of solvent within protein crystals, allowing biophysical measurements in the presence of liquid solvent at temperatures down to ~200 K. When internal ice forms it is stacking disordered, consistent with nucleation within deeply supercooled solvent, and ice does not form in solvent within ~5 Å of the protein surface.

Abstract Ice formation within protein crystals is a major obstacle to cryocrystallographic study of protein structure, and has limited studies of how a protein's structural ensemble evolves with temperature in the biophysically interesting range from ~260 K to the protein-solvent glass transition near 200 K. Using protein crystals having solvent cavities as large as ~70 Å, we use time-resolved X-ray diffraction to study the response of protein and internal solvent during rapid cooling. Solvent nanoconfinement suppresses freezing temperatures and ice nucleation rates so that ice-free, low-mosaicity diffraction data can be reliably collected down to 200 K without use of cryoprotectants. Hexagonal ice (I_h) forms in external solvent, but internal crystal solvent forms stacking disordered ice (I_{sd}) with a near random stacking of cubic and hexagonal planes. Analysis of powder diffraction from internal ice and single crystal diffraction from the protein lattice shows that the maximum crystallisable solvent fraction decreases with decreasing crystal solvent cavity size, and that a ~5 Å thick layer of solvent adjacent to the protein surface cannot crystallize. These results establish protein crystals as excellent model systems for study of nanoconfined solvent. By combining fast cooling, intense X-ray beams, and fast X-ray detectors, complete structural data sets for high-value targets including membrane proteins and large complexes may be collected at ~220-240 K that have much lower mosaicities, comparable B factors, and that may allow more confident identification of ligand binding than in current cryocrystallographic practice.

3.1 INTRODUCTION

Ice formation and its prevention are key issues in many areas of bioscience and biotechnology, including the cold hardiness of microorganisms, animals, and agriculturally-relevant plants; the cryopreservation of cells, tissues, and organs; the cold storage of proteins and biologics; and in biomolecular and cellular structure determination using electrons and X-rays.

X-ray crystallography is our primary tool for probing biomolecular structure. In its early days, crystallography was performed using protein crystals at or near room temperature. Once synchrotron X-ray sources became widely available in the 1990s, data collection shifted to near-exclusive use of cryogenically cooled crystals. Cryocooling to ~100 K reduces the rate at which diffraction properties degrade with X-ray dose by a factor of ~50, increasing the amount of data that can be collected per crystal, and reduces thermal motions, often increasing resolution (Rupp, 2009; Pflugrath, 2015). In favourable cases, crystallography beamlines can now collect diffraction data sets sufficient for protein structure determination in less than one second.

However, proteins have complex, multi-tiered energy landscapes, and biologically relevant information is lost when crystals are cryocooled due to thermal freeze-out of conformational motions (Fraser *et al.*, 2009; Keedy *et al.*, 2015; Halle, 2004) and due to steric hindrances imposed by increased molecular packing densities – the same factors responsible for improved diffraction resolution. Only a handful of crystallographic studies have examined the temperature evolution of protein structure in the biophysically interesting regime down to the protein-solvent glass (or dynamical) transition near ~200 K, where most non-harmonic motions are kinetically quenched and enzymatic activity ceases (Frauenfelder *et al.*, 1979; Tilton *et al.*, 1992; Teeter *et al.*, 2001). Recent studies (Keedy *et al.*, 2015) enabled by advances in electron density interpretation and modelling (Lang *et al.*, 2014; Fraser *et al.*, 2011; van den Bedem *et al.*, 2009) have illustrated the unique potential of variable-temperature crystallography to provide all-

atom, atomic resolution information about protein conformational ensembles, solvent structure, and energy landscapes and their connection to function.

A primary challenge in cryo- and especially variable-temperature crystallography is ice formation in crystal solvent, which disrupts the protein lattice and leads to loss of ordered diffraction (Rupp, 2009; Pflugrath, 2015). Cryoprotectants such as glycerol, PEGs, and alcohols are added to crystallization solutions or used in post-crystallization soak solutions to suppress ice formation (Pflugrath, 2015). For data collection at 100 K, typical concentrations are 20-30% v/v (Pflugrath, 2015), rising to 60% v/v or more for high solvent content (>80% v/v) crystals. For data collection between bulk water's homogeneous nucleation temperature $T_h \approx 235\ \mathrm{K}$ and 180 K, crystals have been soaked in 75% v/v methanol (Tilton *et al.*, 1992). Cryoprotectants can stabilize proteins, but they can also perturb protein structure, degrade crystal diffraction, and displace or be difficult to distinguish from weakly bound ligands in active sites (Pozharski *et al.*, 2013). Even when ice does not form, cryocooling to 100 K degrades long-range lattice order: mosaicities increase from <0.01° to 0.3° or more, leading to diffraction peak overlap when crystal unit cells are large. The challenges posed by cryocooling and ice formation are growing as the focus of structural studies shifts from smaller soluble proteins to membrane proteins, large biomolecular complexes, and to weakly packed, large solvent content crystal forms that are most likely to reveal native-like conformational ensembles and responses to optical, chemical, or thermal perturbations.

Related challenges are encountered in the cryopreservation of protein solutions, cells and tissues (Fahy & Wowk, 2015). However, in protein crystals the solvent is nano-confined within a periodic protein structure. Studies of water confined within nanoporous inorganic (primarily silica) matrices over the last two decades have shown that nanoconfinement dramatically modifies ice formation (Morishige & Kawano, 1999; Schreiber *et al.*, 2001; Jähnert *et al.*, 2008; Suzuki, Steinhart *et al.*, 2015; Taschin *et al.*, 2015; Mascotto *et al.*, 2017; Moore *et al.*, 2010). Deficiencies of the available matrices have complicated

interpretation of experiments, especially on non-equilibrium aspects such as nucleation, and made studies with biophysically relevant solvent compositions difficult.

Here we examine solvent behaviour and ice formation in protein crystals between 180 K and 260 K, using data from over 400 crystals of three proteins having solvent cavities as large as ~7 nm. We show that protein crystals enable new quantitative approaches to probing the effects of nanoconfinement on ice formation. Nanoconfinement strongly modifies the form and formation of internal ice in protein crystals, and enables biophysical measurements of the conformational evolution and dynamics of proteins in the presence of liquid solvent at temperatures down to ~200 K.

3.2 METHODS

3.2.1 Crystal growth, soaking, and X-ray data collection

Our studies focused on crystals of cubic apoferritin and tetragonal thaumatin with additional measurements with tetragonal lysozyme (Sec. 6.2.1). All crystals were grown by the hanging drop vapor diffusion method in 24-well plates.

Crystals of equine spleen apoferritin (Sigma A-3641) were grown in hanging drops of 2 µL of 10 mg /mL protein in 0.1M sodium acetate buffer at pH 6.5 and 2 µL of a well solution of 2% w/v $CdSO_4$ and 15% w/v $(NH_4)_2SO_4$ in the same buffer. Cubic crystals in space group $F4_32$ grew to dimensions of 300-500 µm within a week (Fig. 6.2.1(a)).

Crystals of thaumatin (Sigma T7638) were grown in hanging drops comprised of equal volumes of 40 mg/mL protein in 0.1 M sodium acetate buffer at pH 6.5 and a well solution of 14% w/v potassium-sodium-tartrate in the same buffer. Tetragonal crystals in space group $P4_12_12$ grew to dimensions of 200-300 µm within one week (Fig. 6.2.1(b)).

Crystals of lysozyme (Sigma L6876) were grown in hanging drops comprised of equal volumes of 80 mg/ml protein in 0.1 M sodium acetate buffer at pH 5.2 and a well solution of 2.5% w/v of NaCl in the same buffer. Tetragonal crystals in the space group $P4_32_12_1$ grew to dimensions of 300-800 μm. Crystal appeared within one week and stopped growing within four weeks.

Crystals were used as grown or else cryoprotected by soaking for at least 5 min in glycerol solutions having concentrations of 10%, 20%, and 40% v/v, obtained by adding glycerol to a solution with the same composition as the previously mentioned well solutions. Each crystal was transferred to a separate drop of NVH oil (Cargille) and manipulated until all external solvent was removed from their surface (Warkentin & Thorne, 2010*b*). Crystals were mounted on microfabricated loops encapsulated in NVH oil to prevent dehydration during data collection and stored in MicroRT tubes (MiTeGen) containing mother liquor or cryoprotectant solution for ~1 hour prior to data collection.

X-ray data was collected on station F1 at the Cornell High-Energy Synchrotron Source (CHESS) using a Pilatus 6M detector (Sec. 6.2.2). A cold nitrogen gas stream programmed to the desired final sample temperature was directed at the crystal but was initially blocked using a shutter. Each crystal was placed in the X-ray beam at room temperature, 10 frames totalling 5° in rotation were collected to assess the crystal for damage or dehydration, and then the crystal was rotated back to its initial orientation. The gas stream was then unblocked and collection of frames with 0.5° rotation and 0.1 to 0.2s exposure per frame commenced (Fig. 6.2.3).

3.2.2 Processing and modelling of protein lattice diffraction

Diffraction frames were indexed, integrated, and scaled using XDS in segments of 5 frames (Sec. 6.2.3). Structural models were derived from frames starting after the unit cells reached a stable equilibrium and until the end of data collection. Molecular replacement and model refinement were performed using PHENIX, and results checked using Coot (Sec. 6.2.4). Protein and solvent volumes were then evaluated

using the final refined models (Sec. 6.2.5). Refinement statistics for the 45 apoferritin and 53 thaumatin structures used in the analysis are given in the Supporting Information.

3.2.3 Processing and modelling of ice diffraction

Diffraction frames from the detector were processed using python scripts to remove protein Bragg scattering and background, and azimuthally integrated (Sec. 6.2.6). The resulting intensity vs resolution plots were analysed by embedding the program DIFFaX (Treacy *et al.*, 1991), which calculates diffraction from samples containing stacking faults, in an optimization routine to determine best-fit parameters for stacking disordered ice formed of planes of hexagonal (I_h) and cubic (I_c) ice. These fits were compared with those obtained assuming a simple mix of cubic and hexagonal crystallites (Sec. 6.2.7).

3.2.4 Estimating ice fractions in protein crystals

Structure factors calculated from models of protein lattice and ice diffraction were used to normalize protein lattice and ice diffraction data collected from the same crystal in the same X-ray beam using the same detector, yielding the ratio of ice to protein crystal volume (Sec. 6.2.8). The ice volume was then compared with the protein crystal's solvent cavity volume at the ice observation temperature, with corrections applied based on estimates of the solvent fraction that exited the unit cell on cooling.

3.3 Results and Discussion

3.3.1 Solvent content and solvent cavity size distributions in protein crystals

Figs. 1 and S4 show the size distribution and cumulative distribution, respectively, of the largest solvent cavity within a protein crystal's unit cell versus solvent content and unit cell volume for 17,146 non-redundant protein structures, obtained from the Protein Data Bank (PDB) (Sec. 6.2.9). Maximum solvent cavity size, a primary determinant of ice formation during cooling, tends to increase with both solvent content and unit cell volume. Diffraction resolution at 100 K degrades with increasing solvent

content and solvent channel size (Fig. 6.2.8), due to reduced constraints on atomic displacements from crystal packing interactions and to increased disorder caused by cryocooling. Membrane protein crystals tend to have larger non-protein volume fractions and larger solvent channels than soluble proteins (Fig. 3.1 (c,d) and Fig. 6.2.7) contributing to the difficulty in obtaining high quality structural data sets from these crystals. For pure water in nanoporous silica and alumina, the effects of nanoconfinement on freezing and melting temperatures become pronounced for cavity sizes below ~10 nm (Jähnert, *et al*., 2008; Mascotto *et al.*, 2017; Morishige & Kawano, 1999; Schreiber *et al*., 2001; Suzuki *et al*., 2015; Taschin *et al*., 2015). Fig. 3.1 suggests that the effects of nanoconfinement on solvent behaviour should be pronounced in nearly all protein crystals.

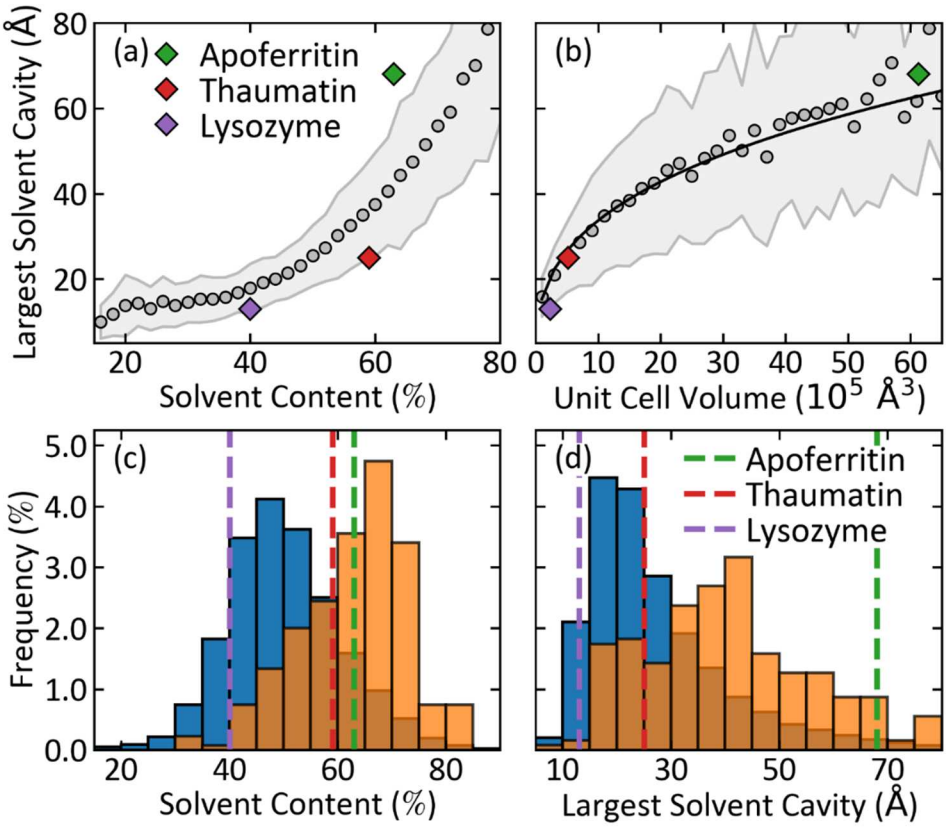


Figure 3.1 (a),(b) Mean size and distribution of the largest solvent cavity within the unit cell versus solvent content and unit cell volume obtained from 17,148 non-redundant protein structures in the PDB, excluding small peptides and viral proteins. Symbols indicate mean values and shading the region within one standard deviation of the distribution's maximum. The solid line fit in (b) has the form (cavity size) $\propto$ (cell volume)$^{1/3}$, so cavity size scales with linear unit cell dimension. (c),(d) Histograms of PDB entry distributions versus solvent content and largest solvent cavity, for soluble proteins (blue) and membrane proteins (orange). Corresponding values for cubic apoferritin, tetragonal thaumatin, and tetragonal lysozyme crystals are marked in each frame.

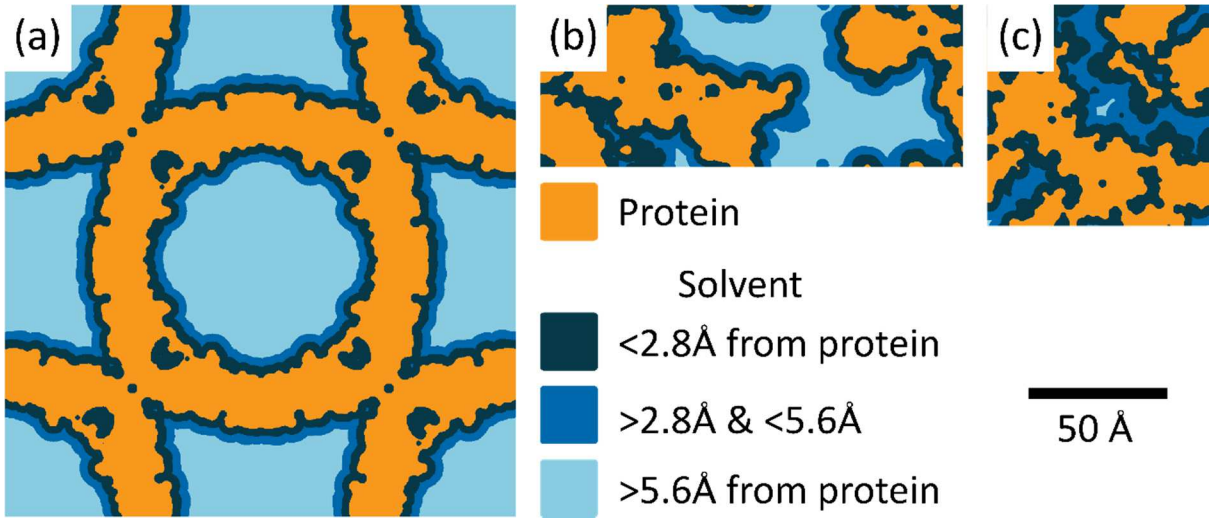


Figure 3.2 Solvent cavity structure in (a) cubic apoferritin, (b) tetragonal thaumatin, and (c) tetragonal lysozyme crystals at room temperature. The van der Waals surface of the protein is shown in orange. Solvent spaces within the first and second hydration shells are shown in dark blue and medium blue, respectively.

We studied cubic apoferritin, tetragonal thaumatin, and (with limited measurements) tetragonal lysozyme crystals, having Matthews-coefficient derived solvent contents of 63%, 59% and 42% v/v and maximum solvent cavity sizes of 68 Å, 25 Å, and 13 Å, respectively, as indicated in Figs. 3.1 and 3.2. These maximum cavity sizes span the range of relevance in protein crystallography (Fig. 3.1), with apoferritin's cavities being larger than those found in ~98% of PDB entries (Fig. 6.2.7). Although apoferritin and thaumatin have similar solvent contents, the fractions of solvent located beyond the protein's first two hydration shells are very different (Figs. 3.2 and 6.2.2).

Time-resolved X-ray diffraction measurements were performed on apoferritin and thaumatin crystals soaked in solutions containing 10%, 20%, and 40% v/v glycerol or harvested as is (0% v/v) and then abruptly cooled (in <1 s) to temperatures between 180 and 260 K (Sec. 6.2.3). Fig. 3.3 shows the fraction of apoferritin crystals that remained free of internal ice and diffracted to high resolution for at least (a) 3 and (b) 20 seconds, times sufficient to collect complete structural data sets on high brilliance synchrotron beamlines, after their unit cell reached its steady state or minimum value. No apoferritin crystals, regardless of glycerol concentration, showed internal ice formation at temperatures above 240 K. Ice eventually appeared below 240 K in crystals with lower glycerol concentrations. But at all temperatures internal solvent within at least a substantial minority of these crystals could be maintained in a supercooled state for at least a few seconds. Ice first became detectable up to ~20 s after crystals reached their steady-state temperature. Similar results were obtained using thaumatin crystals (Sec. 6.2.10 and Fig. 6.2.9). Salt concentrations present within the crystallization solutions and internal crystal solvent suppress bulk freezing temperatures by only a few degrees. When bulk solutions containing these salt concentrations are cooled below the freezing temperature, ice forms in ~10 ms unless the temperature drops below the solvent glass transition temperature T_g first (Sec. 6.2.11). Consequently, the suppression of ice formation in protein crystals with up to ~70 Å solvent cavities must be due to nanoconfinement.

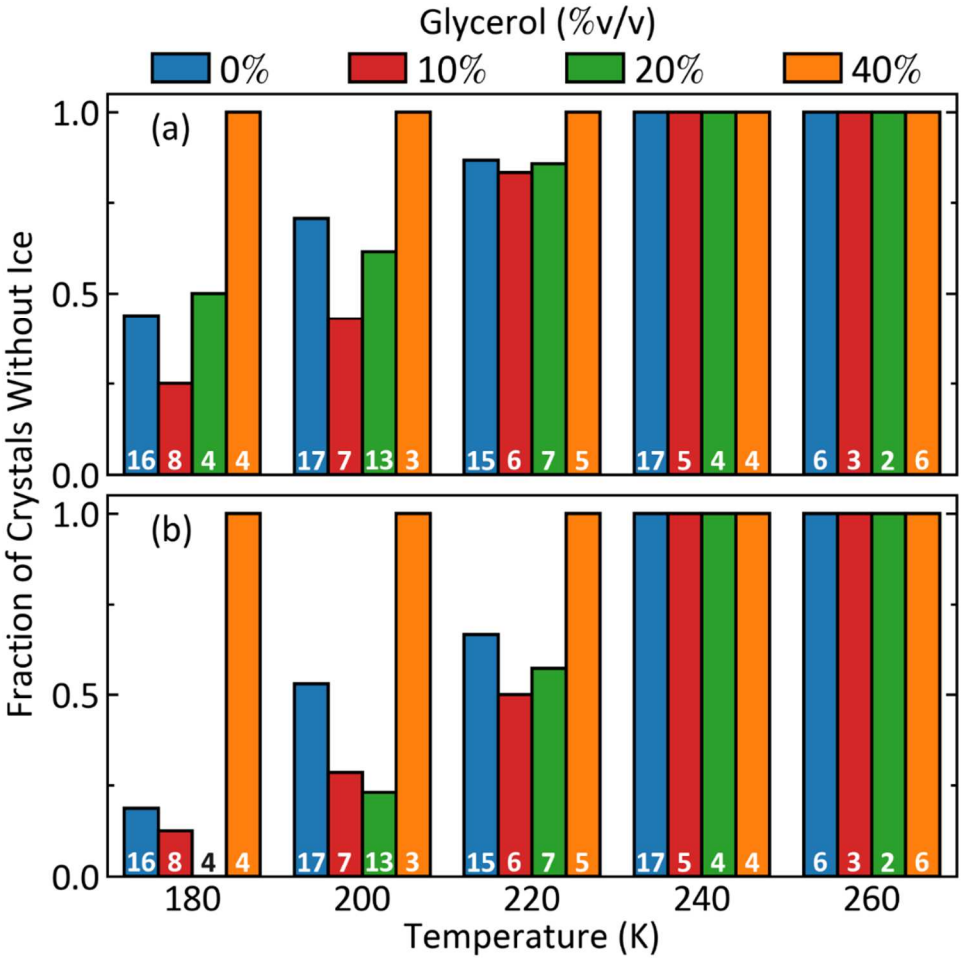


Figure 3.3 Fraction of apoferritin crystals at each temperature and glycerol concentration that remained ice-diffraction free for at least (a) 3 and (b) 20 seconds after the unit cell reached its steady state or minimum value, excluding roughly 25% of crystals that formed hexagonal ice in external solvent. Numbers on each bar indicate crystals examined for each condition. Data for thaumatin are given in Fig. 6.2.9.

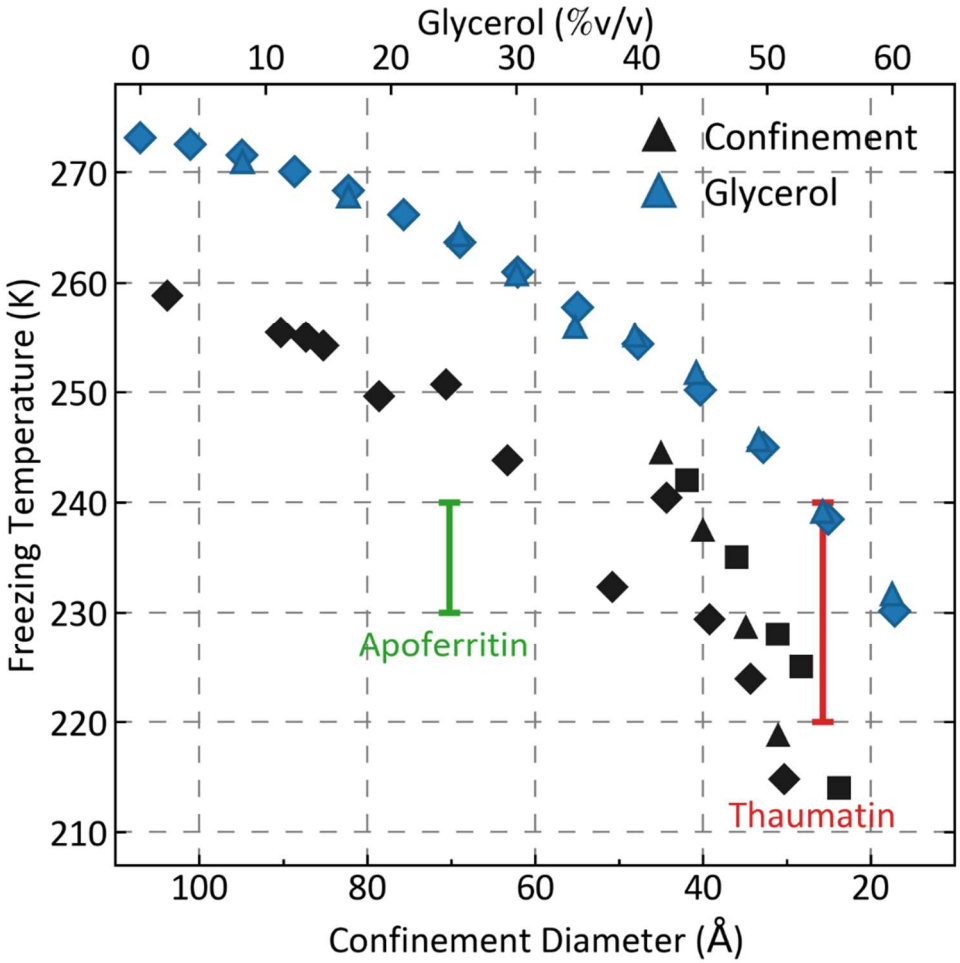


Figure 3.4 Blue points and upper scale: freezing temperature of bulk aqueous glycerol solutions versus glycerol concentration, from Refs. (Lane, 1925) (triangles) and (Segur, 1946) (diamonds). Black points and lower scale: freezing temperature of pure water versus confinement diameter for confinement in cylindrical nanopores formed in silica, from Refs. (Findenegg et al., 2008) (triangles), (Jähnert et al., 2008) (squares), and (Kittaka et al., 2006) (diamonds). Confinement within ~100 Å pores is as effective in suppressing freezing temperatures as is adding 30-35% (v/v) glycerol.

As shown in Fig. 3.4, for pure water in nanoporous silica and alumina matrices, the freezing temperature T_f decreases with decreasing pore diameter (Morishige & Kawano, 1999; Schreiber *et al.*, 2001; Jähnert *et al.*, 2008). For cylindrical pores, T_f ~250 K and 223 K for diameters of 67 and 29 Å, respectively; no phase transition is observed for diameters below ~20 Å (Jähnert *et al.*, 2008). The maximum temperatures at which ice forms in glycerol-free apoferritin and thaumatin crystals are comparable, based on their maximum solvent cavity sizes, to these previous measurements.

Freezing point suppression is an *equilibrium* effect of nanoconfinement. Long delays between cooling to below the freezing point and ice formation – and thus the persistence of metastable supercooled internal solvent – in glycerol-free apoferritin and thaumatin crystals at temperatures as low as 200 K indicate that solvent nanoconfinement within protein crystals dramatically modifies the *kinetics* of ice nucleation and growth. Reported nucleation rates between 193 K and 215 K in supersonic-nozzle-generated water nanodrops of diameters between 60 and 120 Å are of order 10^{24} cm^{-3} s^{-1} (Huang & Bartell, 1995; Manka *et al.*, 2012). Nucleation rates in micrometer-size drops near the bulk homogeneous nucleation temperature $T_h \approx 235\ \mathrm{K}$ are ~10^9 cm^{-3} s^{-1} (Murray *et al.*, 2010). These data spanning 14 orders of magnitude in nucleation rate over $\Delta T \approx 43\ \mathrm{K}$ have been fit with models of homogeneous nucleation (Murray *et al.*, 2010). Ice fractions of ~1% are obtained when micrometer-size water drops are cooled at ~10^5-10^6 K/s (Brüggeller & Mayer, 1980). Assuming 100 water molecules per nucleus (Huang & Bartell, 1995; Moore & Molinero, 2010) and that all ice is due to nucleation (i.e., no post-nucleation growth) gives a nucleation rate of order 10^{23} cm^{-3} s^{-1}, consistent with the peak temperature-dependent rate.

However, for water confined within ~400 μm apoferritin crystals, whose maximum solvent cavity size is 68 Å, the solvent can remain as a metastable liquid on ~10^0-10^2 second timescales following cooling to 200-230 K. When ice eventually becomes detectable, its diffraction intensity saturates in as little as 0.2 to 0.4 s. Assuming ice detection arises from a single nucleation event shortly before that detection, nucleation rates between 200 to 230 K are in the range of 10^6 cm^{-3} s^{-1}, ~10^{10} to 10^{17} times smaller than in

water nanodrops comparable in size to apoferritin's solvent cavities. The most conservative assumptions – that ice is first detectable when the ice fraction reaches 2%, that nucleation occurs steadily until that threshold is reached, that each nucleus involves 100 water molecules, and that nuclei don't grow - give a nucleation rate of ~10^{19} cm^{-3} s^{-1}, roughly four orders of magnitude smaller than in water nanodrops below 215 K. Grain sizes of internal ice in apoferritin crystals deduced from Rietveld refinement of diffraction patterns are in the range of ~200 to 800 Å, spanning many unit cells, so actual nucleation rates likely lie between these limits. The dramatically reduced nucleation rates under nanoconfinement we infer are qualitatively consistent with simulations (Li *et al.*, 2013) showing increasing suppression of nucleation in ~3 nm drops as temperatures increase above ~210 K. However, they contrast with much-larger-than-bulk nucleation rates deduced from NMR experiments on nanoporous silica with 12 nm cavities (Mascotto *et al.*, 2017).

3.3.2 Protein crystal diffraction quality is maximized near T=220 K, in crystals with liquid solvent

For ice-free apoferritin and thaumatin crystals, Wilson *B* factors (a measure of short-range crystal disorder strongly correlated with resolution) decrease to a minimum near ~220 K, and neither cooling to 100 K nor the use of glycerol provide clear improvements (Fig. 6.2.5). For both proteins, crystal mosaicities generally increase with decreasing temperature (Fig. 6.2.6) and are smaller by factors of ~2-6 at 220 K than at 100 K for all glycerol concentrations except 40% v/v. Between 200 K and 260 K, glycerol-free crystals of both proteins tend to have the lowest mosaicities.

3.3.3 Unit cell contraction on cooling is not determined by internal solvent contraction

For ice-free apoferritin and thaumatin crystals with all glycerol concentrations, unit cell volumes measured ~3-5 s after cooling contract monotonically between 300 K and 180 K (Fig. 3.5(a,b)). For both proteins, the contraction of the protein volume on cooling to 180 K is small (~0.5-1%) and nearly

independent of glycerol concentration (Fig. 3.5(c,d)). The solvent cavity volume contraction is much larger, and is only weakly dependent on glycerol concentration (Fig. 3.5(e,f)).

As will be discussed in more detail elsewhere (Moreau *et al.*){Chapter 4}, internal solvent contraction on cooling cannot be the primary driver of these unit cell and solvent cavity volume contractions. A simple model for protein crystal volume changes on cooling from an initial (*i*) to final (*f*) temperature is described by

$$v_{cell,f} + v_{exit} = v_{cell,i}(1+\Delta_{cell}) + v_{exit} \approx v_{p,i}\left(1+\Delta_p\right) + v_{sb,i}\left(1+\Delta_{s,b}\right) + v_{sh,i}\left(1+\Delta_{s,h}\right) \qquad (1)$$

Here, $v_{cell,i}$ and $v_{cell,f}$ are the initial and final unit cell volumes, v_{exit} is the amount of solvent that leaves the unit cell (due to differential thermal contraction of the cell, protein, and solvent) (Juers & Matthews, 2001; Kriminski *et al.*, 2002), $v_{p,i}$, $v_{sb,i}$, $v_{sh,i}$ are the initial volumes of protein, bulk-like solvent, and hydration (strongly perturbed) solvent, and Δ_{cell}, Δ_p, $\Delta_{s,b}$, and $\Delta_{s,h}$ are fractional changes in specific volumes on cooling. Pure bulk water expands by ~6% on cooling from a room temperature liquid to low-density amorphous (LDA) ice at 77 K, whereas a 40% (v/v) glycerol solution contracts by ~5% (Tyree *et al.*, 2018). The largest fractional specific volume changes for both protein and solvent occur between 300 K and 200 K. Assuming that $v_{exit} = 0$, that all solvent in apoferritin crystals has bulk-like volume contraction ($v_{s,h} = 0$) and the same glycerol concentration as the soak solution, and that Δ_p does not depend on glycerol concentration, the unit cell volume at 100 K for 40% (v/v) glycerol crystals should be ~7% smaller than for glycerol-free crystals. In fact, unit cell volumes for 40% glycerol apoferritin crystals are only 0.4% and 1.4% smaller than for glycerol-free crystals at 100 K and 200 K relative to room temperature. Similar discrepancies between expected and measured cell volumes are observed at temperatures between 180 K and 260 K.

The most plausible explanation for these discrepancies is that v_{exit} is not zero (Juers & Matthews, 2001; Kriminski *et al.*, 2002; Juers *et al.*, 2018), and that on cooling a substantial amount of solvent exits

(or enters) the ordered unit cells that contribute to Bragg diffraction (SI Sec. S5). With the most conservative assumptions, measured unit cell contractions to 100 K for apoferritin give v_{exit} ~ 9 % of the room-temperature solvent cavity volume for glycerol-free crystals, and $v_{exit} \sim -1.8\,\%$ for crystals soaked in 40 % (v/v) glycerol solutions (the negative sign implying that solvent must enter the unit cell). Similar results are obtained for thaumatin crystals. Table 3.1 gives estimates of the fraction of crystal solvent that exits the unit cells of glycerol-free crystals of apoferritin and thaumatin on cooling to temperatures between 180 K and 260 K.

These results indicate that unit cell contraction on cooling at rates up to ~1000 K/s is not appreciably modulated by internal solvent contraction or expansion. It is driven by the hydrated protein lattice, by the reduction in protein entropy that accompanies side chain ordering and formation of additional crystal contacts (Juers & Matthews, 2001), and perhaps also by reduced hydration layer solvent entropy.

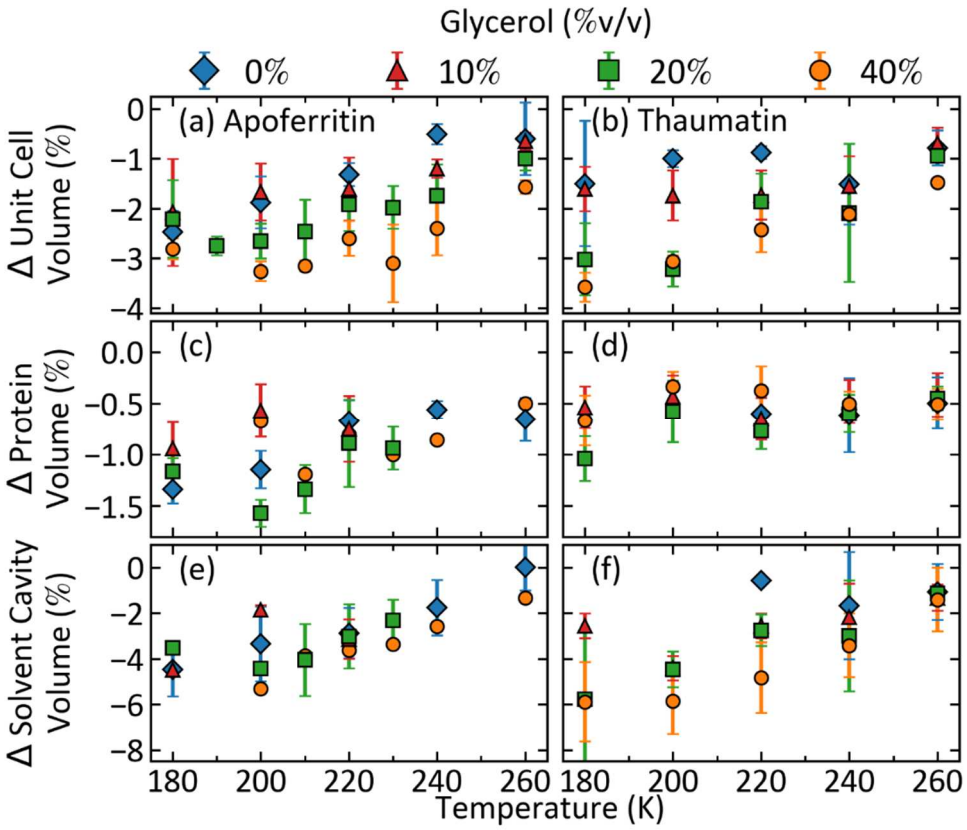


Figure 3.5 Changes in the unit cell, solvent cavity, and protein volumes from their room-temperature values versus temperature for apoferritin (left column) and thaumatin (right column).

Fractional changes from room temperature to temperature T:	Apoferritin – 0% glycerol				
	180	200	220	240	260
Change in solvent cavity volume	-5.3% ± 1.4%	-3.7% ± 1.5%	-2.3% ± 1.3%	0.5% ± 2.1%	0.7% ± 1.4%
Change in solvent volume, "bulk"	4.5% ± 1.5%	5.7% ± 1.6%	6.0% ± 1.6%	4.2% ± 1.5%	1.4% ± 1.5%
Change in solvent volume, "interface-perturbed"	3.3% ± 1.6%	4.4% ± 1.6%	4.6% ± 1.6%	3.3% ± 1.6%	1.0% ± 1.5%
$f_{exit}(180K)$ "bulk"	9.4% ± 1.3%	8.9% ± 1.4%	8.8% ± 1.0%	6.2% ± 1.2%	1.8% ± 1.0%
$f_{exit}(180K)$ "interface-perturbed"	8.6% ± 1.4%	8.0% ± 1.5%	7.7% ± 1.1%	5.4% ± 1.2%	1.5% ± 1.1%
Fractional changes from room temperature to temperature T:	Thaumatin – 0% glycerol				
	180	200	220	240	260
Change in solvent cavity volume	-2.3% *	-1.7% *	-0.6% ± 0.2%	-1.7% ± 2.4%	-1.1% ± 1.2%
Change in solvent volume, "bulk"	4.5% ± 0.2%	5.7% ± 0.2%	6.0% ± 0.2%	4.2% ± 0.2%	1.4% ± 0.2%
Change in solvent volume, "interface-perturbed"	2.7% ± 0.7%	3.5% ± 0.8%	3.8% ± 0.8%	3.6% ± 0.7%	3.3% ± 0.7%
$f_{exit}(180K)$ "bulk"	6.5% ± 0.8%	7.0% ± 0.8%	6.2% ± 0.2%	5.7% ± 2.3%	2.4% ± 1.2%
$f_{exit}(180K)$ "interface-perturbed"	5.0% ± 1.0%	5.2% ± 1.0%	4.2% ± 0.7%	4.3% ± 2.4%	1.9% ± 1.4%

Table 3.3.1 Fractional changes on cooling from room temperature to each indicated temperature in solvent cavity volume and solvent volume (assuming bulk and interface-perturbed solvent contractions), and the fraction of solvent that must exit the unit cell, for apoferritin and thaumatin crystals, calculated as described in Sec. 6.2.5. The volume fractions of the room temperature unit cell occupied by solvent cavities, determined from structural models using the program map_channels are 63.4% and 60.2% for apoferritin and thaumatin, respectively.

3.3.4 Internal ice in protein crystals is stacking disordered

Ice diffraction is routinely observed in protein crystallography, and can arise both from internal solvent and from external solvent surrounding the crystal. An analysis of PDB-deposited data (generally, the best data obtained in a given set of experiments) found evidence of errors in protein crystal structure factors consistent with contamination by (and incomplete modeling of) ice in roughly 20% of entries (Thorn *et al.*, 2017). Ice diffraction from protein crystals has been discussed in terms of ideal hexagonal ice I_h, cubic ice I_c, and low-density amorphous ice I_{LDA} patterns.

In the present experiments, we attempted to remove all external solvent. For glycerol concentrations of 0% and 10% (v/v) and temperatures from 180 to 230 K, roughly 70% of apoferritin crystals (70 of 98) eventually formed ice. Only three showed azimuthally integrated diffraction patterns consistent with pure I_h. Another 26 crystals showed diffraction peaks at all expected I_h positions, but with peak shapes and intensities that were inconsistent both with pure I_h and any simple mixture of I_h and I_c grains (Fig. 3.6(a,d)). For all 29 of these crystals, ice formed within ~2 s of the start of cooling, suggesting that it nucleated during cooling, and many of the raw diffraction patterns had a component that was azimuthally "lumpy" (Fig. 3.6(a)), indicating a small number of large ice grains. In some cases, residual frozen solvent on the crystal surface was clearly visible. We thus attribute the appearance of hexagonal ice as arising from nucleation in residual external solvent, not internal solvent.

By far the most common patterns of ice diffraction, observed in 41 of 98 apoferritin crystals and 36 of 51 thaumatin crystals between 180 K and 230 K, consisted of a strong but broadened peak near d = 3.7 Å, a weaker broad peak near d = 3.0 Å, and broadened peaks near 2.2 and 1.9 Å (Fig. 3.6(e,f)). The I_h peaks near 3.5 Å and 2.1 Å were absent or strongly suppressed, and near the position of the 3.5 Å peak a smooth shoulder was observed instead. Ice diffraction in this case was always uniform and isotropic (Fig. 3.6(b,c)), indicating a large number of small, randomly oriented grains. For the glycerol-free apoferritin crystals that

developed these ice diffraction patterns, the mean time to ice formation was 6.4 s and the standard deviation was 8.3 s, suggesting delayed and stochastic nucleation. These systematics together with estimated freezing temperatures for bulk-like internal solvent indicate that the ice arises from nucleation within deeply supercooled internal solvent. Poor fits to the diffraction patterns obtained in refinement indicate that it is neither cubic nor hexagonal nor a simple mix of the two.

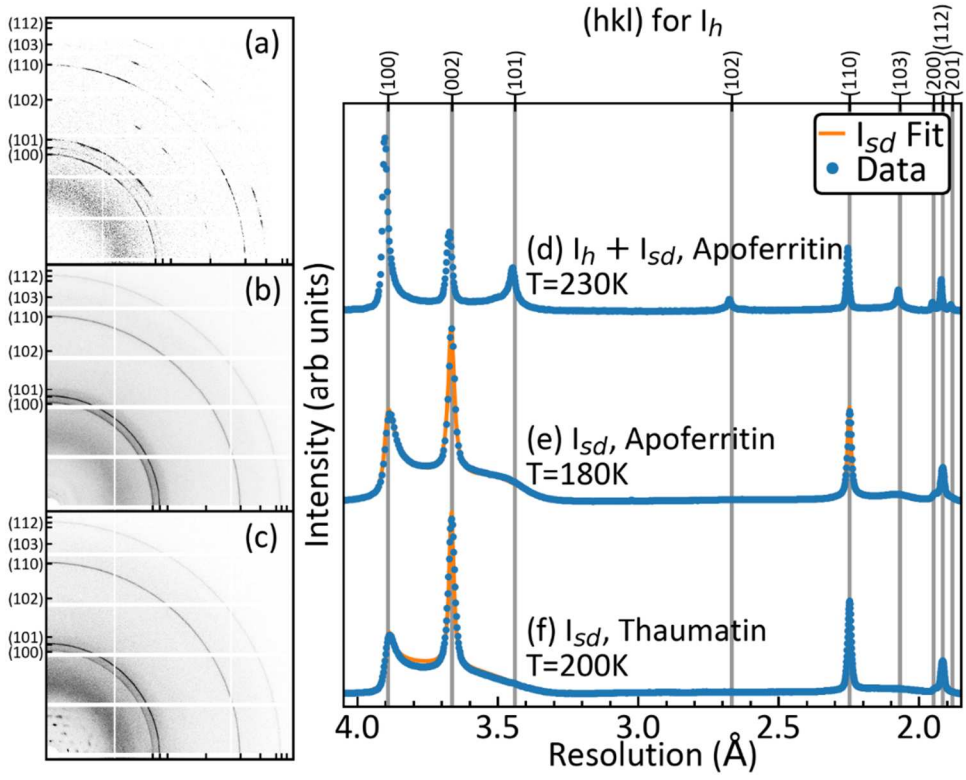


Figure 3.6 (a-c) Examples of detector images showing ice diffraction. (a) Mix of hexagonal ice Ih and stacking disordered ice Isd in a glycerol-free apoferritin crystal at 230 K. (b) Isd in glycerol-free apoferritin at 180 K. (c) Isd in a glycerol-free thaumatin crystal at 200 K. (d-f) Dotted blue lines indicate azimuthally integrated and background subtracted ice ring diffraction profiles calculated from the detector images in (a-c). Solid orange lines are best-fit profiles calculated using the program DIFFaX.

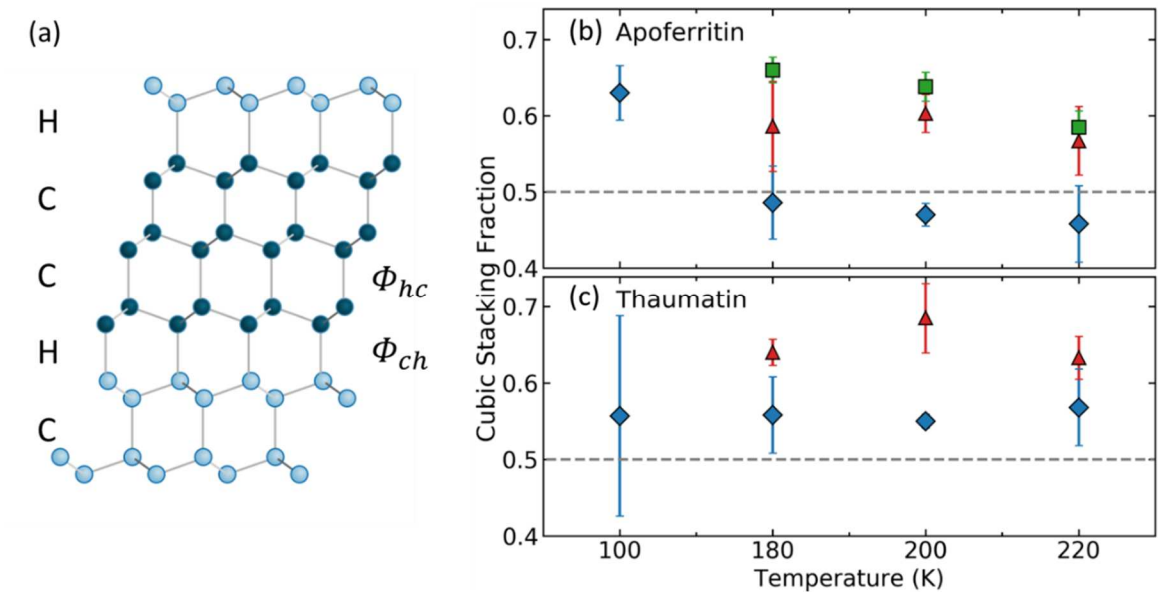


Figure 3.7 (a) Model for stacking disordered ice I_{sd} (adapted from (Malkin et al., 2012)), showing oxygen atoms connected by H bonds. Along the vertical direction (normal to (0001) planes in I_h and to (111) planes in I_c), successive cubic ice planes are horizontally shifted, whereas successive hexagonal planes are mirror-reflected about a horizontal axis. At left, C and H indicate pairs of planes that have cubic and hexagonal stacking, respectively. At right, Φ_{ch} indicates the probability that a cubic stacking is followed by hexagonal stacking, and Φ_{hc} that a hexagonal stacking is followed by cubic stacking. (b),(c) Cubic stacking fraction $\Phi_{hc}/(\Phi_{hc}+\Phi_{ch})$ versus temperature for (b) apoferritin and (c) thaumatin crystals, determined from DIFFaX fits as in Fig. 3.6 (e) and (f). Symbols are samples with different glycerol concentrations as in Fig. 3.5.

Similar diffraction patterns have been observed in experiments on ice formed on cooling in deeply supercooled water micro- and nanodrops (Morishige & Uematsu, 2005; Malkin *et al.*, 2012, 2015; Kuhs *et al.*, 2012), on abrupt warming from a vitrified state, and also in molecular dynamics simulations (Moore & Molinero, 2011; Hudait *et al.*, 2016). They have been attributed to a disordered stacking of cubic and hexagonal ice planes along the hexagonal *c* direction (Fig. 7(a)).

In a simple model for this "stacking-disordered" ice, I_{sd} (Malkin *et al.*, 2015), the probability of a cubic plane being followed by a hexagonal plane is Φ_{ch} and of a hexagonal plane being followed by a cubic plane is Φ_{hc} (Fig. 3.7(a)). The solid lines in Fig. 3.6(e,f) are refined ice diffraction fits assuming this model (Sec. 6.2.7), calculated using the program DIFFaX (Treacy *et al.*, 1991). At all temperatures and all glycerol concentrations at which internal ice formed, the fits provide an excellent account of the observed diffraction. Fig. 3.7 shows that between 180 K and 220 K, the fraction of cubic layers is near 0.5, the value for purely random stacking, for glycerol-free crystals of both proteins, and increases with glycerol concentration. Fit quality depends sensitively on cubic stacking fraction (Malkin *et al.*, 2015), and the small crystal-to-crystal variance (indicated by error bars in Fig. 3.7 (b) and (c)) for the substantial number of crystals analysed (Table 3.1.1) indicates that this stacking fraction is robust. A near random stacking of cubic and hexagonal layers may result because the free energies of nucleating cubic and hexagonal ice planes on an ice crystal surface are similar, so that competitive nucleation under conditions of deep supercooling leads to stacking dominated by kinetics rather than thermodynamics (Malkin *et al.*, 2015). The delayed formation of stacking-disordered ice within protein crystals is thus consistent with the existence of liquid, deeply supercooled internal solvent at temperatures down to ~200 K.

Experiments on water nanoconfined within nanoporous silicas (MCM-41 and SBA-15) (Baker *et al.*, 1997; Morishige & Uematsu, 2005) and ordered nanoporous aluminium oxide membranes (Suzuki, Duran *et al.*, 2015) observed evidence of stacking disordered ice only when pore diameters were larger than ~200 Å and 350 Å, respectively; for smaller pores internal ice was largely cubic. We observe disordered stacking with near equal fractions of cubic and hexagonal planes in apoferritin and thaumatin crystals with

maximum pore diameters of 68 and 25 Å. Accurate modelling of diffraction from internal ice in protein crystals, with its peak-dependent and asymmetric broadening, should improve estimates of protein crystal structure factors and the accuracy of structural models.

3.3.5 Protein crystals enable novel quantitative estimates of crystallizable internal solvent fractions and perturbed interfacial layer thicknesses.

