Improved methods for serial femtosecond crystallography data reduction

Helen Mary Ginn

DPhil in Structural Biology

University of Oxford



Supervisors:

Professor David Ian Stuart
Dr Gwyndaf Evans

Word count: 40,968

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This thesis is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and acknowledgements. This dissertation contains fewer than 50,000 words including appendices, figure legends and tables.

Helen Mary Ginn

September 2017

Abstract

X-ray free electron lasers (XFELs) are a young technology which alleviates a number of issues, including radiation damage, which remain problematic in conventional macromolecular X-ray crystallography. As it turned out, existing methods had been optimised for synchrotron-based rotation data and coped poorly with snapshot XFEL diffraction images. In the first four results chapters, I present algorithms developed in my efforts to contribute to improved data reduction for XFEL analysis. In order of execution, these are: detector geometry refinement, indexing, initial orientation matrix refinement and post-refinement. Geometry refinement has been a field plagued with the problem of correlated parameters. I introduce the "slip-and-slide" algorithm to solve this by re-parameterising detector movements based on their effect on diffraction. I developed an indexing algorithm, TakeTwo, to reduce diffraction pattern wastage by maximising information content on a single image. Initial orientation matrix refinement aims to reduce, without any prior reference, errors of a crystal model from its own information content. When an approximate solution is available, one may implement post-refinement, which, in combination with a model describing the partial illumination of each reflection, drives the model parameters towards a self-consistent data set. In the final results chapter, I present an example of how the combination of all of these algorithms can push the information quality to new levels, illustrated by merging XFEL diffraction collected from scarce quantities of bovine enterovirus type 2 crystals and demonstrating the power of the amplitudes to drive phase extension. I end the thesis with a discussion of the realistic impact of these algorithms, the directions that methods developer communities could take from here onwards, and my personal scientific direction for the future.

Acknowledgements

First and foremost, I would like to thank my supervisor, Professor Dave Stuart, for guiding me throughout my DPhil. I could not ask for a better balance of scientific guidance, ideas, warnings, and a vast wealth of information collected from the early days of crystallography. Even more importantly, I was given the space to run on my own motivational steam, allowing me to explore stupid ideas which I needed to see fail myself. Having a fairly different scientific field from the rest of the lab, Dave has been extremely generous with his time to fulfil the role of post-doctoral supervisor and also principle investigator, and has a wonderful ability to describe shapes of scatter plots in terms of everyday objects such as churches, birds and scattered iron filings.

Despite the fact that my co-supervisor works in a different city, I have had a tremendous amount of support in the form of reading manuscripts, invitation to CCP4 events and a voice of sanity from Dr Gwyndaf Evans and I would like to thank him for all he has done, but especially bringing my attention to my now-ceased habit of flipping between present and past tense in my Methods.

Although not my DPhil supervisor, I must thank Elspeth Garman for introducing me to the world of attending conferences – I do not think I would have approached the wider scientific community without Elspeth dragging my nervous self to my first Gordon Research Conference.

I extend my warmest gratituds to Robert Gilbert. During my DPhil he has supported me in the role of College Advisor. In my undergradute degree he was my College Tutor at Magdalen College. It makes me proud to be able to live up to and fulfil his expectations.

It is impossible to word a complete explanation of how indebted I am to Natalie Watkins, who entered my life as a pole-dancing, snake-charming neuroscientist in 2011. She turned my view of this world upside-down, and in doing so corrected the perception of myself to the right way up. Nobody is complete without a Natalie in their life.

Helen Duyvesteyn has been the most fantastic guinea pig who was subjected to my fledgling teaching skills as my first student. It turns out that we do not just share a first and middle name, but also an outlook, sense of humour and very strong friendship known simply as Helen^2. I wish her the best of luck in her own pursuit of a DPhil, as she flows down the pipeline two years behind me.

I will always have fond memories of drinking Kasteel Donker beer and chatting about crystallography in the early hours of the morning with Jonny Brooks-Bartlett, in a much enjoyable ritual that can only ever be described as 'Kasteelography'.

Sometimes bonds are formed with those outside the immediate environment in Oxford as you meet people at conferences who change the way you see the world and have a big impact on you. And so I found myself clicking with the crystallographic crowd in Hamburg. This includes Thomas White, Anton Barty, Arwen Pearson, Carolin Seuring and many others – an entire building full of them.

I cannot express how much I am reliant on the skills of those who perform magic in the wet lab, or who can write code in Python without screaming with rage, or can build an end-station and not set something on fire, or have the patience to repeatedly perform phase extension on datasets I send promptly, without fail, at the time they just decide to leave, and then continuing to stay past midnight each night for weeks and weeks until it worked. Therefore, in a very particular order, determined by a rand() function, I would also like to thank a miriad of collaborators who I have worked with, on projects both big and small, both directly and indirectly: Rike Müller-Werkmeister, Dwayne Miller, Nicholas Sauter, Alexander Marx, Takanori Nakane, Darren Sherrell, Ren Jingshan, Robin Owen, Danny Axford, Jonathan Grimes, Aaron Brewster, James Parkhurst, Karl Harlos, Alke Meents, Sascha Epp, Valerio Mariani, Geoff Sutton, Richard Gildea, Arwen Pearson, Abhay Kotecha, Philip Roedig, Tom Walter and Charlie Bury.

My parents provided support and cheering from the sidelines throughout my DPhil and celebration of every paper that came out. I would also like to thank Benjamin Krishna for providing moral support and encouragement during the first three years of my DPhil.

The support staff at our division have supported me every step of the way – the administrative staff Eleanor Martin, Agata Krupa and Melissa Matthews have always swooped in to save me from bureaucracy anxiety and form-filling panic, another skill I lack. Jun Dong and Robert Esnouf have been particularly kind in tolerating my many terabytes of data parked on the cluster.

I finally must warmly mention my anonymous peer review referees, some of whom I may have unwittingly thanked already, for providing constructive criticism at the final hour to prevent me from publishing a silly sentence or spelling mistake which slipped through the net.

Contents

1 Introduction 14

1.1 A short history of hard X-ray sources for crystallography 15

1.2 Hard X-ray free electron lasers 15

1.3 Suitable samples for XFEL exposure 18

1.4 Specifics of each XFEL 20

1.4.1 Linac Coherent Light Source, California USA 20

1.4.2 SPring-8 Angstrom Compact Free Electron Laser, Japan 20

1.4.3 European XFEL, Hamburg, Germany 21

1.4.4 SwissFEL, Switzerland 22

1.4.5 Sample delivery methods at XFELs 22

1.4.6 Detectors used at XFELs 26

1.5 Software development for XFEL 29

1.5.1 Online data analysis packages 31

1.5.2 Background subtraction and hit-finding 32

1.5.3 Downstream processing: indexing and data reduction 33

1.6 Structural studies using the XFEL 35

1.6.1 Early XFEL structures 35

1.6.2 Lysozyme and other model systems 36

1.6.3 Radiation-sensitive samples 38

1.6.4 XFEL-based GPCR structures 39

1.6.5 Difficult anomalous phasing experiments . . . 41
1.6.6 Time-resolved studies using XFEL radiation and the use of difference maps. . . . 42
1.6.7 Coherent X-ray diffraction . . . 43
1.7 Focus of this thesis . . . 45

2 Detector geometry refinement. 47
2.1 Note to the reader . . . 47
2.2 Introduction . . . 47
2.2.1 The problem of detector geometry . . . 47
2.3 Methods . . . 52
2.3.1 Selection of spots from images by the droplet algorithm. . . . 52
2.3.2 Converting to reciprocal lattice point positions. . . . 53
2.3.3 Inter-spot vectors from images and idealised unit cell parameters. . 53
2.3.4 Data input. . . . 55
2.3.5 Detector geometry model. . . . 55
2.3.6 Panel movement modes. . . . 56
2.3.7 Refinement of unit cell dimensions . . . 57
2.3.8 Geometry refinement using indexing solutions. . . . 57
2.3.9 Geometry refinement using unit cell information only. . . . 59
2.3.10 Calculation of anomalous signal. . . . 61
2.4 Results . . . 62
2.4.1 Corrections were made to the unit cell parameters. . . . 63
2.4.2 Indexing-driven geometry refinement . . . 63
2.4.3 Pre-indexing geometry refinement uses more images, but circumvents indexing . . . 66
2.4.4 Geometry refinement reveals a tilt in the CSPAD relative to the beam. 66
2.4.5 Improvement in random error reduction and indexing rates . . . 68
2.4.6 Data set quality improvement . . . 71

2.5 Conclusion 74

3 The TakeTwo indexing algorithm. 77

3.1 Note to the reader 77

3.2 Introduction 77

3.2.1 Indexing methods suited to synchrotron crystallography. 78

3.3 Methods 79

3.3.1 Selection of spots from images. 80

3.3.2 One-dimensional pseudo-powder patterns. 80

3.3.3 Matching observed vectors to theoretical vector candidates 80

3.3.4 Seeding an inter-spot vector network 81

3.3.5 Building up an inter-spot vector network 81

3.3.6 Generating an orientation matrix from two identified inter-spot vectors 82

3.3.7 XPLOR metric to determine similarity of two matrices. 83

3.3.8 Indexing additional lattices using the inter-spot vector network. . . 84

3.3.9 Modifications made after publication. 84

3.4 Results 85

3.4.1 Global experimental parameters can be corrected using the pseudopowder pattern. 85

3.4.2 TakeTwo is relatively robust against errors in geometry. 86

3.4.3 Indexing rates frequently exceed 100% for various crystalline samples. 89

3.4.4 The algorithm is frequently able to index multiple lattices on each image. 89

3.5 Conclusions 90

3.5.1 Reception from the XFEL community 91

3.5.2 Incorporation into CrystFEL 92

4 Initial orientation matrix refinement for still images. 93
4.1 Note to the reader . . . 93
4.2 Introduction . . . 93
4.2.1 Orientation matrix determination prior to initial orientation matrix refinement. . . . 94
4.2.2 The cypovirus genus infects insect cells and has a major impact on the silkworm industry. . . . 95
4.2.3 Previous attempts to solve the structure of CPV17 polyhedrin. . . . 96
4.3 Methods . . . 96
4.3.1 Preparation of cypovirus polyhedrin crystals . . . 96
4.3.2 Powder pattern collection at Diamond at room and at cryo-temperatures . . . 98
4.3.3 Data collection at SLAC . . . 98
4.3.4 Extracting data from the SLAC XTC streams (*cctbx.xfel*) . . . 99
4.3.5 Extracting data from the SLAC XTC streams (*Cheetah*) . . . 100
4.3.6 Indexing images using cctbx.xfel . . . 101
4.3.7 Indexing images using DIALS . . . 101
4.3.8 Indexing images using the TakeTwo algorithm . . . 102
4.3.9 Initial orientation matrix refinement . . . 102
4.3.10 Detector geometry refinement . . . 103
4.3.11 Refinement of beam centre, wavelength and detector distance . . . 104
4.3.12 Integration methods . . . 104
4.3.13 Resolving the indexing ambiguity . . . 106
4.4 Results . . . 106
4.4.1 Transition from room-temperature to cryo-temperature involves a 1 Å contraction of the unit cell dimension. . . . 106
4.4.2 Diffraction was obtained to high resolution. . . . 108

4.4.3 Different indexing methods index image hits with varying degrees of success. . . . 108
4.4.4 Fitting a plane to the background subtraction provided no benefit in quality of integration. . . . 109
4.4.5 Initial orientation matrix refinement produced higher quality orientation matrices. . . . 109
4.4.6 Phases were determined by molecular replacement using an average structure from seven related cypovirus polyhedrin structures. . . . 109
4.4.7 Data collection and structural refinement statistics. . . . 110
4.4.8 Phases were transferred to a data set collected at Diamond I24 beam line. . . . 113
4.4.9 A disulphide bond present in the room temperature data set was broken when the crystals were cryo-cooled. . . . 115
4.5 Conclusions . . . 115
4.5.1 Reception by the XFEL community . . . 118

5 Post-refinement for still images. 119
5.1 Note to the reader . . . 119
5.2 Introduction . . . 119
5.2.1 Partiality of individual reflections as a source of error . . . 119
5.2.2 The concept of post-refinement . . . 120
5.2.3 Post-refinement applied to real data . . . 121
5.3 Methods . . . 121
5.3.1 Architecture of post-refinement . . . 121
5.3.2 Calculation of partiality . . . 123
5.3.3 Normalisation factors for partialities . . . 126
5.3.4 Fixed and varied parameters . . . 127
5.3.5 Calculating scale factors . . . 127
5.3.6 Choice of target function . . . 128

5.3.7 Choice of minimisation algorithm . . . 129
5.3.8 Outlier image rejection . . . 130
5.3.9 Merging intensities and error calculations . . . 131
5.3.10 Outlier rejection of reflections . . . 132
5.4 Results and Discussion . . . 133
5.4.1 On each cycle of refinement, the R_{split} values decrease and $CC_{1/2}$ values increase. . . . 133
5.4.2 The target function $Rfactor$ for individual images produces the best merging statistics. . . . 134
5.4.3 CPV17 crystals are best described by a bandwidth-based model rather than a mosaicity-based model. . . . 134
5.4.4 The exponent, bandwidth and mosaicity for CPV17 crystals should remain fixed. . . . 136
5.4.5 The Nelder-Mead algorithm finds a lower minimum than the step-search algorithm. . . . 136
5.4.6 Normalised partialities improve the values of R_{work} and R_{free}. . . . 139
5.4.7 Low resolution reflections are exquisitely sensitive to accuracy of numerical integration of partialities. . . . 139
5.4.8 Post-refined data quality is sufficient to see anomalous signal from sulphur atoms. . . . 141
5.4.9 Optimal outlier rejection is on a per-image basis, not on a merging-stage basis. . . . 141
5.4.10 Overfitting can be protected against by checking independent measurements such as the anomalous signal peak heights. . . . 144
5.5 Conclusion . . . 145
5.5.1 Reception by the XFEL community . . . 145

6 Structure solution of bovine enterovirus (BEV) type 2. 147
6.1 Note to reader . . . 147

6.2 Introduction . . . 147
6.3 Methods . . . 150
6.3.1 Crystal production . . . 150
6.3.2 Data collection . . . 151
6.3.3 Hit-finding . . . 151
6.3.4 Initial detector geometry correction . . . 152
6.3.5 Finding spots on individual images. . . . 152
6.3.6 Correction of detector distance. . . . 153
6.3.7 Indexing images . . . 154
6.3.8 Extra detector geometry correction . . . 155
6.3.9 Integration of images . . . 155
6.3.10 Initial orientation matrix refinement . . . 156
6.3.11 Post-refinement and final merge . . . 157
6.3.12 Phase extension from initial phase information. . . . 157
6.3.13 Refinement of atomic model against merged intensity values. . . . 159
6.3.14 Omit map of the cofactor in the lipid-binding pocket. . . . 159
6.4 Results . . . 159
6.4.1 Diffraction was observed during the course of data collection. . . . 159
6.4.2 Cryo-cooling samples for XFEL beam time is unnecessary or detrimental. . . . 160
6.4.3 Initial hit-finding results from cctbx.xfel were conservative. . . . 162
6.4.4 Geometry refinement revealed a tilt to the detector. . . . 163
6.4.5 Indexing success rates were very high. . . . 163
6.4.6 Initial orientation matrix refinement reduced the error significantly with small tilt corrections. . . . 163
6.4.7 Integration strategy was optimised based on merging statistics. . . . 164
6.4.8 Post-refinement and the partiality model extracted additional information. . . . 164

6.4.9 Data collection and structural refinement statistics 166
6.4.10 Extension from initial phases of a synchrotron structure were possible from 5 Å. 166
6.4.11 Atomic model refinement and omit map showed binding of sphingosine to the viral capsid. 166
6.5 Conclusions 170

7 Discussion 172
7.1 Summary of the progress and the effects on the field of crystallography. . . 172
7.2 Possible directions of focus in XFEL studies by the community. 175
7.3 Personal direction and focus in structural biology. 178

8 Appendix 191

Chapter 1

Introduction

Free electron lasers (FELs) are a technology which is able to produce extremely bright pulses of photons of various wavelengths. The free electron laser was invented in 1971 (Pellegrini, 2012) and was capable of creating infrared, visible and UV-wavelength radiation. Free electron lasers of increasing maximum photon energy have been developed over time, with a landmark being the Linac Coherent Light Source (LCLS) in Stanford, USA, which first lased in April 2009. This laser was capable of producing X-rays to 9500 eV, putting it in the range capable of supporting X-ray crystallography. Since 2009, a large number of research projects have been focusing on collecting diffraction data at X-ray FELs (XFELs) in order to take advantage of its unique properties. However, there was a niche to be filled, as there were (are) very few researchers focusing on the technical challenges involved in processing the type of diffraction data from XFELs. My DPhil focused on overcoming as many of these technical challenges as possible to assist structure solution of biological macromolecules. This thesis is focused on the methods, rather than the biological results of particular samples which were derived from my work.

1.1 A short history of hard X-ray sources for crystallography

Synchrotrons built in the last few decades have been an essential replacement for the previously used home X-ray sources which were used to solve the first protein structures such as myoglobin (Kendrew et al., 1958), haemoglobin (Perutz et al., 1960), lysozyme (Blake et al., 1965) and ribonuclease (Wyckoff et al., 1967; Kartha et al., 1967). The advent of the new Deutsches-Elektronen Synchrotron (DESY) allowed X-rays to be used to probe protein structure on a wider scale, beginning with low-angle X-ray scattering experiments of insect muscle tissue. Synchrotron X-ray radiation was preferable due to its monochromaticity, wider range of wavelengths, improved convergence and focus (Leigh and Rosenbaum, 1974). However, this operation mode was not dedicated to X-ray generation: the first generation of synchrotron sources were off-shoot projects from hardware dedicated to high-energy particle physics. At this time, the inability to easily display or transfer information about atomic coordinates derived from crystallography experiments was addressed by the establishment of the Protein Data Bank (PDB). This was designed as a repository to enable transfer of the atomic coordinates generated from crystallographers describing the three-dimensional structure of solved biomolecules. The PDB has been well-reviewed (Berman, 2008). The second generation of synchrotrons was built for X-ray generation as its primary purpose, and the third generation of synchrotrons was optimised for brightness by producing longer straight sections capable of accommodating undulators and multi-pole wigglers. Maximising the undulator length, taken to an extreme, leads to the fourth generation of X-ray sources: the X-ray free electron laser.

1.2 Hard X-ray free electron lasers

X-ray free electron lasers were first developed for infra-red energy light, and were scaled up in several stages to ultra-violet light, soft X-rays and eventually hard X-ray free electron lasers, the use of which in diffraction studies is the focus of this thesis. The major difference in the layout of the XFEL compared to the synchrotron is the linearity of the XFEL.

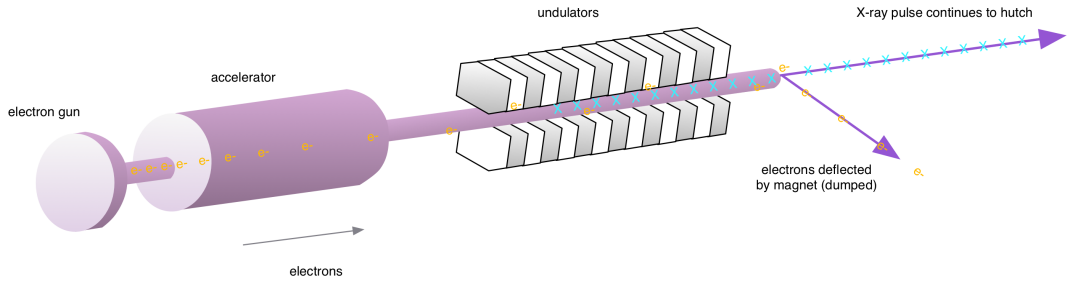


Figure 1.1: Physical layout of the XFEL and rudimentary outline of the strategy used to generate X-ray pulses. Electrons are emitted from the electron gun (paths determined by orange e- symbols) and pass through the accelerator which brings them closer to the speed of light than a synchrotron. Electrons pass through a series of undulators containing magnets on each side with alternating polarities, which causes the electrons to wiggle and consequently produce coherent X-rays which are grouped into very high intensity pulses. The electrons are dumped at the end of the series of undulators so that they do not contaminate the X-ray beam and the X-rays proceed to the hutch.

XFELs cause linear acceleration of electrons along several hundred metres, followed by bunching by undulators (alternating polarity of magnets surrounding the electron beam), which produces a series of femtosecond-duration pulses (Figure 1.1). The electrons in these pulses emit X-rays in the same direction as the electrons, and each of these pulses has a flux $10^9 - 10^{10}$ times higher than that of a typical third-generation synchrotron. Assuming similar beam size properties, this means the nature of the experiment, and the nature of the data processing, is fundamentally changed. Crystals in the path of the XFEL at near-full or full flux are entirely obliterated, but diffraction from the crystal is collected before the vast majority of radiation damage has begun: this effect is named "diffraction before destruction" (Chapman et al., 2014). This ability to outrun radiation damage forms the basis of many of the perceived advantages of the XFEL compared to the synchrotron. However, it produces a significant challenge in terms of data processing, as the data which can be collected from a single shot of a crystal is somewhat limited.

XFELs can theoretically operate in a self-amplified spontaneous emission (SASE) mode, where there is no filtering of the energies which are emitted by the X-ray free electron laser and the energy spectrum is broader and more variable. Some XFELs are also ca-

pable of operating in self-seeding mode; in this the X-rays from the first undulator are filtered by a thin diamond crystal to isolate a single wavelength. The electron beam is simultaneously diverted around this crystal by a relay of magnets. These filter X-rays to a subset significantly closer to a monochromatic wavelength, which then continue through the subsequent undulators to reach amplified peak saturation.

Due to the strong flux causing immediate destruction of the crystal, only a single frame, or "snapshot", can be taken of each crystal. The orientation of the crystals is often essentially random, and certainly cannot be presumed even if biased by the sample delivery method, and generally must be determined from scratch for each individual image (a computational method called indexing). This is complicated by the possibility, at least with several of the sample presentation methods, that multiple crystals may be exposed by the beam simultaneously and increase the number of spots on each image. In cases where large numbers of crystals can be generated but few produce diffraction, or if only few crystals can be produced in total, a rotation series using a synchrotron source may be a better method of collecting Bragg intensities for a static structure solution, especially if the crystal preparation process is particularly laborious.

However, several unique possibilities are presented by the XFEL which are not possible to be achieved at a synchrotron. In cases where diffracting crystals can be produced in bulk, the ability to outrun radiation damage is clearly an advantage for radiation-sensitive crystals, especially those with heavy metal ions or sensitive but fragile structures. Small crystals may be particularly amenable to XFEL analysis, if they were to yield little useful data at a microfocus beam line due to insufficient amplification of Bragg intensities. The intensity of the XFEL pulses theoretically allows a measurable coherent diffraction signal to be collected from a non-crystalline biomolecular source, producing coherent diffraction. The femtosecond nature of the pulses allows researchers to resolve time-dependent structural changes in their samples on the order of bond breakage and formation events; theoretically this could be combined with collecting coherent diffraction but is also suitable for crystalline samples where the structural changes do not cause the loss of crys-

talline order.

1.3 Suitable samples for XFEL exposure

Although any crystal size can be exposed to XFEL pulses, researching using microcrystals or nanocrystals may take advantage of the significant laser power of the XFEL compared to microfocus beam lines at synchrotrons. However, microfocus beam lines are becoming more effective with time. Effects of long-range disorder in the crystal split reflections into microdomains at high resolution, and ultimately prevent high resolution information from being obtained due to spreading the signal over a larger number of pixels, all of which are subject to an approximately constant background contribution. Crystals which have lower long-range disorder, and therefore sharper reflections, may respond more readily to increasing the flux as these reflection intensities more readily rise to useful levels.

The ability to out-run radiation damage has provided a means by which sensitive crystals containing important heavy metal ions can be analysed before the radiation damage has caused site-specific damage to the metal ion sites, potentially rendering the crystal structure useless for biological analysis. Thus, the undamaged manganese ion cluster in Photosystem II has finally been studied with the advent of the first XFEL (Kern et al., 2012, 2014; Kupitz et al., 2014; Suga et al., 2015, 2017; Young et al., 2016). This has also been used to analyse structures which would have otherwise been exposed to too high a radiation dose at a synchrotron and lose the specific and biologically active arrangement of fragile structures. It is claimed that radiation damage disturbs the specific arrangement of water molecules in the M2 influenza channel, provoking work to be undertaken at an XFEL (Thomaston et al., 2015). Note that longer pulse times may introduce radiation damage.

One of the novel opportunities presented by the development of XFELs is the ability to access new timescales on which time-resolved dynamics can be elucidated. Traditional experiments with synchrotrons are mostly only capable of accessing millisecond timescales, although the improved flux at synchrotron microfocus beamlines may man-

age to improve on this. In practice, due to the damage one crystal sustains in the beam, one crystal is used to generate a single frame at a given timepoint in a movie of the entire reaction event. However, each crystal only provides a slice of reciprocal space and therefore data from many crystals must be collected before a full three-dimensional structure can be recovered for a given time point. Furthermore, as a reaction event must be synchronised, this must be initiated by a timed trigger. If the trigger is incapable of exciting the entire crystal, the partial excitement will produce a mixture of different states (in a simple case, one ground and one excited state), which will require more data to increase the signal-to-noise ratio. As XFEL is in its infancy, most time-resolved experiments are performed with a trigger which is straightforward to implement, and therefore the studied reactions of published experiments to date have been initiated using a laser of a given wavelength which aims to trigger the desired reaction.

The initial aim of the XFEL was to do away with the need to create ordered lattices of single particles and image them directly via the coherent diffraction produced. Although there has been some success on this front, the number of diffracted photons has not yet been high enough to push further than on the order of 10 nm resolution at this point. Studies are currently constrained to very large biological particles or other particles with high electron density. As well as barely reaching the minimum number of photons (and possibly too weak at present to image smaller biological particles), sample delivery is an even harder challenge as it is not currently possible to separate diffraction from multiple particles entering the beam simultaneously. Furthermore, the disordered solvent content around the particle will serve to add additional coherent scattering. This can be easily distinguished from the signal in a crystalline environment due to the concentration of diffractive power into Bragg peaks, but becomes indistinguishable from the signal from the protein when considering single particle diffraction.

1.4 Specifics of each XFEL

1.4.1 Linac Coherent Light Source, California USA

The first hard X-ray free electron laser that became operational was the Linac Coherent Light Source, located in Palo Alto, California, USA. The first iteration of this XFEL can operate up to X-ray energies of 10 keV and a pulse rate of 120 Hz. Each pulse is between 30-50 fs duration at shorter wavelengths, and can be between 40-300 fs duration at soft X-ray wavelengths (~15 Å or higher). The beam can be focused to 2-3 μm and contains ~2.0×10^{12} photons per pulse. LCLS has been accepting proposals since 2010. It has six separate end-stations, some of which are able to be operated simultaneously, with specialisations for macromolecular femtosecond crystallography (MFX beamline), coherent X-ray imaging (CXI beamline) and X-ray pump-probe experiments (XPP beamline). At the time of writing, upgrades to LCLS are under development (to produce LCLS-II), and an additional project to allow LCLS-II to reach a higher energy beam. This version of the XFEL will have an increased repetition rate. LCLS-II aims to reach a million pulses per second, and although the final energy range has not been specified, would permit phasing using selenium atoms if it exceeds 12.67 keV. There is a choice of detectors: the Cornell-SLAC hybrid Pixel Array Detector (CSPAD) (Hart et al., 2012) (an in-house segmented detector capable of operation at 120 Hz) or single-panel Rayonix CCD detectors for which data collection must operate at a maximum of 10 Hz. Detectors at the CXI beamline run on rails and modifying the detector distance can occur during the course of an experiment with ease (possibly to the detriment of the data analysis team; an example is detailed in Chapter 6). Detectors at the XPP beamline are suspended on a robotic arm and are moved manually.

1.4.2 SPring-8 Angstrom Compact Free Electron Laser, Japan

The second XFEL, the SPring-8 Angstrom Compact Free Electron Laser (SACLA) is more compact than LCLS, and is currently the only XFEL capable of operating above 12.67 keV,

permitting selenium phasing (Yamashita et al., 2015). It has a lower repetition rate of 30 Hz and uses the MultiPort Charged Coupled Device (MPCCD) (Kameshima et al., 2014) detector which has eight separate segments, consisting of 1024 x 512 pixel panels with a pixel size of $50\mu m$. The pipeline for initial calibration and write-out of data installed at SACLA, which is described below (put in place by Takanori Nakane), recalibrates the data to have a gain of 10 counts per photon. In this setup, pixel saturation occurs at around 10,000 arbitrary detector units. Some of the four hutches have a fixed detector distance of approximately 50 mm, and others can be occasionally varied, but must be done so manually and with greater difficulty than at LCLS. This distance is also not measured automatically. This simplifies some of the pipeline as the distance in certain hutches is almost always within a couple of millimetres of 50 mm. This aids non-expert users of processing software.

1.4.3 European XFEL, Hamburg, Germany

The European XFEL situated in Schenefeld, Hamburg near the Deutsches Elektronen Synchrotron (DESY) campus came online for the first experiments in September, 2017, two beamtimes of which I attended. This XFEL has a different structure of X-ray pulses: they are intending to be eventually grouped into 600 microsecond bunchtrains with 2700 pulses arriving within each bunch train; each second carries 10 of these bunchtrains. Initial experiments are operating on a less ambitious train and pulse sequence. The detector expected to collect the diffraction is the Adaptive Gain Integrating Pixel Detector (AGIPD) detector (Henrich et al., 2011) (1 Megapixel version developed and 4 Megapixel version planned/in preparation) which will be capable of storing 800 values per pixel, stored in individual cells, to write out 800 images simultaneously. Although other detectors, such as the Jungfrau detector (Redford et al., 2016), are in development, the first few European XFEL beam times which took place before submission of this thesis featured the 1M AGIPD detector. This will supply future XFEL data analysts with a steady stream of technical challenges, including incorporation of 800 different pixel masks to flag the ap-

propriate dead pixels for each image write-out, plus 800 calibration settings per pixel per gain setting, of which there are three. Each gain setting is usually modelled as requiring an offset and a multiplication.

1.4.4 SwissFEL, Switzerland

SwissFEL is being developed in Switzerland as part of the Paul-Scherrer-Institut (PSI). At the time of writing, user experiments have not begun but pilot experiments are due to start (year 2017). It is 740 metres long and will be able to deliver X-ray flashes of between 1 and 60 femtoseconds. It began sending X-rays of 1.2 nm wavelength into the optics hutch on 31st August 2017, and will be increasing the energy and improving the beam quality over time.

1.4.5 Sample delivery methods at XFELs

Sample delivery to the X-ray pulses of the XFEL is non-trivial, both in the context of reducing sample consumption and reducing the number of wasted XFEL pulses or beam time. Several solutions have been so far developed, each of which has its own advantages and disadvantages and certain types of sample are suited to each type.

Liquid jet for crystalline solution or single molecules

The most common and established method for sample delivery to the XFEL is via a liquid jet (DePonte et al., 2008) passing in front of the beam buffered with helium gas within a vacuum to reduce dehydration (Figure 1.2). Crystals are passed through an injector nozzle and enter within the liquid jet, some of which then encounter an X-ray pulse and produce X-ray diffraction. Because of the balance between flow rate, sample concentration and X-ray pulse frequency, the proportion of crystals which are hit is often very low, and so a lot of sample is wasted. Although the liquid jet is injected at room temperature and despite the helium gas buffer, the vacuum does cause a drop in temperature during the

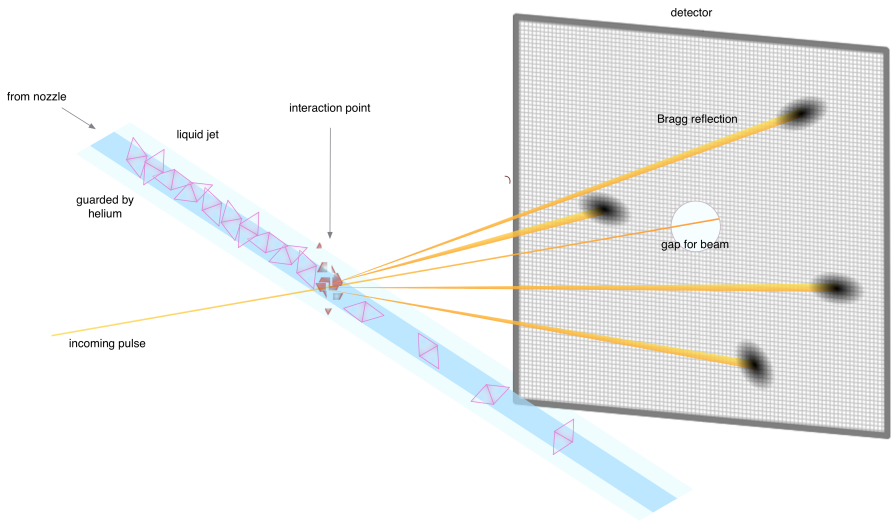


Figure 1.2: Experimental setup for liquid jet experiments. Liquid jet experiments have a reduced tendency to bias the orientation, but increase the sample wastage as most crystals pass by without being hit by an X-ray pulse.

crystals' path to the beam, which should be factored into the analysis of the structures if structure temperature is of importance. This is often the method of choice for crystals which are easily mass-produced and not in limited supply.

Gels with embedded crystals

Lipidic cubic phase gels have been used fairly extensively for most of the membrane protein crystal structures which have been solved at LCLS and SACLA (Kang et al., 2015), especially for the XFEL G protein-coupled receptor (GPCR) structures. LCP and other micro-extrusion media such as vaseline are able to sustain a low flow rate (0.02 - 0.50 mm/min) which reduces sample consumption (Sugahara et al., 2016). However, LCP has some disadvantages: in this format it is opaque, which makes it harder to diagnose problems when faced with the question of why crystals are failing to diffract in the XFEL beam. It also contributes to a higher background, leading to a poorer signal-to-noise. Alternatives are therefore sought after.

An agarose-based medium has been used as an inert delivery material for crystals (Conrad et al., 2015), such as diffraction of crystalline Sindbis virus capsids (Lawrence et al., 2015). This is a transparent medium, which is advantageous as it means the crystals can be observed using optical microscopy. However, one group say that although the agarose gel reduces the background observed, the viscosity of the medium means that the crystals are prone to dissolution (Sugahara et al., 2016). This group have suggested the use of a water-based matrix from water and hyaluronic acid, which produces an oil-free medium which is less susceptible to crystal dissolution. They demonstrate the success of hyaluronic acid and show the dramatic reduction in background and rate of change of background at almost all resolution ranges. They also report structure solution from (a) crystals of lysozyme and (b) proteinase K. However, the signal-to-noise ratio decreases for the former when immersed in hyaluronic acid and has a reported unit cell dimension greater than that in Super Lube (an alternative grease-based matrix which has also been used for sulphur-SAD experiments (Nakane et al., 2016b)). The unit cell dimensions for proteinase K show no reported change, and the signal-to-noise ratio of the merged data are negligibly greater. It is possible that the sample-to-detector distance may have changed due to a change in the position of the crystal extrusion stream, leading to an incorrectly assigned inflation of the unit cell dimension as these parameters are highly correlated (see Figure 1.7 for analysis of XFEL-based lysozyme crystal unit cell dimensions). This may lead to the mis-integration of integration windows and therefore mis-assignment of foreground and background; mixing these two will lead to an apparent reduction in signal-to-noise.

Micropatterned silicon chips for mounting crystals

Several micropatterned silicon chips designed for mounting crystals for exposure to XFEL pulses have been developed (see Figure 1.3). These consist of a silicon crystal base chip into which small holes have been fashioned for insertion of crystals. The silicon crystal base reduces background scattering by funnelling all the chip support diffraction into

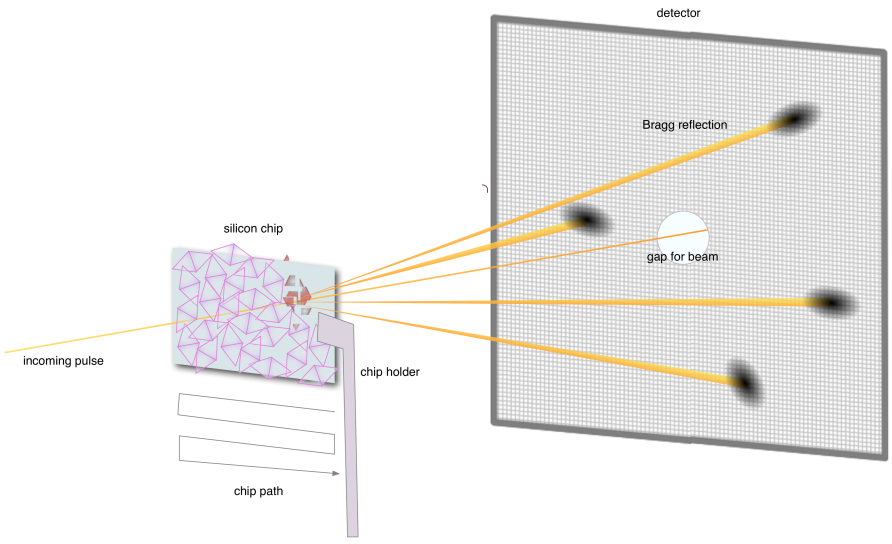


Figure 1.3: Experimental setup for exposing crystals mounted on a supporting chip. Chips have a tendency to bias the orientation, especially for needle-like crystals, but increase the hit rate substantially as all crystals can be targeted individually and can be good for lower yield samples.

Bragg peaks, and the experimental setup can be arranged such that these reflections do not satisfy the Bragg condition so they avoid hitting the detector. In fact, preventing silicon diffraction hitting the detector to avoid damage to the ASICs is of paramount concern when considering the extreme flux sent through a single reflection from an XFEL beam. There are two major manufacturers of chips: the group of Dwayne Miller (Oghbaey et al., 2016) (for which I provided the data processing results, not discussed in thesis) and Alke Meents (Roedig et al., 2015), both of which have gone through several rounds of iterations. The former group adds a layer of mylar film which adds minimal background, and is an impervious layer which prevents the crystals drying out. A bovine enterovirus structure was determined using the combination of the latter chip and all of the algorithms described in this chapter, and is the subject of Chapter 6.

Acoustic droplet delivery method

A method has been used to deliver crystals to the beam at a slower rate with more precision over their timing to reduce sample wastage (Fuller et al., 2017; Roessler et al., 2016). This involves applying a series of drops onto a tape (called the drop-on-tape (DOT) method) which is then passed on a conveyor belt to within reach of the beam path. The drops are propelled onto the tape by a sound wave in a method known as acoustic droplet ejection (ADE). The tape is then passed into the path of the beam, taking care to maximise exposure of the drop and minimise that of the tape. This technique not only reduces sample wastage, but allows timing precise enough and in the correct ranges for analysis of individual proteins by both X-ray emission spectroscopy (XES) and X-ray diffraction (XRD). In the work of Fuller et al. (2017), it had been used to probe the biological mechanisms, and metal reduction state, of the manganese cluster of photosystem II (PSII), a bacterial Mn-Fe ribonucleotide reductase and static room temperature structure solutions of phytochromes.

1.4.6 Detectors used at XFELs

Traditional synchrotron detectors are generally unsuitable for the high volume of photons per pulse generated by an XFEL. This produces many photons in an extremely small time frame. Detectors commonly used on synchrotron beamlines like the PILATUS3 S and X series have a hybrid pixel counting technology. Pixel counting technology measures whether a photon has or has not hit the detector in a very small time window. Over time, a very large number of measurements of small time periods accumulate and produce a total count. If the flux per unit time is too high then this will fail to produce a sensible dynamic range as coincidental photons are uncounted. Alternatively, detectors can use a photon integrating mode, where the detector collects photons which deposit a charge for a certain time frame, and then these are converted into an estimation of the number of photons. These are more suitable for XFEL experiments, but rely on a difficult calibration

of signal to photon count.

CSPAD family of detectors

The Cornell-SLAC pixel array detector (CSPAD) has been the in-house detector developed at LCLS at SLAC, which embodies the technological advances necessary to cope with the data generated by the XFEL diffraction. The CSPAD has been through several iterations of development, but across all versions comprises 64 application-specific integrated circuits (ASICs), pairs of which are bump-bonded to a single rigid surface which means that their relative positions and orientations can be assumed. These are arranged into quadrants with eight ASIC pairs on each quadrant. These quadrants are on guard rails which can be brought further in or out, and the entire detector can be shifted as a unit. This is done manually or on another rail depending on the beamline. The CSPADs at LCLS are periodically dismantled and rebuilt and this rearranges the segments of the detector each time. Rough estimations of the panel positions are calculated by LCLS but additional refinement is inevitably performed by researchers. The pixel size is 110 μm and each ASIC measures 184 x 192 pixels (Figure 1.4). The detector saturates at 14000 counts and typical pixel behaviour is shown in Figure 1.5.

MPCCD at SACLA

The MultiPort Charged Coupled Device is the detector employed at SACLA on four hutches. Like the CSPAD, it is also segmented, but comprises eight panels instead of 32. The pixel size is smaller than that of the CSPAD, 50 μm, and the panels are 512 x 1024 pixels wide (Figure 1.6). The increased number of pixels on the panel axes means that these detectors can be more sensitive to the effect of tilted panels on the spot positions.

Rayonix detector

The Rayonix detector is a commercial CCD detector (supplied by Rayonix) installed at the xpp beamline at LCLS and was capable of operating at 10 Hz in 2015, and therefore

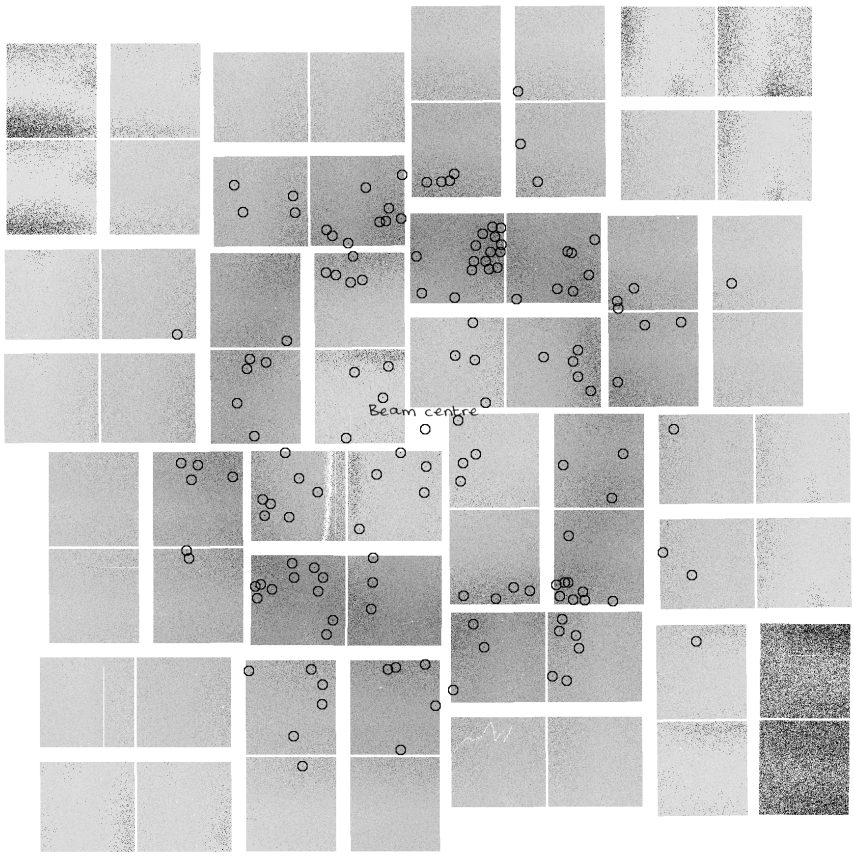


Figure 1.4: An example of a diffraction pattern from a 2013-version CSPAD, which is less well-behaved compared to more recent iterations. Diffraction spots are circled for clarity of differentiation from the background. The diffraction shown is from a CPV17 crystal slurry.

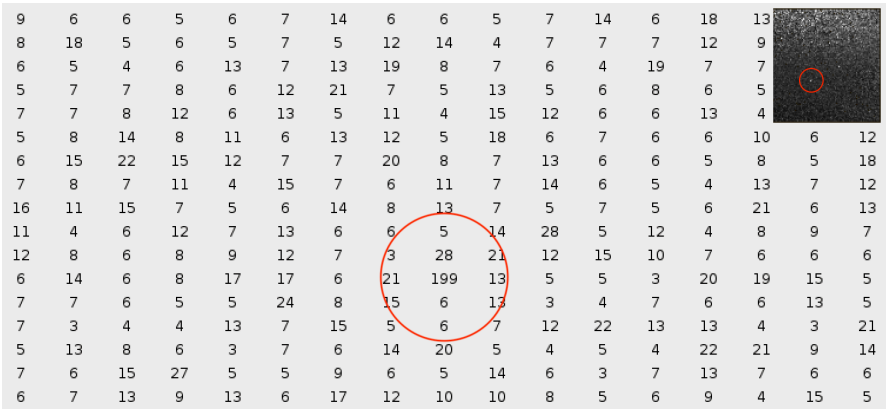


Figure 1.5: One of the middle-resolution panels, shown in both monochrome representation and in terms of pixel counts. The same spot featured on both representations is circled in red. Drawn using the hdfsee executable in CrystFEL (White et al., 2012, 2016).

cannot use the full potential of the XFEL pulse frequency. However, it consists of a single panel, and therefore has highly simplified geometry calibration. The high dynamic range (sensitive up to at least 60,000 8 keV X-ray photons per pixel) allows a wide range of reflection intensities to be integrated, but the large point spread function means that the optimal integration window is difficult to determine. This detector tends to be less widely used due to the reduction in the operation frequency required.

1.5 Software development for XFEL

Many software packages are available for synchrotron crystallographic data processing and have been available in some form for many decades. A large proportion of these are closed source, which protects the intellectual property of the author but is not in the spirit of scientific pursuits which should be open to inspection and criticism. The lack of source code under an appropriate license also prevents these software suites from being modified, and therefore new software packages have to be created. Some of these call binary executables from closed-source packages, but the general trend is to focus on creating open-source code to aid future development.

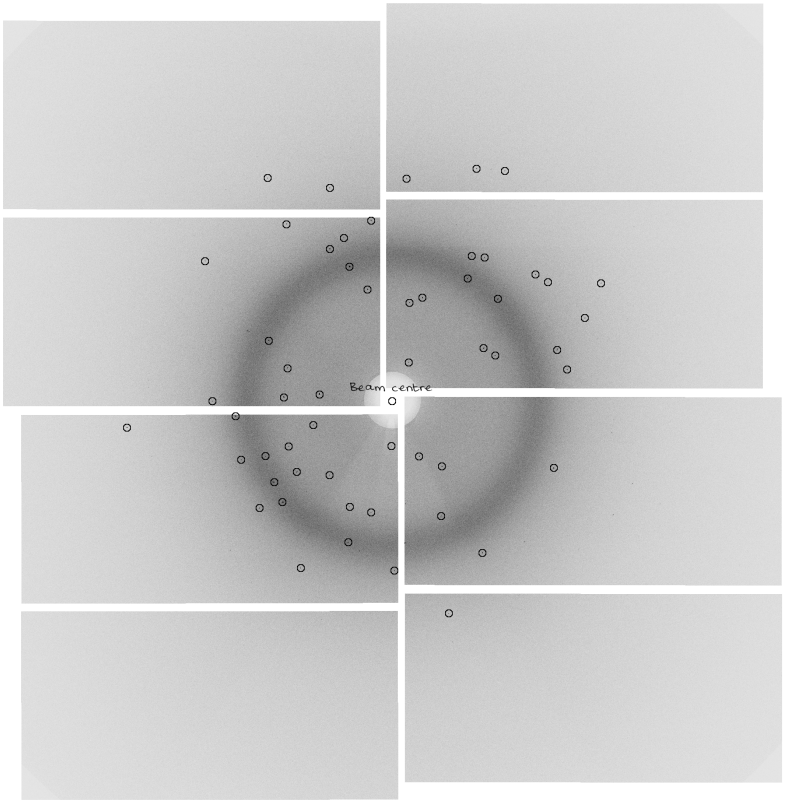


Figure 1.6: An example of a diffraction pattern from the MPCCD detector at SACLA. Diffraction spots are circled for clarity of differentiation from the background. The diffraction shown is from bacteriorhodopsin crystals which have been released on CXIDB (no. 53).

The software designed for processing synchrotron-derived data sets are poorly suited for handling the data from XFEL sources. Snapshot images and the way in which individual Bragg peaks are illuminated with no rotation angle are poorly handled by existing software, despite performing very well for even a small rotation data set. Hence, a few new software suites have been developed, most of which are open-source, which allow real-time online analysis of diffraction events and offline rigorous analysis which leads to merged data sets for structure solution.

The software development has been focused on several stages of the pipeline: online data analysis, detector correction, hit-finding, indexing, integration and merging. After merging, the data can be treated in the same way as synchrotron data with regards to phasing and structure refinement; and so falls back into the regular synchrotron data processing stream after this point.

1.5.1 Online data analysis packages

There are currently two major data analysis packages for getting immediate online feedback for observing diffraction results: OnDA by Valerio Mariani (Mariani et al., 2016) which interfaces directly with the DAQ at the LCLS XFEL and cheetah-dispatcher by Takanori Nakane (Nakane et al., 2016a) at SACLA XFEL. The former of the two has been adapted for use at the European XFEL. These programs are designed to provide simple, real-time feedback during the process of data collection but without producing results from in-depth analysis. Hence, the analysis run on the images must be able to keep up with the data acquisition rate (or limit itself to a subset of the collected images). As a result, these packages tend to produce hit rates in real-time, and may follow up with indexing with CrystFEL (below). Some form of real-time monitoring is essential during beam time. Even a simple "diffraction present or not present" analysis can be fairly helpful in directing the course of an experiment during run-time.

1.5.2 Background subtraction and hit-finding

Cheetah

Cheetah (Barty et al., 2014) is a software suite which interfaces with the *psana* data acquisition library at LCLS (maintained by Anton Barty) and has been modified for use at SACLA (maintained by Takanori Nakane (Nakane et al., 2016a)). I also aided in its modification to read data at the European XFEL from the AGIPD detector. It performs initial background subtraction, gain correction and performs peak-finding in order to identify hits (by determining if the number of peaks hits a certain threshold) and writes the non-empty images to disk in HDF5 format and include the peak-finding results. Cheetah writes the images to disk with no applied geometry (instead arranging the ASIC readouts from the CSPAD in a tiled format) and is well-integrated into the ecosystem of CrystFEL (see below), which then rearranges the images on disk into their appropriate positions using a specified geometry file format.

CASS

Lutz Foucar in Heidelberg has developed an equivalent software suite to the function of Cheetah (Foucar, 2016) but it is less well-used, probably due to having less documentation available than Cheetah.

cctbx.xfel

The early stages of processing using *cctbx.xfel* (which is discussed in more detail below) interfaces with *psana* and writes out images to a Python serialisation format (pickle file). The detector panels are rearranged when written to disk according to an initial baseline geometry (referred to as metrology by the authors) as determined at the beamline, and then additional modifications as required; however, if the geometry is revised then the pickles must be re-written to disk from the original streams at LCLS. The pickles can then be accessed from downstream software such as other modules of *cctbx.xfel*, *DIALS* and

PRIME, but this format is fairly incompatible with processing by CrystFEL.

1.5.3 Downstream processing: indexing and data reduction

CrystFEL

CrystFEL is a processing package created and developed by Thomas White in the group of Henry Chapman in the Centre for Free Electron Laser Science (CFEL) in Hamburg (White et al., 2012). CrystFEL is the software suite with a clean user interface, large volumes of documentation, tutorials and organised workshops, which has been instrumental in its placement as the most widely used software suite. However, due to its streamlined code and the way it has been created, it is more difficult to use as an experimental sandbox for trying new ideas for methods of data processing. At the same time, this simplicity is likely to be one of the major draws for users inexperienced with the theory of diffraction, which makes up (or will make up) a large percentage of the XFEL user base. CrystFEL uses the output of Cheetah (see Section 1.5.2) and performs peak-finding (if necessary to reprocess from Cheetah), indexing, integration and merging (with optional post-refinement). The current release of CrystFEL, v0.6.3, implements my TakeTwo indexing algorithm described in Chapter 3.

cctbx.xfel

cctbx.xfel is a software suite which is based on the Python *cctbx* framework which is developed and maintained by Nicholas Sauter. It was introduced in 2014 with the structure solution of thermolysin and lysozyme (Hattne et al., 2014). The *cctbx.xfel* indexing relies on underlying subroutines in *labelit* (Sauter et al., 2004), and has custom-written integration and merging routines. Until recently, *cctbx.xfel* took a controversial decision to reject negative intensities which resulted from integration, on the basis that this boosts the signal. However, significant pressure from the community was applied due to the effect of causing a systematic inflation of weak reflections. This produces inflated Fourier series

for the weak reflections which must be added to the map with a given phase, and refinement programs will generally add phases which support the purported atomic model, regardless of its accuracy. A recent paper with use of *cctbx.xfel* has included negative intensities (Colletier et al., 2016) stating that it increased the stability of refinement of the atomic model. The software is written in both C++ and Python, and the current focus of Nicholas Sauter's group is to transfer the important algorithms in *cctbx.xfel* to *DIALS*.

PRIME

PRIME is software dedicated to merging and post-refinement led by the group of Axel Brunger and mainly developed by Monarin Uervirojnangkoorn (Uervirojnangkoorn et al., 2015). It performs cyclic refinement using a best set of reference reflections, refining the crystal rotation, linear scale and B factor, reciprocal lattice point size and inflation with resolution and includes a partiality model. PRIME calculates its partiality based on the properties of the reciprocal lattice points and assumes a single Ewald sphere with which the reflections intersect. It was also used to perform post-refinement of intensities in the above paper (Colletier et al., 2016).

DIALS

Diffraction Integration for Advanced Light Sources (DIALS) (Waterman et al., 2016) is a software suite which is based jointly with Diamond Light Source Ltd. in the UK, and Nicholas Sauter's team in Berkeley, USA. This software package aims as a whole to create an open-source framework to process diffraction data from goniometer-based experiments and serial crystallography experiments at both synchrotrons and XFELs, with the former processing methods largely developed at Diamond, and the latter largely developed at Berkeley. DIALS was developed with the aim to provide an interface for new algorithms to be easily added to the framework, and welcomes collaboration from anyone who wishes to contribute to the project. Like *cctbx.xfel*, it is based on the *cctbx* framework and written jointly in Python and C++.

nXDS

nXDS (nanoXDS) is a software suite written by Wolfgang Kabsch for XFEL data analysis, but has not been made publicly available in either source code or binary executable. It is a modification of XDS Kabsch (2010), which is a software suite suited to synchrotron data. Two structures have been reported which used nXDS, one at a synchrotron source (PDB code 4RLN) and one at an XFEL (PDB code 5EJX). The former is a study on room temperature serial crystallography data from a synchrotron source of lysozyme crystals, and reports processing results from CrystFEL and nXDS; their performance is largely similar and both capable of supporting single/multiple isomorphous replacement and anomalous scattering (SIRAS or MIRAS). The latter is a rotation-data set study at an XFEL reporting data processing only from nXDS. A helical scan was used to expose additional portions of the same crystal to the beam for a total of nine crystals.

1.6 Structural studies using the XFEL

1.6.1 Early XFEL structures

Several XFEL structures were published in high impact journals after the LCLS became open to user experiments in 2009. These structure solutions were embarked upon before the full extent of XFEL data processing difficulty was understood; and due to the poor modelling of parameters per image, early data processing used a method which was named the 'Monte Carlo' method (although this method is a misnomer as it does not truly fit the definition of 'Monte Carlo' integration). This assumes that with enough observations, all the various error sources will eventually cancel out and produce a value proportional to the true value with respect to all the other reflections. The first structure determination reported introduced this as a concept in the structure solution of Photosystem I (Chapman et al., 2011). This was of considerable impact because it was a proof-of-principle that structure determination was physically possible from XFEL data.

Being the first structure reported in the literature, the focus was not on efficient data processing: three million diffraction patterns were collected of which 13% were indexed and integrated to create a structure to the detector edge (8.5 Å). More notably, some exceptional results were determined from aspects of the coherent diffraction (see Section 1.6.7).

Cathepsin B is a cysteine protease from the trypanosome parasite *Trypanosoma brucei* which is involved in degradation of host proteins for reuse of the amino acids. This protein is essential for the parasite's survival, and the parasite itself causes Chagas disease and sleeping sickness in humans. The crystal structure of cathepsin B was determined at a synchrotron, but the room temperature structure from an XFEL was reported in 2013 (Redecke et al., 2013) using small needle-like crystals with a minimum dimension of 0.9 μm. There were measurable differences between the cryo-cooled and room temperature structures: in particular a loop which is proteolysed during protein maturation is disordered in the room temperature but more ordered in the cryo-cooled structure. This highlights the importance of collecting data at room temperature: in other cases, the most flexible regions of the protein may often confer the functional properties, but cryo-cooling the crystals forces these regions into a smaller subset of the most stable conformations.

1.6.2 Lysozyme and other model systems

Lysozyme was an early target for structure solution by XFEL crystallography and the structure by molecular replacement was reported in 2012 (Boutet et al., 2012) using very small crystals (1 by 1 by 3 μm). Lysozyme is a frequent model system, chosen due to easy purification in large quantities from eggs and the high propensity for lysozyme to crystallise; in reality lysozyme crystals are actually very heterogenous, varying a lot with their level of hydration and temperature, which is concurrent with changes in the unit cell axes. There are some reasons to be concerned about the choice of unit cell dimensions for lysozyme (Figure 1.7).

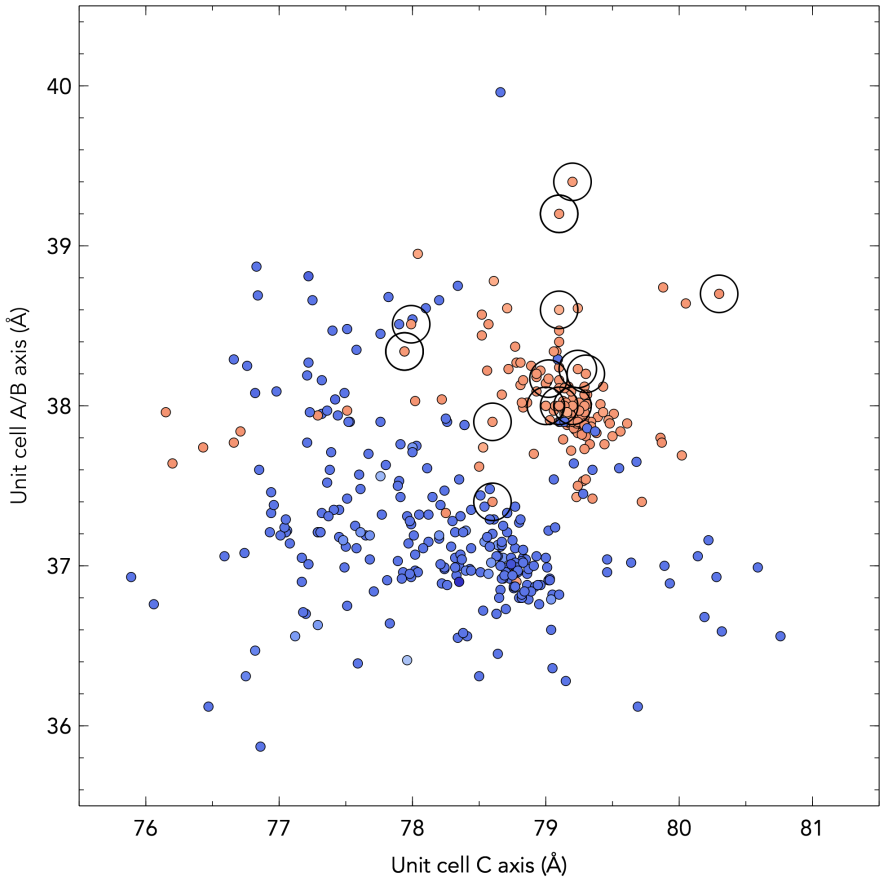


Figure 1.7: All 523 reported lysozyme structures on the PDB within the suitable unit cell dimensions for space group $P4_32_12$. The two main clusters (bottom left, top right) appear to correspond to the temperature of the data collection (plotted from blue at 100 K to red at 298 K). Circled are the structures for lysozyme solved at XFELs, 60% of which are outside of the range ($37.7 < a < 38.7$ & $78.7 < c < 79.7$), compared to 26% of other structures above the melting point of water. The other outlier was a specific examination of the effect of relative humidity on the lysozyme structure. The non-isomorphous nature of lysozyme can make it very difficult to differentiate between various types of errors. Errors in assumed unit cell dimensions can be corrected for by other uncertain parameters (detector layout, detector distance), which are highly correlated with errors in the unit cell dimensions. Although good for guaranteed crystal growth, it may not be the best choice of model system for XFEL analysis.

The ability of XFEL diffraction data to reveal anomalous signals in the intensities across Friedel pairs was established (Barends et al., 2013) two years after the publication date of the first static structure, where the anomalous signal of the sulphur atoms from methionine and cysteine provided real-space anomalous signal peaks in combination with the known phases of lysozyme. Using a stronger anomalous scatterer, gadolinium ions, soaked into the lysozyme crystals, a more easily measurable anomalous difference could be established of 5-15%. This resulted in a landmark study with a protein structure phased using the anomalous differences between Friedel pairs for the first time at an XFEL (Barends et al., 2014).

Another key model system used by the Berkeley laboratories is thermolysin, an extremely heat-stable and robust protein which forms hexagonal crystals and the structure has been reported along with another lysozyme solution in one paper (Hattne et al., 2014).

1.6.3 Radiation-sensitive samples

One of the major systems which has generated considerable interest with regards to studies using XFELs is photosystem II (PSII). Because PSII is the only molecular machine capable of splitting water using visible light, it is considered important to study because it may generate a template from which synthetic water-splitting machines may one day be created. PSII consumes water, evolving oxygen and simultaneously mobilising protons and electrons. The protons are balanced out electrostatically by magnesium ions, resulting in a chemical gradient which is capable of driving ATP synthesis. The electrons reduce plastoquinone, a membrane-bound co-factor, and are transferred to soluble plastocyanin for use in the RuBisCo cycle. The oxygen-evolving centre of PSII is a manganese cluster with four Mn ions, one Ca ion and five oxygen atoms. PSII has been shown to cycle through five distinct states, the indication for which is evolution of oxygen only on every 5th step of each cycle. These are labeled states S0 - S4, with S1 being the 'dark' ground state, and oxygen evolution on the transition between S4 and S0. Studies of PSII at synchrotrons have yielded structures from 3.8 Å (Zouni et al., 2001) to 1.9 Å resolution (Umena et al.,

2011) ten years later. However, these have been plagued by the radiation damage which is sustained while attempting to collect data with high signal-to-noise at a synchrotron. This is evident from the bond lengths present in the structures between manganese ions, which do not agree with those measured previously using extended X-ray absorption fine structure (EXAFS) and which are elongated in radiation-damaged samples.

Photosystem I (PSI) was immediately tackled with XFELs at their advent, with one of the first structures being reported (Chapman et al., 2011). Not long after, diffraction from PSII crystals at the XFEL were reported at room temperature (Kern et al., 2012). The structure reported had poor processing statistics, but is merely indicative of the infancy of XFEL data processing at the time. This was followed by a paper reporting the S1 (dark) state and the S2 state limited to 4.1 Å resolution (Kern et al., 2013). The following year, the same authors report the states between S2 and S3, and between S3 and S4 to similar resolutions (Kern et al., 2014). These are transition states which cannot be captured with a conventional synchrotron experiment due to the timing of the reaction, regardless of the presence of radiation damage. However, there are significant difficulties with interpreting amplitude information to such low resolution due to the poor maximum phase quality which can be obtained. A different study of one of the intermediates was performed at SACLA with data collected to 1.95 Å resolution (Suga et al., 2015) using several diffraction patterns per crystal collected with spatial separation. This study reports a set of bond lengths which agree more closely with the EXAFS data, giving evidence that XFEL studies do indeed manage to outskirt at least the vast majority of radiation damage.

1.6.4 XFEL-based GPCR structures

A number of technical difficulties are encountered when collecting data from GPCR crystals at synchrotrons. GPCR crystals usually are limited in size when grown in the currently preferred medium, LCP, and GPCR crystals which are less than 10 μm in a given dimension are very difficult to collect diffraction from, even with the best synchrotron microfocus beamlines. However, these crystals may be amenable to XFEL studies, due to the increase

in photon flux. Several GPCR structures have thus been reported to date.

The serotonin $5 - HT_{2B}$ receptor is a GPCR which produced crystals exposed to the beam at LCLS of $5 \times 5 \times 5\mu$m, which is a size below that reportedly possible to analyse at a synchrotron. However, a different synchrotron structure was available from a larger crystal form at cryo-cooled temperatures. This structure solution was reported to 2.8 Å collected from room temperature diffraction (Liu et al., 2013). Although absolute B factor comparison will be dominated by the differences in processing, a relative comparison of the B factors between the cryo-synchrotron and RT-XFEL crystals showed significant disparities especially in the intra-cellular and extra-cellular loops, which showed greater movement in the RT-XFEL structure. There was also a comparative reduction in the core of the GPCR (the seven transmembrane-helix) compared to the synchrotron structure. These differences are likely to be ascribed to the thermal motion at room temperature, but may also result from radiation damage to the synchrotron sample. Alternatively, some processing error may be responsible for this. This could be addressed by comparing the fits of each model to each data set and allowing an application of a global B factor to account for residual scaling error. If the differences in the B factors can explain the differences in amplitudes between data sets, this would provide weight to the theory that the B factor differences are genuinely caused by real-world differences between the two crystals.

Structural information from XFEL studies has been used to guide structure-based drug design against GPCRs. One of these studies into the angiotensin receptor type 1 (AT_1R) uses XFEL methodology to provide a novel structure bound to a drug precursor (Zhang et al., 2015). The structures of angiotensin receptors in complex with drugs known to bind to them proved elusive, but this paper reported a complex between AT_1R and a stronger-binding hypothetical drug precursor (ZD7155). Mutagenesis studies confirm four key amino acids which confer the drug's specificity to the angiotensin receptor directly and this structure is expected to guide drug design for treating both hypertension and auto-immune diseases. Another study of the smoothened receptor, a validated drug target in

cancers, has used a combination of thermal shift assays to identify a potential ligand and XFEL studies to guide drug design to a more stable version (Zhang et al., 2017). This drug has been entered into Phase II clinical trials.

Arrestins are binding partners of GPCRs, blocking downstream signalling to intracellular components. Crystals containing rhodopsin in complex with arrestin proved to be a challenging case (Kang et al., 2015), as the crystals were reported to be physically twinned. However, it was still possible to solve with molecular replacement. This structure revealed an interaction point between rhodopsin (transmembrane loop 7 and final helix 8) and a cleft which correspondingly opens up by 20° on a hinge between the N-terminal and C-terminal domains of arrestin.

1.6.5 Difficult anomalous phasing experiments

One of the concerns after the initial reports of Monte Carlo integration (Chapman et al., 2011; Redecke et al., 2013) was whether XFEL data would be capable of supporting phasing *de novo*. Initially, this question was answered using the gadolinium-phased lysozyme structure (Boutet et al., 2012), but more subtle phasing experiments were needed to be shown to be possible. This is not only for the purpose of phasing, but also in effect asking whether XFEL is capable of deriving the accuracy of data necessary for certain types of experiments, including the extraction of time-resolved changes. Hence, a number of papers have been dedicated to showing the capabilities for XFEL to phase using a regularly achievable number of diffraction patterns and also off the native anomalous signal from sulphur atoms. Note that LCLS is incapable of achieving the wavelength required for selenium phasing, but has been shown to be possible at SACLA where the wavelength is achievable (Yamashita et al., 2015). In addition to the study already described in Section 1.5.3 (Botha et al., 2015) which showed phasing by MIRAS and SIRAS, another study has reported SIR and SIRAS with minimal numbers of images (Nakane et al., 2016b) from iodo-detergents bound to bacteriorhodopsin, and also single-wavelength anomalous dispersion (SAD) using the native sulphur atoms in lysozyme (Nakane et al., 2015).

Sulphur SAD has also been shown for thaumatin crystals (Nass et al., 2016). These studies show the feasibility for a number of model systems, however for the methods to be generally useful it is likely that additional work will be needed to reduce the volume of diffraction data necessary to perform these calculations.

1.6.6 Time-resolved studies using XFEL radiation and the use of difference maps.

The suggestion to perform time-resolved studies using femtosecond X-ray pulses was put forward some time ago (Neutze et al., 2000) and recently reviewed by the same lead author (Neutze and Moffat, 2012) in the context of other time-resolved techniques as well. As XFEL was in its infancy at the time that Neutze and Moffat (2012) was published, a lot of the challenges that they specify to be a problem with time-resolved experiments are also problems with static serial femtosecond crystallography as well. The problem of partiality (which I address in Chapter 5) is mentioned and the need for high-multiplicity data. This is mentioned in the context of the static structure solutions which had been published up to that point (Chapman et al., 2011; Boutet et al., 2012; Kern et al., 2012). The major pull to perform XFEL experiments for time-resolved studies is to achieve femtosecond-timescale resolution of chemical events.

At times, even when the full potential for time resolution is not exploited, performing experiments at XFELs is necessary to bypass the concern of radiation damage. The sequence of events in bacteriorhodopsin was analysed using XFEL radiation, which take place on the timescale of nanoseconds (Nango et al., 2016). Three-dimensional images of bacteriorhodopsin were taken in the dark state, 16 ns and 290 ns after a photon-induced proton transfer. Difference maps were used to obtain the changes in the structure as a result of the changes in amplitudes. The chemical changes taking place in carboxymyoglobin when triggered by a photon to release the bound carbon monoxide do, however, take place on the femtosecond timescale, and has been studied with very clear changes (Barends et al., 2015). Difference maps also highlight changes taking place in the pho-

toactive yellow protein (Tenboer et al., 2014).

One thing to note in these time-resolved studies is the use of difference maps, taking the observed amplitudes from two timepoints of the same structure and combining these with the starting phases. A difference map may contain up to half of the full information about the differences between the two structures depending on whether the difference structure factor is in- or out-of-phase with the starting structure factor. Although it will be diluted by the remaining phase error, it is more powerful than attempting to view the structure using macromolecular modelling alone due to the consistency of those phase errors.

1.6.7 Coherent X-ray diffraction

One of the earliest ideas behind the X-ray free electron laser was its ability to image from single particles using coherent diffraction only. This has turned out to be extremely technically challenging as the number of photons scattered is particularly minimal from single particles. X-ray diffraction could be seen to exist as a spectrum from fully crystalline diffraction where all the diffractive power is funnelled into the Bragg peaks, to the opposite extreme of coherently imaging single molecules to obtain a continuous molecular transform. There has been substantial progress also in the middle of this spectrum, where crystals become small enough to observe inter-Bragg diffraction in the form of fringe peaks which supplement the rest of the Bragg information.

The first structure reported (Chapman et al., 2011) showed the power of using the coherent diffraction around the Bragg peaks to determine the shape of the crystal from the Miller indices (1, 0, 0), (0, 1, 0) and (0, 0, 1). The number of 'fringes' between the Bragg peaks (caused by the beginning of the breakdown of the convolution from the crystal and the protein itself) is equal to the number of unit cells present along the beam axis, and a number of diffraction patterns were recorded where this number was minimal enough to be seen (one particular diffraction image reported demonstrated only nine unit cells along the beam axis).

The mimivirus is the fourth largest virus discovered to date, with a capsid diameter of 400 nm and all copies of the virus are suitably similar. This particle has been three-dimensionally reconstructed from single particle X-ray free electron laser data (Ekeberg et al., 2015). This is extremely technically challenging due to the small number of scattered photons, the ambiguity in determining orientation (due to lack of fiducial markers as are produced by protein crystals), and the necessity to detect multiple hits and exclude them. The reconstruction was estimated to be meaningful to a resolution of 125 nm. Compared to the diameter of the mimivirus, this is a reflection on the difficulty of reconstructing such particles from XFELs with the current state-of-the-art. For comparison, a cryo-electron microscopy of another giant virus, Cafeteria roenbergensis virus, reached a resolution of 2.1 nm (Xiao et al., 2017).

Another method of extracting structural information from non-Bragg peaks has been pioneered by Chapman (Ayyer et al., 2016), using PSII as a test case. These crystals, in space group $P2_12_12_1$, exhibit crystalline disorder which is purported to be largely translational. If the disorder in a crystal is driven by translational movement of proteins as rigid bodies, rather than containing a substantial rotational component, then the nature of the Fourier transform will lead to faint coherent scattering between Bragg peaks which is equal to that of the Fourier transform of the protein. This has been exploited in an attempt to extend the resolution by incorporating the coherent scattering which lies around and beyond the measurable Bragg peaks (Ayyer et al., 2016). Use of the Bragg peaks to align the different orientations aids the process of merging the coherent data. This is largely convincing, but suffers from some drawbacks: due to the rotational symmetry in the 222 point group, the coherent background scattering has four-fold averaging, making it almost impossible to solve from an *ab initio* model. Instead, an envelope calculated from 10 Å phases of the PSII was used to re-derive the phases. These caveats, and the lack of an *ab initio* structure solution, has made the results in the paper the subject of some contention. Unfortunately it is not possible to easily refine against such inter-Bragg scattering using existing software, and so was not reported in the paper. The scaling

between Bragg peaks and the coherent diffraction in the background adds a significant technical challenge in the art of reporting R factors for any attempted refinement. As a result, density improvement was driven by cyclical phase extension, constraining to a molecular envelope and the knowledge that all electron density should be positive values. The authors report an improvement in the shape and continuity of side chain density, including a better fit for the non-ring elements of the chlorophyll electron donor.

1.7 Focus of this thesis

In the course of my DPhil in Structural Biology I have tackled a number of problems in the computational methods to relieve the pressure in the data processing bottleneck for XFEL crystallography. The resulting algorithms have been released in a software suite *cppxfel* (Ginn et al., 2016a). This impacts four stages in particular in the data reduction pipeline: geometry refinement (Ginn & Stuart 2017, accepted), indexing (Ginn et al., 2016c), initial orientation refinement (Ginn et al., 2015a) and post-refinement (Ginn et al., 2015b). Four individual chapters focus on and provide additional introductory material for each stage.

Geometry refinement was a mostly unforeseen problem at the start of XFEL: the expectation was that the detector would have good calibration before crystal diffraction data was processed; the reality was that the expected detector segment positions provided at the various beamlines were rarely accurate. Geometry refinement also turned out to be a trickier problem than first thought, and I have written up my own contribution to the problem in this thesis in Chapter 2.

This is followed by my contribution to the algorithms for indexing. The indexing rates (diffraction patterns indexed as a proportion of the number of hits) for XFEL structure solutions were extremely low and have slowly increased over time. Increasing indexing rates n-fold provides a $\sqrt{n}$ increase in the signal over the noise, and so there is an element of diminishing returns for improving the indexing rates. Nonetheless, since crystal supply, plus cost and labour of production, is limiting in many important structural problems,

increasing the indexing rate can aid a useful reduction in the actual number of images necessary to be collected. I have written up my contribution to the indexing problem here in Chapter 3, an algorithm called TakeTwo, and often manage to exceed a 100% increase in the original indexing rate for various data sets, and so produce a non-negligible improvement in the signal-to-noise ratio.

Orientation matrices directly taken from indexing solutions do not fully exploit the information on the images to maximise their accuracy. Solutions derived from the first generation of XFEL software were extremely far off the true result and those generated from the TakeTwo indexing algorithm were also quite divergent. Hence I outline a protocol in Chapter 4, termed initial orientation matrix refinement (renamed 'prediction refinement' when incorporated into CrystFEL), for extracting more information on the image to align the orientation correctly. This is an important step as it reduces the radius of convergence for post-refinement.

After performing an initial merge of the reflections, post-refinement takes the merge as a reference data set and uses it to refine the parameters of each of the crystals used in the merge. This exploits the power of knowing approximate intensities for the reflections, which can then be used to improve the definition of each crystal's orientation, reciprocal lattice point size and the wavelength of the X-ray beam. This, combined with a model for reflection partiality, is the focus of Chapter 5.

I include an example of these elements in practise by showcasing one of the structures which was solved using the combination of all of these algorithms (bovine enterovirus type 2) in Chapter 6. However, additional structures of proteins which have also been published have used various elements of the pipeline as well (Ginn et al., 2015a,b; Oghbaey et al., 2016).

Chapter 2

Detector geometry refinement.

2.1 Note to the reader

We wish to thank Thomas A. White for providing an initial geometry estimate and initial processing with Cheetah. We thank Helen Duyvesteyn for hit-finding with Cheetah for additional CPV17 runs. The contents of this chapter have been adapted from an associated publication (Ginn and Stuart, 2017).

2.2 Introduction

2.2.1 The problem of detector geometry

A key part of data processing for any diffraction experiment, including X-ray diffraction, is determining the parameters of the experiment: those of the sample (e.g., unit cell dimensions, crystal size, mosaicity), those of the probe (e.g. the beam size, profile, divergence, energy and spectral distribution), and the recording device (e.g. detector gain curve, angle relative to beam, placement relative to sample, segment placements relative to each other). Due to well-characterised detectors present in most synchrotron beamlines, researchers benefit from very good calibration of these detectors, including gain

and precise knowledge of the placement of the sensors with respect to each other.

The machines with the most brilliant hard X-ray sources are now X-ray free electron lasers: the Linac Coherent Light Source (LCLS), the SPring-8 Angstrom Compact Free Electron Laser (SACLA) are already in use and the European XFEL (EuXFEL) and SwissFEL have begun lasing. These are accompanied with custom in-house detectors, which provide the basis for technological advances necessary to cope with the demands of recording XFEL data, including: the Cornell-SLAC hybrid Pixel Array Detector (CSPAD) (Hart et al., 2012) at LCLS, the MultiPort Charged Coupled Device (MPCCD) (Kameshima et al., 2014) detector at SACLA, and the upcoming Adaptive Gain Integrating Pixel Detector (AGIPD) (Henrich et al., 2011) and the Jungfrau detector (Mozzanica et al., 2014) at the EuXFEL and SwissFEL respectively. The former two detectors are widely in use at their respective light sources. These detectors are all segmented, and are being upgraded frequently. The CSPAD in particular is a 64-segmented detector of Application Specific Integrated Circuit (ASIC) modules which are bump-bonded into pairs. This detector requires regular dismantling and rebuilding, which inevitably changes the relative positions of the detector segments. This is an important part of the technological development of these detectors. However, the frequent (and unknown) rearrangement of the detector panels has a pronounced effect on the ability to process the data. The resulting panel shifts are largely unknown, and have to be back-calculated from data collected at the XFEL. For the regular crystallographer, the careful calibration of commercial detectors at synchrotron beamlines may have allowed this aspect of detector calibration to drift to the back of the mind.

X-ray free electron laser data provides the biggest challenge for back-calculation of detector geometry. At an XFEL, crystals enter the beam with an unknown orientation, and a portion of X-rays entering a crystal will satisfy the Bragg condition and diffract to produce a measurable spot on the detector. A crucial stage of the data analysis workflow is working out the relative positions of the putative reflections which gave rise to these spots. This dictates how X-rays are traced from spot to sample and determines the

intersection with the Ewald sphere, a critical component of determining crystal orientation. The geometry is also essential for predicting the positions of the diffracted X-rays in preparation for integrating reflection positions. Even a sub-pixel offset in the assigned position of a panel from its reality will disturb the integrated intensities.

Geometry refinement calculations have been carried out by a number of groups. Work in early phases of XFEL data processing noted the requirement for further detector geometry refinement from the initial parameters, being incorporated into *geoptimiser* (Yefanov et al., 2015), a stand-alone software package associated with *CrystFEL* (White et al., 2016), and a geometry (also termed metrology) refinement module in *cctbx.xfel* (Hattne et al., 2014). Aaron Brewster has incorporated a related form of the DIALS detector geometry refinement in the DIALS framework for goniometer-based experiments (Waterman et al., 2016) and this has been adapted to exploit XFEL data (Brewster et al, manuscript in preparation).

The first two XFEL-specific methods both concentrate on the challenging arrangement of the CSPAD detector. The methods of geometry refinement by Hattne et al. (2014) were carried out using an initial GUI-based manual adjustment of the summed-image powder rings to get an initial starting solution, which is also performed automatically using *geoptimiser*. Both of these reported methods then refine the positions and rotations of all ASIC pairs on the CSPAD, the orientation and unit cell dimensions for each crystal, and the beam translation and sample-to-detector distance by least-squares refinement to minimise the discrepancy between observed and predicted spot positions from known indexed lattices. On the other hand, Yefanov et al. (2015) refine three parameters per individual panel: translation in X and Y (axes perpendicular to the beam), and a rotation around the centre of the panel, all kept in plane with no variation in the Z axis. Although this appears to be a less ambitious approach, because certain combinations of parameters can look remarkably similar, refining fewer parameters prevents one parameter from compensating for errors in another correlated parameter. As a result, this method may be advantageous against the chosen target function.

Detector geometry refinement has also been incorporated into DIALS, which aims to refine geometry for both single-shot experiments as per XFEL experiments and goniometer-based rotation experiments at synchrotrons. This is a more ambitious project which includes fully three-dimensional detector models with an optional hierarchical structure, able to model the panels arranged on a cylinder of the Pilatus 12M detector at Diamond, beamline I23 (Wagner et al., 2016). In the DIALS framework, during refinement, panels are rotated with respect to their basis vectors, and also translated along these three vectors (which are not necessarily parallel to laboratory framework axes). However, several combinations of parameters defined in this model are still correlated, which significantly complicates the process of refinement.

This chapter presents a new method of parameterisation of a detector in three dimensions which aims to deconvolute the vast majority of the correlated parameters. These can be split into two major modes of movement: Euclidean movements which affect the relative arrangement of putative reflection coordinates in reciprocal space, and non-Euclidean movements of spherical geometry which rotate a panel around the sample, but do not affect the relative arrangement of putative reflection coordinates (Figure 2.1). This has been termed the slip-and-slide algorithm, as one first 'slips' into the correct orientation relative to the sample (Euclidean movements), and then 'slides' around the sample (non-Euclidean movements). These can be refined against different target functions. Refining these movements separately prevents a one-step movement in the X axis, and must be decomposed into their component modes (Figure 2.2). In addition, we present two types of target function: those which use the traditional approach of comparing observed to predicted pixel positions of diffracted rays through illuminated Bragg peaks, and another function which takes advantage of the "unit cell fingerprint" of the lattice-derived pattern of spots on the detector, with no concern given to the orientation of the lattices.