Water's structure and dynamics are locally perturbed by hydrogen bonding and other interactions with solutes including proteins (Svergun *et al.*, 1998), and with interfaces including the confining walls of nanoporous systems (Liu *et al.*, 2008; Erko *et al.*, 2012; Taschin *et al.*, 2015). Several experimental criteria (Bagchi, 2005) have been used to classify and quantify the fractions of locally perturbed and bulk-like water as a function of, e.g., solute concentration or pore size. Perturbed layer thicknesses of ~3-7 Å, in general agreement with simulations, are typically found, with substantial uncertainties arising from models used to fit, e.g., NMR lineshapes or calorimetric data. These compare with nominal thicknesses of the first and the first two hydration layers of 2.8 Å and 5.6 Å, respectively.

One metric of water's perturbation is its ability to participate in a crystalline network (Sartor *et al.*, 1995; Rault *et al.*, 2003). Protein crystals enable a novel and highly quantitative approach to determining non-crystallizable solvent fractions (Sec. 6.2.8). Unlike in the most widely studied nanoconfining systems, the confining matrix provided by crystals of proteins like apoferritin and thaumatin is nearly perfectly periodic and has excellent long-range order in all three dimensions, indicated by sharp diffraction peaks having negligible strain broadening and mosaic broadening as small as 0.003° (comparable to silicon). Single crystal diffraction from this matrix and powder diffraction from internal ice confined within it can be recorded using the same X-ray beam and detector. Bragg diffraction from the protein matrix can be crystallographically modelled to determine its full atomic structure. Comparison of measured diffraction intensities with refined model structure factors yields a quantity related to the X-ray illuminated volume

of the crystal. Powder diffraction from internal ice, recorded after its intensity reaches steady state, can be modelled using DIFFaX. Comparison of measured ice diffraction intensities with those from a refined ice model yields a quantity related to the total X-ray illuminated volume of ice (Secs. 6.2.8). The ratio of these quantities from ice and protein crystal diffraction then gives the fraction of the X-ray illuminated volume occupied by ice. Using the refined crystallographic models for the protein lattice, the fraction of the unit cell occupied by solvent can be determined. This allows the fraction of internal solvent that forms ice to be determined to high accuracy.

As shown in Tables 3.2 and 6.2.2, the resulting maximum crystallized solvent volume fractions at T=180-220 K in glycerol-free crystals of apoferritin, thaumatin, and lysozyme are ~59%, 35%, and 17%, respectively. Accounting for possible solvent outflow from the unit cell due to differential contraction of solvent and solvent cavities on cooling (Sec. 6.2.5, Table 3.1), these decrease to 49%, 25%, and 8%. The large difference in crystallizable solvent fraction between apoferritin and thaumatin crystals occurs despite only a 4% difference in their solvent contents. This difference may explain why apoferritin crystals lose nearly all ordered protein diffraction when ice forms, while thaumatin (and especially lysozyme) crystals continue to diffract to moderate resolution. The large crystallized solvent fractions for apoferritin and thaumatin confirm that the observed ice diffraction is from internal solvent: an external solvent volume with roughly 1/4 the volume of a ~200-400 μm protein crystal would be naked-eye visible and would immediately crystallize to form hexagonal ice on cooling.

By comparing these crystallizable solvent fractions with the cumulative distribution of solvent distances from the protein surface in each crystal (Fig. 6.2.2), the thickness of the layer of non-crystallizable solvent adjacent to the protein surface can be determined. This thickness is of order 5 Å for all three proteins (Table 3.2), comparable to the thickness of first two hydration shells, and consistent with approximate values estimated from studies of ice formation in hydrated protein powders (Sartor *et al.*, 1995) and in porous inorganic glasses (Rault *et al.*, 2003).

Fraction of solvent exiting the unit cell	Fraction of internal solvent that forms ice		
	Apoferritin	Thaumatin	Lysozyme
(a) $f_{exit} = 0$	59% ± 13%	35% ± 6%	17% ± 5%
(b) f_{exit}, bulk	49% ± 13%	25% ± 6%	8% ± 6%
(c) f_{exit}, perturbed	50% ± 13%	29% ± 6%	12% ± 6%
Fraction of solvent >2.8Å from protein	83%	70%	29%
Fraction of solvent >5.6Å from protein	*55%*	*31%*	*0.5%*

Table 3.3.2 Estimates of the maximum fraction of the solvent cavity space occupied by ice in glycerol-free crystals of apoferritin, thaumatin, and lysozyme at temperatures between 180 and 220 K, as described in Sec. 6.2.8.

3.4 CONCLUSIONS

We have connected the behavior of water and ice in protein crystals to results from previous studies of water in nanoporous inorganic matrices and in micro- and nano-drops. Protein crystals have significant advantages over other nanoconfining systems. The ~100,000 known protein crystal structures offer tremendous variety in pore size, pore geometry (including relatively simple geometries as found in apoferritin), and chemical properties. A majority have excellent long-range order, so the full atomic structure of the confining matrix is known from crystallography and available for simulations.

Ordered inorganic nanoporous matrices are typically synthesized as micrometer-size powders, having large ratios of external surface area to volume, significant (5-10%) pore size variations, and substantial defect densities. Measurements on nanoconfined solvent require use of packed powders, often filled from the vapor phase to minimize internal bubbles and overfilling, introducing substantial uncertainties and restricting study to pure water and other volatile liquids. Ice formation in solvent on the surface of individual grains likely corrupts measurements of ice nucleation rates.

In contrast, protein crystals are typically tens to hundreds of micrometers in size. Single crystals are sufficient for many measurements, and surface solvent can be optically detected and removed. X-ray diffraction methods demonstrated here should allow ice nucleation and growth, grain sizes, and crystallized solvent fractions to be tracked during and following cooling in single crystals having <100 ms thermal response times. The composition of a crystal's internal solvent can be changed by serial soaking in aqueous solutions containing salts, sugars, alcohols, and polyols, including at concentrations (including nearly pure water) which cause crystal dissolution and/or protein unfolding, and often still retain excellent order, and the effects of, e.g., preferential hydration of protein surfaces and solute rejection by growing ice crystals on crystallizable solvent fractions determined. These features make protein crystals attractive model systems for studying the effects of confinement on ice formation, especially under biophysically relevant conditions.

Even in protein crystals with ~70 Å solvent cavities - larger than in ~98% of current Protein Data Bank deposits, nanoconfinement dramatically suppresses freezing temperatures (to ≤240 K) and ice nucleation rates, the latter allowing internal solvent to remain as a (supercooled) liquid for at least several seconds at temperatures between 200 and 240 K. By combining abrupt *in situ* cooling with intense synchrotron X-ray beams and fast X-ray detectors, complete structural data sets for high-value targets including membrane proteins and large complexes may be collected at ~220 K that have much lower mosaicities, comparable B factors, and that may allow more confident identification of ligand binding than in current cryocrystallographic practice. This same crystal-based strategy of abrupt cooling and fast data collection before ice formation may enable a variety of temperature-dependent biophysical studies of protein structure, conformational ensembles, and function, including at temperatures near T_h that are inaccessible when studying proteins in solution or *in vivo*.

Funding information National Science Foundation, Directorate for Biological Sciences (award No. MCB-1330685 to Cornell University); National Institutes of Health, National Institute of General Medical Sciences (award No. T32GM0082567 to Molecular Biophysics Training Grant, Cornell University; award No. R01-GM127528 to Cornell University); National Science Foundation, Division of Materials Research (award No. DMR-0936384 to Cornell High-Energy Synchrotron Source, Cornell University); National Institutes of Health, National Institute of General Medical Sciences (award No. GM-103485 to Macromolecular Diffraction Facility at CHESS, Cornell University).

Acknowledgements All X-ray data collection was performed at the Cornell High-Energy Synchrotron Source (CHESS), which is supported by the NSF under award DMR-0936384, using the Macromolecular Diffraction at CHESS (MacCHESS) facility, which is supported by NIH award GM-103485. We thank James Holton for discussions.

3.5 REFERENCES

Adams, P. D., Afonine, P. V., Bunkóczi, G., Chen, V. B., Davis, I. W., Echols, N., Headd, J. J., Hung, L. W., Kapral, G. J., Grosse-Kunstleve, R. W., McCoy, A. J., Moriarty, N. W., Oeffner, R., Read, R. J., Richardson, D. C., Richardson, J. S., Terwilliger, T. C. & Zwart, P. H. (2010). *Acta Cryst. D*. **66**, 213–221.

Alcorn, T. & Juers, D. H. (2010). *Acta Cryst. D*. **66**, 366–373.

Amaya, A. J., Pathak, H., Modak, V. P., Laksmono, H., Loh, N. D., Sellberg, J. A., Sierra, R. G., McQueen, T. A., Hayes, M. J., Williams, G. J., Messerschmidt, M., Boutet, S., Bogan, M. J., Nilsson, A., Stan, C. A. & Wyslouzil, B. E. (2017). *J. Phys. Chem. Lett.* **8**, 3216–3222.

Ashiotis, G., Deschildre, A., Nawaz, Z., Wright, J. P., Karkoulis, D., Picca, F. E. & Kieffer, J. (2015). *J. Appl. Crystallogr.* **48**, 510–519.

Atakisi, H., Moreau, D. W. & Thorne, R. E. (2018). *Acta Cryst. D*. **74**, 264–278.

Bagchi, B. (2005). *Chem. Rev.* **105**, 3197–3219.

Baker, J. M., Dore, J. C. & Behrens, P. (1997). *J. Phys. Chem. B*. **101**, 6226–6229.

Barbosa, R. D. C. & Barbosa, M. C. (2015). *Phys. A Stat. Mech. Its Appl.* **439**, 48–58.

Bartell, L. S. & Chushak, Y. G. (2003). *Water in Confining Geometries*, Vol. pp. 399–424.

van den Bedem, H., Dhanik, A., Latombe, J. C. & Deacon, A. M. (2009). *Acta Cryst. D*. **65**, 1107–1117.

Brüggeller, P. & Mayer, E. (1980). *Nature*. **288**, 569–571.

Charron, C., Kadri, A., Robert, M., Giege, R. & Lorber, B. (2002). *Acta Cryst. D*. **58**, 2060–2065.

Chen, V. B., Arendall, W. B., Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., Murray, L. W., Richardson, J. S. & Richardson, D. C. (2010). *Acta Cryst. D*. **66**, 12–21.

Chukin, V. V., Pavlenko, E. a. & Platonova, a. S. (2010). *Russ. Meteorol. Hydrol.* **35**, 524–529.

Crichton, R. R. & Declercq, J. P. (2010). *Biochim. Biophys. Acta*. **1800**, 706–718.

Datta, S., Biswal, B. K. & Vijayan, M. (2001). *Acta Cryst. D*. **57**, 1162–1167.

Doster, W. (2010). *BBA - Proteins Proteomics*. **1804**, 3–14.

Douzou, P., Hoa, G. H. & Petsko, G. a (1975). *J. Mol. Biol.* **96**, 367–380.

Ebbinghaus, S., Kim, S. J., Heyden, M., Yu, X., Heugen, U., Gruebele, M., Leitner, D. M. & Havenith, M. (2007). *Proc. Natl. Acad. Sci. U. S. A.* **104**, 20749–20752.

Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. (2010). *Acta Cryst. D*. **66**, 486–501.

Erko, M., Wallacher, D., Hoell, a., Hauß, T., Zizak, I. & Paris, O. (2012). *Phys. Chem. Chem. Phys.* **14**, 3852.

Espinosa, J. R., Navarro, C., Sanz, E., Valeriani, C. & Vega, C. (2016). *J. Chem. Phys.* **145**,.

Espinosa, J. R., Sanz, E., Valeriani, C. & Vega, C. (2014). *J. Chem. Phys.* **141**,.

Fahy, G. M. & Wowk, B. (2015). *Cryopreservation and Freeze-Drying Protocols*, Vol. *1257*, edited by W.F. Wolkers & H. Oldenhof, pp. 21–82. New York: Springer.

Fenimore, P. W., Frauenfelder, H., McMahon, B. H. & Young, R. D. (2004). *Proc. Natl. Acad. Sci. U. S. A.* **101**, 14408–14413.

Findenegg, G. H., Jähnert, S., Akcakayiran, D. & Schreiber, A. (2008). *ChemPhysChem*. **9**, 2651–2659.

Fokine, A. & Urzhumtsev, A. (2002). *Acta Cryst. D*. **58**, 1387–1392.

Fraser, J. S., van den Bedem, H., Samelson, A. J., Lang, P. T., Holton, J. M., Echols, N. & Alber, T. (2011). *Proc. Natl. Acad. Sci. U. S. A.* **108**, 16247–16252.

Fraser, J. S., Clarkson, M. W., Degnan, S. C., Erion, R., Kern, D. & Alber, T. (2009). *Nature*. **462**, 669–673.

Frauenfelder, H., Hartmann, H., Karplus, M., Kuntz, I. D., Kuriyan, J., Parak, F., Petsko, G. a, Ringe, D., Tilton, R. F. & Connolly, M. L. (1987). *Biochemistry*. **26**, 254–261.

Frauenfelder, H., Petsko, G. A. & Tsernoglou, D. (1979). *Nature*. **280**, 558–563.

Garman, E. (2003). *Curr. Opin. Struct. Biol.* **13**, 545–551.

Garman, E. F. & Schneider, T. R. (1997). *J. Appl. Crystallogr.* **30**, 211–237.

Halle, B. (2004). *Proc. Natl. Acad. Sci.* **101**, 4793–4798.

Hansen, E. W., Stöcker, M. & Schmidt, R. (1996). *J. Phys. Chem.* **100**, 2195–2200.

Hansen, T. C., Koza, M. M., Lindner, P. & Kuhs, W. F. (2008*a*). *J. Phys. Condens. Matter*. **20**, 285104.

Hansen, T. C., Koza, M. M., Lindner, P. & Kuhs, W. F. (2008*b*). *J. Phys. Condens. Matter*. **20**, 285104.

Hare, D. E. & Sorensen, C. M. (1987). *J. Chem. Phys.* **87**, 4840–4845.

Holten, V. & Anisimov, M. A. (2012). *Sci. Rep.* **2**, 1–7.

Huang, J. & Bartell, L. S. (1995). *J. Phys. Chem.* **99**, 3924–3931.

Hudait, A., Qiu, S., Lupi, L. & Molinero, V. (2016). *Phys. Chem. Chem. Phys.* **18**, 9544–9553.

Jahn, D. A., Wong, J., Bachler, J., Loerting, T. & Giovambattista, N. (2016). *Phys. Chem. Chem. Phys.* **18**, 11042–11057.

Jähnert, S., Vaca Chávez, F., Schaumann, G. E., Schreiber, A., Schönhoff, M. & Findenegg, G. H. (2008). *Phys. Chem. Chem. Phys.* **10**, 6039.

Juers, D. H., Farley, C. A., Saxby, C. P., Cotter, R. A., Cahn, J. K. B., Holton-Burke, R. C., Harrison, K. & Wu, Z. (2018). *Acta Crystallogr. Sect. D Struct. Biol.* **74**, 922–938.

Juers, D. H. & Matthews, B. W. (2001). *J. Mol. Biol.* **311**, 851–862.

Juers, D. H. & Matthews, B. W. (2004*a*). *Q. Rev. Biophys.* **37**, 105–119.

Juers, D. H. & Matthews, B. W. (2004*b*). *Acta Cryst. D*. **60**, 412–421.

Juers, D. H. & Ruffin, J. (2014). *J. Appl. Crystallogr.* **47**, 2105–2108.

Kabsch, W. (2010). *Acta Cryst. D*. **66**, 125–132.

Kantardjieff, K. A. & Rupp, B. (2003). *Protein Sci.* **12**, 1865–1871.

Keedy, D. A., van den Bedem, H., Sivak, D. A., Petsko, G. A., Ringe, D., Wilson, M. A. & Fraser, J. S. (2014). *Structure*. **22**, 1–12.

Keedy, D. A., Kenner, L. R., Warkentin, M., Woldeyes, R. A., Hopkins, J. B., Thompson, M. C., Brewster, A. S., Van Benschoten, A. H., Baxter, E. L., Uervirojnangkoorn, M., McPhillips, S. E., Song, J., Alonso-Mori, R., Holton, J. M., Weis, W. I., Brunger, A. T., Soltis, S. M., Lemke, H., Gonzalez, A., Sauter, N. K., Cohen, A. E., Van Den Bedem, H., Thorne, R. E. & Fraser, J. S. (2015). *Elife*. **4**, 07574.

Kittaka, S., Ishimaru, S., Kuranishi, M., Matsuda, T. & Yamaguchi, T. (2006). *Phys. Chem. Chem. Phys.* **8**, 3223.

Knudsen, E. B., Sørensen, H. O., Wright, J. P., Goret, G. & Kieffer, J. (2013). *J. Appl. Crystallogr.* **46**, 537–539.

Kriminski, S., Caylor, C. L., Nonato, M. C., Finkelstein, K. D. & Thorne, R. E. (2002). *Acta Cryst. D*. **58**,.

Kuffel, A. & Zielkiewicz, J. (2012). *J. Phys. Chem. B*. **116**, 12113–12124.

Kuhs, W., Bliss, D. & Finney, J. (1987). *J. Phys. Colloq.* **48**, 631–636.

Kuhs, W. F., Sippel, C., Falenty, A. & Hansen, T. C. (2012). *Proc. Natl. Acad. Sci. USA*. **109**, 21259–21264.

Lane, L. B. (1925). *Ind. Eng. Chem.* **17**, 924–924.

Lang, P. T., Holton, J. M., Fraser, J. S. & Alber, T. (2014). *Proc. Natl. Acad. Sci. USA*. **111**, 237–242.

Lee, J., Maj, M., Kwak, K. & Cho, M. (2014). *J. Phys. Chem. Lett.* **5**, 3404–3407.

Li, A. J. & Nussinov, R. (1998). *Proteins Struct. Funct. Genet.* **32**, 111–127.

Li, T., Donadio, D. & Galli, G. (2013). *Nat. Commun.* **4**, 1–6.

Liu, D., Zhang, Y., Liu, Y., Wu, J., Chen, C. C., Mou, C. Y. & Chen, S. H. (2008). *J. Phys. Chem. B*. **112**, 4309–4312.

Liu, X. X., Wang, Q., Huang, X. F., Yang, S. H., Li, C. X., Niu, X. J., Shi, Q. F., Sun, G. & Lu, K. Q. (2010). *J. Phys. Chem. B*. **114**, 4145–4150.

Loerting, T., Bauer, M., Kohl, I., Watschinger, K., Winkel, K. & Mayer, E. (2011). *J. Phys. Chem. B*. **115**, 14167–14175.

Malkin, T. L., Murray, B. J., Andrey, V., Anwar, J., Salzmann, C. G., Malkin, T. L., Murray, B. J., Brukhno, A. V, Anwar, J. & Salzmann, C. G. (2012). *Proc. Natl. Acad. Sci.* **109**, 1041–1045.

Malkin, T. L., Murray, B. J., Salzmann, C. G., Molinero, V., Pickering, S. J. & Whale, T. F. (2015). *Phys. Chem. Chem. Phys.* **17**, 60–76.

Manka, A., Pathak, H., Tanimura, S., Wölk, J., Strey, R. & Wyslouzil, B. E. (2012). *Phys. Chem. Chem. Phys.* **14**, 4505.

Mascotto, S., Janke, W. & Valiullin, R. (2017). *J. Phys. Chem. C*. **121**, 23788–23792.

Matthews, B. W. (1974). *J. Mol. Biol.* **82**, 513–526.

Merzel, F. & Smith, J. C. (2002*a*). *Proc. Natl. Acad. Sci. U. S. A.* **99**, 5378–5383.

Merzel, F. & Smith, J. C. (2002*b*). *Acta Crystallogr. Sect. D Biol. Crystallogr.* **58**, 242–249.

Merzel, F. & Smith, J. C. (2005). *J. Chem. Inf. Model.* **45**, 1593–1599.

Miyatou, T., Ohashi, R., Ida, T., Kittaka, S. & Mizuno, M. (2016). *Phys. Chem. Chem. Phys.* **18**, 18555–18562.

Moore, E. B., de la Llave, E., Welke, K., Scherlis, D. a & Molinero, V. (2010). *Phys. Chem. Chem. Phys.* **12**, 4124–4134.

Moore, E. B. & Molinero, V. (2010). *J. Chem. Phys.* **132**, 1–11.

Moore, E. B. & Molinero, V. (2011). *Phys. Chem. Chem. Phys.* **13**, 20008–20016.

Moreau, D. W., Atakisi, H. & Thorne, R. E. *Acta Cryst. D (to Be Submitt.*

Morishige, K. & Kawano, K. (1999). *J. Chem. Phys.* **110**, 4867–4872.

Morishige, K. & Nobuoka, K. (1997). *J. Chem. Phys.* **107**, 6965–6969.

Morishige, K. & Uematsu, H. (2005). *J. Chem. Phys.* **122**,.

Murray, B. J., Broadley, S. L., Wilson, T. W., Bull, S. J., Wills, R. H., Christenson, H. K. & Murray, E. J. (2010). *Phys. Chem. Chem. Phys.* **12**, 10380.

Nakasako, M. (2004). *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **359**, 1191-1204; discussion 1204-1206.

Oliphant, T. E. (2007). *Comput. Sci. Eng.* **9**, 10–20.

Parsegian, V. A., Rand, R. P. & Rau, D. C. (2000). *Proc. Natl. Acad. Sci. U. S. A.* **97**, 3987–3992.

Persson, F., Söderhjelm, P. & Halle, B. (2018). *J. Chem. Phys.* **148**, 215101.

Petrov, O. & Furó, I. (2011). *Phys. Chem. Chem. Phys.* **13**, 16358.

Pflugrath, J. W. (2015). *Acta Cryst. F*. **71**, 622–642.

Pozharski, E., Weichenberger, C. X. & Rupp, B. (2013). *Acta Cryst. D*. **69**, 150–167.

Rault, J., Neffati, R. & Judeinstein, P. (2003). *Eur. Phys. J. B - Condens. Matter*. **36**, 627–637.

Riechers, B., Wittbracht, F., Hütten, A. & Koop, T. (2013). *Phys. Chem. Chem. Phys.* **15**, 5873.

Ringe, D. & Petsko, G. A. (2003). *Biophys. Chem.* **105**, 667–680.

Rodgers, D. W. (1994). *Structure*. **2**, 1135–1140.

Rupp, B. (2009). Biomolecular Crystallography: Principles, Practice, and Application to Structural Biology. Garland Science.

Saraswathi, N. T., Sankaranarayanan, R. & Vijayan, M. (2002). *Acta Cryst. D*. **58**, 1162–1167.

Sartor, G., Hallbrucker, A. & Mayer, E. (1995). *Biophys. J.* **69**, 2679–2694.

Schirò, G., Caronna, C., Natali, F., Koza, M. M. & Cupane, A. (2011). *J. Phys. Chem. Lett.* **2**, 2275–2279.

Schirò, G., Fichou, Y., Gallat, F.-X., Wood, K., Gabel, F., Moulin, M., Härtlein, M., Heyden, M., Colletier, J.-P., Orecchini, A., Paciaroni, A., Wuttke, J., Tobias, D. J. & Weik, M. (2015). *Nat. Commun.* **6**, 6490.

Schmidt, R., Hansen, E. W., Stoecker, M., Akporiaye, D. & Ellestad, O. H. (1995). *J. Am. Chem. Soc.* **117**, 4049–4056.

Schreiber, A., Ketelsen, I. & Findenegg, G. H. (2001). *Phys. Chem. Chem. Phys.* **3**, 1185–1195.

Segur, J. (1946). *Heat. Vent.* **44**, 86.

Shen, C., Julius, E. F., Tyree, T. J., Moreau, D. W., Atakisi, H. & Thorne, R. E. (2016). *Acta Cryst. D*. **72**, 742–752.

Shimizu, S. & Smith, D. J. (2004). *J. Chem. Phys.* **121**, 1148–1154.

Sinibaldi, R., Ortore, M. G., Spinozzi, F., Carsughi, F., Frielinghaus, H., Cinelli, S., Onori, G. & Mariani, P. (2007). *J. Chem. Phys.* **126**, 235101.

Solveyra, E. G., de la Llave, E., Scherlis, D. a & Molinero, V. (2011). *J. Phys. Chem. B*. **115**, 14196–14204.

Suzuki, Y., Duran, H., Steinhart, M., Kappl, M., Butt, H. J. & Floudas, G. (2015). *Nano Lett.* **15**, 1987–1992.

Suzuki, Y., Steinhart, M., Butt, H. J. & Floudas, G. (2015). *J. Phys. Chem. B*. **119**, 11960–11966.

Svergun, D. I., Richard, S., Koch, M. H., Sayers, Z., Kuprin, S. & Zaccai, G. (1998). *Proc. Natl. Acad. Sci. U. S. A.* **95**, 2267–2272.

Taschin, A., Bartolini, P., Marcelli, A., Righini, R. & Torre, R. (2015). *J. Phys. Condens. Matter*. **27**, 194107.

Teeter, M. M., Yamano, A., Stec, B. & Mohanty, U. (2001). *Proc. Natl. Acad. Sci.* **98**, 11242–11247.

Thorn, A., Parkhurst, J., Emsley, P., Nicholls, R. A., Vollmar, M., Evans, G. & Murshudov, G. N. (2017). *Acta Cryst. D*. **73**, 729–737.

Tilton, R. F., Dewan, J. C. & Petsko, G. A. (1992). *Biochemistry*. **31**, 2469–2481.

Timasheff, S. N. (2002). *Biochemistry*. **41**, 13473–13482.

Toby, B. H. & Von Dreele, R. B. (2013). *J. Appl. Crystallogr.* **46**, 544–549.

Treacy, M. M. J., Newsam, J. M. & Deem, M. W. (1991). *Proc. Roy. Soc. A*. **433**, 499–520.

Tyree, T. J., Dan, R. & Thorne, R. E. (2018). *Acta Cryst. D*. **74**, 471–479.

Vekilov, P. G., Monaco, L. A., Thomas, B. R., Stojanoff, V. & Rosenberger, F. (1996). *Acta Crystallogr. Sect. D Biol. Crystallogr.* **52**, 785–798.

Voss, N. R. & Gerstein, M. (2010). *Nucleic Acids Res.* **38**, 555–562.

Waasmaier, D. & Kirfel, A. (1995). *Acta Cryst. A*. **51**, 416–431.

Wang, Q., Zhao, L., Li, C. & Cao, Z. (2016). *Sci. Rep.* **6**, 26831.

Warkentin, M. A., Sethna, J. P. & Thorne, R. E. (2013). *Phys. Rev. Lett.* **110**, 015703.

Warkentin, M., Badeau, R., Hopkins, J. B., Mulichak, A. M., Keefe, L. J. & Thorne, R. E. (2012). *Acta Crystallogr. Sect. D Biol. Crystallogr.* **68**,.

Warkentin, M., Berejnov, V., Husseini, N. S. & Thorne, R. E. (2006). *J. Appl. Crystallogr.* **39**, 805–811.

Warkentin, M. & Thorne, R. E. (2009). *J. Appl. Crystallogr.* **42**, 944–952.

Warkentin, M. & Thorne, R. E. (2010*a*). *Acta Cryst. D*. **66**, 1092–1100.

Warkentin, M. & Thorne, R. E. (2010*b*). *J. Struct. Funct. Genomics*. **11**, 85–89.

Warren, B. E. (1990). X-ray Diffraction New York: Dover.

Weik, M. (2003). *Eur. Phys. J. B*. **12**, 153–158.

Weik, M., Schreurs, A. M. M., Leiros, H. K. S. K. S., Zaccai, G., Ravelli, R. B. G. & Gros, P. (2005). *J. Synchrotron Radiat.* **12**, 310–317.

Yao, Y., Ruckdeschel, P., Graf, R., Butt, H. J., Retsch, M. & Floudas, G. (2017). *J. Phys. Chem. B*. **121**, 306–313.

4 SOLVENT FLOWS, CONFORMATION CHANGES, AND LATTICE REORDERING IN A COLD PROTEIN CRYSTAL

Synopsis By maintaining internal solvent in a fully liquid state, temperature- and time-dependent crystal disordering and reordering, protein conformational relaxations, and clear evidence for solvent flows following cooling can be observed in apoferritin crystals at temperatures between 220 and 260 K. These results illuminate causes of and remedies for cooling-induced crystal disorder, and suggest the feasibility of studying aspects of cold denaturation under more nearly native solvent conditions at temperatures down to 200 K.

Abstract When protein crystals are abruptly cooled, the unit cell, protein, and solvent cavity volumes all contract, but the volume of bulk-like internal solvent may expand. Outflow of this solvent from the unit cell and its accumulation in defective interior crystal regions has been suggested as one cause of the large increase in crystal mosaicity on cooling. We show that when apoferritin crystals are abruptly cooled to temperatures between 220 and 260 K, the unit cell contracts, solvent is pushed out, and the mosaicity grows. On temperature-dependent timescales of 10 to 200 s, the unit cell and solvent cavity volume then expand, solvent flows back in, and the mosaicity and B factor both drop. Expansion and reordering at fixed low temperature are associated with small amplitude but large scale changes in apoferritin's conformation and packing. These results demonstrate that increases in mosaicity on cooling arise due to solvent flows out of or into the unit cell and to incomplete, arrested relaxation of protein conformation. They indicate a critical role for time in variable temperature crystallographic studies, and the feasibility of probing interactions and cooperative conformational changes that underlie cold denaturation in the presence of liquid solvent at temperatures down to ~200 K.

4.1 INTRODUCTION

Protein crystallography remains the primary tool for determining atomic-resolution information about protein structure and function. Protein crystals are damaged by X-rays, and the rate of damage with X-ray dose drops dramatically on cooling from ~300 K to 200 K, with a smaller additional drop on cooling to 100 K (Warkentin & Thorne, 2010*a*; Warkentin *et al.*, 2013). The ability to collect much more useful diffraction information per crystal and a convenient experimental infrastructure for handling cryocooled crystals have led to near total dominance of T=100 K data collection in the last 20 years. However, increasing recognition of what is lost by cryocooling (Fraser *et al.*, 2011; Keedy *et al.*, 2015) and increasing competition from single-particle cryoelectron microscopy for determination of initial structures (Vinothkumar & Henderson, 2016) will drive a large expansion of both room temperature and variable/multi-temperature X-ray crystallography in the coming decade.

Although cryocooling reduces the amplitude of thermal atomic motions and often increases achievable diffraction resolution, it increases non-thermal crystal disorder. Crystal mosaicities, a measure of lattice orientational order, increase from as little as a few thousands of a degree to tenths of a degree or more; the spread in unit cell parameters measured in reciprocal space mapping (Kriminski *et al.*, 2002) increases; and crystallographic B factors may also increase. This disorder reduces diffraction peak-to-background ratios and limits data quality improvements that can be achieved by reducing crystal rotations per frame (e.g., in fine ϕ slicing). It is particularly limiting in study of biomolecular complexes and viruses, whose crystals have large unit cells and large solvent cavities. These are targets for which single-particle cryoelectron microscopy has made strong inroads.

The same factors that lead to the improved resolution of cryocooled crystals also perturb the protein's conformational ensemble away from its biologically relevant form. Minority sidechain conformations that may be important in molecular recognition or allosteric interactions are depopulated (Lang *et al.*, 2010;

Fraser *et al.*, 2011; Fenwick *et al.*, 2014). New conformations may appear (Fraser *et al.*, 2011), especially for sidechains involved in crystal contacts but also in non-contact solvent-exposed and buried regions (Atakisi *et al.*, 2018).

Variable/multi-temperature crystallography (Tilton *et al.*, 1992; Teeter *et al.*, 2001)(Warkentin & Thorne, 2009, 2010*b*; Warkentin *et al.*, 2012) performed at temperatures near and above the protein-solvent glass transition (~200 K) has great promise as a tool for mapping a protein's conformational ensemble and energy landscape and for identifying and understanding key interactions (Keedy *et al.*, 2015, 2018). Key obstacles to such studies have been the formation of diffraction-destroying crystalline ice, and the need to use large, crystal perturbing cryoprotectant concentrations to prevent it. We have recently shown that ice formation inside protein crystals is dramatically suppressed by nanoconfinement (Moreau *et al.*, 2019). Following abrupt cooling, the internal solvent in cryoprotectant-free protein crystals with solvent cavities as large as ~70 Å can be maintained in a (supercooled) liquid state at temperatures down to 200 K for a time sufficient to collect a complete data set.

Here we report changes in unit cell and solvent cavity volume, protein conformation and packing, and crystal order that occur at constant temperature *after* apoferritin crystals are cooled to temperatures between 220 and 260 K. These results illuminate the roles of solvent transport and of protein relaxations on different timescales in generating cooling-induced disorder, show how this disorder can be reduced, and suggest the feasibility of studying cooperative conformational changes including those associated with cold denaturation at temperatures down to ~200 K.

4.2 Materials and Methods

4.2.1 Structure of cubic apoferritin crystals

This study focused on cubic crystals of apoferritin, chosen because of their very large solvent cavities and large bulk-like solvent fraction. The atomic structure of cubic apoferritin has been extensively studied at both room and cryogenic temperature. Ferritin monomers (Fig. 4.1(a)) are comprised of five alpha helices labelled A-E (Hempstead *et al.*, 1997) roughly aligned within a volume of 45 Å $\times$ 20 Å $\times$ 20 Å. These monomers readily form dimers (Fig. 4.1(b)), with contacts formed between the A and B helices and along the BC loop. The interface between dimers is highly diverse and is stabilized by hydrophobic interactions, hydrogen bonding between the monomers, and bound solvent. Twelve dimers, having a total molecular weight of 476 kDa, assemble into a nearly spherical shell having a 68 Å diameter internal cavity for storage of iron. In the apo form this cavity is filled with solvent. In cubic apoferritin crystals these solvent cavities have a face-centred cubic arrangement. Theses cavities are larger than those found in roughly 98% of PDB-deposited protein crystal structures (Moreau *et al.*, 2019). A second set of large cavities lies between the spherical shells, having a characteristic size of 65 Å (the distance between second nearest-neighbor shells) and together containing roughly 65% of the total solvent volume. In addition to these large cavities inside and between spherical shells, 8 hydrophilic channels, 6 Å long and 3.4 Å in diameter (providing conduits for water, ions, and glycerol when crystals are soaked or cooled) and 6 hydrophobic channels, each 12.5 Å long and 2 Å in diameter, pass through each spherical protein shell.

The spherical shells contact each other at the dimer interface along the BC loops (Fig 4.1(c)). Residues ASP80 and GLN82 protrude from the BC loop and connect to a cadmium atom between the shells. ASP80 has well defined electron density and is always involved in the contact. GLN82 is found in either an '*in*' or '*out*' conformation in a given crystal (Fig. 6.3.1); the specific conformation may depend on the pH of the mother liquor or soak solution (De Val *et al.*, 2012). In the *in* conformation GLN82 is connected to the Cd

atom with well-defined electron density. In the *out* conformation GLN82 protrudes outward into the solvent cavity and an additional water molecule coordinates with the Cd atom. The distance between the spherical shells was slightly different in crystals having each of these conformations. To quantify effects of temperature and glycerol concentration, crystals exhibiting the *in* and *out* conformations were separately analysed.

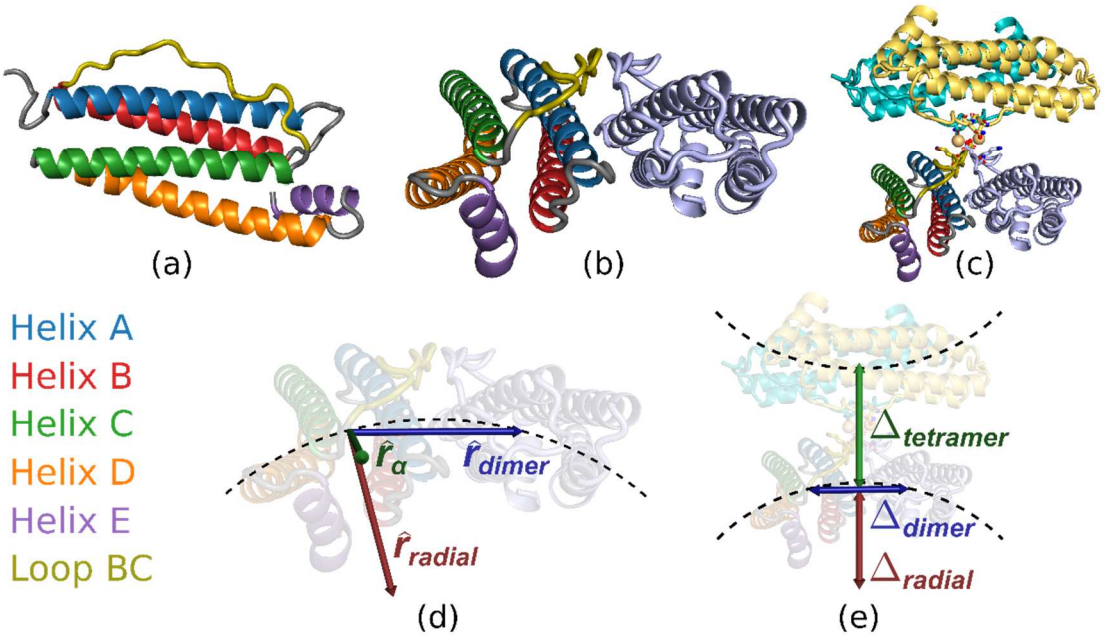


Figure 4.1 (a) A ferritin monomer, comprised of five helices and a long flexible loop. (b) A ferritin dimer. Twelve of these dimers assemble into a spherical shell with an internal cavity of diameter 68 Å. In the apo form this cavity is filled with solvent. (c) In the crystal, the interface between spherical shells involves contacts between the BC loops of adjacent dimers, mediated through Cd atoms. (d) Coordinate reference directions and (e) distances described in the text and used to characterize structural changes during cooling and cold unit cell expansion.

4.2.2 Protein crystallization and harvesting

Cubic crystals of equine spleen apoferritin (Sigma A-3641) were grown by the hanging drop vapor diffusion method. Drops comprised of 2 µL of 10 mg /mL protein in 0.1M sodium acetate buffer at pH 6.5 and 2 µL of a well solution of 2% w/v $CdSO_4$ and 15% w/v (~1.1 M) $(NH_4)_2SO_4$ in the same buffer were equilibrated with well solution in 24-well plates. Cubic crystals in space group F432 having room-temperature unit cell dimensions a = b = c = 183.5 Å and with sizes of 300-500 µm were obtained within a week. The crystals have a Matthews-coefficient-derived solvent content of 63% v/v.

Crystals were used as grown, or else soaked in solutions containing 10%, 20%, or 40% v/v glycerol for at least 5 minutes. Crystals were then transferred to a separate drop of NVH oil (Cargille) and manipulated until all external solvent was removed from their surface, as indicated by a near disappearance of the crystal due to the close match between its refractive index and that of the oil (Warkentin & Thorne, 2010*b*). Crystals were mounted on MicroGrippers (MiTeGen) in a spherical blob of NVH oil to prevent dehydration during data collection; the fingers of the MicroGrippers helped to immobilize both the oil and crystal. Mounted crystals were stored in MicroRT tubes (MiTeGen) containing mother liquor or cryoprotectant solution for ~1 hour prior to data collection. Although apoferritin crystallization solutions and solvent present in crystal cavities contain significant salt concentrations, these only reduce the solution's freezing temperature by 4°C (Clegg 1995). Ice formation is dominated by the effects of nanoconfinement by the protein lattice (Moreau *et al.*, 2019) and, at concentrations of 10% v/v and larger used here, by glycerol.

4.2.3 X-ray data collection

As was discussed in the Supporting Information of Moreau *et al.* (Moreau *et al.*, 2019), time-dependent X-ray diffraction data was collected using the F1 beamline at the Cornell High-Energy Synchrotron Source (CHESS) using the experimental configuration shown in Fig. 6.3.2. Samples were illuminated using a Gaussian beam with a 65 µm FWHM, a divergence of 0.015°, a photon energy of 12.7

keV, and a flux of 2.2×10^9 ph/s , giving dose rates of ~2 kGy/s. Diffraction patterns were recorded by a Pilatus 6M detector using a frame rate of 5 Hz.

Sample temperature was controlled using a nitrogen gas stream generated by an Oxford Cryosystems Cryostream 700 nitrogen gas cryocooler. The gas flow rate was 5 L/min, corresponding to a ~1 m/s velocity through a 1 cm diameter aperture. The gas stream's temperature was varied between 180 K and 260 K using the cryocooler's internal heater and was monitored using both the cryocooler's internal temperature sensor and using a thermocouple that was periodically placed at the sample position.

Before X-ray data collection, with the cryostream retracted and blocked, a sample in its MicroRT tube was manually placed on the beamline goniometer stage, the crystal was allowed to settle in the oil, and then the crystal was translated into the X-ray beam path. The MicroRT tube was then removed, and automated data collection initiated using the beamline's ADX software. In a typical experiment, an initial set of 10 frames (0.5° of sample rotation and 0.2 s exposure time per frame, 5° rotation and 2 s total per exposure), was collected at room temperature. The crystal was then returned to its initial orientation, the cryostream head was extended and unblocked, and a single set of 200 frames (0.5°, 0.1 or 0.2 s exposure time per frame, 100° rotation and 20-40 s total per exposure) acquired. With a dose rate of ~2 kGy/s, total doses ranged from ~40 to ~100 kGy, much less than the half dose at all temperatures studied, so changes in diffraction properties with time were dominated by effects other than radiation damage. Sample cooling rates, as monitored using the crystal's unit cell, varied with sample (crystal + surrounding oil) volume, with maximum cooling rates of ~300 K/s and cooling times to the final sample temperature in the range of 0.5-2 s. As illustrated in Fig. 4.2, the time required for each crystal to cool and the stability of its final temperature were verified by monitoring the positions of the strong, broad peak near 5.7 Å due to scattering from the surrounding NVH oil. Temperature versus time at the sample position (with the sample removed) was also measured using a 250 μm thermocouple (Fig. 6.3.4). All three measurements gave consistent results.

A total of 142 crystals of apoferritin were examined at CHESS on eight different dates between November 2015 and April 2018. Measurement of a large number of samples was essential to drawing robust conclusions because of variations in the extent to which external solvent was removed, the stochastic nature of ice formation within supercooled internal crystal solvent, and because sample response depended to some extent on cooling time, which in turn depended on sample size.

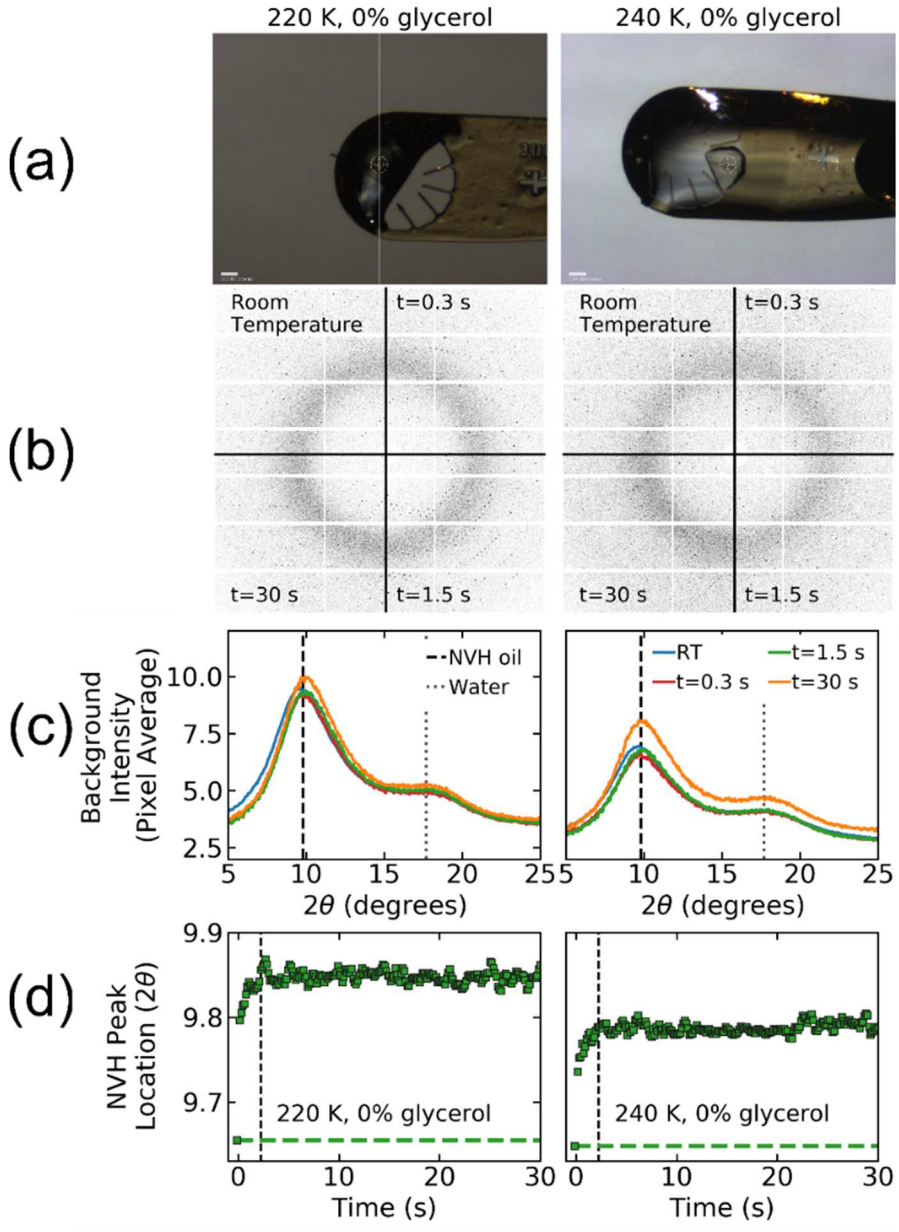


Figure 4.2 (a) Images of the glycerol-free apoferritin crystals in NVH oil at (left) 220 K and (right) 240 K, for which unit cell, mosaicity, and Wilson B factor data are shown in Fig. 4.5. The images were taken at the end of data collection. The crystals are ice free and appear clear. When internal ice forms, crystals become darker and opaque. (b) Diffraction patterns recorded from the apoferritin crystals in (a) at room temperature, and at the indicated times after the start of cooling to 220 K (left) and 240 K (right). The Bragg peaks from the protein lattice have been deleted, leaving diffuse scatter primarily generated from the NVH oil surrounding the crystal and also from internal solvent. (c) Azimuthally averaged diffuse diffraction intensity from the patterns in (b). Peaks from NVH oil and from internal solvent undergo only very small shifts in position during cooling, indicating small changes in density. No evidence for ice or other major changes in solvent structure (e.g., as might occur in a transition between low-density and high-density amorphous ice) is observed. (d) Position of the diffraction peak from NVH oil versus time during and following cooling. The peak position monotonically increases from its room-temperature position (indicated by the dashed line) to its final position at the final temperature. The direction and size of the shifts at left and right are consistent with an increase in the oil's density with decreasing temperature. The timescale of this increase gives the timescale for crystal cooling. The relatively weak diffraction from internal solvent made precise tracking of its peak position with time/temperature difficult.

4.2.4 Processing of protein lattice diffraction

Diffraction frames were indexed, integrated, and scaled using *XDS* (Kabsch, 2010), with an input file acquired from the MacCHESS website and modified for use with the Pilatus 6M detector. The resolution and *XDS*-determined mosaicity at room temperature, averaged over all crystals, were 2.2 ± 0.3 Å and 0.06° ± 0.02°, respectively. Bragg peaks were thus strong and well separated, so data processing using a single initial input file generally proceeded with few issues.

Data sets were processed in segments of 5 frames. A segment's refined outputs of unit cell, beam centre, and sample-to-detector distance were used as inputs in the XDS.INP file when processing subsequent segments. As a check on these results, data were also processed using *XDS* in segments of 2 and 10 frames, and using *HKL2000*). The use of 5-frame segments balanced time resolution with variance in parameter estimates.

Unit cell values were taken directly from *XDS*'s CORRECT.LP output file. Wilson *B* factors were estimated as half the slope of a linear fit to the natural log of the Bragg peak intensities (obtained using *phenix.merging*_stats (Adams *et al.*, 2010)) vs. $\left(\sin\theta/\lambda\right)^2$. Only Bragg peaks having resolutions numerically smaller than 4 Å and worse than a cut-off resolution where I/σ>2 typically ~ 2 Å and no larger than 3 Å were used to determine *B* factors.

The refined crystal mosaicities, calculated using *XDS* as the standard deviation of a Gaussian, had a "floor" of 0.045°. This was imposed by *XDS* and did not reflect actual crystal mosaicities. Diffraction frames were also indexed, integrated and scaled using *HKL2000*, which had a mosaicity "floor" close to the refined beam divergence, estimated by *XDS* to be 0.015°. Mosaicity values reported here were calculated using values from the *HKL2000* *.x integration files, converted from full width at half maximum to standard deviation to allow comparison with *XDS* results.

4.2.5 Protein structure modelling and refinement

Data sets used for structure determination and analysis all had resolutions ≤2.1 Å and an average resolution of 1.96 Å, where the resolution cut-off was set so that the highest-resolution shell had at least 98% completeness and I/σ>2. Molecular replacement and model refinement were performed using *PHENIX* (Adams *et al.*, 2010), using PDB entry 3F32 as a starting model for all crystals/temperatures/glycerol concentrations. An initial refinement cycle using *phenix.refine* was performed, with simulated annealing, rigid body, real-space, xyz coordinates, and individual B-factor options. In *Coot* (Emsley *et al.*, 2010), these models were checked for large peaks in the difference maps, Ramachandran outliers, rotamer outliers, and regions of poor geometry. Cadmium atoms located at the interface of the spherical shells and residues at a partially disordered loop, Gly155 and Ser156, were added into the model. A second round of refinement was performed to identify and add ordered solvent. These models were checked residue by residue for errors and alternative conformations added. A third round of refinement including occupancies and B factors for alternative conformers and target weights was performed. The placement of the water molecules at the interface of the spherical shells were checked for accuracy and a fourth round of refinement was performed with the same parameters as the third round. Final model validation was performed using *MolProbity* in the *PHENIX* software package (Chen *et al.*, 2010).

At room temperature, only 5 degrees of data were collected from each crystal, so no single crystal gave enough data for structure determination. Data from multiple room-temperature crystals, collected on the same day, were merged using *XSCALE* to create 4, 2, 2, and 1 structures at 0%, 10%, 20%, and 40% v/v glycerol respectively.

4.2.6 Protein and solvent volume calculations

The volume of protein within the unit cell $v_{protein}(T)$ was calculated as the volume enclosed by the solvent excluded surface (SES), using a custom "ball rolling" algorithm very similar to that implemented by the program 3vee (Voss & Gerstein, 2010). In 3vee, the SES is calculated by “rolling” a fixed-diameter probe over the surface of a voxelized 3D model of the protein, with atoms modelled as spheres having radii matching their Van der Waals radii (Li & Nussinov, 1998). Our program was designed to also find the volume for multiple copies of the protein extended to fill the unit cell, accounting for crystal contacts and periodicity. Additionally, it uses probe radii of 1.4 Å and 1.7 Å for polar and apolar atoms, respectively (Li & Nussinov, 1998). The program divided the unit cell into voxels of dimension 0.15 Å and recorded all regions where the probe clashes with an atom as part of the protein’s volume.

Solvent cavity volumes $v_{cavity}(T)$ were determined by subtracting the protein volumes $v_{protein}(T)$ from the unit cell volumes $v_{cell}(T)$. At room temperature, all crystals were highly ordered with very small mosaicities. Consequently, the solvent volume within a crystal could be assumed to be fully contained within the solvent cavities of the crystallographically-determined unit cell, and the crystal's solvent volume fraction was equal to the volume fraction of the unit cell occupied by cavities,

$$f_{solvent}(300\ \mathrm{K}) = \frac{v_{cavity}(300\ \mathrm{K})}{v_{cell}(300\ \mathrm{K})}\ . \qquad (4.1)$$

We also separately calculated the solvent cavity volumes located inside and outside of the spherical protein shell. Using the voxel grid created with our ball rolling method, we calculated the fraction of voxels belonging to solvent enclosed within thin spherical shells concentric with the apoferritin shell as a function of the shell radius, starting at radii much smaller than that of the apoferritin shell. For the radius at which this solvent volume fraction reached a minimum, the enclosed solvent volume was taken as the solvent

cavity volume located inside the spherical protein shell. The volume outside the spherical protein shell was then obtained by subtracting this from the total solvent cavity volume in the unit cell.

Crystals with GLN82 in both the "in" and "out" conformations discussed in Sec. 4.2.1 were observed in our experiments, with all crystals measured on a given date having the same conformation. Because the two conformations gave slightly different unit cells and subtle differences in backbone, we only compared structures with the same conformation.

4.2.7 Protein structure analysis

To explore structural changes to apoferritin during initial unit cell contraction on cooling and during subsequent cold expansion, models of the ferritin monomers were constructed using data sets from 29 glycerol-free apoferritin crystals at temperatures between 180 K and 300 K, including 4 initial room-temperature data sets, the 16 data sets between 180 K and 260 K where the unit cell contracted on cooling but did not expand appreciably during the data acquisition time, and 9 data sets collected from crystals at temperatures between 220 K and 260 K that were deemed to have fully expanded by a plateauing of their unit cell in time and a drop in *HKL2000* mosaicity to near the beam divergence limit. One crystal measured at T = 220K, where the time for expansion was longer, allowed separate structures to be determined after initial cooling while the unit cell remained close to its minimum cold value (having expanded by less than 19% of the total expansion), and then after expansion by at least 74%. The modelled structures were overlaid and compared residue-by-residue for differences within the protein monomers, in the monomer packing, and in the interaction between the apoferritin shells. No significant differences, either to backbone structure or side chain conformations with corresponding differences in observable electron density, were observed between any structure, other than the previously discussed variation to GLN82.

To characterize more subtle shifts in protein structure and packing during cooling and cold expansion, several quantities shown in Figs. 4.1(d) and (e) were evaluated from the positions of the backbone C, N,

and C_α atoms of the refined models. Before this analysis, residues 1-3, and 155-158 were removed from the models due to their ill-defined electron density. Refined models were reduced to a single conformer and superposed to the same location in the unit cell by symmetry operations.

In order to characterize changes in protein structure and packing during cooling and cold expansion, several quantities shown in Figs. 4.1(d) and (e) were evaluated from the positions of the backbone C, N, and C_α atoms in the refined models. The centre of volume of the protein shell is a known location in the unit cell. The centre of volume of the monomer was estimated as the average position of its atoms, $\langle \vec{r} \rangle$. The radial distance between the shell centre and a monomer centre was defined as Δ_{radial}, and the unit axis connecting these points as $\hat{r}_{radial}$. A second protein monomer was added in the dimer symmetry location. The distance between the centres of volume of the monomers was defined as Δ_{dimer} and the unit axis connecting these points as $\hat{r}_{dimer}$. This axis is roughly tangent to the protein shell's surface. A third axis is defined perpendicular to these axes, $\hat{r}_\alpha = \hat{r}_{radial} \times \hat{r}_{dimer}$, and is roughly oriented along the monomer's long alpha helices and tangent to the shell. These axes are shown relative to the apoferritin dimer in Fig. 4.1(d). A second dimer belonging to the adjacent protein shell and in contact with the first dimer was added at the symmetry location. The centre of volume of each dimer was estimated, and the distance between their centres $\Delta_{tetramer}$ was interpreted as the distance between protein shells.

The radius of gyration of a monomer was estimated from the distances of its backbone atoms from the monomer's center of volume,

$$R_g = \frac{1}{n}\sum_{i=1}^{n}\sqrt{\left(\vec{r}_i - \langle \vec{r} \rangle\right)^2} \quad , \tag{4.2}$$

where $\vec{r}_i$ is the position of the ith backbone atom. To evaluate how the protein's shape changed, a radius of gyration along each of the axes defined above was estimated as

$$R_{g,axis} = \frac{1}{n}\sum_{i=1}^{n}\sqrt{\left(\left(\vec{r}_i - \langle\vec{r}\rangle\right)\hat{r}_{axis}\right)^2} \quad . \tag{4.3}$$

To image structural changes, two structures at room temperature, three structures at 220 K that did not show post-cooling expansion, and four structures at 220 K that showed post-cooling expansion were converted to a set of average structures, one for each condition, by averaging the backbone atom coordinates of the structures determined at that condition,

$$\left\{\vec{r}_i\right\}^{condition} = \frac{1}{m}\sum_{j=1}^{m}\vec{r_i^j} \quad . \tag{4.4}$$

Index j represents structures obtained from different data sets acquired under the same conditions. The scalar displacement $\varepsilon_i^{a,b}$ of backbone atom *i* between averaged structures at conditions a and b was calculated as

$$\varepsilon_i^{a,b} = \sqrt{\left(r_{i,x}^a - r_{i,x}^b\right)^2 + \left(r_{i,y}^a - r_{i,y}^b\right)^2 + \left(r_{i,z}^a - r_{i,z}^b\right)^2} \, . \tag{4.5}$$

This was normalized by subtracting the average displacement of all backbone atoms,

$$\left\langle \varepsilon^{a,b} \right\rangle = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^{a,b} \, , \tag{4.6}$$

and dividing by the square root of the sum of variances σ_i^2 within each population,

$$\sigma_i^2 = \frac{1}{m}\sum_{j=1}^{m}\left(r_{i,x}^j - \left\langle r_{i,x}\right\rangle\right)^2 + \left(r_{i,y}^j - \left\langle r_{i,y}\right\rangle\right)^2 + \left(r_{i,z}^j - \left\langle r_{i,z}\right\rangle\right)^2 \, , \tag{4.7}$$

to obtain

$$\hat{\varepsilon}_i^{a,b} = \frac{\varepsilon_i^{a,b} - \left\langle \varepsilon^{a,b} \right\rangle}{\sqrt{\left(\sigma_i^a\right)^2 + \left(\sigma_i^b\right)^2}} \, . \tag{4.8}$$

Normalizing in this manner pulls out the more significant displacements by down-weighting atoms that might show large variances due to imprecise placement from weak electron density.

4.3 Results and Analysis

4.3.1 Unit cell, protein, and solvent cavity contraction on cooling

Fig. 4.3 replots our previously reported data (Moreau *et al.*, 2019) for how the unit cell volume, protein volume, and solvent cavity volume of apoferritin crystals, determined immediately after cooling has completed, vary with temperature and glycerol concentration. All three contract on cooling. The protein volume shows a fractional contraction that is 2-3 times smaller than that of the unit cell, so the fractional contraction of the solvent cavity volume is nearly double that of the unit cell. Addition of glycerol increases the unit cell and solvent cavity volumes at room temperature, and has a small effect on the amount these contract on cooling from room temperature to 220 K. In contrast, glycerol has a large effect on the contraction of bulk aqueous solutions. Between 300 K and 77 K, pure water expands by 6% (Loerting *et al.*, 2011) and a 40% v/v glycerol solution contracts by ~6% (Shen *et al.*, 2016), respectively, as they cool into amorphous ice. Between 300 K and 240 K liquid water expands by 2% (Hare & Sorensen, 1987) and a 40% v/v glycerol solution contracts by 2% (Glycerine Producers Association, 1963). This suggests that solvent must be expelled from or flow into the unit cell to account for the difference between solvent cavity and solvent contractions (Juers & Matthews, 2001; Kriminski *et al.*, 2002; Juers & Matthews, 2004; Tyree *et al.*, 2018; Juers *et al.*, 2018). As shown in Fig. 6.3.3, for glycerol-free crystals and crystals soaked in 40% v/v glycerol, 1.7% and 0.9% of the solvent must exit the unit cell on cooling to 240 K. These values assume solvent in the first hydration shell has the same contraction as the protein, and that the remaining solvent has the solvent contraction of the bulk solution. Other reasonable assumptions about the contraction of solvent in the first hydration shell yield similar results

Fig. 4.4 shows how (a) crystal mosaicity and (b) Wilson B factor, determined immediately after cooling has completed, vary with final temperature and with glycerol concentration. For glycerol-free crystals, the mosaicity determined by *HKL2000* remains near the "floor" value of ~0.03°, set by the incident beam divergence, on cooling to temperatures as low as 240 K. On cooling to temperatures of 180 and 200 K, mosaicities increase to ~0.15° to 0.2°. Wilson B factors show at most a small decrease with decreasing temperature down to 180 K. Both mosaicities and Wilson B factors and tend to be somewhat lower for glycerol-free crystals than for those soaked in 20% and 40% v/v glycerol at temperatures above ~200 K.

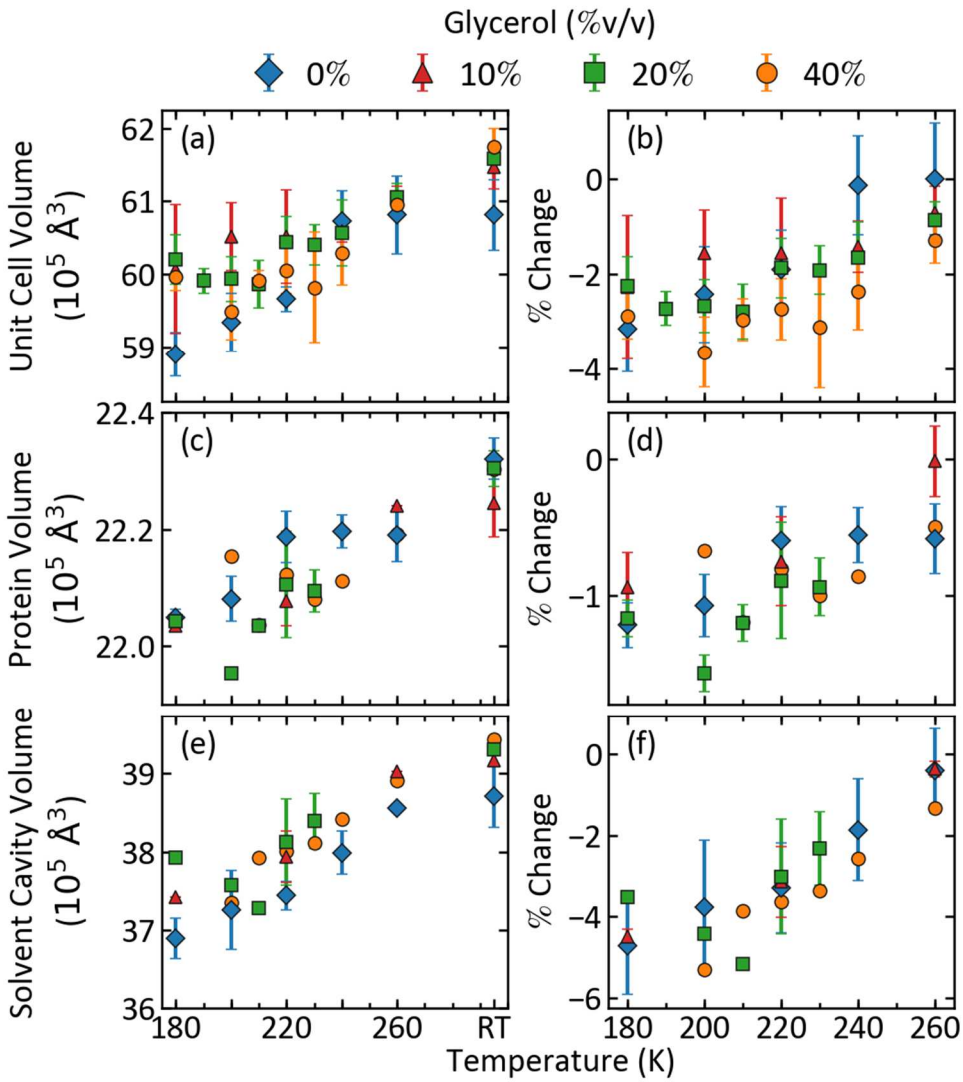


Figure 4.3 Absolute values and percent changes from room temperature, respectively, of (a),(b) unit cell volume, (c),(d) protein volume, and (e),(f) solvent cavity volume, as a function of temperature and glycerol concentration for cubic apoferritin crystals. For 220 K ≤ T ≤ 260 K, where cell volumes increased on long timescales after cooling, minimum cell volumes measured a few seconds after cooling had completed are plotted. Protein volume within the unit cell was deduced from full structural models by evaluating the solvent excluded surface of the protein, and the total solvent cavity volume was obtained by subtracting this volume from the unit cell volume. Error bars indicate cell variations between crystals prepared and measured under nominally identical conditions. The number of crystals examined for each condition is given in Table 6.3.1. Adapted from Fig. 5 of (Moreau *et al.*, 2019).

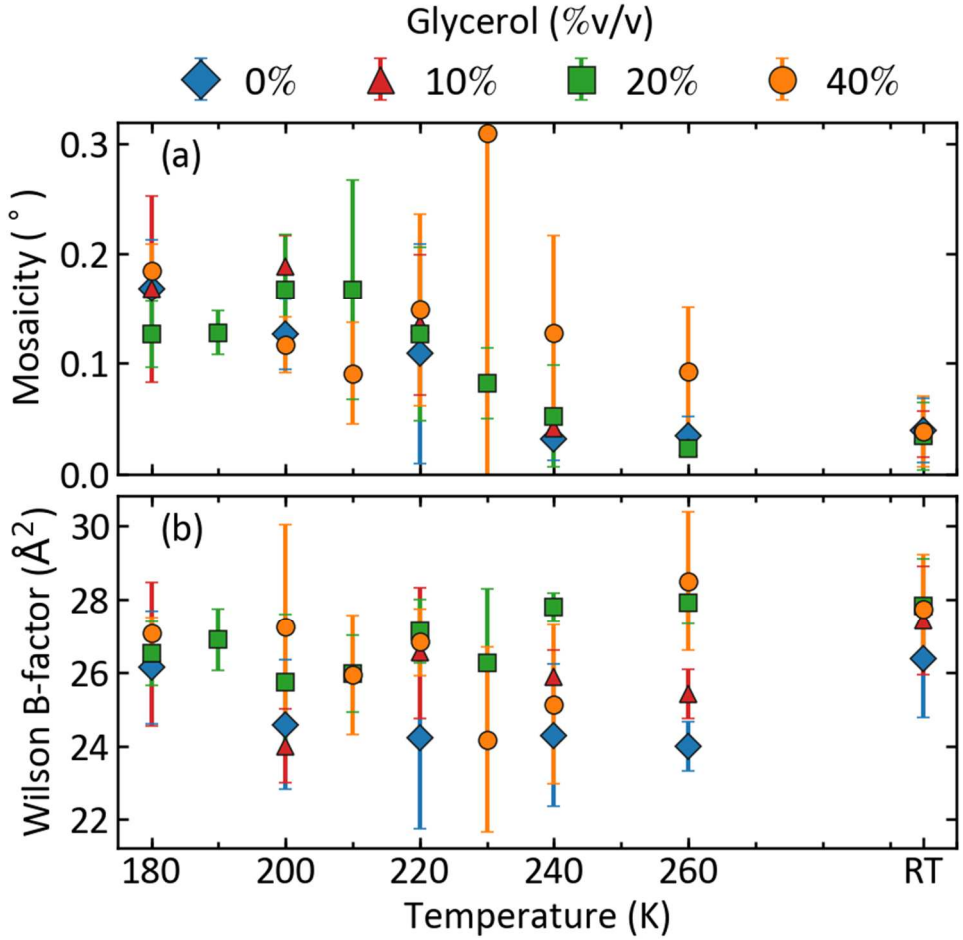


Figure 4.4 (a) Mosaicity and (b) Wilson B factor, determined using data acquired 3-5 s after cooling, versus temperature and glycerol concentration for ice-free cubic apoferritin crystals. Mosaicities have a "floor" determined by the incident X-ray beam divergence of ~ 0.03°. Actual room-temperature crystal mosaicities are likely ~0.01° or less. Significant scatter in values arises from crystal-size-related variations in illuminated volume and background scatter. Adapted from Supporting Information Fig. S11 of (Moreau *et al.*, 2019).

4.3.2 Time-dependent unit cells, mosaicity, and Wilson B factors in cold crystals

When crystals are cooled to ~140 K or below, vitrified internal solvent immobilizes protein atoms and crystal diffraction properties are (in the absence of radiation damage) independent of time. As crystals are warmed from ~140 K to ~190 K, solvent atoms develop increasing diffusive mobility and ice crystals grow (Weik, Kryger *et al.*, 2001). Ice formed within the unit cell's solvent cavities can cause unit cell expansion. Ice formed at crystal surfaces draws water from the rest of the crystal for its growth, and can cause unit cells to shrink (Juers *et al.*, 2018**)**, just as ice formed in intercellular spaces leads to cellular dehydration in biological cryopreservation (Fahy & Wowk, 2015).

How do crystal diffraction and unit cell volume evolve at *constant* temperature following abrupt cooling, under conditions where the internal solvent remains liquid and ice does not form? On cooling, the unit cell volume initially drops and the mosaicity grows, on the timescale of the cooling transient (~0.2-2 s). As illustrated by the examples shown in Fig. 4.5, for roughly 70% of ice-free apoferritin crystals at temperatures between 220 and 260 K, after the final temperature was reached the unit cells then expanded, by an average of ~0.8% at 220 K for glycerol-free crystals and ~2.4% at 240 K for crystals soaked in 40% (v/v) glycerol. As shown in Fig. 4.6, the final unit cell volumes often exceeded their room temperature values. As shown in Fig. 4.7, the unit cell expansion time scale, obtained from an exponential fit, tended to increase with decreasing temperature and increasing glycerol concentration. For glycerol concentrations of 0%, 10%, and 20% but not 40% v/v, cell expansion was often accompanied by major improvements in crystal order as reflected in large drops in crystal mosaicity (Figs. 4.5 and 4.8) and sometimes also substantial decreases in Wilson B factors (Fig. 4.5). As discussed later, these diffraction quality improvements in oil-encased crystals held at fixed low temperature in dry nitrogen gas cryostreams are unrelated to those seen in "cryoannealing", where crystals are briefly warmed (typically to room temperature) so that external solvent melts before re-cooling. No appreciable cell expansion or diffraction quality changes were observed at 180 K or 200 K on timescales of 200 s, mostly like because of

diverging protein relaxation timescales as the protein-solvent glass transition was approached (Ringe & Petsko, 2003; Doster, 2010; Weik, Ravelli *et al.*, 2001; Warkentin & Thorne, 2010*a*). In contrast with these results for apoferritin, similar measurements on thaumatin crystals (Fig. 6.3.4) showed no low-temperature unit cell expansion and/or crystal reordering on <200 s timescales at any temperature between 180 and 260 K.

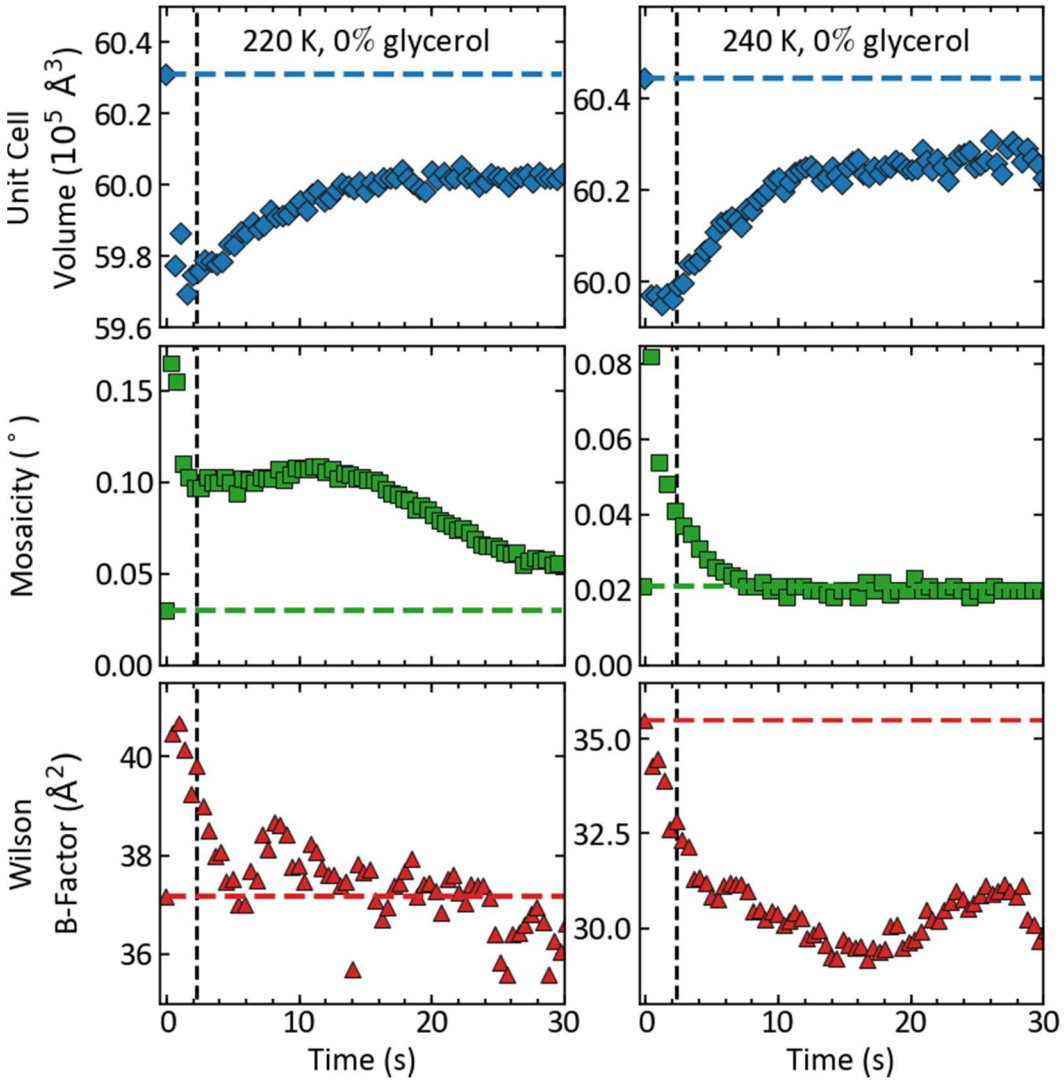


Figure 4.5 Unit cell volume, mosaicity, and Wilson B factor during and following cooling from room temperature to (a) 220 K and (b) 240 K, measured for the two glycerol-free apoferritin crystals shown in Fig. 3.2 that remained ice-free throughout. The unit cell volume decreases sharply during the initial cooling transient, and then expands on a timescale that increases with decreasing temperature. Similarly, the mosaicity increases and then decreases. Dotted blue, green, and red lines indicate the room temperature cell volume, mosaicity, and B factor, respectively. The dashed vertical lines indicate the time at which the NVH oil peak position reached 95% of its final value; at larger times, the sample temperature was constant to within a few degrees.

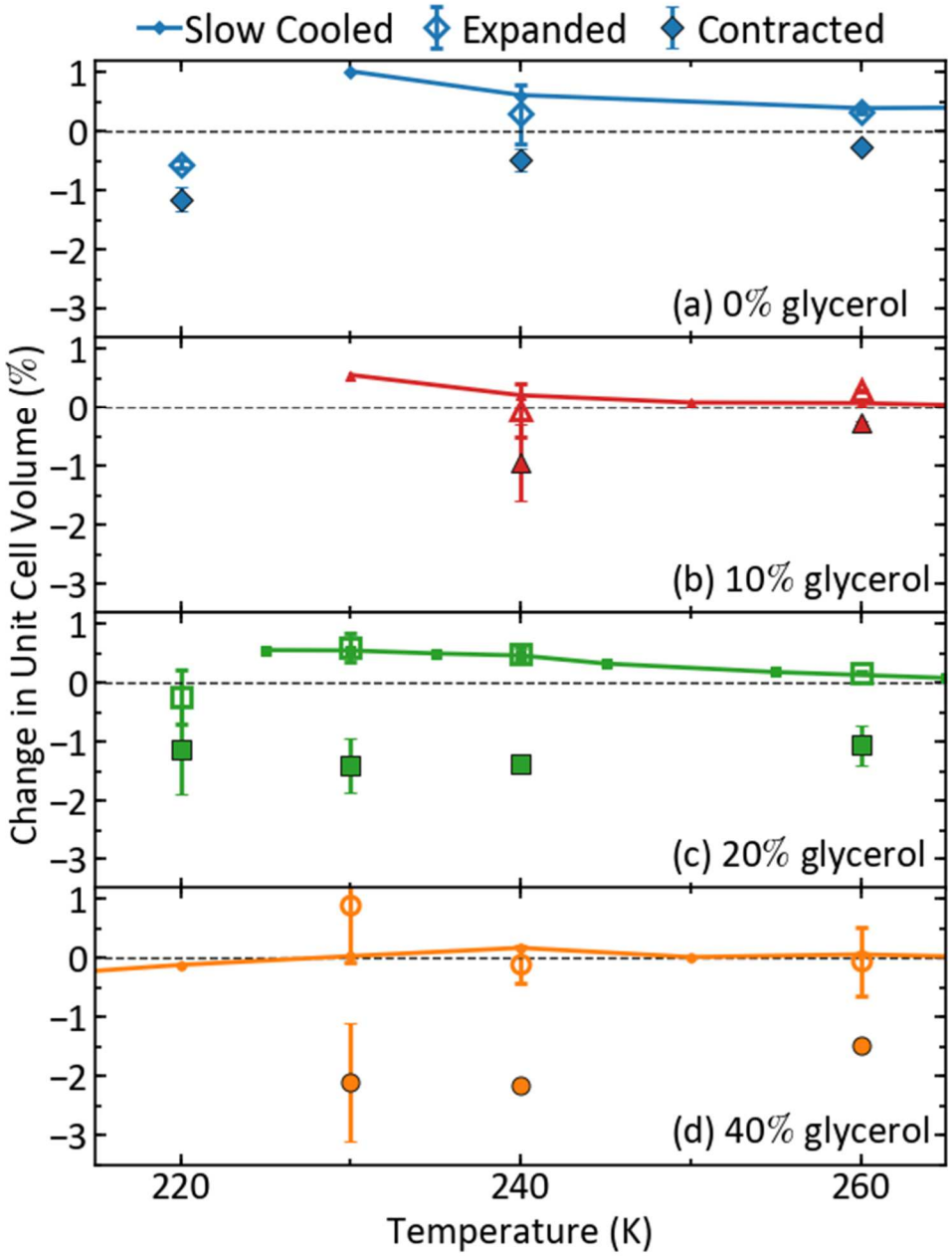


Figure 4.6 Change in unit cell volume relative to room temperature for apoferritin crystals that were slowly cooled at ~0.1 K/s (solid lines and small symbols), and that were abruptly cooled (in <1 s) to each temperature (large symbols), versus temperature. Closed and open symbols indicate unit cell volumes measured just after cooling had completed and after completion of unit cell expansion, respectively. For all glycerol concentrations, the expanded unit cell volumes approach the slow-cooled volumes, which between 220 and 250 K were generally larger than the room temperature volumes. The slow cooled data ends at the lowest temperature where ice diffraction was not observed.

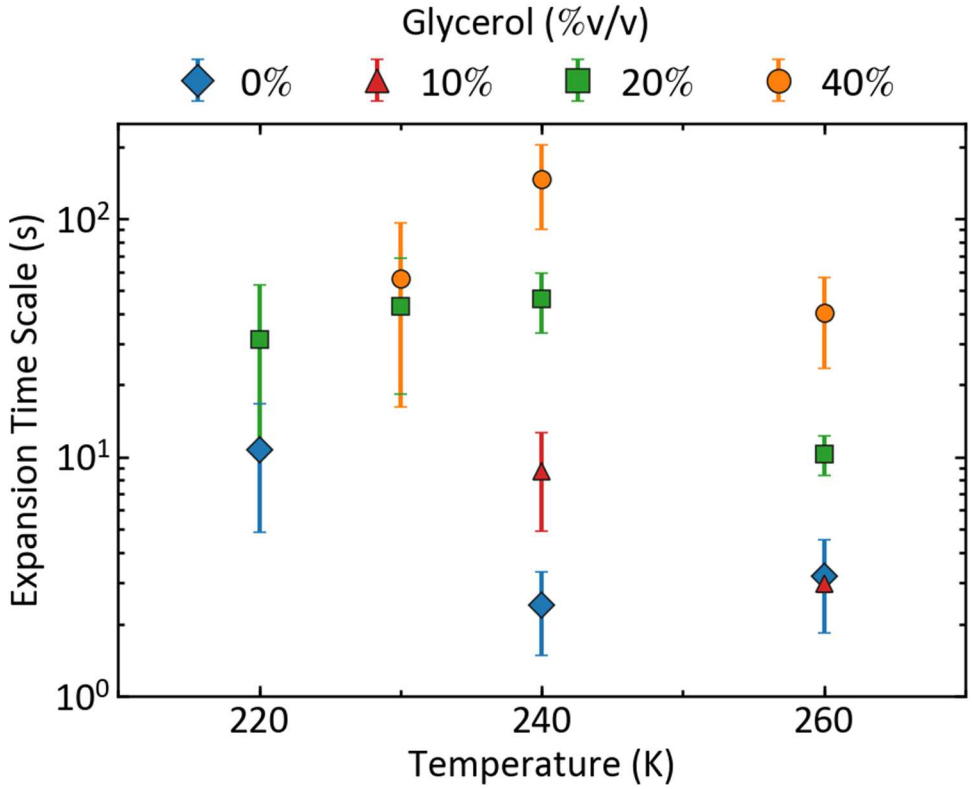


Figure 4.7 Timescale for unit cell expansion at constant temperature following abrupt cooling versus temperature and glycerol concentration. Error bars indicate the standard deviation of values obtained from fits to crystals at each temperature/concentration. The expansion timescale τ was obtained from a fit to the unit cell data of the form

$$V_{cell}(t) = \left(V_{cell}\left(t_{final}\right) - V_{cell}\left(t_{initial}\right)\right)\left(1 - \exp\left(-(t - t_{initial})/\tau\right)\right) + V_{cell}\left(t_{initial}\right),$$

where $t_{initial}$ was the time at which the expansion began.

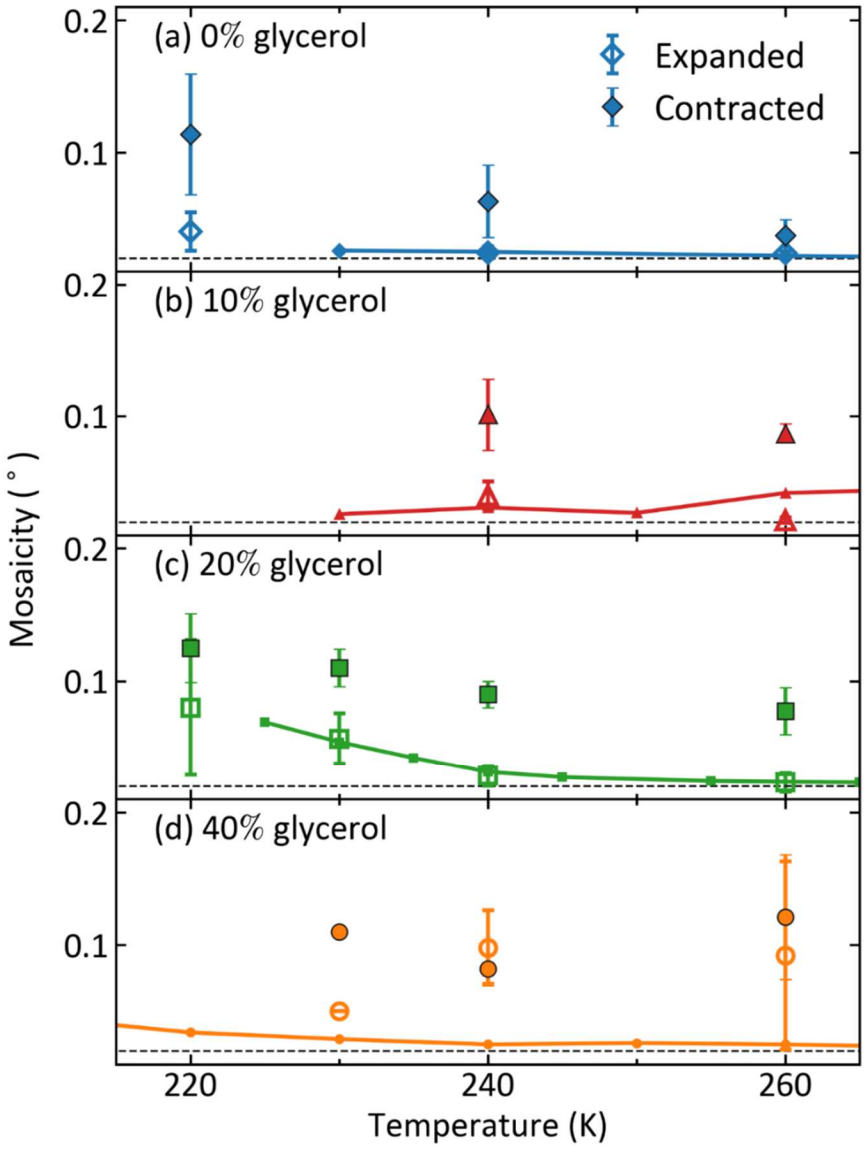


Figure 4.8 Mosaicity before (closed symbols) and after (open symbols) unit cell expansion at constant temperature versus temperature. For glycerol-free crystals and crystals soaked in solutions containing 10% and 20% v/v glycerol, the mosaicity drops dramatically during expansion toward the 0.03° floor (horizontal dashed line) set by the incident beam divergence. For crystals soaked in 40% v/v glycerol, the mosaicity tended to get worse during expansion. The small symbols and solid lines indicate the mosaicity measured when the crystal was slowly cooled at 0.1 K/s.

4.3.3 Unit cells and mosaicities obtained during slow cooling

To gain insight into the possible causes of low temperature unit cell expansion in apoferritin, diffraction data were acquired as apoferritin crystals were slowly cooled – at 0.1 K/s to 200 K (requiring ~1000 s) – by a programmed ramp of the nitrogen gas stream temperature. This very slow cooling allowed a more complete approach toward equilibrium at each temperature. Over the temperature range where each crystal remained ice-diffraction free, crystal mosaicities determined by HKL2000 remained at their room temperature (beam-divergence limited) values. The solid lines in Fig. 4.6 show the change in apoferritin's unit cell volume relative to its room temperature value over this temperature range. For all glycerol concentrations, the initial unit cells obtained after abrupt cooling (closed symbols) are smaller than those obtained by slowly cooling, whereas cells measured after expansion at fixed temperature (open symbols) closely match those obtained by slowly cooling, for temperatures between 230 and 260 K. This suggests that the expanded state more nearly approximates the protein and crystal's equilibrium state at each temperature.

For all glycerol concentrations, the expanded and slow cooled unit cells both show little variation with temperature between 220-235 K and 300 K, even though there is a strong decrease over this temperature range in the initial unit cell after abrupt cooling.

4.3.4 Analysis of cooling and post-cooling structural changes

Structural changes to apoferritin that occurred as the unit cell contracted during cooling and expanded after cooling were subtle, and somewhat challenging to confidently interpret given the resolution of the datasets analysed. We focused on structures obtained using glycerol-free crystals, as these tended to have the highest resolutions and lowest mosaicities. We separately analysed crystals in which the *in* and *out* conformations of GLN82 dominated, as this conformational difference substantially affected some results. Table 6.3.4 lists the number of crystals used for the structural analysis at each condition, and a supplementary spreadsheet provides refinement statistics for all analysed crystal

structures. Figs. 4.9 (a) and (b) show the solvent cavity volume measured following abrupt cooling and following cold unit cell expansion, determined by subtracting the protein volume from the unit cell volume. Figs. 4.9 (c) and (d) show that the volume outside apoferritin's protein shell contracts more during cooling and expands more during the cold expansion than does the solvent volume inside the shell. This suggests that the largest contribution to the unit cell changes comes from interactions between apoferritin shells rather than from changes within them.

Fig. 4.10 shows the fractional changes from room temperature of the unit cell's linear dimension and of Δ_{radial}, $\Delta_{tetramer}$, and $\Delta_{\mathrm{dim}er}$ as defined in Fig. 4.1(e). While each distance metric decreased on cooling and increased on expansion, $\Delta_{tetramer}$ showed by far the largest fractional changes, further implicating the interface between protein shells as the dominant location for changes to the unit cell. As shown in Fig. 6.3.5, fractional changes in R_g of each monomer and its components along $\hat{r}_{radial}$, $\hat{r}_{dimer}$, and $\hat{r}_{\alpha}$ are much smaller.

Figs. 4.11 (a) and (b) show representations of the dimers in the *out* conformation, created in *pymol*. The colours from blue to red indicate increasing values of $\varepsilon_i^{a,b}$ obtained by comparing (a) room temperature and initial cold structures and (b) initial cold structures with final cold structures after expansion. Arrow locations indicate positions where the normalized atomic displacements are at least one standard deviation larger than the mean, arrow directions indicate the direction of the displacement, and arrow lengths are 50 times larger than the actual displacement. During cold expansion (Fig. 3.11(b)) the atomic changes are somewhat localized to the BC loop at the tetramer interface, consistent with post-

cooling expansion occurring primarily between the spherical shells. During cooling (Fig. 3.11(a)), structural changes are more uniformly distributed throughout the protein shell.

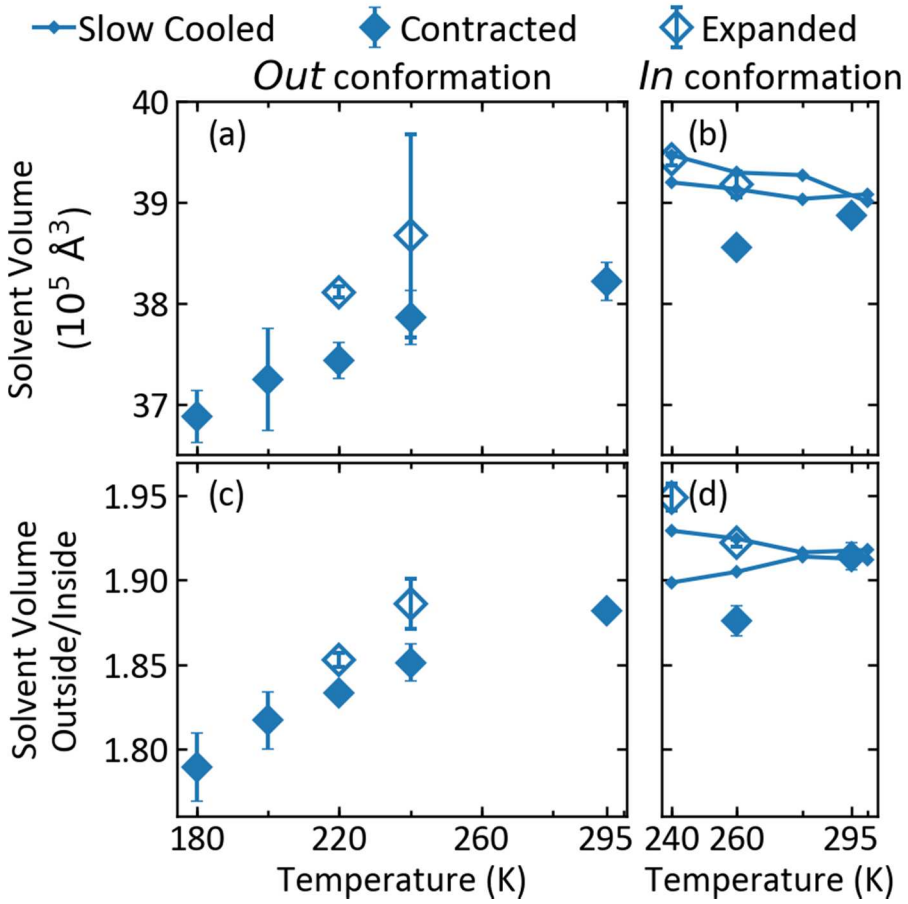


Figure 4.9 (a)(b) Total solvent cavity volume in apoferritin's cubic unit cell and (c),(d) ratio of solvent cavity volume located outside the apoferritin shell to the volume inside, versus temperature for glycerol-free crystals having GLN82 in the "out" configuration (a,c) and the "in" configuration (b,d) respectively. Closed and open symbols indicate values determined before and after unit cell expansion, respectively. Solid lines indicate values obtained from slowly cooled (0.1 K/s) rather than abruptly cooled (~300 K/s) crystals. The solvent cavity volume and the fraction of the total cavity volume located outside the apoferritin shell drops with decreasing temperature, and both increase during cold cell expansion. This suggests that the primary structural changes during cold expansion occur at the interface between apoferritin shells rather than within them.

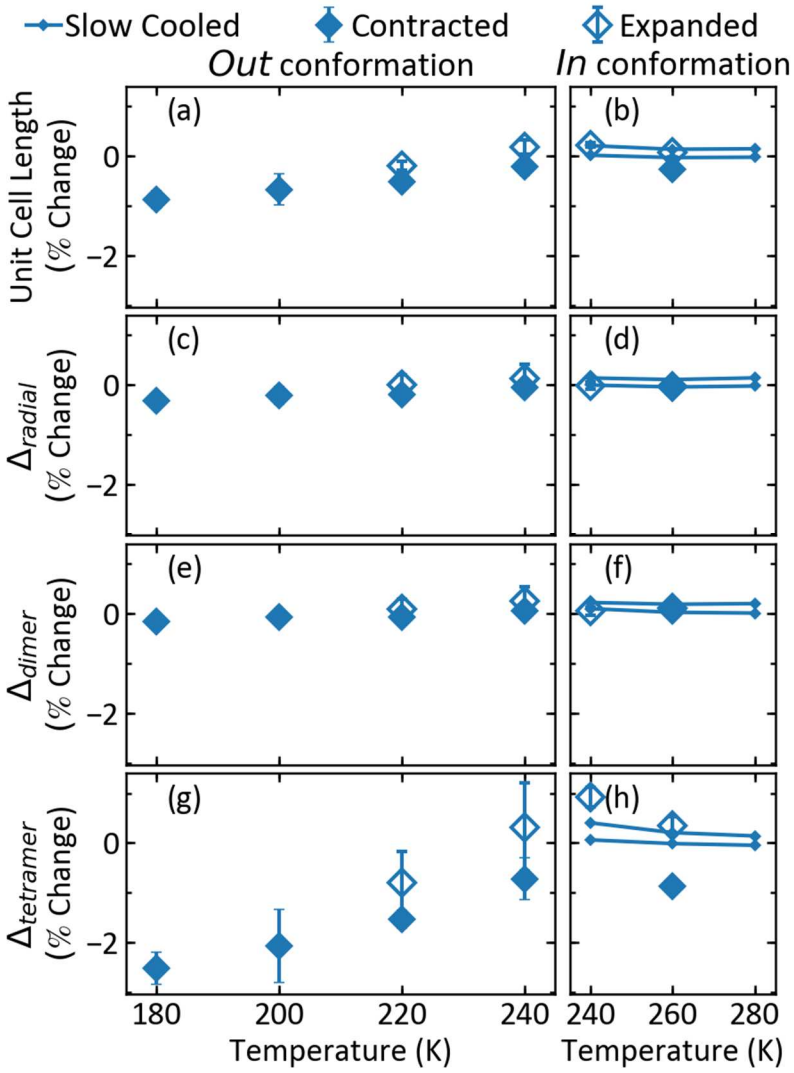


Figure 4.10 Percent change in (a) linear dimension of the cubic unit cell, (b) radial distance from the centre of volume of the apoferritin shell to the centre of volume of a ferritin monomer within the shell, (c) distance between centres of volume of ferritin monomers within a dimer; and (d) distance between centres of volume of adjacent dimers in adjacent shells (roughly, the distance between protein shells), versus temperature for glycerol-free cubic apoferritin crystals. Results for crystals with GLN82 in the "out" and "in" conformations are shown separately. Closed and open symbols indicate values before and after cold expansion, respectively, and solid lines represent values obtained from crystals that were slowly rather than abruptly cooled. Expansion has by far the largest effect on the shell-shell distance, with little effect on dimensions within the shell.

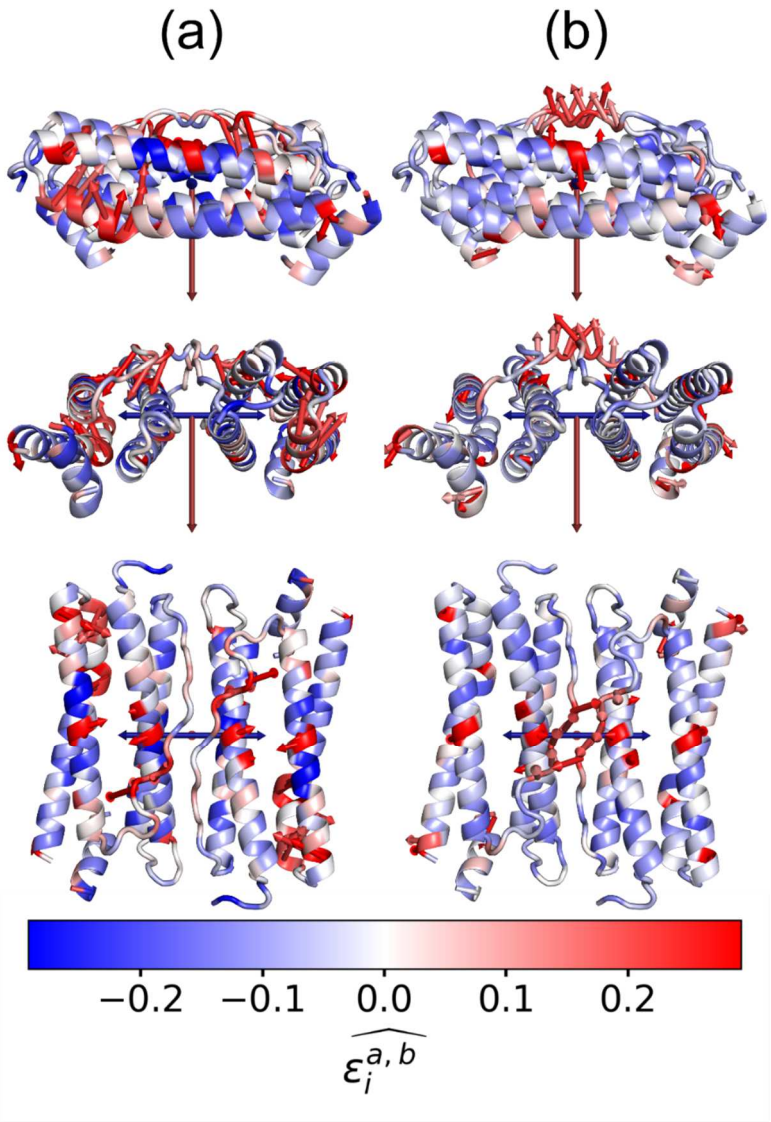


Figure 4.11 Atomic level changes to apoferritin monomers in a dimer pair (a) during cooling from room temperature to 220 K and (b) during post-cooling expansion at T=220K. Only structures with GLN82 in the "out" conformation are used. The blue arrowed line connects the centres of volumes of the monomers and the long red arrow points in the direction of the centre of the spherical apoferritin shell (comprised of 24 monomers). Colouring indicates the normalized displacements $\widehat{\varepsilon_i^{a,b}}$ between atomic positions in the compared structures, calculated using Eq. 4.7. Short red arrows indicate the direction of the displacements, and arrow lengths are 50 times larger than the actual displacements. The overall shift of atomic positions due to the decrease in separation of monomers on cooling and the increase during expansion has been subtracted out and is not reflected in the colour/arrow scheme. The mean and standard deviation of the normalized displacements $\widehat{\varepsilon_i^{a,b}}$ are (a) 0.36 ± 0.15 and (b) 0.21 ± 0.11.

4.3.5 Solvent flows, cooling induced disordering, and cold reordering

The present results provide the clearest evidence to date for the role of solvent flows in causing cooling-induced disorder in protein crystals. During cooling, the unit cell volume generally contracts more than the protein volume, so that the solvent cavity volume shows the largest fractional decrease (Juers & Matthews, 2001, 2004; Juers *et al.*, 2018). Unless the solvent internal to the crystal contains large (~30% v/v or more) cryoprotectant concentrations, its volume will contract less than the solvent cavities. If cooling times are short, the exiting solvent cannot make its way to the crystal surface and the lattice of protein molecules must undergo elastic deformations and plastic failure to accommodate it.

Here we observe a substantial expansion – often to values larger than at room temperature – of the crystal lattice and a concurrent crystal reordering, as reflected in a drop in mosaicity and also sometimes in *B* factor, at *constant* temperatures between 220 and 260 K. All external solvent was removed from the crystals (essential to allow cooling below 260 K without ice formation), and an oil coating prevented all solvent transport between the crystal and the surrounding dry nitrogen gas cryostream. Consequently, the solvent required to fill the expanding solvent cavities must come from inside the crystal and nowhere else. The solvent itself could expand if, e.g., there was a change in water structure toward the open structure of ice, but this would be expected to occur on timescales far shorter than the tens of seconds required for expansion. We thus conclude that the solvent that fills the expanding solvent cavities within the unit cells flows from elastically and plastically deformed crystal regions having excess solvent.

Evidence for the role of solvent transport in unit cell expansion may be provided by the temperature- and glycerol-dependent timescale of the unit cell expansion following cooling, shown in Fig. 4.7. The viscosity of bulk glycerol-water mixtures increases by roughly an order of magnitude, for temperatures between 220 and 260 K, as the glycerol concentration increases from 0% to 40% v/v, and by a similar factor for concentrations in this range on cooling from 260 K to 220 K (Trejo González *et al.*, 2011). These trends are qualitatively consistent with the observed expansion timescales, although viscosities should be

modified by nanoconfinement within the protein structure, and other factors (e.g., thermally activated structural relaxations) could affect expansion timescales.