The effect of this method of geometry refinement on even a high quality initial data set is demonstrated by observing the effect on an independent parameter (anomalous signal from sulphur), which should indicate the relative importance of geometry calibration. This

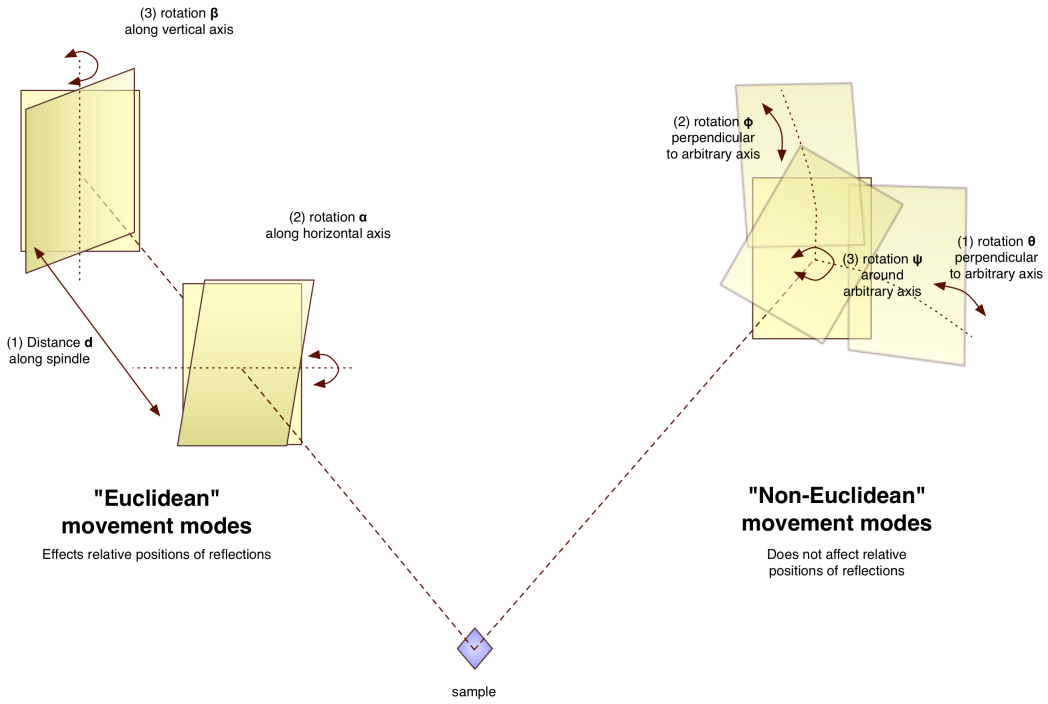


Figure 2.1: Types of movement: Euclidean (left) and non-Euclidean (right). Left: the detector panel can move backwards and forwards on the axis connecting its midpoint to the sample by distance **d**, and rotate by two angles on two axes orthogonal to the connecting axis (α and β). This will affect the predicted spot locations (or conversely, it will affect the relative arrangement of back-projected rays). Right: the detector panel can move around polar coordinates θ, ϕ and ψ around any arbitrary axis passing through the sample. This will not affect the relative arrangement of ray projections back onto the Ewald sphere, only their absolute position relative to the beam centre.

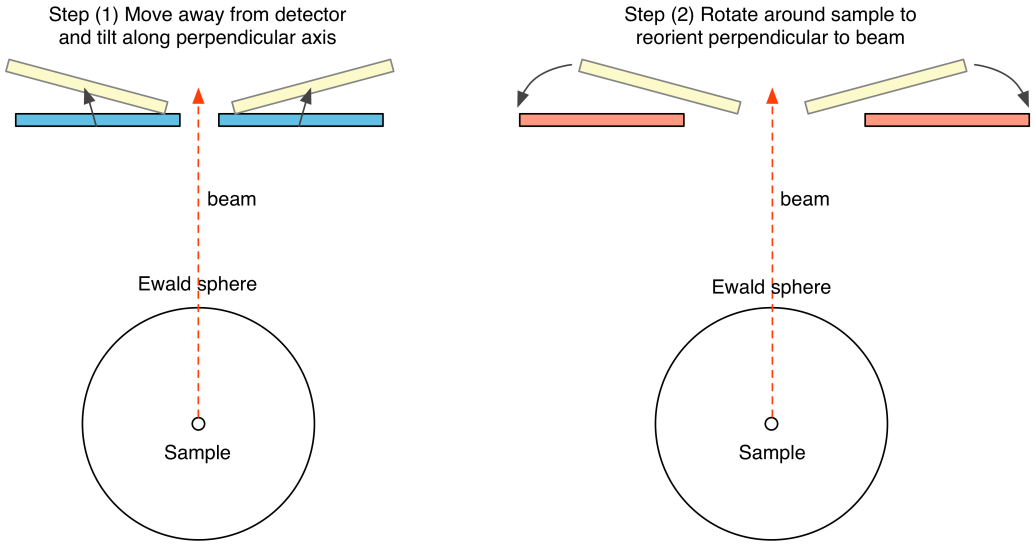


Figure 2.2: A lateral movement in panels must be split into the Euclidean elements (step 1, left) and the non-Euclidean elements (step 2, right). The former aligns the spot projections or predicted reflection positions within a single panel. The latter aligns the panels relative to the beam centre or each other.

corroborates the message from a study on sulphur SAD phasing of thaumatin crystals (Nass et al., 2016) which found that applying further geometry correction to gadolinium-phasing of lysozyme (Barends et al., 2014) could reduce the required patterns from 60,000 in the original study to 7,000.

2.3 Methods

2.3.1 Selection of spots from images by the droplet algorithm.

Potential reflection positions were identified using a novel algorithm (see below) which does not depend strongly on the statistics of the CSPAD being well-behaved. An initial theoretical 5-by-5 pixel model for a spot was created. The outer ring of pixels was set to a minimum value (arbitrarily set to 10 counts) and the central pixel was set to the maximum value (set to 100 counts) with a linear ascent from the outer ring to the middle pixel. This was then placed on an additional buffer of two minimum-value pixels to create a 9-by-9

model with which candidate spot positions on the image data could be compared.

The diffraction image was then searched and pixels which reached a minimum count threshold were considered as potential spots. For each of the potential spots, a correlation coefficient was calculated between the counts of the theoretical spot model and the counts from the data, setting the central pixel of the model as the considered pixel. If this correlation coefficient exceeded a threshold then this pixel was considered to be a spot. The use of a correlation coefficient means the gradient (correlated with the intensity of the spot) is almost irrelevant, and as such, searched for 'blob-like' structures rather than using signal-to-noise ratios. This does have a tendency to reject oblong or mis-shapen spots, and a manual inspection should confirm that the parameters are permissive enough not to reject too many non-ideal spot shapes.

In some cases, the correlation threshold used was not a constant value, but scaled inversely with the intensity, such that higher spot intensities could be accepted despite having a lower correlation threshold.

2.3.2 Converting to reciprocal lattice point positions.

Relative reciprocal lattice point positions were generated with the knowledge of the experimental setup. In this case, a simple model was used: a detector which lies orthogonal to the beam on both axes at a certain detector distance d, with a specified beam centre b_x and b_y, and using an X-ray wavelength of λ. The relative positions of the reciprocal lattice points producing the spots on the detector were calculated approximately by back-projecting the positions of spots on the detector onto the Ewald sphere, as it was known that these reflections must satisfy Bragg's law (Figure 2.3). More complex detector geometry employs ray-tracing to determine relative reflection coordinates.

2.3.3 Inter-spot vectors from images and idealised unit cell parameters.

Difference vectors between reciprocal lattice point positions were calculated up to a certain maximum separation in Å^{-1}, which in this case was set to 0.15 Å^{-1}. This cutoff was

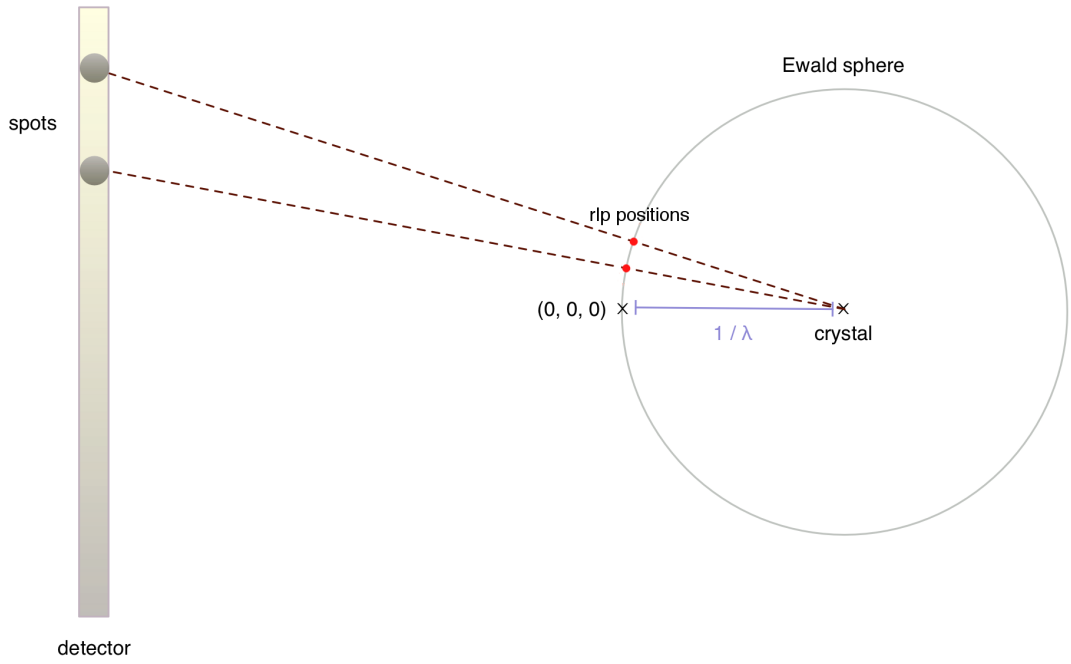


Figure 2.3: The projection to reciprocal space uses the knowledge of the X-ray wavelength (to determine the Ewald sphere) and the initial detector geometry (detector distance, relationship between individual panels etc.) to reconstruct the experiment. This allows the relative intersection points between the X-rays and the Ewald sphere to be determined, to create inter-spot vectors (boxed in red), although it cannot alone determine the absolute positions (i.e. the indexing solution).

named the maximum reciprocal distance. The length of these vectors were calculated and stored for quick access to speed up calculation time. Inter-spot vectors were also generated for the given space group and unit cell dimensions with which indexing was to be attempted, by considering all vectors originating from (0, 0, 0) up to the maximum reciprocal distance, excluding non-axial systematic absences. These were the idealised vectors for the given crystal symmetry. The lengths of these vectors were also calculated and stored.

2.3.4 Data input.

Initial geometry was loaded from the CrystFEL geometry file format for the CSPAD (at LCLS) into the *cppxfel* suite. Files in the CXI format (using the HDF5 specification) were created using *Cheetah* (Barty et al., 2014) and can now be directly loaded into *cppxfel* (Ginn et al., 2016a). CPV17 diffraction patterns from a prior study (Ginn et al., 2015a) *Cheetah* hit-finding results (data deposited in CXIDB (Maia, 2012), code 29) were used, from LCLS proposal LS06. Note that for the CSPAD, the bump-bonded ASIC pairs were treated as a single unit.

2.3.5 Detector geometry model.

The underlying model of the detector geometry is no more complex than that implemented in DIALS. The laboratory basis vectors were chosen such that the Y axis pointed to the ceiling, the Z axis along the beam from sample to detector, and the X axis in the appropriate direction to complete a right-handed orthogonal coordinate system. The origin was defined as the sample position. Each segment of the detector stored three properties: its midpoint coordinate, and the direction vectors which specify the fast-scan and slow-scan axes (from which the orthogonal vector may be derived). The 64 individual ASICs of the detector were paired together to form 32 ASIC pairs. Individual ASICs first inherited the properties from these higher-order ASIC pairs, before applying further

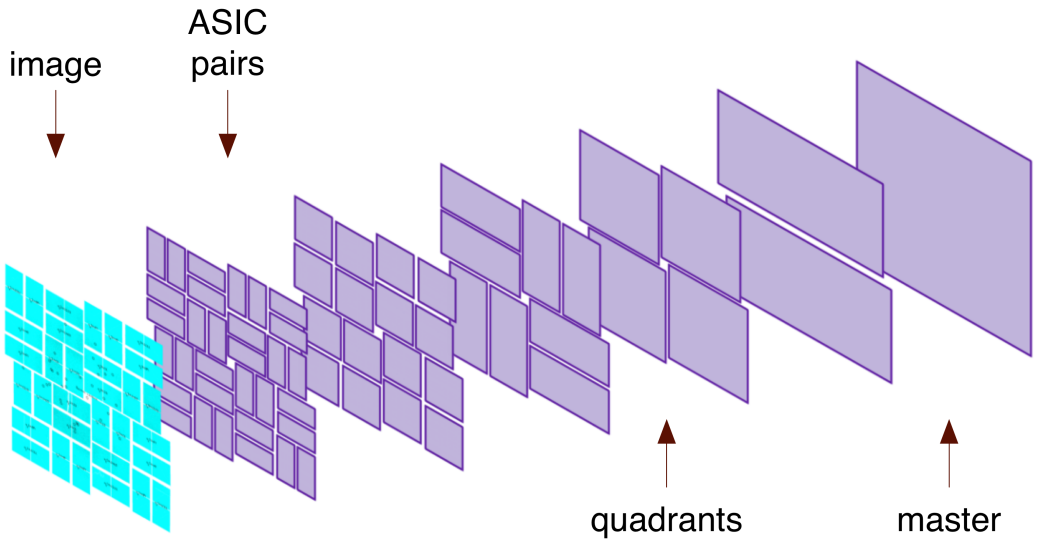


Figure 2.4: Detector hierarchy for a CSPAD image (on left). Subsequent images show the pairs of detector groups joined in a higher-level group from the previous image, finishing with the final 'master' group. Each group will inherit the properties of the parent group recursively. This enables blanket changes to be made to the entire detector or specific sections of it, and also aids the process of refining geometry from unindexed diffraction patterns.

corrections. These 32 ASIC pairs were paired recursively, ending with the definition of a single 'master' detector object (Figure 2.4).

During the course of refinement, small offsets are applied to the detectors in order to minimise the target function, which are applied on-the-fly. At the end of each cycle of refinement of a detector panel, these small offsets (rotation angles and translations as defined in Section 2.3.6) are absorbed into the direction vectors which specify the fast- and slow-scan axes. The small offsets are then set to zero before the next round of refinement. This is merely a detail of the implementation.

2.3.6 Panel movement modes.

Movement modes were divided into Euclidean and non-Euclidean movements. The three Euclidean movements are: (a) movement back and forth along a connecting axis connecting the centre of the panel to the sample, and (b) a rotation along an axis normal to the

connecting axis, and (c) a rotation around the other orthogonal axis normal to the connecting axis. There were three types of non-Euclidean movement, polar coordinates θ, ϕ and ψ around any axis passing through the sample (Figure 2.1).

2.3.7 Refinement of unit cell dimensions

To refine unit cell dimensions independently from detector geometry, a combination of over-prediction and a local search of the diffraction image for the highest peak, followed by integration, could "catch" reflections which were not necessarily assumed to be satisfying the Bragg condition using the parameters of the initial model. This identified illuminated reflections which were located further afield from the initial prediction, which was followed by initial orientation refinement as per Chapter 4. Reflections were registered as either strong or not strong, and thus provide data which were independent of the diffraction geometry. Global unit cell refinement was then performed to bring the unit cell dimensions in line with what was required to focus the reflections around the mean wavelength. Iterations of orientation refinement followed by unit cell refinement were performed until convergence. This involves feedback from work involved in Chapters 3 and 4.

2.3.8 Geometry refinement using indexing solutions.

Diffraction patterns of CPV17 run 4 were indexed using an initial estimate of the detector geometry. The initial offsets in this file between the predicted and observed locations for the vast majority of ASIC pairs were within 1-3 pixels, which was also affected by the unit cell and distance correction described above. Indexing was performed using the TakeTwo algorithm (Ginn et al., 2016c), including solution validation. During indexing, the surrounding area up to five pixels away from the predicted peak was searched, and the highest pixel value stored as an offset. The offset was stored in both pixel coordinates and in reciprocal coordinates when back-projected onto the Ewald sphere. This local search was also used during the error checking process of indexing solutions for the

initial geometry only, which relaxes the restraints on a valid solution. For the validation of indexing rates determined with refined geometry, the local search was not performed. Only the strongest peaks were recorded using an I/SIGI cutoff of 6, making an assumption of a detector gain of 1.0 photons/ADU. The likely gain of the detector will be around 9.0 photons/ADU for the 1.46 Å wavelength pulses used.

The Euclidean movement modes were refined first using the Nelder-Mead algorithm (Nelder and Mead, 1965) against a measure of the spread of the pixel offsets. As there were many incorrectly located peaks, a least-squares offset would put undue emphasis on outliers where the majority of the errors are not similar to those derived from a Gaussian error model. Instead, a "reward"-based target function was developed (rather than penalising for being far from the true value). This shares many similarities with a maximum likelihood distribution.

The equation E which was to be maximised using the Euclidean movement modes only is described in Equation 2.1. This calculated and summed a value for every strong reflection i, which was a sum of an exponential function of the axis offsets x and y of every other reflection j from reflection i. Hence this was independent from the origin of the pixel offsets and would be best maximised by reducing the spread of neighbouring offsets. The non-Euclidean movements were refined against the deviation of the reciprocal offsets (in reciprocal coordinates) from the origin. When X-rays diffract through a finite reciprocal lattice point at a wide angle and hit a detector far from the perpendicular, the rays will be distributed with an increasing spread of incident angles. The choice to calculate in reciprocal offsets guarded against the over-compensation which would occur by trying to centre on the midpoint of the pixel offset distribution. This was calculated from the related Equation 2.2 using non-Euclidean movements to rotate the panels until the offset is zero. Equation 2.2 is less computationally expensive than Equation 2.1. The constant k in both equations is a scale factor to roughly scale the offsets to the appropriate scale (the target function is largely insensitive to this parameter), and was set to 5.0 for pixel offsets and 0.002 $Å^{-1}$ for reciprocal coordinate offsets. The flow of operations are

described in Figure 2.5.

$$E = \sum_i \sum_j e^{-k^2((x_i - x_j)^2 + (y_i - y_j)^2)} \tag{2.1}$$

$$N = \sum_i e^{-k^2(x_i^2 + y_i^2)} \tag{2.2}$$

Each mode of movement was refined independently for each panel. This ensures that any global trends observed in the detector have not been biased by using a hierarchical detector. This means any trends in overall tilt or detector distance can be determined independently.

2.3.9 Geometry refinement using unit cell information only.

Spot-to-spot vectors were determined in the same manner as previously described in Section 2.3.3. "Intra-panel" vectors are vectors which span two spots on the same detector panel, whereas "inter-panel" vectors are those which have two spots each from a different panel.

For a given panel, all pairs of intra-panel vectors sharing a common spot were stored, and on each iteration of refinement, the two reciprocal lengths of the vectors and the angle between them (restricted to 0 <= angle < 90) were recalculated. This produces a three-dimensional histogram of two lengths, an angle and a frequency. When summed over multiple still shots from several angles, a sampling of all potential vector pairs is expected. For each panel, this histogram was refined against the ideal histogram calculated from the unit cell dimensions, using the Euclidean movement parameters. This is refined using the Nelder-Mead algorithm to convergence.

Inter-panel vectors between two chosen panels were paired either with other inter-panel vectors, or intra-panel vectors on one panel or the other, as long as they shared a common spot. The three-dimensional histogram was generated and refined against the non-Euclidean movements of the two panels next to each other. The movement is

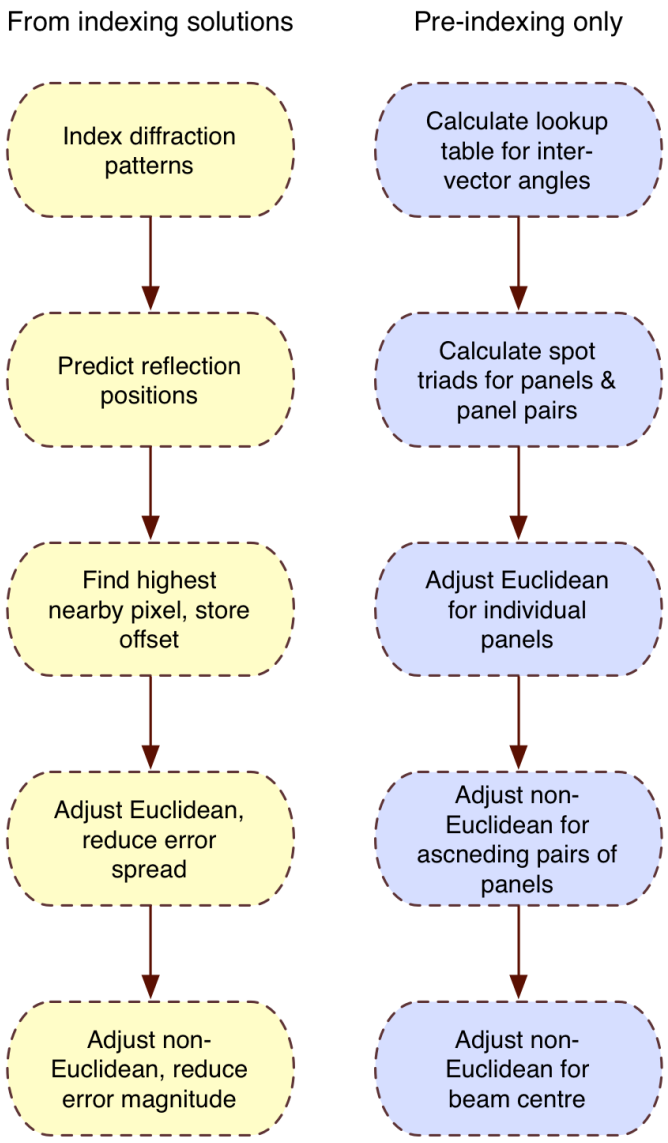


Figure 2.5: The order of events is shown for geometry refinement from indexing solutions for individual crystals (left) and for using the pre-indexing "lattice-like" information only (right).

defined by the polar coordinates θ, ϕ and ψ, the axis for which was defined as the midpoint between the centres of the two chosen panels. The panels are restricted to perform equal and opposite movements (e.g. rotation by θ of one panel must be opposed with a $-\theta$ rotation of the other), to prevent unnecessary drift of panels around the experimental hutch. A hierarchy of panels were formed for both the CSPAD and MPCCD detectors, with ascending pairs of panels up to the entire detector (Figure 2.4). For the CSPAD, adjacent ASIC pairs are paired, followed by paired ASIC foursomes, up to quadrants, halves and then the master detector. Each level of the hierarchy was refined using the Nelder-Mead algorithm to convergence.

The target function for the unit cell fingerprint was calculated from the ideal unit cell dimensions and a reciprocal lattice point (rlp) radius (akin to the profile radius in *CrystFEL* or the crystal domain size for *cctbx.xfel*). These acted as tolerances for the lengths of vectors to be accepted as resulting from the chosen lattice parameters. Vectors were rejected if their magnitude exceeded 0.15 reciprocal coordinate units in Å^{-1}. The angles were calculated from the tolerances on the lengths of the vectors such that shorter vectors had wider tolerances. These are stored in memory with 240 voxels sampling each dimension, producing 240^3 single-precision floating points (occupying 55.3 MB in memory). Each voxel contained a 'score' for a given combination of two vector lengths and the angle. This was a linear ascent from 0 at the edge of the acceptable tolerance to 1 at the most ideal calculated value. Sets of three spots were taken for each target function, and the three vectors between them were stored in memory. For the target function, a sum of the individual scores for each spot triad were calculated. Each pair of the three vectors in a spot triad were given a score between 0 and 1 according to the lookup table, and multiplied together, which can also be calculated in advance.

2.3.10 Calculation of anomalous signal.

Diffraction patterns were indexed using a given geometry file and orientation matrices were refined according to a previous protocol (Ginn et al., 2015a). The peak locations

for these orientations were then predicted, and could be treated as exact, or approximate. For the former, images were integrated at the exact peak locations; for the latter, after a local search up to three pixels away in each dimension for the highest pixel value, on which the integration window was centred. Integration and post-refinement of these diffraction patterns were performed as previously described (Ginn et al., 2015b) and anomalous merges were performed per cycle.

2.4 Results

Geometry movements were categorised into either "non-Euclidean" or "Euclidean" movements. The non-Euclidean geometry movements are those which do not affect the relative positions of diffracted rays on the surface of the panel, i.e., any movement on a sphere around the sample of a constant radius. The other movements are those which only affect the relative positions of the diffracted rays and are classed as Euclidean movements. These two modes of movements are refined against separate target functions, involving no component to which the parameters are insensitive. This means that, apart from the exception of moving an entire single-panel detector forwards or backwards, any single movement along the traditional X/Y/Z axes is disallowed. These movements must be decomposed into the Euclidean and non-Euclidean elements (Figure 2.2) and refined against separate target functions.

Two methods can be used to refine the geometry, depending on whether indexing solutions have been obtained or not. If the geometry is far from the truth, orientation matrices may be extremely difficult to determine. However, the images contain a great deal of information as the spots still obey a "lattice-like nature" defined by the unit cell parameters and space group centring. Hence the Euclidean and non-Euclidean movements can be refined against related target functions, depending on whether the geometry is close to the true positions or needs to be refined from a far starting position.

2.4.1 Corrections were made to the unit cell parameters.

The unit cell dimensions and detector geometry parameters must be carefully separated. Using the same target functions to refine the unit cell dimension as to refine the detector geometry will lead to incorrectly determined parameters. Hence, a method to abstract the information necessary for unit cell refinement from the detector geometry is performed instead (see Methods paragraph 2.3.7). This clusters the prediction of the strong reflections around the mean wavelength recorded by integration of the spectrometer output at the XFEL for both data sets. This method was used to update the unit cell dimensions used for CPV17 from 106.1 Å to 105.5 Å (deflation of approximately 0.6%). An approximate correction was applied to the initial detector distance (inflation by 0.6%) to compensate for the change in unit cell.

2.4.2 Indexing-driven geometry refinement

If the detector geometry is sufficient to index and produce pixel offsets not diverging too far from zero, the Euclidean and non-Euclidean movements can be refined against the appropriately sensitive target functions. The Euclidean movements are refined against the spread of pixel offsets (but need not be equal to zero), and the non-Euclidean movements are refined against the spread of reciprocal coordinate offsets being centred around zero. The method of refinement on individual panels using the pixel offsets is shown in Figure 2.6, showing the two-stage refinement protocol, and the overall effect on the pixel offsets on each quadrant is shown in Figure 2.7.

The TakeTwo algorithm (Ginn et al., 2016c) may be able to index with very poor geometry, in the case where the independent panels are well-oriented in their Euclidean movements but are poorly oriented with respect to either each other or the beam centre. If there are enough panels oriented correctly with respect to the sample, and enough inter-connected panels to find enough vectors consistent with a single indexing solution, identifying the correct basis vectors is very likely to be successful. However, the pixel

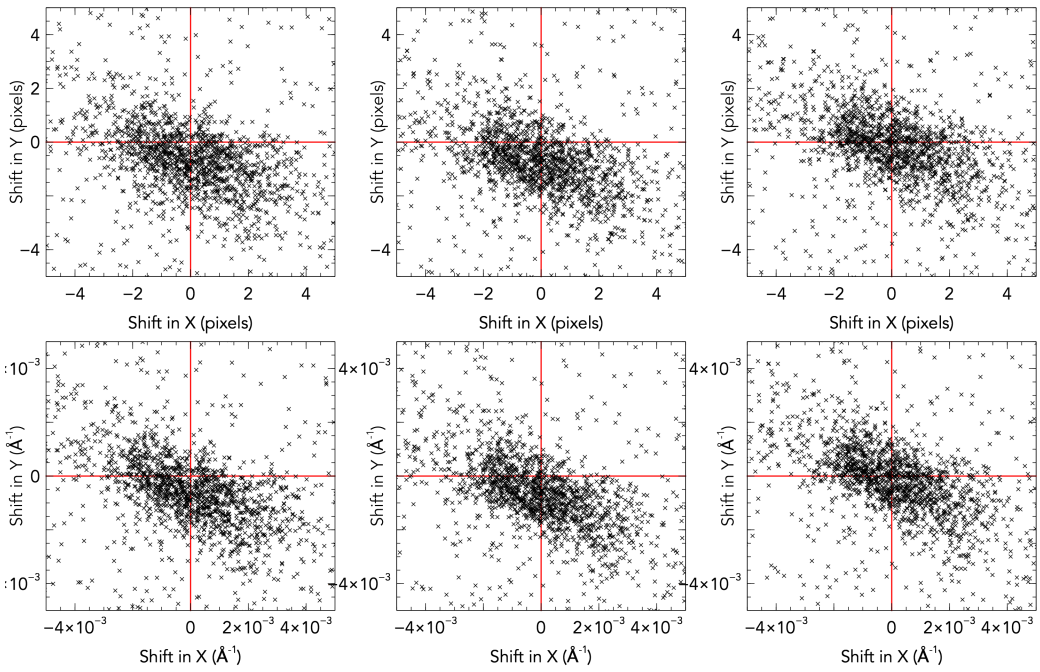


Figure 2.6: The shifts in predicted reflection positions on a mid-resolution panel at the start (left), after Euclidean movements (centre), and after non-Euclidean movements (right). The former is refined against the spread of pixels (top row), whereas the latter is refined against the centre-point of the reciprocal offsets (bottom row).

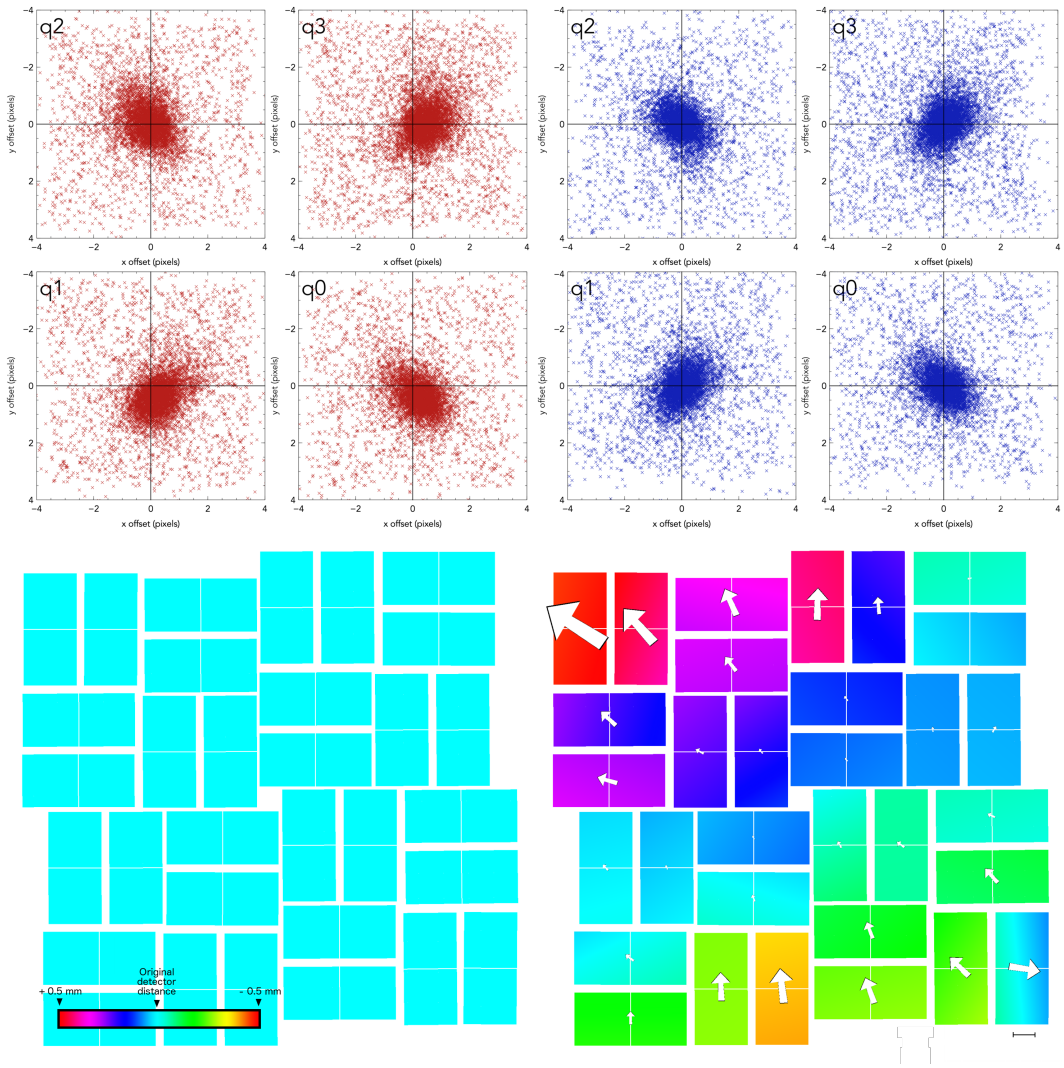


Figure 2.7: The shifts of the pixel offsets from the predicted and observed peak locations in each quadrant of the CSPAD from the original offsets in the geometry file (top, red), and after refinement (top, blue). On the bottom is the before and after CSPAD corresponding to the graphs above, with the Z axis spanning a 1 mm rainbow around the original detector distance. Arrows shown are proportional to the offset in X and Y of each panel from the original position and independent of the Z axis. Small black bar in bottom right corner shows length of a 2 pixel shift (0.22 mm). Geometry refinement performed using 170 indexing solutions.

offsets may be extremely divergent. If the indexing rate is very low to begin with, pre-indexing geometry refinement can be performed to help bootstrap indexing.

2.4.3 Pre-indexing geometry refinement uses more images, but circumvents indexing

By analysing pairs of spot-to-spot vectors, one can essentially produce an orientation-less "unit cell fingerprint" of the space group and centring (disregarding axial systematic absences). This unit cell fingerprint will be perturbed if the non-Euclidean movement parameters of a single panel are incorrect, and so can be used as a target function to lock this information. This is demonstrated graphically in Figure 2.8.

Due to the lack of exact orientation information, it is harder to lock panels in an absolute position relative to the beam centre. However, it is possible to lock panels to one another, starting with pair-wise interactions between independent panels, and then pair-wise interactions between pairs, in an ascending hierarchy, until the whole of the detector has been aligned with itself. The inter-panel vectors contain the information which defines the relative positions of the two participating panels and this forms the basis of this part of the calculation.

Finally, the entire detector must be aligned with the beam centre. However, this is unlikely to diverge far from the original centre, since locking onto an alternative Miller index is likely if only the entire tilt of the detector has exceeded the 2θ scattering angle of the lowest resolution Bragg peak (in the case of the CPV17 unit cell, this is 1.1 degrees).

2.4.4 Geometry refinement reveals a tilt in the CSPAD relative to the beam.

Geometry refinement using the indexing solutions produces a 5.8% boost in indexing rate using information from only 125 images of cypovirus type 17 polyhedrin (CPV17) data used for a previous structure solution (Ginn et al., 2015a). The boost in the indexing rate

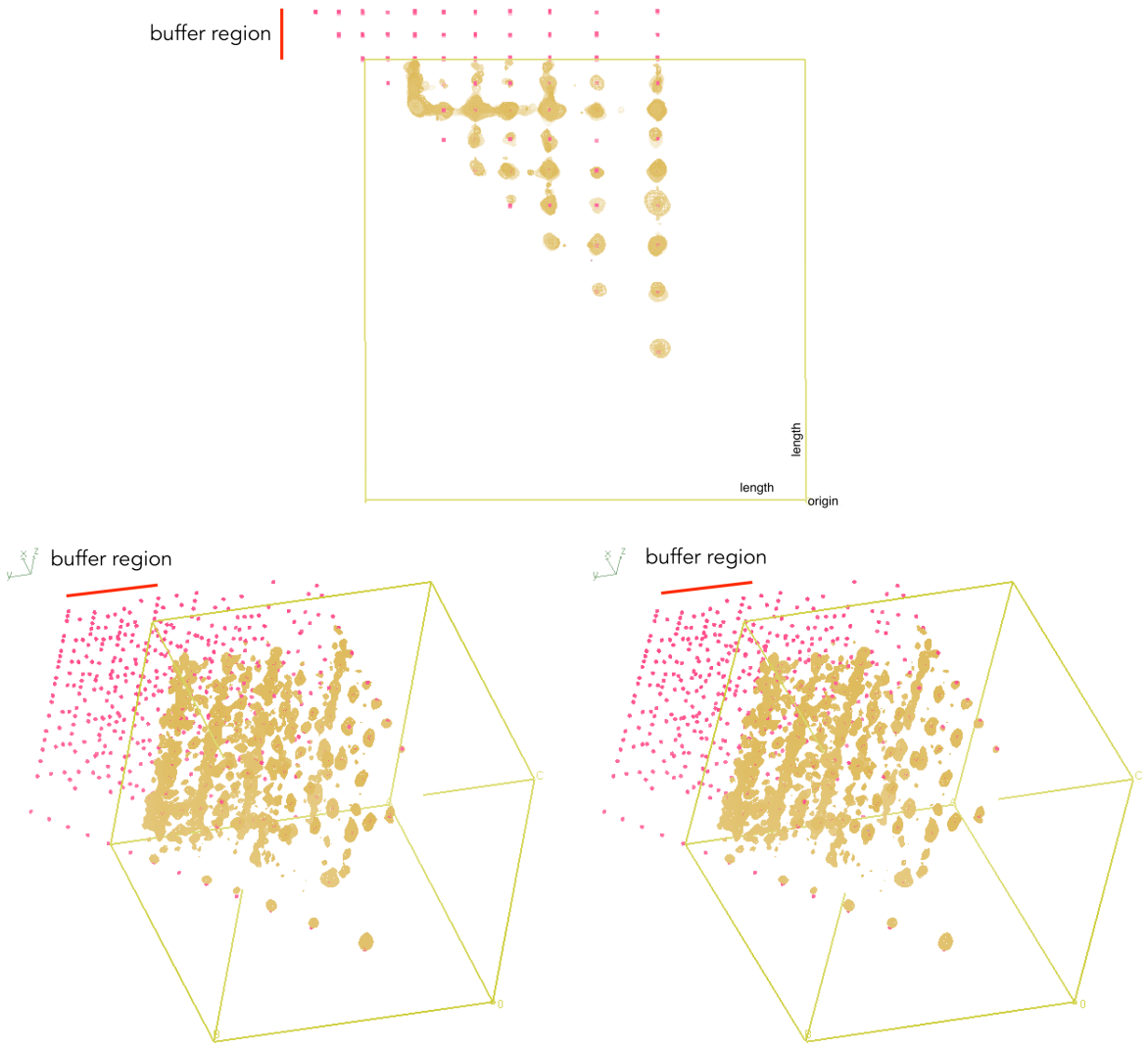


Figure 2.8: Representation of the three-dimensional histogram of inter-spot vectors using Coot. Atoms denote the positions of the expected positions, calculated with a buffer region outside of the observed vector positions (within the bounds of the unit cell). The map density reflects the set of observed vectors, contoured at 6.5 sigma. Covers reciprocal distances from 0 $Å^{-1}$ to 0.04 $Å^{-1}$. Top diagram is shown with all angles projected down perpendicular to the page. Bottom diagram is a stereo image showing two lengths (x, y axes) and angle (z axis) illustrating all dimensions.

is comparable to that using 500 unindexed patterns (Figure 2.9). The former method of refinement can exploit more information and therefore reaches convergence faster than using pre-indexing information only. However, the calculation time scales proportional to the square of the images from indexing solutions (due to Equation 2.1), but is directly proportional to the number of images from the pre-indexing information. So including more data has a reduced impact on the calculation time using pre-indexing information. Furthermore, the latter method cuts out the time required to index beforehand. Therefore in time-constrained conditions, one can approach an approximate solution from a larger number of unindexed diffraction patterns, and fine-tune the geometry using a smaller number of indexing solutions.

Because the geometry refinement is run separately on individual panels using indexing solutions (and for the Euclidean stages of pre-indexing information), any global trends which are observed throughout the detector can be considered to have been reached in an unbiased manner. In this case, both forms of geometry refinement reveal a tilt in the detector (Figure 2.10) of approximately 0.3 degrees, which produces a 1.1% error in the Z axis (a detector distance in the centre of 90.3 mm). Integration and post-refinement confirms that the nominal beam centre passes through the (0, 0, 0) Miller index.

2.4.5 Improvement in random error reduction and indexing rates

In order to judge the reduction in random error, images in the data sets were split into odd and even chunks. Indexing solutions from each half were used to direct geometry refinement using both indexing solutions and pre-indexing information only. The changes in the X, Y and Z coordinates of the four panel corners (all of which are a net effect of both movement modes) were stored for each half after refinement, and the correlation between them measured for increasing numbers of images. This increases with use of additional images and rises faster for indexing information than pre-indexing information (Figure 2.9).

The indexing rates show an immediate improvement after geometry refinement even

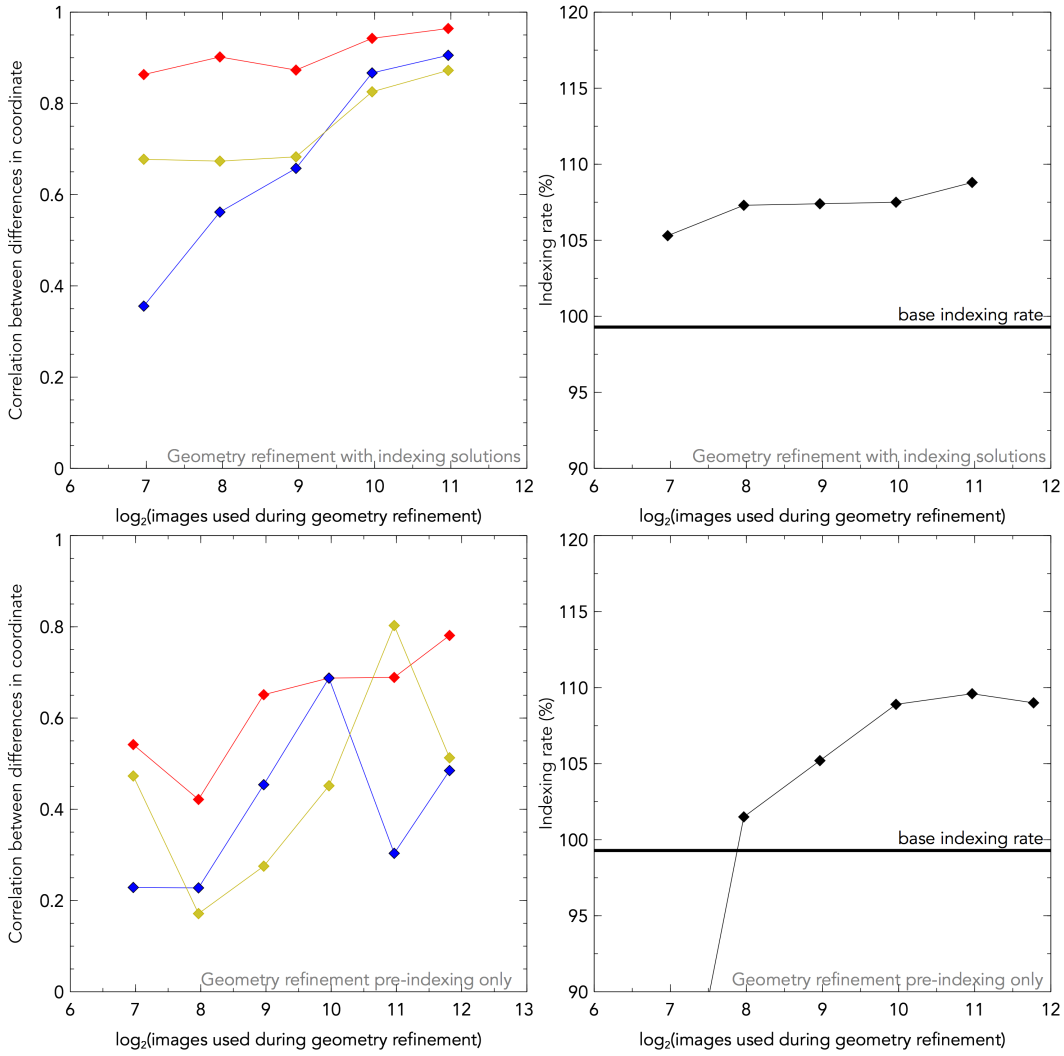


Figure 2.9: From the CPV17 data set: the correlation coefficients of the X (blue), Y (yellow) and Z (red) coordinate offsets of the four panel corners from the original positions are plotted (left), from 125 images to 2000 images (top, with indexing solutions) and from 125 to 3500 images (bottom, pre-indexing only). On the top right, the base indexing rate is shown in black for the original, partially-refined CrystFEL format geometry file for CPV17, with a local search size of 5 pixels for the maximum peak (used for indexing solution verification). After refinement from indexing solutions, with a local search size reduced to 0 in all cases, the improvement in the indexing rates are shown. A noticeable improvement in the indexing rate can be seen from 125 images using indexing solutions. On the bottom right, 125 images fail to produce verifiable indexing solutions, but 250 images produce a small improvement in the indexing rate.

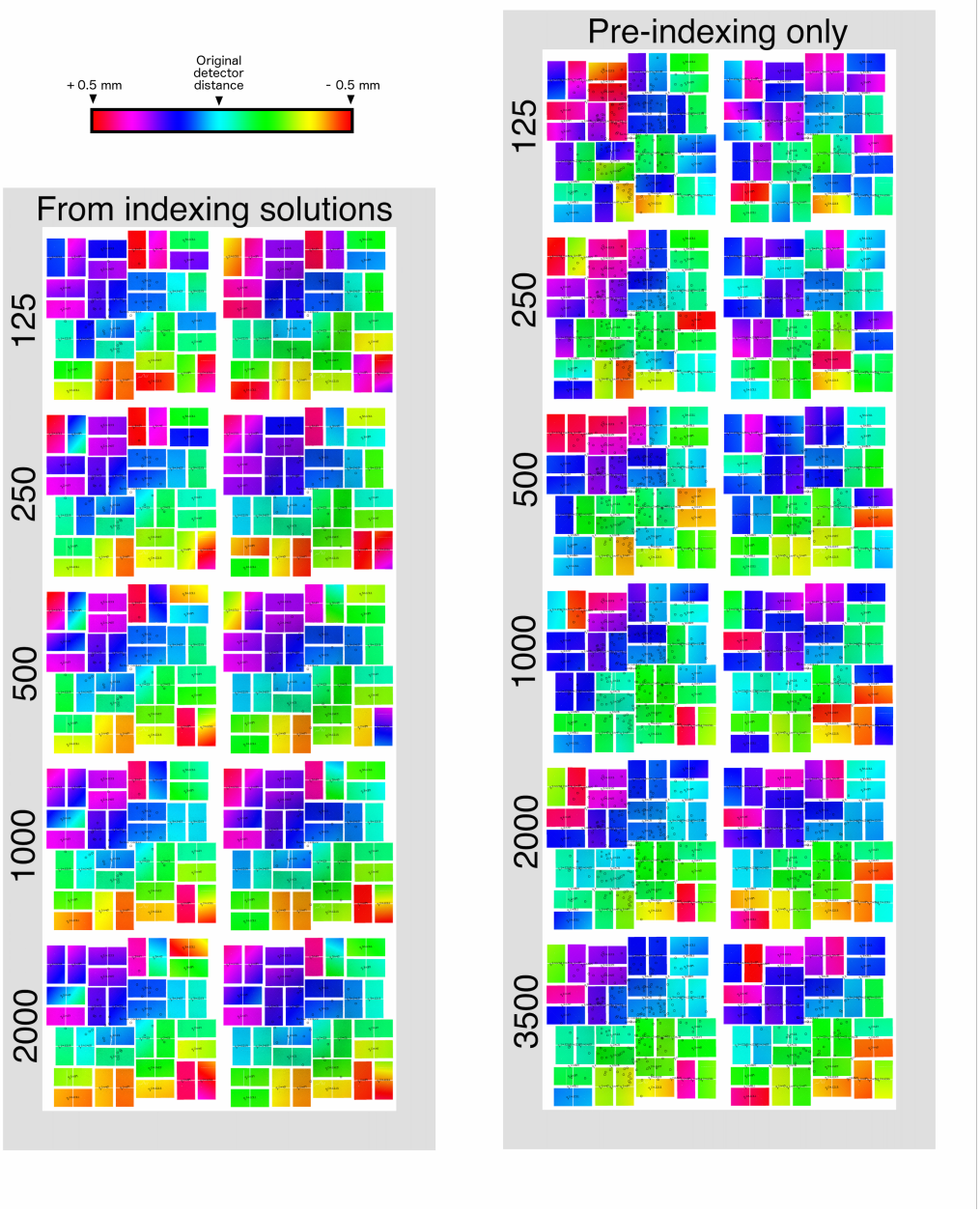


Figure 2.10: Comparison side-by-side of geometry refinement with indexed and unindexed diffraction patterns. Images of the CSPAD after geometry refinement from oriented indexing solutions (left), and using the pre-indexing information contained within the diffraction patterns only (right). Each column is divided into two half-data sets which were used to run the geometry calculation independently. The number of images which were given to each half are displayed on the left hand side.

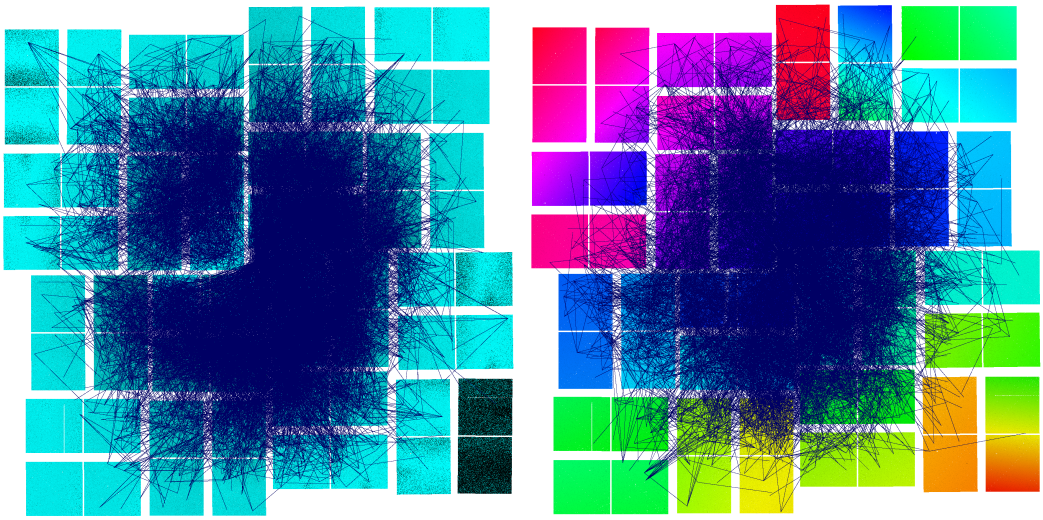


Figure 2.11: All inter-spot vectors used for indexing 500 images using TakeTwo (see Chapter 3) displayed on a single "iron filing" plot, before (left) and after (right) geometry refinement with indexing solutions from 2000 images. The left image had one quadrant deliberately moved by three pixels to the left in order to accentuate a region with poor geometry. Note that very few vectors spanning the gaps between this quadrant and the others can be used to contribute to an indexing solution.

with the minimum number of images (125 images), despite reducing the local search area to zero pixels, which produce more stringent indexing solution checks. The vectors used for indexing via the TakeTwo algorithm can also be plotted on a single image and this may reveal portions of the detector (or vectors between parts of the detector) which are not used during indexing. This "iron filing" plot can highlight poor geometry between two areas or with an area of the detector (Figure 2.11).

2.4.6 Data set quality improvement

In order to judge reduction of systematic error, the anomalous signal was chosen as an independent metric, using ANODE (Thorn and Sheldrick, 2011) to calculate the average sulphur signal from Met and Cys residues after refinement. Anomalous peak heights in real space are output from this program in terms of number of standard deviations above the mean map value (sigma). This was performed for (a) initial geometry without a local

maximum search, (b) initial geometry with a 3-pixel maximum search, and (c) for the best diffraction geometry without a local maximum search. Post-refinement was performed on indexing solutions derived and initially refined with the best geometry, and then post-refined with either initial or refined geometry definitions. Indexing solutions were initially obtained and refined using the best geometry, resulting in 7613 indexed lattices (from 7000 images).

All data sets respond to post-refinement, increasing the anomalous signal per cycle (Figure 2.12). Integrating the 7613-lattice data set using the initial geometry without performing any local search produced the worst anomalous signal, reaching a maximum of 3.21 sigma across all sulphur atoms (increase of 43% from cycle 0). Integrating using initial geometry combined with a local maximum search stabilises the post-refinement process, with a maximum anomalous signal of 3.45 sigma (increase of 41% from cycle 0). However, this will be held back by poorly estimated integration windows for weak reflections and would be unsuitable for careful structural analysis. This causes a systematic inflation of weak reflections which perturbs the twinning H-test by skewing the intensity distribution (Truncate (Winn et al., 2011) reports a twinning fraction of 9.1% for the reflection intensities after a local search, vs. 0% for reflection intensities without). The systematic inflation of weak reflections could leave a structure prone to model bias, as phases of inflated weak reflections can be manipulated to add Fourier waves supporting any correct or incorrect input model in the electron density map. Using the best geometry to predict exact peak locations for the 7613-lattice data set, the anomalous signal reaches a maximum of 3.89 (increase of 49% from cycle 0), showing a significant improvement derived from a more accurate prediction of the spot locations, and does not systematically inflate weak reflections.

In summary, a local search for strong reflections can mitigate the effect of poor geometry definition, improving the anomalous signal over predicting at exact locations. However, it is no substitute for accurate geometry determination.

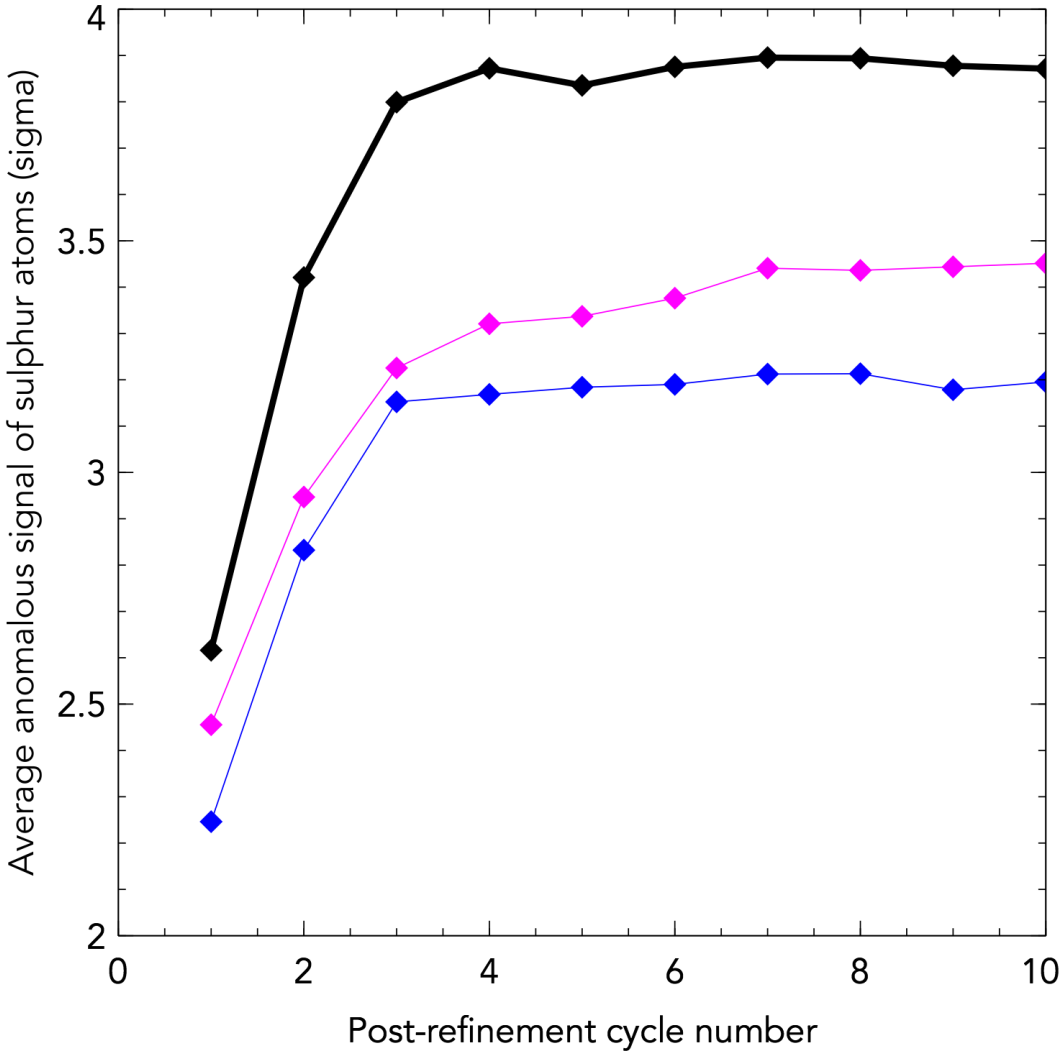


Figure 2.12: Graph showing the average sulphur atom anomalous signal per post-refinement cycle number for various treatments. 7613 orientations were derived from indexing and initial refinement using the best geometry file (2000 images and their indexing solutions). Crystal orientations were used to re-predict peak locations using the initial geometry integrating at the exact location (blue), or with a three-pixel local maximum search (magenta), or using the best geometry file and integrating at the exact location (black).

2.5 Conclusion

Using non-indexed diffraction information relies on good spot-finding information and consistent unit cell parameters from crystal-to-crystal, and computation time scales with the number of images.

Using indexed diffraction information relies heavily on the ability to index, and computation time scales with the square of the number of images. However, it will be more robust for weak data where the knowledge of the predicted peak location can be used to select high value pixels that are more likely to be from a Bragg peak-diffracted X-ray. The geometry refinement described here allowed the structure solution of bovine enterovirus type 2 by using only 352 indexed diffraction patterns Roedig et al. (2017) and this is discussed further in Chapter 6.

This algorithm leaves one type of parameter correlation not particularly well-separated: the radial distance often has quite a similar effect to tilting along both of the perpendicular axes simultaneously. These effects can only be separated if there is a sufficient distortion in the spread of spots due to a tilt: the spots at one extreme end of a panel would be clustered together compared to the other extreme end. If the spread of angles between diffracted rays allows sufficient sensitivity to separate these effects, the geometry algorithm is capable of doing so. However, this becomes harder to separate as the wavelength becomes shorter (leading to a flatter Ewald sphere) and when the detector distance is large. This is a loss of information which is inherent to all geometry refinement experiments. Therefore, experiments to acquire good geometry should be performed with a long wavelength to provide sufficient curvature (perhaps around 1.5-1.7 Å to balance this with absorption effects from air) and short sample-to-detector distance.

Because the DIALS framework adjusts the position and rotation of the detector panels with respect to the vectors which define the panel edges and the cross product of the two, the refinement of the detector at the I23 beam line would heavily emulate the non-Euclidean movement modes described here, as the panels are deliberately arranged in a circle to match the origin of the Ewald sphere (sample position). However, in the vast

majority of experiments the detector panels lie on an approximately flat surface.

Assuming correct unit cell information, one can evaluate the success of the detector geometry refinement by (a) starting from different detector distances and seeing its ability to converge on a solution from different initial positions, (b) comparing the similarity of two half-data sets and their convergence and (c) checking that the spread in pixel or reciprocal offsets has reduced from indexing solutions. The work of Waterman et al. (2016) suggests to perform 'joint refinement', which is similar to (a) but using two data sets to drive refinement rather than perform a validation check. In theory, this is another way of decorrelating parameters because the various 'valleys' in the function landscapes of various data sets will not fully overlap and possibly guide the geometry refinement to a global minimum. However, this feels unnecessary if it is possible to decorrelate the parameters as described in this chapter. If the unit cell dimensions are inflated or contracted from their true value, this may cause the detector to become concave or convex to try and correct for the error in the Ewald sphere curvature. Hence, the unit cell dimensions may need correction beforehand.

This algorithm provides a method to quickly uncover previously unnoticed tilts in the detector using a very small number of images. This is achieved by taking the set of observations and dividing the target function into parts which are sensitive to different types of motions of the detector. These two parts of the target function can converge independently. Using indexing solutions, tilts revealed across the entire detector are unbiased because the hierarchy of the detector is not considered, and each panel is refined independently. The effect of geometry refinement has a positive effect on indexing solutions and anomalous peak heights. The code has been released for use in *cppxfel* (Ginn et al., 2016a).

My personal feeling towards this work is that uncovering this algorithm taught me the most in my DPhil. For the first three years of my DPhil I had encountered this problem of detector geometry mis-alignment for various data sets, but I had fallen into the very human trap of thinking in terms of x, y, z translations and rotations around the princi-

ple axes. It took a very long time to extract myself from this type of thought process and re-parameterise the experiment, and I believed I learned the most about careful presentation of target functions to minimisation algorithms from this exercise. Minimisation algorithms are usually incapable of distinguishing global from local minima, and even if so, their capacity to do so is fairly limited. These minimisers are therefore fundamentally blunt tools and have no protection against mis-fitting using wrong parameters in scenarios where parameters are highly correlated. Therefore, the parameters should be delivered piecewise to the minimisation algorithm against the aspects of the target function to which they are most sensitive, and the parameterisation of the problem is a fundamental component of finding the way to decorrelate the parameters most effectively.

Chapter 3

The TakeTwo indexing algorithm.

3.1 Note to the reader

Some of this chapter has been taken verbatim or been adapted from my own paper (Ginn et al., 2016c).

3.2 Introduction

In many of the structural analyses from XFEL data sets mentioned in this thesis, indexing with the methods designed for synchrotron crystallography was often a difficult or impossible task. When attending the Gordon Research Conference in 2014 and the Ringberg 2nd meeting on Structural Biology with XFELs in 2015, it was surprising, from conversations with others, how contented researchers were with structure solutions where only 10-30% of the images were successfully indexed. Several studies with low indexing rates have already been published (Barends et al., 2013; Liu et al., 2013; Johansson et al., 2013; Ginn et al., 2015b; Chapman et al., 2011). Several of my own analyses (especially bovine enterovirus structural analysis in Chapter 6) led me to develop algorithms for analysing these data and indexing patterns. Thus, the TakeTwo algorithm, named in reference to the methods it uses and the re-examination of indexing that it entails, has been used by

myself and now other researchers for its higher indexing rates for XFEL data sets.

3.2.1 Indexing methods suited to synchrotron crystallography.

Indexing, or deducing the specimen orientation from crystalline diffraction patterns, can potentially be performed with high accuracy and precision due to the integral nature of the Miller indices at which Bragg reflections are located. Indexing algorithms implemented in programs such as XDS (Kabsch, 1993), iMosflm (Powell et al., 2013), DENZO (Otwinowski and Minor, 2006), LABELIT (Sauter et al., 2004), and DIALS (Gildea et al., 2014) are well-established for data collection strategies that involve crystal rotation, which are typically employed at synchrotron sources. For data collected at X-ray free-electron laser (XFEL) sources, where each image represents diffraction from a separate non-rotating specimen, the measure of success is less well-defined. Data analysis pipelines sometimes include a preprocessing "hitfinder" step, which distinguishes images without diffraction (blanks) from hits that exceed a certain threshold number of candidate Bragg spots, typically set to around 20. The indexing rate is therefore defined as the percentage of hits for which crystal orientations can be determined. Reported indexing rates are often quite low, e.g. 10% (Barends et al., 2013), 21% (Liu et al., 2013), 20% (Johansson et al., 2013), 49% (Ginn et al., 2015b) and 13% (Chapman et al., 2011), although the rate depends on a number of factors, including the strength of diffraction on each image and diffraction resolution.

Before the TakeTwo algorithm was created, the indexing programs used for XFEL analysis had originally been created for synchrotron crystallography. The major difference between synchrotron and XFEL crystallography is the large wedge of reciprocal space which is collected at a synchrotron due to a range of oscillation angles; in XFEL crystallography only a snapshot is obtained which covers a range of Ewald spheres and is convoluted with the non-singular reciprocal lattice point shape and size. This also introduces a significant error into the true reciprocal lattice point positions recorded on a detector. Indexing follows spot-finding, where potential reflections on the detector are located. All indexing

algorithms begin by back-projecting the spots onto their relative positions on the Ewald sphere. Indexing algorithms generally can incorporate the prior knowledge of the lattice as part of its algorithm, to restrict the outcoming solutions, or derive that knowledge from the spots as well.

Some algorithms rely on the use of one-dimensional (Hattne et al., 2014; Steller et al., 1997) or threedimensional (Campbell, 1998) fast Fourier transforms (FFTs) from which the peaks can provide the relative orientations and lengths of basis vectors in a three-dimensional wedge of data collected using a goniometer to rotate the crystal during exposure, as for conventional, synchrotron-based crystallography experiments. The failure rates for these methods are extremely low even for virus crystallography experiments where typically only 0.5° of data may be collected before the sample is destroyed. However, Fourier-based methods are problematic if the various vectors are not sampled frequently enough to establish Fourier peaks. Some indexing methods can now avoid the use of Fourier transforms altogether (Gildea et al., 2014), which increases the success rate for XFEL-based data as it performs a brute-force search of the expected spacings for a given set of unit cell parameters and space group centring. However, despite this, diffraction patterns which to an expert eye look like they should be indexable were still failing. This is due to a number of factors: lack of robustness against geometry errors, multiple lattices and noise all contribute to lowered indexing rates. Hence, I set out to create a new algorithm which would be more robust against the reduced information in an XFEL single shot, presence of multiple lattices and geometry errors. It is dependent on prior knowledge of the unit cell parameters, which is usually the case or can be determined by other methods for XFEL data.

3.3 Methods

In this chapter I will be focusing on the indexing results from CPV17 crystals and thermolysin crystals for clarity. In practice, many other protein samples were indexed with this

method and the results for those samples are detailed in the relevant chapters. CPV17 provides an example of a orthonormal lattice (orthonormal lattices are the cubic, orthorhombic and tetragonal space groups), whereas thermolysin crystals provide an example of an oblique lattice (which are the hexagonal, monoclinic and triclinic groups).

3.3.1 Selection of spots from images.

At the time of publication of the TakeTwo algorithm, spots were selected using the DIALS implementation of the XDS spot-finding algorithm using the DIALS program *dials.find_spots*. This has been superseded by an implementation of the blob-finding algorithm as explained in Chapter 2. The spot-finding parameters had to be varied on a case-by-case basis for the various data sets, for which the explanations and values are documented in the associated publication (Ginn et al., 2016c).

3.3.2 One-dimensional pseudo-powder patterns.

A histogram showing the frequency of each length of inter-spot vector was generated, along with the frequencies of the idealised vectors for comparison. Note that the measured relative frequencies may differ from those calculated from the idealised vectors. These plots were referred to in full as one-dimensional pseudo-powder patterns as they follow a similar, but not the same, distribution as a conventional powder pattern. The major differences are (a) pseudo-powder patterns are related to frequencies of inter-spot vectors, rather than summations of intensities; and (b) axial systematic absences will not be present on a conventional powder pattern but vectors of equivalent distance will be sampled by the pseudo-powder pattern.

3.3.3 Matching observed vectors to theoretical vector candidates

In order to use observed space vectors for indexing, they are filtered according to lengths. Observed space vectors are considered for indexing if they match one or more of the the-

oretical space vectors in length, within a certain tolerance, which is a parameter supplied to the algorithm. For the CPV dataset, this reciprocal distance was set to 1.5×10^{-3}Å^{-1}, which is about 11% of the length of the shortest reciprocal distance for the (1, 1, 0) vector.

3.3.4 Seeding an inter-spot vector network

An indexing solution cannot be determined by one inter-spot vector alone as this only anchors reciprocal space in one axis. An indexing solution must be generated by a minimum of two vectors which are not linearly dependent. To try to constrain the potential solution to a single lattice, it is stipulated that the two vectors must share a common spot. Inter-spot vectors were considered a candidate for a solution if the angle between them in the observed space matched the angle between matching vectors in the theoretical space within 1°. For each of these pairs of inter-spot vectors, a matrix was generated which aligns the observed space onto the theoretical space (the opposite transformation was achieved using the inverse). This was combined with the unit cell change-of-basis matrix to generate a potential indexing solution (see Section 3.3.6). Due to the presence of noise and multiple lattices on a single image, the map of orientation matrices produced by this method had a low signal-to-noise ratio. Enlarging this seed to form a network of inter-spot vectors helps ensure the algorithm may escape noise. Seeds were generated from pairs of vectors starting with the shortest pair of matching vectors available; if this fails to seed then the next shortest pair of matching vectors is taken, until either a seed is successful or no seed manages to produce an indexing solution (indexing solution failure).

3.3.5 Building up an inter-spot vector network

Connecting additional self-consistent vectors to the original seed to form an inter-connected network of vectors aimed to identify a single indexing solution with a large number of supporting observations. The network was considered as an indexing solution if the number of vectors reached a certain threshold above the level of 'noise'. This threshold was set to 20 by default. An indexing solution was built up recursively from a starting

vector by adding additional vectors which belong to the same indexing solution. In most cases, a new vector was not added unless it shared at least one spot with the existing network to increase the likelihood of picking spots from only a single lattice. However, this could be toggled in cases where only single lattices were suspected to be present. In order to increase the speed of calculation, vectors were 'pre-screened' to ensure the angles between the pre-screened vector and the current network of existing vectors were consistent. The resulting orientation matrices between the pre-screened vector and existing vectors were then calculated according to Section 3.3.6 and made sure to match within a 8° tolerance of the existing indexing solutions as judged by the XPLOR metric (see Section 3.3.7).

3.3.6 Generating an orientation matrix from two identified inter-spot vectors

A rotation matrix was generated from two identified inter-spot vectors i and j. Reciprocal space was rotated to line up the first observed vector with the corresponding theoretical vector. To achieve this, the cross product and angle between the observed vector i_{obs} and the theoretical vector i_{thr} were calculated.

$$C = i_{obs} \times i_{thr} \tag{3.1}$$

$$cos(\alpha) = \frac{i_{obs}.i_{thr}}{|i_{obs}|\,|i_{thr}|} \tag{3.2}$$

We then generated a rotation matrix Q_a, rotating by the angle α around the axis C, which aligns one axis in reciprocal space. The second axis must be aligned using the second vector j_{obs}. The rotation matrix Q_b rotates j_{obs} by an angle β around the axis i_{thr} such that the angle between $Q_b.j_{obs}$ and j_{thr} is minimised. Let Q_c be Q_bQ_a. The inverse of Q_c, Q_c^{-1}, equals Q_c^T, and is now equivalent to the U matrix defined by Busing & Levy (Busing and Levy, 1967). The unit cell transformation matrix is generated from the unit

cell parameters, which is equivalent to the Busing-Levy B matrix. The entire orientation matrix R is calculated as follows:

$$R = UB \tag{3.3}$$

This matrix R can be applied to integer (h, k, l) values to map them into rotated Cartesian coordinates in reciprocal space.

3.3.7 XPLOR metric to determine similarity of two matrices.

Two matrices are checked to judge similarity when building up an inter-spot vector network or when screening out duplicate lattices, determined by a reimplementation of the matrix similarity metric contained within XPLOR (Brünger, 1990; Brunger, 1992). This has been reproduced here for clarity. For two rotation matrices P and Q, and n symmetry operators where O_s is the rotational component of symmetry operator s, the metric $m(P, Q)$ is defined as in Equation 3.4.

$$m(P, Q) = \min_{s=1,n} \sqrt{Tr[(P - O_s Q)(P - O_s Q)^t]} \tag{3.4}$$

If this metric has a value below the threshold of 0.20, this corresponds to a rotation of 8° or less and the two are therefore treated as identical matrices.

For oblique lattices (where the unit cell angles are not all 90°) care must be taken to define the 'real space' rotation and not the Miller index relationship. For this it is necessary to perform a change of basis into the lattice basis vectors using the orthogonalisation matrix, followed by multiplication by the expected Miller index rotation matrix, and then multiply with the fractionalisation matrix to return to a pure rotation. This will, for example, translate the six equivalent rotations in a hexagonal space group from apparent 45° and 90° rotations (if one were to consider the Miller indices to be on an orthonormal lattice) to the 60° rotations as expected for the oblique hexagonal lattice.

3.3.8 Indexing additional lattices using the inter-spot vector network.

The indexing of additional lattices using the inter-spot vector network is made possible by a previous method (Gildea et al., 2014; Powell et al., 2013) where spots which are predicted by an existing solution are removed from the dataset and indexing is resumed with the remaining spots. Inter-spot vector networks which lead to a previously determined orientation matrix, or a geometrically equivalent matrix, are rejected. Solutions are scored for their equivalence by the XPLOR method as mentioned previously. This repeats until the maximum number of indexing solutions has been achieved. For images with large numbers of spots, it relies on an 'accidental indexing' event to begin to reduce the noise, after which successive indexing solutions become more and more easy to acquire for a particularly crowded image.

3.3.9 Modifications made after publication.

Some major modifications were made after publication. It was noted that for images with a particularly long unit cell dimension for which rows of closely spaced reflections satisfied the Bragg condition, the algorithm could only latch onto a single axis and essentially leave the final one randomised, by picking two starting vectors which had an angle between them of 0° or 180°. Hence, the algorithm was modified to explicitly exclude starting vectors which were co-linear before elongation of the inter-spot vector network.

Another modification was made to ensure that, for a starting solution from two vectors, that all three pairs of two vector magnitudes and the angle between the vectors closely matched the theoretical values. This excludes a small number of incompatible solutions before growth of the inter-spot vector network.

3.4 Results

3.4.1 Global experimental parameters can be corrected using the pseudopowder pattern.

For the publication of the TakeTwo algorithm (Ginn et al., 2016c), I had already noticed the scope of using the pseudo-powder pattern to correct for experimental errors such as global detector distance and, in some cases, the beam centre. Although this was superseded by significantly more involved work on full detector geometry refinement in Chapter 2, the potential was noticed early on in order to correct for detector distance in Chapter 6.

We find that well-defined pseudo-powder patterns (generated as described in the chapter on geometry refinement (Chapter 2) can be obtained even when maximum-pixel powder patterns cannot be usefully calculated. For instance, when few patterns are available or rings cannot be separated on individual images, pseudo-powder patterns can still be produced. These are not only useful for diffraction patterns with small unit cell dimensions (Figure 3.1(a)), but are still interpretable for diffraction patterns from crystals with large unit cells and very poor reflection separation, as for BEV crystals (Figure 3.1(c)).

An example of the vectors used in a single image to contribute to the overall pseudo-powder pattern is shown for CPV crystals (Figure 3.1(a)). The conventional powder patterns generated by overlaying multiple images may show fewer peaks than the pseudo-powder patterns due to axial systematic absences. The corresponding distances will not be present in a conventional powder pattern, but will appear in the pseudo-powder pattern as this vector does occur between Miller indices where neither index is the origin.

As the number of BEV crystals was very small, a conventional powder pattern could not be generated by superimposition of the images, and if it had been possible, the rings would be impossible to distinguish: the unit cell dimension is 437 Å ; even the longest reciprocal space distance in the pseudo-powder pattern (50 Å) only appears 3 pixels from the beam intersection with the detector, and was therefore lost in the unrecorded low

angle region. Only after a one-dimensional pseudo-powder pattern had been generated, was it revealed that the original detector distance reading was misaligned by 20 mm. This value could be refined manually by observing the effect on the pseudo-powder pattern (Figure 3.1(c-d)). This enabled indexing of these diffraction patterns in a situation where the presence of this anomaly was otherwise difficult to establish. The BEV images had very poor separation of individual reflections, with the closest Miller index vector (1, 1, 1) corresponding to only three pixels. If the error in a spot position was $\sim$0.5 pixels, the error in the vector will be about 23% when considering the error over two spots, having a major impact on the pseudo-powder pattern especially at small reciprocal distance values.

3.4.2 TakeTwo is relatively robust against errors in geometry.

To test the robustness of the algorithm, experimental parameters were altered and the effect on the indexing success rate was observed, using CPV17 data. As the detector distance and beam centre are the most likely to be incorrectly measured at an XFEL beamline, these parameters were varied. The detector distance was varied from its true value of 101.2 mm by 1.0 mm in either direction, and similarly for the beam centre X position. This algorithm had a good tolerance to experimental errors, and although indexing rates decreased when the model parameters deviated from their true values, this was in the form of a gradual decay. Indexing success only decreased by about 45% after moving approximately 1.0 mm from the optimal beam centre (Figure 3.2). The fall-off was sharper for the detector distance, dropping by 45% after moving approximately 0.5 mm from the optimal detector distance. For comparison, the shortest vector for this space group would have spanned a minimum of 1.77 mm across the detector if it had been located near the beam centre.

Furthermore, a major problem with Fourier-based methods is that it requires the geometry of the detector to be broadly correct across the image. As a non-FFT-based method, TakeTwo is capable of indexing off a minor fraction of a correctly aligned image regardless of the positioning of other parts. However, it is still likely to be rejected

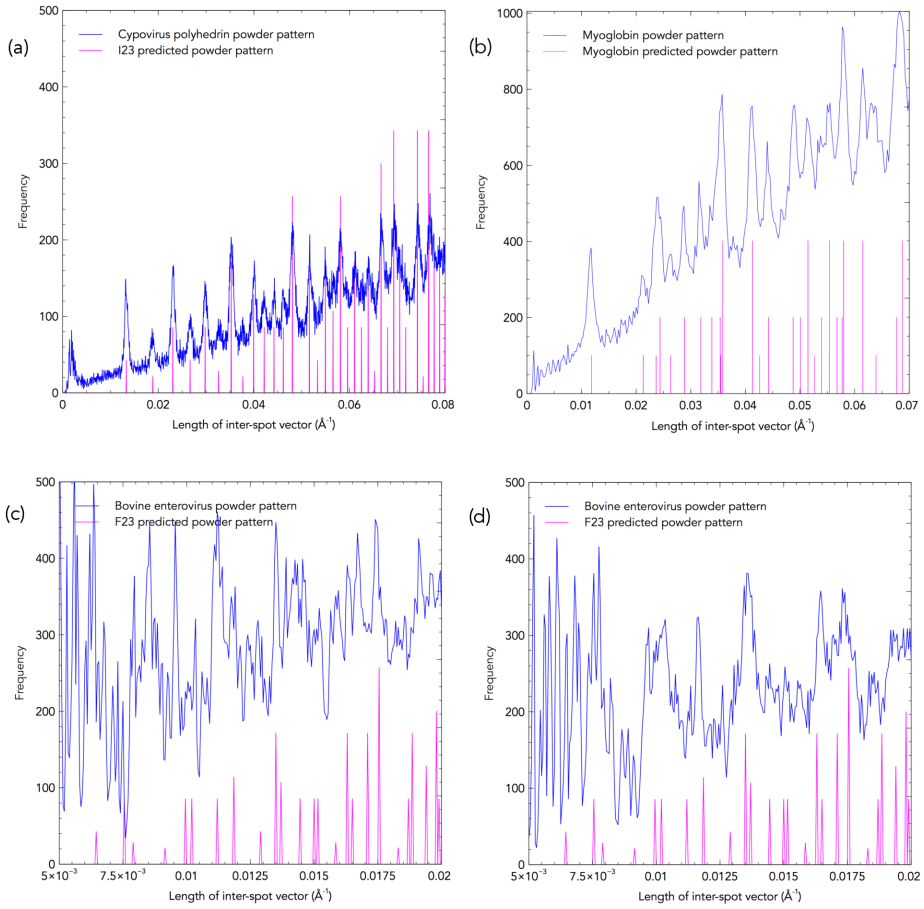


Figure 3.1: CPV17 at 101.2 mm detector distance (a), myoglobin (b) and BEV pseudo-powder patterns (c and d) generated using optimised experimental parameters without spot vector filtering. Blue line is generated from experimental data whereas the pink line is the expected pseudo-powder pattern given the unit cell and space group information. The height of the magenta line denotes the relative number of Miller indices which give rise to the particular inter-spot vector length; the absolute values are not shown for clarity. CPV17 powder pattern was generated from 262 images. Due to the higher mosaicity of the myoglobin crystals (464 images), the experimental peaks for each powder ring in (b) are broadened compared to the CPV powder pattern (a). For the BEV pseudo-powder patterns (304 images), the pattern in (c) was generated from the recorded detector distance of 105 mm. However, pattern (d), showing a better fit to the pseudo-powder pattern, is achieved at a detector distance of 85 mm. The peaks are less well separated due to the high error associated with measuring spot vectors between extremely close spots.

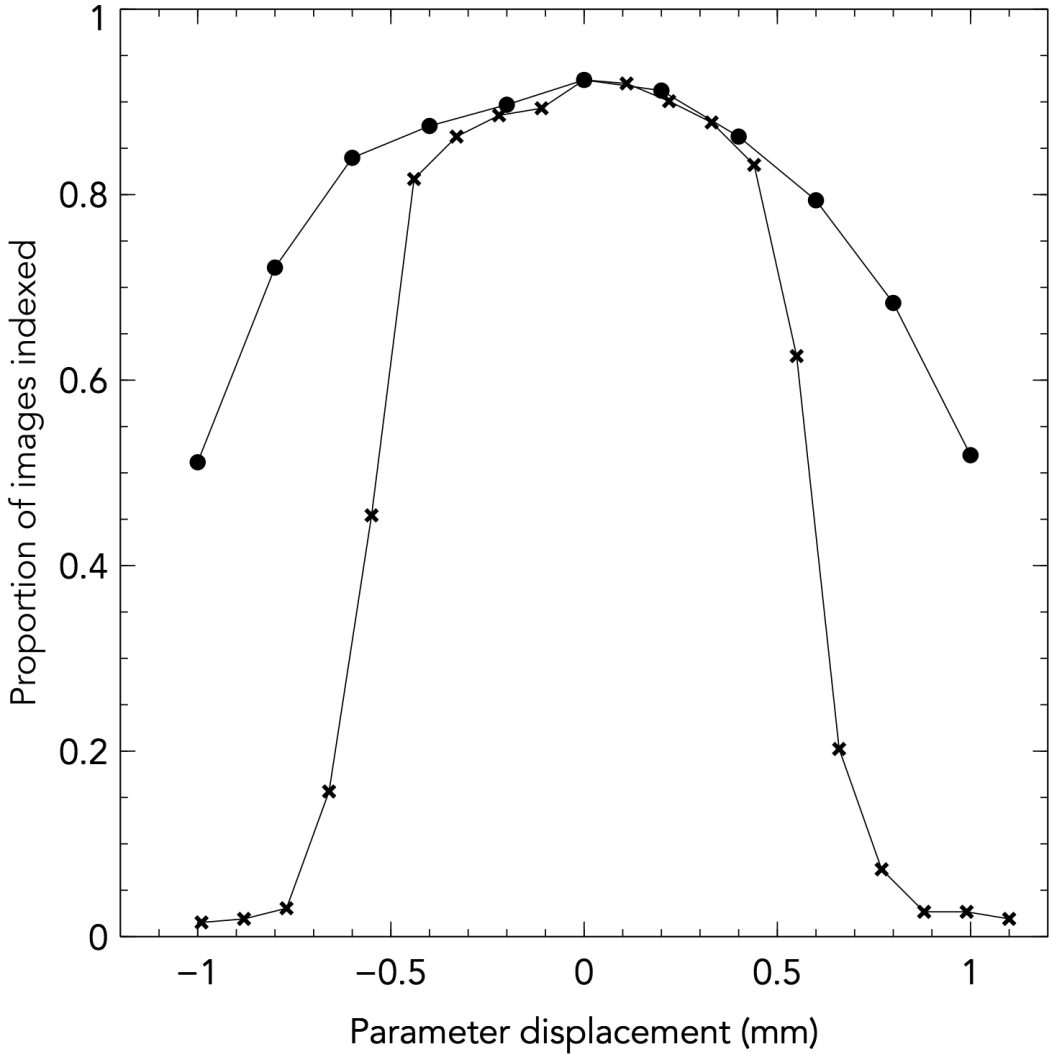


Figure 3.2: Tolerance of the indexing algorithm to the detector distance parameter (filled circle) and beam centre horizontal pixel offset (cross) are shown. Indexing success decreases significantly after 0.5 mm displacement from the optimal detector distance. Calculated from CPV17 sample at 101.2 mm detector distance.

by any indexing solution validation algorithms which may lie downstream of the indexing stage.

3.4.3 Indexing rates frequently exceed 100% for various crystalline samples.

Due to the orthorhombic nature of myoglobin, powder rings will overlap in ways which are not observed in cubic space groups. This will lead to a degree of uncertainty during the indexing stage when assigning theoretical vectors to observed vectors according to their distances. The pseudo-powder pattern for the myoglobin data set is shown in Figure 3.1(b).