Additional evidence for the role of solvent transport comes from how cold expansion affects mosaicities. For crystals soaked in glycerol solutions with concentrations of 20% and below, cold cell expansion causes mosaicities to drop; for crystals soaked in 40% solutions, mosaicities frequently increased or showed no change during expansion. For glycerol concentrations below ~20-30% v/v the bulk solvent contraction is smaller than that of the solvent cavities occurring during abrupt cooling, so solvent should initially be expelled from the unit cells, increasing mosaicity, and then re-enter during cold expansion, decreasing mosaicity. For crystals soaked in 40% v/v glycerol, the bulk solvent contraction is larger than the initial solvent cavity contraction, so during cooling solvent flows into the unit cell, depleting it from other crystal regions and disordering them. During cold expansion, even more solvent must flow into the unit cell, further depleting other crystal regions and disordering them, so the mosaicity should increase. The actual flows into and out of the unit cell are somewhat uncertain because both the internal glycerol concentration and solvent contraction are modified by preferential hydration of and nanoconfinement by protein surfaces.

Fig. 6.6.3 shows estimates of the volume fraction of internal solvent present at room temperature that cannot be accommodated within the solvent cavity volume deduced from refined structures of the initial cold crystal and the crystal after unit cell expansion. The solvent volume at each temperature was estimated as described in (Moreau *et al.*, 2019) by assuming the first hydration layer does not contract and that the remaining solvent has the same contraction as the bulk liquid. The volumes of solvent transported during both cooling and cold expansion are at least a few percent of the crystal volume. Since the crystals were typically a few hundred micrometres in size, were that much solvent to be present at the crystal surface, it would be easily visible and would (for glycerol concentrations of 20% v/v and below) crystallize immediately. We are thus confident that most of the solvent transported during cooling and

cold expansion remains within the crystal, in defective regions that contribute to the observed rise in mosaicity.

Indirect evidence for the role of solvent transport in creating protein crystal disorder has been provided by X-ray topography, which has directly imaged the mosaic domain structure that forms on cooling (Kriminski *et al.*, 2002). The occasional success of "cryoannealing" (Harp *et al.*, 1999), in which brief warming of a cold crystal to room temperature and then re-cooling sometimes leads to reductions in mosaicity and improvement in overall crystal order, may be due in part to adjustment of the internal water content of the crystal via transport of solvent to or from the substantial bulk solvent that surrounds crystals in typical mounting practice (Juers & Matthews, 2004). The brief warming may also allow relaxation of elastically deformed lattice regions and healing of smaller plastically deformed regions, leaving a smaller number of plastically deformed regions to take up excess solvent on re-cooling and thereby reducing average disorder.

Although solvent transport into and out of unit cells correlates with changes in crystal mosaicity during cold expansion, we cannot conclude based on the present evidence that they are solely responsible for these changes, or for mosaicity increases during cooling. Kinetic trapping of incompletely relaxed protein and lattice conformations during abrupt cooling should, even in the absence of differential contraction between solvent cavities and internal solvent, create disorder that increases mosaicity relative to both its room temperature and fully relaxed cold crystal states, and that increases B factors relative to the relaxed cold state. However, the magnitudes of the estimated solvent flows – at least a few percent of the total solvent volume – are so large that their contribution to disorder must be substantial.

4.3.6 Correlation between solvent contraction, unit cell contraction, and crystal mosaicity

Direct efforts to demonstrate a relation between cold crystal mosaicity and the expected contraction of the internal solvent on cooling between room temperature and T=77 K or 100 K have so far yielded

unconvincing results. Juers *et al.* (Juers *et al.*, 2018) found no clear variation of mosaicity with bulk solvent contraction for ice-free crystals of at least seven out of nine different protein crystal systems; only α-lactalbumin showed weakly suggestive evidence of a mosaicity minimum versus unit cell contraction. However, in those careful experiments any trends may have been obscured by the relatively large lower bound on measurable mosaicity due to the large incident beam divergence of the laboratory X-ray source used. Our recent data for crystals plunge cooled in liquid nitrogen and measured at T=100 K suggests a decrease in mosaicity with increasing glycerol concentration for both thaumatin and apoferritin, with crystals soaked in 40% v/v glycerol giving the lowest values (Moreau *et al.*, 2019). But for both apoferritin (Fig. 4.4) and thaumatin crystals cooled in gas streams to temperatures between 180 K and 260 K, glycerol-free crystals typically show the smallest cold mosaicities. Disorder due to kinetic trapping of incompletely relaxed protein conformations may impose a lower bound on low-temperature mosaicities, and this disorder might grow as glycerol concentrations increase.

There is a reliable correlation between unit cell contraction on cooling and the bulk contraction of the solution in which the crystals are grown or soaked. Juers *et al.* observed a near linear relation between unit cell contraction between room temperature and T=100 K and the bulk solvent contraction in eight different protein crystal systems (Juers *et al.*, 2018). The present data for apoferritin in Fig. 3 and data for thaumatin (Moreau, *et al.*, 2019) show that both the unit cell and solvent cavity volume contractions increase with increasing glycerol concentration at fixed temperature between 200 and 260 K.

However, as shown in Fig. 6, for apoferritin crystals that are slowly cooled to temperatures above 200 K or that have undergone cold expansion, the net unit cell contraction on cooling is independent of glycerol concentration and thus solvent contraction.

The different behaviours versus glycerol concentration / bulk solvent contraction observed during abrupt cooling, slow cooling, and cold expansion suggest one possible reason why a simple correlation between mosaicity and solvent contraction has not been observed. Protein crystals can be considered as

poroelastic materials (Biot, 1941). As discussed in detail by Juers *et al.* (Juers *et al.*, 2018), excess solvent expansion or contraction relative to the solvent cavities generates excess pressure that, on short time scales, will drive cavity expansion and, on long time scales, will drive solvent flow. Consequently, the strongest correlation between solvent cavity contraction and solvent contraction (and between unit cell and solvent contractions) on cooling should be observed when crystals are cooled so fast to temperatures well below the internal solvent's glass transition temperature (e.g., well below ~200 K) that little flow can occur before the solvent vitrifies. When crystals are cooled very slowly or held at fixed temperatures above ~200 K, cavity and solvent contraction should be uncorrelated. This neglects any chemical effects of different solvent compositions on, e.g., protein conformation and contraction behavior.

4.3.7 Why does the unit cell contract and then expand?

For cubic apoferritin crystals, initial unit cell volumes obtained following fast cooling are substantially smaller than those obtained via slow cooling to the same temperature, or by allowing crystal relaxation at fixed temperature above 200 K. This implies that the initial fast-cooled state reflects a kinetically favoured but non-equilibrium configuration. During cooling, amplitudes of atomic motions within local energy minima decrease and minor conformers separated by small barriers from major conformers are depopulated. The timescales for these relaxations are short compared with the time interval during fast cooling that the sample remains above the protein-solvent glass transition. These local relaxations should in general lead to contraction of the protein and of the crystal lattice.

However, larger scale, cooperative changes in protein conformation and lattice packing toward new equilibrium configurations will in general occur much more slowly. These changes will be driven in part by temperature dependent changes in interaction strengths. For example, as temperature decreases, pH, pKas of side chains, and water activity all change. The hydrophobic interaction largely responsible for protein folding and for the formation of some crystal contacts weakens, due to increased tetrahedral ordering of water and a reduction in the entropic cost of solvent “caging” around hydrophobic residues,

and this can lead to cold unfolding/denaturation of proteins in solution (Privalov, 1990). Cooperative structural changes have much slower kinetics than the quenching of local motions, especially within the constrained environment of the crystal, and so limited evolution toward new minima can occur during fast cooling. The present results show that, as long as the internal solvent remains liquid, substantial cooperative evolution of the protein and its lattice toward a new temperature-dependent equilibrium can occur at temperatures as low as 220 K.

Several other mechanisms that could give rise to cold unit cell expansion can be ruled out. First, unit cells can expand when internal ice forms. For glycerol concentrations below 20% v/v, diffraction from internal ice is sharp and easily detected and its intensity rapidly saturates after nucleation. Internal ice was not observed (Fig. 4.2), and cannot explain observed expansions in nominally ice-diffraction-free crystals. For larger glycerol concentrations, ice grows more slowly and the ice grain size becomes much smaller, so that initially weak and diffuse ice diffraction may be difficult to detect in the diffuse diffraction background.

Second, migration of excess solvent initially present at the surface of the crystal or from ambient humid gas to the crystal interior is known to cause crystal expansion at and near room temperature. This is believed to be the most important mechanism involved in "cryo-annealing" protocols (Harp *et al.*, 1999; Juers & Matthews, 2004). However, for all crystals reported here, external solvent was carefully removed and replaced with oil. Any substantial volumes of surface solvent (estimated using optical measurements to be much less than 0.1% of the crystal volume) rapidly formed ice I_h, especially for glycerol-free crystals, for which we observed the greatest reductions in mosaicity and *B* factor during cell expansion. The surrounding oil provides an effective barrier to water diffusion to or from the ambient gas. The ambient gas was the cold, dry, flowing N_2 of the cryostream that, when inadequate oil was present, rapidly dehydrated the crystals and caused the unit cells to shrink, not expand.

Third, initial solvent contraction might occur if the pressure within the crystal during cooling became sufficient to drive the solvent into a high density amorphous (HDA) ice (at low temperature) or a high density liquid (HDL) phase. The observed position of the solvent (and NVH oil) diffraction peaks (Fig. 4.2) remain constant during the cell expansion, and in glycerol-free crystals is consistent with the expected diffraction from normal (low density) water or LDA. The solvent peak position does not shift and is unambiguously inconsistent at all times with the peak position expected for HDA (Kim, *et al.,* 2008).

Fourth, a temperature-dependent chemical potential for glycerol or salt within the crystal could lead to transient changes in unit cell following cooling. For example, if preferential hydration and glycerol exclusion increase with decreasing temperature, then diffusion of glycerol out of and water into the crystal following cooling could lead to changes in unit cell. The sign of such changes is not obvious.

Finally, we note that not all apoferritin crystals examined at temperatures between 220 K and 260 K showed cold expansion on the 200 s timescale of our experiments. For those that did not, both the unit cell contraction and the mosaicity increase on cooling were much larger (e.g. a factor of 2 at 240 K) than for nominally identically prepared crystals that showed cold expansion. The difference in behaviour could be due to slight dehydration by the dry N_2 gas stream during cooling if, e.g., the oil thinned near a crystal corner or edge, or due to some difference in crystallization conditions.

4.4 CONCLUSIONS

The present results provide the clearest evidence to date for internal solvent transport in protein crystals during cooling, and insight into the roles of this transport and of kinetically quenched structural relaxations in generating crystal disorder during cooling. During abrupt cooling, local vibrations and small conformational motions freeze out, leading to contraction of the unit cell and of the solvent cavities within it. At ambient pressure, the equilibrium volume of the internal solvent generally contracts by a different percentage than the equilibrium volume of the solvent cavities themselves. For extremely large (>10^4 K/s)

cooling rates, solvent flow during the cooling time before vitrification should be limited, internal pressure within the solvent cavity should grow, and solvent cavity and solvent contractions should track. For cooling at rates typically achieved using cold gas streams or in hand plunging in liquid nitrogen (~10^2 – 10^3 K/s), solvent flows into or out of unit cells until it vitrifies, reducing or eliminating the correlation between solvent cavity and solvent contractions, but creating elastic and plastic deformations in regions where excess solvent accumulates (or from which it is depleted) that increase crystal disorder. At much slower cooling rates (<1 K/s), solvent may have time to flow to or from the crystal surface without substantial accumulation within / depletion from internal regions and without associated increases in crystal disorder.

Abrupt cooling quenches slower timescale, larger length scale cooperative relaxations of the protein-solvent system as it attempts to evolve toward its temperature-dependent equilibrium, and this also creates crystal disorder. However, at final temperatures near and above the protein-solvent glass transition temperature (~200 K) and as long as the internal solvent remains liquid, substantial evolution of both conformation and lattice packing can occur, on timescales of seconds to tens of seconds that increase with decreasing temperature. In the case of apoferritin, these lead to a reversal of solvent flows and a dramatic reordering of the crystal at fixed low temperature, including in crystals containing no cryoprotectants other than salt present in the mother liquor.

Perhaps the most interesting – if obvious - conclusion from the present work is the importance of time in variable/multi-temperature studies of protein structure. Although short equilibration times (e.g., a few seconds) at a given temperature should be sufficient to examine freeze-out of side chain conformers, exploring cooperative conformational changes will generally require much longer equilibration times.

The potential of protein crystals for study of cold denaturation is particularly intriguing. Most proteins are expected to unfold at sufficiently low temperature, and this may be a major issue in cold storage and recovery of proteins and biologic medications and in cryopreservation of cells and tissues. Studies of unfolding in solution have been constrained by the formation of ice near 273 K, and so have often required

the addition of denaturants like urea to promote unfolding at higher temperatures. Nanoconfined water within protein crystals can be maintained in a (supercooled) liquid state at temperatures down to ~200 K for tens of seconds or more using at most small cryoprotectant concentrations. While full unfolding cannot occur within the constraints of the protein lattice, cooperative conformation changes associated with weakening of the hydrophobic interaction that gives rise to unfolding should be observable, and may allow differences between proteins in low temperature stability to be evaluated.

Acknowledgements All X-ray data collection was performed at the Cornell High Energy Synchrotron Source (CHESS), which is supported by the National Science Foundation under award DMR-1332208, using the Macromolecular Diffraction at CHESS (MacCHESS) facility, which is supported by award GM-103485 from the National Institute of General Medical Sciences, National Institutes of Health. This work was supported by the NSF under award MCB-1330685. DWM acknowledges additional support from Cornell University's Molecular Biophysics Training Grant (NIH T32GM0082567) and from the NIH (R01-GM127528). RET acknowledges a financial conflict of interest, as some of the tools used in this work were provided by MiTeGen, LLC, in which he has a significant financial interest.

References

Adams, P. D., Afonine, P. V., Bunkóczi, G., Chen, V. B., Davis, I. W., Echols, N., Headd, J. J., Hung, L. W., Kapral, G. J., Grosse-Kunstleve, R. W., McCoy, A. J., Moriarty, N. W., Oeffner, R., Read, R. J., Richardson, D. C., Richardson, J. S., Terwilliger, T. C. & Zwart, P. H. (2010). *Acta Cryst. D*. **66**, 213–221.

Atakisi, H., Moreau, D. W. & Thorne, R. E. (2018). *Acta Cryst. D*. **74**, 264–278.

Biot, M. (1941). *J. Appl. Phys.* **12**, 155–164.

Chen, V. B., Arendall, W. B., Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., Murray, L. W., Richardson, J. S. & Richardson, D. C. (2010). *Acta Cryst. D*. **66**, 12–21.

Doster, W. (2010). *BBA - Proteins Proteomics*. **1804**, 3–14.

Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. (2010). *Acta Cryst. D*. **66**, 486–501.

Fahy, G. M. & Wowk, B. (2015). *Cryopreservation and Freeze-Drying Protocols*, Vol. *1257*, edited by W.F. Wolkers & H. Oldenhof, pp. 21–82. New York: Springer.

Fenwick, R. B., van den Bedem, H., Fraser, J. S. & Wright, P. E. (2014). *Proc. Natl. Acad. Sci. U. S. A.* **111**, E445-54.

Fraser, J. S., van den Bedem, H., Samelson, A. J., Lang, P. T., Holton, J. M., Echols, N. & Alber, T. (2011). *Proc. Natl. Acad. Sci. U. S. A.* **108**, 16247–16252.

Glycerine Producers Association (1963). Physical Properties of Glycerine and Its Solutions Glycerine Producers' Association.

Hare, D. E. & Sorensen, C. M. (1987). *J. Chem. Phys.* **87**, 4840–4845.

Harp, J. M., Hanson, B. L., Timm, D. E. & Bunick, G. J. (1999). *Acta Cryst. D*. **55**, 1329–1334.

Hempstead, P. D., Yewdall, S. J., Fernie, A. R., Lawson, D. M., Artymiuk, P. J., Rice, D. W., Ford, G. C. & Harrison, P. M. (1997). *J. Mol. Biol.* **268**, 424–448.

Juers, D. H., Farley, C. A., Saxby, C. P., Cotter, R. A., Cahn, J. K. B., Holton-Burke, R. C., Harrison, K. & Wu, Z. (2018). *Acta Crystallogr. Sect. D Struct. Biol.* **74**, 922–938.

Juers, D. H. & Matthews, B. W. (2001). *J. Mol. Biol.* **311**, 851–862.

Juers, D. H. & Matthews, B. W. (2004). *Acta Cryst. D*. **60**, 412–421.

Kabsch, W. (2010). *Acta Cryst. D*. **66**, 125–132.

Keedy, D. A., Hill, Z. B., Biel, J. T., Kang, E., Rettenmaier, T. J., Brandão-Neto, J., Pearce, N. M., von Delft, F., Wells, J. A. & Fraser, J. S. (2018). *Elife*. **7**, 1–36.

Keedy, D. A., Kenner, L. R., Warkentin, M., Woldeyes, R. A., Hopkins, J. B., Thompson, M. C., Brewster, A. S., Van Benschoten, A. H., Baxter, E. L., Uervirojnangkoorn, M., McPhillips, S. E., Song, J., Alonso-Mori, R., Holton, J. M., Weis, W. I., Brunger, A. T., Soltis, S. M., Lemke, H., Gonzalez, A., Sauter, N. K., Cohen, A. E., Van Den Bedem, H., Thorne, R. E. & Fraser, J. S. (2015). *Elife*. **4**, 07574.

Kriminski, S., Caylor, C. L., Nonato, M. C., Finkelstein, K. D. & Thorne, R. E. (2002). *Acta Cryst. D*. **58**, 459–471.

Lang, P. T., Ng, H.-L., Fraser, J. S., Corn, J. E., Echols, N., Sales, M., Holton, J. M. & Alber, T. (2010). *Protein Sci.* **19**, 1420–1431.

Li, A. J. & Nussinov, R. (1998). *Proteins Struct. Funct. Genet.* **32**, 111–127.

Loerting, T., Bauer, M., Kohl, I., Watschinger, K., Winkel, K. & Mayer, E. (2011). *J. Phys. Chem. B*. **115**, 14167–14175.

Moreau, D. W., Atakisi, H. & Thorne, R. E. (2019). *IUCrJ*. **(in press)**,.

Privalov, P. L. (1990). *Crit. Rev. Biochem. Mol. Biol.* **25**, 281–305.

Ringe, D. & Petsko, G. A. (2003). *Biophys. Chem.* **105**, 667–680.

Shen, C., Julius, E. F., Tyree, T. J., Moreau, D. W., Atakisi, H. & Thorne, R. E. (2016). *Acta Cryst. D*. **72**, 742–752.

Teeter, M. M., Yamano, A., Stec, B. & Mohanty, U. (2001). *Proc. Natl. Acad. Sci.* **98**, 11242–11247.

Tilton, R. F., Dewan, J. C. & Petsko, G. A. (1992). *Biochemistry*. **31**, 2469–2481.

Trejo González, J. A., Longinotti, M. P. P., Corti, H. R., Trejo González, J. A., Longinotti, M. P. P. & Corti, H. R. (2011). *J. Chem. Eng. Data*. **56**, 1397–1406.

Tyree, T. J., Dan, R. & Thorne, R. E. (2018). *Acta Cryst. D*. **74**, 471–479.

De Val, N., Declercq, J. P., Lim, C. K. & Crichton, R. R. (2012). *J. Inorg. Biochem.* **112**, 77–84.

Vinothkumar, K. R. & Henderson, R. (2016). *Q. Rev. Biophys.* **49**, e13.

Voss, N. R. & Gerstein, M. (2010). *Nucleic Acids Res.* **38**, 555–562.

Warkentin, M., Badeau, R., Hopkins, J. B. J. B. & Thorne, R. E. (2012). *Acta Crystallogr. Sect. D*. **68**, 1108–1117.

Warkentin, M., Hopkins, J. B., Badeau, R., Mulichak, A. M., Keefe, L. J. & Thorne, R. E. (2013). *J. Synch. Rad.* **20**, 7–13.

Warkentin, M. & Thorne, R. E. (2009). *J. Appl. Crystallogr.* **42**, 944–952.

Warkentin, M. & Thorne, R. E. (2010*a*). *Acta Cryst. D*. **66**, 1092–1100.

Warkentin, M. & Thorne, R. E. (2010*b*). *J. Struct. Funct. Genomics*. **11**, 85–89.

Weik, M., Kryger, G., Schreurs, A. M. M., Bouma, B., Silman, I., JL, S., Gros, P. & Kroon, J. (2001). *Acta Cryst. D*. **57**, 566–573.

Weik, M., Ravelli, R. B. G., Silman, I., Sussman, J. L., Gros, P. & Kroon, J. (2001). *Protein Sci.* **10**, 1953–1961.

5 ICE IN BIOMOLECULAR CRYOCRYSTALLOGRAPHY

Abstract Diffraction data acquired from cryocooled protein crystals often includes diffraction from ice. Analysis of ice diffraction data from crystals of three proteins shows that ice formed within solvent cavities during rapid cooling is comprised of a stacking-disordered mixture of hexagonal and cubic planes, with the cubic plane fraction increasing with increasing cryoprotectant concentration and increasing cooling rate. Building on the work of Thorn et al., we define a revised metric for detecting ice from deposited protein structure factor data, and validate this metric using full frame diffraction data from the Integrated Resource for Reproducibility in Macromolecular Crystallography (IRRMC). Using this revised metric, our analysis of structure factor data from a random sample of 83,938 PDB entries indicates that roughly 16% show evidence of ice contamination, and that this fraction increases with increasing solvent content and maximum solvent cavity size. By examining the ice diffraction peak positions at which structure factor perturbations are observed, we find that roughly 20% of crystals exhibit ice with primarily hexagonal character, indicating that inadequate cooling rates and/or cryoprotectant concentrations were used, while the remaining 80% show ice with a stacking disordered or cubic character. Refined electron density maps deduced using model data with added ice contamination show increased noise, higher R factors and increased number of modelled solvent atoms.

5.1 INTRODUCTION

Ice diffraction frequently contaminates diffraction data collected from biomolecular crystals at cryogenic temperatures. Ice may form during crystal cooling, in solvent present within the crystal's solvent cavities or in residual solvent on the crystal surface. Ice may also appear as contaminating frost on the sample or sample holder surface, due to exposure to moist ambient air during handling or data collection, or from accumulated frost in the liquid nitrogen used to initially cool and to store the crystals. Ice that forms from solvent confined to the crystal's solvent cavities, from bulk-like solvent containing substantial cryoprotectant, or from bulk-like solvent that is rapidly cooled is typically highly polycrystalline, producing continuous and largely isotropic ice rings. Ice that forms from bulk-like solvent containing little cryoprotectant or that is cooled slowly tends to be comprised of fewer, larger crystals, producing discrete ice diffraction peaks or "lumpy", anisotropic, quasi-continuous diffraction rings. Frost is typically comprised of large, dendritic single crystals.

When only a small number of large ice crystals are present in the X-ray beam (as is often the case with frost) ice diffraction may manifest as discrete, isolated ice diffraction peaks. These peaks are the dominant type of "zinger" seen on diffraction frames. When such a peak overlaps a protein diffraction peak, the measured intensity will be larger than of symmetry related reflections and/or will be larger than predicted from Wilson statistics. This allows them to be identified as outliers and rejected by standard diffraction frame merging or scaling software (Read, 1999).

When many ice crystals are present in the X-ray beam, ice diffraction consists of continuous or quasi-continuous rings that overlap the protein Bragg peaks and interfere with the background subtraction process when the Bragg peaks are integrated. The integration procedure sums the pixel values selected as being associated with the Bragg peak, which includes X-ray counts from diffraction and scattering sources other than the long-range ordered component of the protein crystal. For each individual protein

Bragg peak, background subtraction algorithms estimate the X-ray counts from these sources based on pixel values near to but not associated with that Bragg peak. Each integration program estimates the background counts from these pixel values in a slightly different manner. XDS assumes a constant background beneath each protein Bragg peak equal to the average value of the neighbouring pixels (Kabsch, 2010), MOSFLM fits a plane to the neighbouring pixels (Leslie, 2006) and DIALS provides options to model the logarithm of the background as a constant or a plane (Parkhurst *et al.*, 2016). When ice rings overlap or are near the protein Bragg peak, the true background underneath the protein Bragg peak is no longer adequately modelled by a constant value or by a linear plane. In the case of a direct overlap, the ice diffraction will be weaker, or not present, in the neighbouring pixels used to estimate the background. The estimated background will then be considerably smaller than the true background, leading to an overestimate of the protein Bragg peak's true value. If the ice ring is near to but does not overlap the protein Bragg peak, its can still adversely affect the background estimation procedure. In this case, ice diffraction in neighbouring pixels used to estimate the background will be stronger than ice diffraction overlapping the Bragg peak. This leads to an overestimate of the background and in turn, an underestimate of the protein Bragg peak. Parkhurst et al. (2017) illustrates these two scenarios in their Fig. 2 and thoroughly explains how ice rings lead to incorrect estimates of the background.

A background subtraction algorithm providing much improved management of ice rings – the Global Background Model (GBM) – has been developed and implemented within the DIALS package (Parkhurst *et al.*, 2017). Instead of a constant or planar background, the GBM method does a pixel-by-pixel average of the background across all diffraction frames in a data set (excluding those frames where the pixel is part of a protein Bragg peak but including frames where it is part of ice diffraction), median filters these averaged values in azimuthal rings at each resolution, and then scales this average background image to match the observed background around each individual Bragg reflection during integration of that Bragg reflection. The use of a single scaling parameter (for each Bragg reflection) allows for fast integration and

could prevent overfitting of the background. This background-subtraction algorithm greatly reduces ice-related biasing of ice ring intensities. It works particularly well when the ice diffraction has the form of homogeneous, isotropic rings, and less well when the ice rings are “lumpy” or in general have structure that varies azimuthally and between frames.

Here we examine the nature and effects of ice on diffraction data, modelling and refinement in protein cryocrystallography. We first examine ice diffraction in diffraction frames we have collected from crystals of three proteins and in reference diffraction frame data obtained from the Integrated Resource for Reproducibility in Macromolecular Crystallography (IRRMC), (Grabowski *et al.*, 2016). Fits to ice diffraction from ice formed inside protein crystals indicate that the ice is not purely hexagonal, purely cubic, or simple mix of the two. Instead, the predominant form of ice is a stacking disordered mix of cubic and hexagonal planes, with the cubic fraction increasing from ~50% with increasing cryoprotectant concentration (Moreau *et al.*, 2019). In contrast, ice diffraction from ice formed in drops of aqueous cryoprotectant solutions at lower cryoprotectant concentrations and lower cooling rates and/or at higher temperatures has a primarily hexagonal character; as cryoprotectant concentration increases, cooling rate increases and/or the temperature of ice formation decreases, the diffraction develops a stacking disordered character with increasing cubic fraction.

We then extend methods developed by Thorn *et al.* (2017) for detecting the presence of ice based on experimental protein structure factor data alone (i.e., without access to the full diffraction frames) and obtain more accurate ice detection with far fewer false positives and false negatives. Using our revised ice detection metric, we find that roughly 16% of PDB entries show structure factor perturbations from ice contamination, consistent with results from Thorn *et al*. Of these, roughly 20% show evidence of ice contamination at the positions of all hexagonal ice peaks, and for the remaining 80% the observed peaks are consistent with stacking disordered ice with a substantial cubic fraction. To investigate the effects of structure factor biasing by ice, we begin with ice-free detector diffraction images, add varying amounts of

ice diffraction directly to these images, and then determine refined models using these "contaminated" diffraction images. Comparison of these "contaminated" models and corresponding maps with those from the original ice-free data indicates an isotropic addition of noise that can lead to additional modelled waters and possibly also errors in ligand identification.

5.2 METHODS

5.2.1 Crystal growth and preparation

Crystal growth and preparation of equine spleen apoferritin, thaumatin and tetragonal hen egg white lysozyme crystals for X-ray diffraction measurements are described in (Moreau *et al.*, 2019). Crystals of equine spleen apoferritin (Sigma, catalog No. A-3641) were grown in hanging drops consisting of 2 ml of protein at 10 mg ml^{-1} in 0.1 M sodium acetate buffer pH 6.5 and 2 ml of a well solution consisting of 2%(w/v) $CdSO_4$ and 15%(w/v) $(NH_4)_2SO_4$ in the same buffer. Cubic crystals in space group F432 grew to dimensions of 300–500 µm within one week.

Crystals of thaumatin (Sigma, catalog No. T7638) were grown in hanging drops comprised of equal volumes of protein at 40 mg ml^{-1} in 0.1 M sodium acetate buffer pH 6.5 and a well solution consisting of 14%(w/v) potassium sodium tartrate in the same buffer. Tetragonal crystals in space group $P4_12_12$ grew to dimensions of 200–300 µm within one week.

Crystals of lysozyme (Sigma, catalog No. L6876) were grown in hanging drops comprised of equal volumes of protein at 80 mg ml^{-1} in 0.1 M sodium acetate buffer pH 5.2 and a well solution consisting of 2.5%(w/v) NaCl in the same buffer. Tetragonal crystals in space group $P4_32_12_1$ grew to dimensions of 300–800 µm. Crystals appeared within one week and stopped growing within four weeks.

Crystals were used as grown, or else soaked in solutions containing 10%, 20%, or 40% v/v glycerol. Crystals were then transferred to a separate drop of NVH oil (Cargille) and manipulated until all external

solvent was removed from their surface, as indicated by a near disappearance of the crystal due to the close match between its refractive index and that of the oil (Warkentin & Thorne, 2009). Crystals were mounted on microfabricated loops in a spherical blob of NVH oil to prevent dehydration during data collection and stored in MicroRT tubes (MiTeGen) containing mother liquor or cryoprotectant solution for ~1 hour prior to data collection.

5.2.2 X-ray diffraction data collection

X-ray data was collected on station F1 at the Cornell High-Energy Synchrotron Source (CHESS) using a PILATUS 6M detector. At the time of the experiments, the F1 station delivered a Gaussian beam with a 65 µm FWHM, a divergence of 0.03°, a photon energy of 12.7 keV, and a flux of 2.2×10^9 ph/s. A cold nitrogen-gas stream (Oxford Cryosystems Cryostream 700) with a flow rate of 5 L/min and programmed to the desired final sample temperature was directed at the crystal. The cold gas stream was initially blocked using an air blade shutter, each crystal was placed in the X-ray beam at room temperature, and ten frames totalling 5^o in rotation were collected to assess the crystal for damage and dehydration. The gas stream was then unblocked, cooling the crystal to its final temperature in ~1 s or less. Additional diffraction data at T = 100 K was collected at the CHESS F1 station from crystals plunge cooled into liquid nitrogen. For all crystals, solvent initially present at the crystal surface was carefully removed by transfer to and manipulation in high viscosity NVH oil (Cargill), as discussed in in (Warkentin & Thorne, 2009), until the crystal almost disappeared in the oil (due to the close refractive index match between the crystal and oil) indicating complete removal of the external solvent

Additional X-ray diffraction data on cryocooled glycerol / water solutions was collected using the same CHESS station and experimental protocols. Samples were prepared by injecting ~10 nl of solution into a 250 µm diameter, ~2 cm long thin-walled polyester tube. The length of tubing filled with solution was

approximately 200 µm. The tube was then affixed to a goniometer base (Mitegen GB-B1A) and centered in the X-ray beam. Fig. 6.4.1 shows an image of the sample in the beam path.

5.2.3 Diffraction image sourcing

A total of 30 sets of raw diffraction frames, each showing clear visual evidence of ice diffraction, were downloaded from the IRRMC. All data sets were recorded on Dectris Pilatus 6M detectors. Of these data sets, 21 were recorded using beamline BL11-1 at SSRL, 5 using beamline BL12-2 at SSRL, 2 using beam line 19-ID at APS, 1 using beamline 5.0.1 at ALS and 1 using beamline I04-1 at Diamond. These 30 data sets include the 13 data sets used by (Parkhurst *et al.*, 2017).

5.2.4 Zinger Identification

"Zingers" were identified in diffraction frames as pixels, not associated with protein diffraction, that had values approximately five times larger than the local background intensity. Potential protein Bragg peaks were masked within a 6-pixel radius circle based on peak locations reported in the XDS output file XDS_ASCII.HKL. Low resolution Bragg peaks and Bragg peaks near the beam stop shadow, rotational axis, and detector panel segments were occasionally excluded in the XDS output file. Regions corresponding to resolutions numerically larger than 10Å, near the beam stop shadow, along the rotational axis and within 15 pixels of the segment edges were thus masked. *pyFAI's separate* program (Ashiotis *et al.*, 2015) was used to remove Bragg peaks from each diffraction image by azimuthal median filtering and then backfilling the pixels associated with the Bragg peaks. All images in the data set were then averaged to produce a mean background image. This background was linearly scaled (multiplied by a constant) to match each frame. Pixels with intensities larger than five times the scaled mean background were identified as zingers and their resolutions recorded. A histogram was constructed of the number of zingers binned against resolution with a bin spacing of 0.01 Å, and the number of observed zingers was normalized to total zingers per crystal rotation angle to facilitate comparison of histograms between datasets. Table 1 lists

the resolutions of the 11 diffraction rings of hexagonal ice, where zingers due to ice should be located, based on 100 K unit cell parameters of a=4.497Å, b=7.322Å (Fortes, 2018).

5.2.5 Ice diffraction modeling

Ice within and on the surface of protein crystals may be hexagonal I_h, cubic I_c, a mix of these crystal forms (I_h + I_c), or stacking disordered (I_{sd}), in which cubic and hexagonal planes are randomly stacked, as shown in Fig. 6.4.2 (adapted from Moreau *et al*. 2019 Fig. 7 and Malkin et al. 2015 Fig. 9) Diffraction images collected at CHESS and those obtained from the IRRMC archive were fit by models of stacking disordered ice or of a mixture of cubic and hexagonal ice using the methods described in (Moreau, *et al.* 2019). Diffraction images were loaded into python using the package FabIO (Knudsen *et al.*, 2013) and the pyFAI integration package (Ashiotis *et al.*, 2015) was used to remove the protein Bragg peaks and azimuthally average the frames within resolution bins.

Powder diffraction patterns of stacking disordered ice were generated using the program DIFFaX (Treacy *et al.*, 1991). The model consists of (0001) planes of hexagonal ice randomly stacked with (111) planes of cubic ice. The probability of a cubic plane being followed by hexagonal plane is Φ_{ch} and of a hexagonal plane being followed by a cubic plane is Φ_{hc} (Kuhs *et al.*, 2012; Malkin *et al.*, 2015). These stacking probabilities, the unit cell parameters and instrumental broadening were optimized to fit the DIFFaX models to the observed azimuthally averaged diffraction using the scipy.optimize.minimize program in the SciPy python library (Virtanen *et al.*, 2020). The background was determined by evaluating a 10th order polynomial fit to the difference $I_{bg}(q) = I_{\exp}(q) - I_{DIFFaX}(q)$ between the DIFFaX models described below and our experimental diffraction patterns.

Azimuthally averaged diffraction versus resolution bin data were also fit assuming pure hexagonal ice, pure cubic ice, and a mix of hexagonal and cubic ice crystallites, using the same methods implemented to fit the stacking disordered ice diffraction. For pure hexagonal ice (cubic ice) , Φ_{hh} and Φ_{cc} were fixed to

1 (0) and 0 (1) respectively, and the unit cell and broadening parameters were allowed to vary during optimization. For the mixture of hexagonal and cubic ice, a linear combination of pure hexagonal and cubic ice diffraction patterns that were generated using equivalent unit cell and broadening parameters were optimized to fit the observed diffraction. No parameters were held fixed between cases.

5.2.6 Estimation of ice crystallite sizes

The size of the ice crystallites formed within protein crystals was estimated from the data of (Moreau *et al.*, 2019) using the observed ice diffraction ring breadths after azimuthal averaging. Prior to freezing, all external solvent was removed from the protein crystal in this data set. Any observed ice then formed from solvent that originated from within the protein crystal's interior, not its surface.

If a crystal has a finite size, the peaks in the reciprocal space representation of its electron density are radially broadened by an amount inversely proportional to the crystal size. Similarly, the breadth, β, of the crystal's diffraction peaks in 2θ, measured using a large, monochromatic non-divergent beam, increases with decreasing crystal size and is given by Scherrer's equation

$$\beta = \frac{\lambda}{\delta \cos(\theta)}. \tag{5.1}$$

Here, the breadth β of a peak is estimated as the integrated area beneath the peak divided by its maximum amplitude (Stokes & Wilson, 1942), λ is the wavelength, θ is the Bragg angle, and δ is the apparent size of the crystallite. The apparent crystallite size in Scherrer's equation is related to the actual crystallite size, p, (the cube root of the crystal volume) by $p = K\delta$. The proportionality constant K is known as Scherrer's constant (Langford & Wilson, 1978), and depends on crystallite shape, crystal symmetry, Miller index and the definition of peak breadth. Scherrer's constant for a cubic crystal's (100), (110) and (111) reflections are 1.0000, 1.0607 and 1.1547 respectively (Langford & Wilson, 1978) and is assumed to be 1 for this analysis.

5.2.6.1 Estimate of instrumental diffraction peak broadening

To estimate the intrinsic peak broadening β due to crystallite size and shape, the broadening due to instrumental effects was estimated from measurements of the powder diffraction pattern from hexagonal ice in a sample where peak broadening due to crystallite size was assumed to be insignificant. The hexagonal ice sample was generated *in situ* at the F1 CHESS station as follows. First, a 30% w/w polypropylene glycol 425 solution in a 500 µm diameter PET tube was abruptly cooled to T=200 K by unshuttering a cold gas stream. Abrupt cooling to a temperature where the ice nucleation rate was high resulted in an ideal powder pattern with a large width (indicating a small crystallite size) and having a stacking disordered intensity profile. As this sample was slowly warmed, the intensity profile evolved from stacking disordered to purely hexagonal (Kuhs *et al.*, 2012) and the peak breadths decreased (Fig. 6.4.3). Above 250 K, the uniformity of the diffraction rings was lost, and each ring eventually broke up into a collection of individual Bragg peaks. The diffraction pattern recorded at 250 K (Fig. 6.4.4(a)), before the loss of ring uniformity, was used to estimate the instrumental broadening.

The use of this ice sample as a calibrant does not correct broadening due to uncertainty in the experimental geometry, specifically in the assumed beam center and angle between the incident X-ray beam direction and detector normal; these modify the apparent angle of each ice peak and thus the apparent angular width. These sources of broadening were estimated by optimizing these parameters to minimize the sum of the breadths of the all the observed ice rings. For the hexagonal calibrant, this yielded a detector tilt of 0.1° from the incident X-ray beam normal. The calibrant's corrected breadths are shown in Fig. 6.4.4(b) and have an average of 0.037°. This breadth was then taken as an estimate of the instrumental broadening due to the incident X-ray beam divergence and energy dispersion.

The same method for correcting detector tilt and beam center offset were applied to all the other ice diffraction patterns to obtain corrected experimental peak widths, which then were assumed to only have

contributions from crystal size and shape and from instrumental broadening from beam divergence and energy dispersion.

The broadening due to beam divergence and energy dispersion are estimated for comparison. Beam divergence α adds linearly to the peak widths, $\Delta 2\theta \approx \alpha$. The broadening due to beam dispersion is estimated from Bragg's Law through uncertainty propagation,

$$\begin{aligned} \Delta 2\theta &\approx 2\frac{1}{\sqrt{1-\left(\lambda/2d\right)^2}}\frac{\lambda}{2d}\frac{\Delta\lambda}{\lambda} \\ &\approx 2\cdot\tan\left(\theta\right)\frac{\Delta\lambda}{\lambda} \end{aligned} . \qquad (5.2)$$

The beam divergence is $\alpha = 0.028^o$ and the dispersion listed for the F1 station is $\Delta\lambda/\lambda = 0.00173$. These correspond to approximate peak breadths of 0.028° and 0.027° for divergence and dispersion. The observed hexagonal ice peaks of the calibrant sample appeared Gaussian, indicating these two factors should add in quadrature. This gives a combined broadening of 0.04°, on par with the measured peak breadth of 0.037°.

5.2.6.2 Estimate of intrinsic ice diffraction peak broadening and ice crystallite size

The broadening of the observed (detector angle and beam-stop position corrected) ice diffraction peaks is a convolution of the intrinsic and instrumental broadenings. The increase in breadth of a peak from a convolution depends on the shape of the functions involved. For convolutions of Gaussians, the resulting profile is Gaussian, and the breadths add in quadrature; for Lorentzians, the breadths add linearly. The convolution of a Gaussian with a breadth of β_{gauss} and a Lorentzian width a breadth of β_{lor} is a Voight profile with a breadth of (Olivero & Longbothum, 1977)

$$\beta_{voigt} \approx 0.5346\beta_{lor} + \sqrt{0.2166\beta_{lor}^2 + \beta_{gauss}^2} . \qquad (5.3)$$

Peak profiles of the hexagonal ice calibrant were best modelled by Gaussians and are likely dominated by instrumental factors. Peak profiles for ice internal to the protein crystals were best modelled by Lorentzians. This implies that the intrinsic broadening from crystallite size and shape is not Gaussian. While we cannot be certain of its shape, we assumed it is Lorentzian because it is the dominant broadening factor. In this case, the observed breadths are β_{voigt}, the instrumental broadening estimated from the hexagonal calibrant peaks is β_{gauss}, and the desired breadth from the internal ice deconvolved from the instrumental effects is β_{lor} and is given by

$$\beta_{lor} \approx \frac{\beta_{voigt}^2 - \beta_{gaus}^2}{\beta_{voigt}}. \tag{5.4}$$

The Scherrer Equation, Eq. 5.1, suggests that a plot of the $\beta_{lor} \cdot \cos(\theta)$ vs diffraction angle 2θ should be constant, if crystallite size is the only source of intrinsic peak broadening. In all cases, $\beta_{lor} \cdot \cos(\theta)$ increased with angle θ suggesting that crystallite size was not the only source of broadening. Strain and dislocations (Baker, 2002; Thürmer & Bartelt, 2008) are potential sources of broadening, though limited prior knowledge of the underlying broadening mechanisms prevents the determination of a reliable model to deconvolve these factors (Ungár *et al.*, 1998). In a simple case of isotropic strain, ε, the Williamson-Hall model gives a crystallite size broadening satisfying (Scardi *et al.*, 2004)

$$\beta_{lor} \cos(\theta) = \frac{\lambda}{\delta} + 2\varepsilon \sin(\theta). \tag{5}$$

The size δ can be determined using a plot of $\beta_{lor} \cos(\theta)$ vs $\sin(\theta)$.

Modified Williamson-Hall models can account for anisotropic strain, dislocations and planar faults. The non-size broadening factors increase with angle in these models, so the broadening of the lowest angle Bragg peak will have the largest relative contribution from crystallite size. For this reason, only the

lowest angle Bragg reflection, the (002) reflection of hexagonal ice occurring at $2\theta = 15.3^{o}$, which is not broadening by stacking disorder, was used to determine the crystallite size.

5.2.7 Analysis of deposited PDB structure factors for ice

When the effects of ice are incorrectly/incompletely accounted for during background subtraction, there is a systematic biasing of the measured integrated protein Bragg peak intensity / structure factor values that can often be readily observed. To automatically detect this biasing, we followed the procedure used by Thorn *et al.* (2017) in their AUSPEX ice detection method, extended it and then did a benchmark comparison. Our algorithm is based on three separate metrics that compare the distribution of measured protein structure factors at the expected 2θ values / resolutions of ice diffraction with those observed nearby, but off the ice peaks. These three metrics monitor changes in the mean structure factor intensity, the fraction of structure factors with low intensities, and the number of measured structure factors.

5.2.7.1 Ice Finder Score

Fig. 5.1(a) shows a diffraction frame for PDB entry 4H3W taken from the IRRMC. Fig. 5.1(b) shows a scatter plot of the corresponding I_{obs} values deposited in the PDB. Thorn *et al.* used visual inspection of structure factor vs resolution patterns as in Fig. 5.1(b) to assess if ice contamination was present in PDB-deposited structure factors corresponding to 156 IRRMC data sets, and validated these conclusions by examining the raw diffraction frames from the IRRMC as in Fig. 5.1(a). They then took 200 random structure factor data sets from the PDB, used the same visual scoring (based on plots as in Fig. 5.1(b)) as for the IRRMC data to determine if they displayed ice contamination, and developed an algorithm for detecting ice contamination in deposited structure factors based on this data set. Thorn *et al*. then benchmarked their algorithm with a second set of 200 random PDB entries.

Thorn *et al.*'s ice detection algorithm, AUSPEX, looks for changes in the measured background-subtracted protein crystal Bragg reflection intensities within bins of fixed width in inverse resolution. They calculate an Ice Finder Score, IFS, for each inverse resolution bin as

$$IFS = \sqrt{N}\left(\left\langle I_{obs}\right\rangle / \sigma - f\right). \quad (6)$$

Here N, $\left\langle I_{\mathrm{obs}}\right\rangle$ and σ are the number, mean and standard deviation of the measured Bragg intensities in a given inverse resolution bin. f is the expected normalized mean of the intensities (i.e., the mean divided by the standard deviation), estimated by the general trend of intensity values in the bins away from the ice rings. To establish f, the mean and standard deviation of the intensities in coarse inverse resolution bins of size 0.01 Å^{-1} were calculated and these were then linearly interpolated through regions containing ice rings and to reduce the bin size to 0.002 Å^{-1}, shown as a green line in Fig. 5.1(b). These interpolated values were used to calculate a normalized mean for each 0.002 Å^{-1} bin, which was then smoothed with a Gaussian filter with a standard deviation of 0.01 Å^{-1} to give a final estimate of f. In our calculations we used a final bin size of 0.002 Å^{-1}, as this reduced fluctuations, and was still smaller than the ~0.005 Å^{-1} width of the regions biased by ice. Fig. 5.1(c) shows the Ice Finder Score calculated for PDB entry 4H3W.

5.2.7.2 Depletion Score

We define a second metric, the Depletion Score, which detects ice by looking for depletion of low intensity Bragg peak values at the ice ring resolutions. It tracks the fraction of Bragg peaks in an inverse resolution bin with intensities smaller than a fraction, chosen to be 20%, of the mean intensity. Based on Wilson statistics (Wilson, 1949), for acentric reflections the probability of observing a reflection with intensity I_{obs} in a resolution bin with an average intensity $\left\langle I_{\mathrm{obs}}\right\rangle$ is given by:

$$P(I_{\mathrm{obs}}) = \frac{1}{\left\langle I_{\mathrm{obs}}\right\rangle}\exp\left(-I_{\mathrm{obs}} \Big/ \left\langle I_{\mathrm{obs}}\right\rangle\right) \quad (7)$$

This distribution assumes that the atoms in the unit cell are randomly distributed, which is a fair approximation at resolutions better than 4 Å (where all ice diffraction peaks lie), as the contribution of solvent cavities there is negligible. Based on this distribution, the expected fraction of Bragg reflections with intensities smaller than $m\langle I_{\text{obs}}\rangle$ is given by $1-\exp(-m)$, independent of resolution and average intensity $\langle I_{\text{obs}}\rangle$.

The depletion score was calculated using the same binning scheme as the IFS. The expected fraction of Bragg reflections with intensities smaller than 20% of the mean, $\bar{D}$, was estimated by calculating the fraction *D* of Bragg reflections with intensities smaller than 20% of the mean in coarse inverse resolution bins of 0.01 Å^{-1}, and linearly interpolating this through the regions containing ice rings and to increase the sampling to 0.002 Å^{-1}. *D* was then recalculated in finer inverse resolution bins of size 0.002 Å^{-1} and subtracted from the expected fraction $\bar{D}$ to estimate the Depletion Score

$$DS = \frac{\bar{D} - D}{\text{STD}\left(\bar{D} - D\right)_{\text{no ice}}}. \qquad (8)$$

Here, the normalization is by the standard deviation of the difference calculated in regions away from the ice rings. Fig. 5.1(c) shows the Depletion Score calculated for PDB entry 4H3W.

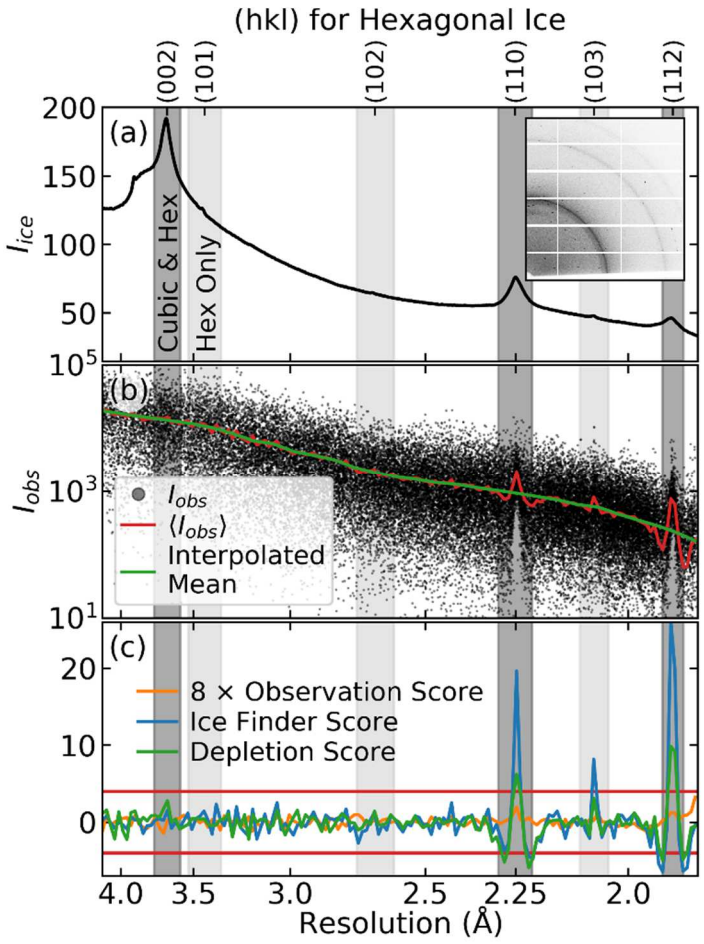


Figure 5.1 Identification of ice in the deposited structure factors of PDB entry 4H3W. (a) The 2D diffraction pattern and azimuthally averaged 1D background suggest that the ice is primarily stacking disordered, I_{sd}. The resolution ranges examined for structure factor biasing using our Ice Contamination Score, corresponding to shared peak positions for hexagonal ice I_h, cubic ice I_c, and stacking disordered ice I_{sd} are indicated by darker vertical shading. The resolution ranges examined to establish if hexagonal ice is present, corresponding to I_h peak positions absent from I_c and that are strongly suppressed in I_{sd} for typical cubic/hexagonal stacking fractions, are indicated by lighter vertical shading. (b) Measured diffraction peak intensities integrated by XDS from PDB entry 4H3W at T = 100 K. The red line shows the average I_{obs} value in bins of width 0.002 $Å^{-1}$. The green line shows the mean intensities calculated from coarser bins and then interpolated between bins and through the dark vertically shaded regions. (c) Ice Finder, Depletion, and Observation Scores calculated for this data set. The red line is drawn at ±4.

5.2.7.3 Observation Score

We define a third metric, the Observation Score, which looks for a reduction in the number observations within a resolution bin. This could occur if the experimenter excludes resolution regions overlapping the ice rings or if the integration and scaling algorithms throw out peaks as outliers. This metric is formed by counting the number of Bragg reflections *N* in inverse resolution bins of size 0.002 Å^{-1} and comparing to an expected number of Bragg reflections, $\bar{N}$,

$$N_{diff} = \frac{\bar{N} - N}{\bar{N}}. \tag{9}$$

The expected number of Bragg reflections is generated from N by removing the regions overlapping ice rings, linearly interpolating through these regions and smoothing with a Gaussian filter with a standard deviation of 0.01 Å^{-1}. Fig. 5.1(c) shows the Observation Score calculated for PDB entry 4H3W.

5.2.7.4 A combined metric: Ice Contamination Score

These three metrics were combined to generate a final score, which we call an Ice Contamination Score (*ICS*) that is calculated at each potential ice ring resolution. For a given ice ring, the maximum Ice Finder Score, Depletion Score, and Observation Score from the 0.002 Å^{-1} bins in the ice ring region are combined to obtain

$$ICS = \frac{1}{8}\left(IFS^{\max} + D_{diff}^{\max} \right) + \begin{cases} 1; N_{diff} \geq 0.5 \\ 0; N_{diff} < 0.5 \end{cases} \tag{10}$$

A final score above 1 for any ice ring location is used as an indicator of ice. Thorn *et al.* suggest the Ice Finder Score loosely resembles a z-score. The Depletion Score is defined as a z-score. If these metrics are both interpreted as z-scores, when their average values have surpassed 4 (corresponding to 4 standard deviations from the mean), then the data set is flagged for the presence of ice. The piecewise addition of one if less than half of the expected number of observations is observed, $N_{diff} \geq 0.5$, automatically flags

the presence of ice; the choice of $N_{diff} = 0.5$ and the use of a binary Observation Score are arbitrary but adequate. The Ice Finder and Depletion scores are correlated but are different methods for detecting larger-than-expected Bragg peak values in a resolution bin. When combined, they increase the Ice Contamination Score's resilience to false positives / negatives.

5.2.7.5 Analyzing PDB entries for the presence and type of ice

These algorithms and metrics were applied to detect the presence of ice and its correlation with protein crystal properties such as solvent content and solvent cavity size. To explore the correlation of ice with solvent cavity size, the same set of 16,940 PDB entries analysed in (Moreau *et al.*, 2019) were used. To explore correlations with unit cell size, solvent content, and year of data collection (which were computationally much less complex to determine), a larger set of 83,938 PDB entries was used. The lowest resolution ice ring is at 3.9 Å. Ice detection metrics become noisy and give false positives when there are too few Bragg reflections. No ice is expected at higher temperatures. Consequently, PDB entries with resolution worse than 3.9 Å, less than 15,000 Bragg reflections, or reported data collection temperatures above 240 K were excluded from the analysis. The "year" metric was taken first from the reported year of data collection; if this was unavailable, the year of deposition was used instead.

Pure hexagonal ice has 11 diffraction rings between 4 Å and 1.5 Å (Table 5.1) and cubic ice has three (Table 5.1, dark shading), located at the positions of the (002), (110), and (112) hexagonal ice rings. Stacking disordered ice in general can have peaks at all 11 hexagonal locations (for large hexagonal stacking fractions). However, only the peaks at the positions of cubic ice are not broadened by the stacking disorder, and these three peaks typically had the largest pixel counts. For initial scoring and detection of ice, we chose to focus only on these three ice ring regions common to all forms of ice, in the resolution ranges displayed by dark shading in Fig. 5.1 and explicitly listed in Supplementary Table 6.4.1.

(hkl)	Resolution (Å)
(100)	3.895
(002)	3.661
(101)	3.438
(102)	2.667
(110)	2.249
(103)	2.068
(200)	1.947
(112)	1.916
(201)	1.882
(202)	1.719
(203)	1.522

Table 5.1 Miller indices of hexagonal ice rings at resolutions numerically larger than 1.5 Å. The darker shaded rows are not affected by stacking disorder rand are the only ice rings present for cubic ice.

This choice to focus on these three ice ring locations common to all forms of ice, as opposed to all 11 locations of hexagonal ice, improved the robustness of interpolating the mean, standard deviation, number of observations and the *D* statistic through the ice rings and from the 0.01 to 0.002 Å^{-1} bins. This choice also made the algorithms less prone to false positives and more sensitive to weaker signals of ice biasing. After binning the I_{obs} statistics in 0.01 Å^{-1} bins, the bins near the ice peaks where ice is being searched for are removed from further consideration. The remaining bins are used to interpolate through these excluded regions. As the total number of regions where ice is being searched for increases, there are more regions that need to be interpolated through and fewer neighbouring bins to guide the interpolation. When all 11 ice regions were used, this often led to erroneous interpolations. Focusing on only the three ice ring peaks common to all ice forms improved interpolations. In the case of cubic-like or stacking disordered ice, only these three peaks tend to be strong enough to bias the I_{obs} values, so there is no down-side to this choice. Hexagonal ice produces much narrower and taller peaks at all locations that strongly bias the I_{obs} values, so ice is easily detectable even when only three locations are examined.

To distinguish cubic / stacking disordered ice from hexagonal ice, PDB entries that were flagged for ice were analysed a second time. Scoring was now performed at three hexagonal ice peak locations not common to cubic ice and that are largely suppressed in typical stacking disordered ice – the (101), (102), and (103) peaks. Three regions were chosen to match the number of regions searched for cubic ice, and to obtain comparably reliable background interpolations. Of the five other hexagonal ice peak locations at resolutions worse than 1.5 Å, the (100) peak is broadened by stacking disorder but is still distinguishable unless the cubic stacking fraction is very high, and so is suitable. Biasing effects of the (200) and (201) peaks would be difficult to separate from that of the unbroadened (112) peak due to their close proximity and relatively weak intensity. The (202) and (203) hexagonal peaks are at 1.72 and 1.52 Å, higher resolutions than the high-resolution cut-off for many PDB entries. If ice was detected at any of one of the

(101), (102), or (103) peak locations, the ice was classified as hexagonal; if no ice was detected at all of these additional positions it was classified as stacking disordered. Note that these classifications are not precise, as stacking disordered ice with a large (>75%) hexagonal fraction will show peaks at all hexagonal locations (Malkin 2015).

5.2.8 Analyzing the effect of ice on atomic models of proteins

To understand the impact of biasing of structure factors by ice on the typical crystallography pipeline, an ice-free lysozyme data set was acquired, and additional data sets were generated by adding different amounts of ice diffraction to the detector frames. These data sets were then analyzed, and refined structures obtained using the ice-perturbed diffraction data and using the original data were compared.

A lysozyme crystal was soaked in 2M NaCl, mounted on a loop, and then enclosed in a MicroRT tube (MiTeGen, Ithaca, NY) containing a plug of the soak solution at one end. A room temperature diffraction data set comprised of 800 frames, each with a 0.25° rotation per frame, was collected at the CHESS F1 station.

Ice diffraction was added to the diffraction images as follows. 1D ice diffraction profiles were generated using DIFFaX for 50% / 50% cubic/hexagonal stacking disordered ice with a FWHM of $\sqrt{2}\tan(\theta/2)$, which increased from 0.2° to 0.4° for the relevant ice rings. These 1D profiles were converted to 2D images using *pyFAI.calcfrom1d* and the refined experimental geometry from XDS. The original lysozyme 2D diffraction images were first converted to *.img format using adxv; this was necessary because XDS would not accept the ice-modified files in *.cbf format. These images were loaded into numpy arrays using Fabio. The 2D ice diffraction images were directly added to these images, with a scale factor chosen so that the mean counts/pixel of the ice diffraction image was between 0 and 3.5 times the mean of the lysozyme crystal image, in intervals of 0.5. The mean of 0 was used as a control so that any errors (and resulting changes in refined structure) introduced by the file conversion could be

detected. The resulting images were saved in .img format and indexed, integrated and scaled using XDS to a resolution of 1.18 Å. Four additional data sets were generated by omitting reflections within 0.013, 0.025, 0.037 and 0.050 Å of the resolutions of the cubic-only ice rings (Table 5.1), to represent data sets where Bragg reflections have been removed to account for ice. Fig. 5.7 shows the resolution range of the removed Bragg reflections for the different data sets.

The scaled native and ice-perturbed data sets produced by XDS were refined to determine the effects of ice on final models. PDB model 4BS7 was used for molecular replacement in phenix.phaser (Liebschner *et al.*, 2019). The first round of refinement used rigid body refinement, individual atom's position and b-factors, and simulated annealing. During this round, 5% of the Bragg reflections of the control (ice-free) data set were randomly selected for R-Free estimates. This same set of reflections was used to estimate R-Free for the other data sets. A Na atom was added at a highly coordinated water site. The second round of refinement used individual atom's position and b-factors and ordered solvent was added. The third round of refinement used individual atom's position and anisotropic B-factors, TLS refinement and ordered solvent was updated. A fourth round repeated the third round with more conservative requirements. The minimum and maximum H-bond distances were set to 2.2 and 3.2 Å, the Fo-Fc map is the primary map used by*phenix.refine* for water placement, and the cutoff value for water placement was increased to 3.5. The 2Fo-Fc map is the secondary map and the cutoff value for water placement was increased to 1.2. The structures were checked in coot between each refinement cycle for Ramachandran outliers, rotamer outliers and large positive or negative peaks in the Fo-Fc maps.

The ice-contaminated models were compared with the ice-free control by estimating RMSD differences between the backbone atoms positions, differences between the isotropic B-factors, and total number of modelled water molecules. The maps were compared with that of the control by generating isomorphous difference maps. Coefficients for the difference maps were generated using *phenix.fobs_minus_fobs_maps*, where the merged reflections of the control and the ice-contaminated

data set were subtracted, and where the refined structure of the control provided the phasing. This difference map represents the difference between just the observations, i.e., the experimental structure factor amplitudes, and does not account for differences in model phases. The difference map coefficients were converted to volume scaled maps using the *phenix.mtz2map* and the RMS deviation was read from the map's header using the ccp4 program *mapdump*.

Rather than use maps generated just from observations, crystallographers generally use 2Fo-Fc maps. Difference maps were thus created between the control's 2Fo-Fc map and the 2Fo-Fc map of the ice-biased datasets. To do this, 2Fo-Fc maps were generated from the *.mtz file obtained from the final refinement of each ice-biased set using *phenix.mtz2map.* The control map was subtracted from the 2Fo-Fc map by first multiplying the control map by -1 using *mapmask* and then adding to it to the 2Fo-Fc map using *mapmask.* To quantify the spatial distribution of the difference density caused by ice, the RMS deviations in the regions of the protein's core, the protein's surface, and the solvent cavities were separately estimated. To do this, masks were generated of these regions (1 in region, 0 out of region) and the difference maps were multiplied by these masks using *mapmask*. The RMS deviations of the product maps were taken from *mapdump.* The RMS deviation grows with the square root of the map's volume, so the RMS deviation of the individual regions were normalized by dividing by the square root of the fraction of the unit cell they occupied. The mask of the protein core and surface was generated by first taking the refined control *.pdb file and separating it into two files, one with the ordered waters and surface side chains, and one with the rest of the atoms. Masks of these regions were then generated using *ncsmask*. The bulk solvent mask was generated by taking the refined control *.pdb model and setting the occupancy of all the atoms to zero. This file was then used as the input to *phenix.fmodel*, with the bulk solvent correction parameters set to *kmask* = 1 and *Bsol* = 0, and the bulk solvent mask was set to not ignore zero occupancy atoms. The resulting map coefficients were converted to a map using *phenix.mtz2map* and then converted to a mask using *mapmask.*

5.3 RESULTS

5.3.1 Types of ice diffraction from bulk cryoprotectant solutions and from protein crystals

Ice diffraction in protein crystallography can arise from ice in internal crystal solvent, from ice formed in residual cryoprotectant-containing solvent on the crystal surface, and from frost. Fig. 5.2 shows 2D diffraction images and 1D azimuthally averaged diffraction patterns of ice formed in glycerol solutions in 250 µm diameter, thin wall polyester tubing, cooled in a nitrogen cryostream to temperatures between 180 and 240 K. For solutions with 20% v/v glycerol and lower, diffraction obtained when cooling to all temperatures indicates that the ice is largely hexagonal, and the diffraction patterns are azimuthally lumpy, indicating a relatively large grain size and a small number of grains within the X-ray illuminated volume. For 30% glycerol solutions, ice diffraction at 240 K is again azimuthally lumpy and largely hexagonal. But ice diffraction obtained by cooling to 180 K is isotropic, indicating a small grain size, and consistent with stacking disordered ice with a substantial cubic fraction. For 40% glycerol solutions, no ice forms at 240 K. At 180 K, ice diffraction is isotropic and consistent with stacking disordered ice with a largely cubic character. The integrated intensity of the ice diffraction is much smaller than for 30% and 20% glycerol solutions, indicating that a significant fraction of the illuminated sample volume has vitrified.

Fig. 5.3 shows examples of diffraction from internal ice in crystals of apoferritin, thaumatin, and lysozyme, soaked in solutions containing between 0% and 20% v/v glycerol, and cooled to temperatures of 220 K and 100 K. As discussed in Moreau *et al.* (2019) in all cases, the observed ice diffraction is neither cubic nor hexagonal nor a simple mix of the two, but exhibits the characteristic selective and anisotropic peak broadening of stacking disordered ice. As glycerol concentration increases, the cubic stacking fraction increases. The cubic fraction is largest for cooling to 100 K, which gives the largest average cooling rate between the internal solvent's freezing and glass transition temperatures.

Fig. 5.4 shows examples of ice diffraction patterns – which may contain internal ice, surface ice, and frost – taken from data sets in the IRRMC. In the example shown in the top row (PDB: 4HF7), the ice diffraction is nearly cubic ice; in the example shown in the middle row (PDB: 4PUC), the ice is mostly stacking disordered with an additional hexagonal component; and in the example shown in the bottom row (PDB: 5UBA), the ice is almost entirely stacking disordered. For each example, the left panel shows the 2D diffraction patterns and the 1D azimuthally averaged diffraction profile, along with fits to a mixture of hexagonal and cubic ice and to a stacking disordered ice model. The upper right panel shows a scatter plot of the I_{obs} values vs. resolution, and the lower right panel shows the calculated ICS. The histogram at the top of the figure shows the cubic stacking fraction calculated from 22 IRRMC data sets along with the cubic stacking fraction of the three examples. The highly cubic diffraction from the 4HF7 is more typical of ice in the deposited data sets.

The ICS scores vary greatly between these three examples and are representative of the biasing that is observed for each type of ice. The nearly cubic ice of the top example is undetected in the ICS score. The broad diffraction peaks typical of cubic-like ice cause less variation in background intensity between the Bragg peaks and neighboring regions and so have a smaller effect on the integrated intensities. Hexagonal-like ice generates ice rings that are much narrower and, for the same integrated intensity, much taller, causing large biasing of integrated intensities. The stacking disordered ice of the bottom example gives detectable ice biasing at the higher resolution (110) and (112) ice ring positions but not at the (002) position. Ice diffraction typically has B-factors an order of magnitude smaller than those of the protein diffraction. The protein's diffraction strength decreases much faster with increasing resolution, and so biasing of integrated intensities at lower resolutions, where protein diffraction is relatively stronger, is less.

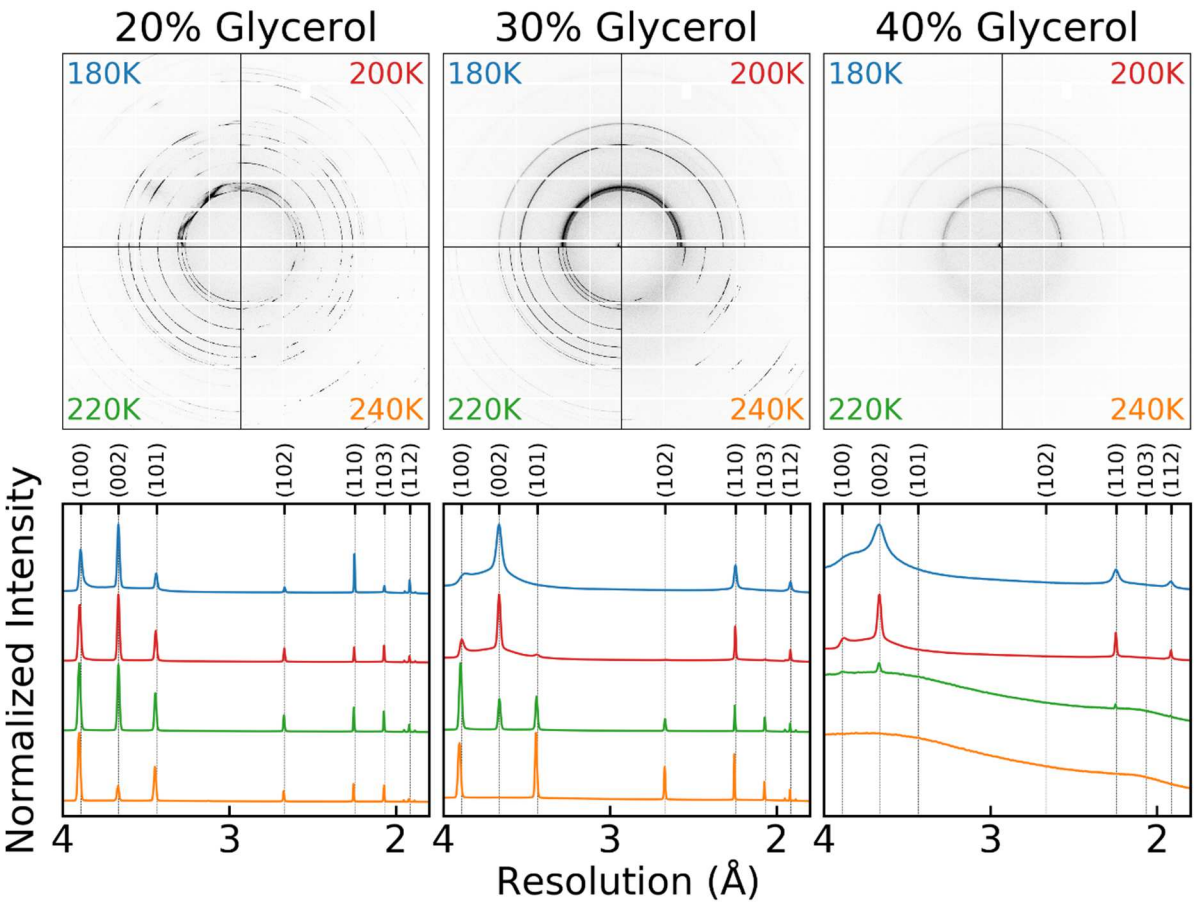


Figure 5.2 Ice diffraction from glycerol / water mixtures at different concentrations and temperatures. The top row shows 2D diffraction patterns, all plotted with the same pixel counts to greyscale calibration. The bottom row shows 1D diffraction patterns obtained from the 2D patterns by azimuthal averaging. The intensity scales of the 1D patterns are individually normalized. Similar trends are observed with other common cryoprotectants including 2-methyl-2,4-pentanediol (MPD), sucrose and polypropylene glycol.

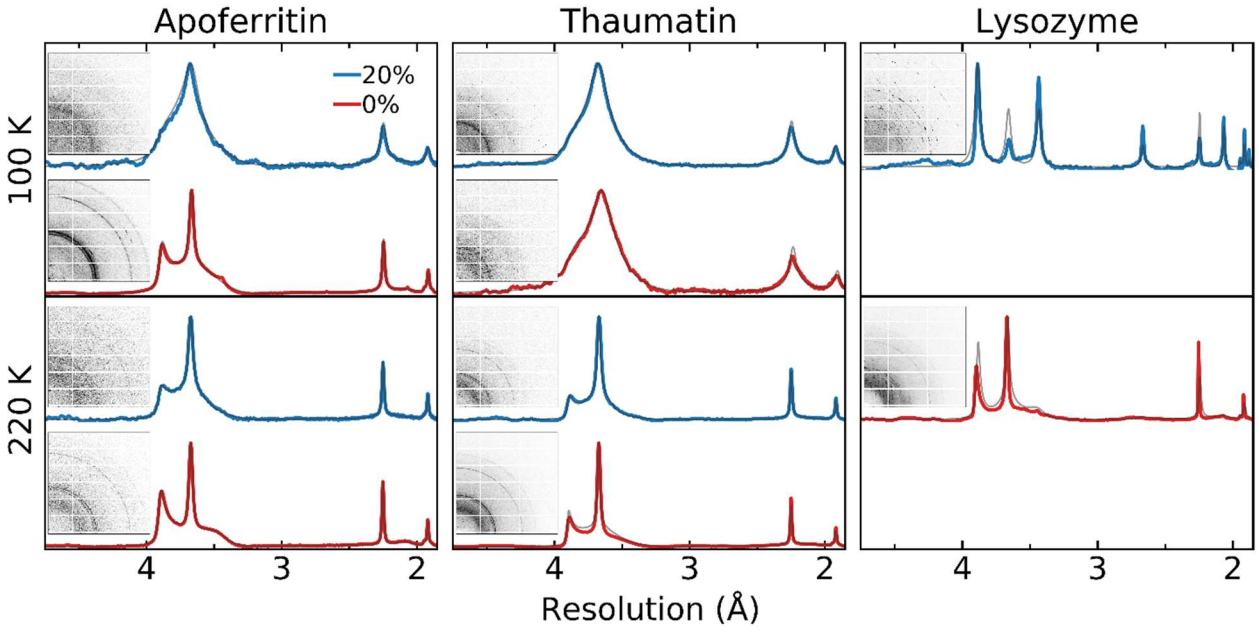


Figure 5.3 Example Bragg-peak subtracted, azimuthally averaged diffraction patterns from apoferritin, thaumatin, and lysozyme crystals for which ice formed in internal crystal solvent. Crystals were either used as grown (0%) or soaked in solutions containing 20% glycerol and cooled to 220 K (top) or 100 K (bottom).

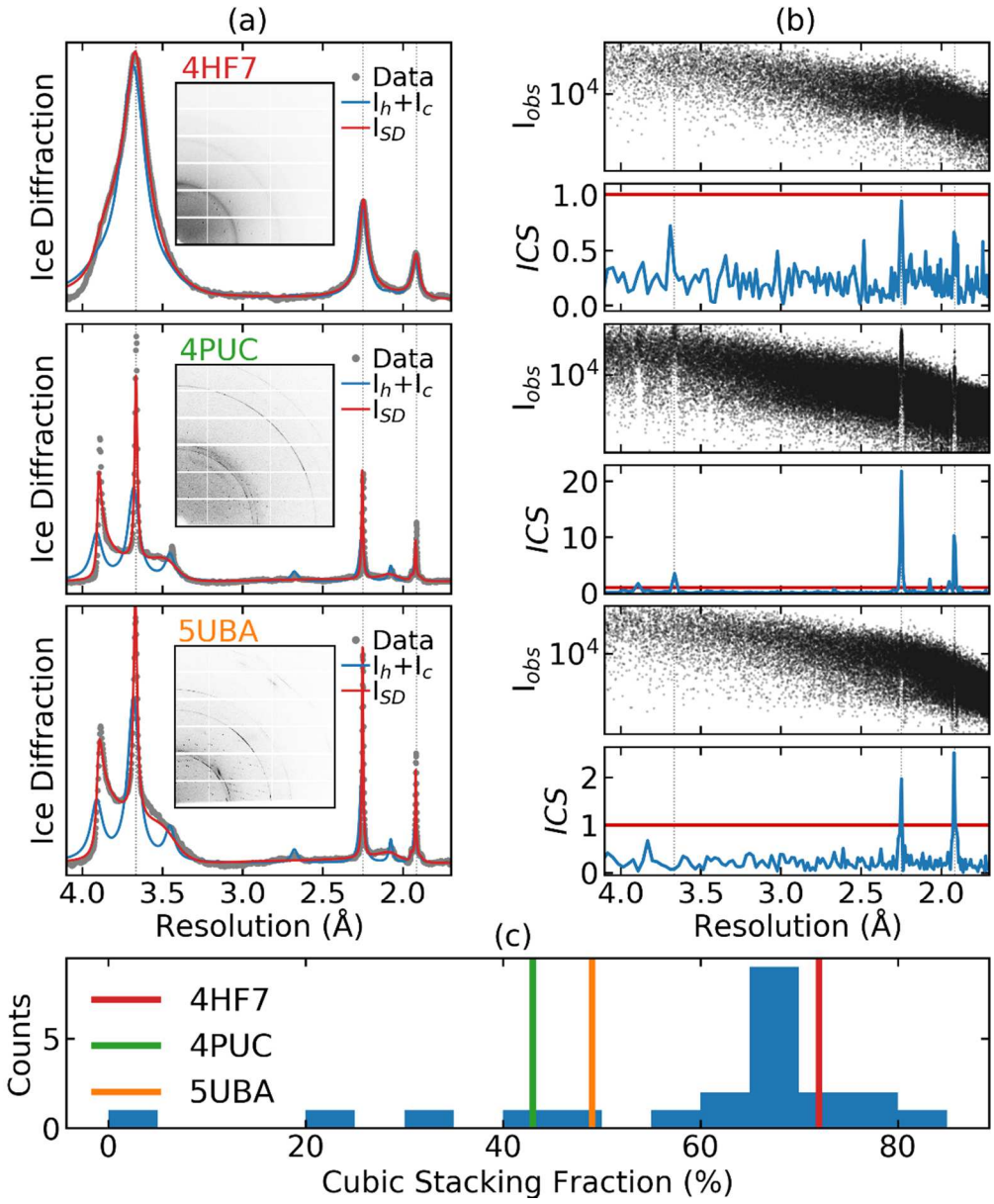


Figure 5.4 (a) Ice diffraction from three data sets taken from the IRRMC. Top example shows ice diffraction from nearly cubic ice (PDB: 4HF7), the middle example shows ice that is mostly stacking disordered with some additional hexagonal component (PDB: 4PUC) and the bottom example shows stacking disordered ice (PDB: 5UBA). (b) corresponding deposited structure factor distribution and calculated Ice Contamination Score (ICS) vs resolution. (c) Histogram of the cubic stacking fraction calculated from 22 IRRMC data sets along with the cubic stacking fraction of the three examples.

5.3.2 Estimates of ice crystallite sizes

Figure 5.5 shows estimated crystallite sizes for ice formed from solvent internal to apoferritin crystals, for a range of temperatures and glycerol concentrations. The left axis displays the approximate crystallite sizes determined from ice ring widths corrected for instrumental broadening. The tick locations on the left and right axes are the same, and the tick labels on the right axis are the crystallite sizes corresponding to the tick labels on the left axis, calculated using the full experimental peak widths, uncorrected for instrumental broadening. These latter values serve as a lower bound for the crystallite size. These crystallite size estimates have considerable uncertainty, because of the simplicity of the model used to determine them. But they are an order of magnitude larger than the solvent cavities within apoferritin (68 Å). This suggests that when ice forms within the crystal, the crystal lattice must be disrupted to make space for the ice crystals as they grow during cooling. This is consistent with large increases in mosaicity and dramatic loss of ordered diffraction from the protein lattice when significant internal ice diffraction develops.

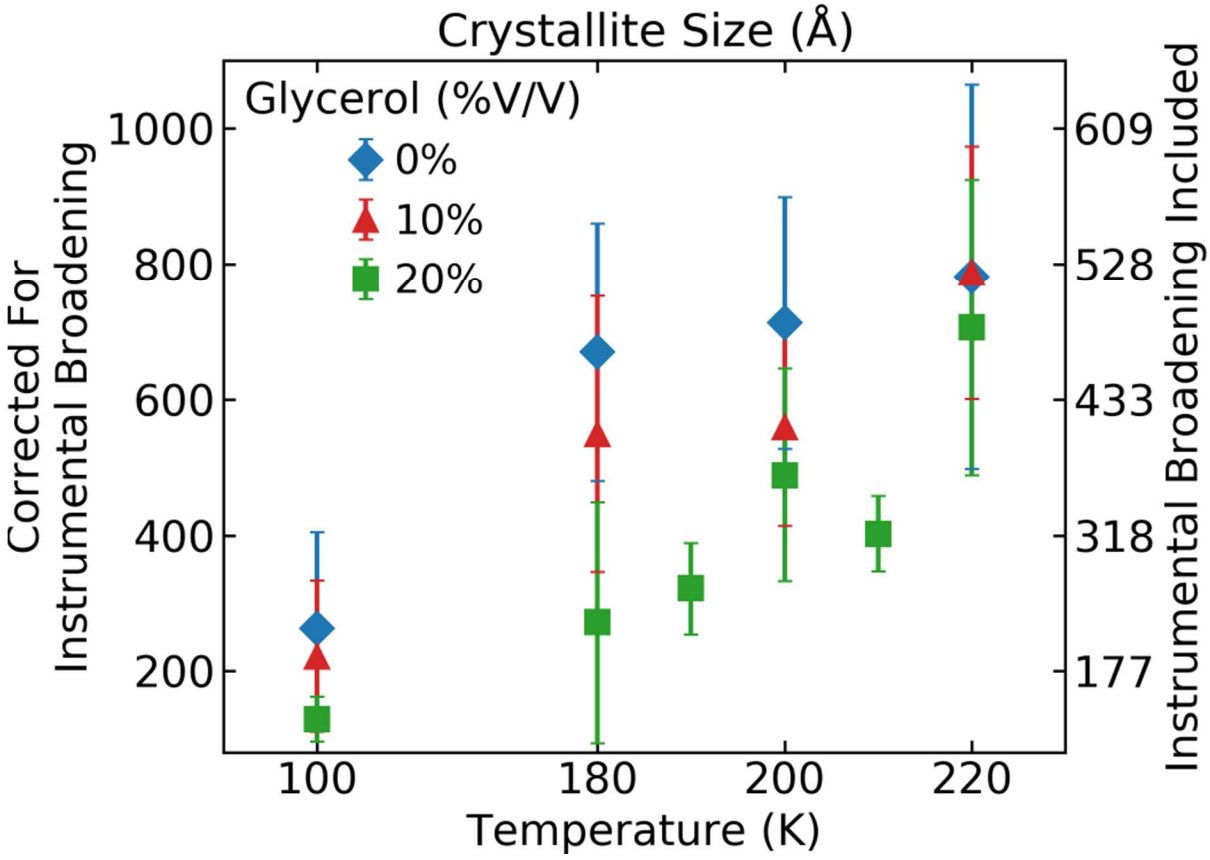


Figure 5.5 Ice crystallite sizes for internal ice in apoferritin crystals soaked in different concentrations of glycerol and cooled to temperatures between 100 K and 220 K. The left axis shows the crystallite size with correction for instrumental broadening. The right axis shows the crystallite sizes corresponding to the left axis tick locations if instrumental broadening is not accounted for; these values give a lower bound on the crystallite sizes.

5.3.3 Zinger analysis.

Histograms of zinger counts vs. resolution covering the resolution range of 10 – 1.4 Å were generated for 60 data sets from the IRRMC archive, including 26 datasets having visible ice rings and an additional 34 datasets without visible ice rings. Details and results of these datasets are in Table 6.4.2. Fig. 5.6 shows an example histogram for PDB entry 4EXR, a single diffraction frame from this entry showing zingers, and azimuthally integrated backgrounds for several frames spanning the full angular range of the data set. Even though the diffraction frames and integrated backgrounds show no evidence of ice, roughly 82% of the zingers are located within 0.01 Å of one of 11 hexagonal ice ring resolutions, which comprises ~10% of the detector area in the 10 – 1.4 Å range where zingers were searched for.

Of 34 data sets from the IRRMC archive that showed no visible ice rings in the azimuthally averaged backgrounds, 22 showed significant numbers of zingers at hexagonal ice ring locations. Zingers were observed at all hexagonal ice ring resolutions, not just those that are common to hexagonal and cubic ice, indicating the ice responsible was largely hexagonal. For these 22 data sets, the total number of zingers per oscillation degree at hexagonal ice resolutions ranged between 1 and 53 with an average of 9 ± 12, and the fraction of zingers observed at hexagonal ice locations ranged from 21% to 84%. For the remaining 12 IRRMC data sets, which were both ice-ring-free and ice zinger free, the average number of zingers per oscillation degree was 3 ± 4. These zingers could be attributed to large, but statistically plausible, background pixels and were more prevalent in datasets with weaker backgrounds.

Another 20 data sets taken from the IRRMC archive showed stacking disordered or cubic-like ice rings. Of these, 12 showed an elevated number of zingers at ice ring resolutions completely suppressed by the stacking disorder and having no visible diffraction peaks at those resolutions in the 2D or 1D diffraction patterns.

Thus, for all 35 data sets showing substantial numbers of zingers at ice ring resolutions, the observed ice-generated zingers are consistent only with hexagonal ice. Furthermore, the hexagonal ice zingers are typically single pixels with very high count rates (e.g., much larger than the count rates observed in ice rings when they are present), indicating that they are generated by relatively large ice crystals: to produce a (112) hexagonal ice peak at its resolution of 1.916 Å that is only a single pixel wide, the ice crystal needs to be a minimum of 4,000 Å in linear dimension (Calculation in Section 6.4.2). These crystals are at least order of magnitude larger than the crystallites producing the diffraction from cryoprotected solutions within the protein crystals. These large hexagonal ice crystals are likely frost crystallized from moist ambient air that accumulated in the liquid nitrogen used to cool the crystals, or that formed on the crystal during post-cooling handling or during data collection at the beamline.

A similar analysis was performed using the X-ray diffraction images we collected from apoferritin, thaumatin, and lysozyme crystals as described in Section 5.2.2, which were prepared with all external solvent removed and cooled *in situ* in the nitrogen gas cryostream. No evidence of hexagonal ice zingers was observed in any of the 433 datasets from apoferritin, lysozyme or thaumatin.

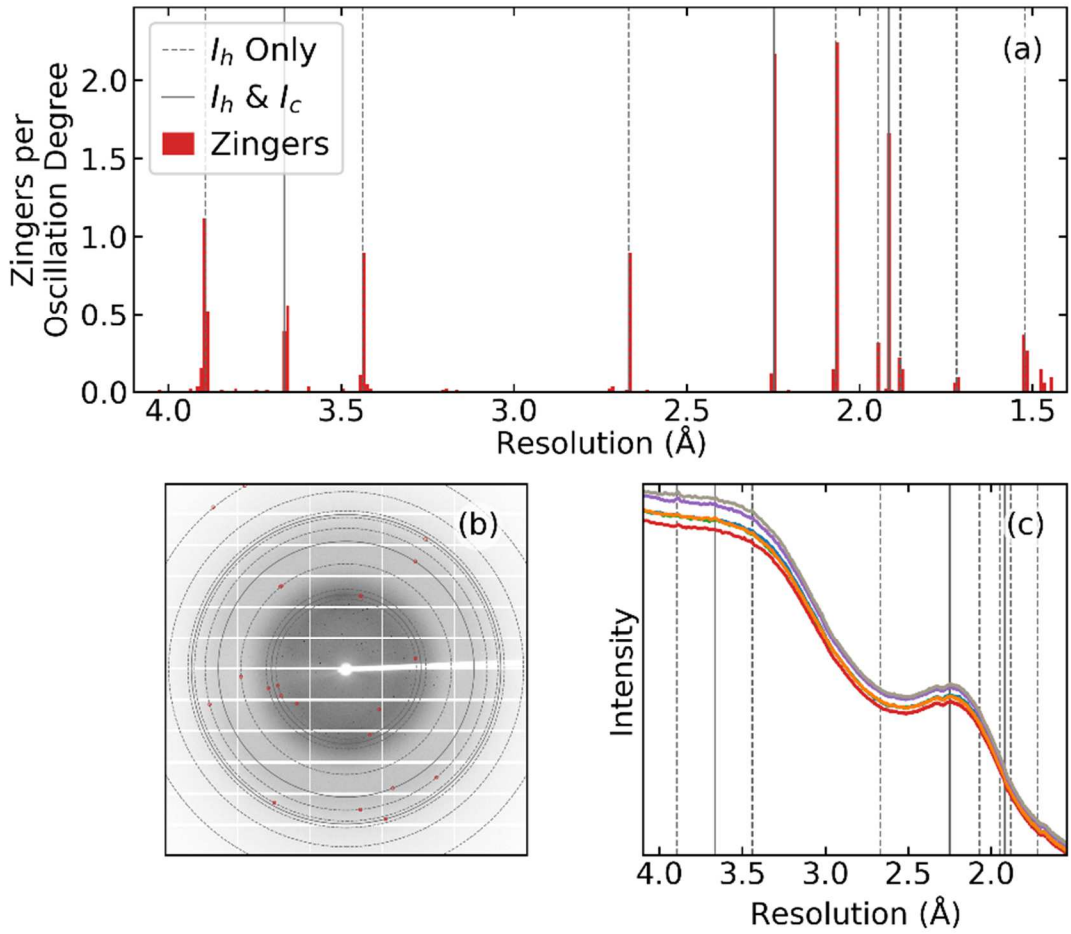


Figure 5.6 (a) Average number of zingers per oscillation degree vs. resolution for PDB entry 4EXR. Dashed vertical lines show resolutions where ice rings exclusive to hexagonal ice and to stacking disordered ice with a large hexagonal fraction occur, and solid lines show resolutions common to hexagonal, cubic, and stacking disordered ice. The 2D diffraction frames taken from the IRRMC show no evidence of ice rings. (b) A 2D diffraction image (0.5° oscillation) for this PDB entry taken from the IRRMC, with red circles drawn at the location of zingers and solid guidelines drawn at ice ring resolutions. (c) 1D azimuthally averaged diffraction intensity versus resolution, calculated using every 30th 2D frame in the data set, confirming the absence of ice rings.

5.3.4 Comparison of metrics for detecting ice in PDB-deposited structure factors

Table 5.2 compares the ice detection performance of our Ice Contamination Score with that of the AUSPEX Ice Finder Score, for the 200 randomly selected PDB data sets Thorn *et al.* used to benchmark their algorithm. Visual inspection of the I_{obs} values (Fig. 5.1b), looking for a depletion of low intensity Bragg peaks or a spike in high intensity peaks at ice ring resolutions, was used to identify datasets that containing ice. False positives are datasets that do not have ice by visual inspection but that are flagged as containing ice with the ice detection algorithm; false negatives are similarly evaluated. Compared with the AUSPEX algorithm, our algorithm reduces the rate of false positives from 10.4% to 2.5% and the rate of false negatives from 42% to 14%.

	ICS	AUSPEX
False positive rate	2.5%	10.4%
False negative rate	13.5%	40.5%

Table 5.2 Our ICS was benchmarked against the AUSPEX algorithm using the same 200 randomly selected PDB entries used by Thorn et al. (2016). Ice biasing of the structure factors was observed by visual inspection in 37 entries. The percentage of entries visually observed to be ice-free but that were flagged as having ice is listed as the false positive rate, and the percentage of entries visually observed to have ice contamination but were not flagged as having ice is listed as the false negative rate

5.3.5 Prevalence and types of ice in the PDB

Fig. 5.7 shows the distributions of ice and the relative prevalence of "hexagonal" to "stacking disordered" ice (defined and determined as described in Section 5.2.7) versus (a) solvent cavity size, (b) unit cell volume, (c) solvent content and (d) deposition year, for (a) 16,940 and (b-d) 83,938 randomly selected PDB entries, where data was collected at cryogenic temperature (typically ~100 K). The presence of ice was determined using our Ice Contamination Score, and the character of the ice determined as described in Section 5.2.7.5. The horizontal dashed lines represent an average over the 83,938 PDB entries used in (b)-(d). Roughly 16% of these entries display ice contamination. Of this subset, roughly 21% show ice that we labelled as hexagonal and so likely had frost as a major component. The remaining 74% showed "stacking disordered" ice that arose solely from residual cryoprotected solvent on the crystal surface or from internal crystal solvent. The prevalence of ice increases with increasing solvent channel size, cell volume, and solvent content. The fraction of annual deposits with ice has been relatively constant. But since 2010 the prevalence of hexagonal ice among samples that exhibit ice has almost doubled.

The 83,938 PDB entries analysed in generating Fig. 5.7 were a subset of a larger dataset with data collection temperatures, as indicated in their PDB header files, below 240 K. A second subset of 3,697 PDB entries was taken from the larger data set that had listed data collection temperatures above 240 K and were excluded from the analysis. However, 205 of these "high-temperature" data-sets showed the presence of ice, initially detected by our Ice Contamination Score and then verified by visual inspection of structure factors as in Fig. 5.1(b). Journal publications associated with 30 of these PDB entries were randomly selected and examined. Of these publications 15 indicated data collection was at cryogenic temperature, 8 did not state the data collection temperature, 3 stated that data was collected at room temperature and 4 entries did not have an associated publication. In many cases, the depositor likely used the crystallization temperature instead of the data collection temperature. Given that 5.5% of these “room-temperature” entries exhibited ice, and that 16% of PDB entries with listed data collection

temperatures below 240 K show ice, roughly 1/3 of the 3,683 PDB entries with listed experimental temperatures above 240 K were probably collected at 100 K.

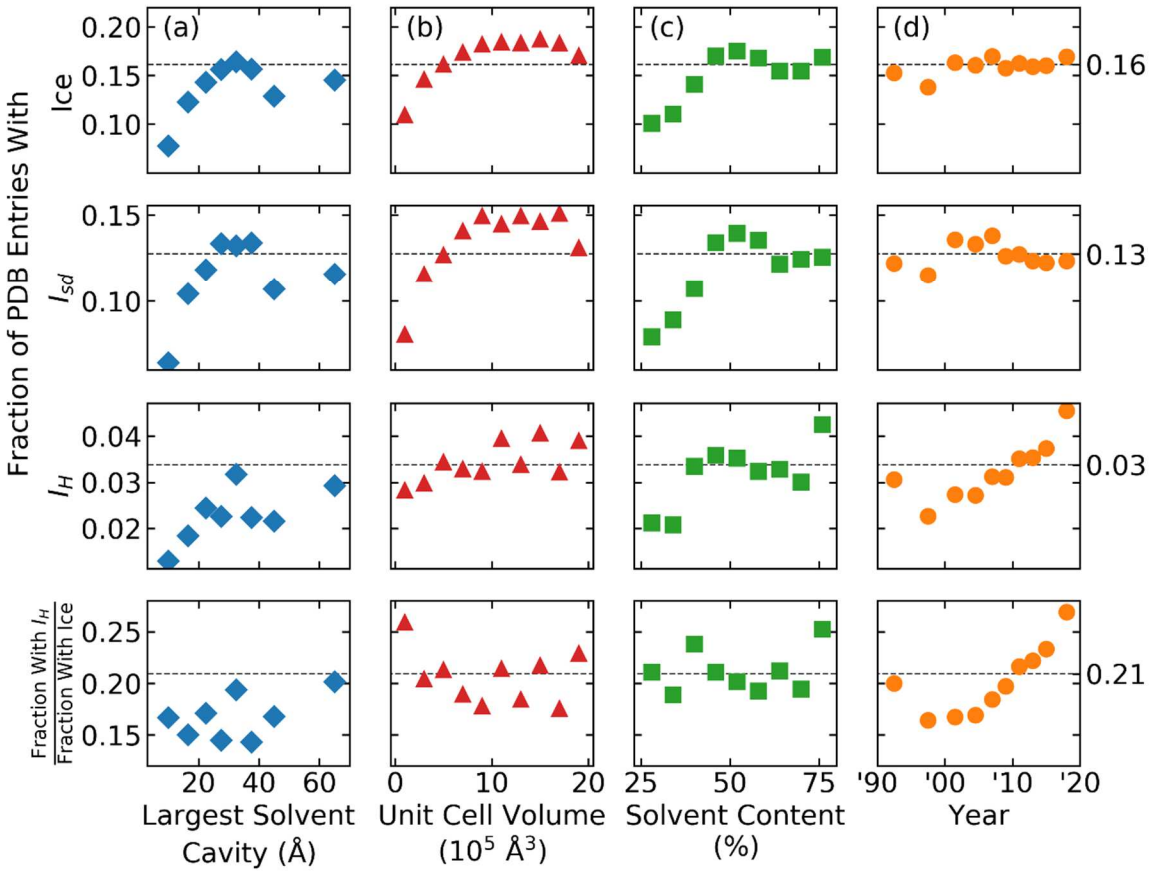


Figure 5.7 Fraction of PDB entries showing ice (first row), stacking disordered ice Isd only (second row), and hexagonal ice Ih (third row), and the fraction of all entries with ice contamination that show hexagonal ice (fourth row), as determined using the Ice Contamination Score and methods described in Section 5.2.7. Column (a) shows the variation with largest solvent cavity size, using the 16,940 PDB entries analysed in (Moreau et al., 2019). Columns (b), (c) and (d) show the variations with unit cell volume, solvent content and the year data was collected from a random set of 83,938 PDB entries. Dashed horizontal lines represent averages within an entire data set.

5.3.6 Effects of ice-related structure factor errors on atomic models of proteins

Fig. 5.8 shows how adding ice rings to diffraction images of a lysozyme crystal affects the resulting models and electron density maps, compared to those obtained using the original ice-free diffraction frames. Fig. 5.8a demonstrates the degree of biasing of the integrated intensities through the calculated ICS scores. Fig. 5.8b shows that with more ice and more biasing, R-free and R-work increase, as does the number of reflections excluded from the analysis. With more ice, the number of Bragg reflections flagged as “Aliens” by XDS and later excluded from analysis by phenix.refine increases. Fig. 5.8c shows that the RMS deviation between ice-free and ice-biased observation maps Fo increases almost linearly with increasing ice biasing. The RMSD between 2Fo-Fc maps (the maps most frequently used for visualization) also increases with ice biasing but is much smaller than for the observation maps. This indicates that model bias is effectively removing the ice bias from the maps.

Inspection and analysis of the difference maps generated from the biased and unbiased data ($\{F_o\}_{Biased} - \{F_o\}_{Control}$, and $\{2F_o - F_c\}_{Biased} - \{2F_o - F_c\}_{Control}$) shows that the differences are largely isotropic, and the RMS deviation of the difference maps is uniformly distributed throughout the unit cell. These map deviations are thus essentially "noise" that have limited impact on the protein models. There are no obvious changes to the overall structure and no reoriented side chains, and the RMSD for backbone atoms is on the order of a hundredth of an Angstrom, a rather insignificant amount. The only appreciable difference between the models was in the number of ordered solvent molecules. The number of modelled waters depends on refinement parameters, and more conservative water placement reduces, but does not eliminate, this difference. These results are consistent with current understanding of the effects of ice on structural models.

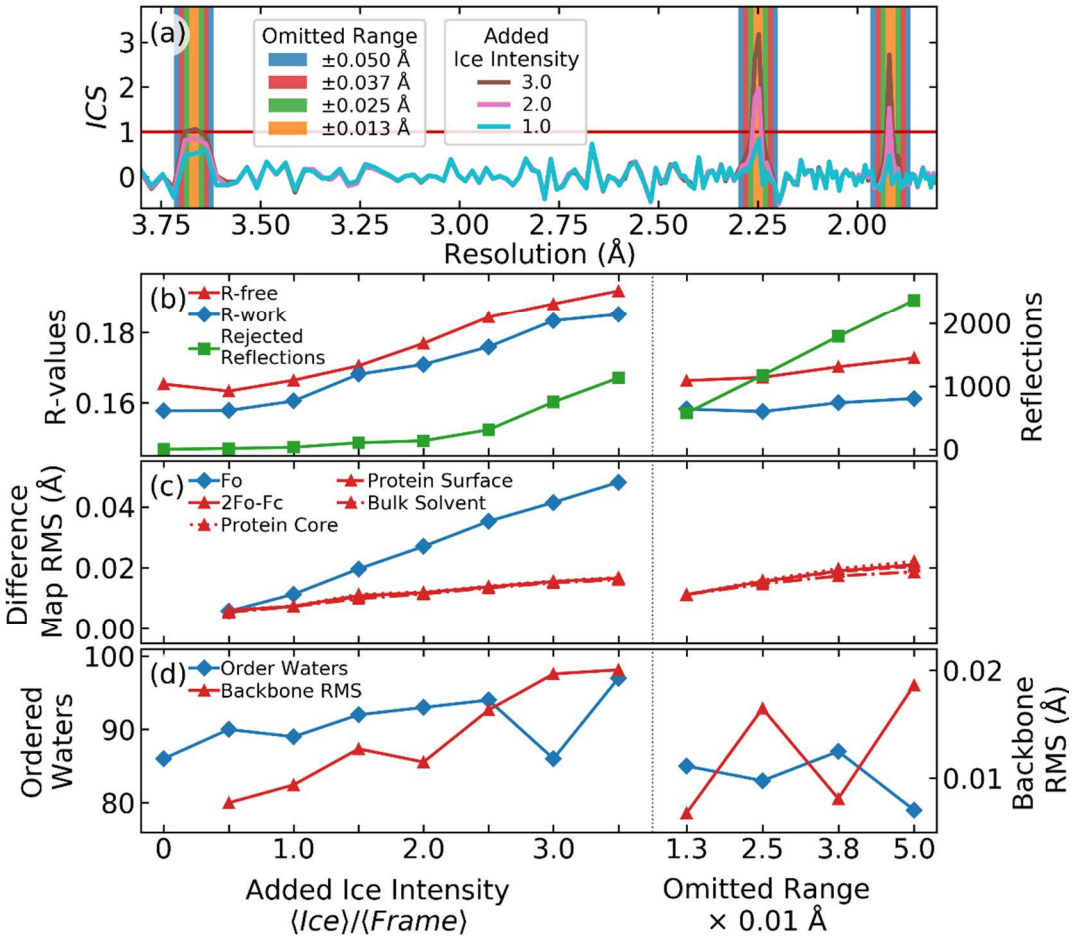


Figure 5.8 Effect of adding ice diffraction to an ice-free lysozyme data set on the refined structural models and electron density maps. (a) ICS scores from data sets with three different amounts of added ice corresponding to the left side of the plots below. Shading indicates the resolution ranges excluded from analysis, corresponding to the resolutions of ice rings that are common to all forms of ice, for data sets in which ice-contaminated reflections were omitted before modelling and refinement. (b) R-free, R-Work and total reflections excluded from refinement by Phenix. (c) RMS difference between electron density maps determined using ice-contaminated and ice-free data. (d) Total number of modeled water atoms added during Phenix refinement using ordered solvent parameters described in Section 5.2.8, and RMS difference between backbone atom positions. In (b)-(d), results in the left-hand frame are determined using all available reflections, for different ratios of average intensity in ice diffraction to average intensity in protein lattice reflections; results in the right-hand frame are determined using the largest amount of added ice diffraction, with reflections omitted within different resolution ranges around each of the ice diffraction peak locations in (a).

5.4 Discussion and Conclusions

5.4.1 Types and origins of ice diffraction in protein cryocrystallography

In an ideal protein cryocrystallography experiment, X-ray diffraction from all sources besides the protein crystal are minimized. This diffraction, typically diffuse scatter from liquid water, oils or polymers associated with the crystallography mounts, varies slowly with angle and is largely isotropic, facilitating accurate subtraction of this background from the protein lattice Bragg peaks.