At the time of publication, the lowest symmetry crystal data available to me were orthorhombic crystals of myoglobin crystals from a collaboration with Dwayne Miller. Indexing with the TakeTwo algorithm was very successful for this crystal system and led to its structure determination (Oghbaey et al., 2016). After publication, a collaboration with Yun-Xing Wang and Jason Stagno on RNA crystals of the adenine riboswitch aptamer provided access to a monoclinic crystal system of heterogenous and poorly diffracting crystals, and further collaborations with Dwayne Miller and Oliver Ernst provided access to defluorinase diffraction patterns (space group P_12_1) during beam time in SACLA, Japan. So far, I have not had access to any triclinic crystal diffraction. Indexing rates for various samples are shown in Table 3.1.

3.4.4 The algorithm is frequently able to index multiple lattices on each image.

The highest indexing rate which was achieved from a single run was that for CPV18 loaded on a micropatterned silicon chip from the July 2015 data collection at LCLS, where there were significant numbers of multiple lattices. This had an average of 2.25 lattices per image with a cap of four lattices to index per image, but it was noted that nevertheless a

Table 3.1: Indexing rates for crystalline samples of various crystal systems.
* This data set was processed indexing a maximum of only one crystal per image.
** This was real-time indexing during a 60 hour beam time at SACLA and not continued afterwards; indexing rate is approximate due to a poor ability to exclude salt diffraction from the hit-finding results. (Fixed Target A (Roedig et al., 2015), Fixed Target B (Oghbaey et al., 2016))

Crystalline sample	Delivery system	Crystal system	Average lattices per image
CPV (US17)	Liquid jet	Body-centred cubic	1.51
CPV (OB18)	Fixed target A	Body-centred cubic	2.25
Thermolysin	Liquid jet	Hexagonal	1.07
Bovine enterovirus *	Fixed target A	Face-centred cubic	0.96
Myoglobin	Fixed target B	Orthorhombic	1.14
Defluorinase * **	Fixed target B	Monoclinic	Approx. 0.80

large number of spots went unexplained. Also, images which had at least one indexed diffraction pattern tended to lead more easily to further indexing solutions. 85.4% of images had at least one image indexed, and of those the average was 2.66 ± 0.61 lattices per image. The presence of multiple lattice crystals and crystals with large unit cells lead to very large numbers of spots measured by the detector. In these cases, it was observed that traditional indexing algorithms have a tendency to generate and accept a solution due to the relatively high probability of a spot being close by to a reflection predicted by an orientation matrix. However, this happens regardless of whether this matrix corresponds to a true crystal in the pulse path or not.

3.5 Conclusions

The TakeTwo algorithm has an in-built tolerance for the distances accepted for generating indexing solutions, which allows the algorithm to tolerate errors due to the finite size of the reciprocal lattice points and the non-monochromatic nature of the XFEL pulses (referred to as a "pink" beam, a reference to the multi-wavelength nature of visible pink light). The spectral dispersion for these pulses is about 0.25%, from my own estimation

when optimising the bandwidth value in Chapter 5. As there is a negligible rotation component of the crystals during the exposure, monochromatic light has a dual effect of increasing the precision of estimated vector distances, but also reducing the number of Bragg peaks which are illuminated. However, if the spread of wavelengths in the XFEL pulses is too great, the various distances will overlap and the algorithm will be diverted down too many "false" solutions when relaxing the tolerance to compensate. There is likely to be a mid-point where these two features balance for both acceptable indexing rates and numbers of observations.

3.5.1 Reception from the XFEL community

I was invited to the 3rd BioXFEL conference in Puerto Rico (January 2016) to give a talk to the XFEL community, for which I presented this indexing algorithm. It was received well and I was invited to give the same talk at the 3rd meeting for Structural Biology with XFELs, in Schloss Ringberg (February 2016). It also highlighted the prevalence of multiple lattices to the community, which had not been noticed, or was not considered to be as concerning before this point. Assuming a Poisson distribution, one would expect multiple hits to be quite rare. In reality, crystals have a tendency to stick together and therefore hitting crystals are not independent events. Previously, multiple lattices were seen as something to avoid due to the difficulty of indexing them; however, the merits of this algorithm and its ability to index multiple crystals opens the possibility of exploiting certain assumptions which can be made about multiple lattice images. One assumption, for example, is that the energy profile of the beam is shared between all crystals exposed on the same image. Therefore, parameters fit to the beam profile may be shared among multiple crystals resulting in an effective increase in the observation to parameter ratio. The BioXFEL talk also triggered a collaboration with Yun-xing Wang and Jason Stagno due to their struggle to index their highly variable RNA crystal diffraction. They are now able to obtain additional indexing results from the TakeTwo implementation in CrystFEL (see section 3.5.2). I also presented this in poster form at the European Crystallography

Meeting (September 2016), generating conversations about its potential applicability to wide-bandwidth energy spectra. I presented the same poster at the Users Meeting for the European XFEL. Coincidentally, I had discussions with Adam Round at the European XFEL about the feasibility of indexing multiple lattices off summed images using the TakeTwo algorithm, as they were looking for potential solutions for the extreme volumes of data expected to be collected when the European XFEL increases its pulse rate.

3.5.2 Incorporation into CrystFEL

The TakeTwo algorithm was incorporated into my published software suite (Ginn et al., 2016a), but use of the indexing algorithm in this software locks users to the cppxfel pipeline. Due to the number of requests for use of the algorithm, and being rather unwilling to be the gatekeeper to the algorithm, I have been working with Thomas White at the Centre for Free Electron Laser science to incorporate this algorithm into the widely used CrystFEL suite, in order to increase accessibility to the TakeTwo algorithm. This was released in version 0.6.3 as an experimental version on 24th July 2017.

Chapter 4

Initial orientation matrix refinement for still images.

4.1 Note to the reader

The work contained within this chapter was the primary focus of a publication (Ginn et al., 2015a). This publication reported the initial orientation matrix refinement and a rudimentary model of partiality. The partiality model in the paper was modified significantly and provided significant additional benefit to merit a second publication, (Ginn et al., 2015b), which is the focus of the next chapter.

I would especially like to thank those who attended the data collection trip prior to the start of my DPhil course. The data collection trip was carried out by Geoff Sutton, Danny Axford, Armin Wagner, Aaron Brewster, Nicholas Sauter, Marc Messerschmidt, Jonathan Grimes and Dave Stuart.

4.2 Introduction

Cypovirus polyhedrin crystals are a very good test case for building a data processing pipeline for XFEL analysis, due to their ease of growth in quantity, their homogenous

nature and cubic crystal system. In this chapter, the downstream data processing after indexing but before post-refinement will be outlined. This considers a major stage: initial orientation matrix refinement, where images are refined without a reference data set, which reduces the error and places the orientations of each crystal in the radius of convergence for the following post-refinement algorithm. The aim of this, and the following chapter on post-refinement, is to improve the quality of the data by maximising the use of information from a given number of indexed patterns.

4.2.1 Orientation matrix determination prior to initial orientation matrix refinement.

The software suite CrystFEL was published in 2012 (White et al., 2012) as the first (and only) user-friendly pipeline for XFEL structure solution, but initially lacked an XFEL-oriented approach to the data analysis. As such, CrystFEL passed images or reciprocal lattice point positions to existing indexing algorithms such as those implemented in Mosflm and Dirax to determine orientation matrices. CrystFEL has since moved onto replacing stages of the pipeline with XFEL-specific algorithms. However, indexing algorithms (whether suited to rotation data or still images) tend to produce a close but imperfect solution. Further refinement is crucial. Orientation matrix refinement is not new. The strategy employed prior to the algorithm presented in this chapter was to take the spot-finding results and refine the orientation matrix such that the reflections which predicted these spots were as sharply centred on an Ewald sphere of a certain radius as possible. This algorithm is essentially extremely similar, with a minor but significant modification: instead of using only the original spot-finding results for orientation matrix refinement, predicted spot positions are integrated, assessed for being 'strong' reflections or 'weak/non-existent' reflections, and the additional strong reflections are added to the refinement target function. This minor alteration in strategy, which does not rely on spot-finding results alone, vastly increases the number of observations and has a tendency to include reflections which are only partially illuminated by the X-ray pulse. Sometimes, reflections on the weaker side

are not picked up by spot-finding algorithms, but can be trusted if prior knowledge of their approximate location is used: this is what this algorithm exploits. This has the effect of improving the accuracy of orientation matrix determination.

4.2.2 The cypovirus genus infects insect cells and has a major impact on the silkworm industry.

The cypovirus is a genus of dsRNA viruses in the family *Reoviridae* which infect insect cells of certain species. They share many similarities with the genus of beta-baculoviruses, which have a dsDNA genome. The polyhedrin of the granulovirus, a species of the beta-baculovirus genus, has also been studied with XFELs and recently published (Gati et al., 2017). In the virus lifecycle, cypoviruses infect insect cells and generate new RNA-filled capsids. Another protein produced alongside is the polyhedrin protein. In all species of cypoviruses and also both alpha- and beta-baculoviruses, the purpose of the polyhedrin is to form crystals which protect the capsids when outside the cell from environmental duress. Cytoplasmic polyhedrosis viruses (CPVs) infect many insects, can be used as insecticides (Kunimi, 2007), and are a problem for the silk industry due to silkworm infection(Goldsmith et al., 2005). These polyhedra have also been repurposed to redirect foreign proteins or complexes into polyhedra using the N-terminal VP3 polyhedrin recognition sequence, which has increased stability against dehydration in microarrays (Ikeda et al., 2006) and is useful for slow release of immobilised growth factors in mammalian tissue culture (Ijiri et al., 2009). Members of our laboratory have focused on the polyhedrin crystal as it is a well-behaved, homogenous crystal with high cubic symmetry, and is well suited to XFEL analysis. They have solved a large number of cypovirus structures over the years before I arrived, and the cypovirus type 17 (CPV17) polyhedrin was the last remaining unsolved crystal structure.

4.2.3 Previous attempts to solve the structure of CPV17 polyhedrin.

The cypovirus polyhedrin crystals grow spontaneously inside Sf9 insect cells when infected using a baculovirus vector, but fail to reach a size greater than 1 μm. Attempts had been made to solve the structure at the microfocus beamline I24 at Diamond Light Source Ltd, and although a merged data set could be generated, it was unable to support molecular replacement. This is in part due to the poor sequence similarity between the different cypovirus strains (Figure 4.1), and significant structural divergence despite maintaining the same space group and similar cell dimension. The attempts were hindered by the extremely small crystal sizes, which meant diffraction quality was poor even at the microfocus beamline of I24. However, significant methodological progress was made as DIALS, one software suite for processing synchrotron data sets, was further developed to support multiple lattices and allowed indexing of up to six lattices per image (Gildea et al., 2014).

4.3 Methods

4.3.1 Preparation of cypovirus polyhedrin crystals

Cypovirus type 17 (CPV17) polyhedrin was expressed, crystallised and purified primarily by Geoff Sutton. The polyhedrin gene for CPV17 was synthesized by GeneArt (Life Technologies) based on the sequence AY876384 from the Gene Bank. This was inserted into the transfer vector pBacPAK9 (Clontech), a baculovirus vector. This was co-transfected into Sf9 cells. On expression of this protein in the cells, the protein forms crystals within the boundaries of the cell, which remain intact. Polyhedra crystals were pelleted at 500 g and resuspended in water for storage.

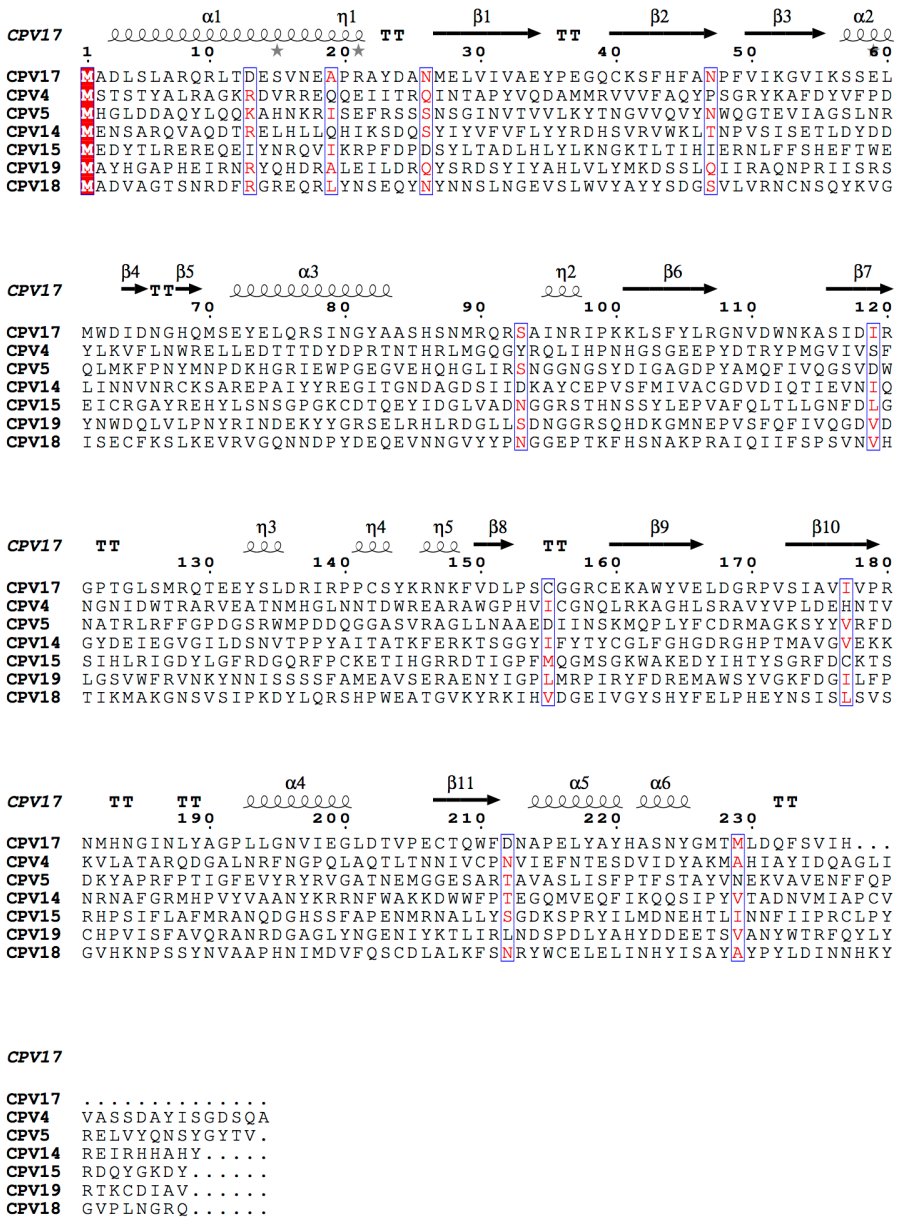


Figure 4.1: Sequence alignment for a number of the cypovirus polyhedrin proteins from a representative sample of the virus species. Note lack of appreciable alignment. The sequence alignment produces no clear homology between different species (those residues marked as potentially highly conserved are in red fill and white text; those marked as potentially mildly conserved are outlined with a blue box). Homology and annotation created using ESpript (Gouet et al., 1999).

4.3.2 Powder pattern collection at Diamond at room and at cryo-temperatures

At the time of data collection at the XFEL, no assumptions were made that the calculation of the mean wavelength of the X-ray pulses was accurate. However, because cryo-temperature crystal diffraction was collected at the synchrotron and room temperature crystal diffraction had been collected at the XFEL, this meant that the relative unit cell dimensions at the two temperatures were not known. To find an independent measurement of this parameter, crystalline slurry of CPV17 were packed into capillary tubes in preparation for powder diffraction collection at Diamond beam line I04 by Geoff Sutton. This boosted the signal to noise ratio by exposing a much larger crystal volume, but still provides well-separated powder rings for which the unit cell could be easily determined for the cubic space group. The crystal slurry was exposed to X-rays at room temperature and then again at liquid nitrogen temperature and separate diffraction patterns were collected. The radius of well-defined rings from the centre of the beam were measured and fitted to Miller indices which would result in a ring at that certain distance from the centre. Due to the space group, I23, all unit cell dimensions are equal. The wet lab and data collection was performed by Geoff Sutton and I performed the data analysis.

4.3.3 Data collection at SLAC

Crystals of CPV17 were taken to SLAC to be exposed to the XFEL in March 2013, prior to my arrival at the laboratory. X-ray pulses of 50 fs duration containing about 2×10^{12} photons per pulse at a wavelength of 1.46 Å (this mean pulse wavelength was initially derived from the tuned energy of the laser). The crystal slurry was diluted to provide 3×10^{9} crystals per ml. The crystals were injected into the XFEL beam in water using a Gas Dynamic Virtual Nozzle DePonte et al. (2008) focused to a diameter of a few mm at a flow rate of 30 ml min. The jet was buffered by a stream of gaseous helium. Data were collected on the CSPAD at a 120 Hz repetition rate and recorded on flash drives for run times of less

than 10 minutes, before being written to hard disk in a LCLS-custom format known as an XTC stream. These gaps are necessary as data cannot be written to disk as fast as it is generated. As the jet speed was set to 10 m/s, a 2 micron beam will only illuminate 240 microns of the liquid jet per second at the LCLS repetition rate, thereby only illuminating one in every 40,000 parts of the jet. Therefore, most crystals are wasted. Data were collected for a total of 20 minutes spread over four runs at 8500 eV wavelength X-rays. A small amount of data were collected at 101 mm distance between the sample and the detector, and then the detector was brought further in to 91 mm to increase the available resolution. The maximum resolution achieved at this wavelength and minimum detector distance was 1.57 Å . After observing that reflections could be seen at the corners of the detector, the XFEL was retuned to a wavelength of 9500 eV in order to access higher resolution reflections. For the remaining runs, the maximum resolution achieved in the corner of the detector was 1.40 Å . In the final runs, the polyhedra crystals were injected at an increased flow rate.

4.3.4 Extracting data from the SLAC XTC streams (*cctbx.xfel*)

The XTC streams created at SLAC are 96 GB files containing a dump of the image data from the detector or detectors in the rawest format. These include a large number of blank images for which no crystal was in the path of the beam. The first stage before any analysis can occur is to extract meaningful images from the XTC streams. This involves the use of the software suite *pyana* which can read the XTC streams. Additional code written by Nicholas Sauter *et al.*, University of Berkeley, California, as part of the *cctbx.xfel* suite, provides XFEL-specific modules which plug into the *pyana* interface. A few helpful events occur at this stage: correction of the positioning of the panels from the optically determined geometry which is calculated on each rebuild of the detector, the application of any improved detector geometry correction in addition to this, and subtraction of a dark image to reduce the background. The dark images are calculated from a 'dark run', where the shutter is closed and the jet is not running. The dark run is taken as the

average of several thousand of images collected during a dark run. For collection of real data, the value of each dark run pixel is subtracted from the corresponding pixel of every subsequent image. It is important to note that this can create images with positive or negative pixel values.

In order to reject blank images from the stream, the *cctbx.xfel* extension to *pyana* scans quickly for hits by applying simple spot-finding algorithms to each image. Those which have a certain number of spots (typically 20 spots or more) are written out as individual files. Optionally, one can also attempt indexing and integration of individual images at the same time, and one can write out image files for indexed images, all hits (whether indexed successfully or not), or all images within a stream.

All hits were extracted in this case and written out in *pickle* file format, which is a generic data storage file format used in Python. These *pickle* files can only be read from Python and the *dxtbx* library is capable of reading these file formats, applying detector geometry corrections embedded in the file and providing the counts for individual pixels. As I have a personal preference for strongly typed languages, I try to write the vast majority of my code in C++ and so distanced myself from the Python environment and worked with dumped binaries of the image files. The disadvantages of this is that these files take up more disk space: each image stores 1765 by 1765 pixel images using a four-byte integer to store the value of each pixel, resulting in 12.5 MB files for each image. In addition, writing out a dumped stream from pickle files represents another step in the pipeline, which makes the software less accessible for other users.

4.3.5 Extracting data from the SLAC XTC streams (*Cheetah*)

There were various limitations of using the *cctbx.xfel* pickle files which drove me to perform new hit-finding with *Cheetah*. The pickle files had imperfect geometry positions applied which were then fixed in the image layout and difficult to modify. Due to the limitations of having to read the file using the *cctbx.xfel* Python library, I also noted that it is not possible (or at least very difficult) to extract the coordinates of the arrangement

of panels exactly, but only an approximate pixel position. Hence I performed new hit-finding with *Cheetah* which dumps into the more flexible HDF5 file format which can be read from any programming language including C++. This does not rearrange the panels into their respective positions, which is preferable when the geometry of the detector is frequently subject to re-calculation.

4.3.6 Indexing images using cctbx.xfel

Images were initially indexed by Nicholas Sauter using the *cctbx.xfel* suite which has been developed by his laboratory. At the time, this software was only capable of indexing one lattice per image. Nicholas Sauter optimised the parameters of the spot finding in order to maximise the indexing rate. These indexing solutions were used for the papers reporting the structure solution of CPV17 (Ginn et al., 2015a,b). Before the advent of my own indexing algorithm, I extracted these indexing solutions from the pickle files for use in my own software and departed from the Python environment.

4.3.7 Indexing images using DIALS

Diffraction Integration for Advanced Light Sources (DIALS) is an open-source, object- and model-oriented software which, at the time of writing, is capable of indexing and integrating synchrotron data for use in scaling and merging programs such as AIMLESS. The spot-finding and indexing parts of DIALS were used to generate orientation matrices for individual XFEL images using automated scripts. This allows the use of the DIALS indexing engine which is capable of indexing multiple lattices by feeding the unindexed spots back into the program after a new lattice is found. The spot-finding parameters were tuned in order to exclude as much noise and poor detector pixels while retaining as many real spot predictions as possible. This was used primarily as a comparison to other indexing algorithm results.

4.3.8 Indexing images using the TakeTwo algorithm

Images were indexed using the TakeTwo algorithm using the spot-finding results of DIALS, long after publication of the initial structure. Up to three lattices were searched for per image, using a search reciprocal lattice point size of 1×10^{-4}Å^{-1}, a bandwidth of 0.26% and a maximum resolution of 0.15Å^{-1}.

4.3.9 Initial orientation matrix refinement

The initial orientation matrices determined by either *cctbx.xfel* or DIALS are not exact. In order to align the orientation matrices to pick up as many spots as possible with the smallest bandwidth, a method was developed to minimise the spread of the apparent nest of Ewald spheres on which the reciprocal lattice points were illuminated, by rotating the crystal in the two dimensions orthogonal to the beam. First, spots were over-predicted by integrating using the initial orientation matrix with a wide bandwidth of 3.5%, and reflections were selected if they had an intensity which exceeded 150 ADU (analogue-to-digital units). In subsequent refinements of the software, this was a changeable parameter which could be expressed in terms of ADU or <I/sigI> ratio. This allowed the program to integrate spots which may not have been picked up by spot-finding from other programs, allowing it to refine the crystal orientation with a larger amount of information than other indexing refinement algorithms. Each reciprocal lattice point sits on an Ewald sphere of a particular radius, corresponding to a given wavelength. This is referred to as the Ewald sphere wavelength and can be calculated from the transformed Miller indices in reciprocal space, as per Equation 4.1, where h_t, k_t and l_t are the transformed coordinates of a given Miller index in reciprocal space.

$$\lambda^{-1} = \frac{h_t^2 + k_t^2 + l_t^2}{-2l_t} \tag{4.1}$$

At this stage rlps were modelled as delta functions. Orientation matrices were refined to minimize the spread of wavelengths required to satisfy the Bragg condition for this ob-

served set. Orientation matrices were rotated about two axes orthogonal to the beam, each initially rotated by 0.5°, then halving the interval step upon finding local minima. For each orientation a histogram was calculated of the number of reflections versus the wavelength at which they were excited. The target function initially maximized reflection count and minimized distribution spread, where the parameters are altered to result in a minimization of the value of the target function. After the interval step had been halved three times, the target function was changed to be the quality of fit to a Gaussian model of wavelength distribution. Refinement was terminated when the rotational step size reached 0.001°. This leads to peaks centred upon the X-ray beam wavelength which are a hallmark of a well-refined orientation matrix.

4.3.10 Detector geometry refinement

A different, significantly more coarse method was used to determine the geometry for the CPV17 structure published later (Ginn et al., 2015b), which is summarised here, however this has now been superseded by the method outlined in Chapter 2.

The CSPAD used in this experiment is as described in Chapter 2. Each of the 32 ASIC panel pairs can be shifted by an X and Y coordinate, creating 64 variables at a bare minimum. On top of this, there could be additional variables for rotation in the three axes. During the initial successful attempt at solving the structure from 8500 eV energy X-rays, the integration step 'locked' onto the nearest pixel with the highest counts. This was sufficient for calculating a structure of CPV17, but would cause an inflation of reflection intensities where the strength of the signal was not much above noise in the image. This would disproportionately affect high resolution reflections which are naturally of weaker intensity. To better treat these high resolution and other weak reflections, it was necessary to better determine the geometry of the detector in order to have greater certainty in the positions of the reflections.

The previous method of correcting for geometry errors was to take the original predicted position from indexing, and examine a window of a determined width and height

for a pixel with the highest intensity around this position. One can take this pixel as the centre of the shoebox for integration. By plotting the average shift of reflections within the examined window, one can choose the position which covers the greatest number of shifts, and this can be assigned as the (x, y) shift for a given panel. Unfortunately although this improves the prediction of integration windows, it resulted in mis-fitting with the wrong parameters, and no strong improvement of the indexing rates.

4.3.11 Refinement of beam centre, wavelength and detector distance

Refinement of the X-ray wavelength and sample-to-detector distance is intrinsically linked to the detector geometry refinement method. If the average shifts are all biased towards one direction, this suggests that the beam centre can be moved to compensate. Similarly, the detector distance can be partially inferred from the pattern of deviations in high resolution panels. If the detector distance used to integrate is actually closer towards the crystal than the true value, the reflections seen at high resolution will shift to a position which appears to be higher resolution than is actually the case. The wavelength itself is intrinsically linked to the scale of the unit cell, and is also closely coupled to the detector distance. However, the detector distance and wavelength correlation break down at higher resolution, as the change of the Ewald sphere radius causes a sharper change in spot position at higher resolution than a change in detector distance. During the initial analysis, the subtle differences regarding geometry refinement were lost on me, and therefore this is an imperfect implementation and should be considered superseded by Chapter 2.

4.3.12 Integration methods

Spot intensities were initially calculated by simple summation integration, where predetermined 'foreground' pixels were summed and the estimated background within the foreground pixels were removed using surrounding background pixels to estimate the background intensity. This is referred to as 'simple summation'. The 'complex shoebox'

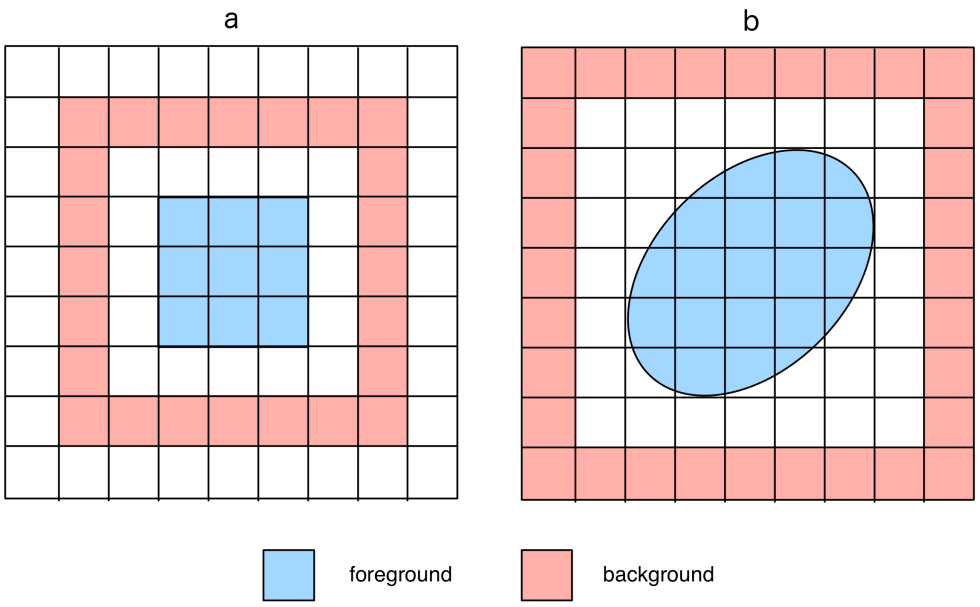


Figure 4.2: Foreground and background masks used to calculate the integrated signal of each reflection, using the simple shoebox during initial orientation matrix refinement (a), and an elliptical shoebox based on the energy bandwidth (b).

is a variation on this which takes into account that the differing spread of wavelengths causes the spots at high resolution to elongate along the vector joining the spot to the beam centre on the image. Based on this observation, integration boxes can be remodelled as ellipses which mimics this spot divergence. The difference in the nature of these integration methods is highlighted in Figure 4.2.

The background plane fitting method was originally developed previously (Fry et al., 1993) although documentation on this method is not extensive. This involves fitting a plane to the background of the image which was re-implemented as a least-squares fit. This produces a plane in the format $p\boldsymbol{i} + q\boldsymbol{j} + r$, where p, q and r are constants and $\boldsymbol{i}$ and $\boldsymbol{j}$ are the horizontal and vertical axis vectors for the image. Outlier rejection can also be performed on the background pixels prior to the calculation of the least-squares fit plane. This removes pixels which are more than a set number of standard deviations away from the mean pixel value for the background pixels. This allows a mathematically derived

estimation of the background to be made by calculating the volume of the geometrical shape produced by this equation within the area of the foreground pixels.

4.3.13 Resolving the indexing ambiguity

The cubic I23 space group has a higher lattice symmetry than the underlying point group. This means two alternative indexing solutions are able to satisfy the same positions of spot-finding results, but will result in differing intensities. For the initial publication of this work, a publication had been recently published on an algorithm to break this indexing ambiguity (Brehm and Diederichs, 2014). This was reimplemented and applied successfully to resolve the indexing ambiguity without the aid of the synchrotron data set.

4.4 Results

4.4.1 Transition from room-temperature to cryo-temperature involves a 1 Å contraction of the unit cell dimension.

Powder diffraction patterns were successfully collected by Geoff Sutton at beamline I04 at Diamond (Figure 4.3), and I carried out the analysis. At room temperature, from a total of eight measured rings, the unit cell dimension was calculated to be 106.1 ± 0.17 Å. For the cryo-cooled crystals, the unit cell was 104.9 ± 0.22 Å, with each mean and standard deviation calculated from eight rings. On the assumption that the unit cell for the room temperature crystals is closest to that which would be observed in the XFEL experiments, the unit cell was observed to contract by about 1.2 Å, which was expected to lead to a reduced amplitude and phase coherence between the two data sets with increasing resolution. However, analysis as detailed in Chapter 2 led me to realise that my unit cell dimension was alarmingly off by 0.5 Å from this initial result. This taught me not to perform data analysis to find subtle shifts using printed sheets, pencils and rulers.

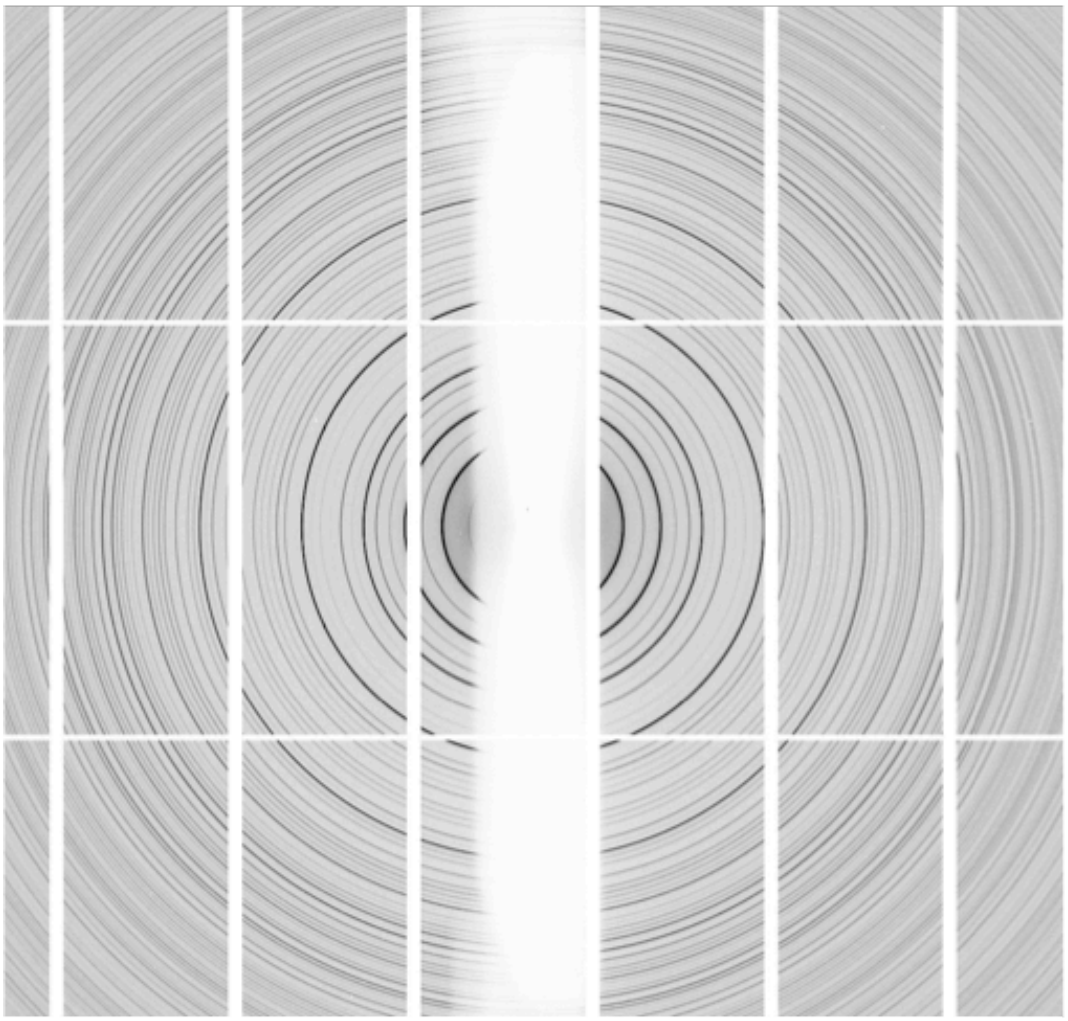


Figure 4.3: Powder diffraction pattern from CPV17 crystal slurry at Diamond beam line I04. Ring radii were estimated with a ruler and fit to the nearest Miller index of incrementing resolution rings. The results were generated from an average of eight powder rings.

4.4.2 Diffraction was obtained to high resolution.

Overall, 144,803 images were collected of which 11768 were selected as hits by the software package *cctbx*.xfel. 262 of these were at the longer detector distance at 101 mm, where data extended to approximately 1.75 Å, and the rest were at the shorter distance of 91 mm with data extending to around 1.6 Å.

4.4.3 Different indexing methods index image hits with varying degrees of success.

Nicholas Sauter performed indexing using *cctbx.xfel* and successfully indexed 49% of crystals (so 5767 crystals). The indexing rate was higher at lower flow rates. Images from runs with a low indexing rate had significantly more spots on each image than those of a higher indexing rate, corresponding to an increase in the number of lattices present in the beam at the time of data collection. The conclusion drawn from this was that the indexing algorithm employed by *cctbx.xfel* did not respond well to images with multiple lattices and the failure rate was higher for these images.

My own analysis with the DIALS software, which is capable of indexing multiple lattices due to the work of Richard Gildea (Gildea et al., 2014), showed an increase in the number of indexed lattices when allowing indexing of multiple lattices. This program successfully found an average of 1.35 crystals per image of those which were indexed, which was an average of 0.65 lattices per image.

During the course of my PhD I developed an indexing algorithm to tackle certain tough data sets, which is explained in Chapter 3. This algorithm found 1.5 crystals per image for the 262 images (the entirety of the hit-finding results from the first run), but was not used for the structure solution shown here.

4.4.4 Fitting a plane to the background subtraction provided no benefit in quality of integration.

Indexing solutions were used to predict the reflection locations on the detector for integration. Unfortunately, at low photon counts, individual pixels on the CSPAD appear to have highly variable gains. Although this does not appear to extrapolate to high photon counts (Anton Barty, personal communication), this variable gain has a strong influence on the background subtraction in low-background experiments where few photons hit the detector outside of Bragg peaks. The non-random nature of the noise means that fitting a plane to the background of the CSPAD is largely unhelpful and the difference in R_{split} between simple summation and plane-fitting background subtraction was negligible.

4.4.5 Initial orientation matrix refinement produced higher quality orientation matrices.

Initial orientation matrix refinement minimising the spread of Ewald spheres on which strong reflections were integrated was successful, resulting in an average shift of 0.075° from the orientation matrices determined by *cctbx.xfel*. An example of distributions of wavelengths before and after initial orientation matrix refinement is shown in Figure 4.4. The prediction of these reflections after refinement is shown in Figure 4.5.

4.4.6 Phases were determined by molecular replacement using an average structure from seven related cypovirus polyhedrin structures.

Seven related cypovirus polyhedrin structures were used for molecular replacement, which were not publicly available at the time of publication of this study. These structures have since been released (Ji et al., 2015). Due to the very poor sequence homology between the structures (the seven structures each have between 9 and 18% sequence homology with CPV17), no one structure was capable of supporting molecular replacement. Super-

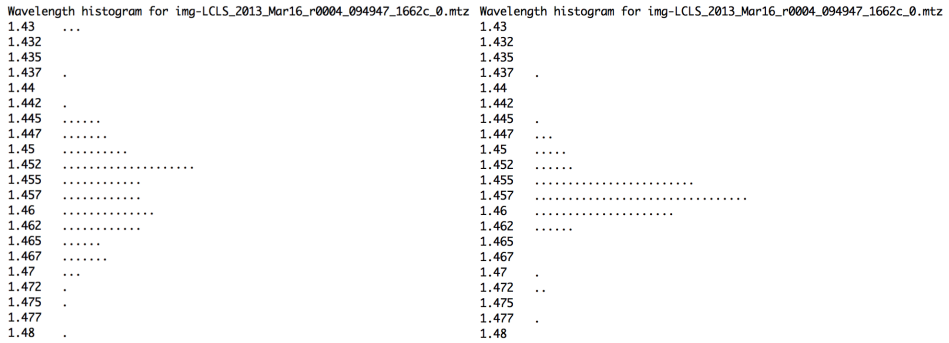


Figure 4.4: Two outputs from the refinement protocol are shown above, with the axis from 1.43 to 1.48 representing the corresponding wavelength of the Ewald sphere in Angstroms. The number of dots after each number represent the number of strong reflections which were recorded at a given wavelength. On the left is the result before additional rotation of the crystal to improve the distribution, and on the right is after an additional rotation.

position of the polypeptide main chain was performed with Coot and the PDB file was manually edited to reduce the occupancy of each atom to 1/7th of the original. This was used as a search model in Molrep. Due to the low solvent content of this crystal (around 35%) the solution could only be obtained by removing the check against packing constraints. The final result without packing constraints had a TFZ score of 7.3, and produced initial R factors of around 50% and was not amenable to automatic rebuilding at first. The quality of the map is shown in Figure 4.6. The majority of the chain had to be rebuilt manually in Coot, after which Buccaneer was used to build in a few remaining residues.

4.4.7 Data collection and structural refinement statistics.

The refinement statistics are reported as they were published in Nature Communications in Table 4.1. Note that the first run, at a detector distance of 101 mm, was excluded from this analysis which led to the number of indexed crystals in this table being smaller, but this effect on the data quality is fairly negligible. These coordinates were published under PDB code 4S1K.

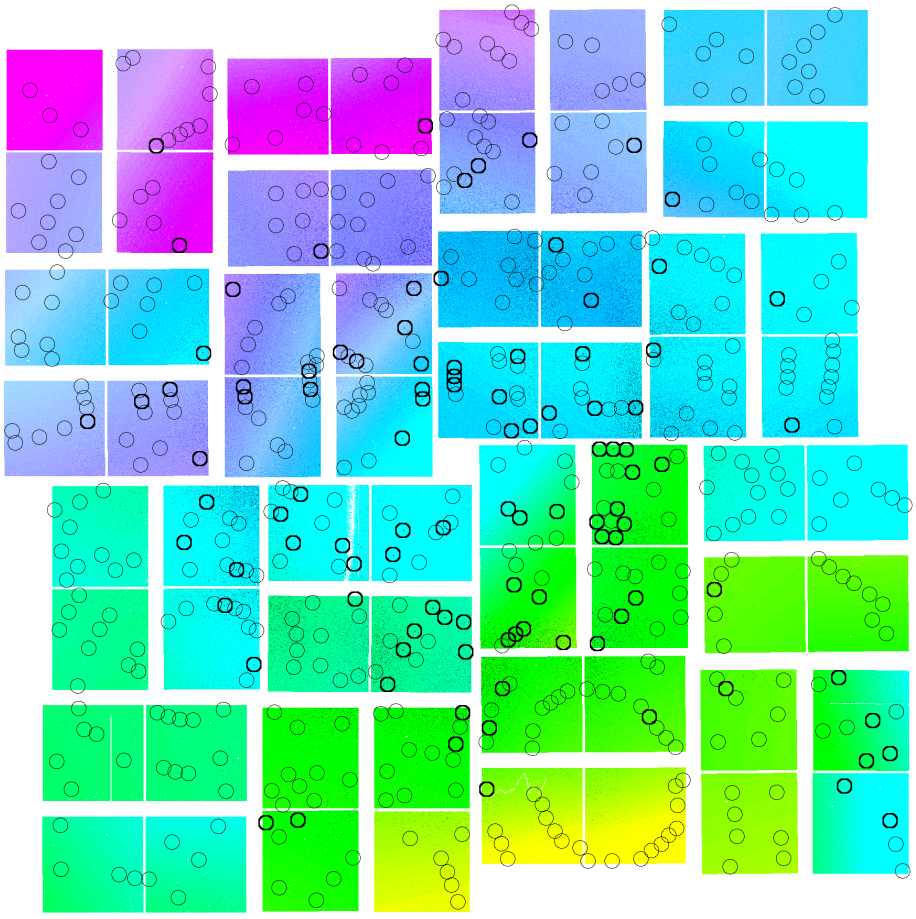


Figure 4.5: Circling of reflections which have been predicted by an indexing solution after initial orientation refinement, using an inflated bandwidth of 0.4% and an inflated reciprocal lattice point size of 2×10^{-4}Å^{-1}. The circles which are thicker are predicted reflections which, when integrated, has an intensity threshold above that which is considered a 'strong' reflection.

Table 4.1: CPV type 17 XFEL data collection and structural refinement statistics.

Data Collection	
Temperature (K)	293
Number of hits	11406
Indexed patterns	5554
Space group	I23
Cell dimensions	
a = b = c (Å)	105.6
$\alpha = \beta = \gamma$ (°)	90
Resolution (Å)	28.36 – 1.75
Completeness (%)	100.0 (100.0)
Redundancy	52.4 (11.1)
Rsplit (%)	11.8 (58.4)
CC 1/2 (%)	98.0 (38.6)
Refinement	
Resolution (Å)	28.36 - 1.75
No. reflections	20,122
Rwork/Rfree	12.2/15.4
No. atoms	
Protein	1914
Ligand/ion	32
Water	174
B factors	
Protein	23.7
Ligand/ion	34.4
Water	32.7
R.m.s. deviations	
Bond lengths (Å)	0.010
Bond angles (°)	1.39

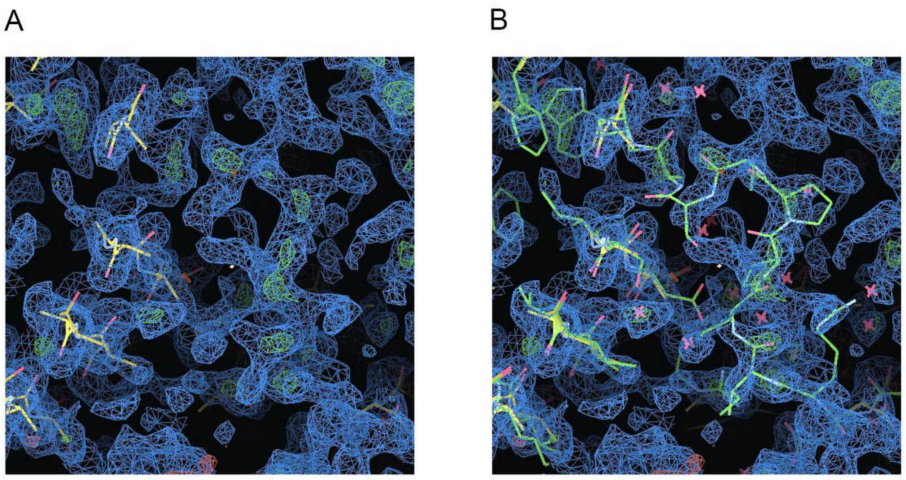


Figure 4.6: On both sides, the map from molecular replacement is shown. (A) one of the initial poly-alanine models is shown, without the additional loop found in CPV17 and (B) after building the final model, superposed to show the additional information contained in the initial electron density map.

4.4.8 Phases were transferred to a data set collected at Diamond I24 beam line.

Once the phases had been determined by molecular replacement using the XFEL data set, the phases could be transferred to the synchrotron data set. It was noted that the amplitudes between both data sets correlated well to high resolution, suggesting the unit cell contraction on cryo-cooling did not significantly alter the fractional coordinates of the atoms within the unit cell (Figure 4.7). This was supported by the observation that, after refinement of the atomic coordinates in the contracted unit cell, the protein appears to have behaved like a spring with a uniform contraction throughout the protein.

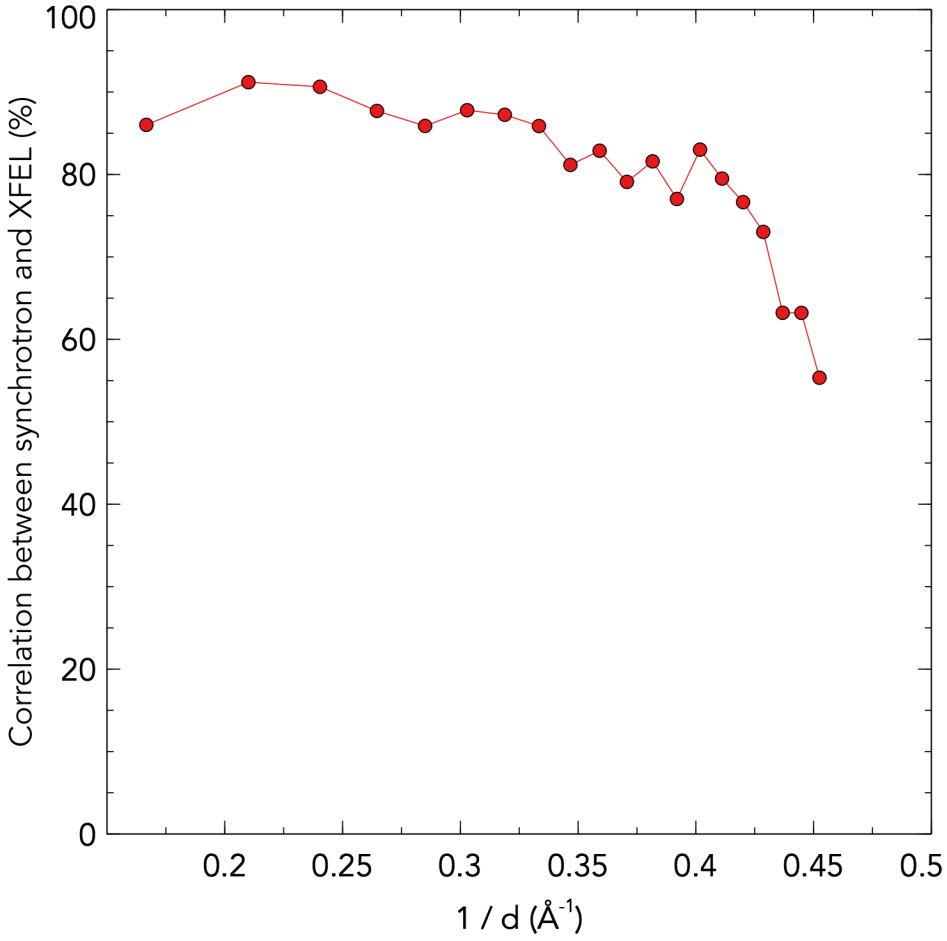


Figure 4.7: Correlation between intensities derived from CPV17 from room temperature diffraction at LCLS vs. diffraction from cryo-cooled crystals at microfocus beamline I24 to the resolution limit of the synchrotron data at 2.2 Å.

4.4.9 A disulphide bond present in the room temperature data set was broken when the crystals were cryo-cooled.

From the comparison of the I24 beam line cryo-cooled Diamond synchrotron structure, and that from the room temperature XFEL experiment, it is appeared that the act of cryo-cooling the sample had caused ablation of the partial occupancy of a disulphide bond between Cys142 and Cys155 of neighbouring symmetry-related molecules (Figure 4.8) and is supported by the anomalous signal in the data set (Figure 4.9). This is corroborated by a gel electrophoresis experiment carried out by Geoff Sutton (Figure 4.8b). The mass of a single cypovirus polyhedrin monomer is 25 kDa. Running denatured cypovirus polyhedrin crystals shows a number of dimers and trimers suggesting that some monomers are covalently linked. The occupancy of dimers and trimers are eliminated by adding β-mercaptoethanol, which reduces disulphide bonds.

4.5 Conclusions

This work was published in Nature Communications and was the subject of an interview and highlight article (Larochelle, 2015).

In the first couple of years of publications concerning XFEL structure determination and prior to the start of my DPhil, the number of patterns used to solve a structure were often in the hundreds of thousands (Chapman et al., 2011; Redecke et al., 2013). This was because merging data was performed using the so-called 'Monte Carlo' approach where very large errors from imperfect models of a diffraction experiment were assumed to average to intensities which would be proportional, relative to intensities of other reflections at least, to the correct value. This provided at least a baseline on which it was known that a structure could be solved, but is not permissible in the long term due to the cost and labour involved in producing enough sample to allow these levels of data collection. The battle to reduce sample cost can be fought on a number of fronts - namely sample delivery and by computational methods.

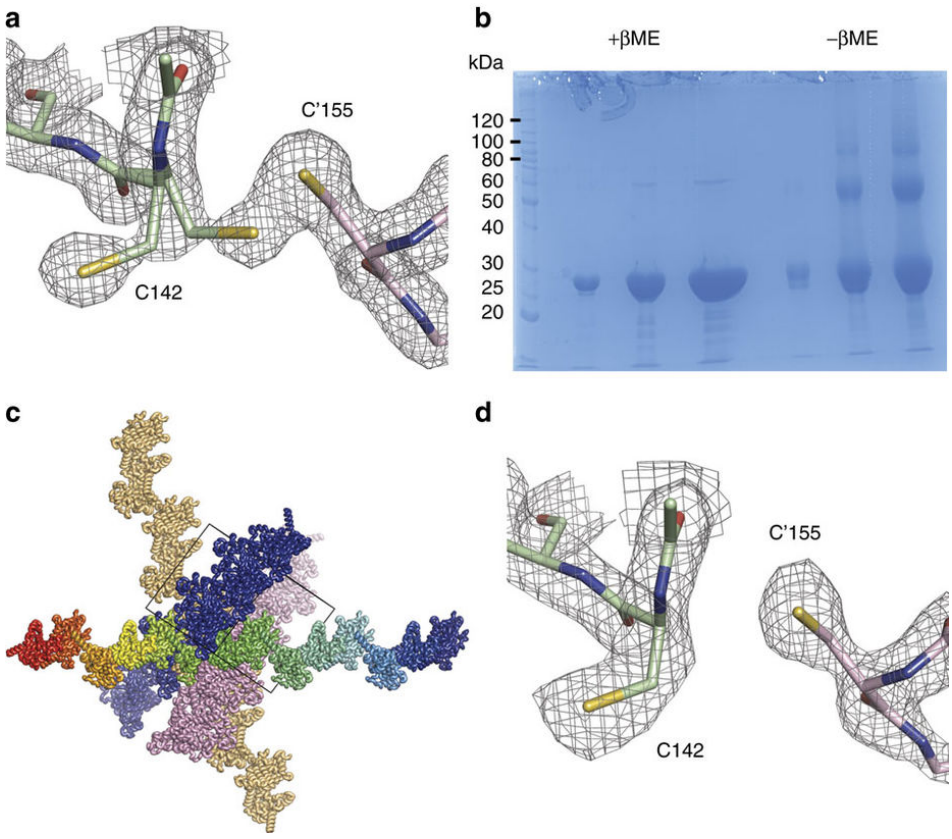


Figure 4.8: (a) Disulphide bond shown in partial occupancy in the room temperature XFEL structure and (d) ablated entirely in the cryo-cooled synchrotron structure. (b) SDS-polyacrylamide gel electrophoresis analysing CPV17 crystals with and without the reducing agent beta-mercaptoethanol. Without the reducing agent (disulphide bonds remain intact), a significant population of dimers and trimers are present in the population. With the reducing agent, disulphide bonds break and the full protein component is present at the monomeric mass. (c) Four helical strings of disulphide-connected monomers are shown. The connections are shown in the left-to-right helical strand with different coloured monomers, and additional helical strands are shown in solid block colours (pink, cream, blue). Figure prepared by Geoff Sutton and used in the publication in Nature Communications (Ginn et al., 2015a).

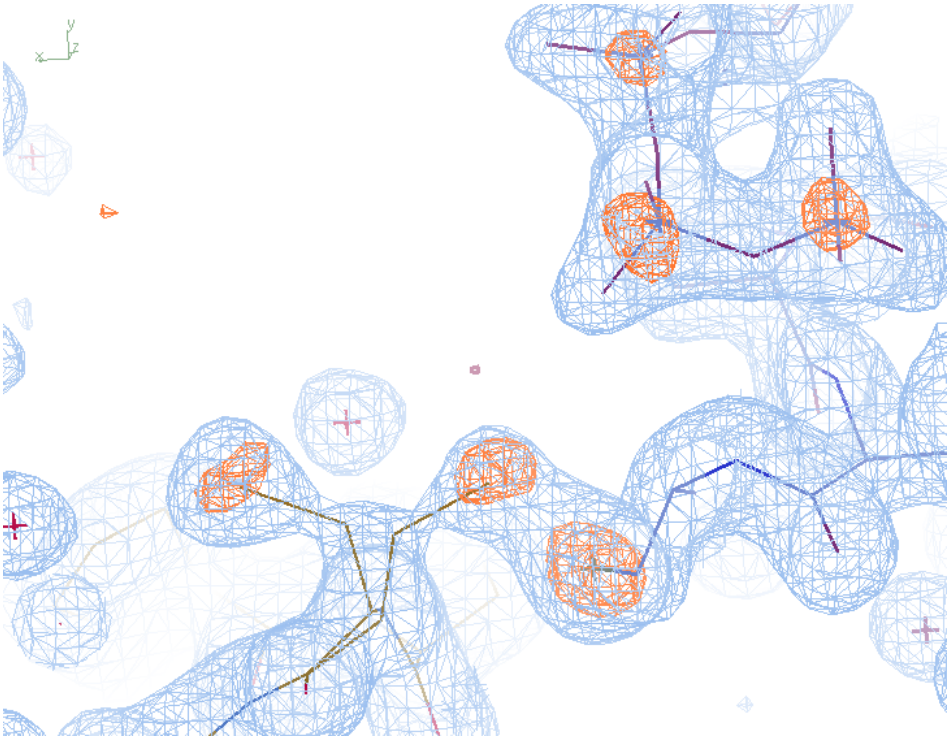


Figure 4.9: Anomalous signal calculated by ANODE (Thorn and Sheldrick, 2011) (drawn at 3.5 sigma in orange, with 1.0 sigma weighted observed map in blue) from 34,000 images supports two conformations for Cys142.

By the end of my DPhil, structure solutions were rarely performed with hundreds of thousands of images unless a time-resolved or anomalous difference was being found - and even then, the work on imaging a time-resolved myoglobin reaction used an average of only 20,000 images per bin to pull out time-resolved differences (Barends et al., 2015), and the work on phasing using combinations of SAD, SIR or SIRAS to phase data sets brought this down to less than 10,000 patterns (Nakane et al., 2016b). This has been mostly due to improvements in the most commonly used software suite over this time, CrystFEL (White et al., 2016). Some of these improvements (either released or in progress) have been inspired by the work in this chapter and the next.

4.5.1 Reception by the XFEL community

I attended the Gordon Research Conference on Diffraction Methods 2014, with the kind support of the lab of Elspeth Garman, presenting this work just after its completion as a poster. It was selected for a 15-minute talk on the final day and seemed to be received well. This also brought the concept of improving computational data interpretation into the limelight. This led me to be invited to the Ringberg conference on Structural Biology with XFELs, organised by Ilme Schlichting, in February 2015, and also invited to give a talk at the 3rd BioXFEL conference in 2016 by Edward Snell. CrystFEL has since incorporated this algorithm and referred to it as 'prediction refinement' (White et al., 2016). In fact, it combines elements of the algorithm above, and also work by Wolfgang Kabsch (Kabsch, 2014), which provides a notable improvement in the implementation. The work by Kabsch, more sensibly, works in minimising the distance to the Ewald sphere in reciprocal coordinates. By contrast, I worked with the differences in the inverse Ewald sphere radii (i.e., the illuminated wavelengths). This is fine for the majority of the reflections, but will break down at low resolution as the calculation of the wavelengths become extremely volatile. This was an oversight in my own implementation. By the time I became aware of this, the essence of my algorithm had been merged into CrystFEL, and so correcting my implementation never reached the top of my priority list and remains unimplemented.

Chapter 5

Post-refinement for still images.

5.1 Note to the reader

The subject of post-refinement quickly followed the result of the first publication on the structure of CPV17 (Ginn et al., 2015a). This represented another step in the process of minimising data required to solve a scientific question, and was the subject of a paper shortly following the first (Ginn et al., 2015b). The combined power of these two chapters thoroughly reduces the number of diffraction images required to pull out the anomalous signal of the sulphur atoms in the protein sequence. Portions of this text have been adapted from the second paper (Ginn et al., 2015b).

5.2 Introduction

5.2.1 Partiality of individual reflections as a source of error

Because of the short duration of an XFEL X-ray pulse and the negligible rotation in the beam during the time of crystal exposure, the Bragg reflections which satisfy the diffraction condition rely heavily on the non-monochromatic nature of the X-ray pulse energy to generate a nest of Ewald spheres which can illuminate Bragg peaks. However, this nest

of Ewald spheres cannot be assumed to be thick enough to fully immerse every reflection. Nor can it be assumed that different shells comprising the nest of Ewald spheres is uniformly bright (the distribution of photon energies is not a top-hat distribution). Hence, every reflection on the XFEL diffraction image represents some fraction of the full intensity of each Bragg peak, as reflections are only partially illuminated by the XFEL beam.

5.2.2 The concept of post-refinement

Post-refinement was coined by Schutt and Winkler (Schutt and Winkler, 1977), as a method of applying post-refinement to synchrotron data in a way which affects the estimation of the intensities where the edge of the rotation series cuts through a reciprocal lattice point. A method was proposed by Rossmann (Rossmann et al., 1979). In practice, the benefit of post-refinement for rotation data sets is less important than for still images as the rotation series usually covers a wide enough angle that the partial reflections can simply be discarded if there are enough fully-recorded reflections. Nevertheless it provides the benefit of a detector distance-independent method to refine the unit cell dimensions and can help recover partial intensities on the cusps of rotation series, and so is implemented in MOSFLM. The principle of post-refinement is that comparison with a reference data set allows the parameters of the diffraction model for a given image (or rotation series of images) to be refined, thereby more accurately predicting the unmerged intensities, which in turn feeds into better prediction of merged intensities.

The proof-of-principle that post-refinement is a potential method which can be applied to XFEL data was published previously (White, 2014), but this was performed using simulated data. This showed that there was sufficient theoretical information to perform recovery of original intensities given an arbitrary choice of partiality model. However, the partiality model chosen was not correct enough to show efficacy on real data; instead it was used to generate patterns from modelled reflections and then recover the full intensities of the reflections.

5.2.3 Post-refinement applied to real data

During 2015, a number of papers came out on post-refinement. This included the post-refinement algorithm behind the software suite PRIME (Uervirojnangkoorn et al., 2015), that behind EVAL (Kroon-Batenburg et al., 2015), that behind nXDS (Kabsch, 2014) (but nXDS has not been made available), and that behind cctbx.xfel (Sauter et al., 2014). My paper representing the contents of this chapter came out around the same time (Ginn et al., 2015b). My criticism of my published paper, with hindsight and experience, would be that I should have taken more care to show the improvement gained on each cycle by using metrics which would not be fooled by systematic errors. In the paper, I only sought to assess metrics such as $CC_{1/2}$ and R_{split} which are only measurements of random error, not systematic error, and would not protect against overfitting the results on a per-cycle basis. A simple thought experiment to show this is that if the reciprocal lattice point size were reduced to the point where only one or two reflections were illuminated per image, one could theoretically achieve an R_{split} of 0%. However, future analysis showed that the post-refinement algorithm was indeed robust and did not cause substantial overfitting, which is demonstrated in Chapter 8.

5.3 Methods

5.3.1 Architecture of post-refinement

In post-refinement, diffraction models for each crystal are refined against a reference data set, where the reference data set is updated after each cycle. The initial reference data set is created by merging all intensities with initial unrefined parameter values. This workflow is shown in Figure 5.1.

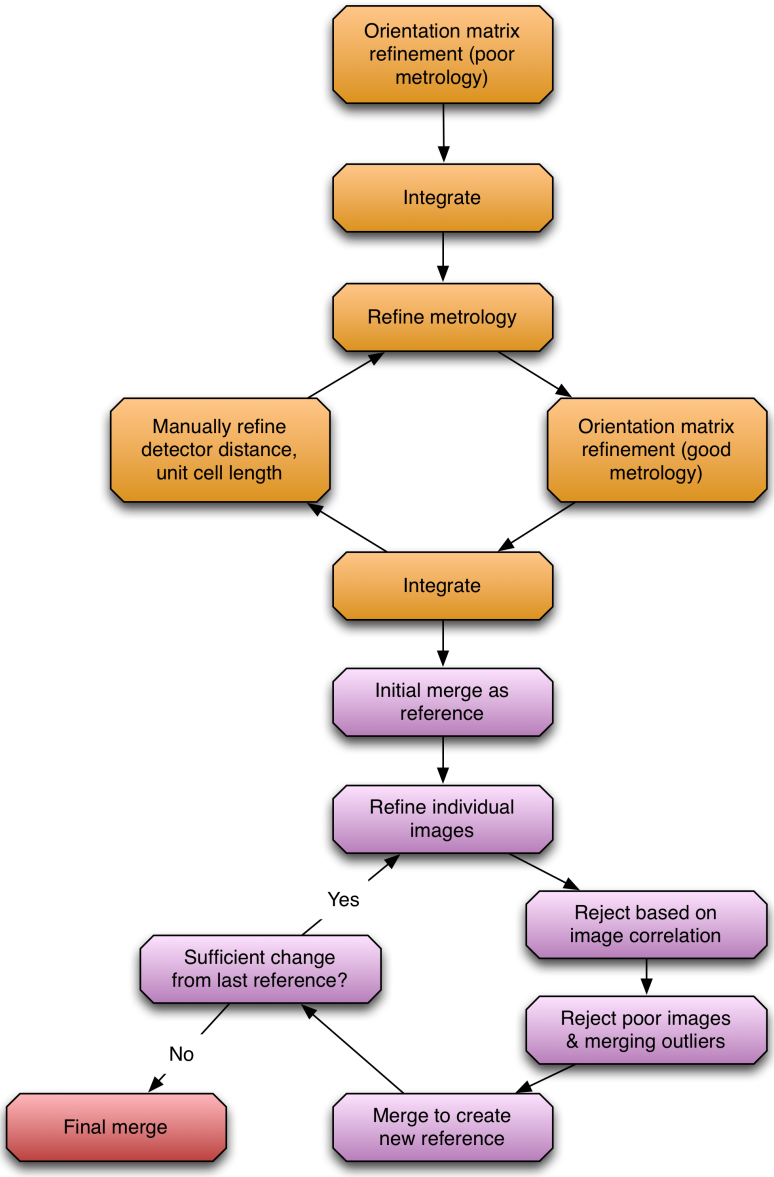


Figure 5.1: Schematic showing the flow of data processing for CPV17 as was carried out for the publication, including steps of orientation matrix refinement in orange boxes. The post-refinement steps are outlined in purple boxes.

5.3.2 Calculation of partiality

A number of variables contribute to the calculation of partiality, some of which are parameters attributed to the crystal, and some of which are parameters attributed to the X-ray beam. The parameters of the crystal include the unit cell dimensions (in the case of polyhedrin this is a single variable), the reciprocal lattice point size, the mosaicity and the orientation in the beam. The beam parameters describe the shape of the energy profile: the average beam energy (or beam wavelength), and two components describing the distribution: a standard deviation (corresponding to the bandwidth) and a super-Gaussian exponent, describing how pointed or flat the distribution is (Figure 5.2). The super-Gaussian distribution (E) calculated to describe the beam profile is shown in Equation 5.1 where σ is the energy bandwidth, μ is the mean wavelength and N is the super-Gaussian exponent. This can be evaluated at λ_p which is the wavelength at position p, a fraction of the length of a reciprocal lattice point ranging from 0 to 1.

$$E = \frac{1}{\sqrt{2\pi\sigma}} exp(\frac{-|\lambda_p - \mu|^N}{2\sigma^N}) \tag{5.1}$$

The reciprocal lattice point itself has a position in fractional coordinates corresponding to its centre in space, and is modelled as a sphere with a radius r. This is a fairly good approximation for CPV17, which produces fairly uniform cubic crystals, where anisotropy of reciprocal lattice points will be low. After discussions with Nicholas Sauter and Gwyndaf Evans, who contributed to my understanding of different ways to treat partiality, a model for the partiality of a reflection is calculated as a product of the beam function and the reciprocal lattice point. The reciprocal lattice point is modelled as a uniform sphere which intersects the Ewald spheres according to its position in reciprocal space. The cross-sectional area P is defined as a function of p, and can be derived from the cross-sectional area of a slice through a sphere at some point p as a fraction of the radius.

$$P = 4p(1 - p) \tag{5.2}$$

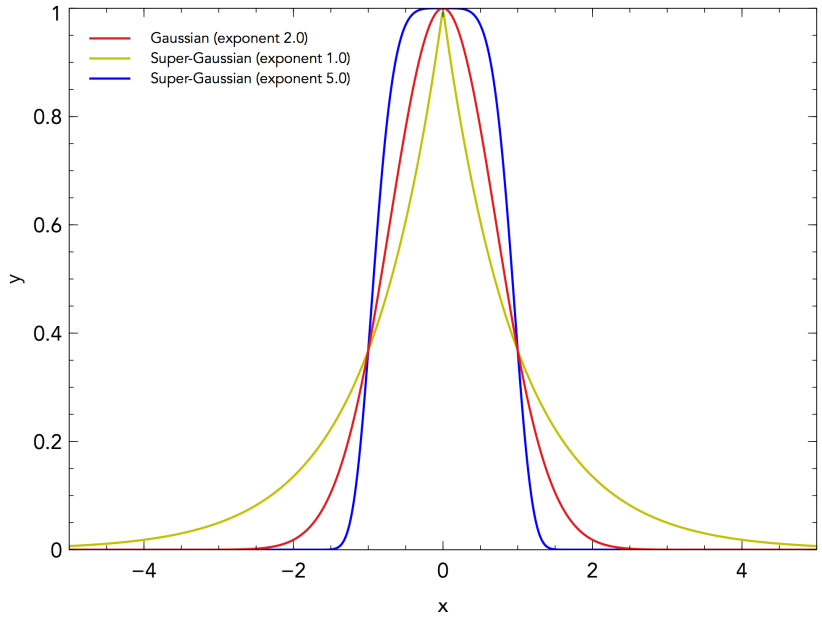


Figure 5.2: Effect of exponent on the equation for the super-Gaussian distribution (see Equation 5.1. Reducing the exponent below 2.0 tends towards a Christmas tree-shaped distribution, whereas increasing the exponent above 2.0 tends towards a top-hat distribution.

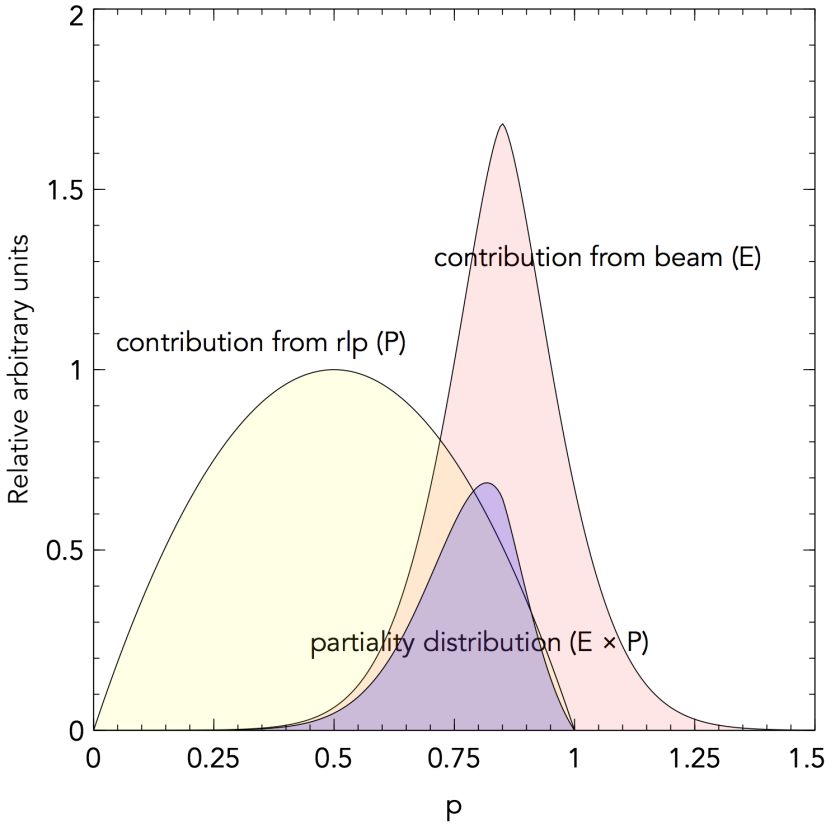


Figure 5.3: Diagram showing the relative contributions of the reciprocal lattice point shape (in yellow) and the contribution from the beam (in red). The multiplied contribution is shown, which is integrated to produce the partiality value (purple). Relative scales of these contributions do not have an effect on the final integration value after normalisation.

Each cross-section of this sphere interacts with the varying densities of Ewald spheres and the partiality ρ is the integral of the product of the two functions E and P. This is demonstrated pictorially in Figure 5.3.

$$\rho = \int_0^1 (E \times P) dp \tag{5.3}$$

5.3.3 Normalisation factors for partialities

At low resolution, the intensities of the reflections are likely to pass across the whole array of Ewald spheres as they are spaced very closely together. At higher resolution, a reflection may be in the centre of the nest of Ewald spheres but only cover a small portion as the Ewald spheres are spaced further apart. This means the maximum partiality that a reflection could have at high resolution will drop as it gets further away from the origin of reciprocal space. This may be a true phenomenon, or it may be more suitable for the partialities to be corrected during refinement. To correct the partialities, the partialities for each reflection are normalised so that the maximum partiality for a reflection at a given resolution is always unity. This is achieved by calculating a 'dummy' reflection centred exactly on the centre of the Ewald spheres. This reflection is calculated by setting the reciprocal space coordinates to $(h_{norm}, k_{norm}, l_{norm})$ where the coordinates are defined as follows:

$$h_{norm} = 0$$

$$k_{norm} = \sqrt{\frac{4d^2 - d^4\lambda^2}{4}}$$

$$l_{norm} = -d^2\left(\frac{\lambda}{2}\right)$$

These coordinates centre the dummy reflection on the Ewald sphere of wavelength λ and are then used to calculate the partiality at the maximum exposure to the beam. The partiality of the given reflection is then divided by this dummy partiality in order to normalise the range of partialities between 0 and 1. This prevents intensities from drifting by inflation over several cycles. The behaviour of the refinement program can be analysed with and without normalisation of partiality intensities, and assessed by comparing the observed amplitudes to model amplitudes.

5.3.4 Fixed and varied parameters

Cypovirus polyhedrin crystals are relatively uniform, with a mostly constant degree of diffraction per crystal, same order of crystal size and little variation in the unit cell dimension. As a result they are a very good test case for the maximum amount of information obtained with varying as few parameters as possible. One must avoid refining too many parameters which may not need to vary, which could introduce local minima away from the true solution due to misbehaviour of a certain parameter, or increase overfitting of the data to the model and decrease the quality of the structure. Some of the model parameters are hard or impossible to distinguish: the unit cell of the crystal and the wavelength are perfectly correlated as expansion of one is equivalent to contraction of the other; as a result the unit cell was fixed to 106.1 Å at the time of publication, and only the wavelength was allowed to vary. The beam bandwidth and crystal mosaicity are also highly, but not perfectly, correlated. If these are to be refined, they must be correctly distinguished. The bandwidth and exponent are also capable of altering to compensate for errors in the orientation matrix of the beam if they are allowed to vary. Small suitability tests were carried out for combinations of components to see which would refine well and which would refine poorly together and success measured by the quality of the $CC_{1/2}$ and R_{split} calculations on each merge.

5.3.5 Calculating scale factors

For certain calculations, including some of the possible target functions (see §5.3.6) and merging data, it is necessary to calculate scale factors between individual crystal data and a reference data set. This can be a linear scale, or incorporate some non-linearity to account for crystal B factors. In general, calculation of B factors leads to significant instability: B factors have a tendency towards a fall-off when calculating agreement against the reference data set. The simplest way to 'chase' the target function to improve the agreement with a noisy reference set is to downweight the poorest resolution shells. Modelling

B factors turned out not to become a focus of my efforts at this stage because CPV17 crystals are more uniform and less affected by crystal-to-crystal variability, so my model was a simple scale factor.

To determine a linear scale factor, the least squares fit gradient between the reference data set and the individual crystal data can be calculated, using the appropriate equation. This can make the assumption that the line of best fit passes through the origin. The gradient is equivalent to a scale factor for a given image. The intensities for an individual crystal can then be multiplied by the scale factor to make them comparable to the reference data set.

For a given chosen image C with corrected intensities I_c and partialities ρ_c, the calculation of the gradient with respect to a reference with merged intensities I_r is shown in Equation 5.4. The resulting scale factor K can then be applied to image C by multiplying all intensities by this value.

$$K = \frac{\sum (I_c)^2 \rho_c}{\sum (I_c)(I_r)\rho_c} \tag{5.4}$$

5.3.6 Choice of target function

Refinement calculated on individual images involves minimisation of a suitable target function. A number of target functions were chosen and tested for their success in producing low R_{split} values and high $CC_{1/2}$ values on merging. These target functions are based on:

- Correlation comparing intensities from an image against the reference data set (identifier: "correl")
- $Rfactor$ comparing intensities from an image against the reference data set (identifier: "rfactor")
- $Rfactor$as defined in the previous option, multiplied by the total illuminated intensity of the reflections (identifier: "r+int")

The target function *correl* between reflections is weighted according to the partiality of each reflection. Thus, a 10% partiality reflection will only count 10% as much towards the correlation as a 100% reflection. This is to reflect the greater errors in measuring low partiality reflections. The *logCC* target function is similar, except the logarithm is taken of intensities on both sides (and thus excluding negative intensities). This is meant to amplify the differences in weak data so that they do not get lost during refinement. The correlation between calculated partiality and the proportion of the merged data set is calculated without weighting, but must be cut off at an appropriate resolution which must also be determined. The $Rfactor$ between intensities in an image requires a pre-scaling event to get the two images on the same scale as each other, and is also weighted according to partiality. Similarly the r+int calculation must also put the image on the correct scale compared to the other images used in the calculation, and can be weighted by the partiality of every considered reflection and the scale of each image used in refinement.

5.3.7 Choice of minimisation algorithm

Various minimisation methods behave in different ways. Two main methods of minimisation of target parameters have been used to post-refine images: a simple step search, and the Nelder-Mead minimisation algorithm (Nelder and Mead, 1965). The step search simply steps forwards or backwards along a parameter with a given step length. If one of those parameters leads to a minimised target function compared to the original, it is changed. If the parameter remains unchanged, the step length is halved. This continues until the step length reaches a pre-determined concluding length, at which point refinement stops. The Nelder-Mead algorithm is more complex and includes (n + 1) test points of an n-dimensional vector, where n is the number of parameters to refine. These test points are evaluated and on each cycle the worst point is moved to try and minimise it. Test points can be reflected, expanded and contracted around the centroid of all test points. This is implemented in *cppxfel* for 25 cycles. The success of the two methods can be compared by measuring the quality of the $CC_{1/2}$ and R_{split} calculations on each

merge.

5.3.8 Outlier image rejection

There are a number of images within the cypovirus polyhedrin data set which are clearly problematic images, having failed to process correctly along some step of the refinement pathway, or were incorrectly indexed by *cctbx.xfel* or *DIALS*. These images should be detected and removed during the process of merging to create a reference data set. Images are rejected based on a number of criteria, which are customisable in the input file of the program. Images can be rejected based on:

- failing to meet a customised minimum correlation with the reference data set
- exceeding a customised maximum $Rfactor$ with the reference data set
- failing to meet a customised minimum number of refined reflections
- failing to meet a minimum correlation between the calculated and experimentally determined partialities, and
- having a rejection flag associated with that particular crystal

For polyhedrin crystals, the only active rejection criteria are meeting a minimum correlation (set to 90% for all merges except the final merge, which is set to 95%) and requiring a minimum of 100 reflections to be present after those below the partiality cutoff have been rejected. Images which contain less than 100 reflections are considered to have incorrectly refined, and the risk of overfitting the parameters of the image is much greater when the reflection count decreases. The reflection count may decrease below this value if the model predicts excitement of reflections which fall outside of the over-predicted region (which is calculated as described in Chapter 4). This could be caused by falling outside of the boundaries of the integrated reflections due to a wandering X-ray wavelength or significantly deviating rotation angles which are applied to the orientation matrix.

5.3.9 Merging intensities and error calculations

In order to create a reference data set for further cycles of refinement, or to produce a finished data set, the intensities for each Miller index from the set of XFEL images need to be merged to a single value (the "I" column in an MTZ file) and an estimate of the error provided (the "SIGI" column or sigma values). To do this, intensities for each image are partiality-corrected and scale factors have to be determined for each image and applied, as described in Section 5.3.5. Merging occurs by taking the weighted average of each partiality- and scale factor-corrected intensity. The weights are a multiplication of the relative exposed fraction of the rlp (so that lower partialities are up-weighted) and the scale factor (so stronger images are up-weighted).

The equation for the corrected intensity I_{corr} per Miller index, for a set of n images, is as described in Equation 5.5.

$$I_{corr} = \frac{\sum_{i=1}^{n} I_i \rho_i K_i \omega_i}{\sum_{i=1}^{n} \omega_i} \tag{5.5}$$

where

$$\omega_i = \rho_i K_i^{-1} \tag{5.6}$$

The sigma values are calculated as the standard deviation of all corrected intensities for a single Miller index, divided by the square root of the multiplicity for that reflection (with correction for sample size). Hence the sigma value calculated here is more akin to a confidence limit than a standard deviation, as in Equation 5.7. However, during the process of refining against the reference data set, these sigmas are not used to weight the calculation as this was empirically found to make data processing worse.

$$\sigma_{corr}^2 = \frac{\sum_{i=1}^{n} \omega_i (I_i - I_{corr})^2}{\sqrt{n-1} \sum_{i=1}^{n} \omega_i} \tag{5.7}$$

The legitimacy of this calculation is questionable, but appropriate calculation of sigma values has not been established at present. It is certainly the case that including sigmas which have been calculated this way results in the vast majority being rejected by SHELXC (Sheldrick, 2010) when preparing data sets for phasing. However, if the sigmas are artificially lowered so as to accept all of them, this can be circumvented, and phasing becomes feasible. Hence, all of these sigma values are treated with caution.

5.3.10 Outlier rejection of reflections

The number of systematic errors which contribute to deviations in the reflections from the true value, beyond that expected for simple counting statistics on the detector is significantly greater for XFEL data than synchrotron data. These errors may be derived from several sources: overlapping reflections from other unindexed patterns on the same image, deviations in the energy spectrum of the XFEL pulse from the super-Gaussian model used to describe them, geometry errors, detector artefacts, and heterogeneity of the crystalline samples.

In the publication of the post-refinement algorithm presented here, there were two presented ways in which outliers are rejected: on a per-image basis, and due to poor merging statistics across multiple images. The per-image basis of outlier rejection is achieved by generating a plot between the post-refined image intensities and the reference intensities and calculating the correlation coefficient. A similar plot is created for each individual reflection removing that reflection from the correlation calculation. For any which result in a significant jump in correlation, the offending reflection is removed, up to three per refinement cycle. This is not carried out for images which correlate better than 99%. The significant jump is defined as reducing the difference between the actual correlation and 100% correlation by 6% or more (for example from 90% to 90.6%). This was chosen by manual inspection of intensity-intensity graphs and the change in correlation when rejecting clear outliers. This was increased to an improvement of 15% some time after publication. This outlier rejection is applied immediately after post-refinement of a

given image in each cycle. The second rejection basis is applied at the stage of merging, and is only applied to reflections with a multiplicity of greater than 5. The standard deviation of the reflections are calculated, and reflections beyond 1.6 standard deviations from the mean are rejected. After a reflection has been rejected, they do not contribute to post-refinement in the next cycle. However, reflections rejected due to merging statistics are reconsidered on every merging cycle. These methods were confirmed to be helpful for the data processing as it stood in 2014. The sigma cutoff for the merging-stage rejection was chosen by assessing the quality of anomalous signal for each cutoff. It is not possible to judge the sensibility of rejection by checking the R_{split} as this will always reduce with rejection of outliers.

5.4 Results and Discussion

5.4.1 On each cycle of refinement, the R_{split} values decrease and $CC_{1/2}$ values increase.

On each cycle of refinement, the value of R_{split} decreases and the value of $CC_{1/2}$ rises. This is most marked when refining against an initial Monte Carlo merge rather than a previously refined set of intensities. The lowest R_{split} and highest $CC_{1/2}$ values are found at low resolution. The resolution cutoff for the CPV17 intensities was set to the beginning of the resolution shell where the $CC_{1/2}$ first dips below 0.3. For the data collected at 8500 eV this resulted in a resolution cutoff of 1.62 Å after post-refinement and the data collected at 9500 eV were cut to 1.44 Å . These are the results which were published in the paper.

For illustrative purposes to show the reduction in model errors, data from a single run of CPV17 crystals from the March 2013 data collection were reprocessed and integrated to 1.6 Å resolution and post-refined using the optimal strategy. Outlier rejection was not used for this illustration, as these would reduce the R_{split} and $CC_{1/2}$ values regardless of

their effectiveness in reducing systematic error. Figure 5.4 shows the refinement process per cycle, the decay in quality per resolution bin, and a before/after plot of the intensities from which the $CC_{1/2}$ figure is calculated.

5.4.2 The target function $Rfactor$ for individual images produces the best merging statistics.

Three target functions were tested for finding the best merging statistics: the correlation coefficient (target *correl*), the R factor with the reference (target *rfactor*), and the R factor multiplied by the total illuminated intensity of the reflections (target *r+int*). The *correl* target produced an R_{split} of 8.27% on the final cycle, the target *rfactor* produced that of 5.69% and the target *r+int* produced an R_{split} of 5.77%. The final merging statistics suggested that use of the $Rfactor$ was the best target function, and was chosen as the default for post-refinement, and that including the absolute intensity as a component in the target function was mildly unhelpful (using *r+int*).

5.4.3 CPV17 crystals are best described by a bandwidth-based model rather than a mosaicity-based model.

The differences between a mosaicity-dominated model and a bandwidth-dominated model are subtle. An example of each of these are shown in Figure 5.5. The two types of models are closer in agreement to each other at mid-to-high resolution than at low resolution (i.e., they have a tendency to converge on similar models at high resolution). The agreement between the estimated partialities and the calculated partialities is poorer at low resolution for the mosaicity-driven model than the bandwidth-driven model. This suggests, that at least for CPV17 crystals, a model with minimal mosaicity is most appropriate, to best emulate the estimated partialities at low resolution in the bandwidth-based model. When collecting synchrotron data on the I24 beamline on these crystals, it was also observed that the refined mosaicity was low, at around 0.03°, but for the XFEL ex-

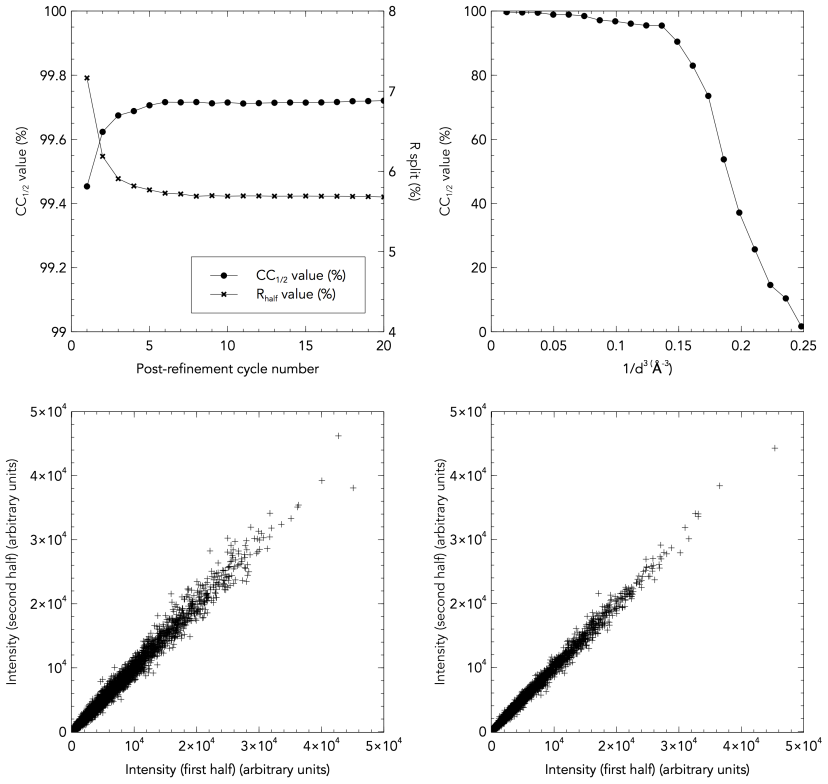


Figure 5.4: Data quality plots for post-refinement of CPV17 diffraction data using a reference to bootstrap the calculation. Top left: $CC_{1/2}$ and R_{split} values per refinement cycle. Top right: $CC_{1/2}$ per resolution bin (ranging to a maximum of 1.6 Å resolution), divided into bins of equal reflection counts. Point at which $CC_{1/2}$ starts to fall significantly is at 2.0 Å. Bottom left: plot of intensities from one half and intensities from the other after the first merge of post-refinement. Bottom right: similar plot of intensities from the 20th merge of post-refinement.

periments the mosaicity was set to 0°, which may reflect an increase in mosaic spread on cryo-cooling for I24.

5.4.4 The exponent, bandwidth and mosaicity for CPV17 crystals should remain fixed.

The energy of an XFEL pulse will vary slightly form shot to shot ('jitter'), and also the mosaicity of crystals may vary. However, allowing variation of the super-Gaussian exponent and the bandwidth of the beam caused both of these parameters to drift to larger values and the crystal rotated to compensate for the expanding size of the energy distribution. Similarly allowing the mosaicity to vary caused an expansion of this value and a rotation of the crystal away from the true value to compensate. It was found that post-refinement was not stable if any of these parameters were allowed to vary, and the R_{work} and R_{free} inflated for post-refinement calculated with these parameters left refineable. As such, the exponent, bandwidth and mosaicity are currently fully constrained during post-refinement for these crystals. The crystals were also observed to be homogenous regarding a lack of mosaic spread.

5.4.5 The Nelder-Mead algorithm finds a lower minimum than the step-search algorithm.

Using the Nelder-Mead algorithm produces a lower minimum in R_{split} than using the step-search algorithm. This was confirmed by running the post-refinement on the 5605 images from the 8500 eV X-rays, selecting the step search algorithm or Nelder-Mead algorithm with the most appropriate parameters for each. These used a well-refined starting data set as a reference created from 8500 eV data. R_{split} follows a similar descent but the evaluation for the Nelder-Mead algorithm is consistently lower than the evaluation for the step-search algorithm on every cycle (Figure 5.6). However, this appears not to be the case for all data sets, and should be checked on a data set-by-data set basis.

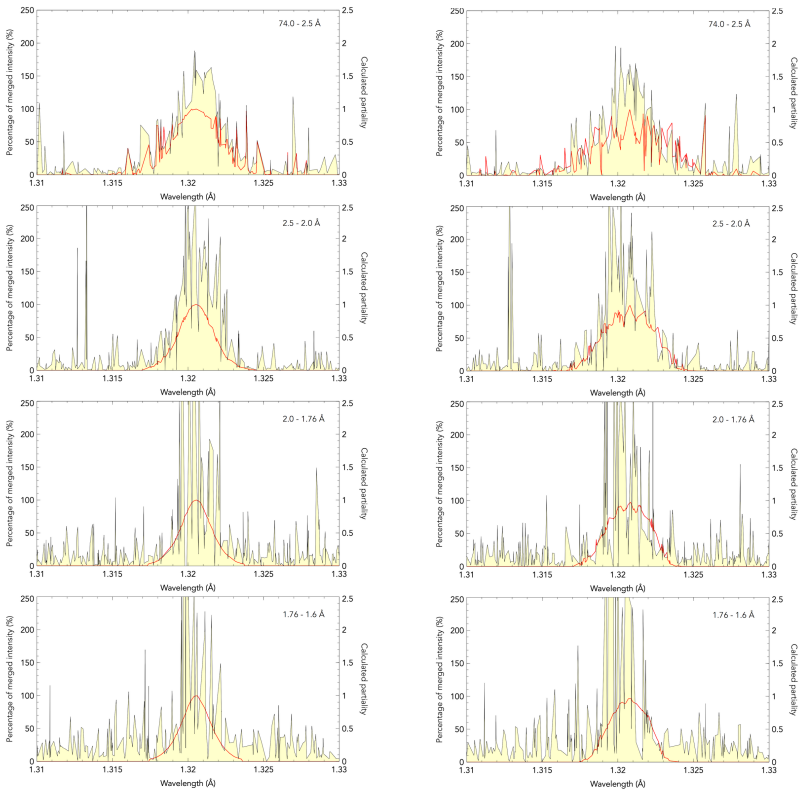


Figure 5.5: Plot of the calculated partiality, represented by a red line, alongside the estimated true partiality, represented by the yellow fill, for a single integrated and post-refined lattice. Each point on the line is a reflection for which the reciprocal lattice point is centred on an Ewald sphere wavelength plotted on the X axis. The estimated true partiality is calculated by dividing the scaled, integrated intensity for a single image against the merged intensity for that reflection. Reflections which are more partial should be a lower percentage of the corresponding merged intensity than reflections which are maximally illuminated. On the left is a bandwidth-based model with a bandwidth of 0.18% and a mosaicity of 0°, ranging from low resolution (top) to high resolution (bottom). The right hand side is a mosaicity-based model with a bandwidth of 0.07% and a mosaicity of 0.03°. Low resolution reflections agree more strongly for the bandwidth-based model than the mosaicity-based model

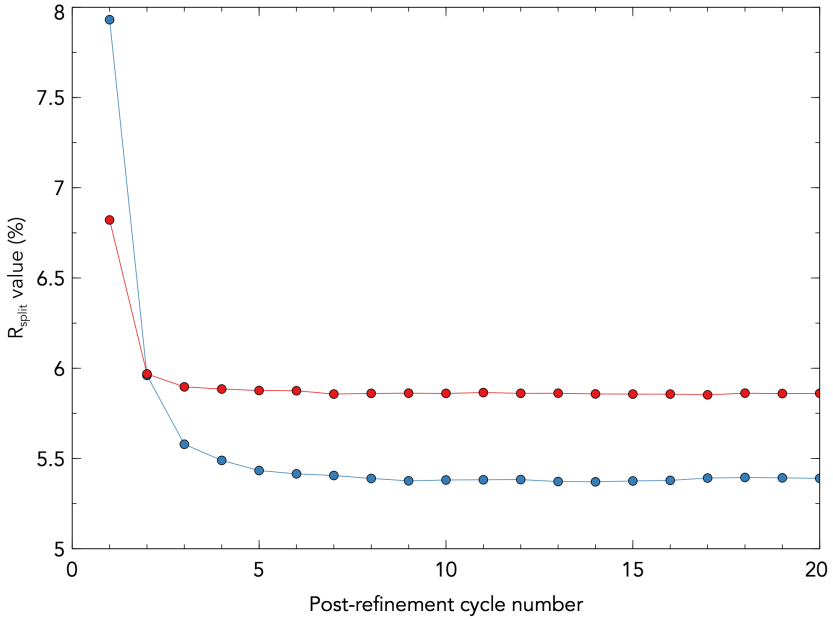


Figure 5.6: Plot of R_{split} for each cycle of post-refinement using the step search algorithm (red circles) or the Nelder-Mead algorithm (blue circles) to minimise the target function for each image, run on the full indexing results from a single run of data collection using 8,924 crystals indexed and integrated using all algorithms to-date. Initial reference is a previously well-refined data set from the 8500 eV data which is published on the PDB (code 4ZQX). The reduction in the R_{split} value is due to inclusion of a significant number of additional lattices than the results published previously, but only uses results from a single run.

5.4.6 Normalised partialities improve the values of R_{work} and R_{free}.

A sample of 1000 images from data collected at 8500 eV energy were refined using the best determined parameters and the Nelder-Mead algorithm, with and without normalisation of partialities for reflections as described in Section 5.3.3. Reflections were sliced into 200 parts at low resolution for calculation of partialities. The resultant data were then used to refine the atomic model in REFMAC5 to the resolution at the corner of the detector with anisotropic B factors and the R_{work} and R_{free} values were compared. The normalised partialities, although hard to justify in scientific terms, nevertheless produced the lower R_{work} and R_{free} results. The greatest improvement was at low resolution but R_{work} and R_{free} are inflated throughout the resolution shells when using non-normalised partialities. The plot of R_{work} and R_{free} per resolution shell is shown in Figure 5.7. The $CC_{1/2}$ value for 1000 images suggests the resolution should be cut at 1.71 Å , which explains why the difference between R_{work} and R_{free} is greater than with isotropic B factor modelling since the fitting of anisotropic B factors is not justified at this resolution. The breakdown of the relationship between R_{work} and R_{free} is evident in the graphs after this maximum resolution due to the reduced signal-to-noise. However, the significant decrease of both R_{work} and R_{free} show that reflections with normalised partialities behave cleanly compared to their non-normalised counterparts.

5.4.7 Low resolution reflections are exquisitely sensitive to accuracy of numerical integration of partialities.

Numerical integration is used in order to determine the partialities of individual reciprocal lattice points, and these are subdivided according to input parameters at low and high resolution. The finer the slicing, the more accurate the partiality will be calculated according to a particular model. It is clear that low resolution reflections required a greater slicing than high resolution reflections due to the sharpness of the distribution of Ewald spheres. A low slicing of partiality (e.g. 10 slices) caused low resolution reflections to be

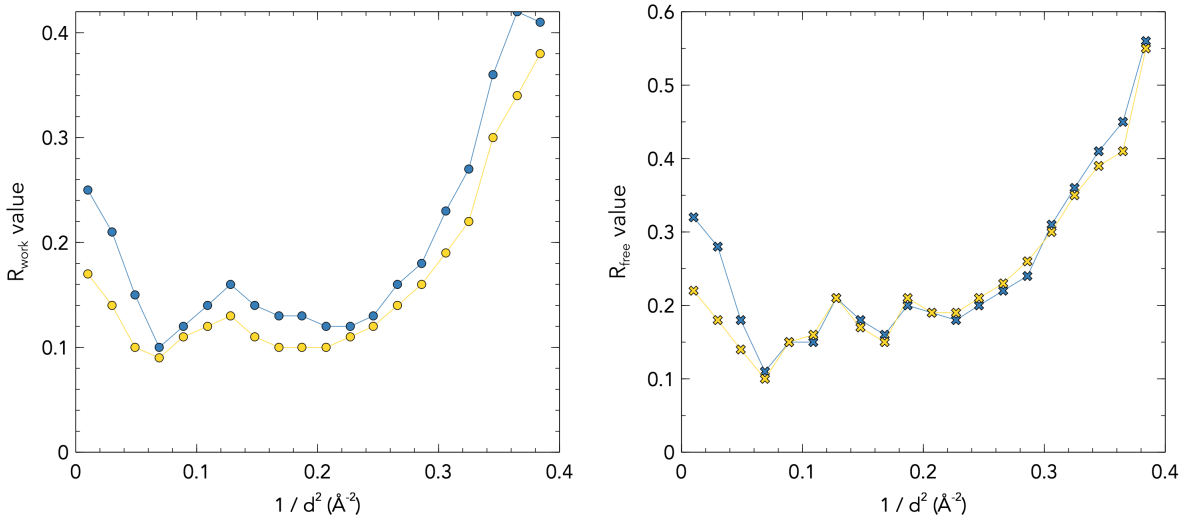


Figure 5.7: Plot of R_{work} and R_{free} across resolution shells for 1000 images of cypovirus polyhedrin, for (a) non-normalised in blue and (b) normalised partialities in yellow. Overall R_{work} and R_{free} for non-normalised partialities is 16.54% and 21.15%, whereas the respective values for normalised partialities is 13.12% and 19.18%. Output from values reported from REFMAC5 in the CCP4 suite.

poorly determined, and caused an extreme inflation of R_{work} and R_{free} at low resolution, rising to over 50%. Notably, these caused an overfitting of parameters as the R_{split} and $CC_{1/2}$ for low resolution were not inflated, and so these metrics cannot be used to determine that this pathology has occurred. A high slicing of partiality (e.g. 200 slices) greatly improved the low resolution reflections such that the R_{work} and R_{free} were closer to 25%. This effect then prompted the re-calculation of partiality by sub-dividing only the illuminated region into ten slices. This is essentially an auto-calculation, which allows the additional parameters for "number of slices" of each reciprocal lattice point at various resolution shells to be removed.

5.4.8 Post-refined data quality is sufficient to see anomalous signal from sulphur atoms.

The improvement of the maps after post-refinement can be seen not only in the anomalous signals, but also by eye (Figure 5.8). Post-refinement of 5787 images at 1.46 Å wavelength data collection or 7225 images at 1.31 Å data collection is sufficient to observe an anomalous signal for twelve of the thirteen sulphur atoms from cysteine and methionine in the crystal. The N-terminal methionine of the crystal structure is disordered and not visible in the anomalous map, and one sulphur atom from Cys142 is found in two conformations, one in a disulphide bond with Cys40 from a crystal contact and the other reduced and facing away from Cys40. The anomalous signal for this sulphur can be seen in both conformations. The signal can also be observed in the structure visually, as shown in Figure 5.9, reproduced from the published paper.

5.4.9 Optimal outlier rejection is on a per-image basis, not on a merging-stage basis.

The measure of R_{split} is a poor value to rely on to assess quality of outlier rejection, as it always decreases with additional rejection. Instead we must ask whether the data points

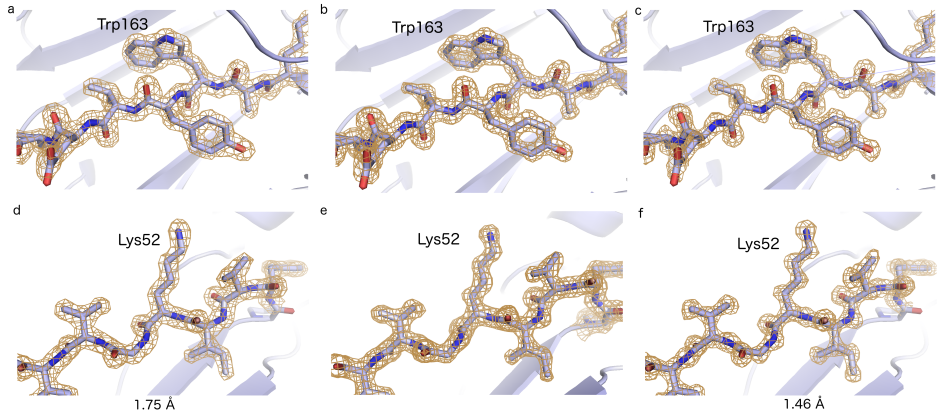


Figure 5.8: Electron-density map ($2mFo - DFc$) from reflections associated with PDB entry 4s1l to 1.75 Å resolution (Ginn et al., 2015a) associated with Trp163 (a) and Lys52 (d), their corresponding electron density in the initial merge for the higher resolution data set, (b) and (e), and the final presented 1.46 Å resolution structure, (c) and (f), at sigma level of 1.5. The N, C and O atoms are distinguishable compared with the 1.75 Å resolution structure. The high-resolution information beyond 1.75 Å and post-refinement of this data set appear to have separate sequential improvements on the quality of the electron density compared with the 4s1l structure.

which were rejected during post-refinement were contributing more signal or more noise, as we wish to improve the overall signal using post-refinement. Instead, the anomalous peak height of sulphur atoms is used as an independent metric to assess outlier rejection. Outlier rejection occurred via two methods: on a per-image basis, and on individual reflections outside a certain multiple of standard deviations away from the mean during merging.

On reprocessing of the data with the latest version of cppxfel in 2017, it was found that merging-stage rejection did not contribute to the quality of the anomalous peak heights, and actually reduced them when rejecting at 2.0 sigma, which was found to be a positive improvement at the time of publishing the paper on this work. With the software in its current state, increasing this number to higher values reduces the effect on the data merging statistics to negligible amounts. This is likely due to a reduction (or at least equalisation) of systematic errors over the course of the program development in the last four years. Correlation rejection added value to the anomalous differences when

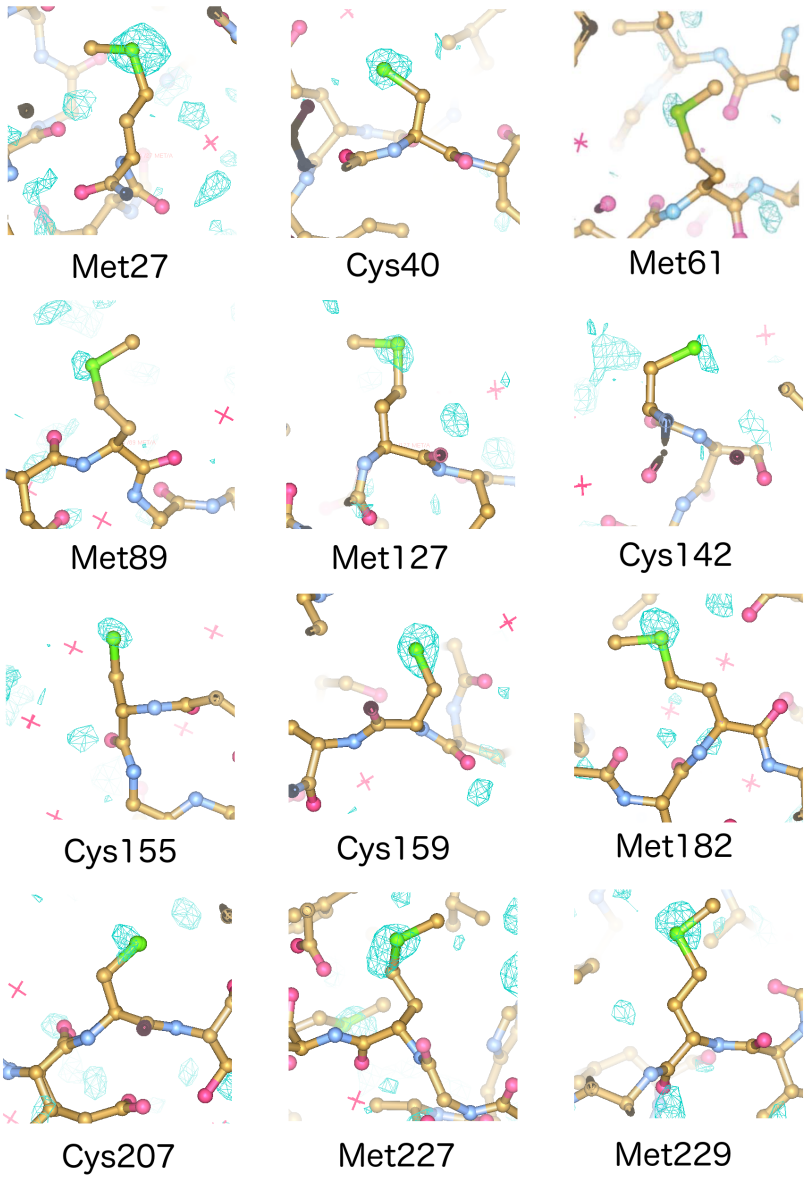


Figure 5.9: Anomalous signal contoured at 2.4 σ highlighted for twelve sulphur atoms as labelled. Highest density is 5.43 σ (Met182) and lowest σ is 2.53 (Met70 and Met89). Average peak σ is 3.87.

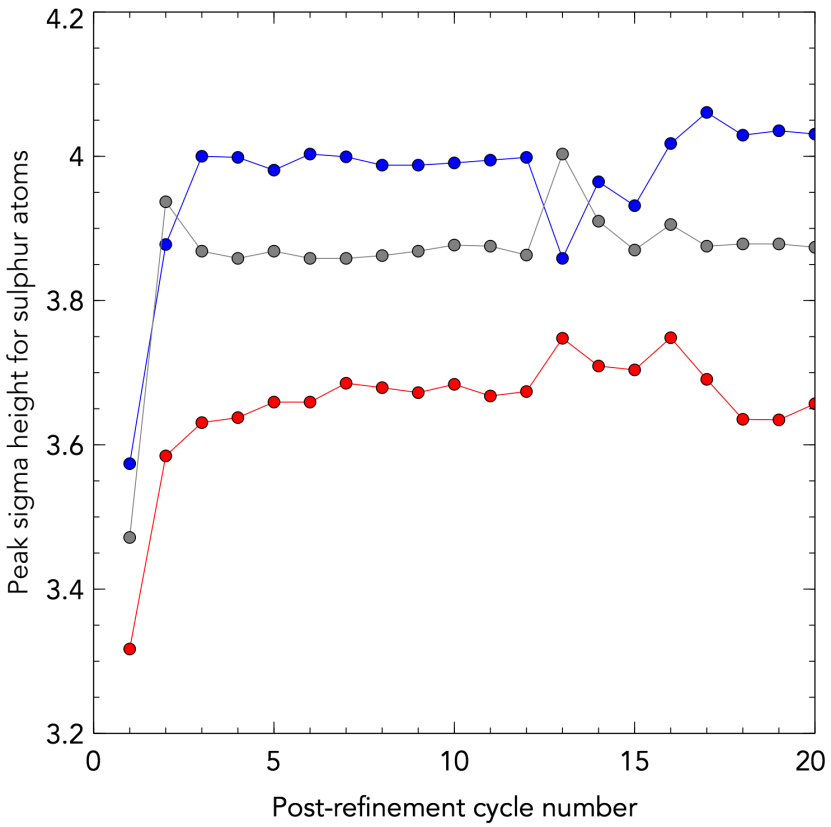


Figure 5.10: Plot of anomalous peak height as calculated using ANODE, part of the SHELX suite, for all 13 ordered sulphur atoms in the structure, for various rejection schemes: no outlier rejection (grey); rejecting beyond 2.0 sigma (red), and rejecting by correlation improvement on a per-image basis (blue). The latter provides an advantage over no rejection strategy.

an improvement of 15% or more is required (e.g., 90% correlation increases to 91.5% by rejection of a single reflection) and would reject 3.89 reflections per image. Anomalous peak height per cycle, for various merging strategies, are shown in Figure 5.10.

5.4.10 Overfitting can be protected against by checking independent measurements such as the anomalous signal peak heights.

Fundamentally, $CC_{1/2}$ and R_{split} measure the agreement between two halves of the dataset. However, it does not assess the agreement between the dataset and the "truth".

Hence, $CC_{1/2}$ and R_{split} can reduce substantially as a result of overfitting, but independent quality metrics would become worse (often by configuring parameters to minimise the total number of reflections illuminated). One such independent quality metric is the anomalous peak heights. Anomalous peak heights are calculated from the phases and the difference in intensities of Friedel pairs, but the post-refinement algorithm has no knowledge of the 'true' Friedel differences and in fact attempts to drive the paired intensities to the same value. Hence, any residual difference in the intensities which is capable of generating a meaningful anomalous signal, which improves from cycle to cycle, shows an improvement in the quality of the data. The anomalous signal from the sulphur atoms in the crystal was used as a demonstration of the quality of the data in a rudimentary way in the associated published paper, by comparing the pre-post-refinement and post-post-refinement values. However, it is much clearer and informative to plot the change in anomalous signal per refinement cycle, as in Figure 5.10. This can also reveal on which cycle the post-refinement begins to overfit, if it does at all. In this case, due to the very small Bijvoet ratio, the I+ and I- values were driven towards the same value. For larger expected differences in the Friedel intensity pairs, it might be more sensible to allow these to drive towards different values.

5.5 Conclusion

5.5.1 Reception by the XFEL community

The work on post-refinement caught the attention of Takanori Nakane from the University of Tokyo, resulting in a collaboration which forms the basis of Chapter 8. It was mentioned in a talk by Nicholas Sauter (LBL) at the European Crystallographic Meeting, and also mentioned favourably in a talk by Thomas White (CFEL) at the same meeting. I presented it in full at the Gordon Research Conference in 2016, and again at a XFEL satellite meeting organised by Allen Orville at Diamond Light Source. Thomas White chose the post-refinement algorithm presented here for inclusion in the more user-friendly software

package CrystFEL, for which I am providing some guidance and reference data to work from.

Chapter 6

Structure solution of bovine enterovirus (BEV) type 2.

6.1 Note to reader

I would like to thank Philip Roedig and Alke Meents as our primary collaborators in this experiment from DESY, Hamburg, as well as the full author list in the publication. The paper (Ginn and Stuart, 2016b) is cited here, but not to be treated as part of the assessment for a DPhil in Structural Biology, as the work within was the subject of my Master's dissertation. This has been published as a joint first author paper with Philip Roedig (Roedig et al., 2017).

6.2 Introduction

Virus crystallography at a synchrotron shares many common elements with general protein crystallography at an XFEL. As the crystals are immensely fragile, the act of cryo-cooling rarely leaves the virus crystals intact and so virus crystallography is performed at room temperature. This means that the radiation dose sustained by a crystal can convert to a greater degree of damage, and so typically the rotation data that can be collected

from such crystals is less than 0.5° before the radiation damage has obliterated the order within the crystal. Due to the large unit cell size, this rarely prevents indexing from being successful. However, this method draws many parallels in merging small fractions of data from multiple crystals, which is the hallmark of XFEL crystallography.

Bovine enterovirus is a virus in the Picornaviridae family (a virus family also containing foot-and-mouth disease virus, hepatitis A, polio virus and rhinoviruses). It infects cattle, and although it is not usually pathological, can cause symptoms including fertility problems, and diarrhoea. Bovine enterovirus is most interesting due to its similarity to other picornaviruses, and the presence of a lipid-binding pocket, which is shared with poliovirus. It is, however, not considered to be a threat to animal or human health, and therefore can be handled without the strict biological safety procedures which govern the handling of many other picornavirus family members such as polio and foot-and-mouth disease viruses. This makes it a good target to study at the LCLS, which does not accept biologically hazardous material at the XFEL beam line.

BEV2, like other picornaviruses, is an icosahedral virus with 60 copies of VP1, VP2, VP3 (which make up the capsid structure) and VP4, which is usually substantially disordered on the inner side of the capsid facing the RNA interior. Icosahedra have 2-fold, 3-fold and 5-fold axes, the latter of which cannot form part of the crystallographic symmetry (as five-fold axes cannot tesselate in three-dimensional space), and therefore enforces a minimum of five-fold non-crystallographic symmetry (NCS) on any icosahedral structure. Such NCS provides a signal-to-noise boost of $\sqrt{5}$ and can help counteract the effects of an incomplete data set. This five-fold NCS axis is also close to where a stabilising lipid cofactor binds to the structure in a number of picornaviruses (including polio virus, but not including foot-and-mouth disease virus). One potential strategy for control of such picornaviruses is to find a common lipid which binds the picornavirus more tightly than the native lipid to stabilise the capsid to the extent that it cannot break apart within cells. This stabilised particle might then be used as an immunogenic vaccine to confer a long-term immune response against the virus. Hence, bovine enterovirus crystals can be used for

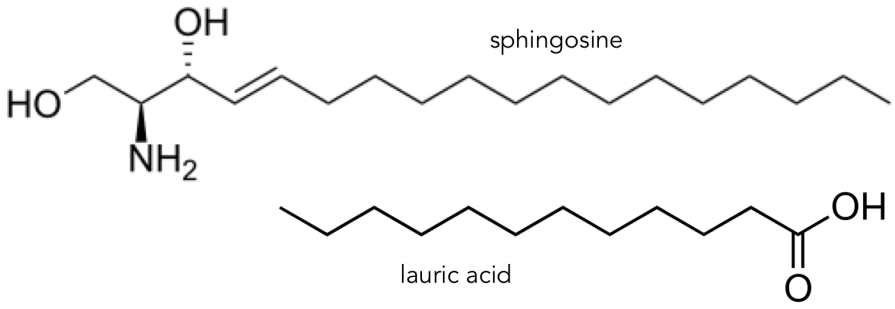


Figure 6.1: Structures of 18-carbon sphingosine and 12-carbon lauric acid, drawn at the same scale to denote relative lengths. Adapted from images in the Public Domain from Wikimedia Commons.

co-crystallisation with various lipids to see if there are alternative lipids that can displace the native version, sphingosine.

The structure solution of BEV2 was previously performed at a microfocus beamline at a synchrotron. The microfocus element of the beamline was necessary as the cubic crystals were only 50 - 60 μm in size. Due to the unit cell dimension of 436.6 Å, the diffractive power is spread out among a larger number of spots, and therefore the additional flux density available from a microfocus beam line is necessary to generate processable datasets. The methodology for the structure solution at a microfocus beamline (albeit not the structure) has been previously published (Axford et al., 2012).

Members of my lab grew bovine enterovirus crystals in the presence of lauric acid (12-carbon chain) to see if lauric acid could displace the natural sphingosine lipid (18-carbon chain) cofactor (Figure 6.1). This would be differentiated in the crystallographic structure as a shortening of the chain of the usual lipid cofactor. The crystal dimension was limited to 8 μm when grown in lauric acid, and therefore they were deemed incapable of being tackled at a microfocus beamline. As BEV2 is considered to be of no danger to human health or agriculture, we are allowed to shoot crystals of BEV2 at LCLS within the safety framework.

Due to the limited number of BEV2 crystals available, we collaborated with Alke Meents

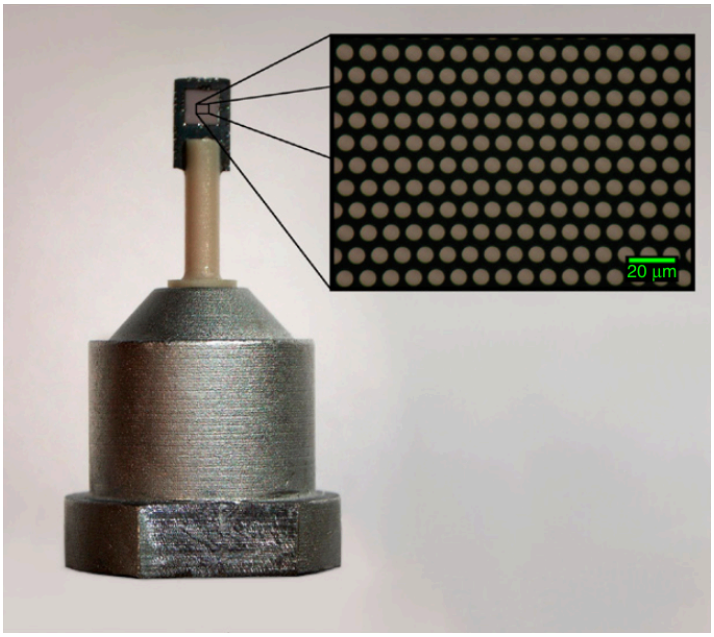


Figure 6.2: Design of the Roadrunner chip in its chip holder with micropatterning shown; figure created by Philip Roedig and reproduced here from the paper published in Nature Methods (Roedig et al., 2017).

at DESY in Hamburg, who had developed the first iteration of the Roadrunner chip with his PhD student Philip Roedig (Roedig et al., 2015). The lauric acid-soaked BEV2 crystals were an interesting sample and model system for picornaviruses for investigation using the XFEL. The Roadrunner chip was a perforated silicon wafer, which provides a low-background surface on which crystalline samples can be mounted and minimises sample wastage. The chip design is shown in Figure 6.2.

6.3 Methods

6.3.1 Crystal production

BEV2 virus particles were produced and purified in the same way as for BEV type 1 (Smyth et al., 1993) and crystallised in nanolitre drops (Walter et al., 2005). Fractions containing BEV2 particles were pooled, incubated with lauric acid, and pelleted by ultracentrifuga-

tion by James Kelly from the Pirbright Institute, UK. Crystallisation plates were set up as explained in the Methods section of (Roedig et al., 2017). The space group and unit cell dimensions of BEV2 crystals were already known due to the work by Axford et al.; space group F23 with a unit cell dimension of 436.6 Å.

6.3.2 Data collection

Bovine enterovirus crystals loaded on chips were taken to LCLS for irradiation at the XPP beamline. I personally was not present at the XFEL at the time of data collection, but was working remotely to process data in the UK to give some real-time results. Around 80% of the chips taken to LCLS shattered during transport and were unusable, severely limiting the sample available. Some chips were prepared on-site from loading fresh samples of bovine enterovirus onto new chips. Transported samples were at cryo-temperatures and fresh samples were at room temperature.

The wavelength was recorded on a per-image basis and averaged 1.3014 ± 0.0011Å. The detector installed at the beamline was a CSPAD (but not the same one used at the CXI beamline), mounted on a robotic arm. This robotic arm was capable of keeping track of relative distances for the most part, but there was no absolute offset and therefore this had to be calculated manually. The detector distance was varied frequently during data collection, with an initial detector distance recorded at 170 mm, and subsequent collections recorded at 150 mm, 110 mm and 105 mm. During data collection, runs were analysed by generating composite images showing the highest pixel value from the complete set of images, using the software suite *cctbx.xfel*, providing real-time feedback on the data collection process.

6.3.3 Hit-finding

Nicholas Sauter processed the initial set of images to search for hits using *cctbx.xfel*, collecting images which registered 20 or more strong spots, for further analysis. A con-

ventional powder pattern was generated for these data by taking the maximum intensity per pixel from the set of all images.

6.3.4 Initial detector geometry correction

Using results from an earlier run showing thermolysin diffraction, Nicholas Sauter calculated initial corrections to the 64 panels on the CSPAD detector, using the positioning of powder rings (from the conventional powder pattern) to guide the position of each panel. This was applied when images were dumped from the XTC streams.

6.3.5 Finding spots on individual images.

The program *dials.find_spots* was used to find spots on individual images with some chosen spot-finding parameters:

```
min_spot_size=2
global_threshold=2000
gain=14
sigma_background=30
sigma_strong=6
mask=mask.pickle
d_max=10 d_min=1.7
```

These parameters are similar to those used in XDS. They indicate that with a gain set to 14, only pixels which exceed 2000 corrected counts will be used for further analysis; the background-sigma and strong-sigma values will determine which pixels are set to background and which ones to foreground; and if these exceed the minimum spot size of 2 pixels then they will be counted as a spot (and the spot position determined as the centroid of all foreground-flagged pixels). No pixels will be considered at resolution worse than 10 Å or better than 1.7 Å.

There were several challenges to spot-finding due to the very weak diffraction spots from these crystals. The background for this experiment registered extremely highly as the detector was moved extremely close to the sample and diffraction was mainly kept to the centre of the image. This caused low resolution background pixels to frequently register as real spots, so low resolution reflections had to be excluded from the analysis by setting a low resolution cutoff of 10 Å (parameter d_max). The high resolution cutoff (parameter d_min) was chosen as 1.7 Å as this excluded erroneous spots picked up in the corners of the detector. A mask had to be manually generated to encompass near-zero pixel intensities for unbonded pixels, which are arranged on every image in a distinctive pattern. These pixels are left unbonded so that they are not responsive to photons, and therefore their value represents the pure electronic offset allowing for common mode correction. However, without a mask, these pixels wrought havoc with background estimation and led to spots being predicted around these areas (Figure 6.3). This was achieved using the *dials.generate_mask* program in the DIALS suite. The parameters global_threshold, sigma_background and sigma_strong were set to higher-than-default values, and although this missed many real spots (as judged by eye), it prevented spots being predicted on the border of CSPAD detector panels. It is notable that sigma_background had to be set at a higher value than sigma_strong, which was empirically derived but may be due to a combination of relatively uncalibrated gains and a potentially non-linear response in gain. At the time of the initial data analysis, there was no method for excluding the empty space around the panels on the image. After indexing the diffraction patterns, the spot-finding results were no longer necessary.

6.3.6 Correction of detector distance.

A pseudo-powder pattern (see Chapter 3) was generated for the bovine enterovirus data, which immediately showed a poor fit to the simulated peaks for the given unit cell dimension and space group. However, correction of the detector distance from 105 mm to 85 mm showed an improved agreement between the observed and the simulated peaks.

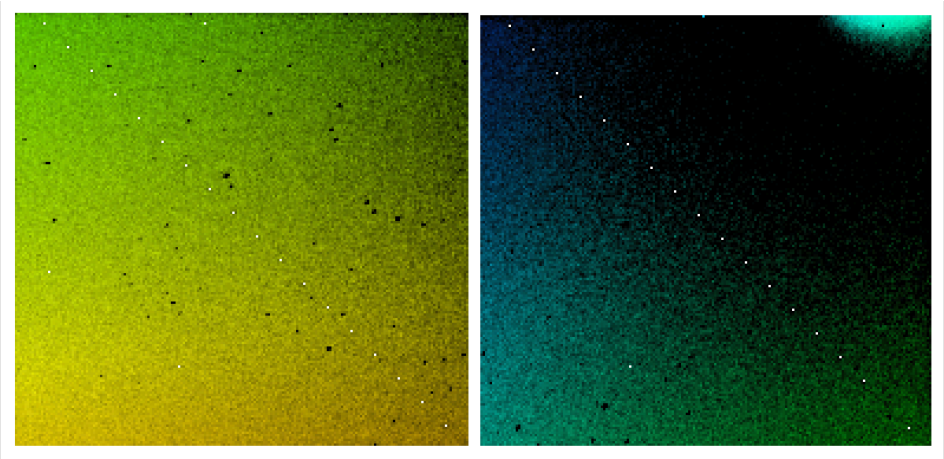


Figure 6.3: Two panels juxtaposing the beam centre (top right corner shows the beam centre shadow) with diagonal pixels which have been unbonded, in steps of $10x + 10y$ pixels. These were manually picked out and added to a list of undesired pixels to be masked during processing.

This discrepancy in the detector distance became a repetitive problem for many data sets, as well as this BEV dataset; I noticed that this degree of error (out by at least 5 mm) was a particular problem for data sets collected during this beam time, and has been incorrect for later beam times as well. This is due to the lack of calibration of the robotic arm on which the detector was mounted. This was a major correction in the diffraction geometry based on the pseudo-powder pattern, and extra geometry correction (detailed below) was also required.

6.3.7 Indexing images

The bovine enterovirus crystals created a difficult challenge for indexing and inspired the creation of the TakeTwo algorithm in order to process these data, explained in Chapter 3. In the associated paper (Ginn et al., 2016c) the TakeTwo algorithm was run multiple times in different combinations of parameters; however, after some small improvements in the algorithm, a good indexing rate could be achieved without needing to modify the default parameters. A quick round of initial orientation matrix refinement (see below) with

only a few cycles was performed in order to verify if the solution would lead to a good wavelength histogram. In order to determine if an indexing solution was successful prior to calculation of good geometry, a 'search size' window of two pixels was allowed to find the nearest highest pixel value as a temporary measure. In other words, the predicted spot location was moved up to two pixels away in any direction in order to settle on the pixel with the highest value.

6.3.8 Extra detector geometry correction

Additional detector geometry correction was deemed necessary, and was carried out according to the protocol outlined in Chapter 2. An earlier attempt using a more rudimentary approach (Ginn et al., 2015b) was insufficient to be able to predict the locations of the spots with enough accuracy to reduce the search size to 0. Due to the way in which the geometry information was stored in the files from *cctbx.xfel*, the geometry calculation was significantly trickier, and but there was no need to restart the process from the output from the alternative software suite Cheetah as the data analysis described here was sufficient to answer the scientific question.

6.3.9 Integration of images

After adequate geometry correction, the 'search size' (explained in Section 6.3.7 above) was reduced to zero. At this point, a very rudimentary interpolation was carried out so that the integration window could fall between two pixels. Spots on the detector could be separated by as little as three pixels, and inspection of individual images showed that spots appeared to extend no greater than 2 pixels in width. For this reason, a two-by-two pixel foreground integration box, followed by a one-pixel padding not used in integration and a surrounding two pixel padded square for background estimation was used, as shown in Figure 6.4. When background pixels from neighbouring reflections overlapped (as per the examples in the Figure), they were accepted; however, if background pixels

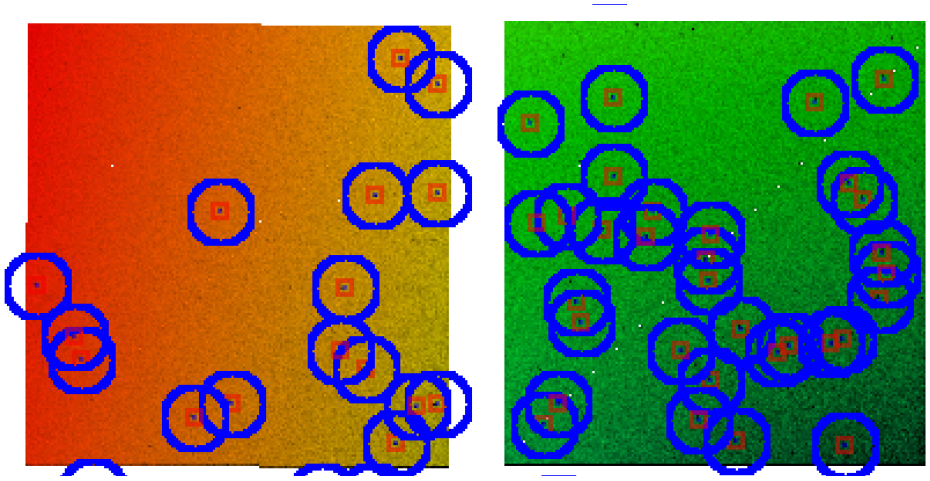


Figure 6.4: A magnification of two panels showing the integration windows which were used during integration of the reflections with refined geometry. Only strong reflections are shown; the opaque blue circle highlights a spot position, and the light blue shade overlaps the chosen integration window for a reflection. The 2-pixel wide red shading indicates the background region chosen for integration. Most of the reflections are covered by their foreground windows; of those which are mis-predicted, the buffer region between the foreground and background windows protects against including the signal in the background integration window.

overlapped with the foreground pixel of another reflection then this reflection was rejected. Care was also taken to minimise the bandwidth of the beam used to generate a small overpredicted zone, to avoid there being so many reflections falling on the single image that all become rejected. Very few images showed evidence of more than one crystal, and multiple lattices were not considered.

6.3.10 Initial orientation matrix refinement

Initial orientation matrix refinement was carried out with an over-predicted bandwidth of 0.8% and over-predicted rlp size of 3×10^{-4}Å^{-1} to generate the sharpest wavelength histogram peaks for all indexed diffraction patterns. The intensity threshold for considering a spot a "strong spot" was set to 18 <I/sigI> cutoff, calculated as the sum of the foreground pixel counts over the square root of the sum of all the pixel counts, assuming

a linear detector gain of 1. The rotation of the crystal around the beam axis was refined to maximise the number of strong spots predicted as per Chapter 4.

6.3.11 Post-refinement and final merge

Post-refinement of the diffraction models of each crystal was carried out with and without a reference data set. The space group F23 has a two-fold indexing ambiguity, which had to be resolved in the case of no reference data set. This was achieved using the Brehm & Diederichs algorithm (Brehm and Diederichs, 2014). This was followed by switching of incorrectly determined indexing solutions during post-refinement if the correlation against the most recent reference data set is higher with the opposite indexing solution. For the with-reference post-refinement, a synchrotron reference data set provided by Jingshan Ren was used to solve the indexing ambiguity of the diffraction models with an initial set of reflections. The partiality cutoff for this data set was set to 0.2. Images which had a correlation coefficient of at least 0.80 with the reference data set were merged, rejecting outlier observations of more than 1.6 standard deviations from the mean merged intensity for each reflection in cases where there were more than five observations per symmetrically equivalent Miller index.

6.3.12 Phase extension from initial phase information.

In order to show that the amplitudes generated from the XFEL data processing had merit, phase extension was performed from (a) phase information from the synchrotron structure no higher than 5 Å resolution, and (b) from a remotely related structure, the FMDV A22 capsid proteins, which has no more than 20% sequence identity. For the latter, phases were calculated to 2.3 Å and no phase extension was performed. The scripts were obtained from much older projects and needed repair. I initially repaired them, and the phase extension experiments themselves were tailored by Dave Stuart and Ren Jingshan, and then performed individually by Ren.

Phase extension comprises a few stages: the starting set of phases to 5.0 Å from the synchrotron structure was provided with all the amplitude information to 2.3 Å (resolution limit chosen for refinement). A Fourier transform was performed on this reflection list to convert this to real space. The General Averaging Program (GAP) performed cyclic averaging according to the five-fold NCS of this virus and performs density modification. This used a grid spacing of 624 voxels per unit cell dimension, which would lead to oversampling of the map at the highest resolution of extension (2.5 Å) by a factor of 1.75.

On the first and every subsequent 5th cycle, a new Wang envelope is calculated based on the working electron density. The Wang envelope is used in phase extension to determine the parts of the map which should be subject to solvent flattening, and which parts are to be subject to NCS averaging. This is primarily calculated on knowledge of the virus placement within the unit cell and the capsid dimensions (an outer radius of 160 Å and an inner radius of 105 Å. The working electron density is used to set voxels below a calibrated electron density value to be flattened, removing internal solvent volumes which are not covered by a basic spherical capsid definition.

The solvent content was increased to the highest value which the phase extension protocol was capable of tolerating (derived empirically as 94%). This is calculated by density convolution and histogramming to remove the bottom 94% of voxels from the protein envelope calculation. A suitable version of the source code for the GAP has been published (Ginn and Stuart, 2016b). After creation of a new map, a Fourier transform returned the information to reciprocal space. The ccp4 programs sortmtz and CAD were used to tidy the reflection file, and the program *shell_scale*, written by Dave Stuart, was used to scale the average amplitudes of the output reflection list to the initial list. The program *shell_scale* was also used to cut the phase information to the same resolution as the previous cycle, plus approximately one tenth of a reciprocal lattice point per cycle. This new reflection list was then used for the conversion to real space.

6.3.13 Refinement of atomic model against merged intensity values.

The atomic model was refined against the merged intensity values by Ren Jingshan, a member of Dave Stuart's lab, using the software package CNS (Brunger, 2007), which is well-suited for crystals with strict NCS such as virus capsid crystals. The atomic co-ordinates were re-refined into the new electron density obtained by phase extension, to reduce model bias from the previous molecular replacement solution. This was performed to satisfy the referees, but it is noteworthy that with 230,000 reflections for 6000 atoms, and five-fold NCS, this equates to 60 observations per atom, which would make overfitting to the wrong solution very difficult.

6.3.14 Omit map of the cofactor in the lipid-binding pocket.

The atoms for the sphingosine cofactor were omitted from the input model by Ren Jingshan, and atomic model refinement with simulated annealing performed in CNS 1.2 (Korostelev et al., 2002). From this, an Fo-Fc difference map was produced.

6.4 Results

6.4.1 Diffraction was observed during the course of data collection.

In total, there were 24 runs collected at a variety of detector distances. The majority were trial runs at low transmittance in varying chip orientations in order to ensure that silicon reflections from the chip were not satisfying the Bragg condition, in order to prevent damage to the detector. For runs with a substantial number of rows of crystals exposed to the beam, there was one run collected at a nominal distance of 170 mm, one at 150 mm, one at 110 mm and seven at 105 mm. There were too few images (less than 10) collected at 150 mm or 110 mm to accurately determine the beam centre and/or detector distance and these were discarded. There were 22 images collected at 170 mm (for which the detector distance was revised to 151.63 mm) and 366 images were dumped from the runs

at 105 mm (for which the detector distance was revised to 85.8 mm). An example of an image collected at the latter distance is shown in Figure 6.5.

Using the wavelength of 1.301 Å and the detector distance of 105 mm, this showed visible diffraction to ~3.1 Å. Due to the miscalculation of the detector distance during beam time, the maximum resolution achieved on the image was incorrectly determined and in the efforts to collect data out to 2.2 Å, data were accidentally collected out to 1.6 Å. The incorrect recording of the detector distance had an impact on the experimentalists' perception of collected resolution. This focused both the spots and the background into the centre of the image, which reduced the data quality (by substantially increasing the background scatter) and made this a very difficult data set to process. Runs which were processed at a unique distance but did not contain enough images were essentially lost, as the cost in effort of calibrating the geometry for these data became too high.

6.4.2 Cryo-cooling samples for XFEL beam time is unnecessary or detrimental.

Cryo-cooled crystals of BEV, maintained at cryo-temperatures before and during data collection, did not diffract well at all, whereas samples kept and exposed at room temperature diffracted well. Also ice crystals produced extremely strong diffraction and had to be avoided by positioning the detector at a point where it could not collect beyond 4 Å diffraction, which severely limits the usefulness of a cryo-cooled data set. The decision to cryo-cool the samples was taken as it is easier than trying to maintain crystal hydration at room temperature; however, the XFEL has enough flux to destroy the sample regardless of its temperature, negating any reasons relating to radiation damage. It should be noted that one of the valuable properties of the XFEL is its ability to solve a structure at room temperature with minimised radiation damage.

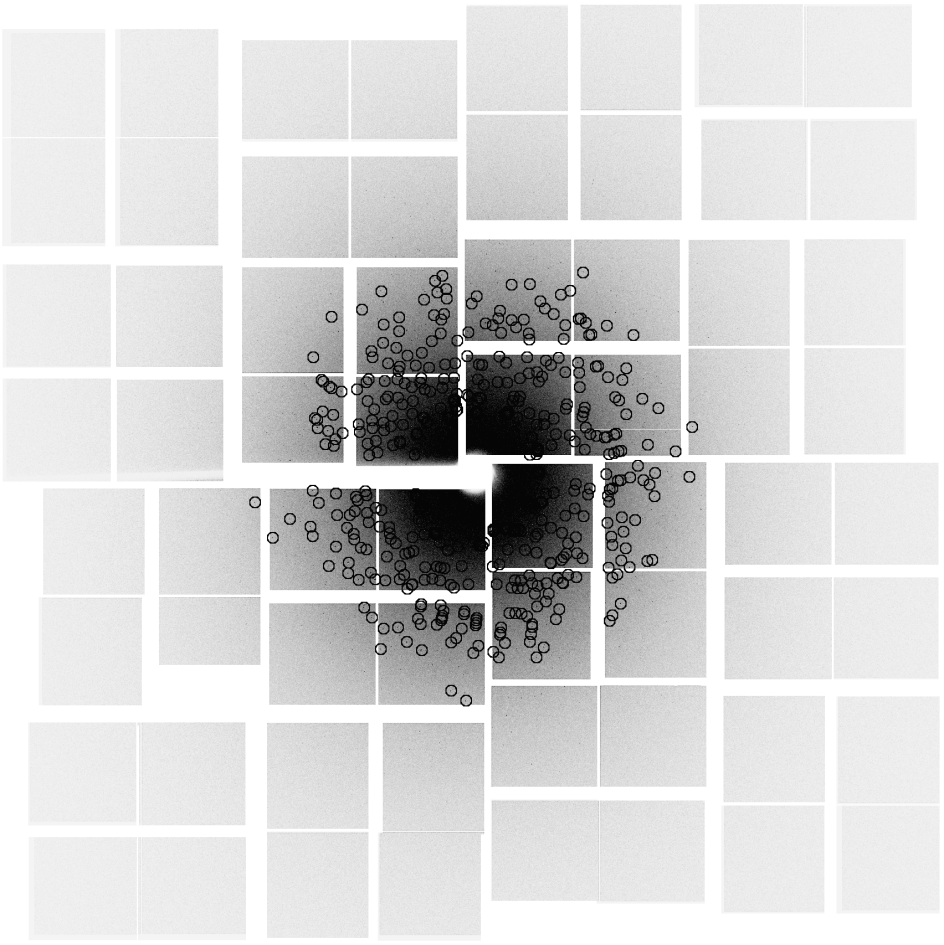


Figure 6.5: An example of a bovine enterovirus diffraction pattern, with spot-finding results shown from the blob-finding algorithm defined in Chapter 2, at a revised distance of 85 mm. Note the high background concentration in the centre of the image most strongly affecting data collected at low resolution, caused by unwittingly bringing the detector too close to the sample.

6.4.3 Initial hit-finding results from cctbx.xfel were conservative.

Of the 366 images collected at a 85.8 mm detector distance, only 304 showed evidence of diffraction from hit-finding results by cctbx.xfel. A number of images appeared to have been dumped for which no or few spots were visible. As this was not enough data to generate a dataset of sufficient quality, the hit-finding process was repeated with cctbx.xfel but without filtering, so all images were written out to disk. Searches for additional lattices based on spots predicted by the DIALS software, and subsequently using the 'blob-finding' algorithm described in Chapter 2, increased the numbers of hits. These hits had 351 ± 183 spots with a maximum spot count of 1368 spots on a single image. Evidence of diffraction was limited to the very centre of the image for the majority of crystals but very few images appeared to have multiple lattices. Otherwise, the DIALS spot-finding routine was robust within a certain resolution range. Due to the small number of images collected at the larger distances, these were abandoned due to insufficient observations to fully determine the detector distance. Many of the additional hits from DIALS and the 'blob-finding' algorithm indexed, but had considerably weaker diffraction and had a proportionally higher error contribution from the background. The total run statistics for the runs included in the structure solution are shown in Table 6.1.

Table 6.1: A table detailing the acquisition rate from BEV chips, including the number of images obtained, hits, hit rates and number of indexed patterns. These were the images used to drive post-refinement.

Run number	Images obtained	Number of hits	Hit rate (%)	Number indexed
294	2280	159	7.0	124
296	180	17	9.4	16
298	1950	76	3.9	69
301	2015	150	7.4	113
303	2387	44	1.8	30

6.4.4 Geometry refinement revealed a tilt to the detector.

The refinement of the geometry for BEV was tricky, as the data do not extend to the edge and therefore the extent of geometry refinement had to be limited to the panels on which enough reflections could be predicted. The geometry refinement was performed as described in Chapter 2 on the predicted reflection positions from indexing solutions. This revealed a tilt corresponding to a 0.5 mm error in the Z axis with panels refined independently. Prior to geometry refinement and reveal of the tilt in the detector, it was not possible to support phase extension (see Section 6.4.10). After geometry refinement, the quality of the intensity estimates was sufficient to support phase extension.

6.4.5 Indexing success rates were very high.

Of the 22 images from the detector distance of 151.63 mm, 15 were indexed successfully. These were also at a lower transmittance (10% to 20%) and so contained very weak diffraction. These were abandoned due to insufficient information to easily determine the detector distance, and therefore a poor effort to reward ratio. The runs at a detector distance of 85.8 mm were recorded at a transmittance of 40% and had a much higher indexing rate: 97.6% of the images from the initial hit-finding results were indexed successfully, using the TakeTwo algorithm. After extending to a wider range of hit-finding results, 79% were indexed across the entire data set. This reduction in indexing rate was due to a decrease in the quality of the images which only passed the more permissive hit-finding parameters.

6.4.6 Initial orientation matrix refinement reduced the error significantly with small tilt corrections.

Initial orientation matrix refinement (as per Chapter 4) resulted in vital, small rotations in the crystal orientations, with the horizontal axis and vertical axis being rotated by angles of (on average) 0.022° and 0.024° respectively. This produced very sharp wavelength his-

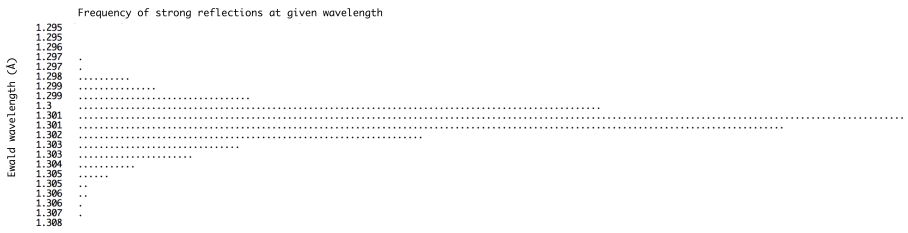


Figure 6.6: BEV2 crystal Ewald sphere wavelength histogram. Every strong reflection is represented by a full stop. The identity of the Miller index and the orientation matrix of the crystal can be used to determine the Ewald sphere radius on which the reflection lies, which is expressed as wavelength (the reciprocal value). The peak wavelength which was recorded when performing the experiment at LCLS was 1.30123 Å.

tograms with a large number of strong reflections, as shown in Figure 6.6. Due to the high number of strong reflections per image, this suggests that these refined initial orientation matrices can be trusted to a high degree of accuracy.

6.4.7 Integration strategy was optimised based on merging statistics.

Reducing the metrology search size once the geometry had been corrected increased the R_{split} significantly (from 28% to 48%) and reduced the $CC_{1/2}$ mildly (from 82% to 76%). Although this sounds like a poor change in statistics, this is a combination of (a) reducing the over-estimation of weak intensities, noting that a higher reflection intensity will reduce the R_{split}, and (b) a tendency for nonsense to correlate with nonsense if the nonsense is non-random, and not 'searching' for the highest peak reduces the non-random nonsense. So in other words, after reduction of systematic error, the statistics which monitor self-agreement get worse.

6.4.8 Post-refinement and the partiality model extracted additional information.

Post-refinement on individual images was extremely stable as approximately 3000 reflections were assigned a significant partiality value per image, which is an excellent

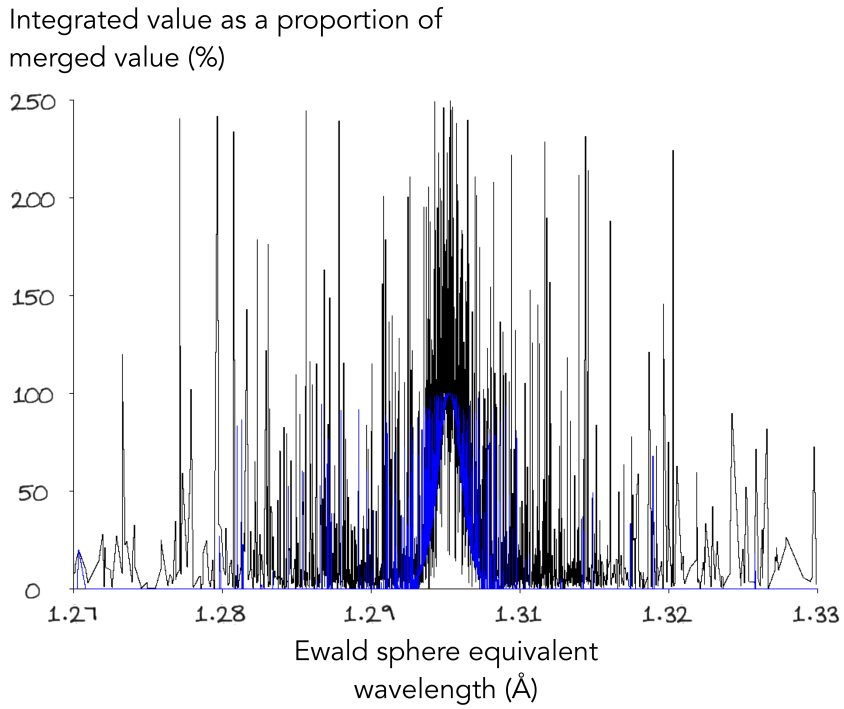


Figure 6.7: Partiality plot for one crystal of BEV2 after post-refinement against the interim merged data set. Every point is a reflection, plotted in black against the inverse of the Ewald sphere radius on the x axis, and displaying the integrated value as a proportion of the merged value from the reference data set which incorporates all images combined. Reflections are drawn from the origin of reciprocal space out to 3.1 Å. The calculated partiality is calculated from the model parameters and plotted separately in blue.

observation-to-parameter ratio. However, it was clear that the integrated intensities were quite variable; this is likely in part due to the highly variable and comparatively strong background signal in the middle of the detector, which can make even low resolution reflections hard to distinguish even by eye. The combination of these factors made for a post-refinement strategy which had low systematic error due to over-fitting and high random error. The resulting partiality plots looked clean given the high noise-to-signal ratio (Figure 6.7). The post-refinement introduced additional offsets in the crystal orientations corresponding to 0.003°, and reciprocal lattice point sizes ranging between 3.0×10^{-5}Å^{-1} and 6.0×10^{-5}Å^{-1}.

6.4.9 Data collection and structural refinement statistics

The data collection and refinement statistics are shown in Table 6.2 and the completeness of the data set is shown on a per-resolution-shell basis in Figure 6.8.

6.4.10 Extension from initial phases of a synchrotron structure were possible from 5 Å.

Phase extension was performed with phases to 5.0 Å from the unpublished synchrotron structure, resulting in an electron density to 2.5 Å resolution into which atomic coordinates could be built in by eye. Phase extension could not be supported from 8 Å. This is likely due to the increased error from a very high background and the volatility of the partiality calculation with respect to model parameters at low resolution, which produce lower quality amplitudes. The phases generated from 5.0 to 2.5 Å can be seen to correlate well with the reference set, though there is a notable subpopulation for which the phase has adopted (180° - ϕ) or (360° - ϕ) as shown in Figure 6.9. A number of reflections have phases which have clustered around the (0, 180) and (180, 0) plot positions and these correspond to the centric reflections, which can only take values of 0° or 180°. Therefore, mis-assignment produces very clear errors in the phase correlation diagram.

6.4.11 Atomic model refinement and omit map showed binding of sphingosine to the viral capsid.

An omit map of the binding pocket factor revealed a density which matches the increased length of sphingosine compared to the shorter hydrocarbon chain of lauric acid (Figure 6.10). This provides some evidence that despite soaking the viruses in lauric acid prior to crystallisation, it is incapable of displacing sphingosine within the crystalline lattice. The lipid cofactor binds inside VP1 which is closest to the five-fold NCS axis in the icosahedral virus structure. It is unlikely that a change in the cofactor could provide a conformational change in the capsid which would spread so far as to substantially alter the crystal contact

Table 6.2: BEV data collection and structural refinement statistics. * Values for high resolution shell in brackets. ** Number of reflections used in R free set.

Data Collection	
Temperature (K)	293
Number of measured chips	5
Number of collected images	8812
Total measuring time (s)	818.6
Average acquisition rate (images/s)	10.79
Number of hits	446
Indexed patterns	352
Number of lattices used in final merge	324
Space group	F23
Cell dimensions	
a = b = c (Å)	436.6
$\alpha = \beta = \gamma$ (°)	90
Resolution (Å)	29 – 2.3 (2.37 – 2.3)*
I/σ(I)	5.4 (2.1)
Completeness (%)	78.0 (83.4)
Redundancy	2.3 (2.4)
Rsplit (%)	48.6
CC 1/2 (%)	74.6
Refinement	
Resolution (Å)	29 - 2.3
No. reflections	223989 (11604)**
Rwork/Rfree	23.3/25.7
No. atoms	6587
Protein	6212
Ligand/ion	37
Water	338
B factors	
Protein	32
Ligand/ion	39
Water	38
R.m.s. deviations	
Bond lengths (Å)	0.011
Bond angles (°)	1.6

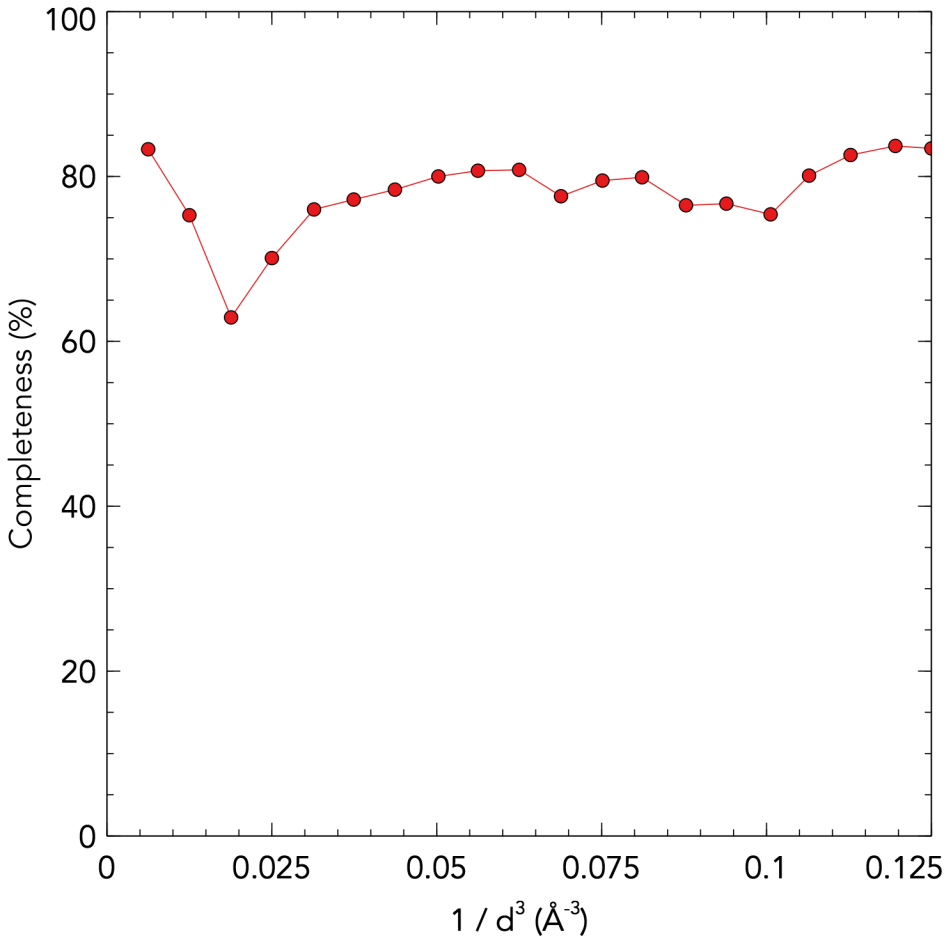


Figure 6.8: Completeness per resolution bin (slices chosen to enclose equal volumes of reciprocal space). The overall completeness is 78%, but is at a minimum between 5.0 and 4.3 Å of 62%, and reaches a maximum of 83% at the chosen edge of accepted diffraction (2.3 Å).

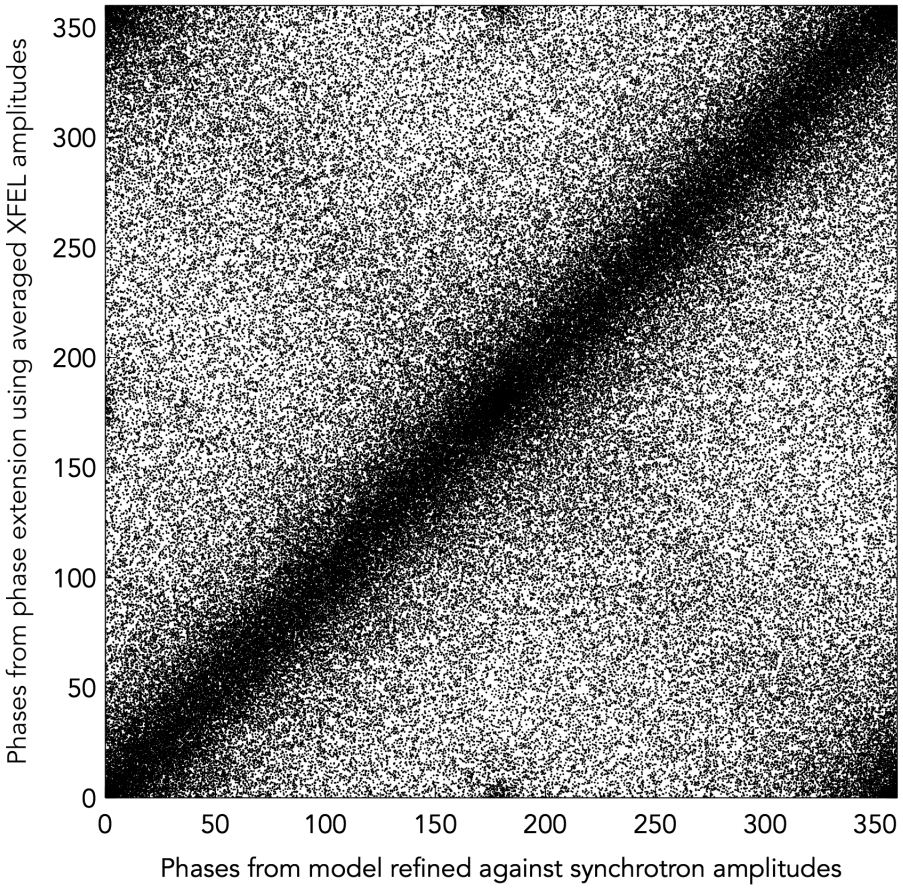


Figure 6.9: BEV extrapolated phases from 5.0 Å to 2.5 Å with agreement against the solved synchrotron structure.

points to prevent them from forming. Therefore we may arguably preclude the possibility that the lauric acid could bind to the capsid but prevent crystal formation. However, because the crystal size was limited in the presence of lauric acid, it may be interfering with the growth of the crystal by some other process.

6.5 Conclusions

As sphingosine is seen to still bind to the BEV despite incubation with lauric acid, it is reasonable to presume that lauric acid is unable to displace sphingosine, and therefore is not worth pursuing as a possible target for stabilising BEV capsids. This is the first virus structure to be reported at the XFEL and is particularly exciting as there were an extremely small number of lattices with which the structure could be reconstructed, and yet it was still achieved. The completeness of this dataset at 78% is mitigated by the five-fold non-crystallographic symmetry, and therefore can tolerate a lower coverage of all reflections in reciprocal space. This is also unprecedently low for successful structure determination by XFEL.

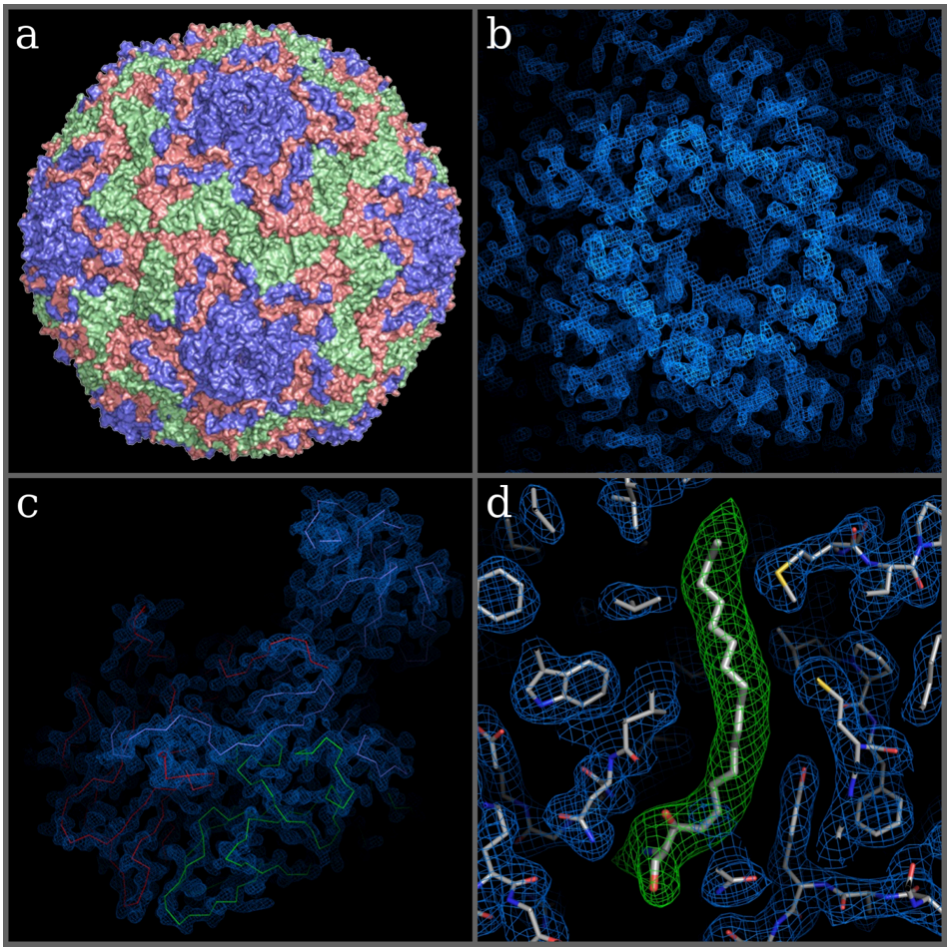


Figure 6.10: Figure prepared by Jingshan Ren. (a) Overall structure of the BEV2 capsid. (b & c) Electron density following the first cycle of cyclic averaging from the initial 5 Å phases. (b) is an image of the electron density facing the five-fold axis, and (c) shows C-alpha tracing of the VP1, VP2 and VP3 protein subunits. (d) Omit map after simulated annealing, to remove phase memory, of the sphingosine lipid cofactor from BEV2. Superimposed on top are the atomic coordinates of sphingosine to show goodness-of-fit.

Chapter 7

Discussion

7.1 Summary of the progress and the effects on the field of crystallography.

The original purpose of releasing the code for *cppxfel* was to provide a reference point for proof-of-principle algorithms. I have been extremely fortunate in that the field has been particularly keen on improving on the initial algorithms used before I entered the field, and therefore I have had many opportunities to aid the incorporation of newly developed algorithms into other software suites; e.g., the "droplet" spot-finding algorithm outlined in Chapter 2 and the TakeTwo indexing algorithm into *CrystFEL* for release in version 0.6.3. Nevertheless, the user-friendliness of the software has increased significantly during the course of my DPhil thanks to the intensive use by Helen Duyvesteyn. Other than myself, Helen Duyvesteyn has been the most heavy user of the software, and she was able to point out a lot of problems with the software design and how the user interface could be improved. However, in retrospect it is much easier to design a clean user interface from the beginning of development, and ideally not in conjunction with learning C++ simultaneously (before starting *cppxfel* my previous software had only been coded in C). Good scientific principles would dictate that any code used to implement an algorithm

should be released open source. As a result, the entirety of the code base for *cppxfel* has been available on GitHub (https://github.com/helenginn/cppxfel) since fairly early on in its development.

Although the code for the slip-and-slide algorithm is released with the code of *cppxfel* and is usable, it is not as accessible as some of the algorithms have become. One advantage of the slip-and-slide algorithm is that it is relatively robust, and only requires two parameters to be modified on a regular basis, with some potential for auto-calculation of these values. This level of robustness is particularly hard to achieve for geometry calibration software. Therefore, it may be considered to be a viable target for implementation in other software suites. However, as the most recent algorithm was only accepted for publication one month prior to the submission of this DPhil thesis, it has not yet had time to reach others in the field of crystallography. I will be delivering at least one talk on the geometry algorithm after the finalisation of my DPhil thesis, and if there is sufficient demand for the algorithm I will be happy to aid someone in implementing it into other software packages which will be maintained by others. The geometry algorithm should not be considered to be XFEL-specific. It can (and perhaps should) be implemented on synchrotron beamlines as well, in particular to detect features like tilts and misaligned detector distances.

Similarly, the indexing algorithm outlined in Chapter 3 should also be employed for serial synchrotron crystallography experiments, which also suffer from a similar lack of information per independent snapshot. Some informal indexing on synchrotron beamlines has proved to be very successful. Having worked with Thomas White, the algorithm has now been released in CrystFEL 0.6.3, which will make it significantly more available for use in the crystallographic community. In the grand scheme of data analysis, this indexing algorithm will likely provide benefits to systems both where sample is abundantly available and also those for which sample is not. In cases where there are insufficient images, additional indexed patterns increase the information content significantly and may be the tipping point between answering the scientific question at hand and being unable to do

so. In cases where it is possible to collect sufficient images to solve a scientific question even with a low indexing rate, the benefit of an increased indexing rate will be to reduce labour hours and production costs of the sample.

The algorithms for initial error reduction in the model parameters (initial orientation matrix refinement) and also post-refinement may have more far-reaching outcomes for other techniques such as electron diffraction, and can theoretically be applied to serial synchrotron crystallography if the definition of the energy spectrum is revised for monochromatic synchrotron beams. It may also be useful if such synchrotron beamlines intend to boost the flux by removing the crystal monochromator and replacing with a multilayer crystal, which would produce a beam more pink even than a SASE beam. The partiality model - a largely separate concept, but a significant mediator of success by post-refinement error reduction - is being incorporated into CrystFEL by Thomas White, and I shall continue to work with him as he seeks to incorporate the model.

The work described here performs best with the assumption that the wavelength has been well-measured. These algorithms work well together because they separate the effects of the experiment and attempt to refine each parameter against the salient information. For example, refinement of the unit cell dimensions described for Chapter 2 is performed on the basis of the strong reflections which have been flagged in reciprocal space, at which point the unit cell refinement has been made independent of the geometry information. However, if one were to fit the unit cell dimension to the spot positions on the detector, although it is possible that one may even find a better fit to the data, it is unclear whether that should have been an error corrected for by the detector geometry. The dangers of correlated parameters is one to take note of, and I have found is rarely worth throwing blindly to a minimisation algorithm. In some cases, the parameters do have the same effect (beam wavelength and unit cell dimension for a cubic space group are perfectly correlated). In other cases, the parameters have divergent effects (e.g., the wavelength and the detector distance in X-ray diffraction). During the course of the DPhil I have seen that the most successful attempts at refining these parameters

involves re-examination of the parameter space to make sure that the salient effects are used to refine the sensitive parameters. This is particularly demonstrated by the work on the slip-and-slide algorithm.

7.2 Possible directions of focus in XFEL studies by the community.

There is considerable importance in continuing to ease the bottleneck on data analysis as part of the XFEL structure solution pipeline. It is unclear how many data sets have been abandoned due to the difficulty of processing, mostly caused by a significant uncertainty in the starting parameters. There are two ways of doing this: there is a significant gain to be made in improving the measurement of the diffraction geometry in the first place. However, there is significant difficulty to organise this, as the ones who will be setting up the ability to make measurements will not have a direct feedback on their success unless they are also involved in the data processing. The other method is to develop algorithms to correct these parameters after the information has been lost; certainly, this thesis features many such algorithms developed to circumvent avoidable problems. I would, however, prefer the experiment to be set up correctly to avoid these computational gymnastics. This adds to the time allocated to data processing by individual users (which is often not considered by many in the process of submitting XFEL proposals), and there are so few developers involved in generating these algorithms that their attention becomes spread quite thinly. It does help users if algorithms are made to be more robust against errors or initial parameter specifications; however, this takes a large amount of effort which should therefore be directed at improving the most pertinent methods.

To talk about the focus of future computational methods development in specific cases, my next target for algorithm development would be to correct for the variable gain at low photon counts. One of our publications (Roedig et al., 2017) and others have been working towards low-background experiments where the non-coherent, scattered

photons hitting the detector are minimised. However, I believe although the technical improvements to achieve these experiments are profound, the benefit of these low backgrounds is effectively lost due to the properties of the CSPAD (the main detector used at LCLS). Finding ways to correct the output of a single detector is likely to be less helpful, especially with the advent of a range of detectors (AGIPD and Jungfrau at the European XFEL, and also consideration of the MPCCD at SACLA), and I decided not to attempt to perform these corrections. The background pixels show an extremely variable response at low numbers of photons (anecdotally, I have seen some adjacent pixels respond completely differently to a single photon with counts of 4 or 40.). There have been some attempts to apply the gain response to the entire photon range, but this increases the errors further because the pixels appear to average to a 'common' gain setting at high intensities. However, I believe there should be some effort made to improve the gain definition at low photon counts in particular, as this causes significant problems for background subtraction. This low-photon gain mode is also not the same as the gain mode when reaching higher photon counts. The detector ADU appear to increase more gradually when subjected to photon numbers more commonly expected for a Bragg peak. This means there is likely to be an over- or under-estimation in the background subtraction (plus a decent error) from an integrated intensity, although the exact effect is hard to predict.

Although self-seeded data produces a much more focused energy distribution, a further improvement for SASE pulses may be made from the incorporation of the energy spectrum into the partiality model as a fixed constant. Speculatively, if this parameter can be measured and therefore removed from the analysis, this gives the process of reference-driven model refinement to be relieved of the burden of some (frequently correlated) parameters and perhaps increase the range of crystal-specific parameters which could be refined in their place. Although, anecdotally, it has been believed that self-seeded data provide no advantage over SASE, this claim originates from processing of data before the advent of incorporation of the contents of Chapter 4 into the mainstream processing

pipelines. It may be worth repeating this calculation periodically as bottlenecks in the pipeline are relieved; prior to orientation matrix refinement, errors in the derived crystal orientations dominated the errors in XFEL analysis.

I also see a significant benefit to making the process more 'cyclical' than has been previously attempted. It would require abandoning the idea of a defined 'pipeline' which forms the basis of a significant number of software packages. This feels like a clean way to approach the problem, but I believe it is not scientifically merited. Improved descriptions of every model in the stages of the pipeline can be used to inform the ones which came before it. For example, an improvement in the description of the crystal orientation obtained during the process of post-refinement can be used to provide better prediction of the spot positions. It is also likely to be an essential component for any correction of the effect of the pink beam on the spot positions. The fact that the beam energy has a non-negligible bandwidth can have an effect on the predicted spot locations, which is particularly problematic at high resolution, but is an example of something which I believe cannot be corrected for without input from the latter stages of the current pipeline (merging and post-refinement). However, the word "post-refinement" at this point will (or already has) become an unhelpful description, as it is used to traditionally describe a very specific point of cyclisation of the pipeline. I would recommend that this term be changed to something more descriptive, such as "reference-driven model refinement", albeit this phrase is too bulky.

One of the types of ambitious projects undertaken at an XFEL is time-resolved crystallography experiments on a femtosecond time scale. These involve the calculation of merged data sets for a series of timepoints based on individual snapshots. At the moment, due to the uncertainty of phase information, it is easiest to see the differences by keeping the phases constant and calculating Fo-Fo maps - but the signal-to-noise is still limited by the quality of the amplitudes in both dark and excited data sets. The concept of binning the data into timepoints, merging and refining in real space as separate data sets will lead to a vast number of correlated parameters being refined independently, and

therefore overfitting and loss of real structural information. Given that femtosecond dynamics is one of the strongest pulls available for people to want to perform XFEL-based studies, this very limited method of using the diffraction data to draw biological conclusions should be on the list for an overhaul.

7.3 Personal direction and focus in structural biology.

I have now spent a small number of years focusing on reducing the errors in a relatively young branch in crystallography. Personally, I feel like it is a ridiculous pursuit to attempt to achieve a precision in the intensities of 5% and then throw away this carefully extracted information by tolerating refinement R factors of the amplitudes of above 20%. I intend to continue not in XFEL research specifically, but I shall attempt to re-parameterise macromolecular model refinement as a whole, to be applied to synchrotron, XFEL crystallography and electron microscopy with the aim to reduce both the R factor gap and over-fitting of models to data.