In fact, intense, highly structured background diffraction generated by ice is frequently observed. This ice arises from three different sources: solvent internal to the protein crystal, solvent external to the protein crystal, and accumulated frost, each producing characteristic diffraction patterns. First, ice may form within the solvent cavities of the protein crystals. As we previously discussed (Moreau *et al.*, 2019), ice formation is strongly suppressed by its nanoconfinement within the protein network and is far less likely to form there than in surface solvent, even in crystals with large solvent cavities. Long range propagation of crystalline ice order within the solvent cavities is inhibited by an intact protein lattice. Ice crystal growth causes disruption of the protein lattice, degrading protein diffraction mosaicity and resolution. As a result, ice grain sizes tend to be small, the number of grains large, and the ice ring diffraction homogeneous and isotropic. The kinetics of ice nucleation and growth in deeply supercooled solutions (Malkin *et al.*, 2015) and within the nanoconfined environment (González Solveyra *et al.*, 2011) favours formation of stacking disordered ice. When the internal solvent has low concentrations of cryoprotectants and/or other solutes, the cubic stacking fraction is near 50%; as cryoprotectants are added and/or cooling rates are increased, the cubic stacking fraction increases (Moreau *et al.*, 2019), and the grain size decreases (Fig. 5.5).

Second, ice may form in residual solvent present on the crystal surface. This solvent typically contains substantial concentrations of cryoprotectants, including salts used for crystallization. Cryoprotectants

lower the freezing temperature and raise the glass transition temperature. They complicate ice nucleation and growth because they must be excluded from the growing ice crystal. As ice crystals grow, cryoprotectants become concentrated in the remaining uncrystallised solvent, further lowering its freezing temperature, raising its glass transition temperature, and inhibiting ice crystal growth. However, when ice forms in residual surface solvent, its growth is unhindered by the protein crystal, allowing ice to grow to a modest grain size, producing lumpy or streaky diffraction rings. For small cryoprotectant concentrations and small cooling rates, this surface ice will have a primarily hexagonal character with a larger grain size, and the lumpiness of the diffraction at the ice ring positions will be pronounced. For larger cryoprotectant concentrations and/or larger cooling rates, the ice becomes stacking disordered. The fraction of cubic planes in the stacking disordered structure increases, the ice grain size decreases, and the ice rings become more homogeneous and isotropic with increasing cooling rate and increasing cryoprotectant concentration. These trends are evident in Fig. 5.2.

Third, frost may be present on the crystal or sample holder/loop surface. Frost may condense from moist air during post-cooling handling and during data collection in a misaligned or otherwise malfunctioning cold gas stream. Frost may accumulate in the liquid nitrogen used for cooling and storage and adhere to the crystal surface. Since it forms from pure water and under modest cooling rates, frost is always pure hexagonal ice. Frost is by far the most common source of zingers in cryogenic temperature diffraction frames, which are generated when only a small number of ice crystals are present in the X-ray beam, too small to generate continuous or quasi-continuous ice diffraction rings.

The presence of one of these forms of ice does not require the presence of another form. Frost may be present in the absence of any other form of ice because it can accumulate on the crystal while stored in liquid nitrogen, at temperatures far below the glass transition temperature of the solvent internal or external to the crystal. Because the protein crystal provides a physical obstacle to the propagation of ice from its exterior and the nanoconfinement of solvent within the crystal raises the glass transition

temperature and lowers the freezing temperature, it is possible for ice to form in the external solvent without penetrating the protein crystal.

5.4.2 Trends in ice formation vs cryo concentration and final temperature

As cryoprotectant concentration grows, the equilibrium freezing temperature T_f drops (from ~266 K at 20% v/v to 255 K at 40% v/v) (Lane, 1925). Assuming a roughly exponential approach to the final (cryostream) temperature T_{final}, the average cooling rate between T_f and, e.g., T_{final} + 5 K decreases with increasing cryoprotectant concentration. This effect is most pronounced in the 240 K data, giving slower ice growth that favours hexagonal stacking. As the final temperature is lowered, the average cooling rate from T_f to intermediate temperatures increases, nucleation occurs at deeper supercooling where growth rates are larger, and stacking disordered ice is generated. The ice growth rate decreases with increasing cryoprotectant concentration. Growing ice rejects cryoprotectant, which becomes concentrated in the remaining uncrystallised solution, raising its glass transition temperature and decreasing the cooling rate required for it to vitrify. As a result, as the initial cryoprotectant concentration increases, a decreasing fraction of the sample will crystallize before the remaining liquid vitrifies.

The crystallite size increases as the temperature to which a crystal is cooled increases. This is consistent with average cooling rates from the solvent freezing temperature to near the final temperature being smaller and the time available for ice to nucleate and grow being longer when the final temperature is higher. It is also consistent with smaller ice nucleation rates and larger ice growth rates at temperatures modestly below the freezing temperature, compared with at much lower temperatures.

As more cryoprotectant is added, the size of the crystallites decreases. Ice crystals contain no or very little cryoprotectant, and cryoprotectant molecules are excluded at the growing ice crystal's surface. Cryoprotectant thus becomes concentrated in the remaining uncrystallised solution, lowering the solution's freezing point, raising its glass transition temperature, and lowering the critical cooling rate

required for it to vitrify. These reduce ice growth rates, increase the fraction solvent that vitrifies without ice formation, and reduce ice crystallite size.

5.4.3 Detection and prevalence of ice in PDB deposits

Both the Ice Finder Score (IFS) of Thorn et al. AUSPEX and our extension, the Ice Contamination Score (ICS), allow automated detection of ice biasing of experimental structure factor amplitudes. Based on comparisons between these scores and visual "scoring" of corresponding 2D diffraction frames, the ICS provides a significantly lower rate of false positives and false negatives. The utility of the two approaches can be scored using their *sensitivity* (the ratio of the number of true positives to the sum of the number of true positives and false negatives), which measures the ability to correctly identify entries having ice biasing, their *specificity* (the ratio of the number of true negatives to the sum of the number of true negatives and false positives), which measures the ability to correctly identify entries that are ice free. For a test set of 200 PDB entries, our ICS-based algorithm improved the sensitivity and specificity relative to AUSPEX from 59% to 86% and from 90% to 98%, respectively.

For the broader PDB, both AUSPEX and our algorithm indicate similar overall levels of ice contamination – 19% and 16% of entries, respectively, a fraction that has remained roughly constant over the last 30 years. Our methods also give information on the type of ice present, which is related to its origin. Of PDB entries with ice contamination, roughly 21% show hexagonal ice, due to crystal/loop frosting and contamination, excess solvent surrounding the crystal, and perhaps also due to crystals with mechanically damaged regions in which pools of solvent larger than those in the ordered lattice's solvent cavities may form. This fraction has increased by roughly 60% over the last 20 years. Diffraction from this hexagonal ice tends to be anisotropic and inhomogeneous, unlike diffraction from the stacking disordered ice that forms in the solvent cavities of reasonably well-ordered crystals. As a result, hexagonal ice

diffraction tends to be much more difficult for advanced background subtraction methods to account for and so has a larger impact on the integrated structure factors.

The increasing prevalence of hexagonal ice, even as beamline cryocooling hardware has evolved to largely eliminate frosting, suggests that cryocooling protocols for an increasing fraction of structural targets have been inadequate. This could reflect a greater focus on challenging targets - crystals with high solvent contents, large solvent cavities, fragile lattices, inconvenient (needles, clusters) growth habits, and/or for which suitable cryoprotectant conditions may be difficult to identify. Time-consuming cryoprotection protocols may have been relaxed to increase throughput, and a larger fraction of crystallographic data is now collected by those for whom crystallography is not a primary focus. The shift to remote data collection may also be a factor. While a dry shipper does an excellent job at keeping the crystals cold in transit to a synchrotron facility (Owen *et al.*, 2004), they can accumulate frost due if the lids are left open or from frequently removed samples.. An increase in non-expert users of crystallography.

5.4.4 Impact of ice on refined electron density maps and structural models

The perturbations of integrated intensities at specific resolutions caused by ice rings occurs in “reciprocal space”. For homogeneous, isotropic ice rings, as are typically generated when ice forms in internal crystal solvent, these perturbations are spread isotropically in real space. As a result, they produce fluctuations in electron density that are largely uncorrelated with the protein structure in the unit cell, and so have the effect of increasing the "noise" in the map. The most obvious effect of this additional noise is degraded refinement statistics (e.g., R-Work and R-Free values). At well-ordered protein atom positions, the noise has negligible effects. But it has sufficient amplitude to obscure fine details of the maps relevant to solvent atom placement and ligand identification and analysis, and to create details that may be erroneously interpreted.

The impact of ice on electron density maps and structural models depends on the relative intensity of the ice rings to the protein Bragg peaks. This is evident in the nearly monotonic increase of the R-values and rms differences between the refined maps and models with increasing ice contamination, in larger ICS scores for high resolution peaks (as in Figs. 5.4 and 5.8a).

5.4.5 Minimizing ice in cryocrystallography

The factors that affect ice formation in cryocrystallography have been extensively discussed elsewhere (Pflugrath, 2015; Garman, 1999; Garman & Doublié, 2003). We conclude by summarizing them.

Cooling rates. Cooling rates in current practice vary by at least three orders of magnitude and is most heavily dependent on the thermal mass of the sample (crystal plus surrounding liquid) and whether gas or liquid cryogens are used (Chinte *et al.*, 2005; Kriminski *et al.*, 2003; Teng & Moffat, 1998; Warkentin *et al.*, 2008; Walker *et al.*, 1998). The cooling rate also depends on plunge speed, choice of liquid cryogen (Teng & Moffat, 1998) and extent of precooling by cold gas present above the liquid cryogen (Warkentin *et al.*, 2006).

Cryoprotectant concentration. In aqueous solutions, the minimum cooling rates required to obtain a sample with no detectable ice increase exponentially with decreasing cryoprotectant concentration (Hopkins *et al.*, 2012; Warkentin *et al.*, 2008).. The addition of cryoprotectants can damage protein crystals or interfere with ligand binding, placing an upper limit to the amount of cryoprotectant that can be used in an experiment.

Amount of solvent surrounding the crystal. Ice formation of the solvent internal to the crystals is strongly suppressed by nanoconfinement (Moreau *et al.*, 2019). For a given cooling rate, the external solvent requires much larger cryoprotectant concentrations to prevent ice formation. This external solvent can be wicked away (Pflugrath, 2015) or replaced with oils (Kwong & Liu, 1999; Riboldi-Tunnicliffe & Hilgenfeld, 1999; Panjikar & Tucker, 2002; Warkentin & Thorne, 2009).

Crystal solvent content and solvent cavity size. Ice formation in internal solvent is strongly suppressed by nanoconfinement, and solvent within the first two hydration layers generally does not crystallize. Ice is thus most likely to form within crystals having large solvent cavities and large fractions of bulk-like internal solvent (Moreau *et al.*, 2019).

Crystal perfection. Growth defects like dislocations, inclusions and vacancies, more general lattice-scale disorder caused by imperfect molecular packing, as well as defects/disorder created by osmotic shock during cryoprotectant soaks, inadvertent crystal dehydration and mechanical damage during handling all can produce solvent pockets within the crystal that are much larger than the solvent cavities in ordered portions of the crystal that are identified by crystallography. Solvent in these relatively less confined regions has a higher freezing point (Findenegg *et al.*, 2008) and ice nucleation rate (Li *et al.*, 2013). Ice may thus be orders of magnitude more likely to first nucleate in these larger solvent pockets, and the ice that forms will have a larger grain size and be more likely to generate anisotropic, lumpy diffraction than ice that forms within (initially) ordered solvent cavities.

Acknowledgements All X-ray data collection was performed at the Cornell High-Energy Synchrotron Source (CHESS), which is supported by the NSF under award DMR-0936384, using the Macromolecular Diffraction at CHESS (MacCHESS) facility, which is supported by NIH award GM-103485.

5.5 References

Ashiotis, G., Deschildre, A., Nawaz, Z., Wright, J. P., Karkoulis, D., Picca, F. E. & Kieffer, J. (2015). *J. Appl. Cryst.* **48**, 510–519.

Baker, I. (2002). *Cryst. Growth Des.* **2**, 127–134.

Chinte, U., Shah, B., DeWitt, K., Kirschbaum, K., Pinkerton, A. A. & Schall, C. (2005). *J. Appl. Cryst.* **38**, 412–419.

Findenegg, G. H., Jähnert, S., Akcakayiran, D. & Schreiber, A. (2008). *ChemPhysChem*. **9**, 2651–2659.

Fortes, A. D. (2018). *Acta Cryst.* **B74**, 196–216.

Garman, E. (1999). *Acta Cryst.* **D55**, 1641–1653.

Garman, E. F. & Doublié, S. (2003). *Methods Enzymol.* **368**, 188–216.

González Solveyra, E., De La Llave, E., Scherlis, D. A. & Molinero, V. (2011). *J. Phys. Chem. B*. **115**, 14196–14204.

Grabowski, M., Langner, K. M., Cymborowski, M., Porebski, P. J., Sroka, P., Zheng, H., Cooper, D. R., Zimmerman, M. D., Elsliger, M. A., Burley, S. K. & Minor, W. (2016). *Acta Cryst.* **D72**, 1181–1193.

Hopkins, J. B., Badeau, R., Warkentin, M. & Thorne, R. E. (2012). *Cryobiology*. **65**, 169–178.

Kabsch, W. (2010). *Acta Cryst.* **D66**, 125–132.

Knudsen, E. B., Sørensen, H. O., Wright, J. P., Goret, G. & Kieffer, J. (2013). *J. Appl. Cryst.* **46**, 537–539.

Kriminski, S., Kazmierczak, M. & Thorne, R. E. (2003). *Acta Cryst.* **D59**, 697–708.

Kuhs, W. F., Sippel, C., Falenty, A. & Hansen, T. C. (2012). **109**, 21259–21264.

Kwong, P. D. & Liu, Y. (1999). *J. Appl. Cryst.* **32**, 102–105.

Lane, L. B. (1925). *Ind. Eng. Chem.* **17**, 924.

Langford, J. I. & Wilson, A. J. C. (1978). *J. Appl. Cryst.* **11**, 102–113.

Leslie, A. G. W. (2006). *Acta Cryst.* **D62**, 48–57.

Li, T., Donadio, D. & Galli, G. (2013). *Nat. Commun.* **4**, 1–6.

Liebschner, D., Afonine, P. V., Baker, M. L., Bunkoczi, G., Chen, V. B., Croll, T. I., Hintze, B., Hung, L. W., Jain, S., McCoy, A. J., Moriarty, N. W., Oeffner, R. D., Poon, B. K., Prisant, M. G., Read, R. J., Richardson, J. S., Richardson, D. C., Sammito, M. D., Sobolev, O. V., Stockwell, D. H., Terwilliger, T. C., Urzhumtsev, A. G., Videau, L. L., Williams, C. J. & Adams, P. D. (2019). *Acta Cryst.* **D75**, 861–877.

Malkin, T. L., Murray, B. J., Salzmann, C. G., Molinero, V., Pickering, S. J. & Whale, T. F. (2015). *Phys. Chem. Chem. Phys.* **17**, 60–76.

Moreau, D. W., Atakisi, H. & Thorne, R. E. (2019). *IUCrJ*. **6**, 346–356.

Olivero, J. J. & Longbothum, R. L. (1977). *J. Quant. Spectrosc. Radiat. Transf.* **17**, 233–236.

Owen, R. L., Prilchard, M., Garman, E., Pritchard, M. & Garman, E. (2004). *J. Appl. Cryst.* **37**, 1000–1003.

Panjikar, S. & Tucker, P. A. (2002). *J. Appl. Cryst.* **35**, 117–119.

Parkhurst, J. M., Thorn, A., Vollmar, M., Winter, G., Waterman, D. G., Fuentes-Montero, L., Gildea, R. J., Murshudov, G. N. & Evans, G. (2017). *IUCrJ*. **4**, 626–638.

Parkhurst, J. M., Winter, G., Waterman, D. G., Fuentes-Montero, L., Gildea, R. J., Murshudov, G. N. & Evans, G. (2016). *J. Appl. Cryst.* **49**, 1912–1921.

Pflugrath, J. W. (2015). *Acta Cryst.* **F71**, 622–642.

Read, R. J. (1999). *Acta Cryst.* **D55**, 1759–1764.

Riboldi-Tunnicliffe, A. & Hilgenfeld, R. (1999). *J. Appl. Cryst.* **32**, 1003–1005.

Scardi, P., Leoni, M. & Delhez, R. (2004). *J. Appl. Cryst.* **37**, 381–390.

Stokes, A. R. & Wilson, A. J. C. (1942). *Math. Proc. Cambridge Philos. Soc.* **38**, 313–322.

Teng, T. Y. & Moffat, K. (1998). *J. Appl. Cryst.* **31**, 252–257.

Thorn, A., Parkhurst, J., Emsley, P., Nicholls, R. A., Vollmar, M., Evans, G. & Murshudov, G. N. (2017). *Acta Cryst.* **D73**, 729–737.

Thürmer, K. & Bartelt, N. C. (2008). *Phys. Rev. B - Condens. Matter Mater. Phys.* **77**, 1–10.

Treacy, M. M. J., Newsam, J. M. & Deem, M. W. (1991). *Proc. - R. Soc. London, A*. **433**, 499–520.

Ungár, T., Ott, S., Sanders, P. G., Borbély, A. & Weertman, J. R. (1998). *Acta Mater.* **46**, 3693–3699.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., Vijaykumar, A., Bardelli, A. Pietro, Rothberg, A., Hilboll, A., Kloeckner, A., Scopatz, A., Lee, A., Rokem, A., Woods, C. N., Fulton, C., Masson, C., Häggström, C., Fitzgerald, C., Nicholson, D. A., Hagen, D. R., Pasechnik, D. V., Olivetti, E., Martin, E., Wieser, E., Silva, F., Lenders, F., Wilhelm, F., Young, G., Price, G. A., Ingold, G. L., Allen, G. E., Lee, G. R., Audren, H., Probst, I., Dietrich, J. P., Silterra, J., Webber, J. T., Slavič, J., Nothman, J., Buchner, J., Kulick, J., Schönberger, J. L., de Miranda Cardoso, J. V., Reimer, J., Harrington, J., Rodríguez, J. L. C., Nunez-Iglesias, J., Kuczynski, J., Tritz, K., Thoma, M., Newville, M., Kümmerer, M., Bolingbroke, M., Tartre, M., Pak, M., Smith, N. J., Nowaczyk, N., Shebanov, N.,

Pavlyk, O., Brodtkorb, P. A., Lee, P., McGibbon, R. T., Feldbauer, R., Lewis, S., Tygier, S., Sievert, S., Vigna, S., Peterson, S., More, S., Pudlik, T., Oshima, T., Pingel, T. J., Robitaille, T. P., Spura, T., Jones, T. R., Cera, T., Leslie, T., Zito, T., Krauss, T., Upadhyay, U., Halchenko, Y. O. & Vázquez-Baeza, Y. (2020). *Nat. Methods*. **17**, 261–272.

Walker, L. J., Moreno, P. O. & Hope, H. (1998). *J. Appl. Cryst.* **31**, 954–956.

Warkentin, M., Berejnov, V., Husseini, N. S. & Thorne, R. E. (2006). *J. Appl. Cryst.* **39**, 805–811.

Warkentin, M., Stanislavskaia, V., Hammes, K. & Thorne, R. E. (2008). *J. Appl. Cryst.* **41**, 791–797.

Warkentin, M. & Thorne, R. E. (2009). *J. Appl. Cryst.* **42**, 944–952.

Wilson, A. J. C. (1949). *Acta Cryst.* **2**, 318–321.

6 APPENDICES

6.1 SUPPORTING INFORMATION FOR CHAPTER 2

6.1.1 Protein Volume Estimations

It is possible to determine the solvent fraction using the calculated solvent excluded volume (SEV) of the refine atomic models. The inability to draw a physically accurate dividing surface between the protein and solvent region makes this approach unreliable for determining the solvent fraction. Ball rolling algorithms are a common method to calculate a proteins SEV from atomic coordinate (Richards, 1977; Connolly, 1985). These algorithms generally assign a radius to each atom of the protein and take a sphere (probe) of constant radii and "roll" it around the protein. The surface traced out by the edge of the probe is defined as the protein's SEV. The volumes estimated by these methods are heavily dependent on the radii assigned to the atoms and the probe. Published parameters for the atomic radii are usually taken from analysis of deposited structures in the PDB (Li & Nussinov, 1998). These atomic radii are underestimated because they can only be determined from the highly ordered water molecules and include analysis of cryogenic structures. More subjectivity is included with X-ray crystallography data due to the lack of modelled hydrogens. Li & Nussinov point out that two probe sizes should be used along with the coulombic or Van der Waals atomic radii for protein atoms used in these ball rolling calculations depending on if the atom is polar or nonpolar, respectively. While this makes sense in principal, adding this functionality is of no practical use. The difference in the SEV volume calculated with one or two probes is comparable to the uncertainty in the estimated SEV due to the true degree of uncertainty in estimated atomic radii. Kundrot & Richards (1988) calculate the density of the solvent within tetragonal lysozyme crystals using protein volumes determined by ball rolling algorithms and their conclusion is that these methods do not give accurate results.

While the absolute volumes determined with ball rolling algorithms should be viewed with skepticism, they provide a powerful method to calculate differences in volumes. Care must still be taken when setting up the grid sizes. Ball rolling algorithms create a three-dimensional grid and grid sizes on the order of 0.05 to 0.1 Å to achieve numerical convergence (Connolly, 1985). The smallest grid size available on the 3Vee web server (Voss & Gerstein, 2010) is 0.5 Å resulting in calculated volumes not reliable for research purposes.

Persson (2018) argues that a more realistic approach to defining this dividing surface is to use a Voronoi tessellation algorithm (Cazals, 2006). While these algorithms are useful in molecular dynamics simulations, they require complete knowledge of the hydration water's positions; a requirement not satisfied by room temperature crystal structures, making them unsuitable for protein volume calculations based on crystallographic structures.

Lysozyme NaCl Soak

% w/v acetate	% w/v Na		% w/v Cl		Density (g / ml)	
0.059	0.605	0.019	0.86	0.005	1.012	0.002
	2.234	0.008	3.40	0.03	1.0431	0.0015
	4.761	0.004	7.33	0.02	1.0799	0.0016
	7.35	0.18	11.35	0.2	1.1183	0.0012
	9.24	0.14	14.14	0.003	1.1595	0.0006
	11.05	0.1	17.22	0.08	1.1936	0.0005

Lysozyme glycerol soak

% w/v acetate	% w/v NaCl	% w/v glycerol		Density (g / ml)	
0.59	3.5	0	0	1.0324	0.0012
		10.08	0.09	1.0528	0.0003
		19.28	0.1	1.0757	0.0009
		29.91	0.1	1.1011	0.0008
		34.33	0.22	1.1083	0.0007

Apoferritin glycerol soak

% w/v acetate	% w/v CdSO4	% w/v AmSO4	% w/v glycerol		Density (g / ml)	
0.59	2	10	0	0	1.0643	0.0007
			23.4	0.3	1.1160	0.0007
			40.4	0.5	1.1524	0.0007
			61.1	1	1.1901	0.0007

Table 6.1.1 Composition and density of soak solutions.

Partial Specific Volume (cm^3/g)	Temperature (C)	Protein Concentration (mg/ml)	Method	Citation
0.732	------	5 to 20	SAXS	Svergun 1998
0.7556	------	7, 10.5, 21, 41	SAXS	Huang 2007
0.7586	------	------	Density	Orthaber 2000
0.712	25	Extrapolated to 0	Density	Gekko 1979
0.7387	20	Extrapolated to 0	Density	Jirasek 2018
0.703	20	13.38 & 27.30	Density	Sophianopoulos 1962
0.714 0.723	25	2.2 to 8.5 & 1	Density	Charlwood 1956
0.688	25	0 to 9	Method of intercepts	Colvin 1952
0.713	25	Extrapolated to 0	Density	Gekko 1998
0.717	20	Extrapolated to 0	Density	Timasheff & Xie 2003
0.703	20	Extrapolated to 0	Density	Lee & Lee 1981
0.702	20	Extrapolated to 0	Density	Lee & Timasheff 1974
0.726	20	Extrapolated to 0	Density	Millero 1976
0.732	25	Extrapolated to 0	Density	Banerjee & Kishore 2006
0.699	18	3	Density	Chalikian 1996
0.725	22 to 26.7	3	Density	Gavish 1983

Table 6.1.2 Measured values of lysozyme's partial specific volume.

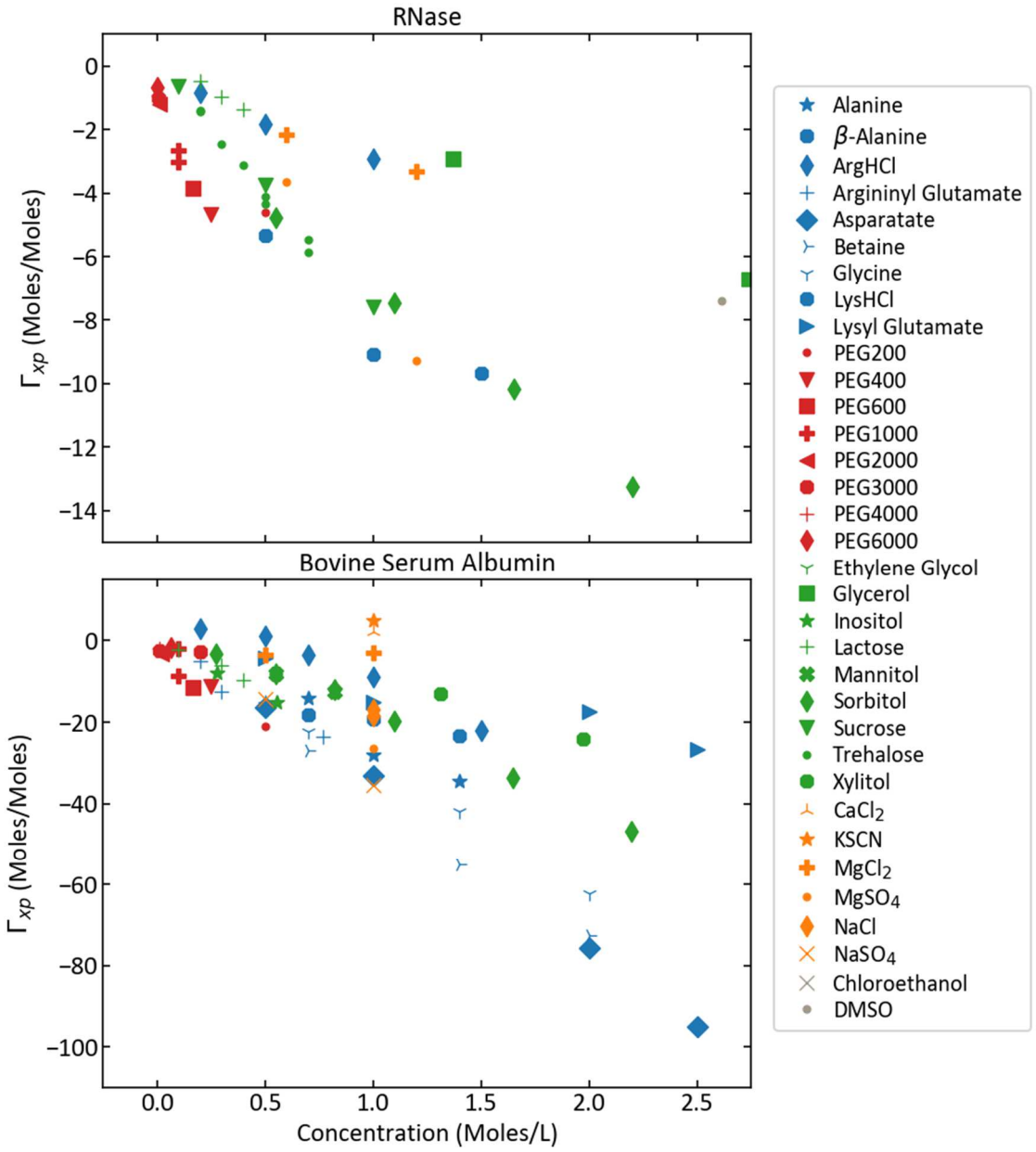


Figure 6.1.1 Experimental measurements of the preferential interaction in RNase and Bovine Serum Albumin.

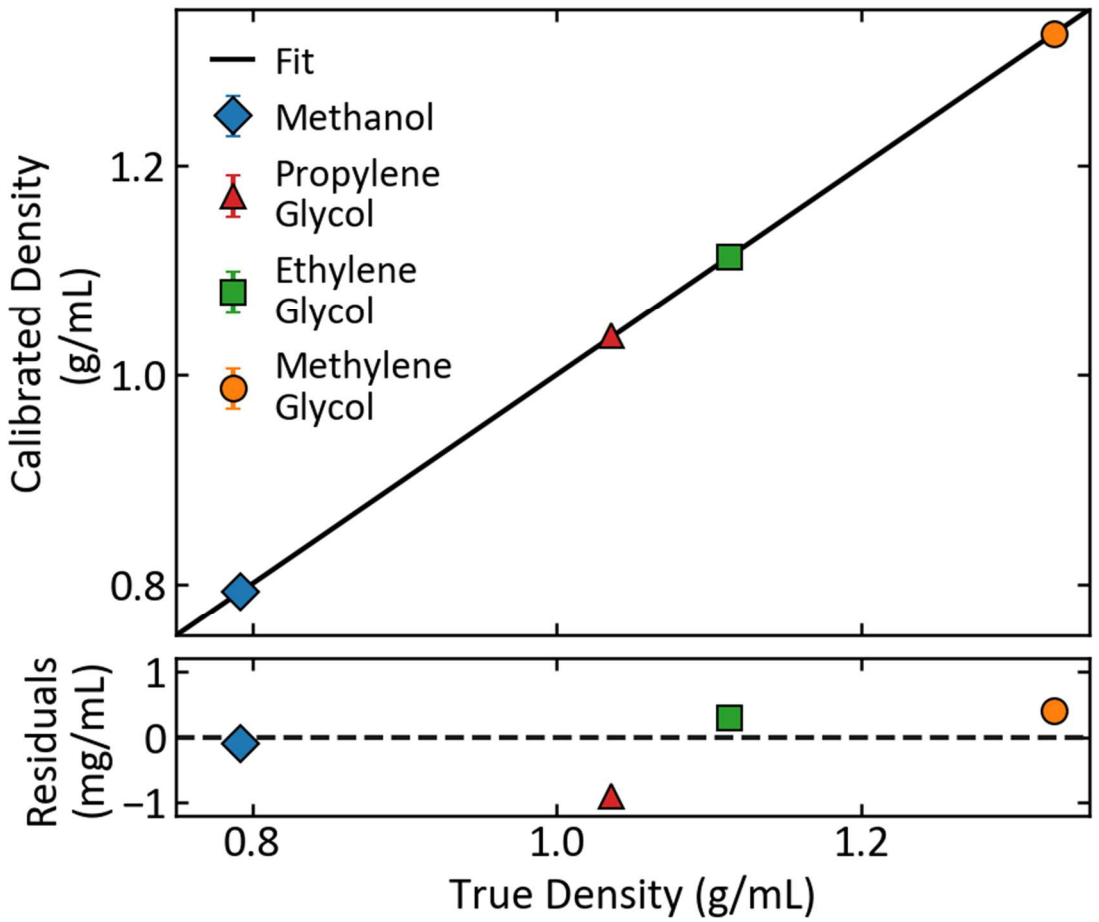


Figure 6.1.2 Calibration of weight used to measure density of soak solutions and 1PO / silicone oil mixtures. The buoyant force on the weight suspended in methanol, propylene glycol, ethylene glycol and methylene glycol was measured. The volume of the mass was determined using these measurements and the known densities of these solutions.

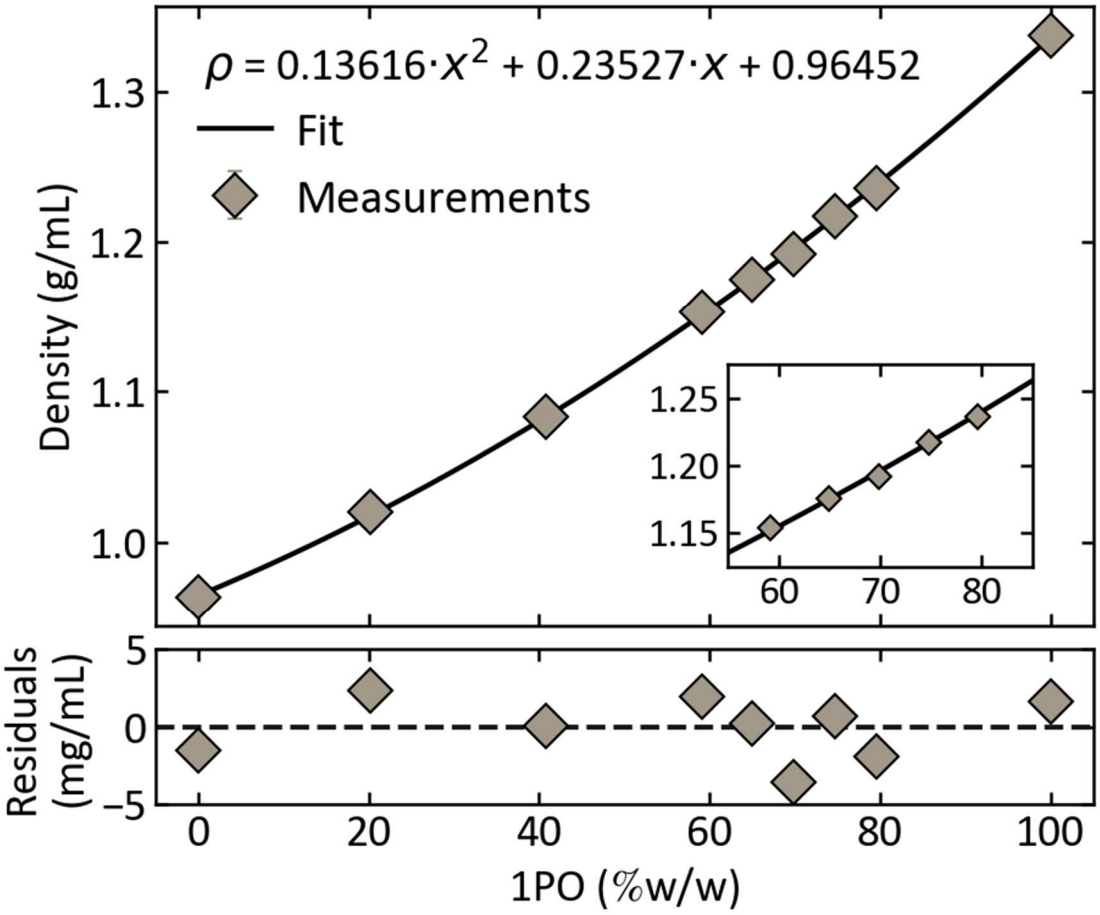


Figure 6.1.3 Density characterization of 1PO / silicone oil mixtures.

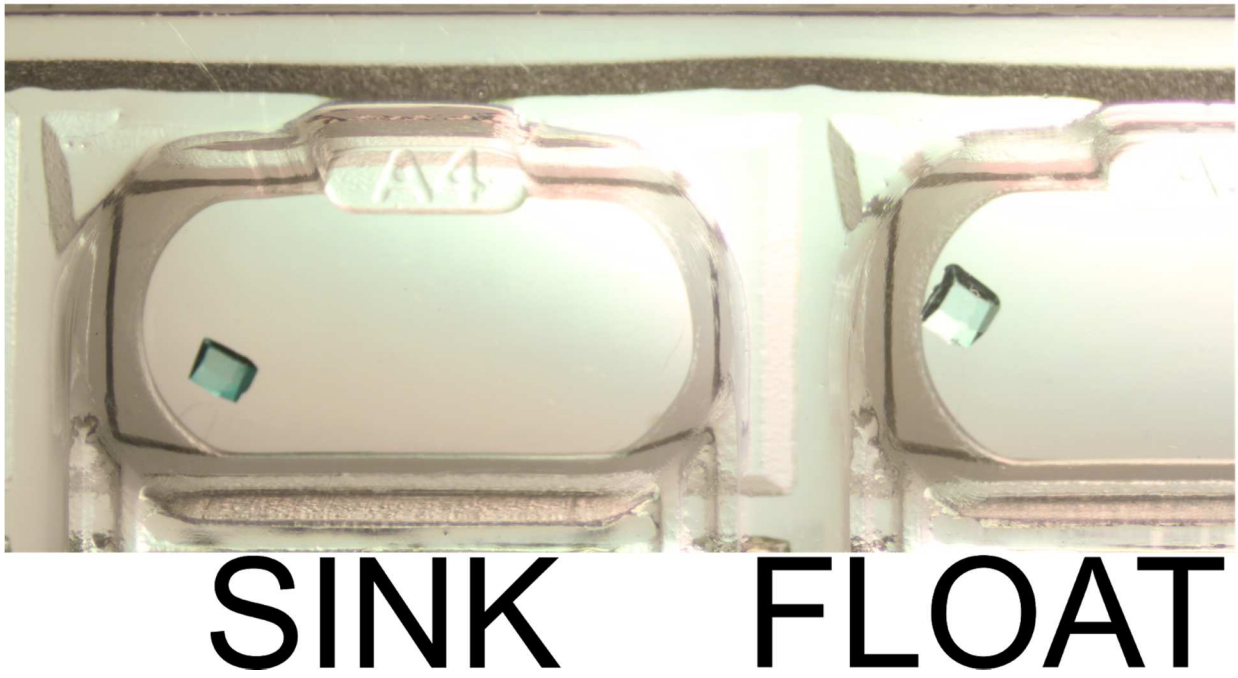


Figure 6.1.4 Example of sink float trials. A lysozyme crystal was soaked in a solution containing blue food coloring to aid in visualization. Oil was added to the both wells, a lower density mixture on the left and a higher density on the right. The crystal sink in the left well and floats in the right.

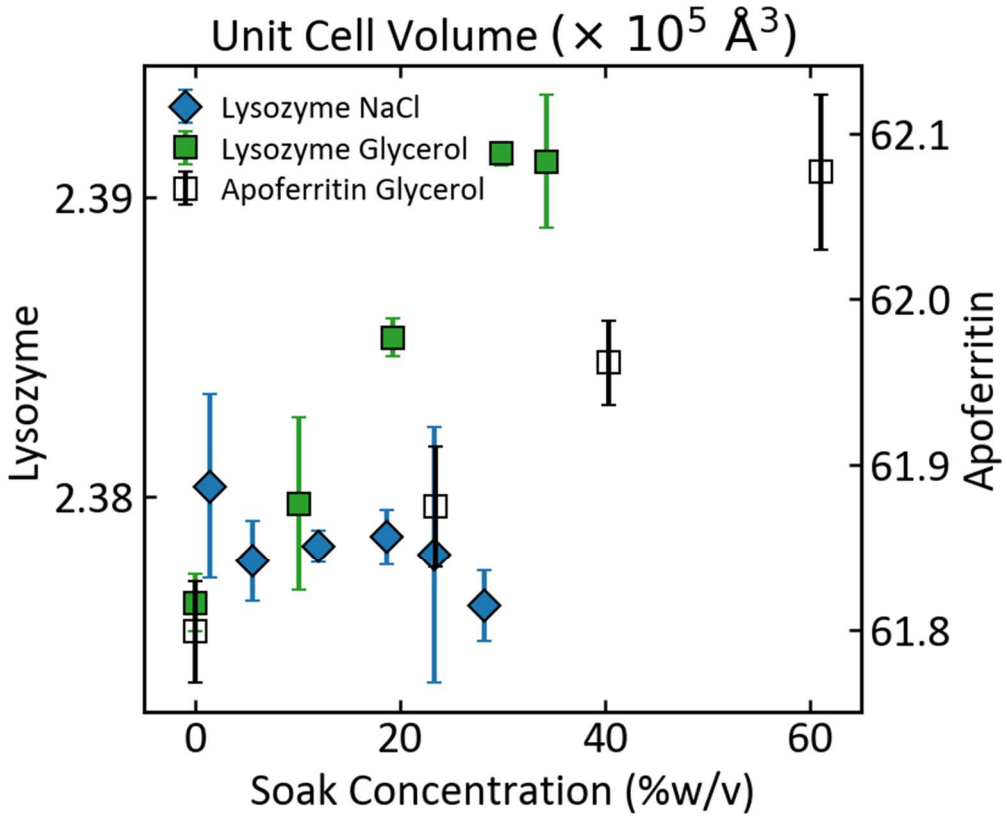


Figure 6.1.5 Measured unit cell volumes of lysozyme and apoferritin crystals soaked in different concentrations of cosolutes.

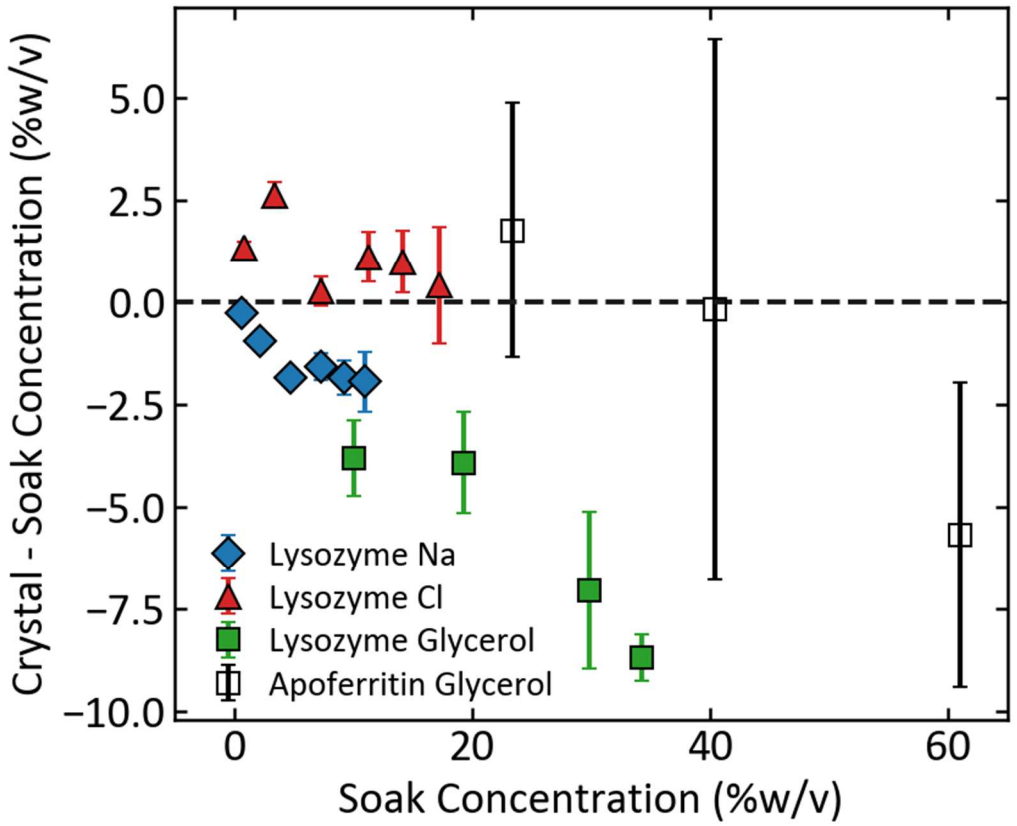


Figure 6.1.6 Difference in the measured composition of the solvent within the protein crystals with the soak solution.

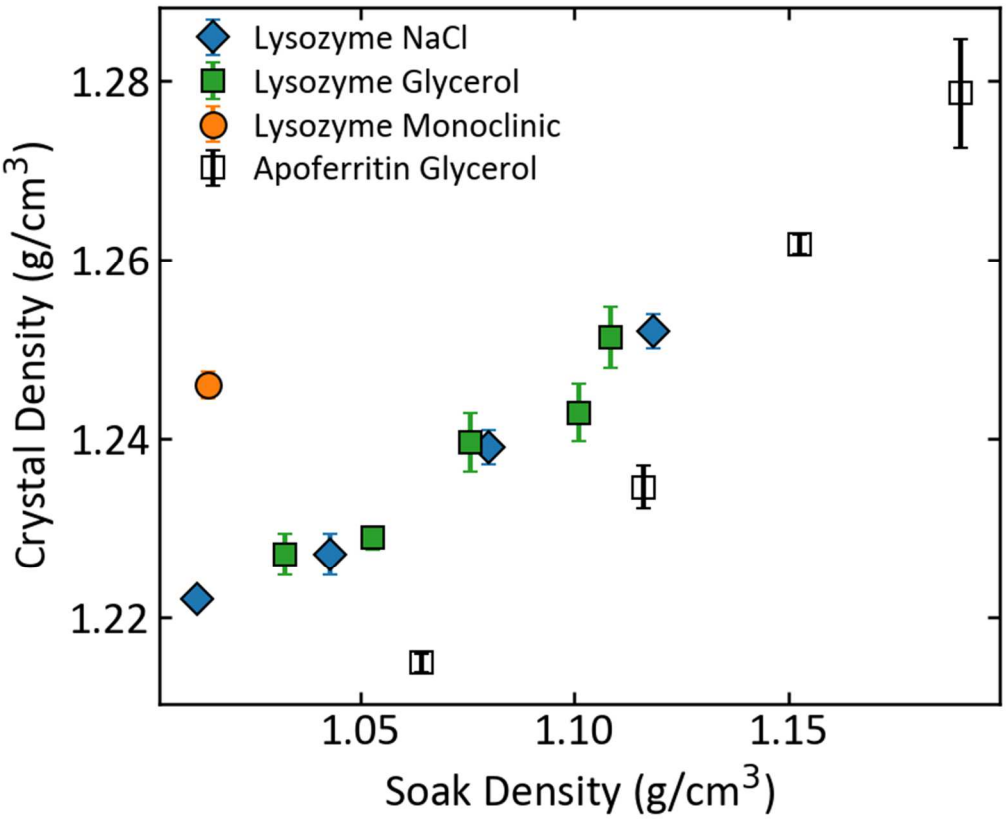


Figure 6.1.7 Measured densities of protein crystals.

6.2 SUPPORTING INFORMATION FOR CHAPTER 3

6.2.1 Properties of apoferritin, thaumatin, and lysozyme crystals

Our studies focused on crystals of cubic apoferritin and tetragonal thaumatin, with additional measurements on crystals of tetragonal lysozyme.

Apoferritin is comprised of twenty-four ferritin monomers, forming a 476 kDa complex having a nearly spherical cavity of diameter 68 Å for storage of iron; in the *apo* form the cavity is filled with solvent (Fig. 3.2(a) and Movie S1).

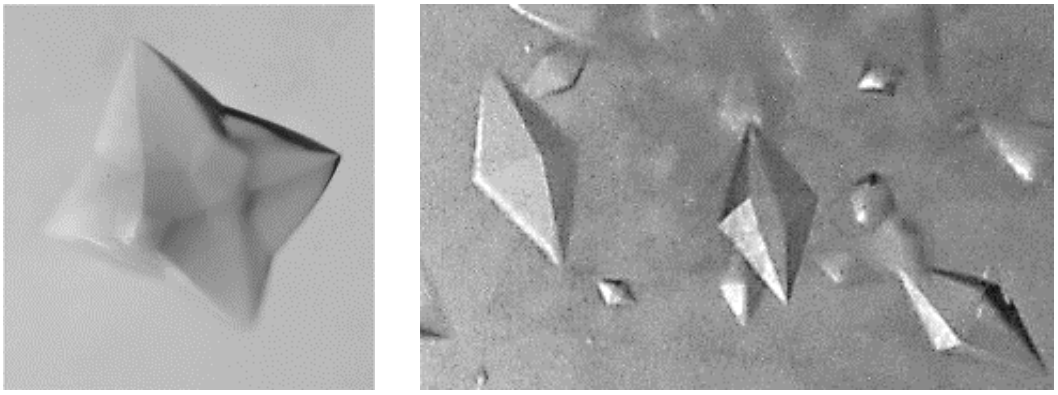
The solvent cavities in cubic apoferritin crystals have a face-centered cubic arrangement. These cavities are shown in Movie S2, where the unit cell origin has been shifted to place an apoferritin molecule in the center of the unit cell (making it appear as a body centered cubic arrangement). These large solvent cavities, their relatively simple geometry, and the resulting large fraction of bulk-like solvent make cubic apoferritin crystals excellent model systems for studies of ice formation under nanoconfinement. These large cavities also in part explain why previous crystallographic studies at T=100 K required relatively large cryoprotectant concentrations to prevent ice formation. Although the solvent cavities appear isolated from one another, when crystals are soaked in solutions containing iron, glycerol, or other small solutes/cryoprotectants, the non-aqueous components diffuse freely into or out of the cavities.

In tetragonal thaumatin crystals (Fig. 3.2(b) and Movies S3 and S4), the solvent is largely contained within twisted channels having a maximum diameter of 25 Å, near the peak of the solvent cavity size distribution in Fig. 3.1(d).

As shown in Fig. 6.2.1(b), thaumatin crystals grew as octahedra with flat faces. Complete removal of external solvent from the crystal facets was straightforward, and none of the crystals studied here showed

evidence of external ice in diffraction. Apoferritin crystals (Fig. 6.2.1(a)) had more complex habits that made external solvent removal more difficult, and roughly 25% of glycerol-free crystals showed diffraction from external ice.

Fig. 6.2.2 shows the distribution of distances between solvent atoms and the nearest protein surface in apoferritin, thaumatin, and lysozyme crystals, calculated using map_channels as discussed in Sec. S9. In lysozyme nearly all solvent molecules are within 6 Å of the protein surface (i.e., roughly within the first two hydration shells), whereas in apoferritin only ~50% of solvent molecules are that close.



(a) (b)

Figure 6.2.1 Crystals of (a) cubic apoferritin and (b) tetragonal thaumatin. Crystals are roughly 300 µm in size.