Bibliography

Axford, D., Owen, R. L., Aishima, J. et al. (2012), 'In situ macromolecular crystallography using microbeams', *Acta Crystallographica Section D: Biological Crystallography* **68**(5), 592–600.

Ayyer, K., Yefanov, O. M., Oberthür, D. et al. (2016), 'Macromolecular diffractive imaging using imperfect crystals', *Nature* **530**(7589), 202–206.

Barends, T. R., Foucar, L., Ardevol, A. et al. (2015), 'Direct observation of ultrafast collective motions in co myoglobin upon ligand dissociation', *Science* **350**(6259), 445–450.

Barends, T. R., Foucar, L., Botha, S. et al. (2014), 'De novo protein crystal structure determination from x-ray free-electron laser data', *Nature* **505**(7482), 244–247.

Barends, T. R., Foucar, L., Shoeman, R. L. et al. (2013), 'Anomalous signal from s atoms in protein crystallographic data from an x-ray free-electron laser', *Acta Crystallographica Section D: Biological Crystallography* **69**(5), 838–842.

Barty, A., Kirian, R. A., Maia, F. R. et al. (2014), 'Cheetah: software for high-throughput reduction and analysis of serial femtosecond x-ray diffraction data', *Journal of applied crystallography* **47**(3), 1118–1131.

Berman, H. M. (2008), 'The protein data bank: a historical perspective', *Acta Crystallographica Section A: Foundations of Crystallography* **64**(1), 88–95.

Blake, C., Koenig, D., Mair, G. et al. (1965), 'Structure of hen egg-white lysozyme: a three-dimensional fourier synthesis at 2 å resolution', *Nature* **206**(4986), 757–761.

Botha, S., Nass, K., Barends, T. R. et al. (2015), 'Room-temperature serial crystallography at synchrotron x-ray sources using slowly flowing free-standing high-viscosity microstreams', *Acta Crystallographica Section D: Biological Crystallography* **71**(2), 387–397.

Boutet, S., Lomb, L., Williams, G. J. et al. (2012), 'High-resolution protein structure determination by serial femtosecond crystallography', *Science (New York, N.Y.)* **337**(6092), 362–364.

Brehm, W. and Diederichs, K. (2014), 'Breaking the indexing ambiguity in serial crystallography', *Acta crystallographica.Section D, Biological crystallography* **70**, 101–109.

Brunger, A. (1992), 'Xplor manual ver 3.1', *Yale University, New Haven* .

Brünger, A. T. (1990), 'Extension of molecular replacement: A new search strategy based on patterson correlation refinement', *Acta Crystallographica Section A: Foundations of Crystallography* **46**(1), 46–57.

Brunger, A. T. (2007), 'Version 1.2 of the crystallography and nmr system', *Nature protocols* **2**(11), 2728–2733.

Busing, W. R. and Levy, H. A. (1967), 'Angle calculations for 3-and 4-circle x-ray and neutron diffractometers', *Acta Crystallographica* **22**(4), 457–464.

Campbell, J. (1998), 'The practicality of using a three-dimensional fast fourier transform in auto-indexing protein single-crystal oscillation images', *Journal of applied crystallography* **31**(3), 407–413.

Chapman, H. N., Caleman, C. and Timneanu, N. (2014), 'Diffraction before destruction', *Philosophical Transactions of the Royal Society B: Biological Sciences* **369**(1647), 20130313.

Chapman, H. N., Fromme, P., Barty, A. et al. (2011), 'Femtosecond x-ray protein nanocrystallography', *Nature* **470**(7332), 73–77.

Colletier, J.-P., Sawaya, M. R., Gingery, M. et al. (2016), 'Mosquito larvicide binab revealed by de novo phasing with an x-ray laser', *Nature* **539**(7627), 43–47.

Conrad, C. E., Basu, S., James, D. et al. (2015), 'A novel inert crystal delivery medium for serial femtosecond crystallography', *IUCrJ* **2**(4), 421–430.

DePonte, D. P., Weierstall, U., Starodub, D. et al. (2008), 'Gas dynamic virtual nozzle for generation of microscopic droplet streams', *J. Phys. D: Appl. Phys.* **41**, 195505.

Ekeberg, T., Svenda, M., Abergel, C. et al. (2015), 'Three-dimensional reconstruction of the giant mimivirus particle with an x-ray free-electron laser', *Physical review letters* **114**(9), 098102.

Foucar, L. (2016), '*CFEL–ASG Software Suite* (*CASS*): usage for free-electron laser experiments with biological focus', *Journal of Applied Crystallography* **49**(4), 1336–1346.

Fry, E., Acharya, R. and Stuart, D. (1993), 'Methods used in the structure determination of foot-and-mouth disease virus', *Acta crystallographica.Section A, Foundations of crystallography* **49**, 45–55.

Fuller, F. D., Gul, S., Chatterjee, R. et al. (2017), 'Drop-on-demand sample delivery for studying biocatalysts in action at xfels', *Nature methods* **14**(4), 443.

Gati, C., Oberthuer, D., Yefanov, O. et al. (2017), 'Atomic structure of granulin determined from native nanocrystalline granulovirus using an x-ray free-electron laser', *Proceedings of the National Academy of Sciences* **114**(9), 2247–2252.

Gildea, R. J., Waterman, D. G., Parkhurst, J. M. et al. (2014), 'New methods for indexing multi-lattice diffraction data', *Acta Crystallographica Section D: Biological Crystallography* **70**(10), 2652–2666.

Ginn, H. M., Brewster, A. S., Hattne, J. et al. (2015b), 'A revised partiality model and post-refinement algorithm for x-ray free-electron laser data', *Acta Crystallographica Section D: Biological Crystallography* **71**(6), 1400–1410.

Ginn, H. M., Evans, G., Sauter, N. and Stuart, D. (2016a), 'On the release of cppxfel for processing x-ray free-electron laser images', *Journal of Applied Crystallography* **49**(3), 1065–1072.

Ginn, H. M., Messerschmidt, M., Ji, X. et al. (2015a), 'Structure of cpv17 polyhedrin determined by the improved analysis of serial femtosecond crystallographic data', *Nature Communications* **6**(6435).

Ginn, H. M., Roedig, P., Kuo, A. et al. (2016c), 'Taketwo: an indexing algorithm suited to still images with known crystal parameters', *Acta Crystallographica Section D: Structural Biology* **72**(8), 956–965.

Ginn, H. M. and Stuart, D. I. (2016b), 'Recovery of data from perfectly twinned virus crystals revisited', *Acta Crystallographica Section D: Structural Biology* **72**(6), 817–822.

Ginn, H. M. and Stuart, D. I. (2017), 'The slip-and-slide algorithm: a refinement protocol for detector geometry', *Journal of synchrotron radiation* **24**(6), 1152–1162.

Goldsmith, M. R., Shimada, T. and Abe, H. (2005), 'The genetics and genomics of the silkworm, bombyx mori*', *Annu.Rev.Entomol.* **50**, 71–100.

Gouet, P., Courcelle, E., Stuart, D. I. and Metoz, F. (1999), 'Espript: analysis of multiple sequence alignments in postscript', *Bioinformatics (Oxford, England)* **15**(4), 305–308.

Hart, P., Boutet, S., Carini, G. et al. (2012), The CSPAD megapixel x-ray camera at LCLS, *in* 'Proc. SPIE', Vol. 8504, p. 85040.

Hattne, J., Echols, N., Tran, R. et al. (2014), 'Accurate macromolecular structures using minimal measurements from x-ray free-electron lasers', *Nature methods* **11**(5), 545–548.

Henrich, B., Becker, J., Dinapoli, R. et al. (2011), 'The adaptive gain integrating pixel detector agipd a detector for the european xfel', *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **633**, S11–S14.

Ijiri, H., Coulibaly, F., Nishimura, G. et al. (2009), 'Structure-based targeting of bioactive proteins into cypovirus polyhedra and application to immobilized cytokines for mammalian cell culture', *Biomaterials* **30**(26), 4297–4308.

Ikeda, K., Nakazawa, H., Shimo-Oka, A. et al. (2006), 'Immobilization of diverse foreign proteins in viral polyhedra and potential application for protein microarrays', *Proteomics* **6**(1), 54–66.

Ji, X., Axford, D., Owen, R. et al. (2015), 'Polyhedra structures and the evolution of the insect viruses', *Journal of structural biology* **192**(1), 88–99.

Johansson, L. C., Arnlund, D., Katona, G. et al. (2013), 'Structure of a photosynthetic reaction centre determined by serial femtosecond crystallography', *Nature communications* **4**(2911).

Kabsch, W. (1993), 'Automatic processing of rotation diffraction data from crystals of initially unknown symmetry and cell constants', *Journal of applied crystallography* **26**(6), 795–800.

Kabsch, W. (2010), 'Xds', *Acta Crystallographica Section D: Biological Crystallography* **66**(2), 125–132.

Kabsch, W. (2014), 'Processing of x-ray snapshots from crystals in random orientations', *Acta Crystallographica Section D: Biological Crystallography* **70**(8), 2204–2216.

Kameshima, T., Ono, S., Kudo, T. et al. (2014), 'Development of an x-ray pixel detector with multi-port charge-coupled device for x-ray free-electron laser experiments', *Review of Scientific Instruments* **85**(3), 033110.

Kang, Y., Zhou, X. E., Gao, X. et al. (2015), 'Crystal structure of rhodopsin bound to arrestin by femtosecond x-ray laser', *Nature* **523**(7562), 561–567.

Kartha, G., Bello, J. and Harker, D. (1967), 'Tertiary structure of ribonuclease', *Nature* **213**(5079), 862–865.

Kendrew, J. C., Bodo, G., Dintzis, H. M. et al. (1958), 'A three-dimensional model of the myoglobin molecule obtained by x-ray analysis', *Nature* **181**(4610), 662–666.

Kern, J., Alonso-Mori, R., Hellmich, J. et al. (2012), 'Room temperature femtosecond x-ray diffraction of photosystem ii microcrystals', *Proceedings of the National Academy of Sciences* **109**(25), 9721–9726.

Kern, J., Alonso-Mori, R., Tran, R. et al. (2013), 'Simultaneous femtosecond x-ray spectroscopy and diffraction of photosystem ii at room temperature', *Science* **340**(6131), 491–495.

Kern, J., Tran, R., Alonso-Mori, R. et al. (2014), 'Taking snapshots of photosynthetic water oxidation using femtosecond x-ray diffraction and spectroscopy', *Nature communications* **5**(4371).

Korostelev, A., Bertram, R. and Chapman, M. S. (2002), 'Simulated-annealing real-space refinement as a tool in model building', *Acta crystallographica.Section D, Biological crystallography* **58**(Pt 5), 761–767.

Kroon-Batenburg, L. M., Schreurs, A. M., Ravelli, R. B. and Gros, P. (2015), 'Accounting for partiality in serial crystallography using ray-tracing principles', *Acta Crystallographica Section D: Biological Crystallography* **71**(9), 1799–1811.

Kunimi, Y. (2007), 'Current status and prospects on microbial control in japan', *Journal of invertebrate pathology* **95**(3), 181–186.

Kupitz, C., Basu, S., Grotjohann, I. et al. (2014), 'Serial time-resolved crystallography of photosystem ii using a femtosecond x-ray laser', *Nature* **513**(7517), 261–265.

Larochelle, S. (2015), 'Structural biology: Solving xfel's image problem', *Nature methods* **12**(5), 392–392.

Lawrence, R. M., Conrad, C. E., Zatsepin, N. A. et al. (2015), 'Serial femtosecond x-ray diffraction of enveloped virus microcrystals', *Structural Dynamics* **2**(4), 041720.

Leigh, J. B. and Rosenbaum, G. (1974), 'A report on the application of synchrotron radiation to low-angle scattering', *Journal of Applied Crystallography* **7**(2), 117–121.

Liu, W., Wacker, D., Gati, C. et al. (2013), 'Serial femtosecond crystallography of g protein-coupled receptors', *Science (New York, N.Y.)* **342**(6165), 1521–1524.

Maia, F. R. (2012), 'The coherent x-ray imaging data bank', *Nature methods* **9**(9), 854–855.

Mariani, V., Morgan, A., Yoon, C. H. et al. (2016), 'OnDA: online data analysis and feedback for serial x-ray imaging', *Journal of applied crystallography* **49**(3), 1073–1080.

Mozzanica, A., Bergamaschi, A., Cartier, S. et al. (2014), 'Prototype characterization of the JUNGFRAU pixel detector for SwissFEL', *Journal of Instrumentation* **9**(05), C05010.

Nakane, T., Hanashima, S., Suzuki, M. et al. (2016b), 'Membrane protein structure determination by SAD, SIR, or SIRAS phasing in serial femtosecond crystallography using an iododetergent', *Proceedings of the National Academy of Sciences* **113**(46), 13039–13044.

Nakane, T., Joti, Y., Tono, K. et al. (2016a), 'Data processing pipeline for serial femtosecond crystallography at SACLA', *Journal of applied crystallography* **49**(3), 1035–1041.

Nakane, T., Song, C., Suzuki, M. et al. (2015), 'Native sulfur/chlorine SAD phasing for serial femtosecond crystallography', *Acta Crystallographica Section D: Biological Crystallography* **71**(12), 2519–2525.

Nango, E., Royant, A., Kubo, M. et al. (2016), 'A three-dimensional movie of structural changes in bacteriorhodopsin', *Science* **354**(6319), 1552–1557.

Nass, K., Meinhart, A., Barends, T. R. et al. (2016), 'Protein structure determination by single-wavelength anomalous diffraction phasing of x-ray free-electron laser data', *IUCrJ* **3**(3), 180–191.

Nelder, J. A. and Mead, R. (1965), 'A simplex method for function minimization', *The computer journal* **7**(4), 308–313.

Neutze, R. and Moffat, K. (2012), 'Time-resolved structural studies at synchrotrons and x-ray free electron lasers: opportunities and challenges', *Current opinion in structural biology* **22**(5), 651–659.

Neutze, R., Wouts, R., van der Spoel, D., Weckert, E. and Hajdu, J. (2000), 'Potential for biomolecular imaging with femtosecond x-ray pulses', *Nature* **406**(6797), 752–757.

Oghbaey, S., Sarracini, A., Ginn, H. M. et al. (2016), 'Fixed target combined with spectral mapping: approaching 100% hit rates for serial crystallography', *Acta Crystallographica Section D: Structural Biology* **72**(8), 944–955.

Otwinowski, Z. and Minor, W. (2006), Denzo and scalepack, *in* 'International Tables for Crystallography Volume F: Crystallography of biological macromolecules', Springer, pp. 226–235.

Pellegrini, C. (2012), 'The history of x-ray free-electron lasers', *The European Physical Journal H* **37**(5), 659–708.

Perutz, M. F., Rossmann, M. G., Cullis, A. F. et al. (1960), 'Structure of haemoglobin: a three-dimensional fourier synthesis at 5.5 Å resolution, obtained by x-ray analysis.', *Nature* **185**(4711), 416–422.

Powell, H. R., Johnson, O. and Leslie, A. G. (2013), 'Autoindexing diffraction images with iMosflm', *Acta Crystallographica Section D: Biological Crystallography* **69**(7), 1195–1203.

Redecke, L., Nass, K., DePonte, D. P. et al. (2013), 'Natively inhibited trypanosoma brucei cathepsin B structure determined by using an x-ray laser', *Science (New York, N.Y.)* **339**(6116), 227–230.

Redford, S., Bergamaschi, A., Brückner, M. et al. (2016), 'Calibration status and plans for the charge integrating jungfrau pixel detector for swissfel', *Journal of Instrumentation* **11**(11), C11013.

Roedig, P., Ginn, H. M., Pakendorf, T. et al. (2017), 'High-speed fixed-target serial virus crystallography.', *Nature methods* **14**(8), 805.

Roedig, P., Vartiainen, I., Duman, R. et al. (2015), 'A micro-patterned silicon chip as sample holder for macromolecular crystallography experiments with minimal background scattering', *Scientific Reports* **5**(10451).

Roessler, C. G., Agarwal, R., Allaire, M. et al. (2016), 'Acoustic injectors for drop-on-demand serial femtosecond crystallography', *Structure* **24**(4), 631–640.

Rossmann, M. G., Leslie, A. G. W., Abdel-Meguid, S. S. and Tsukihara, T. (1979), 'Processing and post-refinement of oscillation camera data', *J. Appl. Cryst.* **12**, 570–581.

Sauter, N. K., Grosse-Kunstleve, R. W. and Adams, P. D. (2004), 'Robust indexing for automatic data collection', *Journal of applied crystallography* **37**(3), 399–409.

Sauter, N. K., Hattne, J., Brewster, A. S. et al. (2014), 'Improved crystal orientation and physical properties from single-shot xfel stills', *Acta Crystallographica Section D: Biological Crystallography* **70**(12), 3299–3309.

Schutt, C. t. and Winkler, F. (1977), 'The rotation method in crystallography, edited by uw arndt & aj wonacott'.

Sheldrick, G. M. (2010), 'Experimental phasing with shelxc/d/e: combining chain tracing with density modification', *Acta Crystallographica Section D: Biological Crystallography* **66**(4), 479–485.

Smyth, M., Fry, E., Stuart, D. et al. (1993), 'Preliminary crystallographic analysis of bovine enterovirus', *Journal of molecular biology* **231**(3), 930–932.

Steller, I., Bolotovsky, R. and Rossmann, M. G. (1997), 'An algorithm for automatic indexing of oscillation images using fourier analysis', *Journal of applied crystallography* **30**(6), 1036–1040.

Suga, M., Akita, F., Hirata, K. et al. (2015), 'Native structure of photosystem ii at 1.95 Å resolution viewed by femtosecond x-ray pulses', *Nature* **517**, 99–103.

Suga, M., Akita, F., Sugahara, M. et al. (2017), 'Light-induced structural changes and the site of O=O bond formation in PSII caught by XFEL', *Nature* **543**(7643), 131–135.

Sugahara, M., Song, C., Suzuki, M. et al. (2016), 'Oil-free hyaluronic acid matrix for serial femtosecond crystallography', *Scientific reports* **6**(24484).

Tenboer, J., Basu, S., Zatsepin, N. et al. (2014), 'Time-resolved serial crystallography captures high-resolution intermediates of photoactive yellow protein', *Science* **346**(6214), 1242–1246.

Thomaston, J. L., Alfonso-Prieto, M., Woldeyes, R. A. et al. (2015), 'High-resolution structures of the M2 channel from influenza A virus reveal dynamic pathways for proton stabilization and transduction', *Proceedings of the National Academy of Sciences* **112**(46), 14260–14265.

Thorn, A. and Sheldrick, G. M. (2011), 'ANODE: anomalous and heavy-atom density calculation', *Journal of applied crystallography* **44**(6), 1285–1287.

Uervirojnangkoorn, M., Zeldin, O. B., Lyubimov, A. Y. et al. (2015), 'Enabling x-ray free electron laser crystallography for challenging biological systems from a limited number of crystals', *Elife* **4**, e05421.

Umena, Y., Kawakami, K., Shen, J.-R. and Kamiya, N. (2011), 'Crystal structure of oxygen-evolving photosystem ii at a resolution of 1.9 Å', *Nature* **473**(7345), 55.

Wagner, A., Duman, R., Henderson, K. and Mykhaylyk, V. (2016), 'In-vacuum long-wavelength macromolecular crystallography', *Acta Crystallographica Section D: Structural Biology* **72**(3), 430–439.

Walter, T. S., Diprose, J. M., Mayo, C. J. et al. (2005), 'A procedure for setting up high-throughput nanolitre crystallization experiments. crystallization workflow for initial screening, automated storage, imaging and optimization', *Acta Crystallographica Section D: Biological Crystallography* **61**(6), 651–657.

Waterman, D. G., Winter, G., Gildea, R. J. et al. (2016), 'Diffraction-geometry refinement in the DIALS framework', *Acta Crystallographica Section D: Structural Biology* **72**(4), 558–575.

White, T. A. (2014), 'Post-refinement method for snapshot serial crystallography', *Philosophical transactions of the Royal Society of London.Series B, Biological sciences* **369**(1647), 20130330.

White, T. A., Kirian, R. A., Martin, A. V. et al. (2012), 'CrystFEL: a software suite for snapshot serial crystallography', *J. Appl. Cryst.* **45**, 335–341.

White, T. A., Mariani, V., Brehm, W. et al. (2016), 'Recent developments in CrystFEL', *Journal of Applied Crystallography* **49**(2), 680–689.

Winn, M. D., Ballard, C. C., Cowtan, K. D. et al. (2011), 'Overview of the CCP4 suite and current developments', *CCP4 Newsletter on Protein Crystallography* **67**(4), 235–242.

Wyckoff, H., Hardman, K. D., Allewell, N. et al. (1967), 'The structure of ribonuclease-S at 6 Å resolution', *Journal of Biological Chemistry* **242**(16), 3749–3753.

Xiao, C., Fischer, M. G., Bolotaulo, D. M. et al. (2017), 'Cryo-EM reconstruction of the Cafeteria roenbergensis virus capsid suggests novel assembly pathway for giant viruses', *Scientific Reports* **7**(1), 5484.

Yamashita, K., Pan, D., Okuda, T. et al. (2015), 'An isomorphous replacement method for efficient de novo phasing for serial femtosecond crystallography', *Scientific reports* **5**, 14017.

Yefanov, O., Mariani, V., Gati, C. et al. (2015), 'Accurate determination of segmented x-ray detector geometry', *Optics express* **23**(22), 28459–28470.

Young, I. D., Ibrahim, M., Chatterjee, R. et al. (2016), 'Structure of photosystem II and substrate binding at room temperature', *Nature* **540**(7633), 453–457.

Zhang, H., Unal, H., Gati, C. et al. (2015), 'Structure of the angiotensin receptor revealed by serial femtosecond crystallography', *Cell* **161**(4), 833–844.

Zhang, X., Zhao, F., Wu, Y. et al. (2017), 'Crystal structure of a multi-domain human smoothened receptor in complex with a super stabilizing ligand', *Nature Communications* **8**(15383).

Zouni, A., Witt, H.-T., Kern, J., Fromme, P. et al. (2001), 'Crystal structure of photosystem II from synechococcus elongatus at 3.8 Å resolution', *Nature* **409**(6821), 739.

Chapter 8

Appendix

Journal of
Applied Crystallography

ISSN 1600-5767

Received 15 June 2014
Accepted 30 July 2014

SynchLink: an iOS app for ISPyB

Helen Mary Ginn,[a,b] Ghita Kouadri Mostefaoui,[a] Karl Erik Levik,[a] Jonathan Mark Grimes,[a,b] Martin Austin Walsh,[a] Alun William Ashton[a]* and David Ian Stuart[a,b]

[a]Diamond House, Diamond Light Source, Harwell Science and Innovation Campus, Didcot, Oxfordshire OX11 0DE, UK, and [b]Division of Structural Biology, The Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford, Oxfordshire OX3 7BN, UK. Correspondence e-mail: alun.ashton@diamond.ac.uk

The macromolecular crystallography (MX) user experience at synchrotron radiation facilities continues to evolve, with the impact of developments in X-ray detectors, computer hardware and automation methods making it possible for complete data sets to be collected on timescales of tens of seconds. Data can be reduced in a couple of minutes and in favourable cases structures solved and refined shortly after. The information-rich database ISPyB, automatically populated by data acquisition software, data processing and structure solution pipelines at the Diamond Light Source beamlines, allows users to automatically track MX experiments in real time. In order to improve the synchrotron users' experience, efficient access to the data contained in ISPyB is now provided *via* an iOS 6.0+ app for iPhones and iPads. This provides users, both local and remote, with a succinct summary of data collection, visualization of diffraction images and crystals, and key metrics for data quality in real time.

1. Introduction

Synchrotron radiation has revolutionized the field of macromolecular crystallography (MX) (Dauter *et al.*, 2010), so that today the vast majority of structure determinations use synchrotron radiation. As a result of advances in detector developments, improved X-ray optics, sample automation and associated software, measurement of diffraction data is now possible in time frames of tens of seconds. There is now also a widespread availability of automated sample mounting (a summary of robotics for MX is given on the RoboSync web site: http://smb.slac.stanford.edu/facilities/hardware/SAM/robosync) and automated pipelines for data integration and scaling (Holton & Alber, 2004; Minor *et al.*, 2006; Monaco *et al.*, 2013; Vonrhein *et al.*, 2011; Winter, 2010). The speed at which experimental data are collected and the rapidity of subsequent online data analysis make experiment tracking *via* databases almost obligatory. However, ensuring efficient use of synchrotron beamtime while dealing with the logistics of user access to the facilities has always been challenging, and developing robust protocols for remotely accessing beamlines has provided many advantages to both the facilities and users (Soltis *et al.*, 2008). To increase the efficiency of beamtime usage while simplifying the logistics of access, many facilities encourage users to form larger collaborative groups which can be geographically disperse, allowing large numbers of samples to be assessed in allocated blocks of time. This idea of categorizing users into block allocation groups (BAGS) originated at the European Synchrotron Radiation Facility (ESRF) and has proved very successful. Key to the success of this access model has been ISPyB (Delageniere *et al.*, 2011; Beteva *et al.*, 2006), a laboratory information system that is in place at the ESRF and Diamond Light Source. ISPyB enables users to manage and track, through a web-based interface, transport of their samples to the synchrotron as well as providing reports of data collections and analyses as they are executed in real time at the beamline.

Extending the reach of ISPyB to smartphones and tablets provides users with the ability to access their synchrotron data from any location on an easily portable device and further enhances the user experience. Therefore, we have developed an iOS 6.0+ app, *SynchLink*, to allow MX users of Diamond to interact with ISPyB to monitor their experiments in real time. We chose iOS in the first instance to mitigate the number of variables, limiting the types of hardware to support the iPhone/iPad (iOS 6.0+), and therefore we have used the native Apple development tools, *i.e.* the Objective-C language and the Xcode integrated development environment (https://developer.apple.com/xcode/).

2. Description

2.1. Design of graphical user interface

To provide ease of use with a small handheld device such as an iPhone, *SynchLink* is arranged as a content drill-down scheme, starting from the highest level screen which lists each visit to the synchrotron, leading to a list of data collections for that visit, and finally down to the associated data and data analysis (Fig. 1). We have developed separate interfaces to use the differing real estate on the phone and tablet devices optimally. In the case of the iPhone a single view is visible and a simple drill-down navigation system allows clutter-free information to be viewed. The crucial information from each data collection has been carefully designed to be presented on the top level page of each synchrotron visit. This allows a complete summary of each data collection and associated data analysis for a synchrotron trip to be reviewed on a single screen from just one screen touch (Fig. 2). This is achieved by use of graphics and colours. For instance, the time at which an experiment was performed has been compressed into an analogue clock face, whose colour reflects the success of the data collection strategy or autoprocessing programs

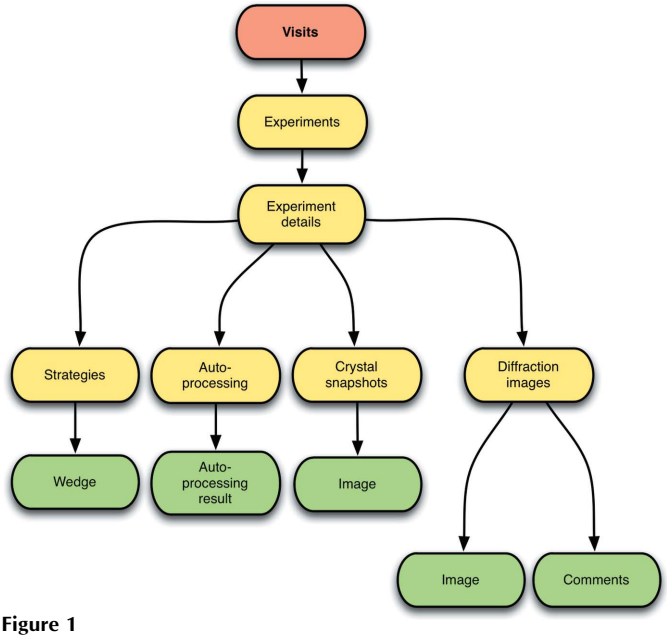


Figure 1
Summary of the individual screens within the navigation system of *SynchLink*. Starting from the top level where a user can see all synchrotron visits and their associated proposal, users can drill down through an experiment to see data collection strategy information, autoprocessing results, crystal snapshots or diffraction patterns.

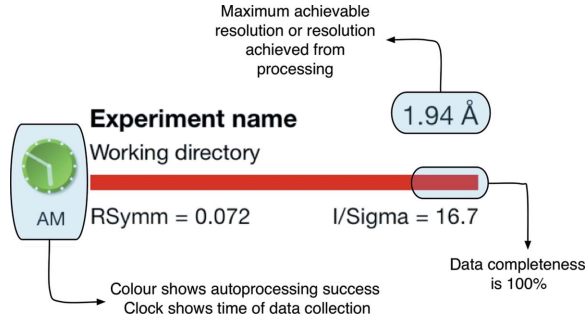


Figure 2
Detail of the information content provided by *SynchLink* through the iPhone screen. The key parameters associated with the data collection and reduction statistics, such as file names, directory paths, data quality parameters $R_{\rm symm}$ and $\langle I/\sigma(I)\rangle$, and data completeness, are presented in a single-screen list for each synchrotron visit to provide fast, direct and informative summaries to the user.

(grey for unknown, red for failed, yellow for indexing success and green for full autoprocessing success). The completeness of the data is represented by the length of the red bar across the entry for the experiment. Finally $R_{\rm symm}$, $\langle I/\sigma(I)\rangle$ and the resolution (user configured to be either that achieved by the autoprocessing or the resolution at the edge of the detector) are shown to provide a quick assessment of data quality (Fig. 2). From this top level summary the user can then easily drill down to view extended details of the experiments, autoprocessing or data collection strategy results, and diffraction and other images associated with the data collection experiments (Fig. 1). In the case of the iPad the larger screen allows a dual page view, which simultaneously displays both the data collection list and a more detailed view of data from a selected data collection (Fig. 3).

A key consideration in the design of the graphical user interface was ease of location of a specific data collection or series of experiments through keyword input. Hence, search functionality has been enabled both from the top level (all synchrotron trips) and from within a specific trip or visit. Searching matches data collections or experiments with a particular name or working directory, or any of the associated comments. In conjunction with a well structured directory and file naming system, this provides an extremely rapid and powerful facility for searching results for particular projects with a bare minimum of information. Furthermore, since comments for data collections are editable from within the app, the use of defined tags can enhance the search facility, enabling searches for particular experiments or experimental outcomes at a later date.

2.2. Information presented

The *SynchLink* app has detailed pages presenting the relevant data from data collections and data reduction (Fig. 1). On drilling down to an individual data collection, experimental parameters are shown (including rotation per image, number of images, % transmission, beamsize), and *SynchLink* provides further drill-down content including crystal snapshots, diffraction patterns, indexing and autoprocessing results, and suggested data collection strategies on the lowest level. Crystal snapshots and diffraction patterns can be seen as thumbnails or clicked on to provide higher-resolution images. One diffraction image per data collection is available at the highest resolution. Multi-touch gestures (zooming and panning) make it

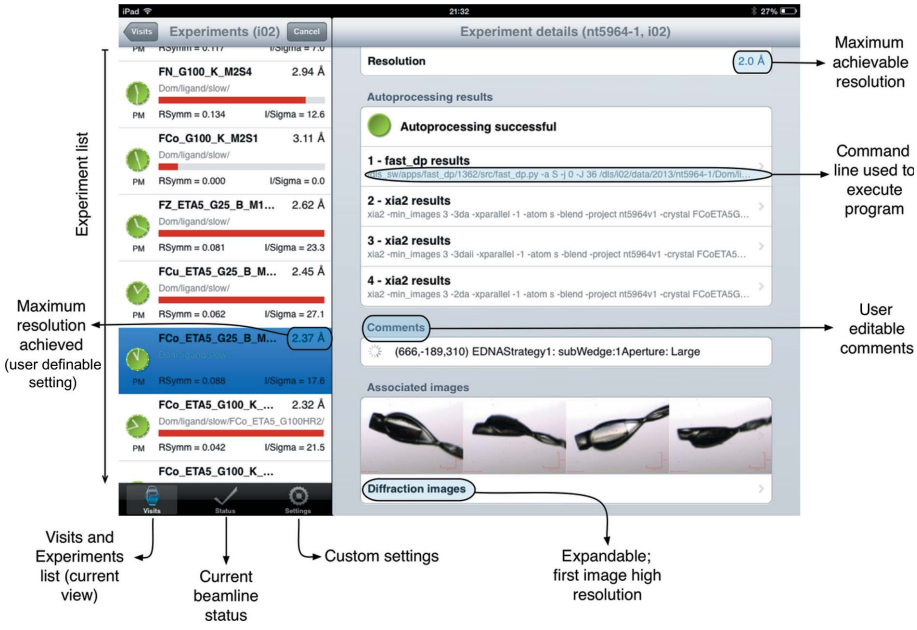


Figure 3
Screenshot from the iPad *SynchLink* screen (for a successful data collection), illustrating how the app exploits the larger screen area of the device. The left panel iPhone view is augmented by the right panel view, which provides both numerical and image data of the selected data collection highlighted in blue. In the above example four data processing runs which were executed automatically are displayed, as well as four snapshots of the mounted sample taken at 90° intervals. Images (*e.g.* crystal images) can be touched to save and to enlarge, whilst the autoprocessing results are also action buttons that bring up a summary of the results. Other touch screen points and features are annotated.

easier to see the potentially tiny reflections produced on low-divergence beamlines with pixel array detectors. For data collections of less than 20 images, which are categorized as data collection strategy experiments, the most likely point group and cell dimensions are presented and a range of data collection strategies are calculated using *EDNA* (Incardona *et al.*, 2009). Each of these strategies can be expanded to see the individual rotation wedges suggested. For full data collections (20 or more images), autoprocessing results (if successful) provide data quality measurements using *FastDP* and *xia2* (Winter, 2010): completeness, $\langle I/\sigma(I)\rangle$ and $R_{\rm symm}$, both overall and in the highest resolution bin, the maximum resolution integrated, the estimated space group, and unit-cell dimensions.

SynchLink also provides a service to show the current status of the storage ring and each beamline, prioritizing MX beamlines at the top of the list.

3. Discussion

Two important issues we have addressed when developing *SynchLink* are the speed of access and minimizing usage of mobile data networks.

Owing to the nature of internet access on smartphones and tablets, often using mobile data networks with data limits, care has been taken to ensure that *SynchLink* downloads the minimum possible data in XML format to avoid compromising responsiveness and to minimize loading times. Furthermore, *SynchLink* has been designed with 'one step ahead of the user' in mind. Data for the next screen in the drill-down hierarchy is loaded while the user is on the previous screen, as long as the content size is sufficiently low to keep traffic to a minimum on mobile data networks. This is aided by the ability of the web service to execute database queries that join information from multiple tables, thereby reducing the total number of calls the iOS device must make to retrieve information.

To avoid draining the battery charge on a small mobile device, it is inadvisable to keep an open connection to a server. *SynchLink* therefore uses multiple single connections that are rapidly opened and closed to save battery life. However, this means that every connection must be authenticated. Full authentication is time consuming as it involves communication with an authentication server, so once completed, the client then obtains a time-limited token which can be used for subsequent authentication. After one hour of inactivity, full authentication is once again required.

4. Conclusions

We hope that *SynchLink* will enable users with iOS devices to access their data more simply and efficiently than by using the ISPyB web interface, and will be particularly valuable when a computer and/or wired internet connection is unavailable. Access to real-time results from the beamline *via* a personal mobile device or tablet should improve productivity for researchers following live data collection remotely. *SynchLink* has been made available to users on the iOS App Store and can be freely downloaded in any of the countries where the Apple App Store is available.

The authors would like to thank the Diamond MX village scientists for feedback during beta testing and Francois Cesmat and Bill Pulford at Diamond for technical assistance.

References

Beteva, A. *et al.* (2006). *Acta Cryst.* D**62**, 1162–1169.
Dauter, Z., Jaskolski, M. & Wlodawer, A. (2010). *J. Synchrotron Rad.* **17**, 433–444.
Delageniere, S. *et al.* (2011). *Bioinformatics*, **27**, 3186–3192.
Holton, J. & Alber, T. (2004). *Proc. Natl Acad. Sci. USA*, **101**, 1537–1542.
Incardona, M.-F., Bourenkov, G. P., Levik, K., Pieritz, R. A., Popov, A. N. & Svensson, O. (2009). *J. Synchrotron Rad.* **16**, 872–879.
Minor, W., Cymborowski, M., Otwinowski, Z. & Chruszcz, M. (2006). *Acta Cryst.* D**62**, 859–866.
Monaco, S., Gordon, E., Bowler, M. W., Delagenière, S., Guijarro, M., Spruce, D., Svensson, O., McSweeney, S. M., McCarthy, A. A., Leonard, G. & Nanao, M. H. (2013). *J. Appl. Cryst.* **46**, 804–810.
Soltis, S. M. *et al.* (2008). *Acta Cryst.* D**64**, 1210–1221.
Vonrhein, C., Flensburg, C., Keller, P., Sharff, A., Smart, O., Paciorek, W., Womack, T. & Bricogne, G. (2011). *Acta Cryst.* D**67**, 293–302.
Winter, G. (2010). *J. Appl. Cryst.* **43**, 186–190.

Journal of
Synchrotron Radiation

ISSN 1600-5775

Received 19 September 2014
Accepted 30 November 2014

PDB References: 4x4b; 4x4c; 4x4d; 4x4e; 4x4f; 4x4g; 4x4h; 4x4i

Radiation damage to nucleoprotein complexes in macromolecular crystallography

Charles Bury,[a] Elspeth F. Garman,[a] Helen Mary Ginn,[a] Raimond B. G. Ravelli,[b] Ian Carmichael,[c] Geoff Kneale[d] and John E. McGeehan[d]*

[a]Laboratory of Molecular Biophysics, Department of Biochemistry, University of Oxford, South Parks Road, Oxford OX1 3QU, UK, [b]Institute of Nanoscopy, Maastricht University, PO Box 616, Maastricht 6200 MD, The Netherlands, [c]Notre Dame Radiation Laboratory, University of Notre Dame, Notre Dame, IN 46556, USA, and [d]Molecular Biophysics, Institute of Biomedical and Biomolecular Sciences, University of Portsmouth, King Henry 1st Street, Portsmouth PO1 2DY, UK. *E-mail: john.mcgeehan@port.ac.uk

Significant progress has been made in macromolecular crystallography over recent years in both the understanding and mitigation of X-ray induced radiation damage when collecting diffraction data from crystalline proteins. In contrast, despite the large field that is productively engaged in the study of radiation chemistry of nucleic acids, particularly of DNA, there are currently very few X-ray crystallographic studies on radiation damage mechanisms in nucleic acids. Quantitative comparison of damage to protein and DNA crystals separately is challenging, but many of the issues are circumvented by studying pre-formed biological nucleoprotein complexes where direct comparison of each component can be made under the same controlled conditions. Here a model protein–DNA complex C.Esp1396I is employed to investigate specific damage mechanisms for protein and DNA in a biologically relevant complex over a large dose range (2.07–44.63 MGy). In order to allow a quantitative analysis of radiation damage sites from a complex series of macromolecular diffraction data, a computational method has been developed that is generally applicable to the field. Typical specific damage was observed for both the protein on particular amino acids and for the DNA on, for example, the cleavage of base-sugar N_1-C and sugar-phosphate $C-O$ bonds. Strikingly the DNA component was determined to be far more resistant to specific damage than the protein for the investigated dose range. At low doses the protein was observed to be susceptible to radiation damage while the DNA was far more resistant, damage only being observed at significantly higher doses.

Keywords: macromolecular crystallography; radiation damage; protein–DNA complexes; specific damage.

1. Introduction

Since the advent of powerful third-generation synchrotron sources, significant progress has been made in the field of X-ray crystallography regarding the analysis of X-ray induced radiation damage to proteins during both 100 K and room temperature (RT) diffraction data collection. A collection of careful systematic studies has increased awareness of the issues and has served to provide practical solutions, from optimized data-collection strategies (Zeldin *et al.*, 2013*a*; Flot *et al.*, 2010) through to the addition of free-radical scavengers, to help mitigate the destructive effects [for a summary of scavenger studies see Allan *et al.* (2013)]. As we strive to solve the structures of macromolecules with ever-increasing complexity, it is also important to consider the effects of radiation damage to fundamental non-protein biological components. A wealth of radiation damage studies for nucleic acids has been provided by a strong community of radiation chemists, and mechanisms have been deduced from experiments on individual nucleotides in isolation through to irradiation of whole cells and tissues. The latter studies underpin the development of current radiotherapies in the treatment of a range of cancers, but the full mechanistic X-ray damage landscape from atoms to organisms is far from complete. The atomic resolution that can be provided by X-ray crystallography has great potential to help link the fields of radiation chemistry and radiation biology by providing an atomistic view of radiation damage to intact biological complexes, particularly those involving nucleoproteins.

Obtaining an understanding of damage to these complexes is important because DNA is rarely naked in a cell, operating instead in a dynamic environment interacting with a plethora of proteins such as the nucleosomal histones through to high-order replication, transcription, modification and repair machinery. In contrast with X-ray radiation damage investigations on isolated crystallized proteins or nucleic acids, there has been relatively little work focused on the quantitative comparison of damage observations between the two. Radiation damage solution studies of biological complexes, including the lac repressor-operator (Begusová *et al.*, 2001; Charlier *et al.*, 2002; Eon *et al.*, 2001) and estrogen response element-receptor complexes (Stisova *et al.*, 2006), suggest a significant bias towards protein damage as compared with the nucleic acid component (reviewed by Spotheim-Maurizot & Davídková, 2011). Previous studies have highlighted the importance of differential radiation induced damage and its effect on the interactions required to form nucleoprotein complexes. It has been shown that bound DNA affords a substantial protective effect in the complex between the DNA glycosylase Fpg and its abasic DNA target site (Gillard *et al.*, 2004) and that the naked protein has relatively higher sensitivity to irradiative damage. The authors note that this may be a common feature in DNA repair enzymes and may make a significant contribution to biological damage processes. Furthermore, subsequent studies by the same group on the lac-operator model system were able to identify damage to specific amino acids of the DNA binding protein component using a combination of spectroscopy and mass spectrometry (Gillard *et al.*, 2007). In the current study, we aim to build on this information by employing techniques that will allow us to investigate these radiation induced damage processes at the atomic level.

In macromolecular crystallographic investigations, direct quantitative comparison between damage to different proteins in different crystalline conditions is intrinsically challenging and this is made yet more difficult when one considers nucleic acid crystals. DNA crystals, for example, often form pseudo-continuous helical lattices that are far from representative of their solution state. On the other hand, the ability to generate crystals of protein–DNA and protein–RNA complexes provides a tractable basis to compare relative damage profiles between the two components under the same controlled experimental conditions.

X-ray radiation damage is generally recognized as the major limitation in macromolecular crystallography (MX), resulting in loss of diffraction data quality from 'global radiation damage', observed in reciprocal space as loss of diffraction intensity, loss of resolution, unit-cell expansion and, usually, increased mosaicity (Murray & Garman, 2002; Owen *et al.*, 2006; Ravelli & McSweeney, 2000; Burmeister, 2000; Garman, 2010). Furthermore, radiation damage can be the result of localized chemical changes to the protein structure, 'specific radiation damage', which at 100 K data collection temperatures is observed to follow a reproducible general trend. These effects include reduction of metallo-centres, disulfide bond breakage, acidic residue decarboxylation, hydroxyl group loss from tyrosines and methylthio group damage, all of which can compromise biological data interpretation (Weik *et al.*, 2000; Ravelli & McSweeney, 2000; Burmeister, 2000; Fioravanti *et al.*, 2007; Garman & Owen, 2006; Ravelli & Garman, 2006). Radiation damage can similarly result in specific site damage to nucleic acids such as base-sugar N_1-C or sugar-phosphate C—O bond cleavage in DNA (Théodore *et al.*, 2006).

The major underlying cause of radiation damage is the photoabsorption in the crystal (solvent and macromolecule) of incoming X-rays. The resulting high-energy photoelectron, typically with a range of 3 to 4 μm for the incident X-ray energies used in MX (Nave & Hill, 2005; Sanishvili *et al.*, 2011) as well as associated Auger electrons, can produce a cascade of many hundreds of further ionization and excitation events (O'Neill *et al.*, 2002) releasing secondary electrons which then lose energy to the medium and gradually thermalize. Inelastic Compton scattering of incoming X-rays provides another source of ionizing electrons which are also mobile in the crystal at 100 K and help spread the damage. Direct damage refers to ionization events occurring within the protein or DNA molecule, and indirect damage is that transferred from the effects of absorption in the surrounding solvent.

There have been numerous studies devoted to understanding the radiation damage mechanisms in DNA, involving resonant attachment to either the bases or the sugar-phosphate backbone, by which secondary sub-ionization-inducing low-energy electrons (0–15 eV) inflict this damage (Michael & O'Neill, 2000; Boudaïffa *et al.*, 2000; Alizadeh *et al.*, 2014; Ptasińska & Sanche, 2007; Alizadeh & Sanche, 2014). This attachment can result in covalent bond cleavage of the DNA backbone, observed as single strand breaks (SSBs) (Boudaïffa *et al.*, 2000; Simons, 2006; Barrios *et al.*, 2002; Berdys *et al.*, 2004; Michael & O'Neill, 2000). Furthermore, other investigations have shown suspected preferential cleavage of sugar-phosphate C—O bonds over base-sugar N_1-C bonds and other bonds present within the DNA structure (Théodore *et al.*, 2006; Ptasińska & Sanche, 2007).

Other X-ray induced DNA damage studies have demonstrated the importance of the secondary hydroxyl radicals ($OH^{\bullet}$) produced by water radiolysis within solvent channels throughout the crystal structure, which then can proceed to abstract H atoms from deoxyribose and bases (Michael & O'Neill, 2000; Cadet *et al.*, 1999). Alternatively hydroxyl radicals can add to unsaturated linkages within DNA bases, resulting in SSBs and modified bases and sugars (Spotheim-Maurizot & Davídková, 2011). In fact it has been estimated that two-thirds of DNA damage in dilute aqueous solution at room temperature can be attributed to 'indirect' secondary free radical damage from solvent around the DNA (Michael & O'Neill, 2000).

In this paper, a well characterized bacterial protein–DNA complex (C.Esp1396I/DNA) is employed to probe the effects of X-ray induced radiation damage and its quantitative distribution among the various components at the atomic level.

The C-protein C.Esp1396I is a member of a large class of helix-turn-helix proteins that act as transcriptional regulators

of gene expression in both prokaryotic and eukaryotic systems. C.Esp1396I finely regulates a bacterial restriction-modification system, providing a delay in endonuclease expression *via* a concentration-dependent 'on' and 'off' switch. This type of transcriptional regulation is a cornerstone of all biological systems, and requires the recognition and intimate interaction of proteins and specific DNA target sites to form productive complexes.

The C.Esp1396I nucleoprotein tetramer complex [PDB ID: 3clc, resolution 2.8 Å (McGeehan *et al.*, 2008)] was chosen as a model due to the relatively large 35 bp dsDNA component, resulting in the number of protein and DNA atoms being of the same order (protein: 2496 non-hydrogen atoms; DNA: 1429 atoms). This allowed a statistically significant comparison between the frequencies of detected specific damage within the DNA and protein components. In addition, having four identical subunits in the complex as well as pseudo-symmetry in the DNA provides a further indicator of statistical significance.

2. Materials and methods

2.1. Crystallization and X-ray data collection

Expression, purification and crystallization of the native C-protein–DNA operator complex C.Esp1396I was performed as described earlier (McGeehan *et al.*, 2008). Pure C.Esp1396I protein at a concentration of 0.7 mg ml^{-1} was mixed with purified double-stranded 35 bp DNA at a 4:1 molar ratio and then crystallized using a vapour-diffusion sitting-drop method at 293 K with 100 nl nucleoprotein solution and 100 nl mother liquor drop. Crystals grew within one month in 50 m*M* 2-(*N*-morpholino)ethanesulfonic acid (MES), pH 7.5, 25% 2-methyl-2,4-pentanediol (*v*/*v*) and 40 m*M* $MgCl_2$. They were cryoloop mounted and vitrified directly under a 100 K N_2 gas stream, with no additional cryoprotectant agent being added to the mother liquor.

Data were collected at 100 K at beamline ID29, ESRF, using a wavelength of 0.932 Å (13.3 keV) and an ADSC Q315R mosaic CCD detector. A pinhole designed by R. Ravelli and F. Felisaz (EMBL Grenoble) was utilized to produce a 25 µm circular low-divergence beam at the sample. The beam size before the pinhole was 0.212 mm (vertical) × 0.279 mm (horizontal) (Gaussian profile FWHM). A crystal of C.Esp1396I (30 µm × 30 µm × 10 µm) was oriented with the beam direction parallel to the smallest crystal dimension (10 µm). The crystal–detector distance was fixed to 390 mm throughout.

Eight datasets were collected from this C.Esp1396I crystal, each consisting of 100 frames of 1° rotation, each over the same 100° wedge of the crystal, with exposure times per frame and beam attenuation being varied according to the values shown in Table 1. The unattenuated beam (100% beam transmission) used for dataset 7 was included to damage the crystal at a higher rate than in datasets 1–6, and ensure that data had been collected for a large range of absorbed X-ray doses by the last dataset.

Table 1
Data collection details for each data set; all data sets were collected with a $\Delta\varphi = 1°$ per image.

Dataset	Exposure time per frame (s)	Transmission (%)	Cumulative dose (DWD) (MGy)
1	1	30	2.07
2	1	30	6.19
3	1	30	10.31
4	1	30	14.43
5	2	30	20.62
6	1	30	26.78
7	1	100	35.73
8	1	30	44.63

2.2. Dose calculation

In order to interpret the X-ray radiation damage data it is crucial to have an accurate estimate of the absorbed X-ray dose. Based on the image of the crystal mounted on the beamline (Fig. 2*a*), RADDOSE-3D (Zeldin *et al.*, 2013*b*), which now includes a recently developed capability to handle polygonal crystals, was used to calculate the absorbed dose distributions after the first (Fig. 2*b*) and eighth (Fig. 2*c*) datasets. The intensity profile of the 25 µm-diameter beam was modelled as a Gaussian with FWHM values as detailed above. The photon flux prior to attenuation was estimated to be 5×10^{10} photons s^{-1} throughout the experiment, and the resulting diffraction weighted dose (DWD) values (Zeldin *et al.*, 2013*a*) for each dataset are shown in Table 1. The flux value was obtained from the in-beamline ionization gauge I1 which had been cross-checked with a calibrated diode (Owen *et al.*, 2009) the day before the data collection.

The calculation of the crystal absorption coefficients in RADDOSE-3D included the water and the heavy-atom content from the crystallization conditions (the sulfur in the MES buffer and the magnesium and chlorine from the $MgCl_2$). All the results on radiation progression were plotted against the resulting DWD values. A plot of the mean intensity values per whole dataset showing the radiation damage induced decay can be found in Fig. S1 of the supporting information.[1] The dose to half mean intensity, $D_{1/2}$, is approximately 45 MGy, in good agreement with the experimental $D_{1/2}$ dose limit of 43 MGy (Owen *et al.*, 2006).

2.3. Data processing

Each dataset was integrated using *iMOSFLM* (Leslie & Powell, 2007) and scaled using the CCP4 program *AIMLESS* (Winn *et al.*, 2011). For the first dataset, molecular replacement was then performed using the program *PHASER* (McCoy *et al.*, 2007) with the previously deposited C.Esp1396I complex structure PDB entry 3clc (McGeehan *et al.*, 2008) as the search model.

The structure obtained from the first dataset was refined, initially using rigid-body refinement in *REFMAC5* (Murshudov *et al.*, 1997), followed by repeated cycles of restrained, TLS and isotropic *B*-factor refinement. This was

[1] Supporting information for this paper is available from the IUCr electronic archives (Reference: XH5044).

coupled with manual inspection and refinement in *Coot* (Emsley *et al.*, 2010), involving solvent molecule removal and addition. In *REFMAC5*, non-crystallographic symmetry (NCS) restraints were used throughout for the four protein subunits present, except in the flexible loop region corresponding to residues 43–46 (Ball *et al.*, 2009). Final rounds of TLS, restrained and isotropic *B*-factor refinement were then performed for the first dataset using *phenix.refine* in the *PHENIX* program suite (Adams *et al.*, 2010), utilizing the integrated *MolProbity4* (Chen *et al.*, 2010) structure validation within *phenix.refine* to manually treat Ramachandran, rotamer and bond angle/length outliers.

2.4. Automated specific damage location and large dataset collection

To observe the real-space specific damage dynamics throughout the C.Esp1396I complex, the CCP4 program *CAD* was used to create a series of seven merged files combining the observed structure factor amplitudes for the first dataset $F_{o,1}$ with each later dataset ($F_{o,n}$, for $n = 2, \ldots, 8$) individually, which were then all scaled using the CCP4 program *SCALEIT*. A set of seven Fourier difference maps $F_{o,n} - F_{o,1}$, for $n = 2, \ldots, 8$, were then calculated using the CCP4 program *FFT* (Ten Eyck, 1973) in order to determine differences in electron density distribution within the protein–DNA complex between the first dataset and each later dataset, which could then be visualized in *Coot* (Emsley *et al.*, 2010). Thus real-space observations of specific radiation damage progression within the C.Esp1396I crystal with respect to increasing dose could be made.

Fourier difference map peaks coincident with the protein–DNA complex indicated potential specific damage regions which were investigated on a case-by-case basis. For a given dose, points within the unit cell situated on an electron density σ-contouring level (s) have calculated electron densities (in electrons Å^{-3}) of s times the standard deviation (σ) from the mean electron density of the Fourier difference map. Thus the higher the $\pm\sigma$-contouring level in *Coot* at which a difference map peak remained present, the greater the electron density gain/loss from the mean map density was present at that specific point in space.

To provide a large dataset of electron density difference map peaks for each dose (seven in total), representing potential specific damage sites within the C.Esp1396I structure, the *FFT* program was utilized to output the coordinates of all difference map peaks above a given threshold in terms of gain/loss of electrons per Å^3. Initial observations of the difference maps in *Coot* led to a chosen threshold of ± 0.04 e Å^{-3} as being suitable to generate large datasets of specific damage sites for which damage was visible above background noise within the Fourier difference maps. In the following analysis this electron gain/loss threshold will be referred to as the *specific damage onset*. The *FFT* program also output the maximum σ-level threshold, s, for each located difference map peak, which could be converted into a spatially local maximum value for the gain/loss of electrons per Å^3 for each peak.

The seven large datasets of difference map peaks were then processed using a custom-made script, written in the object-oriented scripting language *Python*. This was designed to filter the difference map peak datasets, to (*a*) remove positive peaks relating to electron density increases between the first and a given later dataset (to isolate sites of electron density loss resulting from X-ray radiation damage), and (*b*) remove electron density peaks further than a given threshold distance (Å) from the protein–DNA complex (specified by the user). This threshold was initially chosen to be 2 Å since filtering with threshold distances of 2.5 Å or greater was observed to include peaks which were clearly noise, and a 'by-eye' investigation of the remaining difference map peaks for the first two difference map datasets (for which the number of filtered peaks was not too high) led to the observation that all the difference map peaks corresponding to clear specific damage were below the 2 Å threshold in these datasets.

Furthermore, with a detection radius of 2 Å around each atom, electron density changes between atoms spaced at distances of up to 4 Å apart could be detected (2 × the radius). In the case of covalently bonded atoms, carboxyl acids have C—O bond lengths of ∼1.36 Å, with other C—O bond lengths being ∼1.43 Å, the C—S bond length has been reported to be ∼1.82 Å, the C—N bond length ∼1.47 Å, and the C—C bond length ∼1.54 Å (Engh & Huber, 1991); additionally van der Waals distances for interatomic contacts have been reported to typically reside in the range ∼2 Å (for H···H) to ∼3.2 Å (for $CH_2 \cdots CH_2$) (Ramachandran & Sasisekharan, 1968). Hence the chosen threshold should be sufficient to cover both atomic van der Waals radii and roughly one covalent bond length of three-dimensional space around each atom of the protein–DNA complex, to account for potential non-uniform electron distributions around each atom present. Additionally, if the detection threshold is set to approximately the atom–atom covalent bond lengths for the complex, there is a level of symmetry, since for every atom in the complex the threshold covers the atom itself and approximately the covalent bond length (or lengths if the atom is connected to multiple others) of the bond to its adjacent atom.

The command line operated script allowed a more efficient systematic sweep through the remaining difference map peaks (alongside manual *Coot* inspection) to facilitate the quick removal of those which were not suspected to be specific damage to the molecules in the C.Esp1396I crystal. For each accepted difference map peak, the closest protein–DNA complex atom was calculated, and the spatially local maximum electron loss per Å^3 (as discussed above) for that peak was assigned to this selected atom.

Thus specific damage site datasets were constructed for each Fourier difference map $F_{o,n} - F_{o,1}$, for $n = 2, \ldots, 8$, and further processed, involving the deletion of suspected specific damage sites that were detected in an earlier Fourier difference map, but then not detected in all subsequent maps. This was designed to reduce the inclusion of difference map peaks corresponding to noise, which falsely appear to represent

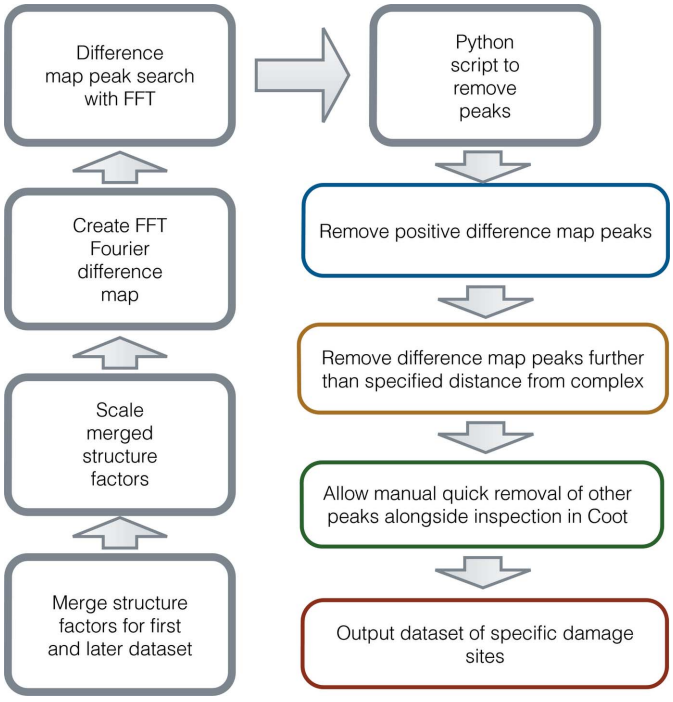


Figure 1
Flow chart for the difference map peak reduction process by systematic analysis, for a given later dataset $n \in \{2, \ldots, 8\}$. Steps performed in the *Python* scripts developed for this work are coloured blue/yellow/green/red.

specific radiation damage in the Fourier difference maps. For each remaining site of specific damage, the dose-dependent dynamics could then be investigated. Fig. 1 provides a flow chart representation of this analysis, in which detected *FFT*-output difference map peaks are filtered, with only suspected specific damage sites remaining.

For a given dataset n, the magnitude of the observed structure factor $F_{o,n}(000)$ was not known since it is equal to the total number of electrons in the unit cell, which could in principle be estimated from the amino acid composition and the m*M* buffer concentration. However, due to the uncertainty in the true solvent concentration within the crystal solvent channels, such a calculation would carry an unknown systematic error. The FFT-output electron density at real-space crystal points is given on an arbitrary fixed scale (*i.e.* multiplied by some constant scale factor k). For crystals differing in protein, DNA or unit-cell solvent composition, this scale factor would be different (Lang *et al.*, 2014) but, since here we are comparing structures all derived from the same crystal, the electron content of the unit is the same and thus the value of k is constant across the dose series and can thus be neglected in this study. Each electron density difference map peak σ-level threshold was converted into a loss/gain of electrons per Å^3 between the first and each later dataset *via*

$$\text{Electron density loss/gain} = \text{Mean electron density} + (\sigma \times s)$$

where σ is the standard deviation from the mean electron density for the difference map, and s is the number of standard deviations of a Fourier difference map peak from the mean electron density. Hence, for each detected difference map peak, a local electron density maximum loss/gain value was determined. These difference peaks were then illustrated using *PyMol* (Schrödinger, LLC) for structural representation and *CCP4mg* (McNicholas *et al.*, 2011) for difference maps showing specific damage examples.

3. Results

3.1. C.Esp1396I protein–DNA complex crystallography

The triangular-shaped crystal used for data collection (Fig. 2*a*) was found to belong to space group $P6_5$ with unit-cell dimensions as detailed in Table 2. In agreement with the original structure, each asymmetric unit contained one tetrameric protein–DNA complex and had a solvent content of 68.7%. The final model for the first dataset was refined to a resolution of 2.8 Å.

To investigate the distribution of specific damage throughout the C.Esp1396I protein–DNA structure with respect to increasing radiation dose, eight successive datasets on the same C.Esp1396I crystal were collected, each exposing the same 100° wedge of the crystal to X-ray radiation. For the later datasets (2 to 8), molecular replacement was performed with *PHASER* using the refined first dataset final model as the search model. The later datasets were refined only using isotropic *B*-factor refinement in *phenix.refine* and were similarly refined to 2.8 Å. Final statistics for model refinements are given in Table 2.

The C-protein–DNA operator complex consists of two 'controller' C-protein dimers (chains *A* and *B*, chains *C* and *D*) bound to a 35 bp DNA operator sequence (chains *E* and *F*) (Fig. 3) (McGeehan *et al.*, 2008). There is a pseudo-dyad axis between bp 17 and 18 of the dsDNA such that chain *A* rotates ∼180° around this axis onto chain *D* (and chain *B* onto chain

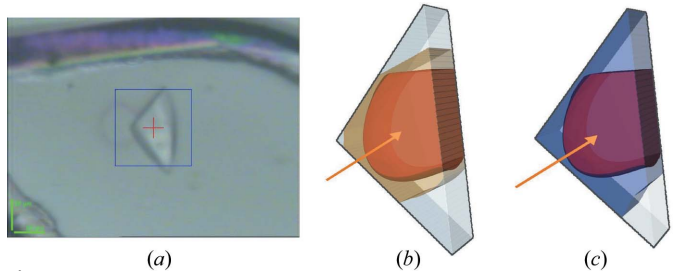


Figure 2
(*a*) The C.Esp1396I crystal within the rayon fibre loop during data collection. Crystal dimensions were x = 30 µm and y = 30 µm. The crystal is positioned in the loop such that dimension z = 10 µm is directly into the page. The blue box shown around the crystal is 50 µm × 50 µm. A faint purple ring, formed by the production of solvated electrons, can be seen to the left of the crystal. This was created from a test shot of the X-ray beam prior to crystal centring and confirms the circular 25 µm beam profile generated from the pinhole. (*b*, *c*) RADDOSE-3D (Zeldin *et al.*, 2013*b*) calculation of dose distributions in the triangular C.Esp1396I crystal after (*b*) the first dataset and (*c*) the eighth dataset. In (*b*) the dose isosurfaces represent 0.16 MGy (white), 3.3 MGy (brown) and 4.2 MGy (orange) and in (*c*) 0.16 MGy (white), 20 MGy (blue) and 47.5 MGy (dark red). The direction of the X-ray beam is indicated with an arrow and the crystal was rotated about a horizontal axis.

Table 2
Data processing and refinement statistics.

Values in parentheses are for the highest-resolution shells. The *B*-factor is estimated from the Wilson plot. For observed F_{obs} and calculated F_{calc} structure factors, $R_{work} = \Sigma|F_{abs} - F_{calc}|/\Sigma F_{obs}$ and R_{free} is the R_{work} formula calculated from a small (5%) test set of randomly selected reflections, output by *phenix.refine* (Adams *et al.*). Unit-cell angles in $P6_5$ are α, β and γ and 90°, 90°, 120°, respectively. The resolution range is 69.5–2.8 Å with 2.95–2.8 Å for the outer shell for all datasets. A plot of the mean intensity values showing the radiation damage induced decay can be found in Fig. S1 of the supporting information.

Dataset	1 4x4b	2 4x4c	3 4x4d	4 4x4e	5 4x4f	6 4x4g	7 4x4h	8 4x4i
Data processing								
Cell dimensions								
a = *b* (Å)	104.29	104.38	104.33	104.43	104.45	104.45	104.33	104.43
c (Å)	139.18	139.26	139.20	139.21	139.19	139.09	139.01	139.08
No. of observations	128219	129102	128743	128973	128476	129426	126479	126715
Unique reflections	21172	21231	21198	21250	21207	21246	21052	21244
Multiplicity	6.1 (6.2)	6.1 (6.2)	6.1 (6.2)	6.1 (6.2)	6.1 (6.2)	6.1 (6.2)	6.0 (6.1)	6.0 (6.1)
Completeness (%)	99.9	99.9	99.9	99.9	99.7	99.9	99.4	99.9
R_{merge}	0.05 (0.30)	0.05 (0.33)	0.05 (0.41)	0.05 (0.47)	0.05 (0.39)	0.06 (0.82)	0.05 (0.63)	0.08 (1.92)
I_n/I_1	1.000	0.988	0.835	0.842	0.729	0.661	0.589	0.514
Refinement								
R_{work}	0.2045	0.2338	0.2346	0.2395	0.2366	0.2336	0.2459	0.2639
R_{free}	0.2616	0.2794	0.2824	0.2848	0.2811	0.2747	0.2929	0.2956
Mean *B*-factor (Å^2)	62.08	65.23	62.93	59.59	66.42	72.59	71.96	81.63
No. non-H atoms								
Protein	2496	2496	2496	2496	2496	2496	2496	2496
DNA	1429	1429	1429	1429	1429	1429	1429	1429
Solvent	5	0	0	0	0	0	0	0
RMSD bond length (Å)	0.011	0.011	0.011	0.011	0.011	0.011	0.011	0.011
RMSD bond angle (°)	1.338	1.338	1.338	1.338	1.338	1.338	1.338	1.338
Ramachandran								
Favoured	93.0	93.0	93.0	93.0	93.0	93.0	93.0	93.0
Outliers	6.7	6.7	6.7	6.7	6.7	6.7	6.7	6.7
Allowed	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3

C). Non-symmetrical binding between protein chains *B* and *C* at the dimer–dimer interface, and a pseudo-symmetrical DNA sequence, prevent true NCS in the complex. The dsDNA is distorted in the complex, due to binding of protein dimers to each operator site, resulting in minor groove compression (leading to a 50° bend at each binding site) and large major groove expansion at the DNA sequence centre (McGeehan *et al.*, 2008).

3.2. Specific damage observations

The automated scripts provided a means to successfully filter difference-map noise peaks from the system under investigation (Fig. 4) but they are also generally applicable for the analysis of other radiation damage datasets. With increasing radiation dose there is an associated increase in the background noise present in the Fourier difference maps due to greater dose-related disorder and non-uniformity in the configuration of each unit cell, which results in the locations of true specific site damage being obscured in high dose datasets. Filtering allowed significant reduction in the level of included noise. Indeed, Fig. 4 shows that at a dose of 45 MGy the number of difference peaks was reduced to around 1% of the

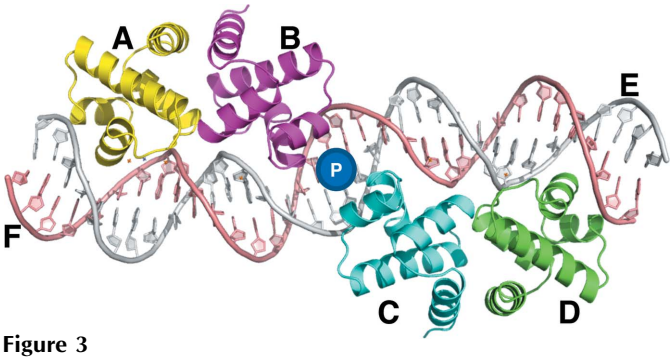


Figure 3
Visual representation of the C.Esp1396I complex rendered using *PyMOL*. Protein (*A*–*D*) and DNA (*E*, *F*) chains are labelled, and the 180° near-NCS symmetry axis is shown by marker *P*.

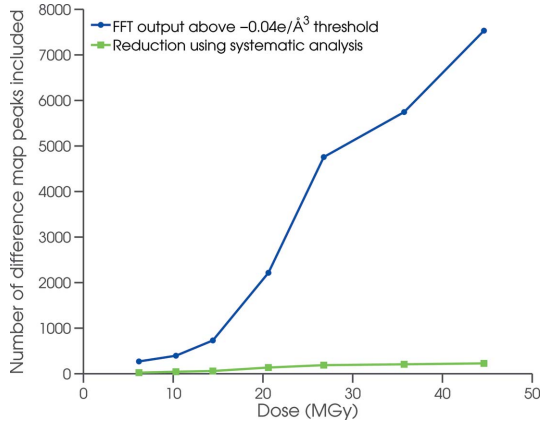


Figure 4
Reduction in the number of detected difference map peaks corresponding to potential specific damage obtained following a systematic analysis using the custom-made script for each dose (MGy).

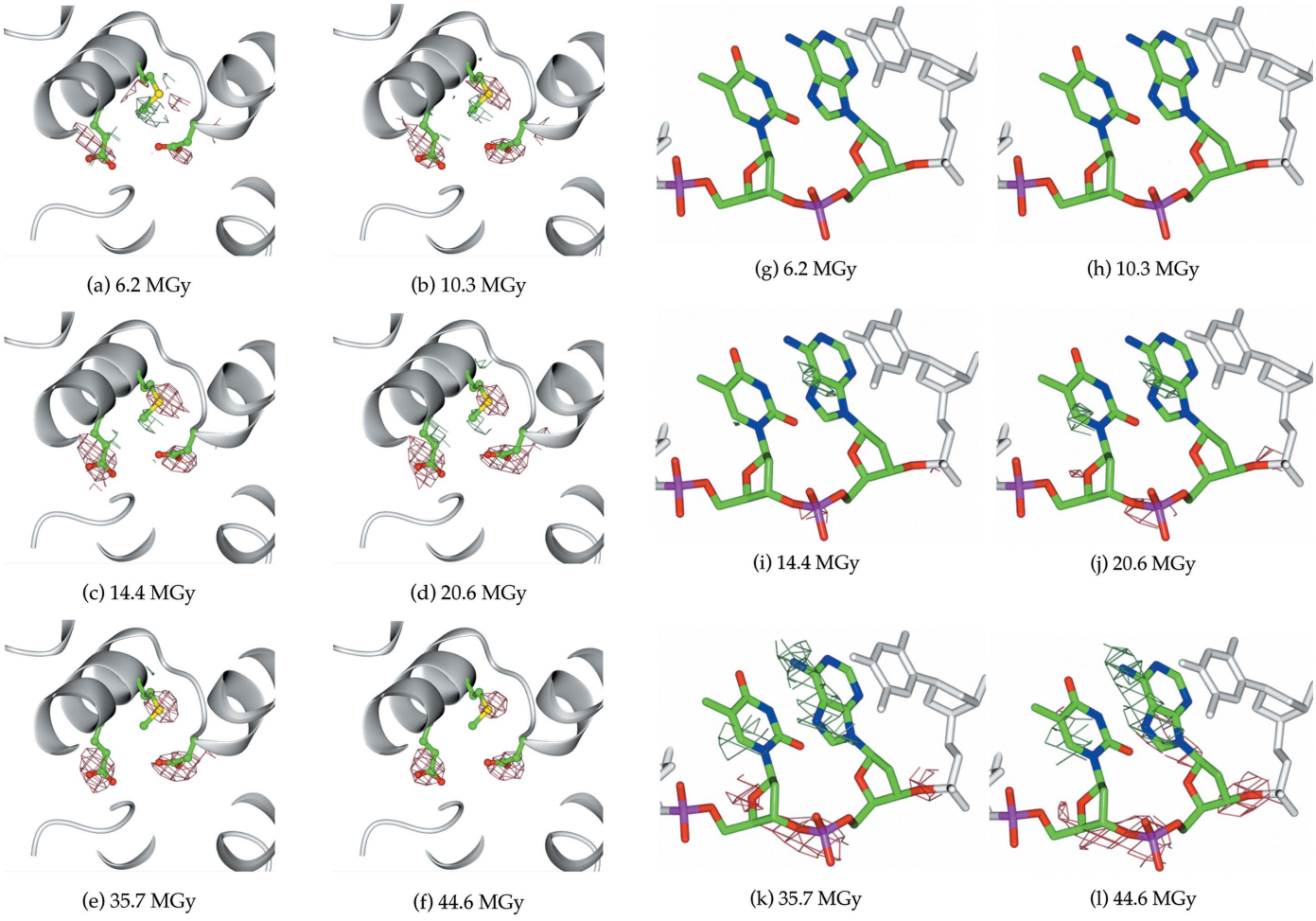


Figure 5
Protein and DNA damage sites in C.Esp1396I. (*a*)–(*f*) Visual representation of specific damage within protein chain *D* to Glu54, Met57 and Asp64 (green, left to right), displaying Fourier difference maps $F_{o,n} - F_{o,1}$, $n = 2, \ldots, 7$, over six increasing doses. Fourier maps are contoured at $\pm 3.0\sigma$ in green/red. (*g*)–(*l*) Visual representation of specific damage within DNA chain *F* (with 5′ to 3′ end from left to right in each image) to T24 and A25 (green, left to right), displaying Fourier difference maps $F_{o,n} - F_{o,1}$, $n = 5, \ldots, 8$, over six increasing doses. All Fourier maps are contoured at $\pm 3.0\sigma$ in green/red.

original 8000 observed. Without filtering, the subsequent quantitative analysis of site-specific damage would not have been achievable over the large dose range considered here. A selection of results from this analysis is illustrated as difference maps over two regions of the complex: amino acid residues remote from the DNA binding interface on helices 4 and 5, and two nucleotides at the highly compressed TATA site between palindromic recognition sequences on the DNA. Specific damage was observed to develop throughout the C.Esp1396I complex with respect to radiation dose for both protein [Figs. 5(*a*)–5(*f*)] and DNA [Figs. 5(*g*)–5(*l*)].

In terms of protein, Figs. 5(*a*)–5(*f*) reveal the specific damage dynamics of Glu54, Met57 and Asp64 in chain *D* (left to right) with respect to absorbed dose. Clear loss of electron density localized around the carboxyl groups, due to decarboxylation, is shown for Asp64 and Glu54, and the rate of carboxyl electron density loss with dose appears marginally greater for Glu54 than for Asp64. Figs. 5(*a*)–5(*f*) also show specific damage to Met57, where electron density loss and side chain disorder can be observed localized on the methylthio group over increasing doses. Although generally fewer difference peaks are observed on the DNA, there are several locations of specific damage on it too [Figs. 5(*g*)–5(*l*)]. The possible sugar-phosphate C—O bond cleavage between the T24 and A25 nucleotides of DNA chain *F* would generate a single-strand break with significant biological consequences. Additionally, at higher doses [Figs. 5(*k*) and 5(*l*)] positive electron density build-up is observed in close proximity to the T24 and A25 bases.

3.3. Chemical and topological distribution of specific damage

These data allow us to investigate the location, frequency and severity of specific damage sites on a range of scales from individual chains down to residues, nucleotides and specific atoms. Fig. 6 details the distribution of detected specific damage throughout the overall C.Esp1396I complex components for each absorbed dose. For each residue type, the damage frequency is heavily dependent on the overall number of that residue present within the four protein monomers.

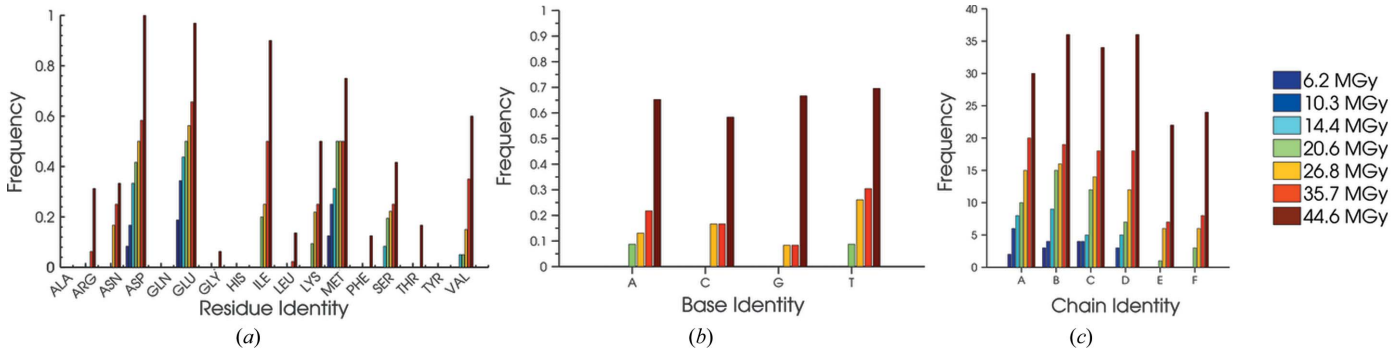
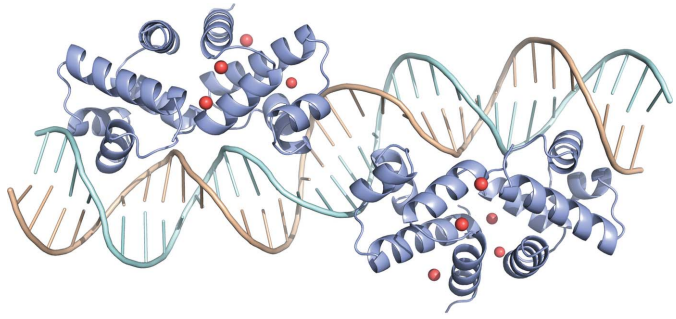


Figure 6
(*a*) Normalized frequency of detected specific damage against protein residue type (normalized to the frequency of occurrence of that residue in the structure) for each dose (MGy). Protein chains *A* to *D* are treated together. (*b*) Normalized frequency of detected specific damage for each DNA base type (normalized to the frequency of occurrence of that base in the structure), for each dose (MGy). DNA chains *A* to *D* are treated together. (*c*) Detected specific damage frequency for each C.Esp1396I chain (protein: *A*–*D*, DNA: *E*–*F*) for each dose.

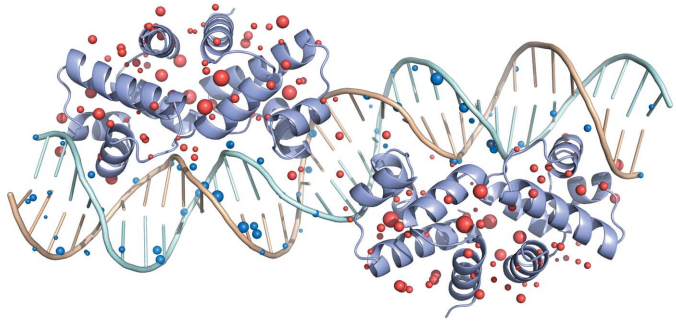
Thus the values have been normalized by the frequency of occurrence of that particular residue in the structure (Fig. 6*a*). Residues such as Asp, Glu, Met and Ser can be seen to accumulate significant damage (loss of electron density around the Ser side chain –OH) even at the lowest doses. At higher doses, damage is observed on Arg and Asn (electron density loss/disorder to the Arg/Asn main chain carboxyl group associated oxygen), Ile (partial loss of density around the side chain Cδ) and Lys (potential damage or disorder to the lysyl side chain), whilst the remaining amino acids in the protein have minimal specific damage even at very high dose. Note that C.Esp1396I has no Cys, Pro or Trp residues. There is significant heterogeneity in the rate of damage accumulation with increasing dose for each residue type. For example, by observing the step increases in damage frequency with increasing dose for each residue type, there is a clear difference in the peak detection rate between glutamate and aspartate decarboxylation for doses in the range 6.2 MGy to 35.7 MGy.

The distribution in DNA specific damage between the four nucleotide types is also shown and it is immediately apparent upon comparison with the protein data [Figs. 6(*a*) and 6(*b*)] that the *specific damage onset* is at significantly higher doses for DNA than for protein within the complex, and that specific damage is more evenly distributed between the four base types than amongst the protein residues.

Overall, there is a clear differential distribution in the dose-dependent intensities of specific damage between different protein residue types and also between the DNA and the protein. A comparison of the four protein chains *A*–*D* with the DNA chains *E* and *F* shows initial damage accumulation at >6.2 MGy *versus* >20.6 MGy, respectively (Fig. 6*c*). In addition to the later onset of specific damage to the DNA chains, the frequency of detected damage peaks in them is generally lower. This observation is made more striking when these data are compiled as a visual representation of the locations of specific damage sites (above the *specific damage onset*) across the protein–DNA complex with respect to accumulating dose (Fig. 7). It is seen that, with increasing dose, (*a*) the damage site frequency increases, and (*b*) the average electron density loss magnitude increases, as expected. Furthermore, it is clearly apparent that lower-dose (Fig. 7*a* and supporting Fig. S2) damage sites are localized on the protein (predominantly chains *B* and *C*), and that even at the highest dose (Fig. 7*b*), when damage sites are more homogeneously distributed throughout the protein monomers, there are still significantly fewer damage sites detected within the dsDNA component, suggesting slower time-scale damage dynamics



(*a*) Dose: 6.2 MGy



(*b*) Dose: 44.6 MGy

Figure 7
Representation of specific damage distribution throughout the C.Esp1396I complex for structures derived from the (*a*) first and (*b*) last dataset. Specific damage sites are represented as spheres, with radii proportional to electron density loss (electrons per Å^3). Spheres closer/further than 2 Å to/from the DNA strands are coloured blue/red. Similar representations of the structures derived from the five datasets suffering intermediate doses can be found in the supporting information (Fig. S2).

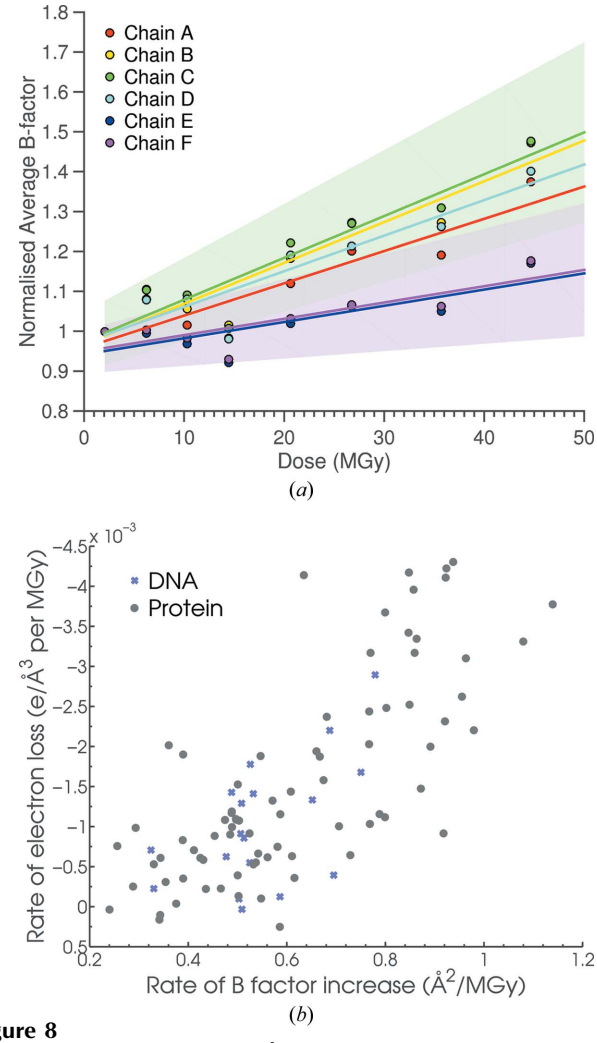


Figure 8
(*a*) Mean isotropic *B*-factor (Å^2) for each chain of the C.Esp1396I complex (*A* to *D* for protein monomers, *E* and *F* for DNA single strands) against dose (MGy), normalized to the *B*-factor of the first dataset for each chain, respectively. Data points are linearly fitted for each chain. Green and purple shaded areas denote 90% confidence intervals on the intercepts and slopes for chains *C* and *F*, respectively (other chain confidence intervals are not shown in the interests of clarity). (*b*) Correlation of the average rate of isotropic *B*-factor increase (Å^2 MGy^{-1}) with the average rate of electron density loss (electrons Å^{-3} MGy^{-1}). Grey/blue scatter points indicate protein/DNA specific damage sites, respectively.

and the existence of different damage mechanisms for the DNA.

3.4. Mean isotropic *B*-factor analysis

Fig. 8 shows an analysis of the 'normalized' average *B*-factor for each chain against dose, where for each individual protein–DNA chain the 'normalized' *B*-factor is defined as the *B*-factor at a given dose divided by the *B*-factor at the lowest dose investigated; consequently the average *B*-factor for each chain at the lowest dose plotted is set to 1. Linear fitting for each nucleoprotein chain gave a larger rate of normalized average *B*-factor increase for each of the protein monomers than for the DNA strands (0.0081 and 0.0102 Å^2 MGy^{-1} for protein chains *A* and *B*, respectively, and 0.0041 Å^2 MGy^{-1} for DNA strands *E* and *F*). Furthermore, close correspondence is observed in the average isotropic *B*-factor dose-dynamics between protein chains *A* and *D* (0.0081 and 0.0089 Å^2 MGy^{-1}, respectively), protein chains *B* and *C* (0.0102 and 0.0105 Å^2 MGy^{-1}, respectively) and DNA strands *E* and *F* (both 0.0041 Å^2 MGy^{-1}), indicating the rotational near-NCS around the DNA 35 bp sequence centre.

3.5. Specific damage dose-dynamics

Analysis of the electron loss per Å^3 with respect to accumulated dose was performed for different residue types. For each subplot shown in Fig. 9, the legends detail the nearest atom of the protein–DNA complex to which specific damage has been assigned.

It is evident that clear differential specific damage rates are present not only between different residue types but also within a given residue type. For example, in Fig. 9(*a*) there is variation in the electron density loss rate with dose for the Met57 methylthio group, with the specific damage onset at lower doses for chains *B* and *C* than for the corresponding residue in chains *A* and *D*. Furthermore the dynamics of electron density loss for Met57 are qualitatively similar for chains *B* and *C*, and also for chains *A* and *D* reflecting the near non-crystallographic symmetry within the complex.

Comparing Fig. 9(*d*) with Figs. 9(*a*)–9(*c*) provides a quantitative example of the greater specific damage resistance of DNA than protein in the complex, since the specific damage onset for DNA base T is detected at significantly larger dose values ($\geq$20 MGy) than is the specific damage for the methionine, glutamate and aspartate residue case studies which are already damaged in the first difference maps (6.2 MGy).

Analysing the average gradient for each detected damage site for electron density loss against dose (Fig. 9) and also similarly produced *B*-factor change against dose plots (not shown), the expected correlation was found between the rate of *B*-factor increase and the rate of electron density loss with respect to dose (Fig. 8*b*). There appeared to be an underlying approximately linear relationship between the two metrics, reinforcing the fact that the *B*-factor increase rate with respect to absorbed dose would serve as a suitable substitute measure to monitor the heterogeneity in specific damage dynamics throughout the protein–DNA complex.

In Fig. 8(*b*) both the DNA and protein damage sites appear to follow an approximately linear trend between the average rate of *B*-factor increase and the average rate of electron density loss. However, most DNA damage sites are observed to have both low average *B*-factor increase and electron density loss rates, providing additional evidence of differential specific site dose-dynamics between the protein monomers and DNA strands.

A comparison of the heterogeneous spatial distribution in both protein residue average *B*-factor rates and protein electron density loss rates with the far more homogeneous clustering of DNA damage sites shown in Fig. 8(*b*) emphasizes that the dose-dependent dynamics of specific damage is

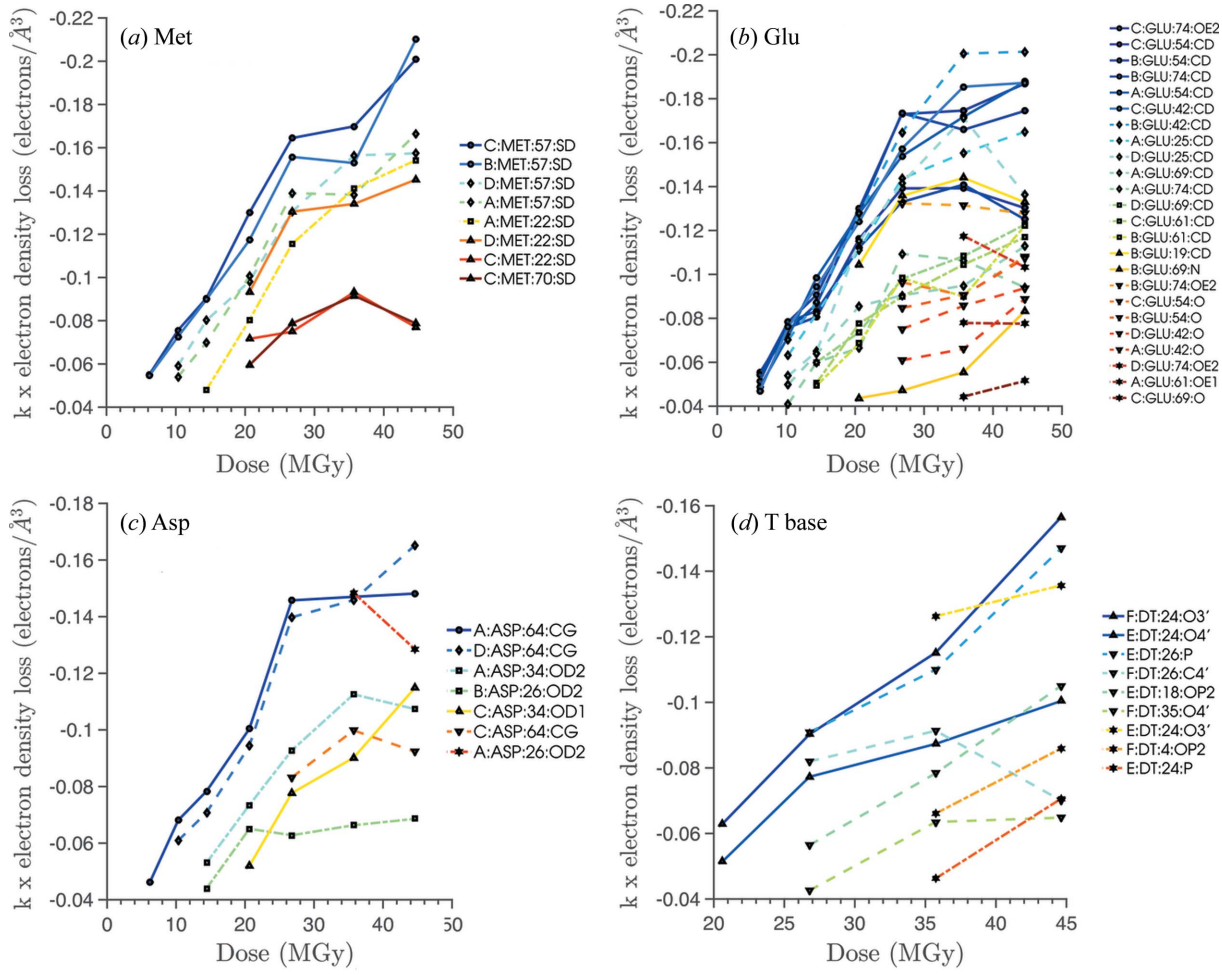


Figure 9
Electron density loss per Å^3 (scaled by constant factor k; see §2.4) against accumulated dose (MGy) for (*a*) methylthio group damage for methionine residues, decarboxylation of (*b*) glutamate and (*c*) aspartate groups, (*d*) DNA base T damage including base-sugar N_1—C bond and sugar-phosphate C—O bond damage.

far more uniform for DNA than for protein residues in C.Esp1396I.

4. Discussion

By using an innovative highly streamlined and automated pipeline for the identification of X-ray induced structural damage patterns, we have established the existence of differential specific damage rates between the protein and DNA components of a model complex C.Esp1396I with respect to dose at 100 K. Whereas other work has studied specific protein and DNA damage in isolation (Spotheim-Maurizot & Davídková, 2011; McGeehan *et al.*, 2007; Simons, 2006; Weik *et al.*, 2000; Ravelli & McSweeney, 2000; Burmeister, 2000; Cadet *et al.*, 1999), this work investigated a large dataset of specific damage sites within a protein–DNA complex in order to produce statistically significant observations on specific damage dynamics.

The modes of action for specific damage to the protein components follow similar patterns to those documented in other studies such as decarboxylation of acidic residues and localized disruption of sulfur-containing residues, the chemistries of some of which are relatively well understood (Burmeister, 2000). We note, however, that many of the mechanisms referenced in that work were deduced from experiments carried out under quite different circumstances from those used here. For example, most radiation chemical investigations have been pursued in dilute aqueous solution. In this environment the ionizing radiation is primarily deposited in the solvent, namely water, and the holes, H_2O^+, rapidly deprotonate to form hydroxyl radicals, while the released electrons cause many further excitations and ionizations in the surrounding medium in the course of thermalization and solvation (Spinks & Woods, 1990). It is also important to recall that many of these aqueous radiation chemistry experiments were carried out at room temperature where both solvated electrons and hydroxyl radicals are diffusively mobile. At 100 K thermalizing electrons can still tunnel freely; however, the range explored by hydroxyl radicals may be much curtailed. The fate of ionization events directly impacting the protein would presumably be much less influenced by temperature.

An interesting example of the complexities involved might be found in the specific damage to methionine residues. In

dilute aqueous solution, hydroxyl radical attack can proceed *via* $^{\bullet}$OH addition, loss of hydroxide (coupled to a water proton), forming a radical cation which subsequently deprotonates at an adjacent carbon to give a relatively persistent neutral carbon-centred radical. One possible decay channel for this species involves so-called β-cleavage, eliminating the terminal methyl, which is driven by the formation of a stabilizing CS double bond (Wisniowski *et al.*, 2002). If the medium is acidic, atomic hydrogen radicals (H$^{\bullet}$), formed by the rapid neutralization reaction of the aqueous electron with ambient protons, will also be active and can, for example, also displace the terminal methyl group in an S_H2 process (Wisniowski *et al.*, 2004). It might be expected that deprotonation at a neighbouring carbon and subsequent β-cleavage would also follow a direct hit on the methionine sulfur either from the incoming X-ray or from a released photoelectron from a nearby site.

Note that the most radiation damage susceptible linkages in proteins, namely disulfide bonds, are not present in the particular macromolecule studied here, and indeed not often present in DNA binding proteins found in the reductive intracellular environment.

In the present study, specific damage was also observed in the nucleic acid component, including evidence of a potential single-strand break (SSB) in the DNA. The location of this SSB is interesting as it correlates with a region of the DNA that is both AT-rich and under significant strain as a consequence of large-scale deformation due to protein binding. It is possible that such strained geometries enhance the radiation damage effects in DNA. This observation merits further investigation since it would have major biological consequences, as the wrapping of eukaryotic DNA around histones to form nucleosomes in part relies on the distortion of DNA around AT-rich sites such as this. The DNA radiation damage observed here could well result from previously suggested mechanisms such as low-energy secondary electron attachment to bases [a suggested precursor to SSBs (Simons, 2006)] or base modification by bonding close proximity solvent free radicals (for example, hydroxyl radical binding to carbon 6 in thymine 24) (Cadet *et al.*, 1999). However, once again it is worth noting the limited mobility of hydroxyl radicals at low temperatures, suggesting that only those formed in the immediate vicinity of the base could participate. Electrons, on the other hand, will maintain considerable mobility at 100 K.

It is again important to note that the low-energy electron damage to DNA components reported by Sanche and coworkers (Boudaïffa *et al.*, 2000; Huels *et al.*, 2003) was initially observed from DNA and components isolated on surfaces under ultrahigh-vacuum conditions. While mechanisms acting in the present protein/DNA crystalline environment might not be identical to those postulated in that investigation, recent work on low-energy electron DNA interactions is moving closer to much more relevant conditions (Alizadeh & Sanche, 2014).

We have also demonstrated that the normalized mean *B*-factor change of a particular chain gives a measure of the real-space averaged disorder present per protein–DNA chain between copies of the C.Esp1396I complex in different unit cells of the crystal. This metric indicates that, on average, the protein components become relatively more disordered with dose than does the DNA, suggesting that the protein is damaged by X-ray radiation at a faster rate than is DNA. Since any global radiation damage effects present would affect the protein and DNA components to the same extent, this could provide a suitable measure to compare specific damage susceptibility between protein and DNA. Our methodology is further validated by the identification of similar damage at palindromically equivalent sites in both protein and DNA components for this particular complex.

Further studies on a range of protein–DNA and protein–RNA complexes would allow these metrics to be tested rigorously and would reveal if our observations provide some general rules for X-ray radiation damage to biological nucleoprotein complexes. To aid this goal the custom-made scripts which allow efficient specific damage searching and enable consistent noise peak filtering could be utilized with Fourier difference maps generated from other systems. These investigations are only tractable through the use of a robust semi-automated pipeline such as the one developed here.

We can speculate that the molecule that holds our genetic blueprint has evolved to be more radiation-resistant than other cellular components, and this may not be so surprising since sacrificial proteins can be more easily replaced than lost genes. Studies on a wide range of nucleoprotein complexes utilizing these partially automated methods should provide further insight into these intriguing observations.

We thank Drs Simon Streeter and Sarah Thresh at the University of Portsmouth for the supply of purified protein and DNA samples, Dr Neil Ball for help in X-ray data collection and the Biotechnology and Biological Research Council (BBSRC) for a research grant to GK (BB/E000878/1). We would like to thank staff at the EMBL for the provision of crystallization facilities and also beamline staff at the ESRF for their advice and assistance during X-ray data collection. We gratefully acknowledge the Engineering and Physical Sciences Research Council (EPSRC) in the UK for funding through a studentship in the Systems Biology programme of the University of Oxford Doctoral Training Centre (CB). HMG is funded through the Wellcome Trust graduate program in Structural Biology. IC is supported by the US Department of Energy Office of Science, Office of Basic Energy Sciences under Award Number DE- FC02-04ER1553.

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.

Alizadeh, E. & Sanche, L. (2014). *Eur. Phys. J. D*, **68**, 97.

Alizadeh, E., Sanz, A. G., Madugundu, G. S., García, G., Wagner, J. R. & Sanche, L. (2014). *Radiat. Res.* **181**, 629–640.

Allan, E. G., Kander, M. C., Carmichael, I. & Garman, E. F. (2013). *J. Synchrotron Rad.* **20**, 23–36.

Ball, N., Streeter, S. D., Kneale, G. G. & McGeehan, J. E. (2009). *Acta Cryst.* D**65**, 900–905.
Barrios, R., Skurski, P. & Simons, J. (2002). *J. Phys. Chem. B*, **106**, 7991–7994.
Begusová, M., Eon, S., Sy, D., Culard, F., Charlier, M. & Spotheim-Maurizot, M. (2001). *Int. J. Radiat. Biol.* **77**, 645–654.
Berdys, J., Anusiewicz, I., Skurski, P. & Simons, J. (2004). *J. Am. Chem. Soc.* **126**, 6441–6447.
Boudaïffa, B., Cloutier, P., Hunting, D., Huels, M. A. & Sanche, L. (2000). *Science*, **287**, 1658–1660.
Burmeister, W. P. (2000). *Acta Cryst.* D**56**, 328–341.
Cadet, J., Delatour, T., Douki, T., Gasparutto, D., Pouget, J. P., Ravanat, J. L. & Sauvaigo, S. (1999). *Mutat. Res.* **424**, 9–21.
Charlier, M., Eon, S., Sèche, E., Bouffard, S., Culard, F. & Spotheim-Maurizot, M. (2002). *Biophys. J.* **82**, 2373–2382.
Chen, V. B., Arendall, W. B. III, 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.
Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. (2010). *Acta Cryst.* D**66**, 486–501.
Engh, R. A. & Huber, R. (1991). *Acta Cryst.* A**47**, 392–400.
Eon, S., Culard, F., Sy, D., Charlier, M. & Spotheim-Maurizot, M. (2001). *Radiat. Res.* **156**, 110–117.
Fioravanti, E., Vellieux, F. M. D., Amara, P., Madern, D. & Weik, M. (2007). *J. Synchrotron Rad.* **14**, 84–91.
Flot, D., Mairs, T., Giraud, T., Guijarro, M., Lesourd, M., Rey, V., van Brussel, D., Morawe, C., Borel, C., Hignette, O., Chavanne, J., Nurizzo, D., McSweeney, S. & Mitchell, E. (2010). *J. Synchrotron Rad.* **17**, 107–118.
Garman, E. F. (2010). *Acta Cryst.* D**66**, 339–351.
Garman, E. F. & Owen, R. L. (2006). *Acta Cryst.* D**62**, 32–47.
Gillard, N., Begusova, M., Castaing, B. & Spotheim-Maurizot, M. (2004). *Radiat. Res.* **162**, 566–571.
Gillard, N., Goffinont, S., Buré, C., Davidkova, M., Maurizot, J.-C., Cadene, M. & Spotheim-Maurizot, M. (2007). *Biochem. J.* **403**, 463–472.
Huels, M. A., Boudaïffa, B., Cloutier, P., Hunting, D. & Sanche, L. (2003). *J. Am. Chem. Soc.* **125**, 4467–4477.
Lang, P. T., Holton, J. M., Fraser, J. S. & Alber, T. (2014). *Proc. Natl Acad. Sci.* **111**, 237–242.
Leslie, A. G. W. & Powell, H. R. (2007). *Evolving Methods for Macromolecular Crystallography*, edited by R. J. Read & J. L. Sussman, pp. 41–51. Dordrecht: Springer.
McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C. & Read, R. J. (2007). *J. Appl. Cryst.* **40**, 658–674.
McGeehan, J. E., Carpentier, P., Royant, A., Bourgeois, D. & Ravelli, R. B. G. (2007). *J. Synchrotron Rad.* **14**, 99–108.
McGeehan, J. E., Streeter, S. D., Thresh, S. J., Ball, N., Ravelli, R. B. & Kneale, G. G. (2008). *Nucleic Acids Res.* **36**, 4778–4787.
McNicholas, S., Potterton, E., Wilson, K. S. & Noble, M. E. M. (2011). *Acta Cryst.* D**67**, 386–394.
Michael, B. D. & O'Neill, P. (2000). *Science*, **287**, 1603–1604.
Murray, J. & Garman, E. (2002). *J. Synchrotron Rad.* **9**, 347–354.
Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* D**53**, 240–255.
Nave, C. & Hill, M. A. (2005). *J. Synchrotron Rad.* **12**, 299–303.
O'Neill, P., Stevens, D. L. & Garman, E. (2002). *J. Synchrotron Rad.* **9**, 329–332.
Owen, R. L., Holton, J. M., Schulze-Briese, C. & Garman, E. F. (2009). *J. Synchrotron Rad.* **16**, 143–151.
Owen, R. L., Rudiño-Piñera, E. & Garman, E. F. (2006). *Proc. Natl Acad. Sci. USA*, **103**, 4912–4917.
Ptasińska, S. & Sanche, L. (2007). *Phys. Rev. E*, **75**, 031915.
Ramachandran, G. N. & Sasisekharan, V. (1968). *Adv. Protein Chem.* **23**, 283–438.
Ravelli, R. B. & Garman, E. F. (2006). *Curr. Opin. Struct. Biol.* **16**, 624–629.
Ravelli, R. B. & McSweeney, S. M. (2000). *Structure*, **8**, 315–328.
Sanishvili, R., Yoder, D. W., Pothineni, S. B., Rosenbaum, G., Xu, S., Vogt, S., Stepanov, S., Makarov, O. A., Corcoran, S., Benn, R., Nagarajan, V., Smith, J. L. & Fischetti, R. F. (2011). *Proc. Natl. Acad. Sci. USA*, **108**, 6127–6132.
Simons, J. (2006). *Acc. Chem. Res.* **39**, 772–779.
Spinks, J. W. T. & Woods, R. J. (1990). *An Introduction to Radiation Chemistry*, 3rd ed. New York/Toronto: John Wiley and Sons.
Spotheim-Maurizot, M. & Davídková, M. (2011). *Mutat. Res. Fundam. Mol. Mech. Mutagen.* **711**, 41–48.
Stisova, V., Goffinont, S., Spotheim-Maurizot, M. & Davidkova, M. (2006). *Radiat. Prot. Dosim.* **122**, 106–109.
Ten Eyck, L. F. (1973). *Acta Cryst.* A**29**, 183–191.
Théodore, M., Sobczyk, M. & Simons, J. (2006). *Chem. Phys.* **329**, 139–147.
Weik, M., Ravelli, R. B. G., Kryger, G., McSweeney, S., Raves, M. L., Harel, M., Gros, P., Silman, I., Kroon, J. & Sussman, J. L. (2000). *Proc. Natl Acad. Sci.* **97**, 623–628.
Winn, M. D., Ballard, C. C., Cowtan, K. D., Dodson, E. J., Emsley, P., Evans, P. R., Keegan, R. M., Krissinel, E. B., Leslie, A. G. W., McCoy, A., McNicholas, S. J., Murshudov, G. N., Pannu, N. S., Potterton, E. A., Powell, H. R., Read, R. J., Vagin, A. & Wilson, K. S. (2011). *Acta Cryst.* D**67**, 235–242.
Wisniowski, P., Bobrowski, K., Carmichael, I. & Hug, G. L. (2004). *J. Am. Chem. Soc.* **126**, 14468–14474.
Wisniowski, P., Carmichael, I., Fessenden, R. W. & Hug, G. L. (2002). *J. Phys. Chem. A*, **106**, 4573–4580.
Zeldin, O. B., Brockhauser, S., Bremridge, J., Holton, J. & Garman, E. F. (2013*a*). *Proc. Natl. Acad. Sci. USA*, **110**, 20551–20556.
Zeldin, O. B., Gerstel, M. & Garman, E. F. (2013*b*). *J. Appl. Cryst.* **46**, 1225–1230.

nature COMMUNICATIONS

ARTICLE

Received 5 Nov 2014 | Accepted 28 Jan 2015 | Published 9 Mar 2015

DOI: 10.1038/ncomms7435 OPEN

Structure of CPV17 polyhedrin determined by the improved analysis of serial femtosecond crystallographic data

Helen M. Ginn[1], Marc Messerschmidt[2,†], Xiaoyun Ji[1,3], Hanwen Zhang[1], Danny Axford[4], Richard J. Gildea[4], Graeme Winter[4], Aaron S. Brewster[5], Johan Hattne[5], Armin Wagner[4], Jonathan M. Grimes[1,4], Gwyndaf Evans[4], Nicholas K. Sauter[5], Geoff Sutton[1] & David I. Stuart[1,4]

The X-ray free-electron laser (XFEL) allows the analysis of small weakly diffracting protein crystals, but has required very many crystals to obtain good data. Here we use an XFEL to determine the room temperature atomic structure for the smallest cytoplasmic polyhedrosis virus polyhedra yet characterized, which we failed to solve at a synchrotron. These protein microcrystals, roughly a micron across, accrue within infected cells. We use a new physical model for XFEL diffraction, which better estimates the experimental signal, delivering a high-resolution XFEL structure (1.75 Å), using fewer crystals than previously required for this resolution. The crystal lattice and protein core are conserved compared with a polyhedrin with less than 10% sequence identity. We explain how the conserved biological phenotype, the crystal lattice, is maintained in the face of extreme environmental challenge and massive evolutionary divergence. Our improved methods should open up more challenging biological samples to XFEL analysis.

[1] Division of Structural Biology, The Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford, Oxfordshire OX3 7BN, UK. [2] SLAC National Accelerator Laboratory, 2575 Sand Hill Road, Menlo Park, California 94025, USA. [3] Molecular Biophysics and Biochemistry, Yale School of Medicine, 333 Cedar Street, New Haven, Connecticut 06510, USA. [4] Diamond House, Diamond Light Source, Harwell Science & Innovation Campus, Didcot, Oxfordshire OX11 0DE, UK. [5] Physical Biosciences Division, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, California 94720, USA. † Present address: National Science Foundation BioXFEL Science and Technology Center, 700 Ellicott Street, Buffalo, New York 14203, USA. Correspondence and requests for materials should be addressed to D.I.S. (email: dave@strubi.ox.ac.uk).

© 2015 Macmillan Publishers Limited. All rights reserved.

Cytoplasmic polyhedrosis viruses (CPVs) parasitize many insect species, are registered microbial insecticides[1] and remain a problem for the silk industry[2]. CPV are protected by a robust case (polyhedron), each a tiny crystal of viral polyhedrin, whose volume varies from 10^{-3} to $10^{3}\,\mu m^{3}$ depending on the CPV type. Fossilized polyhedra suggest that CPV have changed little over the last 100 million years[3], however, as expected for such ancient viral proteins, polyhedrins vary enormously in amino-acid sequence between different types of CPV. Larger polyhedra have been analysed at microfocus synchrotron beamlines[4–6], but the smallest have proved refractory to such analysis, as discussed below.

Here, we measure diffraction from nanocrystals of the smallest CPV polyhedra yet characterized, using serial femtosecond crystallography (SFX) at an X-ray free-electron laser (XFEL)[7]. The very short pulse length of the XFEL beam allows us to collect useful high-resolution diffraction data at room temperature before the crystals are destroyed by the extreme peak power densities. XFEL structural determinations have been performed before, but have required very large numbers of crystals to obtain reliable high-resolution data analysis[8–11]. These analyses have used a Monte Carlo method whereby data from the individual crystals are simply averaged together. We apply more sophisticated methods of SFX data analysis and show that the quality of the measurements can be improved, allowing us to obtain one of the highest-resolution structures published to date (1.75 Å) from an XFEL, using 5,787 crystals of CPV type 17 (CPV17), fewer than previously required for high-resolution analysis[8–11]. We are able to determine the structure by molecular replacement, which in our hands fails on a lower-resolution 100 K data set of CPV17 collected at a synchrotron. However, given the XFEL analysis we are able to solve the 100 K structure and we find, unexpectedly, that the crystals remain isomorphous between 100 and 293 K, despite a ~1% alteration in the cell length, although low temperature ablates covalent disulphide bonds, which brace the room temperature lattice. The structural core of the molecule is similar to that observed for a related polyhedrin, despite less than 10% sequence identity. However, we find that, in addition, there are a number of smaller more variable non-core structural modules. The polyhedrin building blocks and their arrangement in these crystals remain similar to those observed before, but sequence variation in the structural modules allows plasticity and evolutionary robustness. Our methods provide a new physical model for XFEL diffraction data, which allows a more accurate estimation of the actual signal. The need for such improved methods is indicated by the recent papers from White[12] and Kabsch[13]. White devise a model for XFEL diffraction data, which has similarities with the one we describe below, but only describe its application to simulated data. Kabsch defines a term 'Ewald offset correction', which is similar to our 'partiality', and applies it successfully to synchrotron still images. We here demonstrate the power of such a method tailored to and applied to real XFEL data, revealing the promise of such improved models to allow XFELs to tackle more challenging biological samples.

Results

XFEL data collection and analysis. CPV17 polyhedrin nanocrystals were grown in *Spodoptera frugiperda* (Sf9) cells, purified and diffraction data were collected at the Coherent X-ray Imaging (CXI) beamline, Linac Coherent Light Source (LCLS). A suspension of crystals was injected into the beam *in vacuo*[14], presumably at close to room temperature, and femtosecond diffraction snapshots were recorded at 120 frames per second. Initial processing used cctbx.xfel[15,16] to identify useful images and provide preliminary crystal orientation matrices[15]. Within 20 min 5,787 indexable images had been collected at ~1.46 Å wavelength that were used for further analysis.

Accurate modelling of the diffraction underpins the huge success of protein crystallography, using sophisticated models developed over many years[17,18]. However, fundamental differences in SFX diffraction experiments prevent the direct application of these methods. Most XFEL analyses have used the so-called 'Monte Carlo' method[19], where potentially observable reflections are summed and the orientation matrix is modelled imprecisely, requiring the use of massive numbers of XFEL images to obtain accurate measurements because the integratable spots are not accurately predicted. We have developed a refinement method that significantly improves the crystal orientations, and a model for partiality prediction that takes account of crystal size, orientation, mosaicity and the spectral properties of the self-amplified spontaneous emission (SASE) X-ray pulse[16,20]. This allows a quantification of the relative fraction of the beam energy captured by each reflection (termed partiality). Here we present a first implementation of these models to the analysis of the CPV17 data. Recent papers by White[12] and Kabsch[13] developed related models, but did not apply those models to the analysis of experimentally obtained XFEL data.

As an essential first step, we refined the orientation matrices of all crystals by an iterative process. Diffraction was modelled by describing the X-ray beam using an inflated energy bandwidth and infinitely small reciprocal lattice points (rlps), providing a 'net' to catch all possible reflections. These potential reflections were integrated, and strong reflections deemed 'hits'. Refinement initially aimed to sharpen the histogram of frequency of hits *vs* wavelength by slightly adjusting the crystal orientation around the two axes orthogonal to the beam. This method worked well, a typical distribution before and after refinement is shown in Fig. 1a. The average combined angular correction was 0.075°. This refinement also optimized the estimate of the mean wavelength for each pulse. In contrast, the Kabsch paper reorients the crystal to minimize the rlps' distance from a nominally monochromatic Ewald sphere. An experimental model was also invoked, accounting for spot size (derived from crystal size), X-ray bandwidth and crystal mosaicity, each refined to best predict the observed selection of hits. For integration, spot positions were centred on the local maximal peak to allow for imperfect metrology. Images were metrology-corrected by cctbx.xfel[16] and integrated using DIALS[21]. The bandwidth of the pulse was derived from the hit count histogram (Methods). The spot diameter refined to 1.8×10^{-4} Å^{-1}, corresponding to a crystal size of ~0.5 µm, plausible based on electron microscopy analysis of the crystals. A mosaicity of 0.03° was used, which is consistent with synchrotron data but may also be an overestimate. Spots were generally well predicted (Supplementary Fig. 1). The limiting resolution for each image was chosen as the highest-resolution shell with $I/\sigma(I)$ of 2.0, which almost always integrated to the edges of the image.

Partiality was calculated for each reflection. This corresponds to the fraction of the total energy of the X-ray pulse contributing to the diffraction compared with the maximum that could have been used (see Methods). This model differs from that used for the rotation method[22], does not require further Lorentz correction and a polarization correction was not applied. For all further statistical analysis and calculations, reflection intensities and weights were divided by their partiality, and ~113 reflections per image with partiality of 0.3 or below were discarded since they were poorly estimated (the large number of reflections discarded reflects the previous overprediction of reflections), leaving a total of 1,093,434 reflections. Statistics are shown in Table 1. The effectiveness of partiality modelling is shown in Supplementary Fig. 2. Without any partiality modelling, the $CC_{1/2}$ parameter[23] is

© 2015 Macmillan Publishers Limited. All rights reserved.

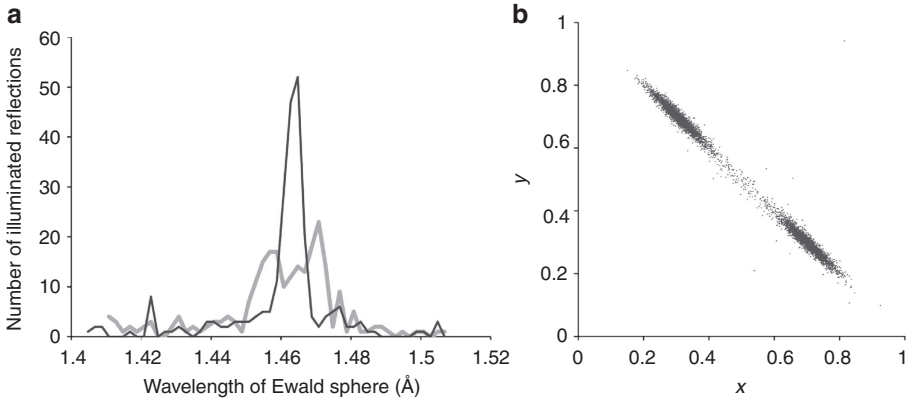


Figure 1 | XFEL data analysis. (**a**) Distribution of Ewald sphere wavelengths for all reflections on a single image, before (grey, thick) and after (black, thin) orientation matrix refinement. Refinement resulted in a 0.117° rotational shift in the orientation model. (**b**) Cluster algorithm by a modified version of algorithm of Brehm and Diederichs[24], showing positions of final artificially defined vectors corresponding to individual images (see Methods). Note the vectors fall into two sets, clearly separating the two indexing choices.

Table 1 | XFEL data collection and refinement statistics.

	CPV17
Data collection	
Space group	I23
Cell dimensions	
a, *b*, *c* (Å)	106.1, 106.1, 106.1
α, *β*, *γ* (°)	90, 90, 90
Resolution (Å)	28.30–1.75(1.79–1.75)*
R_{pim} (%)	7.4 (33.8)
R_{split} (%)	11.8 (58.4)
$CC_{½}$	98.0 (38.6)
Completeness (%)	100 (100)
Redundancy	52.4 (11.1)
Refinement	
Resolution (Å)	28.36–1.75
No. of reflections	20,122
R_{work}/R_{free}	12.2/15.4
No. of atoms	
Protein	1,914
Ligand/ion	32
Water	174
B-factors	
Protein	23.7
Ligand/ion	34.4
Water	32.7
R.m.s. deviations	
Bond lengths (Å)	0.010
Bond angles (°)	1.39

CPV17, CPV type 17; XFEL, X-ray free-electron laser.
Number of crystals used: 5,554.
*Highest-resolution shell is shown in parenthesis.

97.2% across all resolution shells. Using a partiality cutoff of 0.3—but not dividing by partiality—increases $CC_{1/2}$ to 97.4%, while dividing accepted reflections by their partiality increases it to 98.3%.

The crystals belong to space group I23 and are therefore indexed ambiguously—the indexing schemes (h, k, l) and (k, h, −l) are geometrically indistinguishable. We used the method of Brehm and Diederichs[24] to break this ambiguity while including the correlation between the alternative indexing choice, thereby doubling the number of pairwise comparisons (see Methods). Figure 1b shows the clear separation of crystals into the two indexing sets that were transformed to a consistent indexing set and simply merged with crude scale factors applied to give an interim reference data set. Scale factor refinement for each image was then made against this interim reference data set and the weakest images discarded before merging. The $CC_{1/2}$ dropped below 0.3 at a resolution of 1.75 Å, taken as the nominal resolution (Supplementary Fig. 3). The R_{split} value was calculated between the two indexing halves of the data set. Over this resolution range the average R_{split} was 11.7% using 5,554 images. The effect of set size is shown in Supplementary Fig. 3.

Synchrotron cryo-crystallography. In parallel to the XFEL data collection, crystals were analysed at 100 K on a microfocus synchrotron beamline (I24, Diamond Light Source). Data from 768 crystals yielded a moderate resolution data set with a $CC_{1/2}$ of 0.3 at 2.15 Å resolution (Supplementary Fig. 3a)[25]. To accurately determine changes in unit cell dimensions between room temperature and 100 K, powder patterns were collected at Diamond Light Source, which showed a contraction of the unit cell dimension on cooling from 106.1 to 104.9 Å (values used as the cell dimensions for the XFEL and synchrotron analyses, respectively). We consider the unit cells to be highly uniform at each temperature since the powder rings are tightly defined to the highest recorded resolution. Despite this temperature-dependent variation in the crystal lattice, the correlation in intensities between the XFEL and synchrotron data was excellent (90.9% overall, Supplementary Fig. 3a). This surprising isomorphism is discussed below. In the absence of a good molecular replacement model, these synchrotron-source data were not, in our hands, sufficient to support structure determination.

Structure determination from XFEL data. The structure of CPV17 was solved by molecular replacement using the higher-resolution XFEL data. Manual rebuilding and refinement rapidly reduced R_{work}/R_{free} from 56.5%/54.1% to 12.2%/15.4% (Table 1). Beyond 2.0 Å resolution, at which point R_{split} exceeds 20%, the R_{work} and R_{free} increase significantly. However, the quality of the electron density meets that expected for a high-quality 1.75-Å resolution analysis (Supplementary Movie 1), for instance allowing the unambiguous identification of an unexpected bound adenosine triphosphate (ATP) (Fig. 2a). Using this room temperature structure the synchrotron 100 K structure was solved trivially and refined to 2.2 Å (Table 2).

CPV17 polyhedrin shows similarity to CPV1 polyhedrin. Two hundred and thirty-six of the 237 CPV17 polyhedrin residues

© 2015 Macmillan Publishers Limited. All rights reserved.

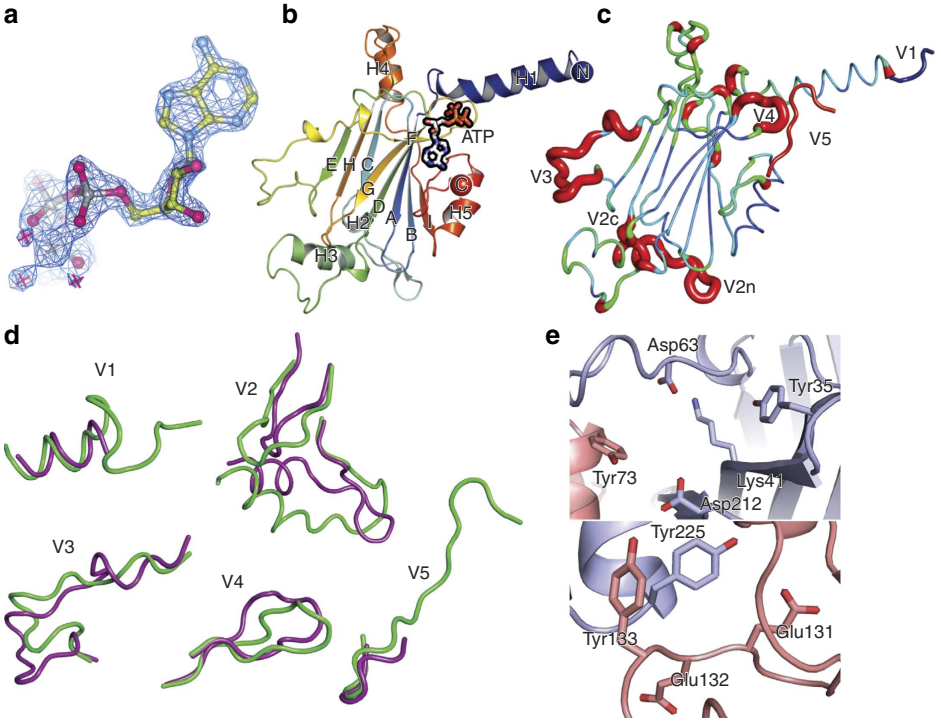


Figure 2 | Structure of CPV17 polyhedrin. (**a**) Electron density around the ATP moiety contoured at 1.3 σ. (**b**) Cartoon of the CPV17 polyhedrin subunit coloured from blue to red, N terminus to C terminus that are denoted by spheres. Secondary structure elements are labelled and the ATP molecule shown as sticks. (**c**) Comparison of the CPV17 and CPV1 polyhedrin structures. The molecules were aligned with program SHP[43]. Both colour and tube thickness represent r.m.s. distance (r.m.s.d.) of equivalent C-alpha atoms (thin, blue: r.m.s.d. $<$1.0 Å, green, thicker: 1.0–2.5 Å, orange, thickest: $>$2.5 Å). Unaligned regions are coloured red and displayed with exaggerated thickness. Variable regions are labelled v1–v5 with the N- and C-terminal extensions of CPV1 drawn. (**d**) Comparison of variable regions between CPV17 (magenta) and CPV1 (cyan). (**e**) The interface between monomers (coloured differently) in the crystal is rich in tyrosines.

were well defined in the XFEL electron density map; only the N-terminal residue could not be seen (Fig. 2b). In addition, an ATP molecule with bound Mg^{2+} was identified (Fig. 2a). Overall, the protein fold is remarkably similar to that of CPV1 polyhedrin with the two protein crystals sharing the same I23 space group[26]. 191 Cα atoms can be superposed with root mean square (r.m.s.) deviation 1.4 Å (Fig. 2c,d), despite less than 10% residues being identical in the sequence (Supplementary Fig. 4). In addition, the molecules are placed almost identically in the crystal lattice (0.6 Å displacement between the cores of the molecules). This is reflected in the similarity in the unit cell dimensions ($a = 104.9$ and 102.8 Å for CPV17 and CPV1, respectively). There are five small structural modules that show significant structural variation between the two molecules (V1–5, Fig. 2c,d). V1 is the N-terminal region, which in CPV1 is extended. V2 comprises some 20 residues forming a 'cap' at one end of the molecule, containing two portions of polypeptide (V2n and V2c, in Fig. 2c). V3 forms a protrusion on the flank of the molecule; V4 interacts with a bound ATP in both CPV1 and CPV17 (described in detail below), and V5 is the C-terminal region, also extended in CPV1. In the alkali conditions of the insect mid-gut (pH 10.5 or greater), the crystal dissolves to release the infectious virus[27]. It has been postulated that the pH susceptibility arises from the deprotonation of tyrosine residues[4]. CPV17 harbours several tyrosines in environments where they could contribute to pH-mediated crystal disassembly (Fig. 2e), and are partly conserved in CPV1 (ref. 6). An unexpected feature of the CPV17 crystal is an incompletely (50%) formed intermolecular disulphide bond (Fig. 3a,b), which assembles helical strings of covalently linked molecules, presumably contributing to the mechanical stability of the crystal (Fig. 3c). Splitting of disulphide bonds is accelerated in the high pH carbonate environment of the insect mid-gut[28], facilitating dissolution. Interestingly, this disulphide is not observed in the 100-K structure (Fig. 3d). To test whether this was due to radiation damage, we assembled two data sets, one derived from the first half of the images collected from each crystal and the other derived from the second half of the data from each crystal. The two electron density maps were essentially identical, indicating that in difference was not due to radiation damage. The B-factors of sulphur atoms in the room temperature structure suggest that XFEL data collection also inflicts little radiation damage. A possible explanation for this temperature-dependent difference in structure is presented below.

Nucleoside triphosphates stabilize the crystal lattice. The polyhedrins of CPV1 and CPV17 possesses a conserved structural core with modules of variable structure associated with it. One of these modules (V4) is involved in ATP association in different ways in CPV1 and CPV17. Crystals of CPV1 polyhedrin harbour various nucleoside triphosphates at monomer interfaces[4]. Unexpectedly CPV17 crystals contain only a single ATP molecule, positioned in a broadly similar position to an ATP found in CPV1. The ATP is well ordered in CPV17, with specific adenine–amine interactions ensuring ATP specificity (GTP is excluded by steric hindrance) (Fig. 4a), however, the triphosphate-Mg^{2+} moiety is the most tightly located, making a set of interactions formed primarily by a ring of backbone amide nitrogen atoms flanked by an arginine side chain (Fig. 4b). This configuration is similar to that locating the corresponding ATP in CPV1 polyhedra (Fig. 4c), although the phosphates are translated by some 6 Å, due to a shift in the position of the V4

© 2015 Macmillan Publishers Limited. All rights reserved.

Table 2 | Synchrotron data collection and refinement statistics.

	CPV17
Data collection	
Space group	I23
Cell dimensions	
a, *b*, *c* (Å)	104.9, 104.9, 104.9
α, β, γ (°)	90.0, 90.0, 90.0
Resolution (Å)	74.16–2.20 (2.26–2.20)*
R_{merge} (%)	66.5 (325.9)
$I/\sigma I$	6.4 (1.4)
Completeness (%)	99.9 (100.0)
Multiplicity	47.8 (25.4)
Refinement	
Resolution (Å)	74.16–2.20
No. of reflections	9,376
R_{work}/R_{free}	14.7%/19.9%
No. of atoms	
Protein	1,907
Ligand/ion	32
Water	146
B-factors	
Protein	22.5
Ligand/ion	50.0
Water	32.2
R.m.s deviations	
Bond lengths (Å)	0.013
Bond angles (°)	1.675

CPV17, CPV type 17.
Number of crystals used: 768.
*Highest-resolution shell is shown in parenthesis.

loop, and the ATP molecule is rotated by ~180°, so that the mechanism of nucleotide selection is quite different. However, despite the variation in such modules (which in CPV17 compensate for the absence of other nucleoside triphosphates found in CPV1), the fundamental assembly of the I23 lattice remains unchanged.

Temperature changes induce lattice swelling/shrinking. Comparison of the structures at room temperature and 100 K demonstrates the response of the crystals to extreme environmental changes. The structures are very similar (r.m.s.d. in Cα 0.26 Å after rigid body superposition). Nevertheless, there is a ~1% reduction in the cell dimensions on cooling. Previous studies[29] indicated that protein crystals accommodate such lattice changes by a number of switches in the intermolecular interactions within the crystal. To see whether this occurs in the CPV17 polyhedra, we artificially increased the unit cell for the 100 K data to match that of the XFEL data (thereby isotropically 'swelling' the structure of the monomer) and performed simple restrained refinement of the resulting structure. Remarkably, these structures were then almost indistinguishable (r.m.s.d. in Cα 0.17 Å). This suggests that the complete crystal has evolved as a robust material able to simply expand or contract slightly in a surprisingly homogenous way in the face of environmental insult, without repacking. This accords with the lattice being a key biological phenotype. Nevertheless, the strain in the (unphysiological) 100-K lattice is enough to sever the partially formed disulphide bond.

Discussion

In summary, orientation matrix refinement, careful selection of illuminated spots, partiality modelling and image scaling have markedly reduced the redundancy required to obtain a high-quality XFEL data set at high resolution, allowing perhaps the most challenging XFEL structure determination reported to date. How general are these results? While our nanocrystals possess high symmetry (point group 23) providing a high multiplicity of unique reflections, they also possess low mosaicity (~0.03°) reducing the number of reflections per image compared with that which might be expected for many protein crystals. We therefore believe that by adopting procedures such as those presented here, the number of XFEL images required for structure determination will be markedly reduced compared with that required previously[8,9]. This, together with improved sample preparation and presentation techniques[30], may render the method attractive for challenging problems, and should greatly facilitate experimental phasing[19]. In addition, less beam time will be required to collect sufficient data, in this case only 20 min. Finally, despite attenuating the beam so that only 2% of the possible X-ray photons were used, the 1.75 Å electron density map was of exceptional quality, suggesting that SFX has unexploited potential for high-resolution analyses.

Methods

Sample preparation. The polyhedrin gene for *Uranotaenia sapphirina* CPV17 was synthesized by GeneArt (Life Technologies) based on the Gene Bank sequence AY876384 (ref. 31). The polyhedrin gene was amplified and inserted into the transfer vector pBacPAK9 (Clontech). Recombinant baculovirus was produced by co-transfection of linearized baculovirus DNA and the transfer vector following a standard protocol[32]. Expression and purification of polyhedra followed the protocol described previously[26]. Cells were lysed, polyhedra pelleted at 500*g* and resuspended in water for storage.

Serial femtosecond crystallography. SFX experiments were carried out at the Linac Coherent Light Source of the SLAC National Accelerator Laboratory (Menlo Park, CA, USA), on the CXI instrument[33]. X-ray pulses of 50 fs duration containing ~2.0×10^{12} photons per pulse at a wavelength of 1.46 Å (this mean pulse wavelength was initially derived from the tuned energy of the laser) were focused by Kirkpatrick–Baez mirrors to 2–3 μm diameter at the interaction point. The sample–detector distance was 90.9 mm. The photon beam parameters are summarized in Table 1. The crystal slurry was diluted to provide ~3×10^9 crystals per ml to provide a hit rate of ~4% (ref. 15)). The crystals were injected into the XFEL beam in water using a Gas Dynamic Virtual Nozzle[14] focused to a diameter of a few μm at a flow rate of 30 μl min^{-1}. During measurement a rotating syringe device[34] was used to avoid crystal settling. The injector was positioned to intersect the X-ray beam, before the Rayleigh break-up of the jet into drops. Single-shot diffraction patterns were recorded at 120 Hz while the liquid jet was flowing. The X-ray beam incident on the jet was attenuated to 2% of its original intensity. Data were recorded on a Cornell-Stanford pixel array detector[10].

Refinement of orientation matrix. A total of 144,803 images from ~20 min of data collection (which used 632 μl of crystal suspension) were analysed with cctbx.xfel[16] to determine initial orientation matrices. A total of 5,787 images could be indexed with the software available at that time. To 'catch' a complete set of observed reflections, the image was integrated with an inflated bandwidth ($\lambda \pm 3.5\%$), and those predicted reflections whose intensity exceeded 150 ADU (analogue-to-digital units) after integration were selected. At this stage rlps were modelled as delta functions. Orientation matrices were refined to minimize the spread of wavelengths required to satisfy the Bragg condition for this observed set. Orientation matrices were rotated about two axes orthogonal to the beam, each initially rotated by 0.5°, then halving the interval step upon finding local minima. For each orientation a histogram was calculated of the number of reflections versus the wavelength at which they were excited. The target function initially maximized reflection count and minimized distribution spread, where the parameters are altered to result in a minimization of the value of the target function. After the interval step had been halved three times, the target function was changed to be the quality of fit to a Gaussian model of wavelength distribution. Refinement was terminated when the rotational step size reached 0.001°.

Partiality model. Like the recent paper from White[12], we treat partiality as arising from the intersection of a rlp with a finite volume with a set of Ewald spheres of different radii, representing the range of incident X-ray wavelengths. Our partiality model rests on defining a partiality of 1 as a reflection that maximally excites all wavelengths of the X-ray pulse. In this context, the extent of excitation is defined as the cross-sectional area of the rlp that intersects the Ewald sphere, relative to the area of a central section of that rlp. Thus, notionally, the 0,0,0 rlp will have a

© 2015 Macmillan Publishers Limited. All rights reserved.

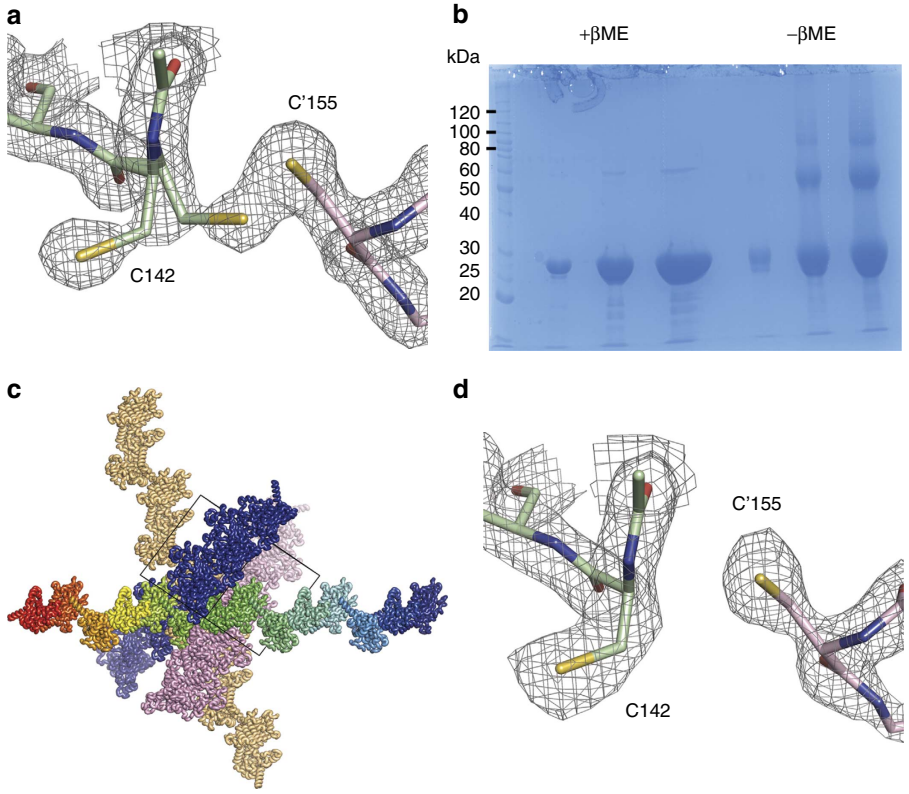


Figure 3 | Disulphide bonds link chains of CPV17 polyhedrin molecules. (**a**) Electron density around C142 (pale green) and C155 (pink) from a related molecule in the crystal contoured at 1.0 σ for CPV17 at 293 K. (**b**) SDS-polyacrylamide gel electrophoresis analysis of CPV17 crystals with and without reducing agent 2-mercapto-ethanol. (**c**) Visualization of the intermolecular disulphide-linked helical strings going through the crystal lattice of CPV17. Four helical strings are drawn, one with individual polyhedrin molecules coloured separately, the others coloured pink, cream and blue. (**d**) Electron density around C142 (pale green) and C155 (pink) from a related molecule in the crystal contoured at 1.0 σ for CPV17 at 100 K.

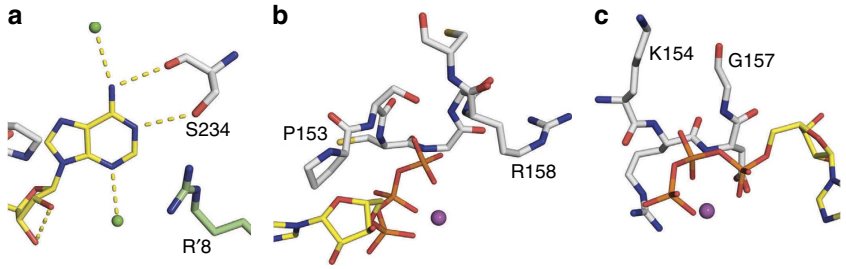


Figure 4 | ATP interactions. (**a**) Interactions of the base moiety of ATP of CPV17. Arg 8 (coloured green) is from a related molecule. Alignment of the ATP molecule for (**b**) CPV17 and (**c**) CPV1.

partiality of essentially 1, since the Ewald spheres are almost flat and pass centrally through the rlp. In practice we split this into two components, we first model the portion of the reflection that lies between the limiting spheres, and then integrate across the wavelengths that the rlp excites, using a Gaussian model for the wavelength distribution of the photons in each pulse. To deconvolute the width of this wavelength distribution from the Gaussian distribution we have fitted empirically for each pulse (which includes the combined effect of bandwidth, crystal mosaicity and finite crystal size), we multiply the s.d. by 0.27, found to be effective after a few manual trials. To model partiality, the Gaussian distribution was set to 0 beyond two s.d. from the mean, thereby defining the limiting Ewald spheres. Rlps were modelled as spherical top-hats with radii derived from crystal size and isotropic mosaicity; the intersection of these with the limiting Ewald spheres was calculated and used to derive appropriately scaled cylindrical approximations (where the cylinder base is in the plane of the tangent to the Ewald sphere, Supplementary Fig. 5). The partiality is then obtained by integrating across the wavelength distribution. The precise unreduced expression for partiality is shown in equation (1),

$$\rho = \left(2p^3 - 3q^2\right)\left(\frac{4\pi r^3}{3}\right)\left(\frac{1}{\pi r^3(p-q)}\right)\int_{x=l}^{h} \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}}\,dx \qquad (1)$$

where p and q are the fractions along the diameter of a spot at which the limiting radii of the Ewald spheres intersect, r represents the radius of the spot after correction for mosaicity, μ is the mean wavelength of the beam, σ is the s.d. of the beam wavelength and l and h are the lower and higher wavelengths corresponding to Ewald radii two s.d. from the mean, respectively. Reflection size was calculated according to a fixed term, spot size (equal to 1.8×10^{-5} Å^{-1}, determined by trial and error to minimize overprediction of spots), plus a mosaicity term (0.03°, as previously defined[22]), which allows variation in spot size with resolution.

The error model for the reflections combined counting statistics from integration divided by both the partiality and the scale factor for each image. Since the gain of the detector was not properly accounted for, these are not reliable.

Integration of each image. Spot positions were computed using the refined orientation matrices and metrology[16]. Any residual errors in metrology were corrected for by centring on the highest count pixel in a 5 × 5-box around this point. Images were integrated using simple two-dimensional (2D) integration using the DIALS software package. This 2D ordinate analysis method[35] will slightly overestimate weak data. The resolution of the image was taken as the highest-resolution shell where $I/\sigma(I)$ reached 2.0. For each image the raw integration count was recorded along with the partiality estimated for that given reflection.

© 2015 Macmillan Publishers Limited. All rights reserved.

Separation of indexing hands by cluster analysis. Each image was assigned an individual 2D vector, $\mathbf{v}_a$, where a is the image number. Both components of the vector are randomly assigned between 0 and 1 according to the cluster algorithm previously defined[24]. The algorithm has been modified to include a second term, corresponding to the inverted indexing hand. r_{ab} is the correlation coefficient between image numbers a and b, (equation 3), while the correlation between the (h, k, l) reflection of one image with the (k, h, l) reflection of the other is represented by r'_{ab}. These vectors were minimized using an L-BFGS algorithm[36] (equation (2)).

$$\varphi = \sum_a \sum_b (r_{ab} - \mathbf{v}_a \cdot \mathbf{v}_b)^2 - \sum_a \sum_b (1 - r'_{ab} - \mathbf{v}_a \cdot \mathbf{v}_b)^2 \qquad (2)$$

Correlation coefficients r_{ab} and r'_{ab} lie between 0 and 1, and the formula is modified to ensure that the regression line passes through the origin (equation (3)). This allows a meaningful correlation coefficient to be generated from just two shared reflections. If the value of r_{ab} or r'_{ab} is imaginary, then it is reassigned a value of 0.

$$r = \sqrt{1 - \sum_n \frac{(I_a - mI_b)^2}{(I_a - \bar{I}_a)^2}} \text{ where } m = \frac{\sum_n I_a I_b}{\sum_n I_a^2} \qquad (3)$$

Correlation coefficients are generated from intensities that have been corrected for partiality. Reflections with a partiality less than 0.3 were rejected, the remaining reflection intensities and sigma values were divided by their partiality. Beyond 1.8 Å resolution the reflection count rapidly dropped due to the smaller angle of observation on the detector.

Scaling images. Scale factors for individual images were generated initially by adjusting the mean intensity of reflections (to 3.5 Å resolution) in each image to equal values. After rejecting 233 (3.6%) of the weakest images (those for which the mean intensity was <333 ADU) and merging, scale factors were then adjusted by calculating the gradient between each image and the initial merged data set, after remerging the data with the new scale factors the process was iterated to convergence.

Powder diffraction. Crystals were mounted into capillary tubes for powder diffraction experiments. Data were collected at 100 K and a second data set collected at room temperature, using otherwise identical experimental conditions (the detector was not moved, the wavelength was identical and the data were collected close together in time). Accurate unit cell dimensions were calculated from the very sharp diffraction pattern rings.

Synchrotron data collection. Polyhedra were purified as for the XFEL sample. Crystals were cryo-protected by mixing 1:1 with ethylene glycol, equilibrated for 60 s, applied to a MicroMesh mount (Mitegen, Ithaca, USA) and allowed to settle for 30 s before excess liquid was wicked away. The crystals were flash-cooled in a stream of nitrogen gas at 100 K. The X-ray beam at beamline I24, Diamond Light Source was trimmed to $\sim 4 \times 4$ µm at the sample using apertures close to the crystal (D.A. and G.E., unpublished). Only 1° of data could be collected per crystal (20 images of 0.05°), due to the long exposure times required. Data collection and processing, summarized in Table 2, have been described previously[25]. The final data set comprised measurements from 768 crystals. Initial attempts at molecular replacement using these data were unsuccessful. The structure was solved by choosing the same indexing choice as the XFEL data and then positioning the XFEL model (see below) by rigid body refinement, followed by standard refinement with REFMAC5 (ref. 37).

Molecular replacement and model refinement. Structures of seven homologues with 9–18% sequence identity with CPV17 (unpublished data and refs 6,26) were superposed, the occupancy of each atom reduced to 1/7 and the coordinates concatenated. This agglomerated coordinate set was used as a single search model in Phaser[38], leading to preliminary phases, and a starting map that clearly showed novel information (Supplementary Fig. 6). The model with closest sequence similarity was set to polyalanine and manually rebuilt in Coot[39]. After initial phase improvement, Buccaneer[40] was used for partial rebuilding of the remaining model. The remainder of the refinement was completed manually, using REFMAC5 (ref. 37) and Phenix[41]. Interestingly, Phaser found a similar molecular replacement solution using the synchrotron data, but we were unable to produce a reliable refined model from that starting point, presumably because, without higher-resolution data, in the absence of non-crystallographic symmetry and with very little bulk solvent it was not possible to escape bias from the starting models. The quality of the final electron density maps is shown in Supplementary Fig. 7.

References

1. Kunimi, Y. Current status and prospects on microbial control in Japan. *J. Invertebr. Pathol.* **95,** 181–186 (2007).
2. Goldsmith, M. R., Shimada, T. & Abe, H. The genetics and genomics of the silkworm, bombyx mori. *Annu. Rev. Entomol.* **50,** 71–100 (2005).
3. Poinar, Jr. G. & Poinar, R. Fossil evidence of insect pathogens. *J. Invertebr. Pathol.* **89,** 243–250 (2005).
4. Coulibaly, F. *et al.* The molecular organization of cypovirus polyhedra. *Nature* **446,** 97–101 (2007).
5. Ji, X. *et al.* How baculovirus polyhedra fit square pegs into round holes to robustly package viruses. *EMBO J.* **29,** 505–514 (2010).
6. Axford, D., Ji, X., Stuart, D. I. & Sutton, G. In cellulo structure determination of a novel cypovirus polyhedrin. *Acta Crystallogr. D Biol. Crystallogr.* **70,** 1435–1441 (2014).
7. Chapman, H. N. *et al.* Femtosecond X-ray protein nanocrystallography. *Nature* **470,** 73–77 (2011).
8. Redecke, L. *et al.* Natively inhibited Trypanosoma brucei cathepsin B structure determined by using an X-ray laser. *Science* **339,** 227–230 (2013).
9. Liu, W. *et al.* Serial femtosecond crystallography of G protein-coupled receptors. *Science* **342,** 1521–1524 (2013).
10. Boutet, S. *et al.* High-resolution protein structure determination by serial femtosecond crystallography. *Science* **337,** 362–364 (2012).
11. Tenboer, J. *et al.* Time-resolved serial crystallography captures high-resolution intermediates of photoactive yellow protein. *Science* **346,** 1242–1246 (2014).
12. White, T. A. Post-refinement method for snapshot serial crystallography. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **369,** 20130330 (2014).
13. Kabsch, W. Processing of X-ray snapshots from crystals in random orientations. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **70,** 2204–2216 (2014).
14. DePonte, D. P. *et al.* Gas dynamic virtual nozzle for generation of microscopic droplet streams. *J. Phys. D Appl. Phys.* **41,** 195505 (2008).
15. Sauter, N. K., Hattne, J., Grosse-Kunstleve, R. W. & Echols, N. New Python-based methods for data processing. *Acta Crystallogr. D Biol. Crystallogr.* **69,** 1274–1282 (2013).
16. Hattne, J. *et al.* Accurate macromolecular structures using minimal measurements from X-ray free-electron lasers. *Nat. Methods* **11,** 545–548 (2014).
17. Greenhough, T. J. & Helliwell, J. R. Oscillation camera data processing: reflecting range and prediction of partiality. 2. Monochromatized synchrotron X-radiation from a singly bent triangular monochromator. *J. Appl. Crystallogr.* **15,** 493–508 (1982).
18. Kabsch, W. Integration, scaling, space-group assignment and post-refinement. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **66,** 133–144 (2010).
19. Barends, T. R. *et al.* De novo protein crystal structure determination from X-ray free-electron laser data. *Nature* **505,** 244–247 (2014).
20. Margaritondo, G. & Rebernik Ribic, P. A simplified description of X-ray free-electron lasers. *J. Synchrotron. Radiat.* **18,** 101–108 (2011).
21. Waterman, D. G. *et al.* The DIALS framework for integration software. *CCP4 Newsletter on Protein Crystallography* **49,** 16–19 (2013).
22. Rossmann, M. G., Leslie, A. G. W., Abdel-Meguid, S. S. & Tsukihara, T. Processing and post-refinement of oscillation camera data. *J. Appl. Cryst.* **12,** 570–581 (1979).
23. Karplus, P. A. & Diederichs, K. Linking crystallographic model and data quality. *Science* **336,** 1030–1033 (2012).
24. Brehm, W. & Diederichs, K. Breaking the indexing ambiguity in serial crystallography. *Acta Crystallogr. D Biol. Crystallogr.* **70,** 101–109 (2014).
25. Gildea, R. J. *et al.* New methods for indexing multi-lattice diffraction data. *Acta Crystallogr. D Biol. Crystallogr.* **70,** 2652–2666.
26. Anduleit, K. *et al.* Crystal lattice as biological phenotype for insect viruses. *Protein Sci.* **14,** 2741–2743 (2005).
27. Rohrmann, G. F. Polyhedrin structure. *J. Gen. Virol.* **67,** 1499–1513 (1986).
28. Donovan, J. W. & White, T. M. Alkaline hydrolysis of the disulfide bonds of ovomucoid and of low molecular weight aliphatic and aromatic disulfides. *Biochemistry* **10,** 32–38 (1971).
29. Juers, D. H. & Matthews, B. W. Reversible lattice repacking illustrates the temperature dependence of macromolecular interactions. *J. Mol. Biol.* **311,** 851–862 (2001).
30. Weierstall, U. *et al.* Lipidic cubic phase injector facilitates membrane protein serial femtosecond crystallography. *Nat. Commun.* **5,** 3309 (2014).
31. Shapiro, A. *et al.* Morphological and molecular characterization of a Cypovirus (Reoviridae) from the mosquito Uranotaenia sapphirina (Diptera: Culicidae). *J. Virol.* **79,** 9430–9438 (2005).
32. Zhao, Y., Chapman, D. A. & Jones, I. M. Improving baculovirus recombination. *Nucleic Acids Res.* **31,** E6–E6 (2003).
33. Emma, P. *et al.* First lasing and operation of an ångstrom-wavelength free-electron laser. *Nat. Photon.* **4,** 641–647 (2010).
34. Lomb, L. *et al.* An anti-settling sample delivery instrument for serial femtosecond crystallography. *J. Appl. Crystallogr.* **45,** 674–678 (2012).
35. Watson, H. C., Shotton, D. M., Cox, J. M. & Muirhead, H. Three-dimensional Fourier synthesis of tosyl-elastase at 3.5Å resolution. *Nature* **225,** 806–811 (1970).
36. Liu, D. C. & Nocedal, J. On the limited memory BFGS method for large scale optimization. *Math. Program. B* **45,** 503–528 (1989).
37. Murshudov, G. N. *et al.* REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallogr. D Biol. Crystallogr.* **67,** 355–367 (2011).

© 2015 Macmillan Publishers Limited. All rights reserved.

38. McCoy, A. J. *et al.* Phaser crystallographic software. *J. Appl. Crystallogr.* **40,** 658–674 (2007).
39. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D Biol. Crystallogr.* **60,** 2126–2132 (2004).
40. Cowtan, K. The Buccaneer software for automated model building. 1. Tracing protein chains. *Acta Crystallogr. D Biol. Crystallogr.* **62,** 1002–1011 (2006).
41. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D Biol. Crystallogr.* **66,** 213–221 (2010).
42. Maia, F. R. The Coherent X-ray Imaging Data Bank. *Nat. Methods* **9,** 854–855 (2012).
43. Stuart, D. I., Levine, M., Muirhead, H. & Stammers, D. K. Crystal structure of cat muscle pyruvate kinase at a resolution of 2.6Å. *J. Mol. Biol.* **134,** 109–142 (1979).

Acknowledgements

A.S.B., J.H. and N.K.S. were supported by US National Institutes of Health grant R01GM102520. G.S. and D.I.S. were supported by the Medical Research Council, grant G1000099. H.M.G. was supported by the Wellcome Trust (studentship ALR00040). M.M. was supported by NSF award 1231306. We are very grateful for the expert support for the operation of the sample injector provided by the group of Ilme Schlichting (Max Planck Institute for Medical Research, Heidelberg, Germany), in particular Sabine Botha, R. Bruce Doak and Robert L. Shoeman. We are very grateful to LCLS-CXI staff, Sébastien Boutet, Garth Williams and Dan Deponte. Portions of this research were carried out at the Linac Coherent Light Source (LCLS) at the SLAC National Accelerator Laboratory. LCLS is an Office of Science User Facility operated for the U.S. Department of Energy Office of Science by Stanford University. Admin support from the Wellcome Trust, grant 090532/Z/09/Z. This is a contribution from the Oxford Instruct Centre. Computer code is available on request.

Author contributions

G.S. prepared samples. All authors performed research. H.M.G. wrote the original software. H.M.G., G.S., N.K.S., G.E. and D.I.S. wrote the manuscript, in discussion with the other authors.

Additional information

Accession codes: The coordinates and structure factors for the XFEL and synchrotron structures of CPV17 have been deposited with the RCSB under accession codes 4S1L and 4S1K, respectively, and raw data at CXIDB.org[42].

Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: Ginn, H. M. *et al.* Structure of CPV17 polyhedrin determined by the improved analysis of serial femtosecond crystallographic data. *Nat. Commun.* 6:6435 doi: 10.1038/ncomms7435 (2015).

This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/

© 2015 Macmillan Publishers Limited. All rights reserved.

BIOLOGICAL CRYSTALLOGRAPHY

ISSN 1399-0047

A revised partiality model and post-refinement algorithm for X-ray free-electron laser data

Helen Mary Ginn,[a] Aaron S. Brewster,[b] Johan Hattne,[b] Gwyndaf Evans,[c] Armin Wagner,[c] Jonathan M. Grimes,[a,c] Nicholas K. Sauter,[b] Geoff Sutton[a] and David Ian Stuart[a,c]*

[a]Division of Structural Biology, Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford OX3 7BN, England, [b]Physical Biosciences Division, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, USA, and [c]Diamond House, Harwell Science and Innovation Campus, Fermi Avenue, Didcot OX11 0QX, England. *Correspondence e-mail: dave@strubi.ox.ac.uk

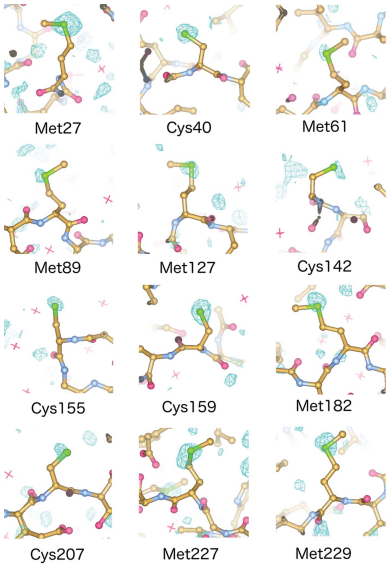
Received 12 March 2015
Accepted 6 April 2015

Keywords: post-refinement; free-electron laser; partiality.

PDB reference: 1.46 Å resolution XFEL structure of CPV17, 4zqx

Supporting information: this article has supporting information at journals.iucr.org/d

Research towards using X-ray free-electron laser (XFEL) data to solve structures using experimental phasing methods such as sulfur single-wavelength anomalous dispersion (SAD) has been hampered by shortcomings in the diffraction models for X-ray diffraction from FELs. Owing to errors in the orientation matrix and overly simple partiality models, researchers have required large numbers of images to converge to reliable estimates for the structure-factor amplitudes, which may not be feasible for all biological systems. Here, data for cytoplasmic polyhedrosis virus type 17 (CPV17) collected at 1.3 Å wavelength at the Linac Coherent Light Source (LCLS) are revisited. A previously published definition of a partiality model for reflections illuminated by self-amplified spontaneous emission (SASE) pulses is built upon, which defines a fraction between 0 and 1 based on the intersection of a reflection with a spread of Ewald spheres modelled by a super-Gaussian wavelength distribution in the X-ray beam. A method of post-refinement to refine the parameters of this model is suggested. This has generated a merged data set with an overall discrepancy (by calculating the $R_{\rm split}$ value) of 3.15% to 1.46 Å resolution from a 7225-image data set. The atomic numbers of C, N and O atoms in the structure are distinguishable in the electron-density map. There are 13 S atoms within the 237 residues of CPV17, excluding the initial disordered methionine. These only possess 0.42 anomalous scattering electrons each at 1.3 Å wavelength, but the 12 that have single predominant positions are easily detectable in the anomalous difference Fourier map. It is hoped that these improvements will lead towards XFEL experimental phase determination and structure determination by sulfur SAD and will generally increase the utility of the method for difficult cases.



OPEN ACCESS

1. Introduction

A number of structures have been solved using serial femtosecond crystallography (SFX) at an X-ray free-electron laser (XFEL; Redecke *et al.*, 2013; Liu *et al.*, 2013; Boutet *et al.*, 2012; Tenboer *et al.*, 2014; Kern *et al.*, 2013, 2014; Ginn *et al.*, 2015), which have benefited from a large number of indexable snapshots from the crystalline samples. The recent structure of photoactive yellow protein (Tenboer *et al.*, 2014), capturing time-resolved high-resolution intermediates, used between 22 678 images and 64 496 images to generate structures, achieving a discrepancy of 6.5% between two half data sets ($R_{\rm split}$) for the latter. Such studies benefit from an abundant supply of crystalline sample, allowing the use of Monte Carlo integration (Kirian *et al.*, 2010, 2011). This method has been successfully used to observe the anomalous signal from the S atoms in lysozyme (Barends *et al.*, 2013) using 43 840 indexed patterns collected at a wavelength of 1.7 Å. Difference Fourier

peaks associated with a methionine S atom were observed to a maximum of 4.5σ at 3.5 Å resolution using the *CrystFEL* software suite (White *et al.*, 2012; Boutet *et al.*, 2012).

The structure of photosystem II (Kern *et al.*, 2013) was solved from data processed using the *cctbx.xfel* software suite, which is used, in part, in this study. The pipeline in *cctbx.xfel* has since been improved to include crystal-orientation refinement, used in the later processing of photosystem II (Kern *et al.*, 2014; Sauter *et al.*, 2014; Sauter, 2015). In systems where the supply of crystals is restricted, a high indexing rate and careful modelling of XFEL-derived data is of paramount importance (Hattne *et al.*, 2014).

Several algorithms and models have been developed, including an algorithm to minimize the distance of modelled reflections from the Ewald sphere (termed the Ewald offset correction; Kabsch, 2014), which improves the orientation matrices. This was also achieved by Sauter *et al.* (2014), successfully improving crystallographic *R* factors during refinement. The problem of assigning accurate orientation matrices to diffraction data is not unique to XFEL studies, and is helped by applying post-refinement, which was first developed by Rossmann *et al.* (1979) and Winkler *et al.* (1979) for oscillation data and essentially uses a reference set of intensities (often obtained by a preliminary merging of the current data set) as a target for the improved modelling of partially recorded reflections. White (2014) developed a post-refinement and partiality model and applied it to simulated data, seeking to assign partialities to match the fact that no reflection is fully recorded, resulting in a large improvement in merging statistics. In a separate method, Sauter (2015) applied a partiality model and several post-refinement algorithms to thermolysin data, refining a simple scale factor, linear isotropic *B* factor and orientation matrix angles all together, which improved the anomalous signal of the Zn atom in the structure. Changes to aspects of the experiment including the treatment of multiple lattices, resolution-cutoff consideration and better spot-shape models have improved the *cctbx.xfel* pipeline, without even needing to consider partiality (Hattne *et al.*, 2014). We have also previously reported a method of improving the orientation matrices and describing these partialities, which provided high-quality data (an $R_{\rm split}$ of 11.7% from 5787 crystals) and a structure determination at 1.75 Å resolution (Ginn *et al.*, 2015). However, the data used to solve this structure were calculated with only refinement of the orientation matrix, without the help of a reference data set. Here, we largely discuss refinement of the parameters which contribute to the partiality model using a reference data set.

We argue that the value of the modelled partiality is acutely dependent on accurate definition of the orientation matrices to describe XFEL data, which have been gradually improving as processing techniques progress. Here, we present an updated partiality model and a post-refinement algorithm applied to 7225 images; together these lead to more accurately defined orientation matrices and more reliable data. This partiality model is distinguished by taking into account both the size of the reciprocal-lattice point (rlp) and the multiple Ewald spheres which are intersected, which are in turn defined by the wavelength of the beam and its energy spread. We have used images generated from cypovirus polyhedra, previously solved to 1.75 Å resolution (Ginn *et al.*, 2015), as a well behaved test case. In the 1.46 Å resolution structure obtained, the differences in peak density for O, N and C atoms are better resolved and data quality is sufficient to clearly see the weak anomalous signal of 12 of the 13 ordered S atoms within the asymmetric unit, despite the X-ray energy being very far from the sulfur *K* edge.

2. Materials and methods

2.1. Data collection

SFX experiments were carried out at the Linac Coherent Light Source of the SLAC National Accelerator Laboratory, Menlo Park, California, USA using the CSPAD detector on the CXI instrument to solve the structure of CPV17 at approximately 1.46 Å wavelength (Ginn *et al.*, 2015). Data collected at a later stage of the same visit (18 March 2013) used the same experimental setup and beamline parameters except that the wavelength was decreased to 1.3 Å, allowing a resolution of 1.65 Å at the edge and 1.42 Å in the corner of the detector. The latter data were used in this study. The transmittance of the beam was set to 3%. A total of 66 564 diffraction patterns were recorded, of which 7225 were indexed using the *cctbx.xfel* pipeline. The flow rate of the injector was increased by 33% relative to the data collected at 1.46 Å, leading to an enrichment in images for which multiple crystals were within the beam compared with the previously reported data set.

2.2. Detector geometry

The CSPAD detector comprises 64 pixel-array panels, and is periodically rebuilt. Over the course of an experiment, when the detector is rebuilt or the physical positions of the XFEL components are changed the relative positions and rotations of the panels may change, but it is difficult to accurately measure and incorporate this information into the image metadata at the time of collection. We refer to the correction of these variations as detector geometry, but it is known in *cctbx.xfel* as metrology. *cctbx.xfel* uses a method of metrology correction based on information from multiple images to correct the metrology to subpixel accuracy (Hattne *et al.*, 2014), but in this case the metrology correction was not extracted from the images, and an alternative method to correct for geometry errors is described here.

Integration (see §2.5) was initially carried out over a subset of 500 images. The integration mask was shifted to the local maximum intensity within a 21 × 21 window centred on the initial predicted spot position. We recorded, for each panel, the average shift for every spot on the detector which falls within the coordinates of that panel. These shifts were then plotted, as in Fig. 1(*a*), showing a clear preference for spots to shift in a characteristic way for each panel. Streaking of the plot (Fig. 1*b*) occurs when there are errors in the detector

distance or integration wavelength, which is discussed in §2.3. To calculate the translational shift for each panel, a moving window of a 2 × 2 pixel box scanned the plot as shown in Fig. 1(*a*), with a step size of 0.1 pixels. The coordinates of the window that lies over the largest number of shifts were stored for future use. The rotations of the panel along the *X* and *Y* axes in the plane of the detector were also refined: for the *X* axis the position of a reflection on the panel can be expressed as a fraction of the panel width, and we found that the recorded shift for a reflection is partially dependent on this fraction. If a panel is rotated on the *X* axis, the left-hand side of the panel may be closer to the crystal than the right-hand side. This will manifest by pushing reflections on both the left-hand and the right-hand sides of the panels towards the edges of the panel. *X* and *Y* rotation parameters take into account the extra shift caused by the dependence on the predicted position on the panel along the *X* or *Y* axis. This was calculated empirically by plotting the horizontal axis shift against the horizontal coordinates of the predicted position as a fraction of the panel width and fitting a regression line, the gradient of which is the *X* rotation parameter. This was similarly calculated for the *Y* rotation and also stored for future use. Later integration events then applied the most common shift to each reflection and applied the rotation parameters based on their position on the panel using the geometry-corrected shifts for all spot integration, locking weak reflections to the correct coordinates.

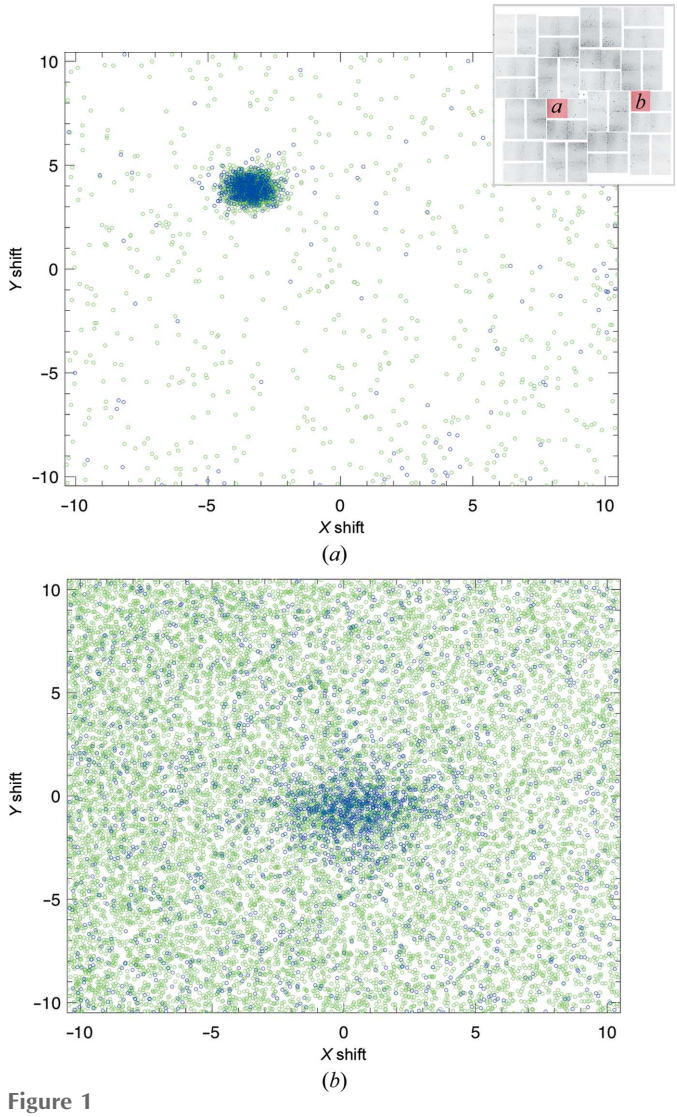


Figure 1
Spot positions were allowed to migrate around a 21 × 21 box centred on the original predicted position. The pixel shift in *X* and *Y* coordinates from the original starting point was retained for each reflection. These pixel shifts were aggregated for each panel. The *X* and *Y* pixel shifts of each reflection on a single panel were plotted against each other, as above. The average intensity counts were calculated for each panel. Green reflections are those below the average, whereas blue reflections are those above the average intensity. (*a*) This panel is the third panel from the left and five panels down from the top of the detector. The most common shift in this panel was easily resolved. The applied shift for this panel was (−3.2, 4.4) pixels in *X*, *Y* coordinates. (*b*) This panel is the second panel from the right and five panels down from the top of the detector, plotted at an incorrect detector distance to show streaking along the *X* axis, but the most common shift could still be resolved and was at (0.2, −0.5) pixels. Because the data were weaker than those in (*a*), more spots incorrectly focused on noise and deviated from the common shift of (0.2, −0.5).

2.3. Refinement of global parameters

The unit-cell length was set to 106.1 Å, which had been determined previously from room-temperature powder diffraction studies (Ginn *et al.*, 2015). Any residual variation in the unit cell was accounted for by refinement of the mean wavelength of each pulse, which is perfectly correlated with the dimensions of a cubic lattice, $a = b = c$.

If the proposed distance from the crystal to the detector was too short, reflections appeared closer to the centre of the image than they actually were, which was reflected in the shift plot. A shift of an individual reflection on a panel was greater if the reflection was furthest away from the centre of the image, whilst lower resolution reflections were less affected by errors in the proposed detector distance. This caused a broadening of shifts along the vector from the centre of the image to the midpoint of a particular panel. This occurred whether the proposed detector distance was too close or too far. Because the streaking of the plot (Fig. 1*b*) was accentuated when the error in detector distance was greater, minimizing this streaking effect could be used to manually adjust the detector distance by eye. The optimum wavelength for integration could then be extracted after initial orientation-matrix refinement (see §2.4). This was an iterative process, with successive changes in detector distance and wavelength progressively improving the panel shift plots.