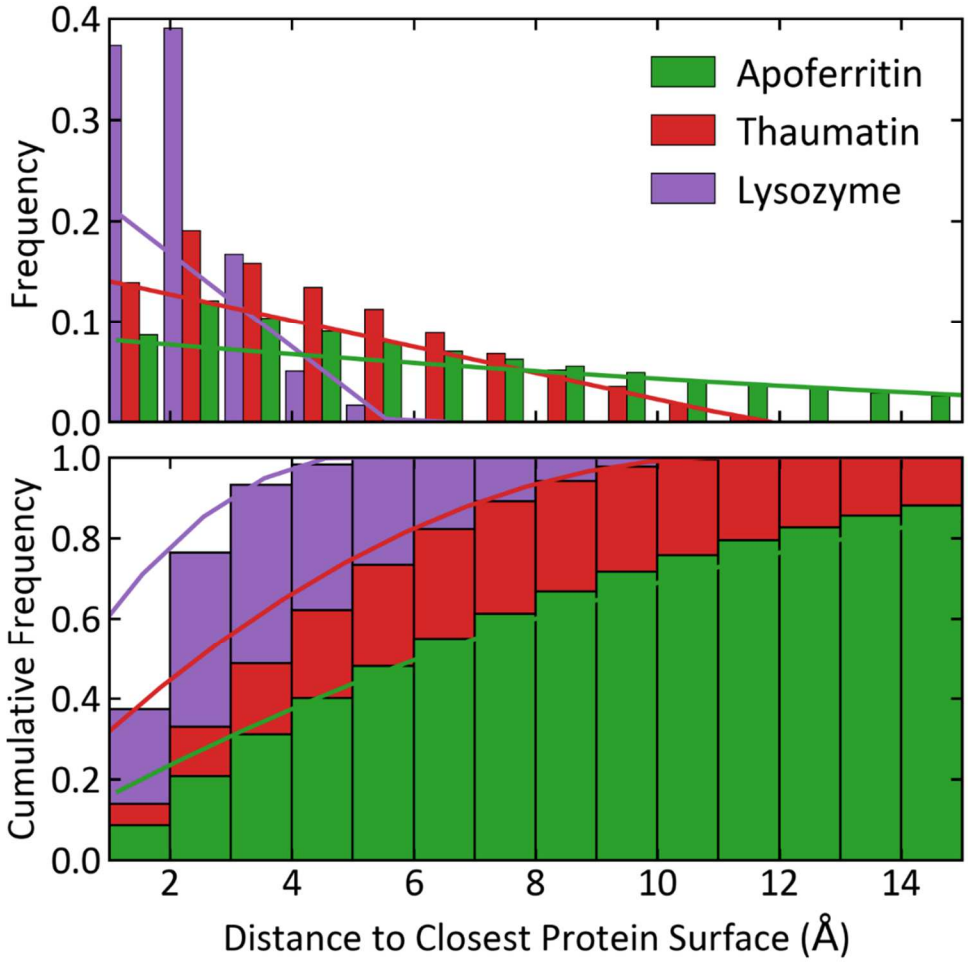


Figure 6.2.2 Distribution (top) and cumulative distribution (bottom) of distances from grid locations within the solvent channels to the nearest protein surface in cubic apoferritin, tetragonal thaumatin, and tetragonal lysozyme crystals at room temperature, determined using the results from map_channels. Lines represent analytical results for the distance to the closest surface from inside of a spherical shell (apoferritin) or cylindrical shell (thaumatin and lysozyme) of diameter equal to the maximum solvent cavity size in each protein crystal.

6.2.2 X-ray data collection

Time-dependent X-ray diffraction data was collected using the F1 beamline at the Cornell High-Energy Synchrotron Source (CHESS) using the experimental configuration shown in Fig. 6.2.3. The F1 station uses a horizontally focusing monochromator using a single bent triangular Si(111) crystal. The monochromator Bragg angle for the (111) reflection of silicon at this energy is 9.0°. Samples were illuminated using a Gaussian beam with a 65 µm FWHM and a divergence of 0.03°. The photon flux of 2.2×10^9 ph/s and photon energy of 12.7 keV gave dose rates of ~2 kGy/s. Diffraction patterns were recorded by a Pilatus 6M detector using a frame rate of 5 Hz.

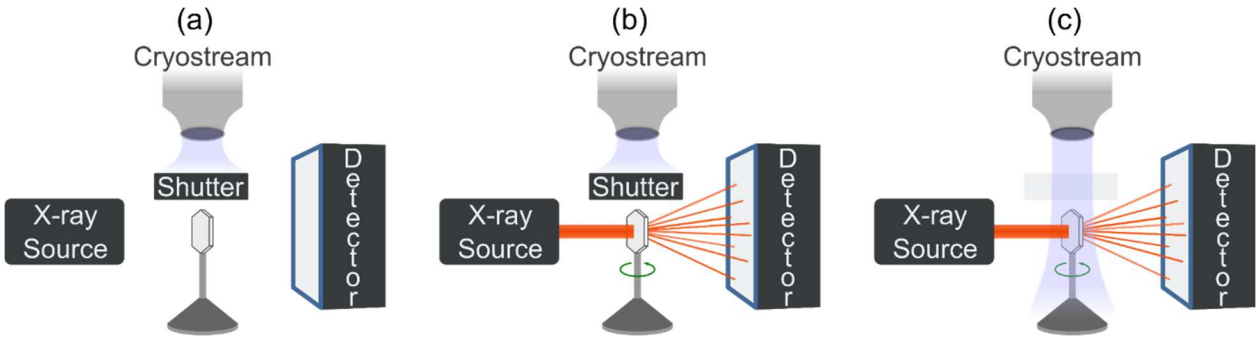


Figure 6.2.3 Experimental configuration used in time-resolved X-ray diffraction experiments at the Cornell High-Energy Synchrotron Source (CHESS). (a) sample with its external solvent removed is placed in the X-ray beam path, (b) room temperature data is collected, and (c) the cryostream is unshuttered and data collected while the crystal is cooled.

Sample temperature was controlled using a nitrogen gas stream with a flow rate of 5 L/min generated by an Oxford Cryosystems Cryostream 700 nitrogen gas cryocooler. The gas stream's temperature was varied between 100 K and 260 K using the cryocooler's internal heater, and was monitored using both the cryocooler's internal temperature sensor and using a thermocouple that was periodically placed at the sample position. For room temperature measurements, the cryocooler head was retracted and the gas stream was blocked using an air blade shutter. To commence cooling, the cryocooler head was extended and the air blade shutter was turned off.

Fig. 6.2.4 shows the temperature at the sample position recorded using a 250 µm bead thermocouple, when the air blade shutter was turned off at *t*=0 and the gas stream temperature was set to 200 K. The maximum cooling rate is ~300 K/s. Actual sample cooling times, deduced from the thermal contraction of the unit cell, varied somewhat with crystal size and amount of surrounding oil but were in the range of 0.5-2 s.

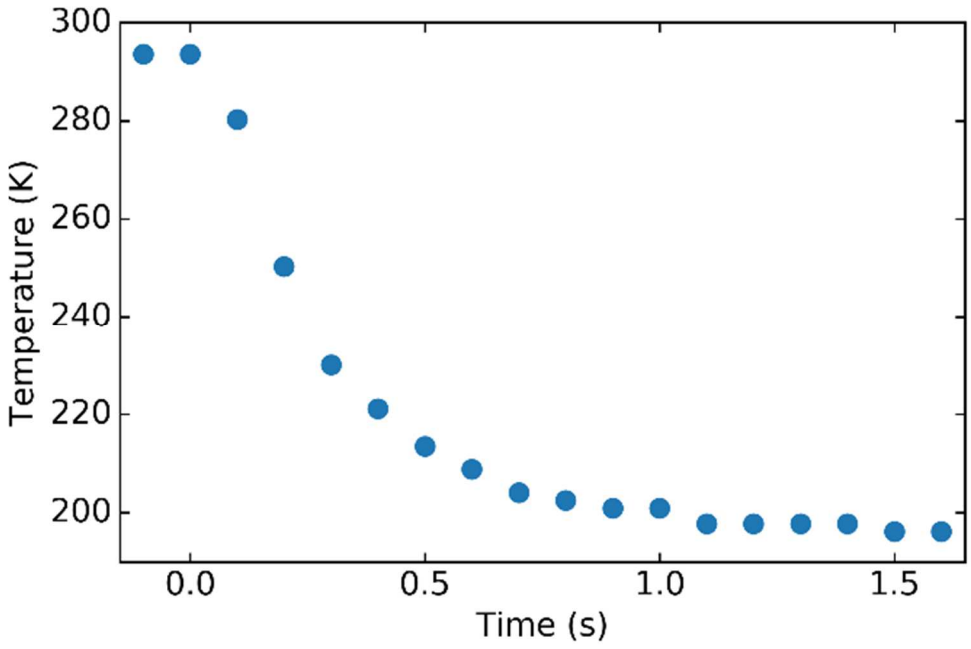


Figure 6.2.4 Temperature recorded at the sample position using a 250 µm bead thermocouple when the air blade shutter in Fig. 6.2.3 was turned off at t=0. Thermal response times of crystals could be faster or slower, depending on the volume of crystal and surrounding oil.

Diffraction data were acquired before, during, and after sample cooling as follows. With the cryostream retracted and blocked, a sample in its MicroRT tube was manually placed on the beamline goniometer stage, the crystal was allowed to settle in the oil, and then the crystal was translated to the position of the X-ray beam. The MicroRT tube was then removed, and automated data collection initiated using the beamline's ADX software. In a typical experiment, an initial set of 10 frames (0.5° of sample rotation and 0.2 s exposure time per frame, 5° rotation and 2 s total per exposure), was collected at room temperature. The crystal was then returned to its initial orientation, the cryostream was extended and unblocked, and a single set of 200 frames (0.5°, 0.1 or 0.2 s exposure time per frame, 100° rotation and 20-40 s total exposure) acquired. With a dose rate of ~2 kGy/s, total doses ranged from ~40 to ~100 kGy, much less than the half dose at all temperatures studied, so changes in diffraction properties with time were dominated by effects other than radiation damage.

A set of early experiments used a slightly different protocol. An initial set of room temperature frames were collected as described above. A set of 20 frames was collected, the crystal was rotated back to its starting orientation, and 8 additional sets of 5 frames each, recorded using the same starting orientation, were collected. This protocol was abandoned for subsequent experiments because ice could form while the samples were being returned to the initial orientation for the start of each set, and so diffraction frames recording ice formation would not be recorded. Most of the data collected by this protocol was from apoferritin crystals with glycerol concentrations of 20% and 40% v/v, for which ice formation was rare in any case.

Using this protocol, diffraction data was collected from crystals cooled to temperatures between 180 K and 260K. Cooling of each crystal was monitored using the time-dependent lattice parameters deduced from successive diffraction frames. Additional T=100 K data sets were collected using crystals that were plunge cooled in liquid nitrogen.

A total of 261 crystals of apoferritin, 168 crystals of thaumatin, and 4 crystals of lysozyme were examined at CHESS on seven different dates between November 2015 and March 2018. Measurement of a large number of samples was essential to drawing robust conclusions, because ice formation was studied versus both temperature and glycerol concentration; ice formation is stochastic; complete removal of external solvent was not always achieved, so external ice sometimes formed; the sample response depended to some extent on cooling time, which varied with crystal size and volume of surrounding oil; and because some samples developed cracks and other defects during post-growth handling that affected their response.

6.2.3 Processing of protein lattice diffraction

Diffraction frames were indexed, integrated, and scaled using *XDS* (Kabsch, 2010), with an input file acquired from the MacCHESS website and modified for use with a Pilatus 6M detector. Diffraction was generally strong to beyond 3.0Å for apoferritin and to beyond 1.7Å for thaumatin, and mosaicities were generally low and peaks were well-separated, so processing was usually straightforward. Data sets were processed in segments of 5 frames. A segment's refined outputs of unit cell, beam center, and sample to detector distance were used as inputs in the XDS.INP file when processing subsequent segments. As a check on these results, data were also processed using *XDS* in segments of 2 and 10 frames, and using *HKL2000*). The use of 5 frame segments balanced time resolution with variance in parameter estimation.

Unit cell values were taken directly from *XDS*'s CORRECT.LP output file.

Wilson *B* factors were estimated as half the slope of a linear fit to the natural log of the Bragg peak intensities (obtained using *phenix.merging*_stats (Adams *et al.*, 2010)) vs. $(\sin\theta / \lambda)^2$. Bragg peaks having resolutions better than (i.e., numerically smaller than) 4 Å were used to determine *B* factors.

The refined beam divergence and crystal mosaicities, calculated using *XDS* as the standard deviation of Gaussians, were 0.015° and 0.045°, respectively. This latter value was a "floor" imposed by

XDS that did not reflect actual crystal mosaicities. Diffraction frames were also indexed, integrated and scaled using *HKL2000*, which had a mosaicity "floor" close to the refined beam divergence. Mosaicity values reported here were calculated using values from the *HKL2000* *.x integration files, converted from full width at half maximum to standard deviation to allow comparison with *XDS* results. Figs. 6.2.5 and 6.2.6 show Wilson B factors and mosaicities, respectively, for apoferritin and thaumatin crystals vs temperature and glycerol concentration.

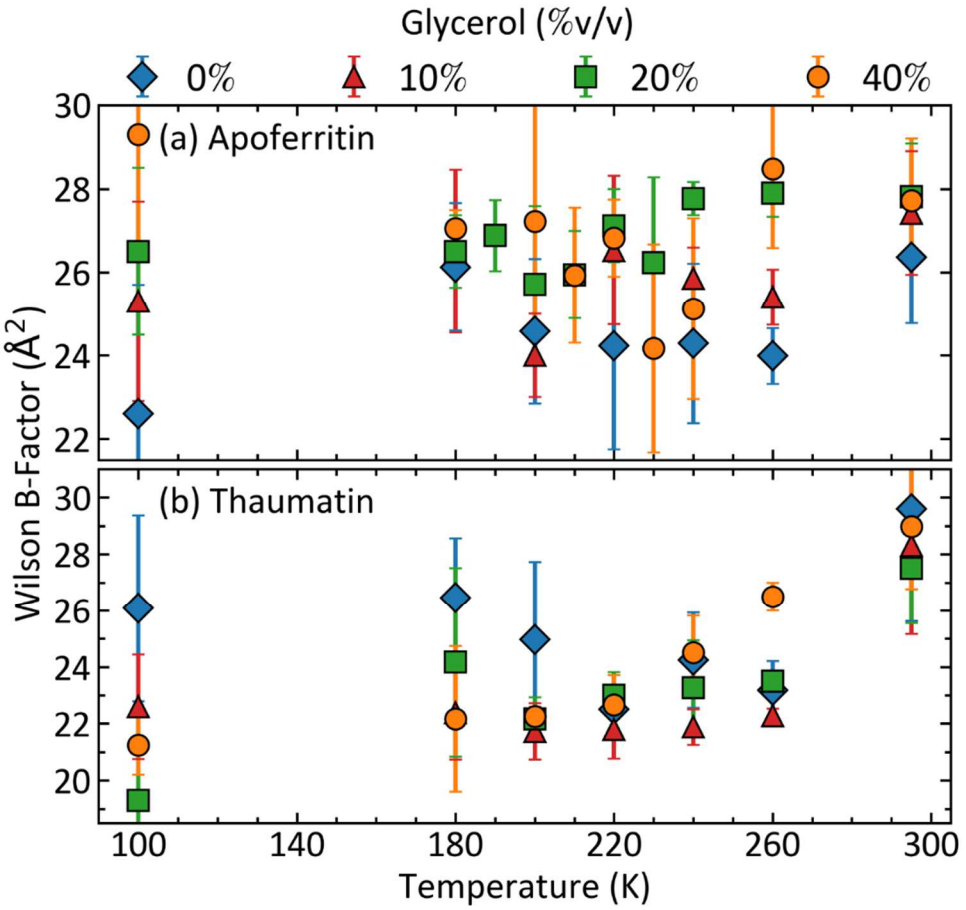


Figure 6.2.5 Wilson B factors for (a) apoferritin crystals and (b) thaumatin crystals that cooled without ice formation, versus temperature. For both proteins, cooling to T=100 K produces no clear improvement in B factors relative to 220 K. Significant scatter in values arises from crystal-size-related variations in illuminated volume and background scatter.

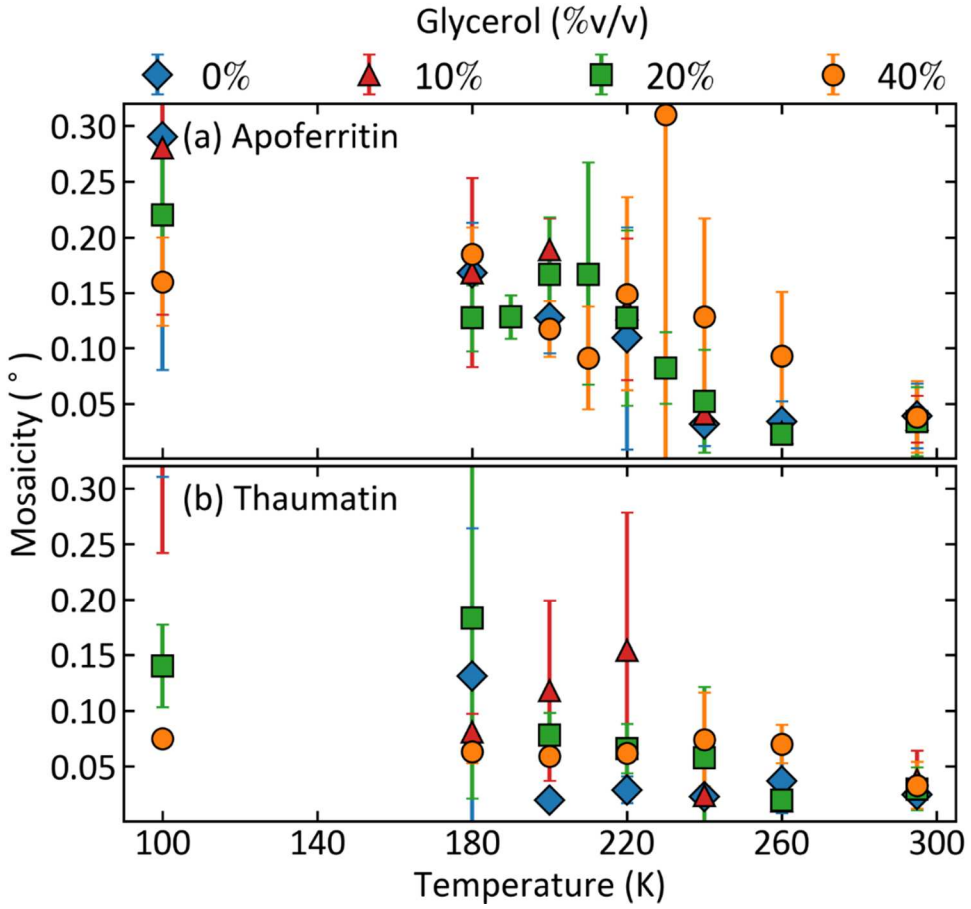


Figure 6.2.6 Minimum mosaicities as determined by XDS during the time following cooling that crystals remained ice-free, for (a) apoferritin and (b) thaumatin crystals versus temperature and glycerol concentration. No glycerol-free apoferritin crystals cooled without ice formation to 100 K. XDS-determined mosaicities floored (e.g., for room temperature crystals) at 0.045°. HKL2000-determined mosaicities were roughly 1.4 times larger than XDS mosaicities, but floored at 0.018°, comparable to the incident X-ray beam divergence. Actual room-temperature crystal mosaicities are likely ~0.01° or less for crystals of both proteins.

6.2.4 Protein structure modelling and refinement

Data frames were indexed and scaled using *XDS*, and molecular replacement and model refinement were performed using *PHENIX* (Adams *et al.*, 2010). PDB entries 3F32 and 3ZEJ were used as initial molecular replacement models for apoferritin and thaumatin, respectively. An initial refinement cycle using phenix.refine was performed, with simulated annealing, rigid body, real-space, xyz coordinates, and individual B-factor options. In *Coot* (Emsley *et al.*, 2010), these models were checked for large peaks in the difference maps, Ramachandran outliers, rotamer outliers, and regions of poor geometry. For apoferritin, cadmium ions and residues at a partially disordered loop, Gly155 and Ser156, were added into the model. For thaumatin, a well-ordered tartrate molecule was added. A second round of refinement was performed to identify and add ordered solvent. These models were checked residue by residue for errors and alternative conformations added. A third round of refinement including occupancies and B factors for alternative conformers and target weights was performed. Final model validation was performed using *MolProbity* in the *PHENIX* software package (Chen *et al.*, 2010). Structures with less than 98% completeness and resolutions worse than 2.1Å and 1.8 Å for apoferritin and thaumatin were excluded from analysis. Most structures were determined using a single crystal, but when that was not possible, data from several crystals having the same glycerol concentration and data collection temperature was merged using *XSCALE*. Refinement statistics for 45 apoferritin data sets and 53 thaumatin data sets are given in the Supporting Information.

6.2.5 Protein and solvent volume calculations

Protein volumes within the unit cell $v_{protein}(T)$ were calculated as the volume enclosed by the solvent excluded surface (SES), using a custom "ball rolling" algorithm very similar to that implemented by the program 3vee (Voss & Gerstein, 2010). In 3vee, the SES is calculated by “rolling” a fixed-diameter probe over the surface of a voxelized 3D model of the protein with atoms assigned Van der Waals radii according

to Ref. (Li & Nussinov, 1998). Our program was designed to also find the volume for multiple copies of the protein extended to fill the unit cell, accounting for crystal contacts and periodicity. Additionally, it uses probe radii of 1.4 Å and 1.7 Å for polar and apolar atoms, respectively (Li & Nussinov, 1998). Grid sizes were 0.15 Å for apoferritin and 0.10 Å for thaumatin and lysozyme.

Solvent cavity volumes $v_{cavity}(T)$ were determined by subtracting the protein volumes $v_p(T)$ from the unit cell volumes $v_{cell}(T)$. At room temperature, all crystals were highly ordered with very small mosaicities. Consequently, the solvent volume within a crystal could be assumed to be fully contained within the solvent cavities, and the crystal's solvent volume fraction was equal to the volume fraction of the unit cell occupied by cavities,

$$f_{solvent}(300\ \mathrm{K}) = \frac{v_{cavity}(300\ \mathrm{K})}{v_{cell}(300\ \mathrm{K})}\ . \qquad (6.1.1)$$

At temperatures of 240 K and below, crystals generally had larger mosaicities indicating significant lattice-scale disorder. The total solvent volume within a crystal (per unit cell) $v_{solvent}(T)$ was assumed to be equal to the sum of the solvent cavity volume $v_{cavity}(T)$ (determined from crystallographic analysis as above) and the volume of solvent that exited the unit cell during cooling $v_{exit}(T)$. The exiting solvent is assumed to accumulate in disordered crystal regions that did not contribute to ordered Bragg diffraction.

As a first approximation, the total solvent volume at temperature *T* can be obtained from the room temperature solvent volume by scaling this volume by the thermal expansion of bulk solvent having the same composition,

$$v_{solvent}(T) = v_{solvent}(300\ \mathrm{K})\left(1 + \Delta_{solvent,bulk}(T)\right), \qquad (6.1.2)$$

where $\Delta_{solvent,bulk}(T)$ gives the solvent volume expansion or contraction relative to room temperature. The volume fraction of solvent that exits the unit cells is then

$$f_{exit}(T) = \frac{v_{solvent}(T) - v_{cavity}(T)}{v_{solvent}(T)}. \quad (6.1.3)$$

The thermal expansions of water and aqueous glycerol solutions can be obtained from density measurements (Hare & Sorensen, 1987; Loerting *et al.*, 2011; Shen *et al.*, 2016), and by extrapolations based on measurements and MD simulations (Holten & Anisimov, 2012; Jahn *et al.*, 2016) at temperatures where measurements are not available. The effects of salts at concentrations present within crystal solvent on expansion behaviour should be modest.

However, this "bulk" solvent approximation assumes that the internal solvent has the same expansion behaviour and composition as the crystallization or soaking solution. Disruption of water's hydrogen bonding network in the immediate vicinity of the protein surface increases its density there (Svergun *et al.*, 1998; Merzel & Smith, 2005; Kuffel & Zielkiewicz, 2012; Barbosa & Barbosa, 2015; Persson *et al.*, 2018); the increase from bulk density is estimated to be 6% and 1%, respectively, for the first and second hydration shells (Persson *et al.*, 2018). Glycerol concentrations in solvent cavities are likely smaller due to preferential hydration of protein surfaces (Timasheff, 2002; Parsegian *et al.*, 2000; Shimizu & Smith, 2004; Sinibaldi *et al.*, 2007; Datta *et al.*, 2001; Saraswathi *et al.*, 2002; Charron *et al.*, 2002). To give a rough estimate of uncertainties due errors in solvent density, the total solvent volume at temperature *T* was estimated assuming that solvent in the first hydration shell remained at a constant density on cooling, and that the remaining solvent had bulk-like composition and thermal expansion behaviour. Defining $f_{1^{st}\ \mathrm{shell}}$ as the fraction of solvent in the first hydration shell, this "interface-perturbed" solvent volume estimate,

$$v_{solvent}(T) = \left(1 - f_{1^{st}\,shell}\right) \cdot v_{solvent}(300\ \mathrm{K})\left(1 + \Delta_{solvent,bulk}(T)\right) + f_{1^{st}\,shell} \cdot v_{solvent}(300\ \mathrm{K}), \quad (6.2.4)$$

was then used in Eq. 6.2.3 to estimate the fraction of solvent that exited the crystal.

Table 3.1 lists the room temperature solvent volume fraction $f_{solvent}(300\ \mathrm{K})$, the fractional solvent cavity volume change relative to room temperature at temperature *T*, and "bulk" and "interface-

perturbed" estimates of the fractional solvent volume change and exiting solvent fraction $f_{exit}(T)$ for apoferritin and thaumatin crystals.

6.2.6 Determining ice diffraction intensities

Diffraction images were loaded into python using the package *FabIO* (Knudsen *et al.*, 2013), which allowed the data to be handled as a numpy array, and were azimuthally averaged using the *pyFAI* integration package (Ashiotis *et al.*, 2015). Ice formation usually caused major loss of protein diffraction resolution, so protein lattice Bragg diffraction at the 2θ positions of the ice rings was generally negligible. When Bragg peaks persisted, they were removed using the *separate* function of *pyFAI*'s azimuthal Integrator module. *pyFAI* reduced the 2-dimensional Bragg-peak-masked diffraction data from the detector (counts at pixel coordinates *x*, *y*) to 1-dimensional diffraction data (average counts in a bin of angular width $\Delta(2\theta)$) versus angle 2θ or resolution $d = \lambda / 2\sin(\theta)$, correcting for beam polarization (90%, in the horizontal direction) and for variations in solid angle recorded by each pixel with 2θ. Accurate powder diffraction analysis relies on the sample being a 'perfect powder', which should generate diffraction rings that are smooth and symmetric about the beam polarization direction. Raw diffraction images were carefully inspected, and ice ring diffraction intensity in each ring versus azimuthal angle ϕ (determined as the angle between the *x* and *y* detector coordinates relative to the direct beam position) was plotted in 1° increments. Only samples for which these plots were uniform and symmetric in ϕ were used in quantitative analysis and fitting.

The non-Bragg X-ray background had contributions from the sample — NVH oil scattering, protein diffuse scattering, scattering from unfrozen water or low-density amorphous ice — and from the experimental set-up (including air scattering). The background was determined by evaluating a 10th order polynomial fit to the difference $I_{bg}(q) = I_{\exp}(q) - I_{DIFFaX}(q)$ between the *DIFFaX* models described below and our experimental diffraction patterns (similar to the method used in the powder diffraction analysis

program *GSAS-II* (Toby & Von Dreele, 2013)), and then optimizing until the *DIFFaX* model and background converged. The resulting background fits were slowly varying curves (on the 2θ or resolution scale of the ice diffraction peaks) with a consistent shape between samples that captured the diffracted intensities from water and NVH oil.

6.2.7 Modeling ice diffraction

Observed azimuthally integrated 1D ice diffraction patterns versus 2θ or resolution d for apoferritin and thaumatin crystals varied with temperature and glycerol concentration. The ice diffraction patterns were of four basic types: (a) azimuthally uniform and broad diffraction consistent with low-density amorphous ice I_{LDA}; (b) azimuthally lumpy, radially very narrow diffraction with intensities roughly consistent with I_h; (c) azimuthally lumpy diffraction with an azimuthally uniform component, the latter component consisting of a mix of narrow and broad peaks; and (d) azimuthally uniform patterns showing a mix of narrow peaks and broad, asymmetric peaks characteristic of stacking disordered ice I_{sd} (a disordered stacking of (001) planes of I_h and (111) planes of I_c). Of these, patterns of type (d) were by far the most common at temperatures between 180 K and 240 K.

Azimuthally uniform ice diffraction patterns were analyzed using the program *DIFFaX* (Treacy *et al.*, 1991), which models diffraction from crystals containing one or more phases that may be separated by coherent planar defects including twins and stacking faults. An input file for stacking disorder was obtained from the supplementary material of Malkin *et al*. (Malkin *et al.*, 2012). The model (Fig. 3.7(a)) consists of (0001) planes of hexagonal ice randomly stacked with (111) planes of cubic ice. The probability of a cubic plane being followed by hexagonal plane is Φ_{ch} and of a hexagonal plane being followed by a cubic plane is Φ_{hc} (Kuhs *et al.*, 2012; Malkin *et al.*, 2015).

To optimize the fits to the diffraction data, *DIFFaX* fitting and background subtraction were embedded in an optimization routine using a bounded, limited-memory BFGS algorithm implemented utilizing the

scipy.optimize.minimize program in the *SciPy* python library (Oliphant, 2007). The results of this simulation were used to determine the ice crystal's unit cell parameters, stacking probabilities, instrumental broadening, and structure factors. To verify the accuracy of the structure factors, they were also calculated using the explicit equations previously reported (Hansen *et al.*, 2008*b*). As shown in Fig. 9 of (Malkin *et al.*, 2015), calculated ice diffraction profiles show strong variation with the stacking parameters Φ_{ch} and Φ_{hc}, so that values obtained by fit optimization are robust. A single *B*-factor was used for the oxygens and hydrogens in the *DIFFaX* model and was fixed to 1.5, which corresponds to $\langle u^2 \rangle = 0.019$ Å^2. The instrumental broadening was assumed to be purely Lorentzian and parameters *u*, *v*, *w* were optimized to estimate the FWHM broadening as a function of angle.

Fig. 3.6 (e,f) shows example DIFFaX fits to ice diffraction data from apoferritin and thaumatin crystals. Fig. 3.7 (b),(c) shows the cubicity parameters – the fraction of all planes that are cubic – deduced from these fits versus temperature and glycerol concentration. Table 6.2.1 lists of the number of crystals whose ice diffraction data was fit for each condition. The diffraction patterns and cubicity parameters were highly consistent between crystals at the same temperature and with the same glycerol concentration.

	Apoferritin			Thaumatin	
T	Glycerol Concentration (v/v)				
	0%	10%	20%	0%	10%
100 K	17			7	
180 K	7	6	4	2	3
200 K	5	6	10	4	2
220 K	2	5	12	1	3

Table 6.2.1 Number of apoferritin and thaumatin crystals that developed internal ice following cooling, and whose ice diffraction patterns were fit at each temperature and glycerol concentration using DIFFAX to obtain the cubicity parameter characterizing stacking disorder.

In order to quantitatively estimate ice volume fractions within protein crystals, we required quantitative values of ice diffraction peak intensities. Radial integration of ice peaks and background subtraction to determine these intensities was complicated due to overlap of broadened and unbroadened peaks from I_{sd}. Diffraction peaks with I_h Miller indices such that $(h-k)/3$ is an integer (Malkin *et al.*, 2012), i.e., the (002), (110), and (112) peaks, are not broadened by stacking disorder, and the other peaks are greatly broadened. The (002) peak is overlapped by the broadened (100) and (101) peaks, the (112) peak is overlapped by the broadened (200) and (201) peaks, while the (110) peak is not overlapped.

To determine the intensities of unbroadened peaks, ice diffraction data were first modelled using *DIFFaX* without including instrumental broadening. Peaks that were not affected by stacking disorder then appeared as sharp spikes (delta functions), and were eliminated from the model's diffraction by interpolating from either side through each peak, leaving only the peaks broadened by stacking disorder. Instrumental broadening was then applied to this model pattern. The sum of the final background and modelled diffraction from the stacking disorder broadened peaks was then subtracted from the measured pattern to determine the intensities of the unbroadened peaks.

6.2.8 Estimating ice fractions in protein crystals

Canonical diffraction equations were used to determine the volume fraction of solvent within each crystal that contributes to the crystalline ice diffraction. The measured diffraction intensities from an initially ice-free protein crystal at room temperature and from the ice that forms within the crystal during and/or after cooling are determined, and relative diffracting volumes of the protein crystal and of ice within the crystal are estimated using the corresponding structure factors for the protein crystal and for the ice, with corrections for unit cell and solvent contractions and solvent expulsion from the unit cell on cooling.

In diffraction data collection from an initially ice-free protein crystal and from the same crystal after ice formation, the X-ray illuminated volume V is the same. The fraction of the illuminated volume that is

filled with ice is $f_{ice} = V_{ice} / V$. The fraction of the illuminated volume filled by solvent $f_s = V_s / V$ can be determined by crystallographic analysis of the unit cell. The crystallized solvent fraction is then given by $f_c = f_{ice} / f_s$. Corrections are needed if the ice-free data and solvent volume are determined at room temperature rather than the temperature at which ice forms.

Quantifying protein crystal diffraction. A single protein crystal of length $L_{pc}\,(\mathrm{m})$ along the incident beam path, with a unit cell volume $v_{pc}\,(\mathrm{m}^3)$, that is being rotated at an angular velocity $\omega(\mathrm{rad/s})$ while being illuminated by an X-ray beam with flux $\Phi(\mathrm{photons/s})$ and wavelength $\lambda\,(\mathrm{m})$ produces Bragg peaks whose integrated intensities recorded by the detector $I_{pc,hkl}$ (photons/s) at angles $2\theta_{pc,hkl}$ are given by (Warren, 1990)

$$I_{pc,hkl} = \frac{1}{\omega}\Phi r_e^2 \lambda^3 \frac{L_{pc}}{v_{pc}^2}\left|F_{pc,hkl}\right|^2 \cdot \exp\left[-2B_{pc}\left(\frac{\sin\theta_{pc,hkl}}{\lambda}\right)^2\right] \cdot LP_{pc,hkl} \cdot A^{pc} \cdot A^{air}_{pc,hkl} \cdot A^{detector}_{pc,hkl} \quad . \quad (6.2.5)$$

Here, $r_e = 2.82\times10^{-15}$ m is the classical electron radius, $\left|F_{p,hkl}\right|$ (electron equivalents) are the crystal's structure factors (including the effects of individual atomic *B* factors), and B_{pc} is a scaling *B*-factor. $LP_{p,hkl}$ is the dimensionless Lorentz-Polarization factor,

$$LP_{pc,hkl} = \frac{1+\alpha\cos^2(2\theta_{pc,hkl})}{(1+\alpha)\sin(2\theta_{pc,hkl})}\ , \quad (6.2.6)$$

where $\alpha = \cos^2(2\theta_m)$ is the Bragg angle of the monochromator crystal from Sec. 6.2.2. A^{pc} and $A^{air}_{pc,hkl}$ are the fractional attenuations of the x-ray intensities due to absorption and non-Bragg scattering by the crystal and the air, and respectively. The air absorption model is

$$A^{air}_{pc,hkl} = \exp\left[-\frac{D\mu_{air}\rho_{air}}{\cos 2\theta_{pc,hkl}}\right]. \quad (6.2.7)$$

Here *D* is the sample-to-detector distance, μ is the x-ray absorption constant at the incident X-ray energy, and ρ is the air density. The product $\mu\rho$ is estimated by *XDS* (whose values agree with those in NIST tables) and is listed in the CORRECT.LP file.

X-ray absorption by the sample is due to absorption by the crystal and the NVH oil. The absorption by the protein crystal is

$$A^{pc}(T) = \exp\left[-\mu_{pc}\rho_{pc}(T)L_{pc}(T)\right]. \qquad (6.1.8)$$

The mass absorption coefficient can be assumed constant, and the product of the crystal density and path length varies with cell volume as $\rho L \propto v^{-2/3}$. X-ray absorption increases as the sample contracts because more material is pulled into the beam path. However, assuming a protein crystal density of ~1.3 g/cm^3, a mass absorption coefficient of ~1.2 cm^3/g, and a crystal size of ~400 μm, the 2.5% reduction in apoferritin cell volume observed on cooling to 180 K increases absorption by only ~0.1%, a result that is only slightly modified when absorption by the NVH oil is included.

The Pilatus 6M detector measures photons that are absorbed in a $h = 320\,\mu\text{m}$ thick silicon layer. At $2\theta = 0°$, 75% of the incident photons are absorbed, and at higher diffraction angles the path length through the silicon and the absorption increase. Since only absorbed photons are detected and counted, the measured intensities at small angles are reduced relative to those at large angles. This effect is accounted for using the factor

$$A^{detector}_{pc,hkl} = 1 - \exp\left[-\frac{h\mu_{Si}\rho_{Si}}{\cos 2\theta_{pc,hkl}}\right]. \qquad (6.2.9)$$

The diffraction equation (6.2.5) can be rewritten as

$$\ln\left[\omega\frac{I_{pc,hkl}}{LP_{pc,hkl}\cdot A^{air}_{pc,hkl}\cdot A^{detector}_{pc,hkl}\left|F_{pc,hkl}\right|^2}\right] = \ln\left[\Phi r_e^2\lambda^3\frac{L_{pc}}{v_{pc}^2}A^{pc}\right] - 2B_{pc}\left(\frac{\sin\theta_{pc,hkl}}{\lambda}\right)^2. \qquad (6.2.10)$$

The slope of a linear fit to a plot of the left side of Eq. 6.2.10 versus $2(\sin\theta/\lambda)^2$ can then be used to determine the scaling *B* factor, and the exponential of the *y* intercept gives a constant

$$C_{pc} = \Phi r_e^2 \lambda^3 \frac{L_{pc}}{v_{pc}^2} A^{pc}, \qquad (6.2.11)$$

determined by the incident flux and the illuminated protein crystal length along the beam path L_{pc}.

In the present experiments, the 10 frames covering 5° of diffraction data collected from each crystal at room temperature were processed using *XDS*, correcting for the Lorentz polarization factor but with other intensity corrections and scalings turned off (by including 'CORRECTIONS=!' in the XDS.INP file) to obtain absolute intensities in the XDS_ASCII.HKL file. The air and detector absorption corrections were applied manually, since the air absorption correction in *XDS* is relative to rather than absolute and since the formula used for detector absorption correction was not clear, to obtain corrected intensity values $I_{pc,hkl}^{obs}$. Using the default *XDS* intensity corrections and scalings changed final ice volume fractions by less than 1%.

The 5° angular range of the room temperature data was not sufficient for structure determination. To determine the structure factors required to normalize the observed intensities and calculate C_{pc}, we used complete room temperature data sets obtained from crystals that were identically prepared (with identical glycerol concentrations) to those in which ice formed, and processed this data and refined models as described in Section S8 above. Hydrogens were added to the starting PDB model using *phenix.reduce*, and initial structure factors were determined from this model using *phenix.fmodel*. For the flat bulk solvent model assumed in *phenix.fmodel*, the k_{sol} parameter was chosen to be the estimated electron density of the crystal soak solutions, and b_{sol} was set to 50 (Fokine & Urzhumtsev, 2002). To obtain the most accurate structure factor estimates, the structure factors were determined directly using the wk1995 tables (Waasmaier & Kirfel, 1995) rather than using an fft method. The resulting structure

factors from *phenix.fmodel* were converted from an .mtz file to a human-readable .csv file that could be compared to the measured intensities from XDS_ASCII.HKL using *phenix.mtz.dump*.

Plots of Eq. 6.2.10 were then applied to the region between 10 Å, 5 Å, and 4 Å for apoferritin, thaumatin, and lysozyme, respectively, and the resolution where $I / \sigma = 2$ to determine a scaling *B* factor and the constant C_{pc}. Beyond the low-resolution cut-off for each protein, the ratio of measured intensities to structure factors tended to become nonlinear.

To assess the sensitivity of the values of C_{pc} to structural model parameters, several different PDB files (3 apoferritin, 3 thaumatin, and 4 lysozyme) were used, and the resulting variations in calculated C_{pc} values were 2%, 1.5% and 1.2% for apoferritin, thaumatin, and lysozyme, respectively. Variation of the bulk solvent parameters had little effect on C_{pc}.

Finally, to compare ice and protein crystal diffraction for cooled crystals, the room temperature value of C_{pc} must be scaled from room temperature to the temperature at which ice is observed. This scaling was calculated as

$$\begin{aligned} \frac{C_{pc}(T)}{C_{pc}(300\ \mathrm{K})} &= \frac{L_{pc}(T)}{L_{pc}(300\ \mathrm{K})} \frac{v_{pc}^2(300\ \mathrm{K})}{v_{pc}^2(T)} \frac{A^{pc}(T)}{A^{pc}(300\ \mathrm{K})} \\ &\approx 1 + \frac{5}{3} \frac{\Delta v_{pc}}{v_{pc}} \end{aligned}, \qquad (6.2.12)$$

where $\frac{V_{ice}}{v_{pc}} = \frac{L_{ice}}{L_{pc}} = \frac{C_{ice}}{C_{pc}} \frac{v_{ice}^2}{v_{pc}^2}$ $v_{exit}(T) = f_{exit}(T)\, v_{solvent}(T)$ $\Delta v_{pc} / v_{pc}$ is the fractional change in unit cell volume on cooling from room temperature to temperature *T*, and is positive if the unit cell contracts. For the temperatures and protein crystals studied here, the scale factor Eq. 6.2.12 is of order 1.05, again a small correction.

Quantifying ice diffraction. For a powder sample of ice of length L_{ice} (m) that is exposed for a time *t* (s), the azimuthally averaged and radially integrated diffracted intensity recorded in a diffraction ring at a given angle $2\theta_{ice,hkl}$ is (Warren, 1990)

$$I_{ice,hkl} = \frac{t}{4}\Phi r_e^2 \lambda^3 \frac{L_{ice}}{v_{ice}^2} m_{ice,hkl} \left|F_{ice,hkl}\right|^2 \times \exp\left[-2B_{ice}\left(\frac{\sin\theta_{ice,hkl}}{\lambda}\right)^2\right] \cdot LP_{ice,hkl} \cdot A^{pc} \cdot A^{air}_{ice,hkl} \cdot A^{detector}_{ice,hkl}, \quad (6.2.13)$$

where $m_{ice,hkl}$ is the multiplicity of the diffraction ring, determined by the number of *hkl* values that produce diffraction at the same 2θ, and all other parameters have the same meaning as for a single crystal in Eq. 6.2.5. For an area detector with a sample-detector distance *R* and pixel size δ, the azimuthally integrated intensity per unit length of a pixel is

$$I_{ice,hkl,pixel} = \frac{t\delta}{8\pi R \sin\left(2\theta_{ice,hkl}\right)}\Phi r_e^2 \lambda^3 \frac{L_{ice}}{v_{ice}^2} m_{hkl} \left|F_{ice,hkl}\right|^2 \times \exp\left[-2B_{ice}\left(\frac{\sin\theta_{ice,hkl}}{\lambda}\right)^2\right] \cdot LP_{ice,hkl} A^{pc} \cdot A^{air}_{ice,hkl} \cdot A^{detector}_{ice,hkl}. \quad (6.2.14)$$

This assumes the detector pixels are on a sphere and subtend a fixed solid angle, which corresponds to our data after *pyFAI* integration and correcting for solid angle. For each ice ring (*hkl* value), Eq. 6.2.14 can be rewritten as

$$C_{ice} = \Phi r_e^2 \lambda^3 \frac{L_{ice}}{v_{ice}^2} A^{cryst} = \frac{8\pi R \sin\left(2\theta_{ice,hkl}\right)}{t\delta} \frac{I_{hkl}}{LP_{ice,hkl} \cdot A^{air}_{ice,hkl} \cdot A^{detector}_{ice,hkl} \cdot m_{ice,hkl} \cdot \left|F_{ice,hkl}\right|^2} \times \exp\left[2B_{ice}\left(\frac{\sin\theta_{ice,hkl}}{\lambda}\right)^2\right]. \quad (6.2.15)$$

Values of C_{ice} from the three diffraction rings that were not broadened by stacking disorder (the hexagonal (002), (110), and (112) rings) were averaged to obtain a final estimate for C_{ice}.

Quantifying the crystallizable fraction of crystal solvent. The ratio of the ice and protein crystal constants C_{ice} / C_{pc} at the same temperature *T* determines the ratio of the volume of ice to the volume of ordered unit cells within the crystal, or equivalently, the volume of ice per unit cell,

$$\frac{V_{ice}}{v_{pc}} = \frac{L_{ice}}{L_{pc}} = \frac{C_{ice}}{C_{pc}} \frac{v_{ice}^2}{v_{pc}^2} \ . \tag{6.2.16}$$

Here lower-case *v* denotes volumes directly associated with the unit cell, determined by crystallography, and an upper case *V* denotes a volume per unit cell at temperature *T,* for a quantity that can be present both inside and outside the unit cell.

Studies of water at interfaces and of nanoconfined water indicate that there is a layer of interfacial water that is strongly perturbed, and this perturbation should inhibit crystallization. To determine the fraction of a protein crystal's solvent that forms ice, the fraction that cannot be crystallized, and the thickness of this interfacial layer requires careful accounting of both the ice and solvent, both of which can be located within unit cells and within disordered regions of the crystal where solvent exiting the unit cell during cooling accumulates.

The primary quantities of interest are the fraction of the solvent cavity volume within the protein crystal's unit cells that is occupied by ice,

$$f_{cavity,ice} = \frac{V_{ice}^{uc}}{v_{cavity}} , \tag{6.2.17}$$

and the fraction

$$f_{cavity,solvent\,(\ell)} = 1 - f_{cavity,ice} = \frac{V_{solvent\,(\ell)}^{uc}}{v_{cavity}} \tag{6.2.18}$$

that is occupied by uncrystallized liquid solvent. The total crystal volume per unit cell is the sum of the volumes of ice, liquid solvent, and protein.

An upper bound on the fraction of the unit cell's solvent cavities occupied by ice can be obtained by assuming that no solvent exits the unit cells on cooling, so that all ice and solvent remain confined within ordered unit cells. In that case, $V_{ice} = V_{ice}^{uc}$, $v_{cavity} = V_{ice} + V_{solvent\,(\ell)}$ so that

$$f_{cavity,ice} = \frac{V_{pc}}{v_{pc}} \left(\frac{1}{1 - v_{protein} / v_{pc}} \right), \tag{6.2.19}$$

where the first term is obtained from Eq. 6.2.17 and the second from crystallographic analysis of the unit cell. This is estimate (a) in Table 3.1.

During cooling, solvent is squeezed out of the unit cell and accumulates at defects and disordered crystal regions, forming pools that are likely larger than the solvent cavities within the unit cell. A second, lower bound on the fraction of the unit cell's solvent cavities occupied by ice can be obtained by assuming that all solvent that exits the unit cell forms ice, and that all liquid solvent is found within the ordered unit cells. In this case, $V_{ice} = V_{ice}^{uc} + V_{ice}^{exit}$, so Eq. 6.2.19 becomes

$$\begin{aligned} f_{cavity,ice} &= \frac{V_{ice}^{uc}}{v_{cavity}} = \frac{V_{ice} - V_{ice}^{exit}}{v_{cavity}} \\ &= \frac{V_{ice}}{v_{pc}} \left(\frac{1}{1 - v_{protein} / v_{pc}} \right) \left(1 - \frac{V_{ice}^{exit} / v_{pc}}{V_{ice} / v_{pc}} \right) \end{aligned} . \tag{6.2.20}$$

V_{ice}^{exit} can be approximated by the solvent volume that exits the unit cell, $v_{exit}(T) = f_{exit}(T)\, v_{solvent}(T)$, estimated as described in SI Sec. 6.2.5.

Tables 3.2 and 6.2.2 give estimates of the volume fraction of solvent cavities occupied by ice for apoferritin, thaumatin, and lysozyme crystals, based on Eq. 6.2.19 (estimate (a)), and on Eq. 6.2.20 assuming bulk-like or interface-perturbed solvent contraction (estimates (b) and (c), respectively) as discussed in SI Sec. 6.2.5 and given in Table 3.1.

Ice has a larger volume than the liquid solvent that forms it, but when using the density of supercooled liquid at temperature T this error is at most a few percent. Ice formed in exited solvent, within disordered crystal regions, can draw liquid solvent from within the ordered unit cells for its growth, leading to shrinkage of the unit cell, but such shrinkage was not observed here; only a few of the crystals that formed surface ice exhibited cell shrinkage with time.

Fraction of solvent exiting the unit cell	Fraction of internal solvent that forms ice					
	Apoferritin			Thaumatin		Lysozyme
	0% glycerol	10% glycerol	20% glycerol	0% glycerol	10% glycerol	0% glycerol
(a) $f_{exit} = 0$	59% ± 13%	63% ± 18%	35% ± 9%	35% ± 6%	19% ± 7%	17% ± 5%
(b) f_{exit}, bulk	49% ± 13%	58% ± 18%	33% ± 9%	25% ± 6%	12% ± 8%	8% ± 6%
(c) f_{exit}, perturbed	50% ± 13%	58% ± 18%	33% ± 9%	29% ± 6%	13% ± 8%	12% ± 6%

Table 6.2.2 Estimates of the maximum fraction of the solvent cavity space occupied by ice in apoferritin and thaumatin crystals at 180<T<220 K. Estimate (a) assumes all solvent is confined to unit cells so all ice forms inside unit cells. Estimate (b) assumes that all solvent has bulk thermal expansion behavior so that solvent exits the unit cells, that all exiting solvent forms ice, and that the volume of ice is the same as the volume of supercooled solvent from which it forms. Estimate (c) instead assumes the solvent density in the first hydration layer is that of bulk, room-temperature solvent, and the remaining solvent has bulk thermal expansion behavior. No systematic variation in ice fractions with temperature between 180 K and 220 K was observed, so values represent averages over all crystals.

6.2.9 Analysis of protein structures from the Protein Data Bank

Unit cell sizes, solvent contents, resolution, and solvent channel sizes were determined from Protein Data Bank entries for 17,148 non-redundant protein structures, excluding small peptides and viral proteins. 96.8% of entries were determined at temperatures below 180 K; 1.1% were determined at temperatures between 180 and 260 K, and 2.1% were determined at temperatures above 260 K (as indicated by the PDB header files, which can be inaccurate.) Unit cell size, resolution, and solvent content (calculated using the Matthews coefficient (Matthews, 1974; Kantardjieff & Rupp, 2003)) were extracted from the PDB header file. The maximum solvent cavity size was determined using the PDB model coordinates and the program *map_channels* (Juers & Ruffin, 2014) as the diameter of the largest sphere that could be entirely contained within the unit cell's solvent space (whose surface was determined by the Van der Waal's radii of the protein atoms.) The radius of this sphere determines the maximum distance of solvent molecules from the protein surface, the relevant metric for solvent perturbations by that surface. The results of these calculations are shown in Figs. 3.1, 6.2.7, and 6.2.8.

For apoferritin, thaumatin, and lysozyme crystals at room temperature, *map_channels* was also used to determine the distribution of solvent distances from the protein surface. The *nc.pdb output file reports the distance between each grid point in a voxelized unit cell and the closest protein surface. The solvent distance distribution, shown in Fig. 6.2.2 was obtained by generating a histogram of those distance values belonging to solvent cavity voxels.