2.4. Initial orientation-matrix refinement

Initial orientation-matrix refinement was carried out as described previously (Ginn *et al.*, 2015) for each image in order to produce an initial orientation matrix as close as possible to the true matrix without requiring a reference data set. Reflections were classed as strong if they reached an $I/\sigma(I)$ of greater than 12 (rather than using an absolute intensity threshold, as was used previously), but the counting errors were not included in further refinement. Strong reflections

were used in initial refinement of the orientation matrix. An $I/\sigma(I)$ of 12 is indeed a high threshold, but was established after a few manual trials as a value which balances obtaining a large number of strong reflections with excluding noise.

2.5. Integration

Integration of reflections involved summation of foreground pixel photon counts minus background pixel photon counts according to the masks outlined in Fig. 2(*a*) during initial orientation-matrix refinement (see §2.4). Starting spot positions were derived from the *cctbx.xfel* orientation matrices. The midpoint of the mask was initially centred on the highest intensity pixel within a 21 × 21 window, before better detector geometry had been calculated, after which the spots were not allowed to wander but were focused on the most common shift for each individual panel (see §2.2). Counting statistics were disregarded owing to potential interference from the uncertain and nonlinear gain of the detector and we found, empirically, that including counting statistics in the merging process reduced the quality of the final data. Errors in individual reflections were set to unity, divided by the partiality and multiplied by the scale factors for individual images.

After initial orientation-matrix refinement and immediately prior to generating the reflection list which continues into post-refinement, unique integration boxes were generated for individual reflections. The broadening on one axis for a given spot was caused by the broad range of wavelengths, and therefore range of Ewald sphere centres, which will cause illumination of a spot. Spots were stretched along the broadened axis by a number of pixels calculated by ray-tracing from limiting Ewald sphere centres to the detector and finding the resultant pixel shift. Each individual pixel can contribute a proportion between 0 and 1 towards the foreground signal for a given spot. The background reading was taken from a square padded by one pixel away from the ellipse, as shown in Fig. 2(*b*). The mean intensity of the background pixels was calculated and this value was subtracted from each foreground pixel. If any part of the integration shoebox or background pixels fell off the edge of a panel it was excluded from integration. This rejected approximately 30% of reflections.

2.6. Creating an initial data set to seed refinement

For the initial merge, each image was overpredicted by using an artificially wide energy bandwidth of 3.0%, and an initial partiality correction was applied to reflections within this range. The parameters for this partiality correction were set as the pre-determined initial starting values, or in the case of the mean wavelength by choosing the Ewald sphere wavelength at which the largest number of strong reflections were illuminated from the initial orientation-matrix refinement. Approximately 2000 reflections were predicted for each image using this method. These overpredicted reflections were kept in memory in order that they could contribute to future alterations in the orientation matrix during post-refinement. There was an indexing ambiguity in space group *I*23 which was broken by using a modified version of the algorithm of Brehm & Diederichs (2014), as has been described previously (Ginn *et al.*, 2015). Intensities were weighted by the unrefined partiality correction and scale factors to bring the average intensity of each image to 1000 ADU (analogue-to-digital units) before merging.

2.7. Nature of the partiality model

We have used a partiality model based on that described in Ginn *et al.* (2015) with some alterations. Two functions were defined, one describing the profile of the energy spectrum, E, and one describing the profile of the rlp, P. The final partiality ρ for each reflection is defined in (1), where p is the fraction of the rlp along the curve of constant length from the origin, from 0 (lowest wavelength) to 1 (highest wavelength), as defined in Rossmann *et al.* (1979),

$$\rho = \int_0^1 (EP)\, dp. \qquad (1)$$

The function E was defined as a Gaussian distribution in our previous version of the partiality model, but this has been changed to a super-Gaussian model of mean wavelength μ and standard deviation σ with an exponent N (Decker, 1995) set to 1.5, closer in resemblance to a Christmas tree, as shown in (2). λ_p represents the Ewald sphere wavelength of the reciprocal-space coordinate at fraction p along the rlp. μ is the mean wavelength of the beam and σ is the standard deviation of the beam wavelength (the bandwidth of the pulse). The shape of

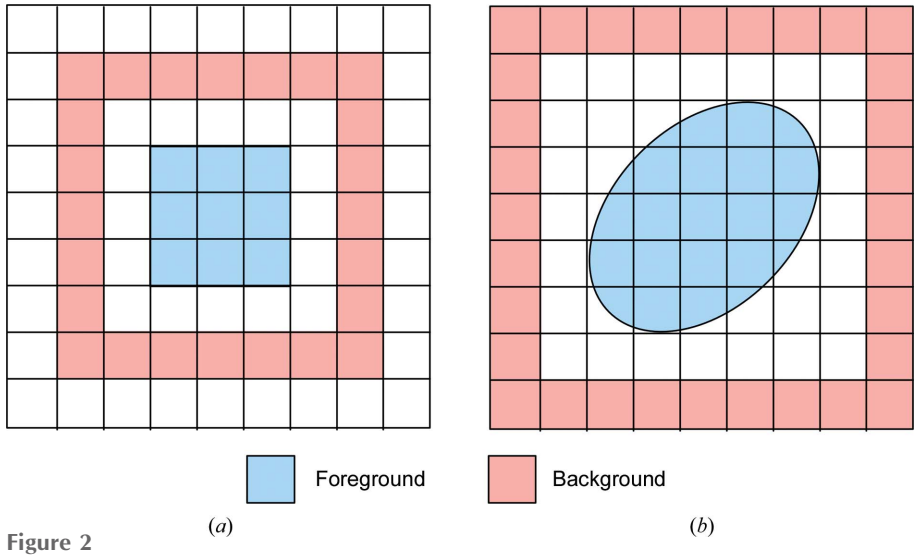


Figure 2
Foreground and background masks used to calculate the integrated signal of each reflection, using the simple shoebox during initial orientation-matrix refinement (*a*) and an elliptical shoebox based on the energy bandwidth (*b*).

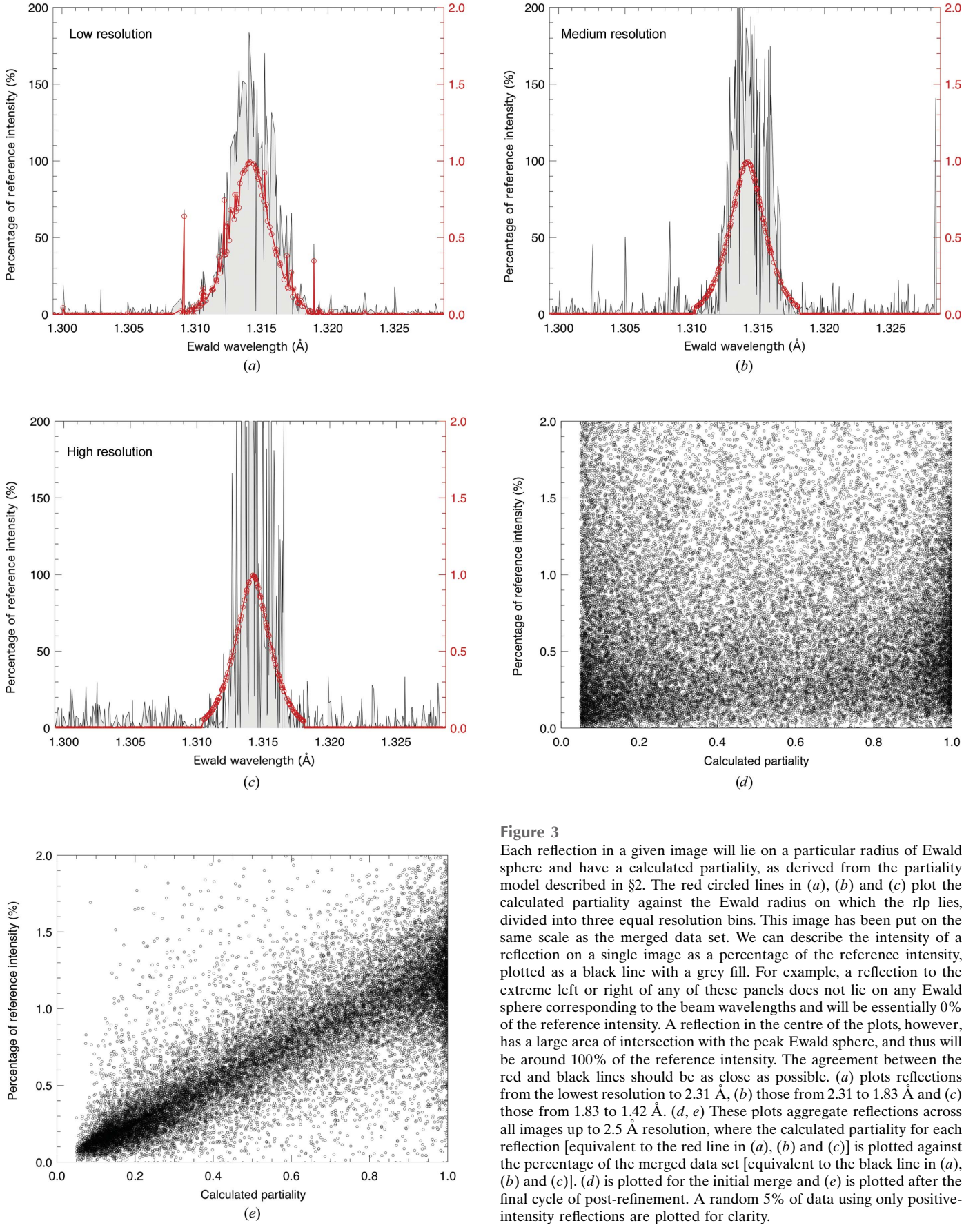


Figure 3
Each reflection in a given image will lie on a particular radius of Ewald sphere and have a calculated partiality, as derived from the partiality model described in §2. The red circled lines in (*a*), (*b*) and (*c*) plot the calculated partiality against the Ewald radius on which the rlp lies, divided into three equal resolution bins. This image has been put on the same scale as the merged data set. We can describe the intensity of a reflection on a single image as a percentage of the reference intensity, plotted as a black line with a grey fill. For example, a reflection to the extreme left or right of any of these panels does not lie on any Ewald sphere corresponding to the beam wavelengths and will be essentially 0% of the reference intensity. A reflection in the centre of the plots, however, has a large area of intersection with the peak Ewald sphere, and thus will be around 100% of the reference intensity. The agreement between the red and black lines should be as close as possible. (*a*) plots reflections from the lowest resolution to 2.31 Å, (*b*) those from 2.31 to 1.83 Å and (*c*) those from 1.83 to 1.42 Å. (*d*, *e*) These plots aggregate reflections across all images up to 2.5 Å resolution, where the calculated partiality for each reflection [equivalent to the red line in (*a*), (*b*) and (*c*)] is plotted against the percentage of the merged data set [equivalent to the black line in (*a*), (*b*) and (*c*)]. (*d*) is plotted for the initial merge and (*e*) is plotted after the final cycle of post-refinement. A random 5% of data using only positive-intensity reflections are plotted for clarity.

this function can be seen in Figs. 3(*a*), 3(*b*) and 3(*c*), where the higher resolution panels (Figs. 3*b* and 3*c*) show minimal interference from the function *P*,

$$E = \frac{1}{k(2\pi\sigma)^{1/2}} \exp\left(\frac{-|\lambda_p - \mu|^N}{2\sigma^N}\right). \quad (2)$$

The scale factor k is the maximum achievable partiality for a reflection of the given rlp radius, if it were centred in the middle of the beam. This normalized partialities and made them comparable between resolution shells so they always fell between 0 and 1, and prevented per-cycle inflation of high-resolution reflection intensities during post-refinement. No Lorentz correction was applied to the data, and as a result the normalization applied above led to higher resolution reflections being somewhat underestimated, which was approximately corrected for by applying a *B* factor to the data. The function *P* describes the cross-section of the rlp as a proportion of the maximum cross-sectional area as in (3),

$$P = 4p(1-p). \quad (3)$$

Intensities and errors were divided by the partiality to inflate their values to an estimate of the true value. A minimum partiality cutoff of 0.05 was defined, and intensities with a partiality of less than 0.05 were not included in refinement. Applying this partiality model resulted in an average of 280 accepted reflections above the cutoff being merged for the images which refined correctly. Below a partiality of 0.05 the contribution of each reflection was so low, as the reflections were weighted by their partiality, that their contribution to the final data set would be negligible.

2.8. Refining individual images

There were four parameters to refine per individual image: two rotation angles, which allow correction of the orientation matrix along two axes perpendicular to the beam and to each other, the rlp size (governed by the size of the crystal exposed in the beam) and the mean wavelength of the SASE pulse. The target function was defined as the *R* factor between the image and the reference, which we aimed to minimize by altering the values of these parameters. Initial starting values were taken to be 0° for the two rotation angles and 1×10^{-4} Å^{-1} for the rlp size, which is equivalent to a 1 µm crystal size. The initial wavelength was taken as the mean of all Ewald sphere wavelengths where the intensity count was greater than 200 ADU (analogue-to-digital units).

The rlp size and mean wavelength parameters were altered repeatedly by testing the current value and the current value plus or minus a defined step length. The new value for the parameter was taken as the value corresponding to the lowest value of the target function. If the value remained unchanged from the previous alteration, the step length was divided by two. The rotation angles were altered in a combined, two-dimensional grid search, where nine possible values were tested and both step lengths were reduced simultaneously. The convergence criteria were 1×10^{-4} degrees for rotation angles, 1×10^{-5} Å for the mean wavelength and 1×10^{-5} Å^{-1} for the rlp size. The bandwidth was set to 0.26% (the spread of four standard deviations), equivalent to 25 eV for the experimental wavelength used. Mosaicity was set to zero, as it is very low for these crystals (from analysis of synchrotron data processing; Gildea *et al.*, 2014) and separating the effects of bandwidth expansion and increased mosaicity is difficult when the mosaicity parameter is so low. In this case, the effect of mosaicity was subsumed in the bandwidth effect. Correlation using each indexing choice was checked in order to make sure images were not mis-indexed during initial separation of indexing choices, and images were corrected if necessary.

2.9. Rejecting outliers

Outlier rejection occurred in both the individual image-refinement stages and at the merging stages. For images with a correlation weaker than 99% with the reference data set, up to three reflections per cycle were rejected if the removal of these reflections caused an upwards shift in the correlation coefficient CC, where the value of (1 − CC) decreased by more than 6% (for example, if a correlation coefficient of 90.0% increases to 90.6% or more). Once reflections had been rejected by this method, they were not reintroduced into the system until the final merge, in which the rejected reflections were recalculated afresh. This resulted in an average of 21.5 rejected reflections per image over the course of the full refinement process. The second stage of outlier rejection occurred just prior to merging. The mean and σ for the intensity of each reflection were calculated from the independent observations weighted by the individual partialities and scale factors. A rejection cutoff of 1.8σ was chosen in order to reject less than 10% of reflections, assuming a Gaussian distribution of observations per unique reflection. Reflections which were more than 1.8σ away from the mean were rejected, but these could be reincluded in subsequent rounds of refinement.

2.10. Merging

Observations were reduced to the unique index for the asymmetric unit of space group *I*23. All images above a certain threshold of correlation with the reference data set were included in the final merge for each macrocycle. This threshold was set to exclude images which failed to refine correctly, in this case 0.9 for all merges apart from the final merge, which was made more stringent (this was set to 0.95). Images were also rejected if the final number of reflections above the partiality cutoff of 0.05 was 100 or fewer. Scale factors were generated for each individual image. Friedel pairs were not separated for the purposes of generating scale factors for each image nor for outlier rejection, as the anomalous differences were considered to be negligible. Friedel pairs were maintained separate only to generate anomalous data on the final merge. The shared reflections between the individual image and the reference data set were plotted, and the best-fit gradient forced through the origin was calculated. The scale factor was set to the reciprocal of this gradient, so that following application of this scale factor the gradient would be

recalculated as unity. For each unique reflection, individual intensity observations were corrected for partiality and the mean and σ were calculated, excluding rejected reflections. The σ value is calculated from the distribution of observations for a given reflection.

Refinement was allowed to continue for at least seven cycles until the R factor between the most recent reference data set and the previous reference data set was 1% or less. After finishing refinement, the data were merged three times using scale factors generated using the previous best merge. The last merge was taken as the final data set.

2.11. Generation of anomalous difference Fourier map

In the case of merging to preserve the anomalous data, which occurred at the final merge, the rejection criteria were set to more stringent values at 1.0σ as these improved the spherical shape of the anomalous difference intensity for these atoms. The resolution for anomalous signal was cut arbitrarily to 1.8 Å. Anomalous difference Fourier maps were generated using phase estimates from *ANODE* (Thorn & Sheldrick, 2011).

3. Results

cctbx.xfel (Sauter *et al.*, 2013; Hattne *et al.*, 2014) identified 7225 hits from a 9 min run of data collection from CPV17 polyhedrin crystals on the CXI instrument at the LCLS (Emma *et al.*, 2010). These data were collected with pulses of X-ray wavelength 1.3 Å, which allowed spots to be integrated up to 1.42 Å resolution in the corner of the detector. Initial orientation matrices were generated by *cctbx.xfel* and were refined according to the protocols defined in Ginn *et al.* (2015). There was no uneven sampling of reciprocal space detected (*Appendix A*). Integration and resolution of indexing ambiguity was followed by a post-refinement algorithm to remove residual errors in the orientation matrix and refinable experimental parameters. The post-refinement algorithm comprises macrocycles, each composed of refinement against a reference data set followed by merging of a new reference data set, as outlined in §2. The overarching program architecture is illustrated in Fig. 4.

3.1. Integration boxes

Owing to the highly reproducible nature of CPV17 crystals, the unit cells were fixed to dimensions $a = b = c = 106.1$ Å, allowing the detector distance to be easily refined manually. Changing the shoebox to an ellipse decreased the highest resolution (1.485–1.460 Å) $R_{\rm split}$ from 51.4 to 42.5%. By fixing the unit cell to 106.1 Å and refining only the detector distance and wavelength this was further reduced to 40.6%.

3.2. Detector geometry

The diffraction images were not automatically corrected for detector-geometry errors, and our previous method corrected for geometry errors by centring on a local maximum photon count. This was a significant contribution to high $R_{\rm split}$ values at high resolution, as it overestimated the weak data. It was also found that by increasing the flow rate of the injector the number of multiple lattice images was increased compared with the data used in Ginn *et al.* (2015); this in turn increased the danger that a spot allowed to 'wander' would focus on an incorrect local maximum such as noise or a neighbouring lattice point. We applied a new type of geometry correction, defined in §2.2, and found that this increased the quality of the data. Without geometry correction, the $R_{\rm split}$ over all resolution shells is 3.53%, whereas with geometry correction this decreases to 3.15%.

3.3. Initial merge

The initial merge was carried out by applying the partiality model, in the absence of any post-refinement of parameters, to each individual image. The $R_{\rm split}$ for the initial merge was 9.20% and the $CC_{1/2}$ was 0.986. These values are similar to the statistics for our earlier 1.75 Å resolution structure, where 5787 crystals were merged to generate an $R_{\rm split}$ of 11.7%, when the small increase in image number is taken into account. However, this included the rejection of 69.3 reflections per

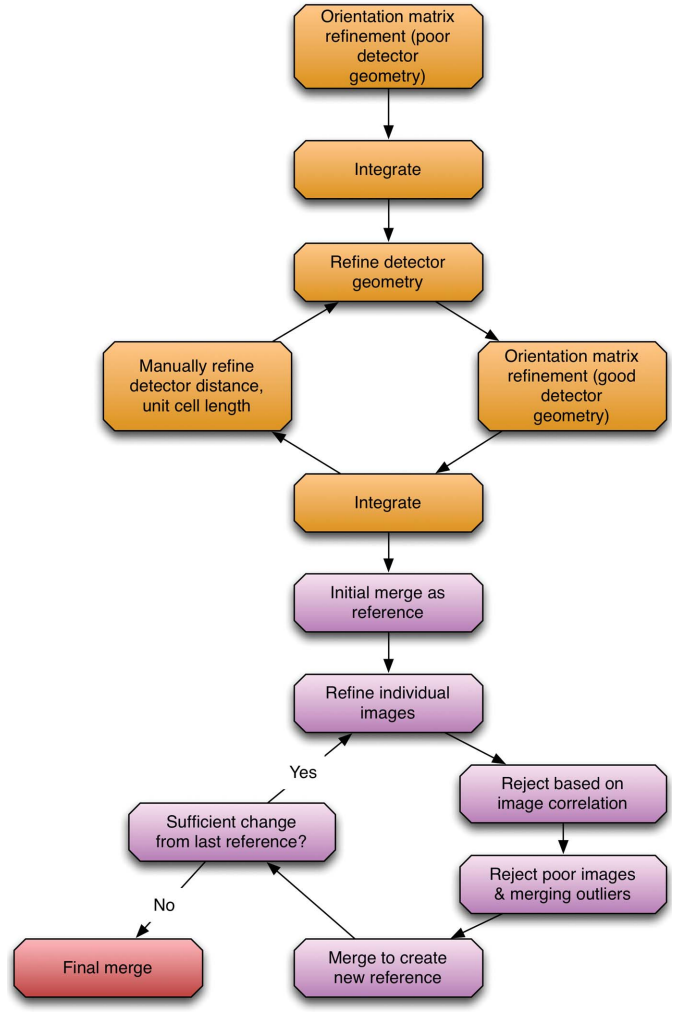


Figure 4
Diagram showing the flow of software during the post-refinement of XFEL data.

Table 1
Calculated $CC_{1/2}$ and R_{split} for each cycle of refinement.

The final merge uses a correlation threshold of 0.95 and fully recalculates rejected reflections.

Cycle No.	$CC_{1/2}$	R_{split} (%)	Rejected reflections per image
Initial merge	0.9860	9.20	69.3
Cycle 1	0.9987	4.76	22.2
Cycle 2	0.9989	4.04	22.9
Cycle 3	0.9990	3.85	23.6
Cycle 4	0.9990	3.78	24.1
Cycle 5	0.9991	3.73	24.4
Cycle 6	0.9990	3.71	24.6
Cycle 7	0.9991	3.70	24.8
Final merge	0.9995	3.15	21.5

image, indicative of the unrefined models producing large numbers of outliers. The major differences between these two processing strategies were the super-Gaussian modelling of the bandwidth, compared with a Gaussian model for the 1.75 Å resolution structure, and the fixing of the energy spread in the 1.46 Å resolution structure, compared with a variable parameter in the previous processing. This initial merge provided a good starting reference against which the data set was refined, even for low numbers of images (see §3.8).

3.4. Refining individual images

The rotation angles, rlp size and mean X-ray wavelength were refined according to §2.8. The average rotation in the orientation matrix was 0.026°. A successfully refined image had a good match between the predicted partiality and the integrated counts as a percentage of the reference data set, which was considered to be a measure proportional to the true partiality of the reflection. This is shown in Fig. 3. The super-Gaussian exponent of 1.5 was chosen on the basis that it became clear from plots such as Fig. 3 that there was some variation in the shape of the energy bandwidth which was not accounted for by an exponent of 2. After refining these parameters, the calculated partiality can be seen to agree well with the estimate of the true partiality. In the higher resolution shells the quality of the data was acutely dependent on the orientation matrix, detector-geometry correction and spot integration being as accurate as possible. The combination of these effects and the weaker data explains the diminishing $CC_{1/2}$ and increasing R_{split} at high resolution.

3.5. Merging after each cycle

In the first cycle after the initial merge, 4.5% of images had their indexing choice corrected. During the final merge, 9.5% of images were rejected owing to individual image-exclusion criteria. For the images which were included in the final merge, the average correlation coefficient with the reference was 0.988. Rejection of individual reflections based on individual image correlation alone rejected an average of 6.3 reflections per image, whereas rejections based on merging statistics rejected 20.9 reflections prior to the final merge. Inspection of individual diffraction patterns showed that these rejections were largely a result of overlapping or close reflections from multiple crystals in the beam. Although excluded reflections were recalculated on each merge, these were also excluded from individual images in the next round of image refinement. To calculate the progress of refinement, $CC_{1/2}$ and R_{split} were calculated on each merge and these values converged in seven cycles, as shown in Table 1, to a final R_{split} of 3.15%.

Table 2
Crystallographic refinement data for new processing of CPV17.

Values in parentheses are for the highest resolution shell or are standard deviations (s.d.) where specified.

Total diffraction patterns	65564
No. of indexed diffraction patterns	7225
No. of patterns used in final merge	6537
Space group	*I*23
Unit-cell parameter (Å)	$a = b = c = 106.1$
Resolution (Å)	25.0–1.46 (1.485–1.460)
Completeness (%)	99.5 (92.8)
Multiplicity	45.4 (9.14)
Unique reflections	34369
Total observations	1830360
Reflections per image	280
No. of reflections rejected/mean No. per image	8.9/21.5
Mean wavelength (s.d.) (Å)	1.3150 (0.00211)
Mean crystal dimension (s.d.) (µm)	1.22 (0.43)
Mean rotation correction (s.d.) (°)	0.026 (0.025)
R_{split} (%)	3.15 (40.6)
$CC_{1/2}$	0.9993 (0.3311)
R_{meas} (%)	19.0 (71.4)
$R_{p.i.m.}$ (%)	1.13 (7.78)
No. of atoms	2243
Protein residues (total/observable)	237/236
R_{work}/R_{free} (%)	11.1/15.8

3.6. Properties of the final data set

Including post-refinement produced a much stronger processing algorithm than that used for the previous solution of CPV17. The initial merge, without post-refinement, achieved an R_{split} of 9.20%, as shown in Table 1, albeit with a large number of rejections. After post-refinement, the R_{split} value was 3.15%, which means that each image has a 12-fold improvement in lowering the R_{split} value, and the number of rejected reflections was reduced. This increase is the squared ratio between the prior R_{split} and the best R_{split} when corrected for number of images, assuming that R_{split} is largely proportional to N, where N is the number of images. Crystallographic refinement statistics for this data set are shown in Table 2. The final resolution cutoff was chosen as the highest resolution shell for which $CC_{1/2}$ was greater than 0.3 (1.46 Å). The increase in reflection number allowed anisotropic *B*-factor refinement, resulting in crystallographic *R* factors of R_{work} = 11.1% and R_{free} = 15.8% using *PHENIX* (Adams *et al.*, 2010).

The improved quality of the high-resolution information can be seen in the R_{work} and R_{free} values for the high-resolution shell. In the post-refined data set, the highest resolution shell, 1.485–1.46 Å, had an R_{work} and R_{free} of 29.0 and 33.9%, respectively, whereas the same shell in the non-post-refined data (cycle 0) had an R_{work} and R_{free} of 34.8 and 40.8%, respectively, and the highest resolution shell for the previous 1.75 Å resolution structure solution, 1.81–1.75 Å, had an R_{work} and R_{free} of 43 and 46%, respectively. Thus, the added

quality of the data was strongly reflected in the markedly improved model-refinement statistics.

3.7. Electron-density maps

Electron-density maps with *B*-factor sharpening for the post-refined 1.46 Å resolution data set clearly showed a marked improvement over the 1.75 Å resolution structure and over the 1.46 Å resolution data set before post-refinement (the initial merge), with visibly distinguishable atomic numbers of C, N and O atoms. The peak density for O and N atoms was sufficiently different to resolve the conformations of glutamine and asparagine residues. The difference in density is shown in Fig. 5. Extra water molecules, as well as alternative conformations in the main chain, were added to the atomic model.

This data set was collected using a 1.3 Å wavelength beam, which was very far from the sulfur *K* edge of 5 Å, and would result in an f'' of 0.42 electrons per S atom. This would produce an anomalous signal of an average 0.5% change in amplitude, compared with a background intensity $R_{\rm split}$ of 3.15%. Nevertheless, an anomalous difference Fourier map calculated on the post-refined data clearly highlighted 12 of the 13 ordered S atoms present in the asymmetric unit of the crystal, as shown in Fig. 6. The only missing sulfur anomalous density was for Cys142 (not shown), which was in two alternative conformers each with half occupancy. One conformation formed a disulfide bond with Cys155. The mean anomalous peak for the 12 anomalous S atoms was 3.87σ, with a maximum of 5.43σ for Met182. The anomalous difference map was generated using *ANODE* (Thorn & Sheldrick, 2011) to produce cleaner maps, although using a basic estimate of the native phase shifted by 90° was also sufficient to observe all of these peaks. Seven of the 12 sulfur anomalous densities were at least 4.2σ at the peak and within half an angstrom of the corresponding S atom.

This anomalous density was compared with the anomalous density achieved by the initial merge with no application of post-refinement. In this case there was barely any density associated with methionine S atoms. The closest peaks to S atoms occurred 0.7–2.8 Å away from the sulfurs at an average peak density of 2.62σ, and the highest peak associated with sulfur was the 244th highest peak in the entire electron-density map. Hence, these were lost within noise and were not specifically associated with S atoms, whereas the application of post-refinement reduced the errors sufficiently to observe the anomalous density, as demonstrated above.

3.8. Effect of the number of images

In order to observe the effect of number of images on $R_{\rm split}$, the above procedure was repeated on subsets of 200 to 4000 images (see Fig. 7). Although the low-image-number subsets were hampered by poor initial reference data sets, they still recovered a substantial amount of information. Using the Monte Carlo relationship $R_{\rm split} \propto 1/N^{1/2}$ (Kirian *et al.*, 2010) and the $R_{\rm split}$ of 24.2% for 500 images, we would expect an $R_{\rm split}$ of 6.36% for 7225 images. However, owing to the post-refinement method using information across many images to improve the parameters for individual images, the $R_{\rm split}$

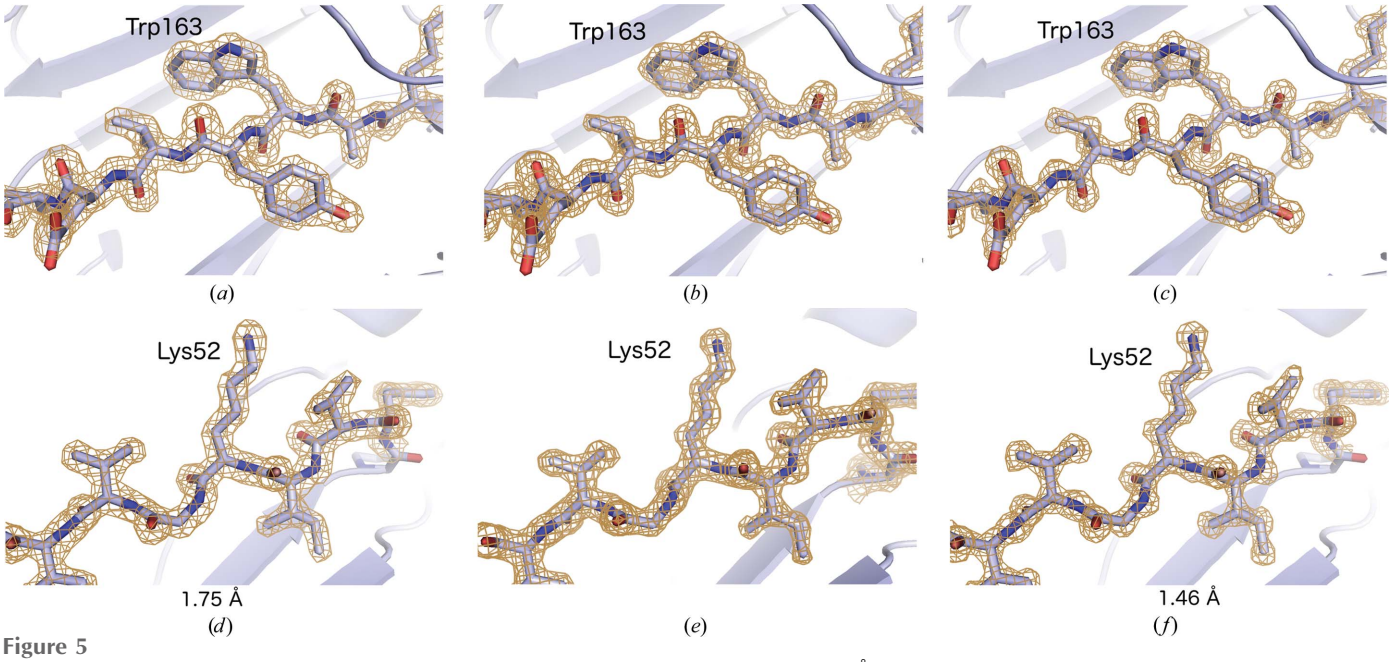


Figure 5
Electron-density map ($2mF_{\rm o} - DF_{\rm c}$) from reflections associated with PDB entry 4s1l to 1.75 Å resolution (Ginn *et al.*, 2015) associated with Trp163 (*a*) and Lys52 (*d*), their corresponding electron density in the initial merge for the higher resolution data set, (*b*) and (*e*), and the final presented 1.46 Å resolution structure, (*c*) and (*f*), at a σ of 1.5. The Lys52 side chain shows the prominence of the H atoms on the methylene groups, which are not pronounced on the perpendicular profile. The N, C and O atoms are distinguishable compared with the 1.75 Å resolution structure. The high-resolution information beyond 1.75 Å and post-refinement of this data set appear to have separate sequential improvements on the quality of the electron density compared with the 4s1l structure.

achieved was actually 3.70% (prior to the final merge), thereby recovering information beyond the effects of simply adding more data to reduce nonsystematic errors and essentially producing a data quality equivalent to what might be expected from a data set three times the size.

4. Discussion

The choice of an appropriately shaped integration box markedly improved the high-resolution information for this set of images when analysed using the methodology described in Ginn *et al.* (2015), but errors remained at high resolution, most likely owing to residual errors in the detector geometry and the comparative weakness of the data, as well as subtle aberrations enhanced at high resolution caused by errors in detector distance and wavelength. Nonetheless, the resolution of the analysis was limited by the geometry of the detector. Outlier rejection was instrumental in lowering the $R_{\rm split}$ owing to the presence of multiple lattices on a large number of images. In order to reject more rigorously, the multiple lattices on a given image should all be indexed, as has been performed to generate a synchrotron structure for crystals of this polyhedrin (Gildea *et al.*, 2014) and as has been carried out on images of thermolysin (Hattne *et al.*, 2014), to identify neighbouring or overlapping reflections from separate lattices. Processing multiple lattices would also increase the number of observations for assembling the data set.

Overall, the post-refinement algorithm has resulted in a marked improvement in the extraction of information from the diffraction patterns. The electron-density maps were significantly improved and the weak anomalous signal from the S atoms was revealed, despite the experiment being performed far from the optimal wavelength. We hope that this will lead the way to allowing routine crystal structure solution by sulfur SAD on serial femtosecond lasers, especially when data are collected at an optimum wavelength for sulfur SAD

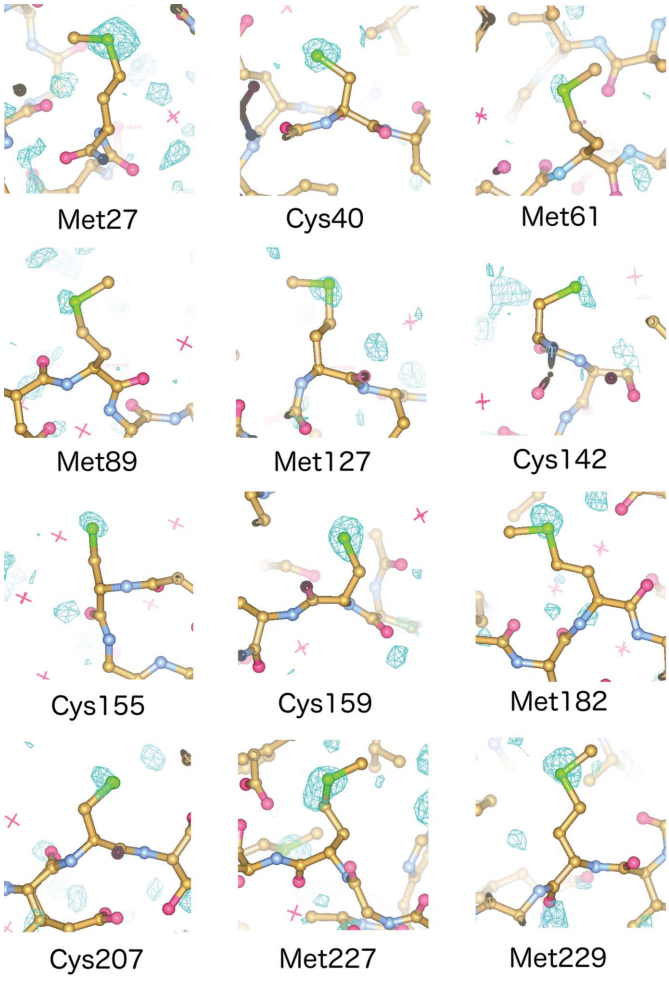


Figure 6
Anomalous signal contoured at 2.4σ highlighted for 12 S atoms as labelled. The highest density is 5.43σ (Met182) and the lowest is 2.53σ (Met70 and Met89). The average peak σ is 3.87.

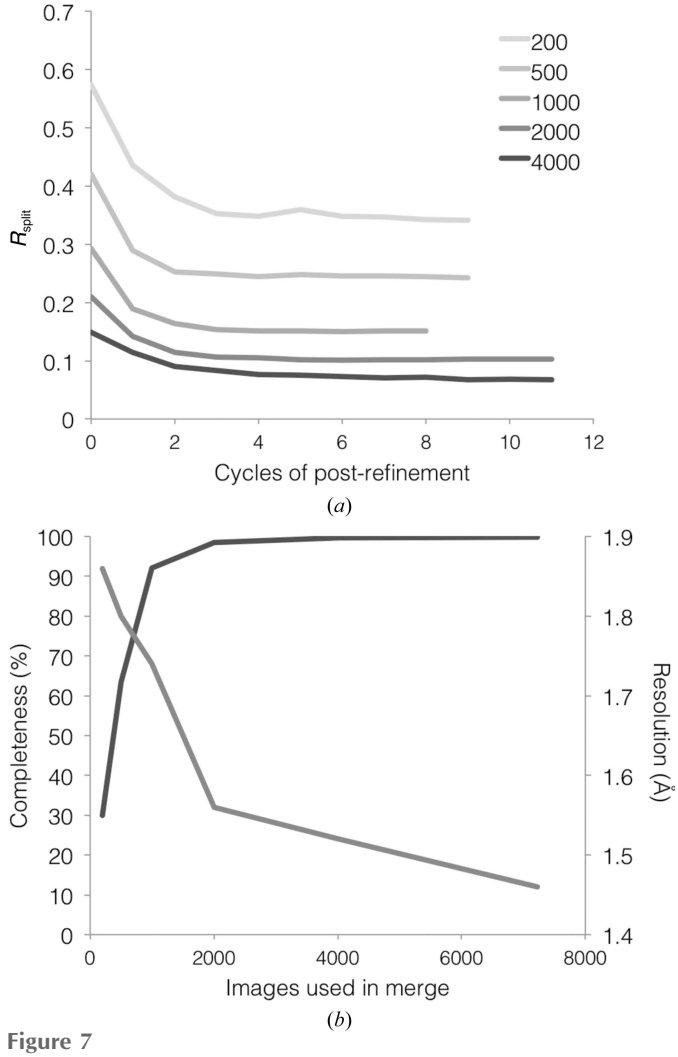


Figure 7
(*a*) Calculated $R_{\rm split}$ values for subsets of images between 200 and 4000 in number, using a final correlation merge threshold of 0.9. (*b*) Maximum resolution (light grey line) denotes the first resolution shell beyond which $CC_{1/2}$ falls below 0.3. Completeness (dark grey line) is calculated from low resolution to the resolution cutoff. This suggests that even 1000 crystals will give a useful data set comparable to that reported for data collected at a synchrotron from a similar number of crystals (Gildea *et al.*, 2014).

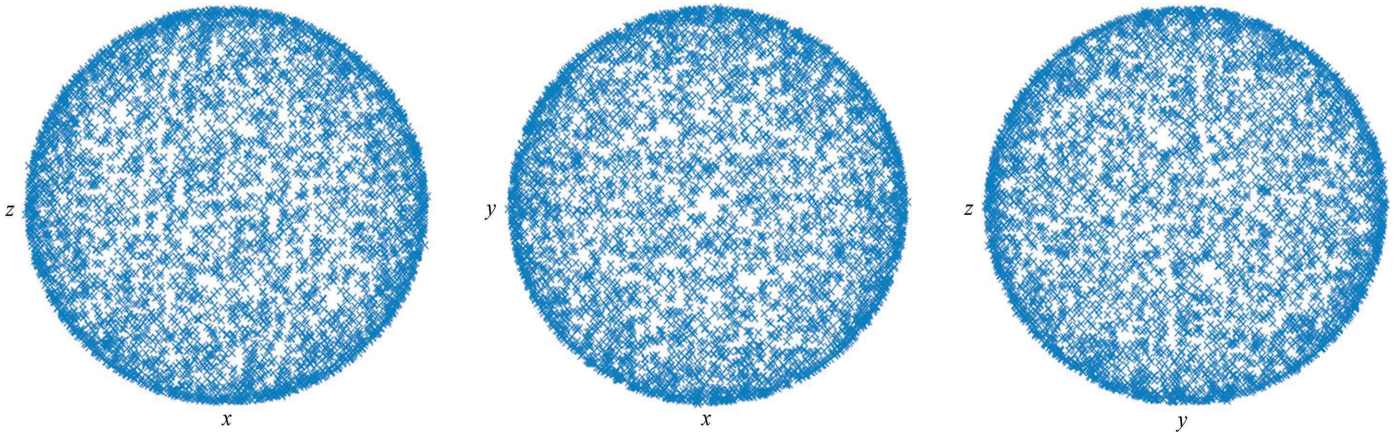


Figure 8
Orientation matrices for 1500 crystals were applied to a (1, 1, 1) coordinate and the point was plotted against the corresponding perpendicular axes, including symmetry-related points owing to cubic space-group symmetry. This shows an even sampling of crystal orientations.

studies. By markedly reducing the number of diffraction images required to provide a high-quality data set, such methods should also open up the method to more challenging problems and increase the efficiency with which the scarce resource of XFEL beam time can be used. We are aiming shortly to fold the code written for this purpose into *cctbx.xfel* and *DIALS* (Waterman *et al.*, 2013).

Note added in proof: An alternate approach to post-refinement of X-ray free-electron laser data has recently been published by Uervirojnangkoorn *et al.* (2015).

APPENDIX *A*
Crystal orientation

Analysis of the orientation matrices for 1500 crystals showed no preferred orientations, as shown in Fig. 8.

Acknowledgements

DIS was supported by the Medical Research Council, grant G1000099. HMG was supported by the Wellcome Trust (studentship 075491/04). ASB, JH and NKS were supported by US National Institutes of Health grants GM095887 and GM102520. Portions of this research were carried out at the Linac Coherent Light Source (LCLS) at the SLAC National Accelerator Laboratory. LCLS is an Office of Science User Facility operated for the US Department of Energy Office of Science by Stanford University. We are very grateful for the expert support for the operation of the sample injector provided by the group of Ilme Schlichting (Max Planck Institute for Medical Research, Heidelberg, Germany), in particular Sabine Botha, R. Bruce Doak and Robert L. Shoeman. We are very grateful to the LCLS–CXI staff, Marc Messerschmidt, Sébastien Boutet, Garth Williams and Dan Deponte. Admin support was received from the Wellcome Trust, grant 090532/Z/09/Z. This is a contribution from the Oxford Instruct Centre. Code is available on request, and will be folded into *cctbx.xfel* and *DIALS*.

References

Adams, P. D. *et al.* (2010). *Acta Cryst.* D**66**, 213–221.
Barends, T. R. *et al.* (2013). *Acta Cryst.* D**69**, 838–842.
Boutet, S. *et al.* (2012). *Science*, **337**, 362–364.
Brehm, W. & Diederichs, K. (2014). *Acta Cryst.* D**70**, 101–109.
Decker, F.-J. (1995). *AIP Conf. Proc.* **333**, 550.
Emma, P. *et al.* (2010). *Nature Photonics*, **4**, 641–647.
Gildea, R. J., Waterman, D. G., Parkhurst, J. M., Axford, D., Sutton, G., Stuart, D. I., Sauter, N. K., Evans, G. & Winter, G. (2014). *Acta Cryst.* D**70**, 2652–2666.
Ginn, H. M., Messerschmidt, M., Ji, X., Zhang, H., Axford, D., Winter, G., Brewster, A. S., Hattne, J., Wagner, A., Grimes, J. M., Sauter, N. K., Sutton, G. & Stuart, D. I. (2015). *Nature Commun.* **6**, 6435.
Hattne, J. *et al.* (2014). *Nature Methods*, **11**, 545–548.
Kabsch, W. (2014). *Acta Cryst.* D**70**, 2204–2216.
Kern, J. *et al.* (2013). *Science*, **340**, 491–495.
Kern, J. *et al.* (2014). *Nature Commun*, **5**, 4371.
Kirian, R. A., Wang, X., Weierstall, U., Schmidt, K. E., Spence, J. C., Hunter, M., Fromme, P., White, T., Chapman, H. N. & Holton, J. (2010). *Opt. Express*, **18**, 5713–5723.
Kirian, R. A., White, T. A., Holton, J. M., Chapman, H. N., Fromme, P., Barty, A., Lomb, L., Aquila, A., Maia, F. R. N. C., Martin, A. V., Fromme, R., Wang, X., Hunter, M. S., Schmidt, K. E. & Spence, J. C. H. (2011). *Acta Cryst.* A**67**, 131–140.
Liu, W. *et al.* (2013). *Science*, **342**, 1521–1524.
Redecke, L. *et al.* (2013). *Science*, **339**, 227–230.
Rossmann, M. G., Leslie, A. G. W., Abdel-Meguid, S. S. & Tsukihara, T. (1979). *J. Appl. Cryst.* **12**, 570–581.
Sauter, N. K. (2015). *J. Synchrotron Rad.* **22**, 239–248.
Sauter, N. K., Hattne, J., Brewster, A. S., Echols, N., Zwart, P. H. & Adams, P. D. (2014). *Acta Cryst.* D**70**, 3299–3309.
Sauter, N. K., Hattne, J., Grosse-Kunstleve, R. W. & Echols, N. (2013). *Acta Cryst.* D**69**, 1274–1282.
Tenboer, J. *et al.* (2014). *Science*, **346**, 1242–1246.
Thorn, A. & Sheldrick, G. M. (2011). *J. Appl. Cryst.* **44**, 1285–1287.
Uervirojnangkoorn, M., Zeldin, O. B., Lyubimov, A. Y., Hattne, J., Brewster, A. S., Sauter, N. K., Brunger, A. T. & Weis, W. I. (2015). *eLife*, **4**, e05421.
Waterman, D. G., Winter, G., Parkhurst, J. M., Fuentes-Montero, L., Hattne, J., Brewster, A., Sauter, N. K. & Evans, G. (2013). *CCP4 Newsl. Protein Crystallogr.* **49**, 16–19.
White, T. A. (2014). *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **369**, 20130330.
White, T. A., Kirian, R. A., Martin, A. V., Aquila, A., Nass, K., Barty, A. & Chapman, H. N. (2012). *J. Appl. Cryst.* **45**, 335–341.
Winkler, F. K., Schutt, C. E. & Harrison, S. C. (1979). *Acta Cryst.* A**35**, 901–911.

Structure-based energetics of protein interfaces guides foot-and-mouth disease virus vaccine design

Abhay Kotecha[1,8], Julian Seago[2,8], Katherine Scott[3], Alison Burman[2], Silvia Loureiro[4], Jingshan Ren[1], Claudine Porta[1,2], Helen M Ginn[1], Terry Jackson[2], Eva Perez-Martin[2], C Alistair Siebert[1], Guntram Paul[5], Juha T Huiskonen[1], Ian M Jones[4], Robert M Esnouf[1], Elizabeth E Fry[1], Francois F Maree[3,6], Bryan Charleston[2] & David I Stuart[1,7]

© 2015 Nature America, Inc. All rights reserved.

Virus capsids are primed for disassembly, yet capsid integrity is key to generating a protective immune response. Foot-and-mouth disease virus (FMDV) capsids comprise identical pentameric protein subunits held together by tenuous noncovalent interactions and are often unstable. Chemically inactivated or recombinant empty capsids, which could form the basis of future vaccines, are even less stable than live virus. Here we devised a computational method to assess the relative stability of protein-protein interfaces and used it to design improved candidate vaccines for two poorly stable, but globally important, serotypes of FMDV: O and SAT2. We used a restrained molecular dynamics strategy to rank mutations predicted to strengthen the pentamer interfaces and applied the results to produce stabilized capsids. Structural analyses and stability assays confirmed the predictions, and vaccinated animals generated improved neutralizing-antibody responses to stabilized particles compared to parental viruses and wild-type capsids.

FMDV is a small-RNA virus whose icosahedral capsid contains 60 copies each of proteins VP1–VP4. FMDV is responsible for a contagious, economically devastating livestock disease endemic in many developing regions of Asia, Africa and South America; developed countries are also at risk of sporadic but serious outbreaks (for example, in the UK in 2001, in Japan in 2010 and in Egypt in 2012 (refs. 1–3)). In areas where FMDV is endemic, disease control is carried out predominantly by vaccination. There are seven serotypes of FMDV (A, O, C, Asia, and Southern African Territories (SAT) 1, 2 and 3)[4], and each contains multiple and constantly evolving subserotype strains. A vaccine against one serotype does not protect against other serotypes or subtypes within a serotype, thus necessitating the continued development of new vaccine strains. The majority of vaccines produced worldwide are type O, whereas SAT2 is the most prevalent serotype in sub-Saharan Africa[5].

FMDV serotypes differ markedly in their capsid stability: A and Asia-1 are relatively stable, whereas O and SAT viruses are more sensitive to heat and pH[6]. Commercial vaccines are produced by chemical inactivation of virus particles, which renders them even less stable, and above 30 °C they rapidly convert into immunogenically incompetent pentameric subunits[6,7]. Thus, vaccine integrity requires expensive, difficult-to-maintain cold chains, and effective protection requires frequent immunizations[5,8]. Recombinant empty capsids show promise as future vaccines because they overcome many of the disadvantages of preparing vaccines from live virus; however, they are even less stable than virus particles[9].

Knowledge of the three-dimensional structure of viral capsids should allow for rational engineering of their properties. Although four different serotypes of FMDV have been analyzed in atomic detail[10–14], few attempts have been made to engineer thermostable capsids[15,16]. The pentameric subassemblies are highly stable, so stabilization of the interpentamer interactions should be sufficient to enhance capsid stability and the effectiveness of vaccine preparations. The introduction of a disulfide bond across the interface between adjacent pentamers has enabled the production of thermostable FMDV empty capsids for an A serotype (A22 Iraq)[17]. However, covalent stabilization may not be compatible with virus viability or with the structure of all viruses[16].

Here, we developed a molecular dynamics (MD)-based strategy for the evaluation of mutations designed to increase capsid stability by increased noncovalent interactions and applied it to the less stable serotypes O and SAT2, and the more stable A viruses. Structural analyses and stability assays confirmed that the method is predictive of increased stability. Furthermore, animals vaccinated with candidate mutant viruses generated improved neutralizing-antibody responses, thus demonstrating potential value for such stabilized vaccine antigens.

RESULTS

FMDV capsids dissociate into pentameric subassemblies

Inactivated FMDV capsids dissociate into pentameric subunits at pH <7.0, at temperatures >30 °C (refs. 1,4,6,15) and after prolonged storage at 4 °C (**Fig. 1a**,**b**). We examined the interface between pentamers for opportunities to stabilize these interpentamer

[1]Division of Structural Biology, University of Oxford, Oxford, UK. [2]Pirbright Institute, Pirbright, UK. [3]Transboundary Animal Disease Programme, Agricultural Research Council-Onderstepoort Veterinary Institute, Onderstepoort, South Africa. [4]Animal and Microbial Sciences, University of Reading, Reading, UK. [5]Merck Sharp & Dohme Animal Health, Cologne, Germany. [6]Department of Microbiology and Plant Pathology, University of Pretoria, Pretoria, South Africa. [7]Diamond Light Source, Didcot, UK. [8]These authors contributed equally to this work. Correspondence should be addressed to D.I.S. (dave@strubi.ox.ac.uk), F.F.M. (mareef@arc.agric.za) or B.C. (bryan.charleston@pirbright.ac.uk).

Received 6 May; accepted 26 August; published online 21 September 2015; doi:10.1038/nsmb.3096

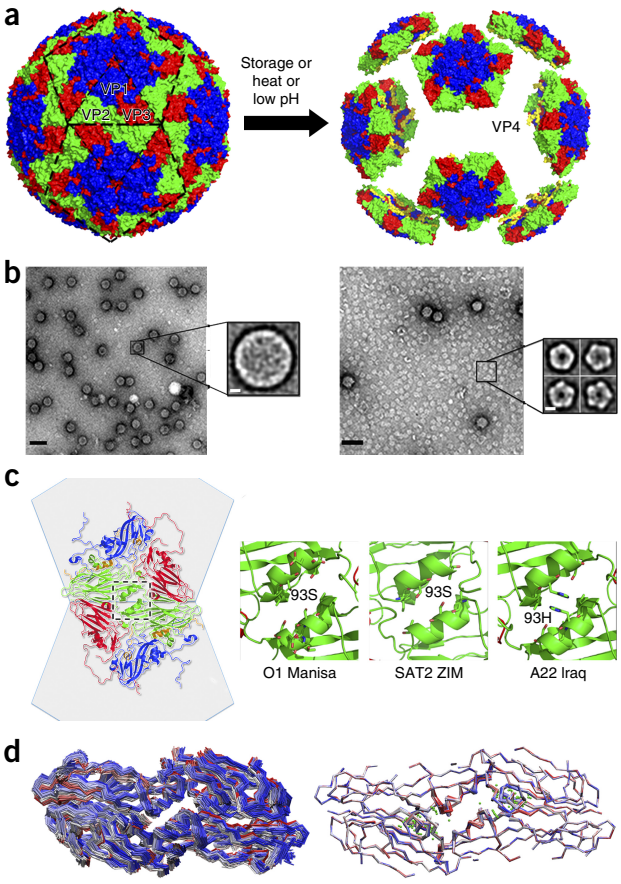


Figure 1 Dissociation of FMDV capsids into pentameric assemblies and design of the models used for MD simulations. (**a**) Surface representation of atomic models of the intact FMDV serotype-O capsid (left, PDB 1BBT) and of its dissociation into 12 pentameric assemblies upon storage, heating or lowering of pH (right). Blue, VP1; green, VP2; red, VP3; yellow, VP4. (**b**) Negative-stain EM images of the inactivated O1M wild-type capsids soon after purification (left, capsids intact with only a few detectable pentamers) or after 10 d of storage at 4 °C (right, 80% dissociated into pentamers). Scale bars, 50 nm. Class averages of the intact capsid or pentamers are shown in zoom view (each calculated from 700 particles with EMAN2 (ref. 38)). Scale bars in zoom view, 5 nm. (**c**) Left, cartoon representation of the atomic structure of the O1M model, showing two icosahedral protomers forming an interpentameric interface. Blue, VP1; green, VP2; red, VP3; yellow, VP4. A truncated model was generated by trimming the protomers to include VP2 and VP3 atoms within 13 Å of the interface, as shown in **d**. Right, residues on two-fold symmetry–related helices for O, A and SAT serotypes. (**d**) 1.5-ns MD trajectory. Left, large r.m.s. deviation from the starting structure for an unrestrained model (average r.m.s. deviation 1.83 Å^2). Right, restrained MD trajectory, with dummy atoms shown in green and placed at the midpoints of the interprotomeric interface to define the restraints. The beginning of the trajectory is shown in red, the middle in white and the end in blue.

contacts by enhancing noncovalent interactions. From four different serotypes[10,11,13,14,18], we identified 85 amino acids contributing to this interface: 57 from VP2 and 28 from VP3. VP2 shows ~80% sequence identity between different serotypes, and we selected nonconserved interface residues inaccessible to antibodies and thus unlikely to affect antigenicity (**Supplementary Fig. 1**) for the introduction of potentially stabilizing interactions (for each residue, a range of mutants was possible). We focused on the region close to the icosahedral two-fold axis because enterovirus uncoating is known to initiate at this point, and disorder of the N terminus of VP2 adjacent to the three-fold axis makes this alternative region less reliable for modeling[19–21].

We chose O1/Manisa/Turkey/69 (O1M), a field strain widely used in current vaccines, and SAT2/ZIM/7/83 (SAT2) as our target antigens, and A22 Iraq 24/64 (A22) as a stable reference strain[17]. In the absence of high-resolution structures, O1BFS (PDB 1BBT)[22] was used as the model for O1M (98% sequence identity in capsid protein VP2), and SAT1 (PDB 2WZR)[14] was used as the model for SAT2 (78% identity in VP2). On the basis of visual inspection of the structures, we generated a panel of 12 potential stabilizing mutants within VP2 for O1M, four of which we also chose for SAT2, and chose the two most promising for A22 (**Table 1**). We modeled these mutants by *in silico* mutagenesis (in COOT[23]) and localized energy minimization.

Table 1 Binding free energy of capsid-stabilizing mutants for O1M, SAT2 and A22

Mutation	O1 Manisa ΔΔ*G* kcal/mol	SAT2 ZIM ΔΔ*G* kcal/mol	A22 Iraq ΔΔ*G* kcal/mol
VP2 93H	−7.7	−5.3	NA
VP2 93Y	−11.8	−12.2	−7.7
VP2 93F	−13.8	−13.2	−10.3
VP2 98F	−10.3	−5.6	–
VP2 97I	−6.8	–	–
VP2 93V	−7.1	–	–
VP2 93L	−1.9	–	–
VP2 93I	−7.2	–	–
VP2 93M	−2.8	–	–
VP2 93W	−9.5	–	–
VP2 97Q	−5.5	–	–
VP2 90N	−3.5	–	–
Negative controls			
VP2 60G	+33.7		
VP2 60L	+24.0		
VP2 57E	+17.7		
VP2 57L	+12.3		

NA, not applicable.

MD simulations and binding free-energy calculations

We devised a new MD-simulation protocol for rapid evaluation of the relative stability potential of mutant viruses. Atomistic dynamics is usually applied to complete biological ensembles[24,25]; however, in the context of small, well-defined changes in a complex virus particle, this approach has a poor signal-to-noise ratio and is expensive computationally. We therefore decreased the complexity of the simulation, initially to two interfacial protomers (**Fig. 1c**), which required a week of processing on eight central-processing-unit cores (AMD Opteron 6174 at 2.2 GHz). Nonetheless, artifacts made it difficult to detect small changes in the binding free energy of the interface. Therefore we devised a protocol in which the model was further trimmed to atoms within 13 Å of the interface, and positional restraints that increased with distance from the interface were applied. To enable this, we placed nonphysical 'dummy' atoms midway between pairs of interacting atoms along the contact interface centered on the icosahedral two-fold axis of symmetry. During simulations, we defined restraints such that atoms within 10 Å of any dummy atom were allowed to move freely, whereas atoms that were more remote, and the dummy atoms themselves, were tightly restrained (**Fig. 1d**). These restraints considerably simplified the simulation (~24 h computation with the same AMD central processing units) and dramatically reduced the noise by 'freezing' unchanged parts of the structure, including the disconnected atoms at the periphery of

© 2015 Nature America, Inc. All rights reserved.

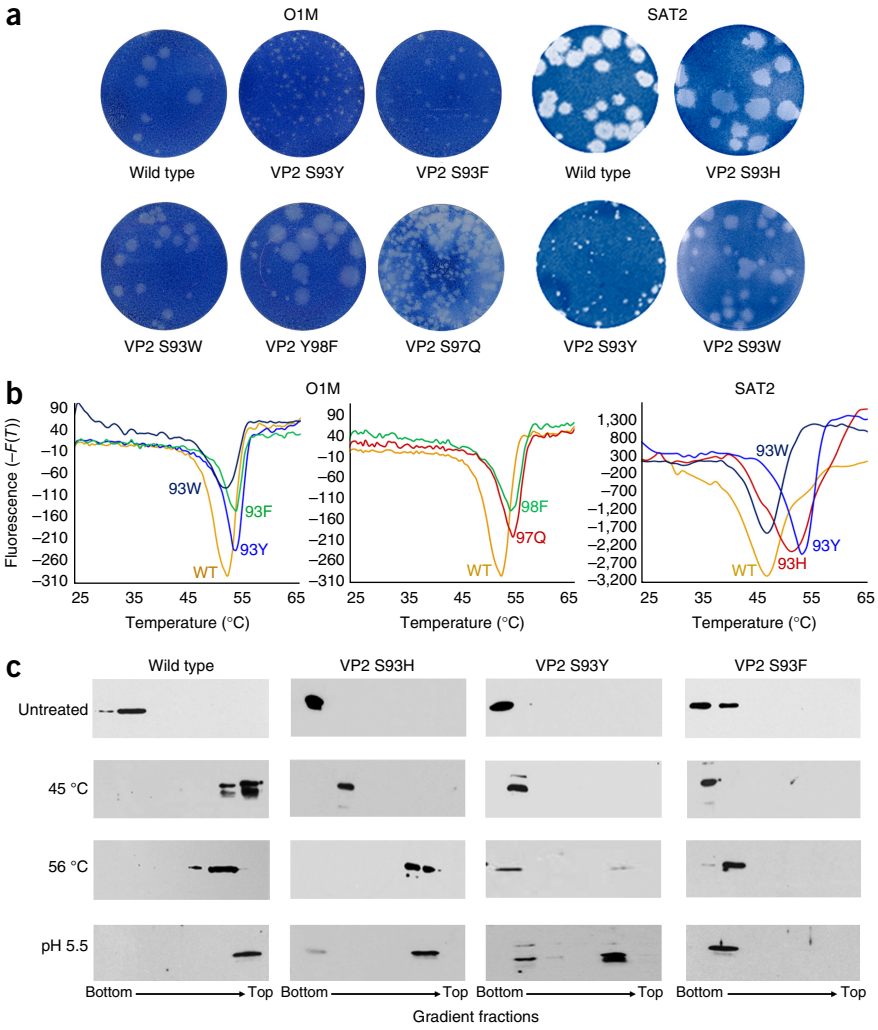


Figure 2 Growth characteristics and stability of the engineered capsids. (**a**) Plaques formed in BHK-21 cells by wild-type and mutant O1M and SAT2 viruses. The patterns of CPE correlated with plaque size. (**b**) Fluorescence assay measuring thermostability of infectious O1M viruses (left and middle) and inactivated SAT2 viruses (right) at pH 7.5 (ref. 29). *F*, fluorescence; *T*, temperature. Wild-type O1M and SAT2 dissociated at 52.0 °C and 47.0 °C, respectively. Mutant S93Y dissociated at 53.5 °C for both O1M and SAT2. O1M mutants S93F, S97Q and Y98F dissociated at 53.5 °C, 54.0 °C and 53.5 °C, respectively. Mutant S93H (SAT2) dissociated at 51.0 °C. S93W dissociation was similar to that of parental virus in both O1M and SAT2. (**c**) Thermostability and pH stability of the recombinant empty capsids assayed by western blotting after sedimentation through 15–45% sucrose density gradients, in which dissociated capsids remain at the top of the gradient, and intact capsids migrate to the bottom. Fractions were probed with anti–FMDV O1M polyclonal antibodies recognizing VP1 (refs. 39,40) (ELISA data are shown in **Supplementary Fig. 3**; uncropped gels are shown in **Supplementary Data Set 1**).

the model. We tested a set of negative controls with the O1M model to validate the protocol, whereby a salt-bridge or hydrogen-bond network was deliberately disrupted (**Table 1**); these mutations resulted in positive $\Delta\Delta G$ values. In one simulation, VP2 Arg60, which forms the salt bridge with the VP2 Glu212 on the opposite protomer, was substituted with either glycine or leucine; as expected, in both cases the interface was destabilized. Similarly, when Gln57, which makes a network of hydrogen bonds across the interface, was substituted with glutamate or leucine, an unstable interface was generated. However, such dynamics-based methods cannot model additional covalent interactions, such as disulfide bonds.

Putative mutations that confer increased stability

The results of the free-energy calculations are shown in **Table 1**. Residue 93 of VP2 is part of an α-helix adjacent to the icosahedral two-fold axis and is largely conserved within but not between serotypes (**Supplementary Fig. 1** and **Fig. 1c**). The relatively stable A serotype has a histidine at this position, whose imidazole ring forms a hydrophobic stacking interaction with the corresponding symmetry-related moiety. Simulation results suggested that this virus could be further stabilized by mutation to tyrosine or, more preferably, to phenylalanine. For O1M and SAT2, the simulations suggested that stability would be enhanced by mutation of the equivalent serine to histidine (S93H) and would be further improved by mutation to tyrosine, whose side chain was predicted to hydrogen-bond to the backbone of VP2 89. Nonetheless, a phenylalanine was predicted to confer even greater stability. Substitution with a bulky tryptophan was less effective, as were residues with nonaromatic hydrophobic side chains (valine, leucine, isoleucine and methionine) (**Table 1**). Some VP2 mutations away from the two-fold axis produced stabilization (S97Q, S97I and to a lesser extent V90N), and Y98F was predicted to be most effective.

Mutant viruses are infectious and stable

The effects of the amino acid mutations were first tested in recombinant virus. Transcripts were electroporated or transfected into BHK-21 cells to generate infectious stocks. These stocks were then used to infect ZZ-R127 goat epithelial cells, which are highly susceptible to FMDV infection[26]. Subsequent passages in BHK-21 cells, which are commonly used in the production of FMD vaccines, were possible because the SAT2 virus was cell-culture adapted, whereas two VP3 mutations (H56R and D60G) were introduced in O1M, thus enabling cell entry via heparan sulfate[27]. These mutations, remote from the pentamer interface, did not affect the particle stability (**Supplementary Fig. 1**) and are often present in vaccine strains. All recombinant viruses were sequenced after four passages, and the results revealed no changes or obvious subpopulations. All produced clear, consistent, cytopathic effects (CPE); however, we observed differences between viruses. The patterns of CPE and plaque sizes varied markedly, and in general stabilized viruses yielded substantially smaller plaques than the parental viruses (**Fig. 2a**). However, the yields of wild-type and mutant viruses were broadly similar (**Supplementary Table 1**). In addition to generating recombinant viruses, we made recombinant empty capsids containing the same amino acid substitutions by using mammalian- and insect-cell expression systems described previously[17,28].

We determined capsid stability by fluorescence thermal stability assays[29] at pH 7.5. For O1M, wild-type virus dissociated at 52.0 °C,

© 2015 Nature America, Inc. All rights reserved.

Table 2 Data collection and refinement statistics

	O1M VP2 S93Y[a]	A22 VP2 H93F[b]
Data collection		
Space group	*I*23	*I*222
Cell dimensions		
a, *b*, *c* (Å)	344.08, 344.08, 344.08	327.6, 341.3 363.6
α, β, γ (°)	90, 90, 90	90, 90, 90
Resolution (Å)[c]	46.0–3.5 (3.56–3.5)	50.0–2.4 (2.49 –2.4)
R_{merge}	0.46	0.27
$I / \sigma I$	1.7 (0.76)	1.23 (0.5)
Completeness (%)	77.7 (71.1)	45.5 (39.8)
Redundancy	1.5 (1.4)	1.4 (1.3)
Refinement		
Resolution (Å)	46.0–3.5	50.0–2.4
No. reflections	65,915	347,924
R_{work} / R_{free}	36.35 / –	20.6 / 21.2
No. atoms		
Protein	5,149	5,192
Ligand/ion	–	–
Water	0	196
B factors		
Protein	49.2	27.0
Ligand/ion	–	–
Water	–	30.6
r.m.s. deviations		
Bond lengths (Å)	0.011	0.013
Bond angles (°)	1.816	1.716

[a]The data for O1M VP2 S93Y were collected from 14 different crystals. [b]The data for A22 VP2 H83F were collected from 19 different crystals. [c]Values in parentheses are for highest-resolution shell.

and S93W dissociated similarly, whereas S93Y, S93F and Y98F dissociated at 53.5 °C. Inactivated SAT2 followed a similar pattern: wild-type and S93W capsids dissociated at a similar temperature (47.0 °C), whereas S93H and S93Y dissociated at 51.0 °C and 53.5 °C, respectively (**Fig. 2b** and **Supplementary Table 2**). At lower pH (6.5), wild-type O1M capsids dissociated at 31.0 °C, whereas mutant capsids were more stable; S93W dissociated at 34.0 °C, S98F and S97Q at 35.0 °C, and S93Y and S93F at 38.0 °C (**Supplementary Fig. 1**).

We assessed the stability of inactivated wild-type and S93Y O1M particles by negative-stain EM. Immediately after purification, ~90% of wild-type capsids were intact compared to almost 100% for S93Y (on the basis of the mean particle count within five randomly selected areas of the grid). After 10 d of storage at 4 °C, only ~20% of the wild-type capsids remained intact, compared to 90% for the mutant (**Supplementary Fig. 2**). Finally, we incubated samples at 37 °C for 5, 10 or 20 min and analyzed them by EM. Remarkably, only ~5% of the wild-type virions remained intact, compared to ~60% for the mutant after 20-min incubation at 37 °C (**Supplementary Fig. 2**). An ELISA on infectious O1M virus produced similar results. We incubated wild-type or S93Y particles at 37 °C for up to 72 h and analyzed samples with llama antibodies (VHH M170, which recognizes intact capsids, and VHH M3, which recognizes pentamers)[30]. As expected, more than 50% of the S93Y capsids were intact after 72 h versus only 15% of the wild type (**Supplementary Fig. 2**). In conclusion, the mutant viruses were consistently more stable at elevated temperatures and at low pH, and the results correlated well with the MD simulations.

Stability of recombinant empty capsids

For empty particles, the RNA fluorescence assay could not be used, so we performed analysis by fractionation on sucrose density gradients and subsequent western blotting. We heated purified wild-type and mutant (S93H, S93Y and S93F) O1M capsids at 45 °C for 1 h. The mutant empty capsids sedimented similarly to intact particles, whereas wild-type empty capsid material stayed near the top of the gradient, thus confirming dissociation (**Fig. 2c** and **Supplementary Fig. 3**). In a more extreme test, we incubated the mutant capsids at 56 °C for 2 h before analysis. Mutant S93H was unstable, S93Y was partially stable (~25% dissociated particles as estimated visually), and S93F was completely stable. As visualized by EM, the peak fraction of heat-treated S93F revealed mostly intact capsids, with few pentamers (**Supplementary Fig. 4**). We obtained similar results with A22 empty capsids: the H93F mutant remained intact after 2 h at 56 °C, whereas wild-type capsids completely dissociated (**Supplementary Fig. 5**). To study the effect of storage at moderate temperature on empty capsids, we incubated O1M wild type and

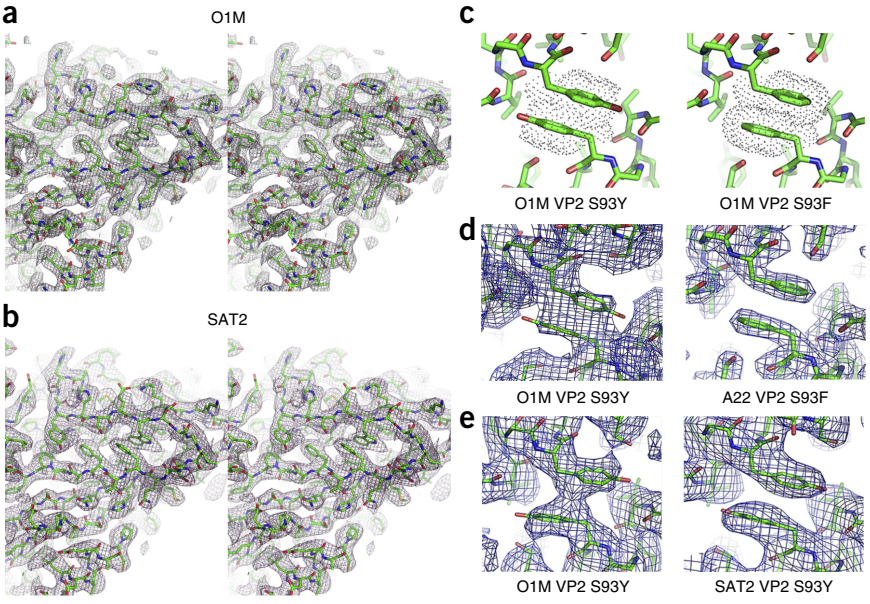


Figure 3 Structural analysis of stable engineered capsids. The structures of engineered mutants O1M S93Y, A22 H93F and SAT2 S93Y were determined by X-ray crystallography and cryo-EM. (**a**,**b**) Stereo views of the density from cryo-EM reconstructions for O1M S93Y at 3.2 Å (**a**) and SAT2 S93Y at 3.5 Å (**b**). The quality of the cryo-EM density maps allowed unambiguous fitting and refinement of the major capsid proteins. (**c**) Predicted structures showing the stacking interactions of tyrosine and phenylalanine at position 93 of O1M VP2 proteins on the two-fold symmetry axis. (**d**) X-ray structures of O1M S93Y and A22 H93F at 3.5-Å and 2.4-Å resolution, respectively. The electron density map clearly showed a stacking interaction of phenylalanine side chains for A22 H93F. Tyrosine density for O1M S93Y was observable but less well defined because of the twinning of O1M crystals. (**e**) Cryo-EM maps showing the density for S93Y for O1M and SAT2 at 3.2-Å and 3.5-Å resolution, respectively.

© 2015 Nature America, Inc. All rights reserved.

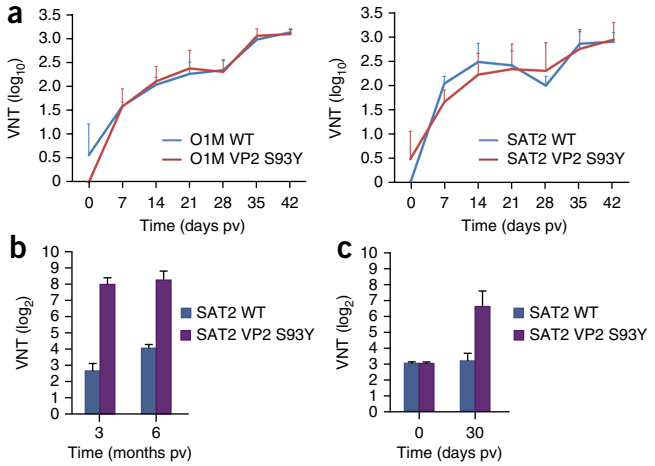


Figure 4 Immunogenicity of inactivated wild-type and stabilized viruses. (**a**) Mean virus neutralizing-antibody titer (VNT (log_{10})) of calves vaccinated with purified inactivated virus of the O1M (left) and SAT2 (right) serotypes. Groups of four calves were vaccinated with either wild-type or mutant S93Y at days 0 and 28; blood samples were assayed at 0, 7, 14, 21, 28, 35 and 42 days pv, error bars, s.d. (**b**) VNTs of guinea pigs to assess immunogenicity of inactivated particles after 1 month of long-term storage. Before immunization, formulated vaccines were stored for 1 month at 4 °C. Two groups of ten guinea pigs were each immunized with either wild-type or S93Y SAT2 antigen, and VNTs were assessed at 3 and 6 months pv. Error bars, s.d. (**c**) VNTs of guinea pigs to assess immunogenicity of inactivated particles after 6 months of long-term storage, showing protective VNTs for only stabilized antigen. Equivalent aliquots of wild-type and stabilized SAT2 vaccines were stored for 6 months at 4 °C before inoculation of two groups of ten guinea pigs. Error bars, s.d.

S93F mutant for 24 h or 96 h at 37 °C. Wild-type capsids dissociated, whereas S93F capsids remained intact (**Supplementary Fig. 6**). To determine pH stability, we incubated O1M capsids at pH 5.5 for 15 min. Wild-type capsids dissociated completely, whereas S93H and S93Y capsids were partially stable (25% remained intact), and S93F remained completely intact (**Fig. 2c**).

Structure analysis of stabilized capsids

We determined structures of the stabilized capsids by X-ray crystallography and cryo-EM. SAT2 crystals did not diffract to high resolution; however, we solved structures of the mutant A22 empty capsid and mutant O1M virus at 2.4-Å and 3.5-Å resolution, respectively, using *in situ* room-temperature X-ray crystallography[17,31]. The O1M crystals were merohedrally twinned, and we developed a protocol to deconvolute the twinned reflections[32] (H.M.G., unpublished data). Refinement using strict noncrystallographic symmetry and real-space averaging yielded reliable maps and models for both structures (**Table 2**). The refined atomic models were very similar to those of the respective natural empty particles and parental viruses[10–12,17], except that, as predicted, the aromatic side chains of VP2 residue 93 stacked at the two-fold axes.

Owing to the lower quality of the O1M analysis, and our inability to obtain diffraction-quality crystals of SAT2 particles, we determined the structures of inactivated O1M S93Y and SAT2 S93Y by cryo-EM (Online Methods). 8,267 particles of O1M and 8,156 particles of SAT2 yielded structures at 3.2- and 3.5-Å resolution (Fourier shell correlation of 0.143), respectively (**Fig. 3a**,**b**, **Supplementary Fig. 7** and **Supplementary Table 3**). We compared the predicted and experimental stacking of the tyrosine and phenylalanine side chains around the two-fold axis in the X-ray and cryo-EM structures (**Fig. 3c**–**e**).

Immunogenicity of stabilized mutants

To compare the immunogenicity of wild-type and stabilized viruses, we vaccinated calves with purified inactivated wild-type or mutant (S93Y) virus of both the O1M and SAT2 serotypes. We maintained inactivated antigen at 4 °C and purified it immediately before use. We measured serum virus neutralization antibody titers (VNTs) at different times postvaccination (pv) (**Fig. 4a**). In all animals, we observed high VNTs on day 28 at the time of boost immunization (mean antibody titers >2 log_{10}, which is considered protective[33]); titers were substantially increased (>3 log_{10}) by day 42 pv.

Because mutant and wild-type antigens had equivalent immunogenic properties when administered directly after purification, we compared their immunogenicity after long-term storage. For this, we stored inactivated wild-type and mutant S93Y FMDV SAT2 particles at doses of 10 μg each, formulated as described above, for 1 month at 4 °C and used them to vaccinate two groups of ten guinea pigs. We assessed VNTs at 3 and 6 months pv (**Fig. 4b**). At 3 months pv, we observed substantially higher VNTs in animals vaccinated with the stabilized SAT2 antigen (>8 log_2 considered protective[34]) than in animals vaccinated with the wild-type SAT2 antigen (<3 log_2). At 6 months pv, the group mean VNT in animals immunized with the stabilized SAT2 antigen remained at >8 log_2, whereas wild type–immunized animals' antibody titers remained below protective levels. Thus, the stabilized vaccine performed markedly better than the wild-type vaccine.

In a further iteration, we stored equivalent aliquots of wild-type and stabilized vaccines for 6 months at 4 °C and again used them to immunize two groups of ten guinea pigs. No animal had a VNT titer at day 0, but by day 30 pv, although the mean VNT titer of the group inoculated with wild-type vaccine remained low (~3 log_2), the mean titer of the group inoculated with stabilized virus was markedly higher and was consistent with protection, >6 log_2 (**Fig. 4c**). Thus, only stabilized particles were able to induce VNT titers, a result consistent with protection after 1 month or 6 months of storage at 4 °C.

DISCUSSION

FMDV vaccines with higher stability have the potential to be more effective. However, very few natural FMDV variants with increased resistance to heat or pH[35,36] have been reported, possibly because overstabilization hinders the release of the viral genome during cell entry. We therefore investigated the rational introduction of mutations to increase thermostability, earlier attempts at which have had only marginal success[16]. We have previously reported a stable disulfide bond at the icosahedral two-fold axis between adjacent pentamers in the A22 virus, thus allowing the production of thermostable and pH-stable recombinant empty particles[17]. Because covalent modification was unlikely to be suitable with engineered viruses and may not be generally applicable, we devised a strategy that allows general mutations conferring noncovalent stabilization to be evaluated *in silico*. We exploited highly restrained atomistic MD simulations focused on the appropriate part of the complete virion to efficiently and reliably calculate the changes in binding free energy for interactions between adjacent pentamers within the FMDV capsid, thus allowing screening and ranking of candidate mutations. We found that a single-residue mutation generating hydrophobic stacking of aromatic side chains at the two-fold axis between adjacent pentamers (position 93 of VP2) was effective in stabilizing the particle, results echoing those from studies in enteroviruses that identify the place at which the particle opens up to start uncoating[20,21]. The predicted degree of stabilization correlated with

© 2015 Nature America, Inc. All rights reserved.

hydrophobicity, thus with O1M, S93H ($\Delta\Delta G$ = −7.7 kcal/mol) < S93Y ($\Delta\Delta G$ = −11.8 kcal/mol) < S93F ($\Delta\Delta G$ = −13.8 kcal/mol). Similarly, when the equivalent (histidine) residue of the stable A serotype, A22 Iraq, was substituted by tyrosine ($\Delta\Delta G$ = −7.7 kcal/mol) or phenylalanine ($\Delta\Delta G$ = −10.3 kcal/mol), the stability of the interface was further increased.

Engineering the stabilization mutations decreased the plaque sizes for both O1M and SAT2 viruses, possibly by delaying viral genome release. Both infectious viruses and recombinant empty capsids carrying these mutations were indeed thermostable and pH stable, and the degree of stabilization mirrored the *in silico* calculations. This noncovalent stabilization of two particularly unstable serotypes produced particles of similar stability to that of covalently cross-linked A22 particles[17]. Formulated vaccines from inactivated wild-type and stabilized O and SAT2 viruses produced equivalent protective levels of neutralizing antibodies in cattle, thus demonstrating the immunogenicity of the engineered particles. The stabilized inactivated vaccine antigen was also intact after 96 h at 37 °C and is expected to be more tolerant of suboptimal cold-chain performance; goals for the stability of SAT2 have been described by Anderson *et al.*[37]. Furthermore, after storage for 1 or 6 months at 4 °C, stabilized particles produced substantially higher VNTs in guinea pigs than did wild-type particles, commensurately with an increase in vaccine shelf life.

In conclusion, we have devised a new restrained MD protocol to drive structure-based design of stabilized FMDV viruses and empty capsids. Such new effective FMDV vaccines would, in the context of empty capsids, eliminate the considerable risks involved in producing vaccine from infectious virus. This strategy may be relevant to many other picornaviruses that are responsible for causing a wide range of mild–to–life-threatening diseases affecting humans and other animals.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. Coordinates and structure factors for the inactivated O1M and A22 mutant particles have been deposited in the Protein Data Bank under accession codes 5DDJ and 5D8A, respectively. The EM structures for O1M and SAT2 mutant particles have been deposited in the Protein Data Bank under accession codes 5AC9 and 5ACA, respectively, and in the EM Data Bank under accession codes 3129 and 3130, respectively.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank the World Reference Laboratory for VNT determination; beamline staff at the Diamond Light Source for assistance; J. Dong and J. Diprose for computing support; P. Afonine and G. Murshudov for advice on Phenix and Refmac; K. Harlos and T. Walter for help with crystallography; D. Goovaerts and E. Rieder for helpful discussions; and B. Haas (Friedrich-Loeffler-Institut) for providing ZZ_R 127 cells. We are grateful to the Wellcome Trust (WT) for a Translation Award to fund this work (grant no. 089755 to B.C., E.E.F., T.J. and F.F.M.). T.J. and B.C. were funded by the Biotechnology and Biological Sciences Research Council Institute Strategic Programme on Livestock Viral Diseases at The Pirbright Institute. The Oxford Particle Imaging Centre electron microscopy facility was founded by a WT Joint Infrastructure Fund award (060208/Z/00/Z to D.I.S.) and is supported by a WT equipment grant (093305/Z/10/Z to K. Grünewald). The WT, UK Medical Research Council (MRC) and Biotechnology and Biology Research Council also support the National EM facility, which provided the K2 detector. B.C. and D.I.S. are supported as Jenner investigators, J.R. and A.K. are WT supported, and E.E.F. and D.I.S. are supported by the UK MRC (grant no. G100099 to D.I.S.). The work of the WT Centre in Oxford is supported by the WT core award 090532/Z/09/Z.