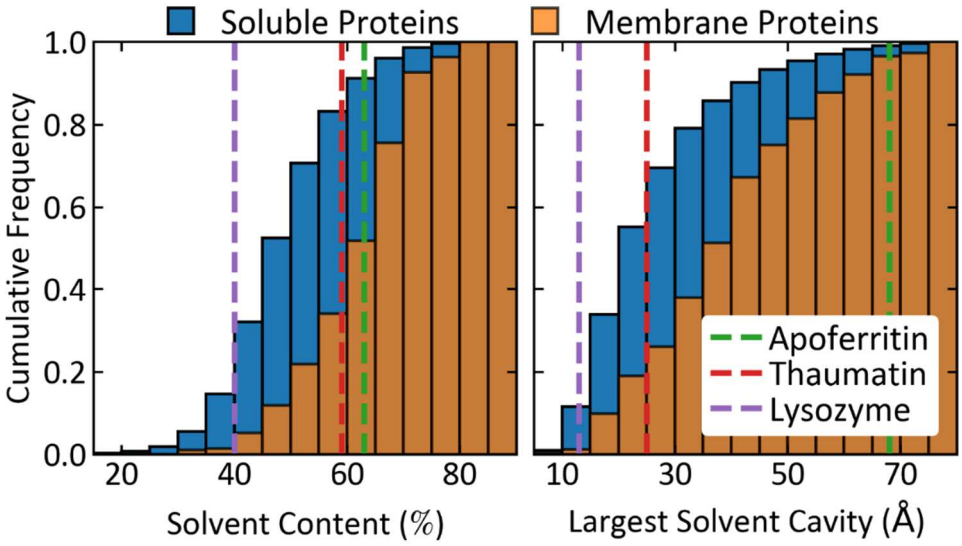


Figure 6.2.7 Histograms of the cumulative frequency of (a) the solvent content and (b) the largest solvent cavity obtained from over 17,000 non-redundant structures deposited in the protein data bank (PDB), for soluble proteins (blue) and membrane proteins (orange). The dashed vertical lines indicate the solvent content and largest solvent cavity for lysozyme (purple), thaumatin (red), and apoferritin (green).

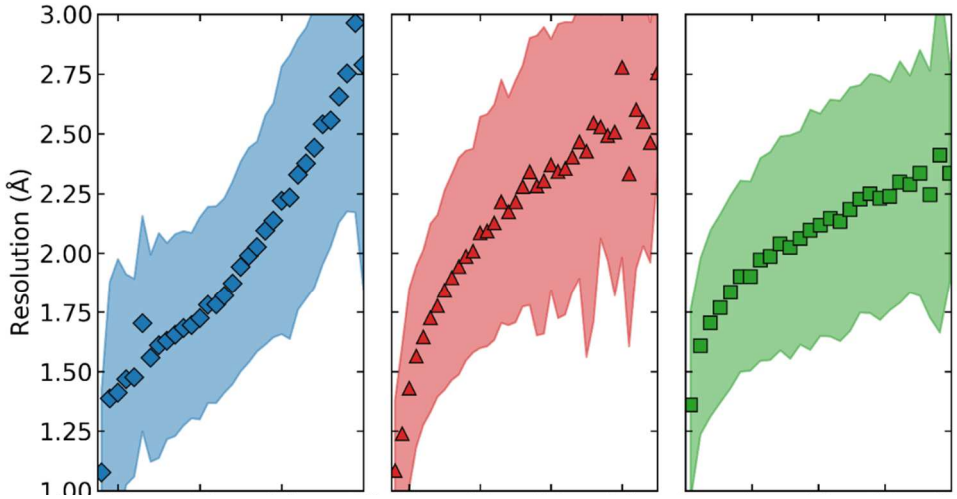


Figure 6.2.8 Distribution of diffraction resolution – which determines the highest spatial frequencies in the electron density map of the protein – versus solvent content (% v/v), largest solvent cavity, and unit cell volume for ~17,000 non-redundant cryogenic temperature protein structures obtained from the Protein Data Bank. Symbols represent the mean value over all structures in a given horizontal axis bin, and the boundary of the shaded region shows where the frequency drops to $e^{-1/2}$ of its peak value within each horizontal axis bin.

6.2.10 Suppression of ice formation in protein crystals

Figs. 3.3 and 6.2.9 show the fraction of apoferritin and thaumatin crystals, respectively, that were cooled to each temperature and remained ice free for at least (a) 3 s and (b) 20 s. Ice nucleation was enormously suppressed relative to rates observed in bulk solutions at all temperatures.

No ice was ever observed above 240 K in either protein, including for glycerol-free crystals. This suggests that 240 K is an upper bound on the freezing temperature of the confined solvent within the glycerol-free crystals (since we cannot be certain that observed ice at lower temperatures did not form in, e.g., larger solvent pockets associated with crystal defects); freezing temperatures decrease to 200-220 K for crystals soaked in 40% w/w glycerol. Fig. 3.4 compares these freezing temperatures with those of bulk glycerol solutions and of pure water confined in inorganic matrices with different confinement (pore) diameters.

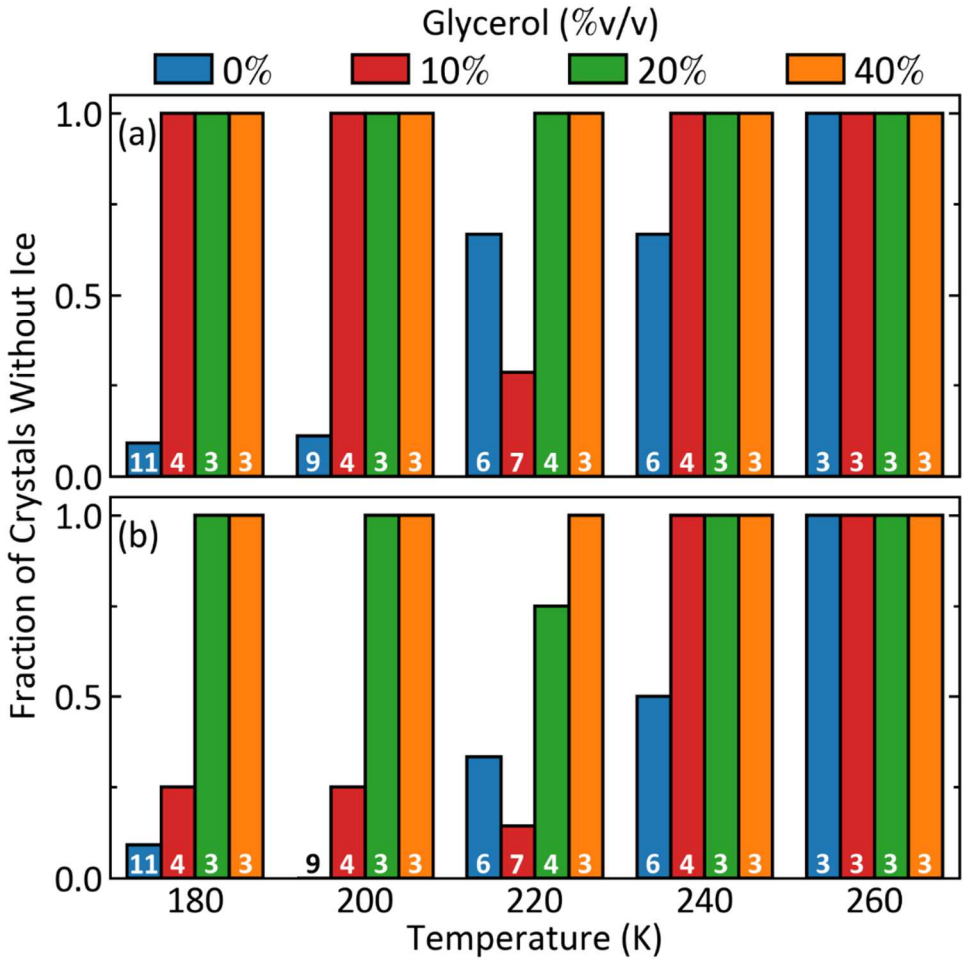


Figure 6.2.9 Fraction of thaumatin crystals that remained ice free for at least (a) 3 s and (b) 20 s following cooing to each temperature. The numbers in each bar indicate the number of crystals examined under each condition. The smaller fraction of ice-free crystals at 180-220 K relative to apoferritin may be due to differences in crystal handling or to a possible effect on nucleation of the greater connectivity of the solvent spaces within thaumatin crystals.

6.2.11 Effects of crystallization salts on ice formation

Any solute that perturbs water's structure and dynamics, including the salts used in crystallization, has cryoprotective abilities (Wang *et al.*, 2016). Apoferritin and thaumatin crystals were crystallized in solutions that, when fully equilibrated with their reservoir solutions, contained ~15% w/v (~1.1 M and 0.5 M) ammonium sulfate and sodium potassium tartrate, respectively. Freezing point suppression at these salt concentrations is at most a few degrees, compared with observed freezing point suppressions in apoferritin and thaumatin crystals of ~30 K. Similarly, measurements on NaCl solutions (Warkentin *et al.*, 2013) suggest that critical cooling rates (the minimum cooling rates required to obtain a glass phase with an ice fraction below ~1% on cooling to 77 K) for these solutions should be ~10^4 K/s. Peak ice nucleation rates occur between ~230 K and 190 K (Manka *et al.*, 2012), and on cooling at 10^4 K/s to 77 K the sample will spend ~10^{-2} s in this temperature range. In the present experiments apoferritin and thaumatin crystals spent 10^1-10^2 s at these temperatures without formation of detectable ice. Lysozyme crystals were grown in solutions containing only 2.5% w/v NaCl, and exhibited a similar or larger freezing point suppression to apoferritin and thaumatin crystals. Concentrations of salts and other solutes within crystals are generally different (typically somewhat lower) than in crystallization or soaking solutions (Vekilov *et al.*, 1996), but these differences are too small to account for our observations. Consequently, salts present in the crystals provide minimal cryoprotection and cannot explain the dramatic reduction of freezing temperatures and ice formation rates at temperatures between 180 and 260 K.

6.2.12 Connectivity of solvent cavities and ice formation

Among the protein crystals studied here, and also among protein crystals in general, cubic apoferritin is unusual in that much of its solvent is contained in apparently isolated cavities rather than in continuous channels (Fig 3.3). However, glycerol, iron, and other solutes freely exchange into these cavities at room temperature through 8 hydrophilic channels 6 Å long and 3.4 Å in diameter (Crichton & Declercq, 2010), so they are less disconnected than the static crystallographic structure in Fig. 3.2 would suggest.

Most studies of nanoconfined water have used nanoporous silica or alumina, in which water is typically confined in a regular matrix of cylindrical channels of nanometer dimensions. Even though the water in adjacent channels is physically separated (except where the matrix has defects or when water extends across the open ends of the channels), experiments suggest that a single ice nucleation event triggers ice formation throughout the matrix (Suzuki, Steinhart *et al.*, 2015). The same may be true in apoferritin, thaumatin, and lysozyme: once ice diffraction is detected, it often saturates to a steady state value in 0.2-0.4 s, except at high (40% v/v) glycerol concentrations, and this steady state value corresponds to conversion of at least a substantial fraction of the crystal's solvent to ice. However, our diffraction measurements have a minimum detectable ice fraction of order 1-2%, and so observed ice may arise from multiple nucleation events.

6.2.13 Cryo- and variable temperature crystallography using cryoprotectant-free protein crystals

For data collection below water's glass transition temperature T_g~136 K, ice formation in most cryoprotectant-free crystals can be outrun by cooling small samples sufficiently rapidly (Warkentin *et al.*, 2006; Pflugrath, 2015). For crystals with smaller unit cells and modest solvent contents (~40% v/v and lower), ice-free diffraction was frequently obtained in the early days of cryocrystallography — before use of penetrating cryoprotectants became standard — by hand-plunging in liquid nitrogen. For crystals with much larger solvent contents and solvent cavities, ice formation on cooling to 77 K can be reliably eliminated by removing external solvent (Warkentin & Thorne, 2009) and using improved plunge cooling methods that increase cooling rates with liquid nitrogen by one to two orders-of-magnitude (Warkentin *et al.*, 2006).

Far more challenging is data collection at temperatures between 180 and 260 K. Above T_g water molecules develop substantial mobility, and in bulk water or aqueous cryoprotectant solutions at concentrations below ~50-70% v/v ice always forms. Attempts at data collection in this temperature range

used crystals soaked in solutions with very large cryoprotectant concentrations (Frauenfelder *et al.*, 1979; Tilton *et al.*, 1992), or else crystals with very low solvent contents and small solvent cavities (Teeter *et al.*, 2001). More recently, careful removal of external solvent has allowed data collection in this temperature range (Warkentin & Thorne, 2010*b*,*a*; Keedy *et al.*, 2015), in some cases without penetrating cryoprotectants. In all of these experiments data collection at each temperature took at least several minutes, and required that crystals remain ice-free during this time. Consequently, they were limited to crystals with modest solvent contents and solvent cavity sizes.

The present results show that cryoprotectant-free data collection can routinely be performed throughout the 180 K-260 K temperature range even with crystals with enormous solvent cavities, by carefully removing or cryoprotecting external solvent, and using intense X-ray beams and fast detectors to record crystal diffraction before ice forms. This dramatically extends the potential scope of variable-temperature crystallographic studies, including to membrane protein crystals with large solvent contents, and to crystals of large complexes having large solvent cavities.

6.2.14 Additional relevant literature references

The present work draws on prior work performed in multiple fields. Space constraints in the main manuscript restricted the number of references we could cite, and in some cases our choice of which references to cite on a given subject was arbitrary. Here we list additional references that informed our work.

Biomolecular cryocrystallography – methods: (Rodgers, 1994; Garman & Schneider, 1997; Garman, 2003; Pflugrath, 2015; Rupp, 2009; Warkentin & Thorne, 2009)

Crystal disorder and degradation of diffraction due to cryocooling: (Kriminski *et al.*, 2002; Juers & Matthews, 2001, 2004*a*,*b*; Alcorn & Juers, 2010; Warkentin *et al.*, 2006; Warkentin & Thorne, 2009)

Temperature-dependent X-ray crystallography of proteins between room temperature and ~180 K: (Frauenfelder *et al.*, 1979, 1987; Douzou *et al.*, 1975; Tilton *et al.*, 1992; Teeter *et al.*, 2001; Warkentin & Thorne, 2010*a*; Warkentin *et al.*, 2012; Keedy *et al.*, 2014, 2015)

Conformational changes in proteins due to cryocooling: (Keedy *et al.*, 2014; Fraser *et al.*, 2011; Atakisi *et al.*, 2018; Halle, 2004; Fraser *et al.*, 2009)

The protein-solvent glass or dynamical transition: (Tilton *et al.*, 1992; Ringe & Petsko, 2003; Schirò *et al.*, 2011, 2015; Doster, 2010; Fenimore *et al.*, 2004)

Measurements of ice nucleation rates in pure water: (Huang & Bartell, 1995; Bartell & Chushak, 2003; Riechers *et al.*, 2013; Chukin *et al.*, 2010; Manka *et al.*, 2012; Murray *et al.*, 2010)

Measurements of ice formation in nanoconfined systems: (Suzuki, Steinhart *et al.*, 2015; Yao *et al.*, 2017; Mascotto *et al.*, 2017; Miyatou *et al.*, 2016; Petrov & Furó, 2011; Morishige & Nobuoka, 1997; Morishige & Kawano, 1999; Findenegg *et al.*, 2008; Jähnert *et al.*, 2008; Schreiber *et al.*, 2001; Taschin *et al.*, 2015; Liu *et al.*, 2010; Rault *et al.*, 2003; Schmidt *et al.*, 1995; Hansen *et al.*, 1996; Bartell & Chushak, 2003)

Simulations of ice formation in nanoconfined systems: (Bartell & Chushak, 2003; Moore & Molinero, 2010; Moore *et al.*, 2010; Solveyra *et al.*, 2011; Li *et al.*, 2013; Espinosa *et al.*, 2014, 2016)

Water and ice in protein crystals: (Weik, 2003; Weik *et al.*, 2005; Persson *et al.*, 2018; Nakasako, 2004)

Water structure and density near interfaces: (Erko *et al.*, 2012; Persson *et al.*, 2018; Svergun *et al.*, 1998; Merzel & Smith, 2002*a*,*b*; Bagchi, 2005; Wang *et al.*, 2016; Ebbinghaus *et al.*, 2007; Lee *et al.*, 2014)

Ice structure and stacking disorder: (Morishige & Uematsu, 2005; Hansen *et al.*, 2008*a*; Malkin *et al.*, 2012, 2015; Moore & Molinero, 2011; Amaya *et al.*, 2017; Kuhs *et al.*, 1987, 2012)

6.3 SUPPORTING INFORMATION FOR CHAPTER 4

Temperature (K)	Number of Apoferritin Crystals Examined			
	Glycerol (% *v/v*)			
	0%	10%	20%	40%
180	8	5	3	4
190	-	-	5	-
200	6	3	10	3
210	-	-	7	4
220	7	8	13	7
230	-	-	7	6
240	7	5	4	4
260	5	3	2	6
RT	61	35	58	34

Table 6.3.1 Number of cubic apoferritin crystals measured at each temperature and glycerol concentration that remained ice-free for long enough to determine the unit cell parameter, Wilson B-factor and mosaicity in a fully cooled state. The number of crystals at room temperature refers to the total number of crystals used in this analysis. Data from these crystals is used to generate Figs. 4.3(a), 4.3(b), and 4.4.

Temperature (K)	Glycerol (% *v/v*)			
	0%	10%	20%	40%
180	4	1	1	-
200	4	-	1	1
210	-	-	1	1
220	3	3	2	1
230	-	-	2	1
240	3	-	-	1
260	2	1	-	1
RT	4	3	4	1

Table 6.3.2 Total number of crystals where a structure could be determined in a fully cooled state without post-cooling expansion. Data from these crystals is used for the protein and solvent cavity volumes in Fig. 4.3 (c-f).

Temperature (K)	Number of Apoferritin Crystals Examined			
	Glycerol (% *v/v*)			
	0%	10%	20%	40%
220	3	-	2	-
230	-	-	2	1
240	6	3	3	3
260	2	2	2	4

Table 6.3.3 Total number of crystals that showed post-cooling expansion and for which the unit-cell parameter, mosaicity, and expansion time-scale was determined. Data from these crystals is used for Figs. 4.6-4.8.

Temperature	*In* Conformation		*Out* Conformation	
	Contracted	Expanded	Contracted	Expanded
180	–	–	4	–
200	–	–	4	–
220	–	–	3	3
240	–	3	3	1
260	2	2	–	–
Room-Temperature	2		2	

Table 6.3.4 Total number of glycerol free apoferritin crystals at each temperature where a structure could be determined before or after cold unit cell expansion, according to whether they showed residue GLN82 in the "*in*" or "*out*" conformation. Data from these crystals is used for Figs. 4.9-4.11 and 6.2.5.

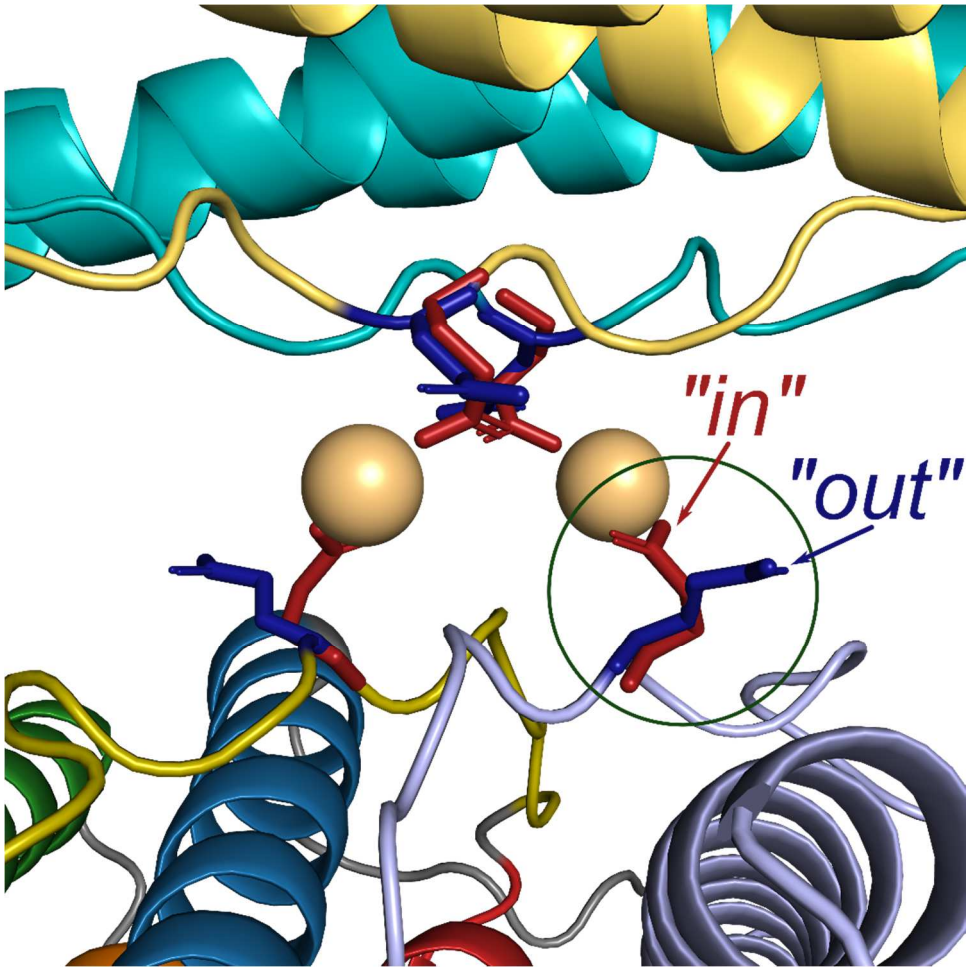


Figure 6.3.1 Interface between apoferritin shells, showing the region of crystal contact between two dimers along their BC loops (See Fig. 4.1). Residues ASP80 and GLN82 protrude from the BC loop and connect to a cadmium atom between the shells. ASP80 has well defined electron density and is always involved in the contact. GLN82 is found in either an '*in*' or '*out*' conformation in a given crystal. In the *in* conformation GLN82 is connected to the Cd atom with well-defined electron density. In the *out* conformation GLN82 protrudes outward into the solvent cavity and an additional water molecule coordinates with the Cd atom.

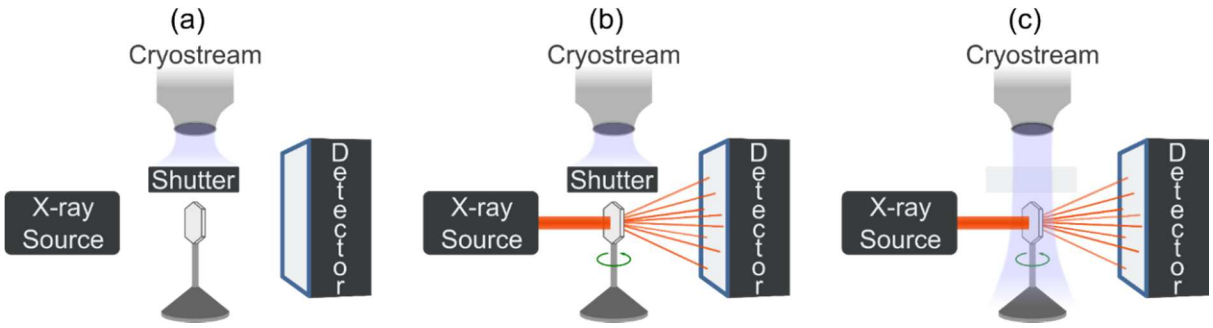


Figure 6.3.2 Experimental configuration used in time-resolved X-ray diffraction experiments at the Cornell High-Energy Synchrotron Source (CHESS). (a) A sample with its external solvent removed is placed in the X-ray beam path, (b) room temperature data is collected, and (c) the cryostream is unshuttered and data collected while the crystal is cooled.

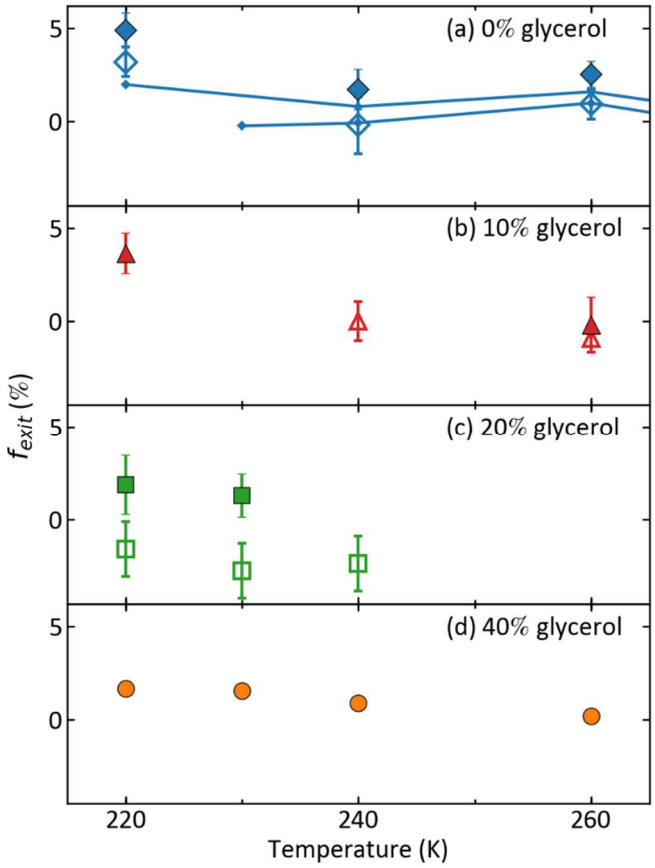


Figure 6.3.3 Estimates of the volume fraction of internal solvent present at room temperature that cannot be accommodated within the solvent cavity volume deduced from refined structures of the initial cold crystal (solid symbols) and the crystal after unit cell expansion (open symbols). Solid lines indicate results for crystals that were cooled at only 0.1 K/s. The total solvent volume at each temperature was estimated as described in (Moreau *et al.*, 2019) by assuming the first hydration layer does not contract and that the remaining solvent has the same contraction as the bulk liquid. Other reasonable assumptions don't qualitatively change the results. This estimate neglects the effects of salts present in the crystallization solution (15% w/v ammonium sulfate, 2% w/v $CdSO_4$) on the solvent's volume change on cooling. This salt concentration suppresses the bulk freezing temperature by only 4 °C and should only modestly reduce the solvent volume expansion on cooling from room temperature to 220-260 K relative to that for pure water. For glycerol-containing solutions at the concentrations studied here, the effects of salts on solvent volume changes should be negligible. Unit cell expansion for crystals with 0% and 10% glycerol brings the excess solvent volume down to toward 0. Complete data sets acquired before and after expansion, needed to determine the change solvent volume, were not obtained for all glycerol concentrations and temperatures.

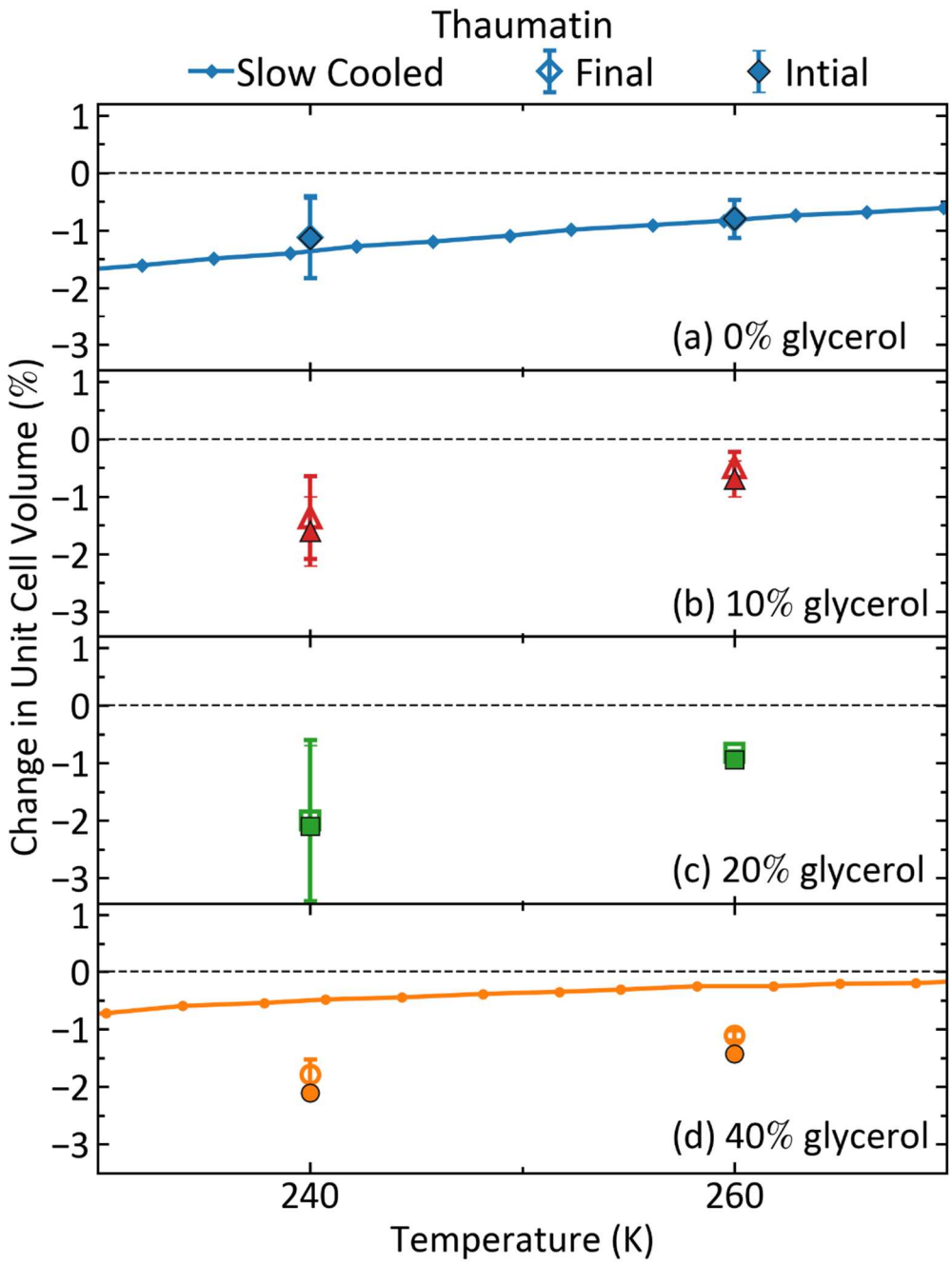


Figure 6.3.4 Change in unit cell volume relative to room temperature for thaumatin crystals that were slowly cooled at ~0.1 K/s (solid lines and small symbols) using data from (Warkentin & Thorne 2009), and that were abruptly cooled (in <1 s) to each temperature (large symbols), versus temperature. Closed and open symbols indicate unit cell volumes measured just after cooling had completed and at the completion of data collection (typically at t=40 s), respectively.

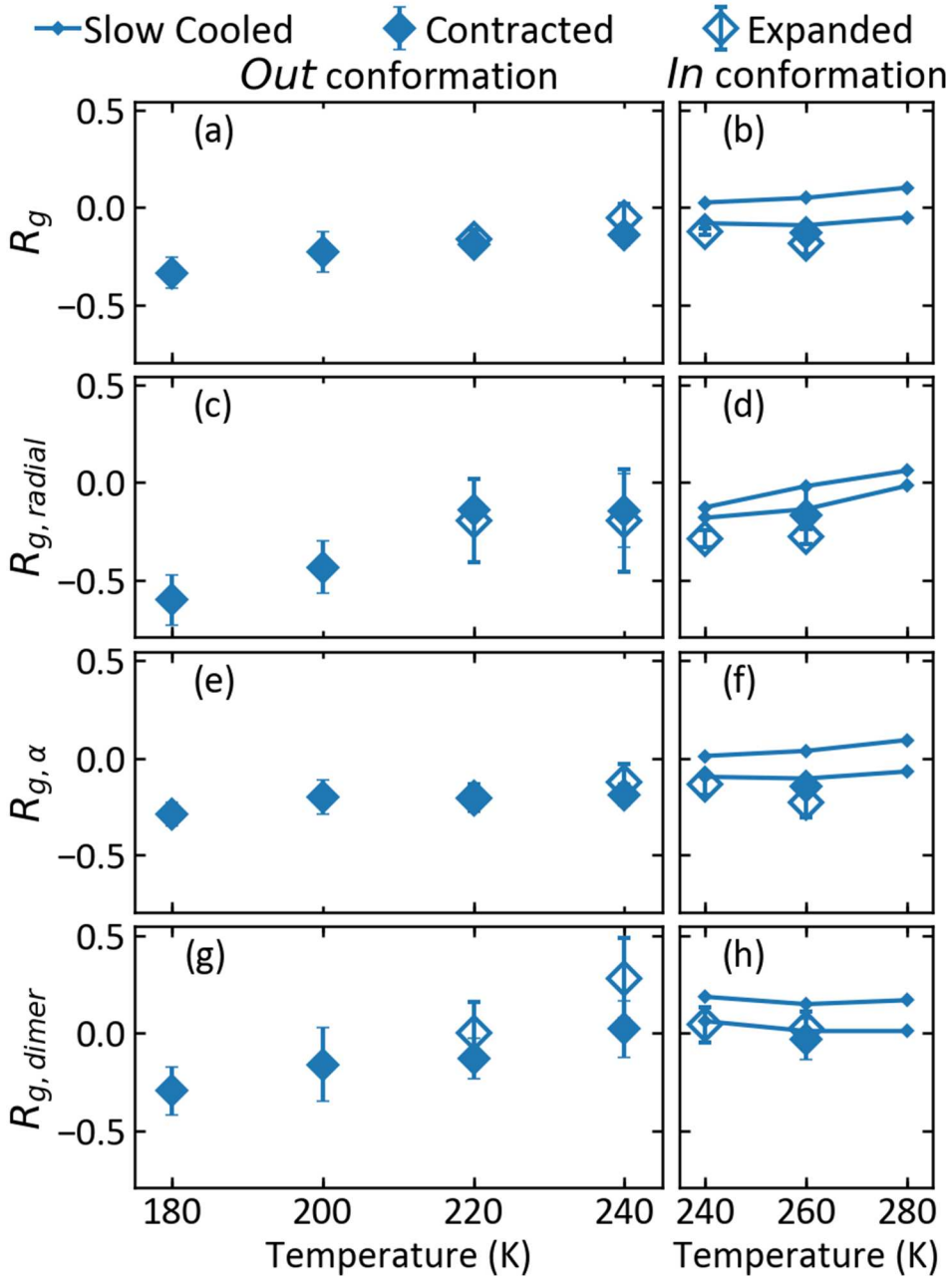


Figure 6.3.5 (a,b) The radius of gyration of a ferritin monomer and (c-h) its components along the axes shown in Fig. 4.1(d). The left and right panels show results obtained from crystals with GLN82 in the "*in*" and "*out*" configurations, respectively. Closed and open symbols indicate values determined just after cooling and after unit cell expansion had completed, respectively, and solid lines indicate results obtain for crystals that were slowly cooled.

6.4 SUPPORTING MATERIAL FOR CHAPTER 5

6.4.1 Number of ice crystallites required for continuous diffraction rings

The minimum number of ice crystallites required to generate azimuthally continuous diffraction rings can be estimated from the illuminated sample volume and the incident beam divergence. This estimate can in turn be used to set a bound on the crystallite size. Assuming the ice has a sufficiently large grain size that its Bragg peaks have a breadth in the azimuthal direction determined solely by the beam divergence, α, the peak's arc length is

$$\Delta = F\tan(\alpha).$$

F is the sample-to-detector distance. The circumference of the diffraction ring is

$$L = 2\pi F\tan(2\theta).$$

If placed end to end, the number of crystallites needed to generate a complete diffraction ring is $N \approx L/\Delta$. This is multiplied by 10 to ensure there is significant overlap between the Bragg peaks to give

$$N \approx 10\cdot 2\pi\frac{\tan(2\theta)}{\tan(\alpha)}.$$

This corresponds to 83,000 crystallites for a diffraction peak at $2\theta = 30^o$ and a measured beam divergence of $\alpha = 0.025^o$ for the CHESS station used in our experiments on apoferritin, thaumatin and lysozyme. For a 100 μm diameter beam and a 500 μm path length, the illuminated volume is $V \approx 3.9\cdot 10^6$ μm^3. Dividing this total volume by 83,000 gives an upper bound on the crystallite volume of 47 μm^3 or, assuming a spherical crystallite, a diameter of 4.5 μm. This crystallite size gives a finite-size broadening of $\beta \approx 0.0015^o$, 45 times smaller than the observed hexagonal ice peak breadth, suggesting that the observed peak breadth is limited by the instrumental broadening, as assumed in this estimate.

6.4.2 Size estimate of ice crystals contributing to ice zingers

A lower bound? on the size of an ice crystal that contributes to a single pixel ice zinger was calculated with Scherrer's equation using the angular breadth of a single pixel. For a diffraction image recorded at a sample-to-detector distance of F, wavelength of λ and pixel size μ, the angle subtended by a single pixel is

$$\beta = \tan^{-1}\left(\left(N+1\right)\mu/F\right) - \tan^{-1}\left(N\,\mu/F\right).$$

Given a diffraction angle of θ, N is the radial distance of the zinger from the detector's beam center location in units of pixels

$$N = \frac{F}{\mu}\tan\left(2\theta\right).$$

The breadth can be used in Scherrer's equation to determine a lower bound on the crystallite size as

$$\delta = \frac{\lambda}{\beta \cdot \cos\left(\theta\right)}.$$

Using $F = 500$ mm, $\mu = 0.172$ mm and $\lambda = 1$ Å, the crystallite size for the (112) hexagonal ice peak at 1.916 Å resolution is 4,000 Å.

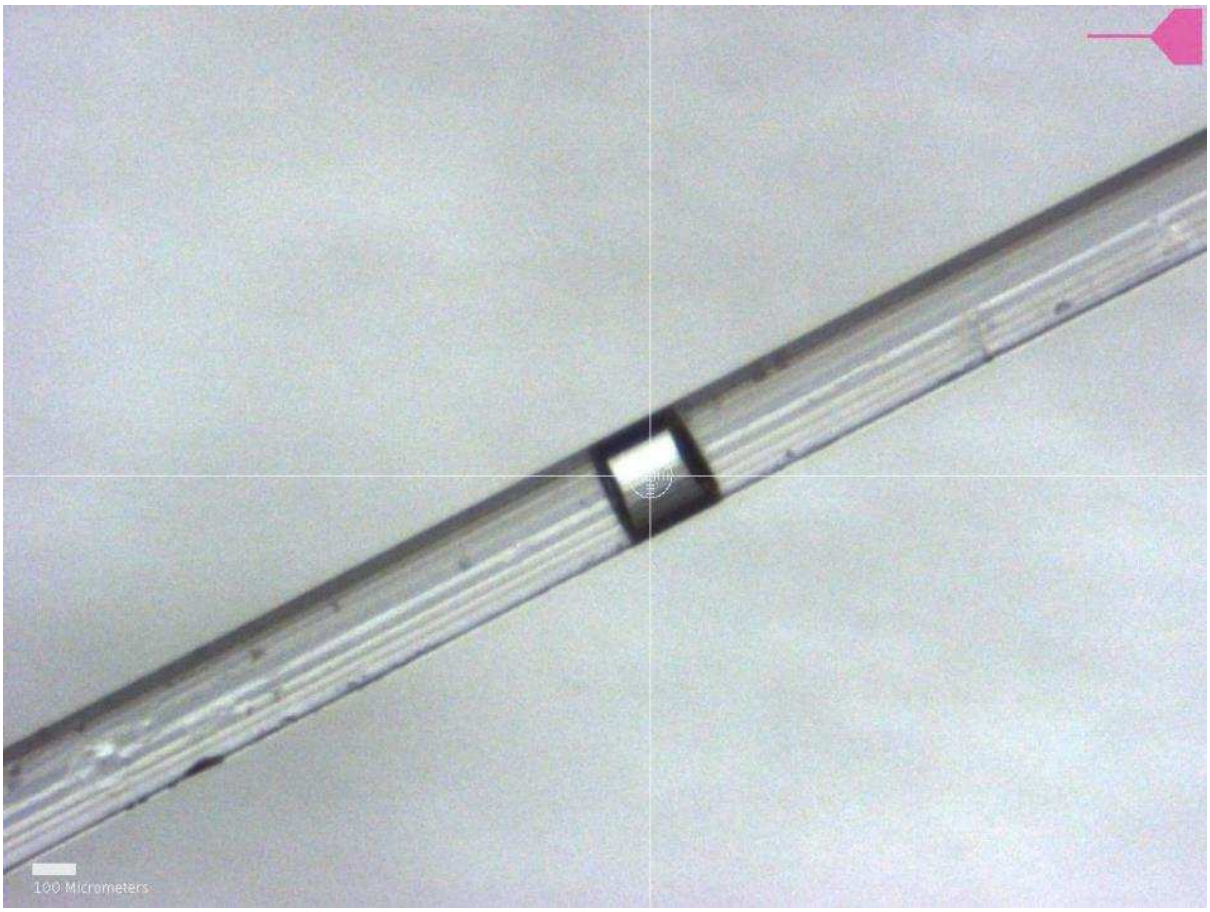


Figure 6.4.1 X-ray diffraction collection from cryocooled glycerol / water solutions. Samples were prepared by injecting ~10 nl of solution into a 250 µm diameter, ~2 cm long thin-walled polyester tube. The length of tubing filled with solution was approximately 200 µm. The tube was then affixed to a goniometer base (Mitegen GB-B1A) and centered in the X-ray beam.

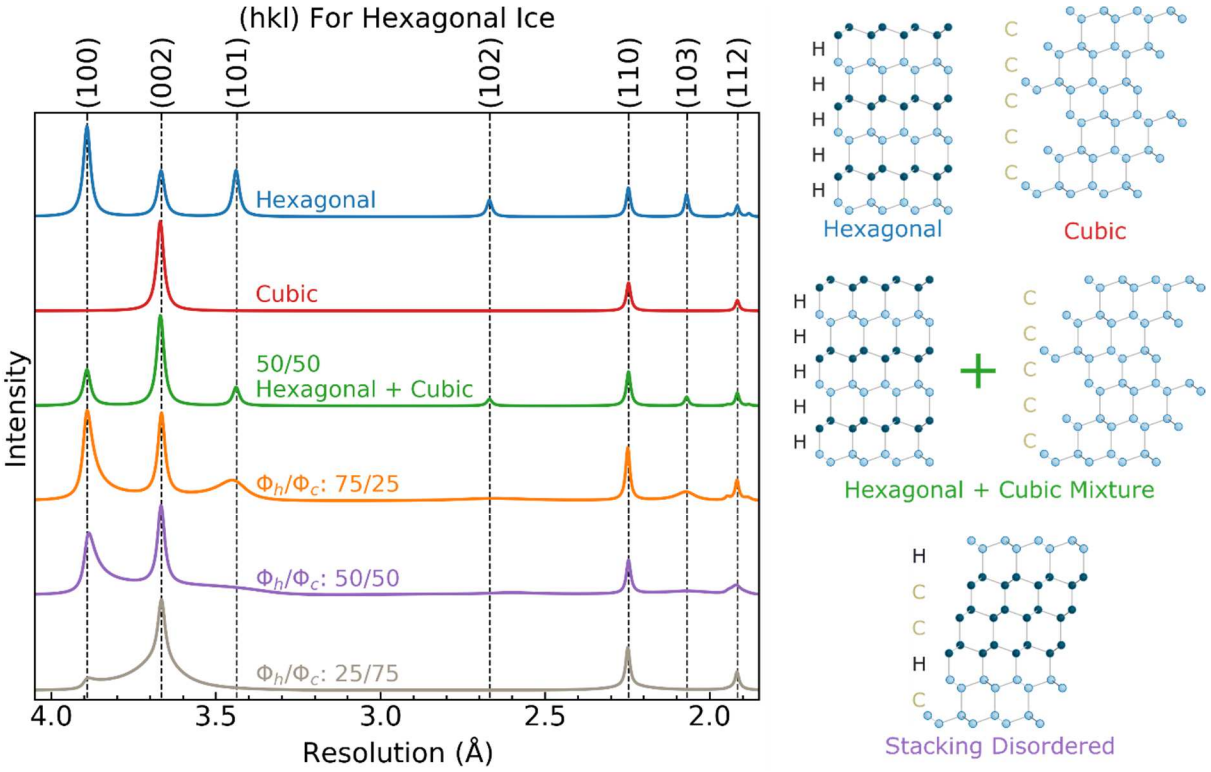

(hkl) For Hexagonal Ice
(100)
(002)
(101)
(102)
(110)
(103)
(112)
Hexagonal
Cubic
50/50
Hexagonal + Cubic
Φh/Φc: 75/25
Φh/Φc: 50/50
Φh/Φc: 25/75
Intensity
4.0
3.5
3.0
2.5
2.0
Resolution (Å)
H
C
Hexagonal
Cubic
Hexagonal + Cubic Mixture
Stacking Disordered

Figure 6.4.2

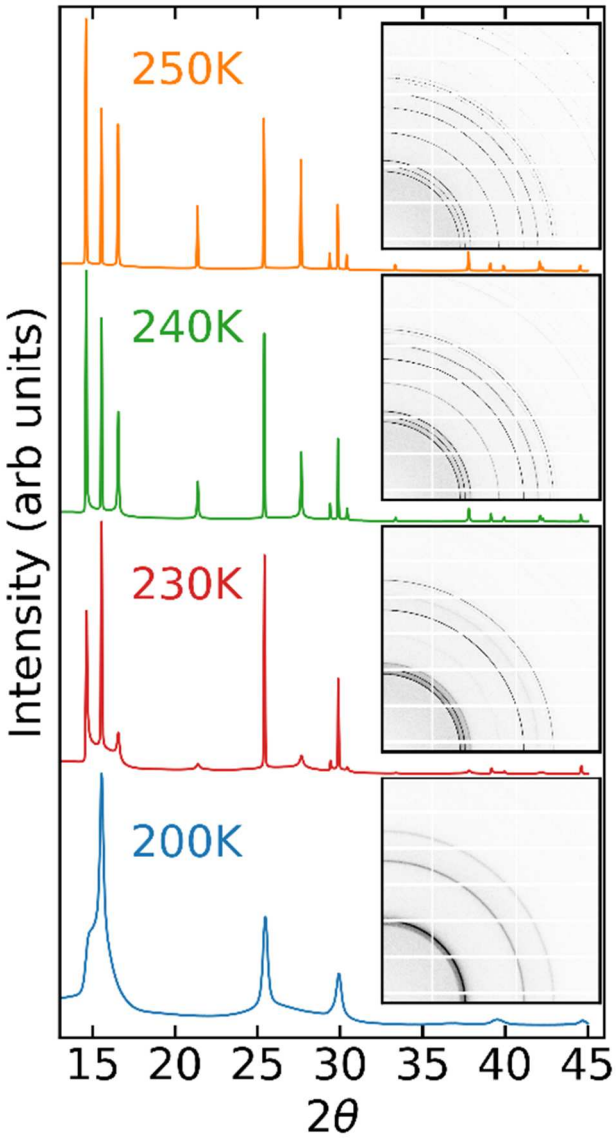


Figure 6.4.3 Generating a pure hexagonal ice sample. First, a 30% w/w polypropylene glycol 425 solution in a 500 µm diameter PET tube was abruptly cooled to T=200 K by unshuttering a cold gas stream. Abrupt cooling to a temperature where the ice nucleation rate was high resulted in an ideal powder pattern with a large width (indicating a small crystallite size) and having a stacking disordered intensity profile. As this sample was slowly warmed, the intensity profile evolved from stacking disordered to purely hexagonal (Kuhs et al., 2012) and the peak breadths decreased

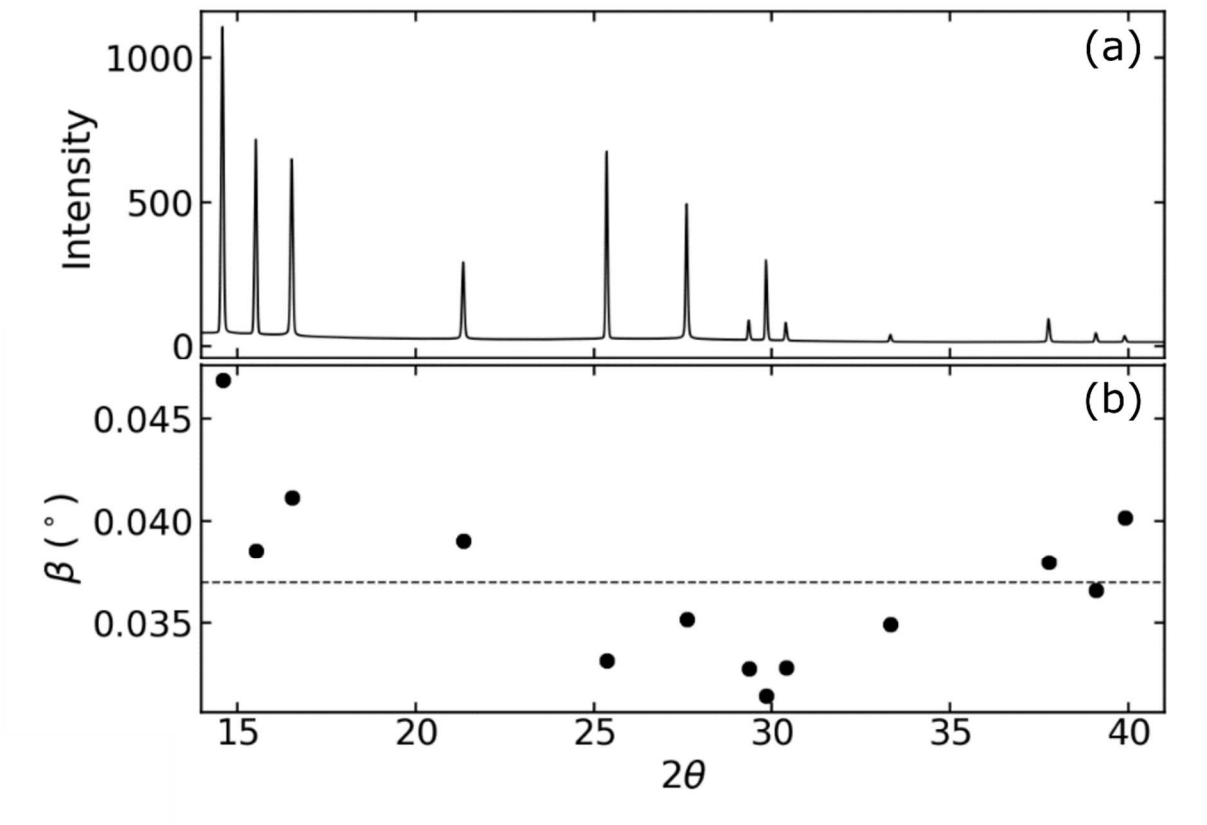


Figure 6.4.4 (a) The diffraction pattern recorded from the hexagonal ice calibration sample at 250 K before the loss of ring uniformity(b) was used to estimate the instrumental broadening.

(hkl)	Resolution Range (Å)
(002)	3.753 – 3.581
(110)	2.294 – 2.209
(112)	1.935 – 1.897

Table 6.4.1