AUTHOR CONTRIBUTIONS

A.K., E.E.F., R.M.E. and D.I.S. developed MD-simulation protocols; A.K., J.S., K.S., A.B., S.L. and C.P. prepared samples; J.R., H.M.G., J.T.H., E.P.-M., G.P., C.A.S., F.F.M. and E.E.F. assisted in research; A.K., J.S., F.F.M., E.E.F., T.J., I.M.J., R.M.E., D.I.S. and B.C. designed the study; all authors analyzed data; and A.K., J.S., E.E.F., B.C. and D.I.S. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Grubman, M.J. & Baxt, B. Foot-and-mouth disease. *Clin. Microbiol. Rev.* **17**, 465–493 (2004).
2. Muroga, N. *et al.* The 2010 foot-and-mouth disease epidemic in Japan. *J. Vet. Med. Sci.* **74**, 399–404 (2012).
3. Kandeil, A. *et al.* Characterization of the recent outbreak of foot-and-mouth disease virus serotype SAT2 in Egypt. *Arch. Virol.* **158**, 619–627 (2013).
4. Bachrach, H.L. Foot-and-mouth disease. *Annu. Rev. Microbiol.* **22**, 201–244 (1968).
5. Doel, T.R. FMD vaccines. *Virus Res.* **91**, 81–99 (2003).
6. Doel, T.R. & Baccarini, P.J. Thermal stability of foot-and-mouth disease virus. *Arch. Virol.* **70**, 21–32 (1981).
7. Doel, T.R. & Chong, W.K.T. Comparative immunogenicity of 146S, 75S and 12S particles of foot-and-mouth disease virus. *Arch. Virol.* **73**, 185–191 (1982).
8. Hall, M.D., Knowles, N.J., Wadsworth, J., Rambaut, A. & Woolhouse, M.E.J. Reconstructing geographical movements and host species transitions of foot-and-mouth disease virus serotype SAT 2. *MBio* **4**, e00591–13 (2013).
9. Basavappa, R. *et al.* Role and mechanism of the maturation cleavage of VP0 in poliovirus assembly: structure of the empty capsid assembly intermediate at 2.9 Å resolution. *Protein Sci.* **3**, 1651–1669 (1994).
10. Acharya, R. *et al.* The three-dimensional structure of foot-and-mouth disease virus at 2.9 A resolution. *Nature* **337**, 709–716 (1989).
11. Curry, S. *et al.* Perturbations in the surface structure of A22 Iraq foot-and-mouth disease virus accompanying coupled changes in host cell specificity and antigenicity. *Structure* **4**, 135–145 (1996).
12. Curry, S. *et al.* Dissecting the roles of VP0 cleavage and RNA packaging in picornavirus capsid stabilization: the structure of empty capsids of foot-and-mouth disease virus. *J. Virol.* **71**, 9743–9752 (1997).
13. Lea, S. *et al.* The structure and antigenicity of a type C foot-and-mouth disease virus. *Structure* **2**, 123–139 (1994).
14. Reeve, R. *et al.* Sequence-based prediction for vaccine strain selection and identification of antigenic variability in foot-and-mouth disease virus. *PLoS Comput. Biol.* **6**, e1001027 (2010).
15. Ellard, F.M., Drew, J., Blakemore, W.E., Stuart, D.I. & King, A.M. Evidence for the role of His-142 of protein 1C in the acid-induced disassembly of foot-and-mouth disease virus capsids. *J. Gen. Virol.* **80**, 1911–1918 (1999).
16. Mateo, R., Luna, E., Rincón, V. & Mateu, M.G. Engineering viable foot-and-mouth disease viruses with increased thermostability as a step in the development of improved vaccines. *J. Virol.* **82**, 12232–12240 (2008).
17. Porta, C. *et al.* Rational engineering of recombinant picornavirus capsids to produce safe, protective vaccine antigen. *PLoS Pathog.* **9**, e1003255 (2013).
18. Berman, H.M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
19. Wang, X. *et al.* A sensor-adaptor mechanism for enterovirus uncoating from structures of EV71. *Nat. Struct. Mol. Biol.* **19**, 424–429 (2012).
20. Ren, J. *et al.* Picornavirus uncoating intermediate captured in atomic detail. *Nat. Commun.* **4**, 1929 (2013).
21. Butan, C., Filman, D.J. & Hogle, J.M. Cryo-electron microscopy reconstruction shows poliovirus 135S particles poised for membrane interaction and RNA release. *J. Virol.* **88**, 1758–1770 (2014).
22. Fry, E., Acharya, R. & Stuart, D. Methods used in the structure determination of foot-and-mouth disease virus. *Acta Crystallogr. A* **49**, 45–55 (1993).
23. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D Biol. Crystallogr.* **60**, 2126–2132 (2004).
24. Karplus, M. & McCammon, J.A. Molecular dynamics simulations of biomolecules. *Nat. Struct. Biol.* **9**, 646–652 (2002).
25. Klepeis, J.L., Lindorff-Larsen, K., Dror, R.O. & Shaw, D.E. Long-timescale molecular dynamics simulations of protein structure and function. *Curr. Opin. Struct. Biol.* **19**, 120–127 (2009).
26. Jackson, T., Sheppard, D., Denyer, M., Blakemore, W. & King, A.M. The epithelial integrin αvβ6 is a receptor for foot-and-mouth disease virus. *J. Virol.* **74**, 4949–4956 (2000).
27. Fry, E.E. *et al.* The structure and function of a foot-and-mouth disease virus-oligosaccharide receptor complex. *EMBO J.* **18**, 543–554 (1999).
28. Porta, C. *et al.* Efficient production of foot-and-mouth disease virus empty capsids in insect cells following down regulation of 3C protease activity. *J. Virol. Methods* **187**, 406–412 (2013).

© 2015 Nature America, Inc. All rights reserved.

29. Walter, T.S. *et al.* A plate-based high-throughput assay for virus stability and vaccine formulation. *J. Virol. Methods* **185**, 166–170 (2012).
30. Harmsen, M.M., Fijten, H.P., Westra, D.F. & Coco-Martin, J.M. Effect of thiomersal on dissociation of intact (146S) foot-and-mouth disease virions into 12S particles as assessed by novel ELISAs specific for either 146S or 12S particles. *Vaccine* **29**, 2682–2690 (2011).
31. Axford, D. *et al. In situ* macromolecular crystallography using microbeams. *Acta Crystallogr. D Biol. Crystallogr.* **68**, 592–600 (2012).
32. Lea, S. & Stuart, D. Deconvolution of fully overlapped reflections from crystals of foot-and-mouth disease virus O1 G67. *Acta Crystallogr. D Biol. Crystallogr.* **51**, 160–167 (1995).
33. Barnett, P.V., Statham, R.J., Vosloo, W. & Haydon, D.T. Foot-and-mouth disease vaccine potency testing: determination and statistical validation of a model using a serological approach. *Vaccine* **21**, 3240–3248 (2003).
34. Bolwell, C., Parry, N.R. & Rowlands, D.J. Comparison between *in vitro* neutralization titres and *in vivo* protection against homologous and heterologous challenge induced by vaccines prepared from two serologically distinct variants of foot-and-mouth disease virus, serotype A22. *J. Gen. Virol.* **73**, 727–731 (1992).
35. Martín-Acebes, M.A., Rincon, V., Armas-Portela, R., Mateu, M.G. & Sobrino, F. A single amino acid substitution in the capsid of foot-and-mouth disease virus can increase acid lability and confer resistance to acid-dependent uncoating inhibition. *J. Virol.* **84**, 2902–2912 (2010).
36. Mateo, R., Luna, E. & Mateu, M.G. Thermostable variants are not generally represented in foot-and-mouth disease virus quasispecies. *J. Gen. Virol.* **88**, 859–864 (2007).
37. Anderson, E.C., Doughty, W.J. & Spooner, P.R. Variation in the thermal stability of isolates of foot-and-mouth disease type SAT 2 and its significance in the selection of vaccine strains. *J. Comp. Pathol.* **92**, 495–507 (1982).
38. Tang, G. *et al.* EMAN2: an extensible image processing suite for electron microscopy. *J. Struct. Biol.* **157**, 38–46 (2007).
39. Ferris, N.P. *et al.* Utility of recombinant integrin αvβ6 as a capture reagent in immunoassays for the diagnosis of foot-and-mouth disease. *J. Virol. Methods* **127**, 69–79 (2005).
40. Ferris, N.P. & Donaldson, A.I. Serological response of guinea pigs to inactivated 146S antigens of foot and mouth disease virus after single or repeated inoculations. *Rev. Sci. Tech. Off. Int. Epiz* **3**, 563–574 (1984).

© 2015 Nature America, Inc. All rights reserved.

ONLINE METHODS

Molecular dynamics methods. These are described in the **Supplementary Note**.

Generation of infectious recombinant cDNA. Infectious FMDV O1K/O1 Manisa (O1M) chimeric clones were constructed with reverse genetics. Briefly, cDNA encoding the VP2, VP3, VP1 and 2A proteins was removed from a derivative of the pT7S3 O1K infectious clone, termed pT7SBmuts, leaving cDNA encoding the Lpro, VP4, 2B, 2C, 3A, 3B, 3C and 3D proteins[41]. The removed cDNA was replaced with the corresponding O1M cDNA from pGEM9zf subclones, encoding the wild-type or a specific amino acid substitution within the VP2 structural protein[42] (UKG/35/2001; GenBank AJ539141). The encoded amino acid substitutions within VP2 were generated with the QuikChange Lightning Mutagenesis kit (Agilent Technologies) according to the manufacturer's instructions. For SAT2, the leader-P1-2A coding region of pSAT2 (ref. 43) was cloned into pBS (Stratagene) with EcoRI and XmaI restriction-enzyme sites. Site-directed mutagenesis was performed with the QuikChange Site-Directed Mutagenesis kit (Stratagene). The pSAT2/pBS clones containing the desired mutations, as verified by nucleotide sequencing, were digested with SspI and XmaI to recover the mutant leader-P1-2A region, which was cloned back into the pSAT2 genetic backbone. All mutations were verified by sequencing.

Preparation of infectious RNA, electroporation and transfection. RNA was transcribed from the infectious clones with the MEGAscript T7 kit (Invitrogen). Infectious RNA was electroporated with a Bio-Rad Gene Pulser (two pulses at 0.75 kV and 25 mF) or transfected into BHK-21 cells (clone 13). After 24 h, the cells were frozen and then thawed in their growth medium. After clarification by centrifugation, the supernatant containing the initial virus stock (termed 'passage 0', P0) was harvested. Fetal goat tongue (ZZ-R 127) or BHK-21 cells were subsequently used to passage the viruses[44]. Cells were infected for 24 h between passages. The yields of wild-type and mutant viruses were similar. BHK-21 cells (clone 13) have been maintained in the Pirbright Institute since 1965. ZZ_R 127 cells were a gift from B. Haas (Friedrich-Loeffler-Institut, (FLI)). All cell cultures were free of mycoplasma and bovine viral diarrhea virus.

Genome amplification and sequencing. Total RNA was extracted with TRIzol (Invitrogen), and the viral RNA genome was reverse-transcribed and PCR-amplified with a One-Step RT-PCR kit (Qiagen). Sequencing reactions were performed with an aliquot of the purified PCR product and the BIG Dye Terminator v3.1 cycle sequencing kit (Applied Biosystems).

Plaque assay. Confluent monolayers of goat epithelial cells or BHK-21 cells were infected with serial dilutions of virus stocks, overlaid with indubiose (Pall Life Sciences) and incubated for 24–48 h at 37 °C. Cells were fixed and stained (4% formaldehyde in PBS containing methylene blue), and the overlay was removed[26]. The diameters of 50 parental (wild-type) plaques were measured with a microscope eyepiece micrometer. These were averaged, defined as 100%, and compared against 50 mutant plaque diameters.

Virus titers. Virus titers were estimated by plaque assays. For O1M, confluent monolayers of BHK-21 cells were infected with serial dilutions of FMDV virus stocks in triplicate, overlaid with indubiose and incubated for 24–48 h at 37 °C. The cells were then fixed and stained (4% formaldehyde in PBS containing methylene blue) before removal of the overlay[26]. For SAT2, BHK-21 cells were infected for 1 h; this was followed by the addition of 2 ml tragacanth overlay[45]. After incubation at 37 °C for 48 h, the infected monolayers were stained with 1% (w/v) methylene blue in 10% (v/v) ethanol and 10% (v/v) formaldehyde in PBS, pH 7.4.

Virus inactivation. Chemical inactivation followed standard operating procedures (SOPs) in line with disease security regulations at The Pirbright Institute. Upon CPE, the cell-culture medium, containing the virus, was harvested and clarified by centrifugation at 2,095*g* for 30 min at 4 °C. Clarified supernatant was inactivated by two consecutive incubations with binary ethyleneimine (BEI). A 10-ml aliquot of 0.1 M BEI solution was added to every 1,000 ml of virus harvested (to yield a final BEI concentration of 0.001 M); this was followed by incubation for 24 h at 37 °C. During incubation, the virus/BEI solutions were mixed by periodic inversion. Innocuity tests were performed by the World Reference Laboratory (WRL) at The Pirbright Institute.

Generation of recombinant empty particles. Expression of recombinant empty capsids used two different systems: O1M capsids were produced with a vaccinia virus (VV) mammalian-cell expression system, and A22 capsids were produced with a baculovirus (BV) insect-cell expression system according to previously described methodology[17]. Briefly, for VV expression cassettes containing synthetic cDNA encoding VP0, VP3 and VP1 and the 2A and 3C nonstructural proteins of wild-type FMDV were synthesized (GeneArt) and cloned into a VV vector, pBG200. Recombinant viruses were generated by transfecting the pBG200 vector into simian kidney-derived CV-1 cells that had been infected with VV. Recombinant VVs (VV-FMDV) were selected by plaque assay and verified by PCR with FMDV-specific primers. FMDV empty capsids were amplified in RK13 or HEK293 cells after dual infection with a recombinant VV expressing T7 polymerase (VV-T7) as previously described[17]. For insect-cell expression, the pTri-EX–derived plasmid pOPINE[46] was used to generate recombinant BVs encoding the FMDV proteins. The pOPINE vector containing the FMDV coding sequence and the AcMNPV (*Autographa californica* nucleopolyhedrovirus) FlashBacGold DNA (from Oxford Expression Technologies) were transfected into Sf9 insect cells (*Spodoptera frugiperda*) with Lipofectin. Recombinant AcMNPV was harvested from supernatant 5 d after infection (P0 virus stock) and amplified by infection of Sf9 cells at 50% confluency with 200 µl recombinant virus per 175 cm^2 flask and collection of the supernatant after 5 d (P1 virus stock). Recombinant empty capsids were expressed by infection of Sf9 cells ($1–2 \times 10^6$/ml) with 1/10 volume of the P1 BV stock. Mutants were generated by PCR mutagenesis (QuikChange Lightning Mutagenesis Kit, Agilent Technologies) according to the manufacturer's instructions. CV-1, RK13 and HEK 293 cells were sourced from the European Collection of Cell Cultures (ECACC). All cell cultures were free of mycoplasma and bovine viral diarrhea virus. Sf9 insect cells were obtained from Sigma-Aldrich.

Purification of viruses. Frozen lysates containing inactivated antigens were thawed on ice and clarified by centrifugation at 3,500*g* for 30 min at 4 °C. Virus was then precipitated by the addition of either 30% w/v ammonium sulfate or 8% w/v PEG 6000 by incubation at 4 °C overnight. Precipitated virus was harvested by centrifugation at 3,500*g* for 1 h at 4 °C and resuspended in 50 mM HEPES (4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid), pH 7.5, containing 200 mM NaCl and 1% v/v NP-40, treated with 0.1% RNase A for 30 min on ice and again clarified by centrifugation at 3,500*g* for 1 h at 4 °C to remove insoluble particulates. Capsids were pelleted over a 3-ml, 30% sucrose cushion (in 50 mM HEPES, pH 7.5, containing 200 mM NaCl) at 105,000*g* for 2.5 h at 4 °C. Pellets were resuspended in HEPES buffer containing 0.5% (v/v) NP-40 overnight at 4 °C, clarified by centrifugation (16,000*g* for 10 min at 4 °C) and purified over a 15–45% sucrose gradient. The gradient was layered with the clarified supernatant containing virus and fractionated by centrifugation at 105,000*g* for 3 h at 4 °C. Fractions were analyzed by SDS-PAGE and subsequent Coomassie staining.

Purification of recombinant empty capsids. Mammalian cells dually infected with VV-FMDV and VV-T7 were harvested by centrifugation at 3,500*g* for 30 min at 4 °C, and pellets were resuspended in PBS. Frozen recombinant BV–infected Sf9 cell pellets containing overexpressed empty capsids were thawed and resuspended in 50 mM HEPES, pH 7.5, 200 mM NaCl and 1% NP-40. Cells were lysed by incubation with detergent for 30 min on ice. Lysates were clarified at 10,000*g* for 30 min at 4 °C, and capsids were pelleted over a 3-ml, 30% sucrose cushion by centrifugation at 105,000*g* for 5 h at 12 °C. Thereafter, recombinant capsids were purified as described for inactivated virus and analyzed by SDS-PAGE and subsequent Coomassie staining.

Negative-stain electron microscopy. Aliquots of purified samples were diluted to 0.2 mg/ml and deposited onto glow-discharged, Formvar/carbon-coated copper grids (Electron Microscopy Sciences). After a 30-s incubation, the excess sample was blotted away, and the grids were washed twice with deionized water. Samples were stained with 1% (w/v) uranyl acetate for 45 s, and excess stain was removed by blotting. Grids were examined on a Tecnai T12 transmission electron microscope operated at 80 kV. Images were acquired

© 2015 Nature America, Inc. All rights reserved.

doi:10.1038/nsmb.3096

on a 4,000 × 4,000 high-sensitivity FEI Eagle camera at 67000× magnification, which corresponded to 1.68 Å/pixel sampling of the specimen.

Thermostability and pH-stability assay for particles. Particle stability thermal release assay (PaSTRy)[29] was performed in 96-well PCR plates with an Agilent MX3005 PCR machine. All assays were performed with 300–500 ng of purified virus (3–15 µl) that had been dialyzed to remove most of the sucrose from purification. 15 µl of SYBR Green-II dye (Molecular Probes, Invitrogen) was diluted 1:100 in the same buffer as used for virus purification, and the volume was brought up to 150 µl with buffer. The temperature was ramped up from 25 °C to 95 °C in 0.5 °C increments at intervals of 10 s. SYBR Green-II fluorescence was read with excitation and emission wavelengths of 490 nm and 516 nm, respectively. The release of RNA, and hence the dissociation of capsids, was detected by an increase in fluorescence signal, and the melting temperature was taken as the minimum of the negative first derivative of the fluorescence curve.

ELISA. Wild-type or S93Y particles were incubated at 37 °C for up to 72 h, and their stability was analyzed by double-antibody sandwich ELISA with llama single-domain antibody fragments ((VHH domains), termed VHH M170, which binds intact FMD viral particles and VHH M3, which recognizes pentamer subunits). VHHs were provided by the Central Veterinary Institute of Wageningen UR)[30]. Microtiter-plate wells were coated with VHHs before the addition of serially diluted samples (in VHH buffer: 1% skimmed milk, 0.05% Tween, 0.5 M NaCl, 2.7 mM KCl, 2.8 mM KH_2PO_4, and 8.1 mM Na_2HPO_4, pH 7.4). After incubation for 1 h at 37 °C, biotinylated VHH M170 was then added and incubated for another 1 h at 37 °C. Plates were washed with PBS-Tween (0.05% Tween 20), and bound biotinylated VHH-M170F was visualized by incubation with streptavidin-HRP (Dako) for 1 h at 37 °C; this was followed by addition of *o*-phenylenediamine dihydrochloride (Sigma) for 15 min at room temperature. The reaction was stopped by addition of 1.25 M sulfuric acid and quantified at 492 nm with a Dynex microplate reader (Dynex Technologies).

Stability assay for recombinant empty capsids. Sucrose gradient–purified samples of FMDV empty capsids were diluted with PBS and incubated in a water bath at either 45 °C for 1 h or 56 °C for 2 h. They were then loaded onto 15–45% sucrose density gradients and sedimented. The gradients were fractionated from the bottom to the top, and an equal volume of saturated ammonium sulfate was added to 300 µl of each fraction to precipitate any empty capsids present. Pelleted precipitates were analyzed by SDS PAGE and western blotting for the VP1 capsid protein, with guinea-pig polyclonal antisera (**Supplementary Fig. 3**). In similar experiments, samples containing empty capsids were diluted with 50 mM sodium acetate buffer, pH 4.6, to produce a pH range of 5.5 to 6.5 (in 0.5 unit increments) and were incubated for 15 min at room temperature. After neutralization with 1/6 volume of 2 M NaOH[47], the empty capsids were loaded onto 15–45% sucrose density gradients and treated as described above before analysis by SDS PAGE and western blotting.

Crystallization, data collection and structure determination. Crystals were grown by the sitting-drop vapor-diffusion method in Crystalquick X plates (Greiner Bio-One) with 100 nl virus/empty capsids plus 100 nl precipitant dispensed with a Cartesian robot as described previously[48]. Microcrystals of inactivated O1M S93Y virus (2.2 mg/ml) and A22 H93F empty capsids (3 mg/ml) grew within 1 week at 294 K with either 1.5 M ammonium sulfate, 100 mM bis-Tris propane, pH 7.0, or 4 M ammonium acetate, 100 mM bis-Tris propane, pH 7.0, respectively. Optimization produced sufficient crystals for structural solution. A 20 × 20 µm^2 or 50 × 50 µm^2 beam (λ = 0.9778 Å; I24 microfocus beamline, Diamond), depending on the crystal size, was used for *in situ*[31] data collection at 294 K on a Pilatus 6 M detector.

The structures of O1M S93Y and A22 H93F were solved by molecular replacement with the coordinates and noncrystallographic symmetry (NCS) operators from O1BFS (PDB 1BBT)[22] and A22 (PDB 4GH4)[11]. Initial phase estimates were obtained by rigid-body refinement with CNS[49]. Twinned data sets were deconvoluted by combining five-fold NCS averaging in real space with rescaling individual reflections according to the twinning operator. This uses redundancy of information in the NCS to recover information lost by the superimposed 90°-rotation twinning operation. Iterative positional and *B*-factor refinement (via CNS) used strict NCS constraints. Phases were further improved by 15-fold (A22) and five-fold (O1M) NCS averaging. The percentages of Ramachandran outliers were 0 and 16.7 for A22 and O1M structures, respectively.

Cryo-EM data collection, structure determination and model building. Aliquots of 4 µl of purified inactivated O1M or SAT2 mutant S93Y virus were added on glow-discharged holey carbon-coated copper grids (C-flat, CF-2/1-2C; Protochips). Grids were blotted for 3 s, in 70% relative humidity, and vitrified in liquid ethane with a plunger device (Vitrobot; FEI). To increase the number of particles in the holes, grids were preincubated with 4 µl of sample for 30 s, and unbound sample was removed by blotting with filter paper.

Images were collected at 300 kV with a Tecnai F30 'Polara' microscope (FEI) equipped with an energy filter (GIF Quantum, Gatan) operating in zero-loss mode (0–20 eV energy selecting slit) and a direct electron detector (K2 Summit, Gatan). Movies (25 frames, each 0.2 s) were recorded at 1.0–3.0 µm underfocus in single-electron counting mode with SerialEM at a calibrated magnification of 37,027×, thus resulting in a pixel size of 1.35 Å. Frames from each movie were aligned and averaged to produce drift-corrected micrographs[50].

Structures were solved with RELION 1.3 according to recommended gold-standard refinement procedures[51], and icosahedral symmetry was applied. Micrographs showing signs of astigmatism or significant drift were discarded and not used for further analysis. Reference-free 2D class averaging was used to discard bad particles. The particle population was further improved by 3D classification. The X-ray structure of native FMDV A22 serotype (PDB 4GH4)[11] was low-pass-filtered to 50 Å and used as an initial template for 3D classification and refinement. A total of 8,267 O1M particles from 210 micrographs and 8,156 SAT2 particles from 321 micrographs were used to solve the final density maps at 3.2-Å and 3.5-Å resolution, respectively, as indicated by Fourier shell correlation and 0.143 cutoff (**Supplementary Fig. 7**). The O1M mutant was fitted in the density map as a rigid body with UCSF Chimera[52]. The fitting was further improved with real-space refinement with COOT[23]. The model of SAT2 was built with the SAT1 structure as a starting model with COOT[23]. REFMAC[53] was used to calculate the difference map, which highlighted the areas where the model was incorrect. Models were further improved by iterative positional and *B*-factor refinement in real space with Phenix[54] and COOT[23] iteratively. Only coordinates were refined; the maps were kept constant. Each round of model optimization was guided by cross-correlation between the map and the model. Statistics of refinement are given in **Supplementary Table 3**.

Vaccination of cattle and guinea pigs with stabilized FMDV vaccines. Four groups of four 100- to 150-kg male Holstein Friesian calves were vaccinated with either inactivated O1M-wt, O1M-(S93Y), SAT2-wt or SAT2-(S93Y) FMDV. Each animal received 15 µg of purified 146S antigen formulated in oil adjuvant ISA206B (SEPPIC) as an intramuscular injection on days 0 and 28 of the study. All 16 animals were bled to collect serum on days 0, 28, 35 and 42, and virus-neutralizing-antibody titers (VNT) were assessed. Animal experimentation was approved by the Pirbright Institute Ethical Review Board under the authority of a Home Office project license (70/7253) in accordance with the Home Office Guidance on the Operation of the Animals (Scientific Procedures) Act 1986 and associated guidelines.

For the guinea-pig experiments, inactivated SAT2 wild-type and stabilized mutant S93Y were formulated as commercial vaccines with oil-based ISA206B (SEPPIC) adjuvant and stored for 1 month at 4 °C. Aliquots containing 10 µg of 146S antigen at the time of formulation of these vaccines were then used to immunize two groups of ten adult inbred guinea pigs, aged 2-4 months. One group received the wild-type vaccine, whereas the other group received the stabilized vaccine. Serum samples were tested on day 0, and at 3 and 6 months pv, bleeds and virus VNT were assessed. The experiments were repeated with the same volumes after the vaccines had been stored for a total of 6 months at 4 °C. Serum samples were collected on days 0 and 30 pv, and the VNTs were analyzed. All studies in the Netherlands complied with the Dutch Experiments on Animals Act, February 5, 1997. All studies complied with the Council Directive 86/609/EEC on the approximation of laws, regulations and administrative provisions of the Member States regarding the protection of animals used for experimental and other scientific purposes.

No statistical method was used to predetermine sample size. Group sizes for the cattle and guinea-pig immunization studies were consistent with those in previously published studies used to determine the induction of antibody titers

© 2015 Nature America, Inc. All rights reserved.

consistent with protection[33,34]. No randomization was performed. The laboratory staff members performing the assessment of antibody titers were blind the animal group allocations.

Titration of neutralizing antibodies. Sera from guinea pigs and cattle were prepared from blood samples. Sera obtained at weekly intervals from day 0 to day 42 pv were assayed by a standard VNT test (OIE, 2012) on porcine kidney IB-RS2 cells with O1 Manisa and SAT2 wild-type viruses. Neutralizing-antibody titers, calculated by the Spearmann-Karber method, were expressed as the last dilution of serum that neutralizes 50% of the homologous virus (100TCID50)[44]. The mean antibody titers of each group of animals was compared with threshold antibody titers that have previously been shown to correlate with protection[33,34].

41. Bøtner, A. *et al.* Capsid proteins from field strains of foot-and-mouth disease virus confer a pathogenic phenotype in cattle on an attenuated, cell-culture-adapted virus. *J. Gen. Virol.* **92**, 1141–1151 (2011).
42. Zibert, A., Maass, G., Strebel, K., Falk, M.M. & Beck, E. Infectious foot-and-mouth disease virus derived from a cloned full-length cDNA. *J. Virol.* **64**, 2467–2473 (1990).
43. van Rensburg, H.G., Henry, T.M. & Mason, P.W. Studies of genetically defined chimeras of a European type A virus and a South African Territories type 2 virus reveal growth determinants for foot-and-mouth disease virus. *J. Gen. Virol.* **85**, 61–68 (2004).
44. Brehm, K.E., Ferris, N.P., Lenk, M., Riebe, R. & Haas, B. Highly sensitive fetal goat tongue cell line for detection and isolation of foot-and-mouth disease virus. *J. Clin. Microbiol.* **47**, 3156–3160 (2009).
45. Rieder, E., Bunch, T., Brown, F. & Mason, P.W. Genetically engineered foot-and-mouth disease viruses with poly(C) tracts of two nucleotides are virulent in mice. *J. Virol.* **67**, 5139–5145 (1993).
46. Berrow, N.S. *et al.* A versatile ligation-independent cloning method suitable for high-throughput expression screening applications. *Nucleic Acids Res.* **35**, e45 (2007).
47. Curry, S. *et al.* Viral RNA modulates the acid sensitivity of foot-and-mouth disease virus capsids. *J. Virol.* **69**, 430–438 (1995).
48. Walter, T.S. *et al.* A procedure for setting up high-throughput nanolitre crystallization experiments: crystallization workflow for initial screening, automated storage, imaging and optimization. *Acta Crystallogr. D Biol. Crystallogr.* **61**, 651–657 (2005).
49. Brünger, A.T. *et al.* Crystallography & NMR system: a new software suite for macromolecular structure determination. *Acta Crystallogr. D Biol. Crystallogr.* **54**, 905–921 (1998).
50. Li, X. *et al.* Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM. *Nat. Methods* **10**, 584–590 (2013).
51. Scheres, S.H.W. RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J. Struct. Biol.* **180**, 519–530 (2012).
52. Pettersen, E.F. *et al.* UCSF Chimera: a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
53. Murshudov, G.N. *et al.* REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallogr. D Biol. Crystallogr.* **67**, 355–367 (2011).
54. Wang, Z. *et al.* An atomic model of brome mosaic virus using direct electron detection and real-space optimization. *Nat. Commun.* **5**, 4808 (2014).

© 2015 Nature America, Inc. All rights reserved.

doi:10.1038/nsmb.3096

Journal of Structural Biology 192 (2015) 88–99

Contents lists available at ScienceDirect

Journal of Structural Biology

journal homepage: www.elsevier.com/locate/yjsbi

Polyhedra structures and the evolution of the insect viruses

Xiaoyun Ji [a,1], Danny Axford [b], Robin Owen [b], Gwyndaf Evans [b], Helen M. Ginn [a], Geoff Sutton [a], David I. Stuart [a,b,*]

[a] Division of Structural Biology, The Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford, Oxfordshire OX3 7BN, United Kingdom
[b] Diamond House, Diamond Light Source, Harwell Science & Innovation Campus, Didcot, Oxfordshire OX11 0DE, United Kingdom

ARTICLE INFO

Article history:
Received 14 May 2015
Received in revised form 7 August 2015
Accepted 17 August 2015
Available online 18 August 2015

Keywords:
Protein microcrystals
Polyhedra
In vivo crystals
Cypovirus
Micro focus crystallography

ABSTRACT

Polyhedra represent an ancient system used by a number of insect viruses to protect virions during long periods of environmental exposure. We present high resolution crystal structures of polyhedra for seven previously uncharacterised types of cypoviruses, four using *ab initio* selenomethionine phasing (two of these required over 100 selenomethionine crystals each). Approximately 80% of residues are structurally equivalent between all polyhedrins (pairwise rmsd $\leqslant$1.5 Å), whilst pairwise sequence identities, based on structural alignment, are as little as 12%. These structures illustrate the effect of 400 million years of evolution on a system where the crystal lattice is the functionally conserved feature in the face of massive sequence variability. The conservation of crystal contacts is maintained across most of the molecular surface, except for a dispensable virus recognition domain. By spreading the contacts over so much of the protein surface the lattice remains robust in the face of many individual changes. Overall these unusual structural constraints seem to have skewed the molecule's evolution so that surface residues are almost as conserved as the internal residues.

© 2015 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

Cypoviruses, members of the family *Reoviridae*, subfamily *Spinareovirus*, possess a single capsid layer with turrets and are commonly embedded in crystalline occlusion bodies called polyhedra, which are formed in the cell cytoplasm and mainly composed of a single virus-encoded protein, polyhedrin (Attoui et al., 2012). Cypoviruses, which generally possess 10 double-stranded RNA genome segments (polyhedrin is encoded by the smallest), are important insect pathogens (Rothman and Myers, 1996), having been isolated from a wide range of insects – Lepidoptera (80%), Diptera (16%), Hymenoptera (3%), Coleoptera and Neuroptera (<1%) (Belloncik and Mori, 1998).

Cypoviruses have been classified into 21 distinct types (Attoui et al., 2012). Within a type the amino acid sequence of polyhedrins are highly conserved, whilst between types there is little conservation. The structure of Bombyx mori cypovirus 1 (CPV1) polyhedrin, determined by micro-focus crystallography, showed the polyhedra to be built from rigid trimeric building blocks, which assemble with the aid of ribonucleoside triphosphates (NTPs) to form a lattice containing very little bulk solvent (Coulibaly et al., 2007). To ensure the delivery of virus particles to the target intestinal cells, allowing the transmission of packages of infectious virus between hosts by oral-faecal routes, polyhedra are disrupted only when they are exposed to the very alkaline pH environment found in insect midguts (Vago and Croissant, 1959). Region-specific ion transport has been found to facilitate the alkalinisation of the anterior midgut lumen with major roles for an H^+ V-ATPase-energised Cl^-/HCO_3^- exchanger and carbonic anhydrase (Boudko et al., 2001), consistent with carbonate ions being used to modulate the pH of the midgut.

We recently reported polyhedrin structures for CPV18, a close relative of CPV1, determined using synchrotron radiation from undisrupted insect cells and from purified polyhedra (Axford et al., 2014), and for CPV17 at room temperature and at 100 K, using an X-ray free electron laser and synchrotron respectively (Ginn et al., 2015b). Here we describe polyhedrin structures for all other types of cypoviruses for which polyhedra could be obtained, giving us a database of nine structures from which we derive rules for crystal formation and dissolution and investigate how preservation of this unusual phenotype (the crystal lattice) has impacted on the variation of the protein over geological time scales.

Abbreviation: SeMet, selenomethionine.

* Corresponding author at: Division of Structural Biology, The Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford, Oxfordshire OX3 7BN, United Kingdom.
E-mail address: dave@strubi.ox.ac.uk (D.I. Stuart).
[1] Present address: Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520, USA.

http://dx.doi.org/10.1016/j.jsb.2015.08.009
1047-8477/© 2015 The Authors. Published by Elsevier Inc.
This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

2. Material and methods

For readability we refer to the cypovirus type and strain by number alone i.e. CPV14 for Lymantria dispar cypovirus 14, as defined in Table S1. For completeness we include CPV1 (Coulibaly et al., 2007) as a reference in our analyses.

2.1. Expression and purification of polyhedra

The cypovirus types studied in this paper are abbreviated as detailed in Supplementary Table S1. cDNA pools of cypovirus genomes for types CPV1, CPV5, CPV14 and CPV15 were a gift from Dr. Shujing Rao (Institute for Animal Health, Pirbright, Surrey, UK). Polyhedrin genes of CPV4, CPV17, CPV18, CPV19 and CPV20 were synthesized by GeneArt (Life Technologies) using sequences obtained from GenBank.

Spodoptera frugiperda (*Sf*9) cells were passaged at 27 °C in Sf-900 II serum-free medium (SFM, Life Technologies) using standard procedures (O'Reilly et al., 1994). The coding regions of genome segment 10 from CPV1, CPV5, CPV14 and CPV15 were amplified and inserted in the multi-cloning site of vector pFastBac1 (Life Technologies). The recombinant plasmids were transformed into competent DH10Bac cells (Life Technologies) to produce bacmid DNA for the generation of recombinant baculoviruses, following the protocol of the Bac-to-Bac® baculovirus expression system (Life Technologies). The polyhedrin genes from CPV4, CPV17, CPV18, CPV19 and CPV20 were sub-cloned into the transfer vector pBac-PAK9 (Clontech). Base domain deletion mutants of CPV1 (CPV1Δ) and CPV4 (CPV4Δ) were produced by three steps of PCR using primers encoding a Gly-Ser-Gly linker (described in (Strachan and Read, 1999)). The recombinant virus was produced by co-transfection of linearized baculovirus DNA and the transfer vector following a standard protocol (Zhao et al., 2003). Polyhedra were expressed and purified as described earlier (Anduleit et al., 2005). SeMet-labelled polyhedra were generated as published previously (Ji et al., 2010) and the incorporation of SeMet was confirmed to be at least 80% using mass spectroscopy. Purified polyhedra were observed and characterised by a Zeiss Axiovert 200 M inverted microscope with fluorescence/phase or DIC (Nomarski) imaging.

2.2. Data collection

Polyhedra were spread on a MicroMesh mount (Mitegen, Ithaca, USA) in a solution of 10 mM HEPES, pH 7.5 and 50% ethylene glycol and flash-frozen in a stream of nitrogen gas at 100 K. X-ray datasets were collected at beamline I24 (Diamond) on either a Rayonix MX300 (Rayonix, Evanston, USA), a PILATUS 6 M or a PILATUS3 6 M detector (DECTRIS, Baden, Switzerland). The typical 10 μm × 10 μm focal spot at the sample position of beamline I24 was reduced with the use of beam defining slits; in the case of CPV17, slits close to sample position and producing a beam size of ~4 μm × ~4 μm with ~1×10^{11} Ph/s; in the case of all other CPVs, slits at the focal point of the first of two pairs of KB mirrors that form the optical configuration of the beamline, producing a beam size of ~6 μm × ~6 μm with ~1.5×10^{11} Ph/s (Evans et al., 2011). Data were processed using various programs including Denzo/Scalepack (Otwinowski et al., 1997), XDS (Kabsch, 1993), Mosflm/Scala (Leslie and Powell, 2007) and XIA2 (Winter, 2010) according to the detector types. The current in the storage ring was maintained around 200–250 mA at Diamond. Each crystal was exposed to 100% transmission of the X-ray beam with a fine sliced oscillation range of 0.05–0.1° for best performance of the PILATUS 6 M detectors. For the smallest crystal only 10–20 frames could be collected. Data from a number of crystals of each cypovirus including native and SeMet-substituted crystals were merged respectively to give a series of datasets with adequate redundancy (Table 1). Details for the data collection strategy of CPV17 at the Linac Coherent Light Source X-FEL (Stanford, U.S.A.) are described in (Ginn et al., 2015a,b).

2.3. Structure determination and refinement

The structure of CPV5 polyhedra was solved using single isomorphous replacement with anomalous scattering (SIRAS). Seven heavy atom sites were found and position refined at 3 Å by SHARP (Fortelle and Bricogne, 1997). SHARP figures of merit were (centric/acentric) 0.37/0.44 (36.0–3.0 Å) and 0.19/0.28 (3.1–3.0 Å). Phasing power (isomorphous/anomalous) was 1.202/0.897. Solvent flattened electron density maps from SOLOMON (Abrahams and Leslie, 1996) and DM (Cowtan, 1994) at 3 Å resolution were used for the initial model building using Coot (Emsley and Cowtan, 2004). Iterative model building and refinement were performed by Coot and BUSTER (Bricogne et al., 2011). Final refinement used PHENIX (Adams et al., 2010) at a resolution of 2.07 Å.

The structures of other cypovirus polyhedrins were solved using a method which combined molecular replacement and SIRAS. The molecular replacement was performed using PHASER (McCoy et al., 2007) with the mainchain atoms of known cypovirus polyhedrin structures as search models against the native data. As we solved additional polyhedrins more starting models could be used for molecular replacement. We also optimised the process by selecting those cypovirus types which bore the highest sequence similarity. The resulting phases were used to calculate anomalous and isomorphous difference Fourier maps using FFT (Read and Schierbeek, 1988; Winn et al., 2011). Potential heavy atom sites were found and positions refined by SHARP (Bricogne et al., 2003) with the phases from molecular replacement not only as a restraint but also helping the refinement. The phases from SIRAS and molecular replacement were finally combined to be the restraint phases in BUSTER (Bricogne et al., 2011) during the process of refinement. Model building was performed in Coot (Emsley and Cowtan, 2004) and final refinement used PHENIX (Adams et al., 2010). All the structures were validated by MolProbity (Chen et al., 2010). Details of the refinement are in Table 2.

Molecule interfaces and oligomerization were evaluated by PISA server at the European Bioinformatics Institute (Krissinel and Henrick, 2007). Three-dimensional structure alignment was performed and analysed by SHP (Stuart et al., 1979). Illustrations were generated by Pymol (Schrodinger, 2010).

2.4. Alkaline disruption

Initially each of eleven polyhedra samples (CPV1, 4, 5, 14, 15, 17, 18, 19, 20, 1Δ and 4Δ) were suspended in 500 μl alkaline buffer (50 mM $NaHCO_3$, pH 10.5) at room temperature. The optical density of the sample was measured at regular time intervals by a spectrophotometer at the wavelength of 600 nm (OD600). The measurements were compared with the starting value to characterize the ratio of dissolved protein to total. In a follow-up experiment aliquots (20 μl) of CPV1, 4, 1Δ and 4Δ polyhedra suspended in water were taken and centrifuged at 20,000×*g* for 1 min. The pellet was dissolved in 50 μl buffer solution (50 mM) at room temperature for 20 min. Five different alkaline buffers were used: Carbonate pH 10.5, Carbonate pH 10, CAPS (N-cyclo hexyl-3-aminopropanesulfonic acid) pH 10, Glycine pH10 and Borate pH 10. The sample was centrifuged again to separate the soluble and insoluble material. The quantities of material dissolved in the alkaline buffers were measured by the absorbance at 280 nm using a ND-1000 Spectrophotometer (NanoDrop).

Table 1
Data collection.

	CPV4 Cell 1	CPV4 Cell 2	CPV5	CPV5 SeMet	CPV14	CPV14 SeMet	CPV15	CPV15 SeMet	CPV17 room temp	CPV17 frozen	CPV18	CPV19	CPV19 SeMet	CPV20	CPV1Δ	CPV4Δ
Number of crystals	18	11	26	128	20	13	40	112	5787	768	21	34	61	51	13	21
Crystal size (μm)	5–7	5–7	3–4	3–4	3–4	3–4	~1.5	~1.5	~1.0	~1.0	4–5	~3	~3	3–4	4–6	5–7
Cell $a = b = c$ (Å)	101.7	104.8	102.3	102.1	102.6	102.8	102.9	103.5	106.1	104.9	102.8	102.7	103.3	103.4	103.9	104.3
Resolution (Å)[a]	29.4–1.72 (1.82–1.72)	28.0–2.00 (2.11–2.00)	51.2–1.61 (1.65–1.61)	72.2–2.07 (2.14–2.07)	72.5–1.91 (2.02–1.91)	36.3–2.24 (2.36–2.24)	51.4–1.82 (1.86–1.82)	51.7–2.48 (2.55–2.48)	28.3–1.75 (1.79–1.75)	74.16–2.20 (2.26–2.20)	72.7–1.48 (1.51–1.48)	51.4–1.50 (1.59–1.50)	42.1–2.03 (2.08–2.03)	51.7–1.82 (1.87–1.82)	42.4–1.88 (1.93–1.88)	33.0–1.71 (1.75–1.71)
Unique reflections	17233 (1580)	11543 (797)	23283 (1726)	10902 (1093)	12952 (1045)	7956 (632)	16418 (1193)	6666 (473)	19096 (1032)	9376 (931)	30415 (2238)	21653 (498)	11977 (870)	16582 (1219)	15023 (1093)	19985 (1319)
R_{merge}	0.199 (0.385)	0.177 (0.330)	0.342 (0.000)**	0.163* (0.387*)	0.199 (0.327)	0.186 (0.363)	0.468 (0.000)**	0.499 (0.000)**		0.665 (0.000)	0.320 (0.000)**	0.368 (0.000)**	0.437 (0.000)**	0.342 (0.000)**	0.226 (0.000)**	0.189 (0.000)**
$I/\sigma I$	9.7 (1.7)	8.1 (1.7)	7.0 (1.4)	7.5 (1.9)	8.4 (1.8)	7.1 (1.6)	4.2 (0.8)	6.4 (1.8)		6.4 (1.4)	7.1 (1.5)	3.8 (0.3)	5.4 (1.0)	5.4 (1.1)	4.8 (1.0)	5.0 (0.6)
Completeness (%)	92.7 (59.0)	88.5 (42.7)	100 (100)	100 (100)	92.3 (52.1)	90.0 (50.3)	99.7 (97.9)	100 (100)	100 (100)	99.9 (100)	100 (99.8)	75.5 (12.2)	98.7 (91.5)	100 (99.7)	98.8 (98.6)	97.3 (88.9)
Redundancy	11.4 (1.9)	7.6 (1.5)	14.8 (10.7)	64.1 (26.1)	7.9 (1.6)	6.7 (1.5)	7.8 (4.5)	19.8 (13.3)	52.4 (11.1)	47.8 (25.4)	12.9 (8.6)	3.8 (1.1)	14.3 (6.4)	10.0 (6.5)	4.3 (4.0)	4.4 (2.7)

CPV17 from (Ginn et al., 2015b).
[a] Values in parentheses are for the highest resolution shell.
* R_{merge} obtained by scaling of two already processed data sets.
** $R_{merge} > = 1$.

Table 2
Refinement statistics.

	CPV4 Cell 1	CPV4 Cell 2	CPV5	CPV5 SeMet	CPV14	CPV15	CPV17 room temp	CPV17 frozen	CPV18	CPV19	CPV20	CPV1Δ	CPV4Δ
Resolution (Å)	29.4–1.72	28.0–2.00	51.2–1.61	36.1–2.07	32.4–1.91	42.0–1.82	28.30–1.75	74.16–2.20	72.7–1.48	41.9–1.51	51.7–1.83	42.4–1.88	27.9–1.71
No. of reflections	17233	11542	23282	10889	11856	16408	20122	9376	30414	20976	16323	15016	19908
R_{work}/R_{free}	0.118/0.166	0.175/0.239	0.159/0.197	0.169/0.228	0.141/0.184	0.162/0.220	0.122/0.154	0.147/0.199	0.130/0.148	0.166/0.228	0.157/0.190	0.189/0.251	0.185/0.216
No. of atoms													
Protein	2028	1637	2002	2002	2022	1952	1914	1907	2001	2058	1979	1686	1628
Ligand (#NTPs)	0	0	0	1 (0)	32 (1)	89 (2)	32 (1)	32 (1)	63 (2)	32 (1)	29 (1)	0	0
Water	176	92	166	112	165	176	174	146	246	235	161	101	136
B-factors													
Protein	13.2	22.8	12.6	14.2	12.7	16.7	23.7	22.5	8.4	14.3	21.4	36.0	26.9
Ligand	–	–	–	51.0	49.8	28.4	34.4	50.0	31.0	18.5	47.1	–	–
Water	20.1	28.4	19.6	16.3	17.3	21.8	32.7	32.2	16.8	20.8	25.6	36.1	37.0
rmsd													
Bond lengths (Å)	0.006	0.006	0.008	0.006	0.006	0.007	0.010	0.013	0.005	0.007	0.003	0.003	0.003
Bond angles (°)	1.09	0.98	1.19	1.02	1.08	1.28	1.39	1.675	1.146	1.12	0.83	0.79	0.74
rmsZ													
Bond length	0.39	0.34	0.41	0.32	0.40	0.37	0.50	0.68	0.30	0.38	0.23	0.26	0.23
Bond angles	0.56	0.52	0.58	0.49	0.57	0.53	0.63	0.82	0.52	0.53	0.43	0.44	0.42
Model quality (Ramachandran plot)													
Most favoured (%)	98.8	98.0	98.0	97.6	96.0	96.6	97.9	97.4	97.6	97.6	97.9	97.1	98.0
Allowed (%)	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
PDB ID code	5A8S	5A8T	5A8U	5A8V	5A96	5A98	4S1L	4S1K	5A9P	5A99	5A9A	5A9B	5A9C

3. Results

3.1. Structure determination

The polyhedrin genes from eight cypovirus types (4, 5, 14, 15, 17, 18, 19 and 20) were cloned and recombinant baculoviruses generated (Material and Methods). The resultant recombinant cypovirus polyhedra varied in size from 6 to 8 μm for CPV4, to ~1 μm for CPV17 (Table 1). The use of a small (~5 μm × 5 μm) beam at the tuneable micro-focus I24 beamline at Diamond Light Source allowed data collection at 100 K and structure solution at 2.2 Å resolution or beyond for all of these polyhedra (except that the initial structure determination of CPV17 required the use of an X-FEL see below, Table 1). Due to the tiny size of the crystals, data from a number of crystals were merged to attain a complete dataset with sufficient signal to noise. In four cases the structures were solved by SeMet labelling, and for two of these over 100 crystals were required to achieve sufficient anomalous signal to solve the structure (Table 1). The final structures were all reliable (Table 2). Data for CPV17 were also collected at room temperature on the CXI beamline at the Linac Coherent Light Source, to a resolution higher than could be achieved on I24. The small differences between the CPV17 100 K and room temperature structures have been described (Ginn et al., 2015b). For CPV1 the resolution was 2.1Å with R_{free} 12.9% and R_{work} 19.6% (Coulibaly et al., 2007).

3.2. The fold of the core structure of polyhedrin is highly conserved

All polyhedra belong to the same *I*23 space group and the native crystals (at 100 K) mostly possess a unit cell of 102–103 Å (102.8 ± 0.4 Å). The exceptions were CPV4, which forms crystals with 2 distinct unit cell dimensions (101.7 and 104.8 Å – the structure of the larger unit cell is an outlier and is discussed below), and CPV17 which has larger cell dimensions (106.1 Å room temperature and 104.9 Å 100 K). The fold of all the polyhedrins is highly similar – a β-barrel core sitting on a base domain, surrounded by a common array of five helices (H1–H5, Fig. 1a). Fig. 1b shows the arrangement of the major secondary structure elements along the polypeptide chain. The β-barrel core is the most conserved part of the structure – 90% of Cαs align with rmsd 1.1 Å (Fig. S1) and non-aligned residues lie at the periphery of the core and the termini. The β-core is reminiscent of the jelly-roll fold seen in the capsids of many icosahedral viruses (Koonin, 2008) (Fig. 1a), however, the arrangement of the nine β-strands in the two β-sheets, IBADGF and CHE, is novel. The base domain lies between H2 and βD, comprises H3 and the loops connecting it to the core and varies in length from 33–41 residues between the different cypovirus types (Fig. S2). The base caps an exposed side of the β-sheets and interacts with the loops connecting the β strands. Despite little amino acid sequence identity (Fig. S2) the 3-D structures of the polyhedrin subunits can be superimposed on one another pairwise such that 203 Cαs (~80%) are aligned with mean rmsd 1.5 Å. The full length polyhedrin structures are generally well ordered, as expected for a crystal lattice which has extensive interactions and low solvent content (19–25%), however some types show regions of flexibility: 13 residues (136–148) of CPV5 and 6 residues (68–73) of CPV14 have high (>40 compared with an overall Wilson B-factor of 21.3) B-factors, whilst 13 residues (66–78) of CPV15 could not be modelled (Fig. S2). These areas are in structurally variable parts of the proteins.

Comparison of the different polyhedrins reveals five variable regions: the N-terminal loop, connections between secondary structures (H2 and H3, βE and βF, βF and βG, βG and βH), and the C-terminal loop, which we designate V1–V5 respectively (Fig. 1).

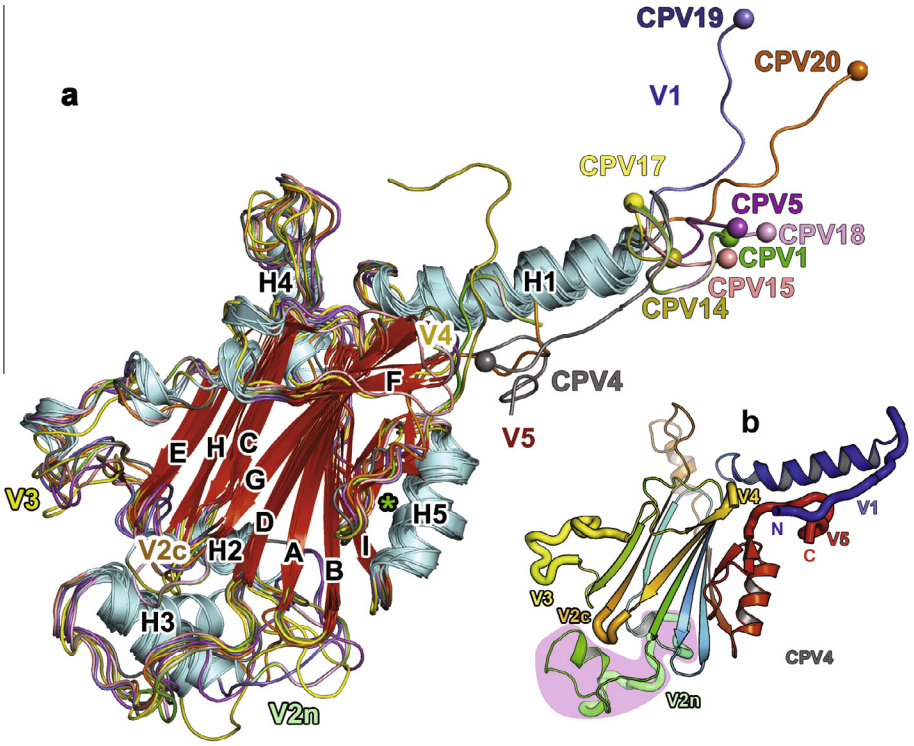


Fig. 1. Overall structure. (a) Cartoon representation of the structures of the polyhedrin subunit for 9 cypovirus types. Structures were aligned using SHP (Stuart et al., 1979). Secondary structure elements are marked and coloured cyan for helices and red for β-sheet with the loops coloured by cypovirus type; CPV1, green; CPV4, grey; CPV5, magenta; CPV14, olive; CPV15, salmon; CPV17 yellow; CPV18, pink; CPV19, slate; CPV20, orange. N-termini are highlighted by a sphere. The variable regions are labelled V1–V5. (b) Cartoon representation of the CPV4 polyhedrin subunit coloured from blue to red, N-terminus to C-terminus. Variable regions are highlighted by thickened "worms" and are labelled. The green star by H5 marks the conserved interaction shown in Fig. 3c.

V2 forms a 'cap' at one end of the protein and is subdivided across two sections of the polypeptide, V2n and V2c (Ginn et al., 2015b). Differences in these regions give each polyhedrin its characteristic appearance. CPV4 features long V1 and V5, which loop back to interact. CPV5 has insertions at V3 and V2c which form two hairpins resembling a clamp. CPV14 essentially lacks a V1 extension whereas a long V5 loop forms a hook and an intra-molecular disulphide bridge helps to stabilise the base domain. CPV15 is like CPV5, but with a smaller clamp, and has a dramatic conformational change in V4. CPV17 is unique in having an incompletely (~50%) formed inter-chain disulphide bond at room temperature, which may confer mechanical stability (Ginn et al., 2015b). Baculovirus polyhedra also contain a disulphide bond, involving a single cysteine, which links to form a dodecamer (Ji et al., 2010). As expected from the 83% amino acid sequence identity, CPV18 looks very similar to CPV1. CPV19 possesses a combination of a long V1 and V2c hairpin. CPV20 has elongated V1 and V5 but unlike CPV4 they do not interact. The V5 regions of CPV5, 15, 17 and 19 differ, but all end in a similar place. Despite a difference in unit cell of only ~0.2 Å, native and SeMet crystals of CPV5 show a conformational change: in SeMet crystals pairs of loops 166–169 and 135–143 protrude to meet and coordinate a Ca^{2+} with Asp138 and Asp168 (Fig. S3), whilst in native CPV5 there is no evidence for the Ca^{2+}, electron density is poor for 136–148, and the 135–143 loop shifts, possibly due to the proximity of methionine/SeMet residues at 136 and 170. There is no obvious significance to this change.

When we finally solved all of the structures of these polyhedra it became clear that they are all closely structurally related. This is remarkable since, although they are not totally novel structures, the low sequence identity, regions of structural variability and extremely low solvent content, rendered molecular replacement was almost impossible.

3.3. Very few amino acids are conserved

Only two residues are conserved across the nine structurally aligned polyhedrins. When CPV17 is excluded this increases to five (highlighted in Fig. S2: Q17, G95, P106, E217 and Y232, using CPV1 numbering), all located in non-variable regions. Q17 lies in H1 and the side chain engages the main chain of a distal part of the molecule – NE2 with the carbonyl group of residue 149, and OE1 with the amino group of residue 165. E217 and Y232 are towards the C-terminal region, outside the core. E217 OE2 forms a hydrogen bond with Y232 OH, and an ionic interaction with R13 in a symmetry related molecule, except in CPV5 and CPV20. E217 OE1 forms a hydrogen bond with the amino group of residue 235. P106 bridges the base domain and strand βD, and its limited flexibility likely helps direct the correct domain organisation. G95 is in the base domain close to H3, indeed only a glycine residue would allow the required packing of H3.

3.4. Crystal contacts and assembly are functionally conserved features

To investigate whether the conservation of the molecular structure and crystal packing arises from conserved interactions, the inter- and intramolecular interactions made by residues were plotted against residue number (Fig. S4). This confirms that both the inter- and intramolecular interactions are broadly conserved across all nine polyhedra. The strongest intermolecular interactions define a trimer composed of a cluster of three β-barrels sandwiched by helices top and bottom (Fig. 2a). Strong hydrophobic interactions and large numbers of hydrogen bonds, mainly involving H3, H4, elements within V3 and sheet-CHE, bind the trimer around the body diagonal threefold axes. Some 38% of the amino acids in each subunit participate in the interface, which buries ~3500 $Å^2$ of surface area (Fig. S4a). Three outstretched arms, formed mainly by H1 mediate, in part, a complex higher-level assembly, with each trimer contacting eight others. These arms stretch to contact another trimer on the far side of the unit cell (translated in the direction of a cell edge). The strongest trimer-trimer interactions are also lateral, involving trimers tightly packed along a cell edge, forming contacts of ~8000 $Å^2$ per trimer (involving ~55% of amino acids, Figs. 2b and S4b). A further interaction of ~28% of the polyhedrin amino acids, burying ~1500 $Å^2$ of surface per trimer (Figs. 2b and S4b) is also formed. The net effect of these lateral interactions is to form layers of molecules in all three directions, stacked alternately (Fig. 2c). As the network of trimers builds up, four trimers are brought close together around (0,0,0) and (1/2, 1/2, 1/2), to form a dodecameric arrangement and there is interaction mainly between the base domains (Figs. 2b and S4b) (involving around 9% of amino acids, and burying ~900–1800 $Å^2$ per trimer). The sum of such interactions, from which the whole network of a polyhedron crystal can be constructed, involves the majority of the surface of the trimer (Fig. S4b).

Overall there is remarkable strength and redundancy in the interactions, with no single weak point, and ambiguity in the assembly pathway may be key to the molecule's function, flexibility in assembly allowing the lattice to remain intact whilst building around virus particles.

3.5. The base domain implicated in viral recognition is dispensable

Native crystals of CPV4 polyhedra grouped into two distinct unit cell dimensions, 101.7 (±0.27) Å and 104.8 (±0.18) Å, whilst all SeMet crystals had the larger size (Table 1). Crystals with both unit cell dimensions were observed on each of several frozen meshes suggesting that this was not an artefact of sample preparation. In the larger-cell structure, residues 1–12, 75–108, 189–192 and 254 were not visible, however SDS–PAGE analysis of crystals showed only a single protein band, corresponding to the full length polyhedrin, suggesting that this is not due to cleavage. Residues 75–108 and 189–192 are clustered around the centre of the dodecameric assembly described above. In the absence of these trimer–trimer interactions the trimers of the dodecamer slide away from teach other, enlarging the unit cell. The volume of the central cavity increases from 7220 $Å^3$ to 18,784 $Å^3$, and the solvent plus disordered protein fraction of the crystal becomes 41% (compared to a solvent content of 19% in the smaller cell crystals). Since residues 75–108 form the base domain we wondered if polyhedra crystals could be formed without the base domain. To test this, polyhedrin proteins were made with the base domain removed by deleting residues 74–110 of CPV4 and 71–103 of CPV1 and replacing them with a flexible Gly-Ser-Gly linker (resulting in CPV4Δ and CPV1Δ respectively). Both formed diffraction capable crystals (Table 1). The CPV4Δ structure was very similar to that of CPV4 with the larger unit cell. Compared to wild-type CPV1, CPV1Δ crystals also had a slightly enlarged unit cell (103.9 *vs.* 103.0 Å). The absence of the base domain did not influence the overall structure of the trimeric building block but increased flexibility such that the electron density for 40–50 residues in the inter-trimer contact areas could not be observed. These disordered regions include the areas in CPV1 which had been shown previously to bind NTPs, so that NTPs are absent from CPV1Δ. The base domain is therefore a region that is neither required for proper folding of the protein, nor for crystal assembly, but fine-tunes the crystal, 'locking-down' the structure, often in conjunction with NTPs. This region is also possibly important for virion recognition and packaging, as discussed below. There are no contiguous channels in the wild-type polyhedra (Fig. S5), although there are significant voids where trimers come together to form dodecamers, and in the heart of the trimers. In CPV1, 14, 15, 17, 18 and 20 solvent molecules can move between these two voids, whereas in CPV4, 5 and 19 they are blocked by

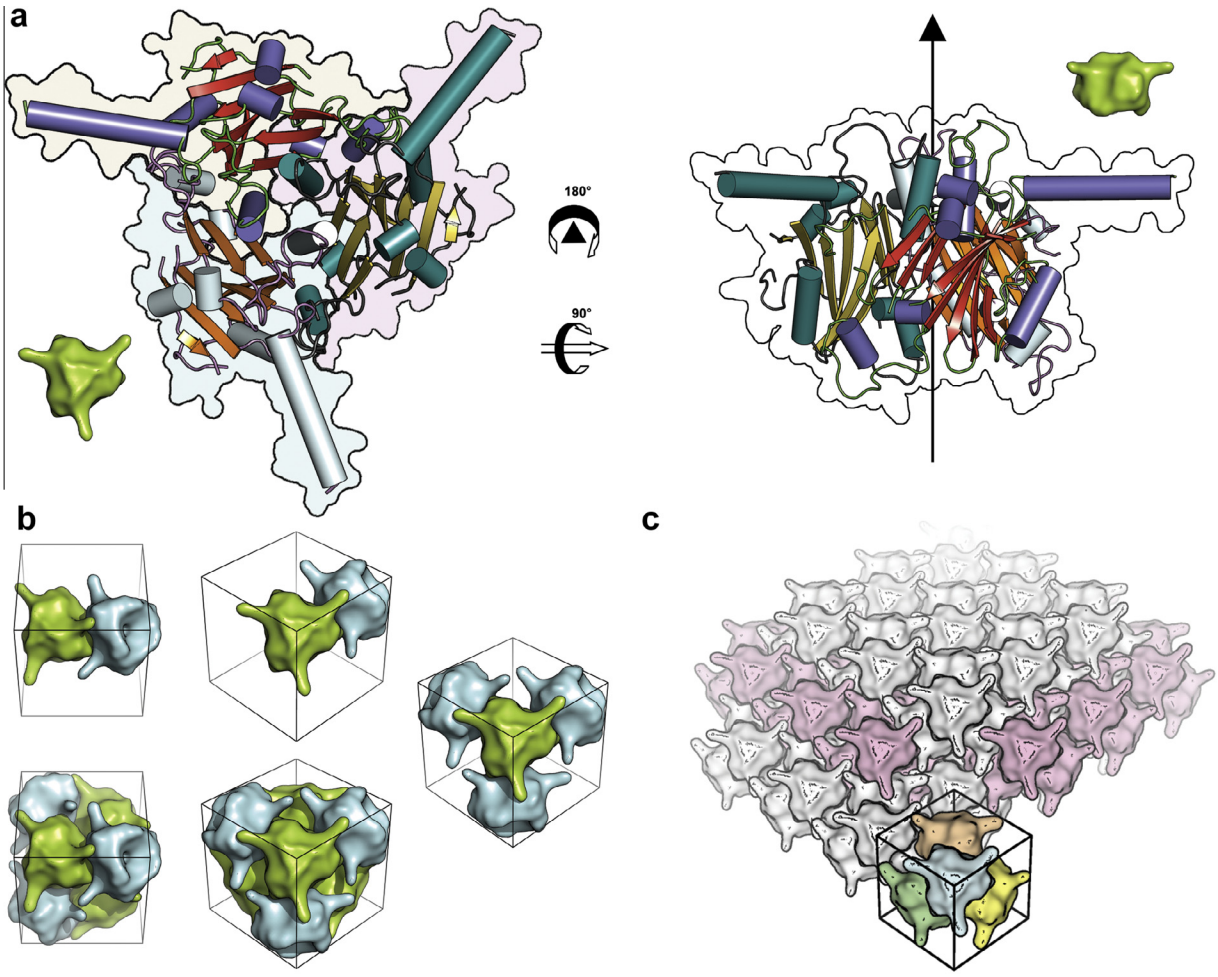


Fig. 2. From trimer to crystal. (a) Two views of a cypovirus trimer coloured by secondary structure. In the left panel (viewing the trimer from the 'top') each polyhedrin monomer is delineated by a different background colour. The arrow on the right panel goes through the threefold axis. The smaller green trimers are for orientation aids for panel (b). (b) Trimers can come together progressively. Trimers facing 'inwards' and 'outwards' have been coloured cyan and green respectively to aid interpretation and show the view. The unit cell is marked for clarity although it is unlikely that assembly will occur solely within it. (c) Multiple copies of the dodecamer are arrayed in the crystal. For one dodecamer each trimer is coloured individually. For clarity the dodecamers at the 1/2, 1/2, 1/2 position are coloured light pink. For all the panels the 'core' structure shown is that of CPV5 using just the residues which are aligned between all the 9 polyhedrin structures. Structural alignments were performed with SHP (Stuart et al., 1979). The surface maps have been progressively smoothed which results in a slightly coarse and stylised rendering but which helps to visualise the complicated interactions.

tyrosine residues. In contrast in the base domain deletion mutants all major cavities are connected, resulting in contiguous solvent channels running through the crystal lattice (Fig. S5).

3.6. Polyhedra act as molecular sieves modulating pH dependent disruption

One characteristic feature of polyhedra is that disruption only occurs when they are exposed to the very alkaline pH environment found in insect midguts, which is physiologically important to ensure the delivery of virus particles to the target intestinal cells (Rohrmann, 1986). To investigate how this critical process is achieved, eleven different cypovirus polyhedra (nine wild-type and two deletion mutants) were incubated in carbonate at pH 10.5 at room temperature for an hour, and the rate of disruption monitored by measurement of OD600 of the insoluble fraction (Fig. 3a). CPV1, CPV19 and CPV20 were resistant such that only a small portion was dissolved in an hour while the other six cypovirus types plus the two deletion mutants dissolved dramatically. To investigate if hindrance of access or buffer chemistry might be a factor in these differences we investigated a subset of polyhedra, CPV1, CPV4 and their deletion mutants, in a series of alkaline buffers at pH 10.0 for 30 min (Fig. 3b). Wild-type polyhedra behaved very differently in these buffers. Carbonate buffer is far more effective at dissolving polyhedra than other buffers, suggesting that the buffer structure as well as pH is important. The deletion mutants are much more sensitive to alkaline buffers, even those such as glycine and borate which do not disrupt crystals of full-length polyhedrin, although CAPS could not dissolve any of the samples effectively. Overall effectiveness is inversely correlated with size of the buffer molecule (Table S2) and the three CPVs (CPV1, 19 and 20) most resistant to alkali buffers have more extensive interactions between the threefold symmetry-related subunits at the CTP binding site, either strengthened by CTP (CPV1) and UTP (CPV20) binding (see below) or achieved by a large number of protein interactions mainly through hydrogen bonding, suggesting that physical access is a major factor in allowing dissolution, and that crystals tend to be resistant to non-physiological alkalis.

It has been proposed that a cluster of tyrosines trigger alkali disruption of both cypovirus (Coulibaly et al., 2007) and baculovirus (Chiu et al., 2012; Coulibaly et al., 2009; Ji et al., 2010) polyhedra, with some supporting mutagenesis data for CPV1 (Y. Ohtsuka et al., 2010). However, of the eight new structures, only CPV18 has a full

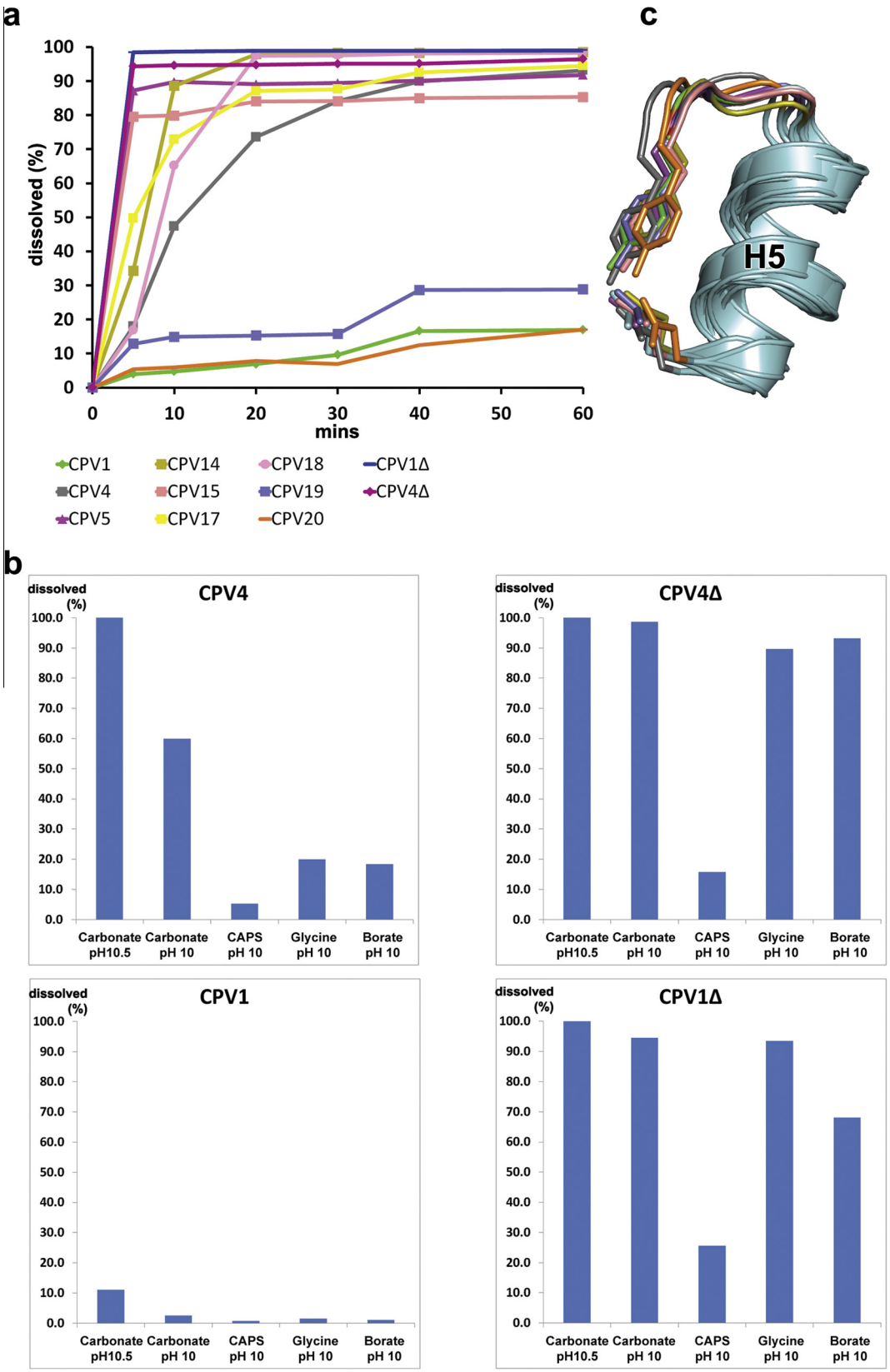


Fig. 3. Polyhedra dissolution. (a) Time course of alkaline disruption of cypovirus polyhedra. The polyhedra of nine cypovirus types plus two deletion mutants were suspended in 50 mM carbonate buffer, pH 10.5 at room temperature and the non-dissolved material monitored by absorbance at 600 nm. This was converted to % disruption by comparison to the starting value. (b) Effect of five alkali pH buffers on polyhedra of 4 cypovirus types. Release of protein from insoluble polyhedra was monitored by A280 after 20 min. As a reference, aliquots of polyhedra were completely dissolved in carbonate buffer pH 12.0, and this was used to convert to % disruption. (c) Tyr-glu interaction conserved in 8 CPV types. The green star in Fig. 1a helps locate this panel to the overall structure (the views are identical). Colours of CPV types as in Fig. 1a.

tyrosine cluster of 8 residues as found in CPV1, the others have between 2 to 5. Nevertheless in all types except CPV17 the hydroxyl-group of conserved tyrosine 232 (CPV1/18 numbering) hydrogen bonds to another conserved residue, glutamate 217 (CPV1/18 numbering) (Figs. 3c and S2). We suggest that upon raising the pH, the tyrosine acquires a negative charge, repelling the glutamate, moving H5 and the C-terminal region and initiating disruption of the crystal lattice. This tyrosine is also vital for polyhedra assembly since mutating it ablates crystal formation (Ohtsuka et al., 2010).

3.7. NTPs and protein chains can exchange places whilst maintaining the same crystal lattice

In CPV1 three NTPs per polyhedrin subunit are specifically trapped within the protein matrix (Coulibaly et al., 2007), which

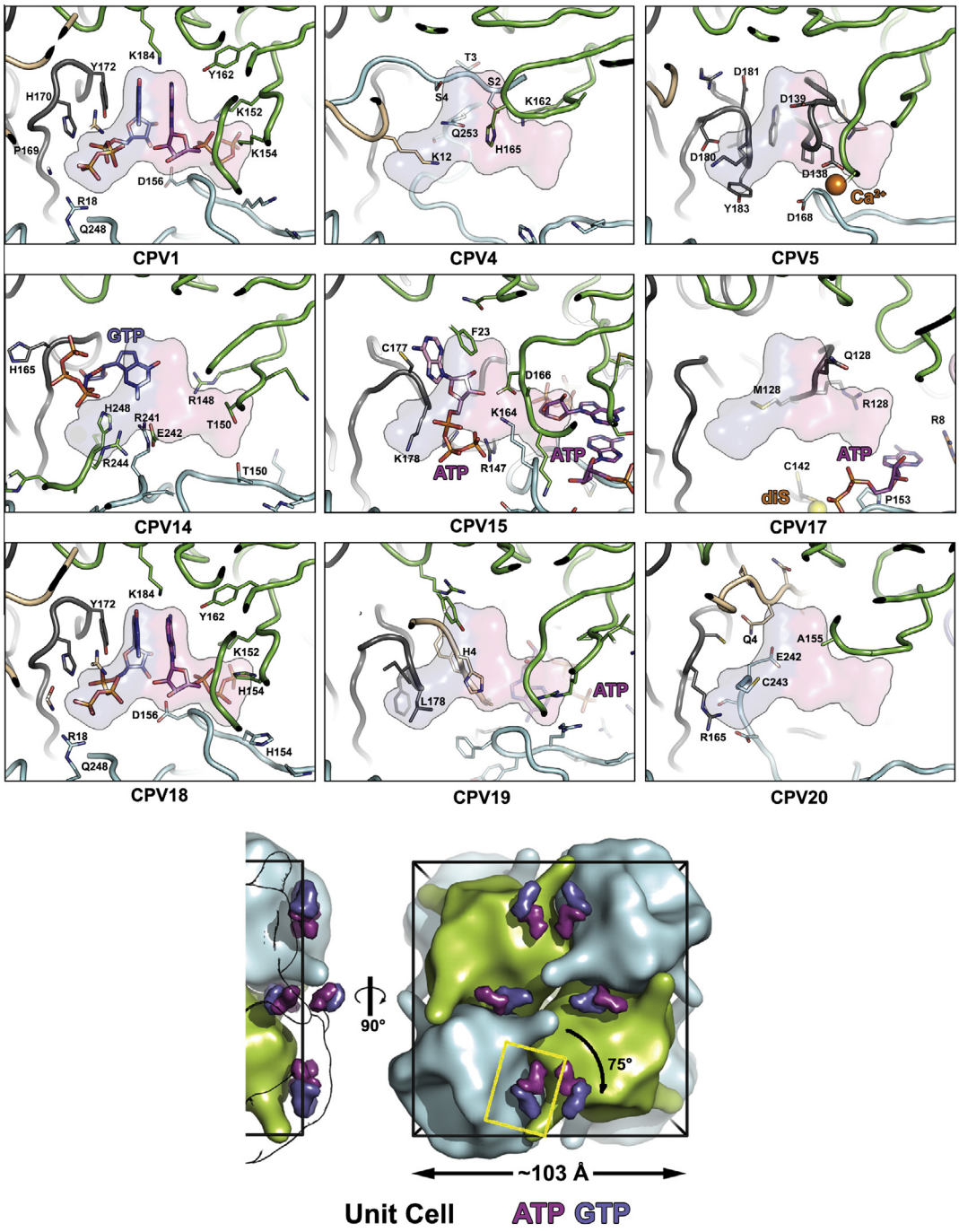
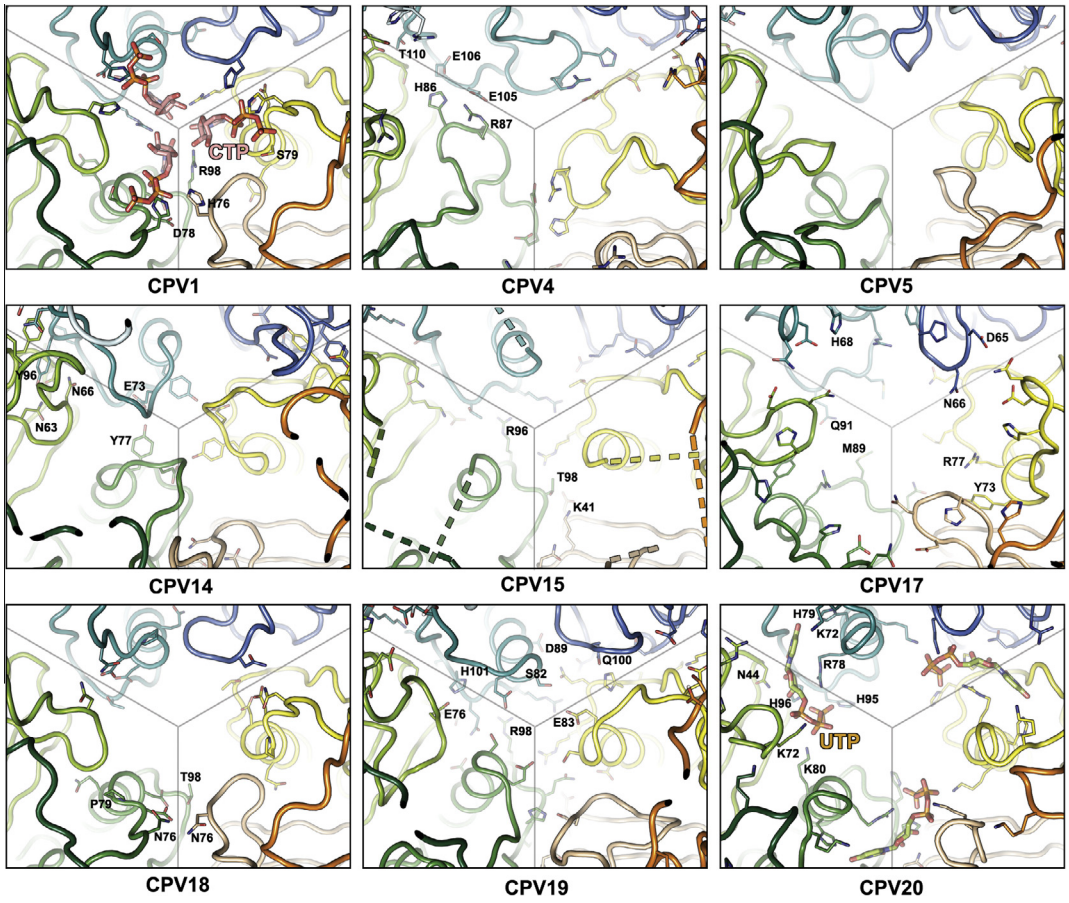


Fig. 4. Purine binding. The bottom panel helps orient the reader to the position of the depicted in the other panels. Side and face-on views of a unit cell are shown, with smoothed maps representing polyhedrin trimers similar to Fig. 2. In the side view the 2 trimers closest to the viewer are shown only as outlines so that the close proximity of the NTPs to the surface of the unit cell can be clearly seen. To improve clarity the positions of the ATP (magenta) and GTP (blue) moieties for just CPV1 are shown in this panel. As a reference the positions of the CPV1 ATP and GTP have been replicated across the other panels. The yellow box shows the area represented in other 9 panels. For the nine cypovirus types the same viewpoint is shown. Protein chains are coloured separately and consistently between types. Side chains of residues which either interact with nucleotides or occupy nucleotide pockets are shown. The view has narrow clipping planes to facilitate interpretation. The positions of the calcium ion in CPV5 and C142 in CPV17 are indicated.

might suggest that these nucleotides are essential for crystal formation and would be conserved during evolution. However, across our database of nine polyhedrin types we can now see that NTPs are dispensable (Table 2). CPV1 in fact bears the most NTPs and so we will relate our results to that type where the NTPs lie in two areas; ATP and GTP bind close to each other to form a purine cluster whilst CTPs amass at the centre of the dodecamer, the pyridine region.

3.7.1. Purine cluster

The purines are found at three 2-fold axes on the facets such that there are eight symmetry equivalent purine clusters per facet (Fig. 4). Each has complex interactions with the neighbouring protein, including variable regions V1, V3, V4, V2c and V5 contributed by four different polypeptide chains, forming diverse interactions between the facets of unit cells in different polyhedra (Fig. S6). In CPV18 purine binding is almost identical to CPV1 in that GTP hydrogen bonds with V2c and V1 while ATP is supported by two V4 regions from different polypeptide chains. The ATP and GTP bases stack with each other and with the phenol ring of Y172. There are two substitutions in the binding pocket, K154H and R155V. Residue 155 interacts via the main chain nitrogen, and is thus sequence independent, whilst at residue 154 both the lysine



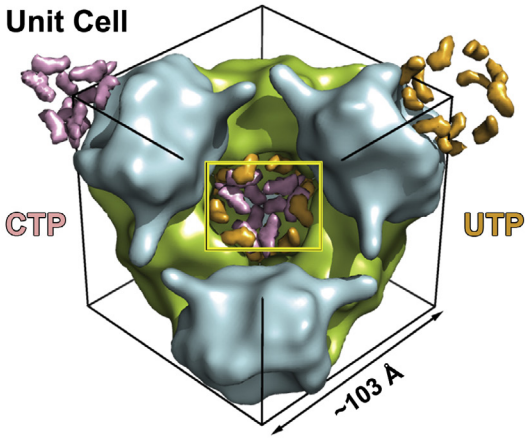


Fig. 5. Pyrimidine binding. The lower panel orientates the reader to the positions of the CTPs and UTPs within the unit cell which cluster around the centre and the apexes. The yellow box shows the area represented in other 9 panels. The view for the nine cypovirus types is from the centre out along a body diagonal. Each trimer is roughly denoted by lines. The 9 subunit protein chains visible in the view are coloured separately with the shading grouped according to which trimer they belong to (blue, green and red-yellow). Nucleotide moieties are shown as sticks along with side chains important in this region. The 13 residues missing from the CPV15 model are shown by dashed lines.

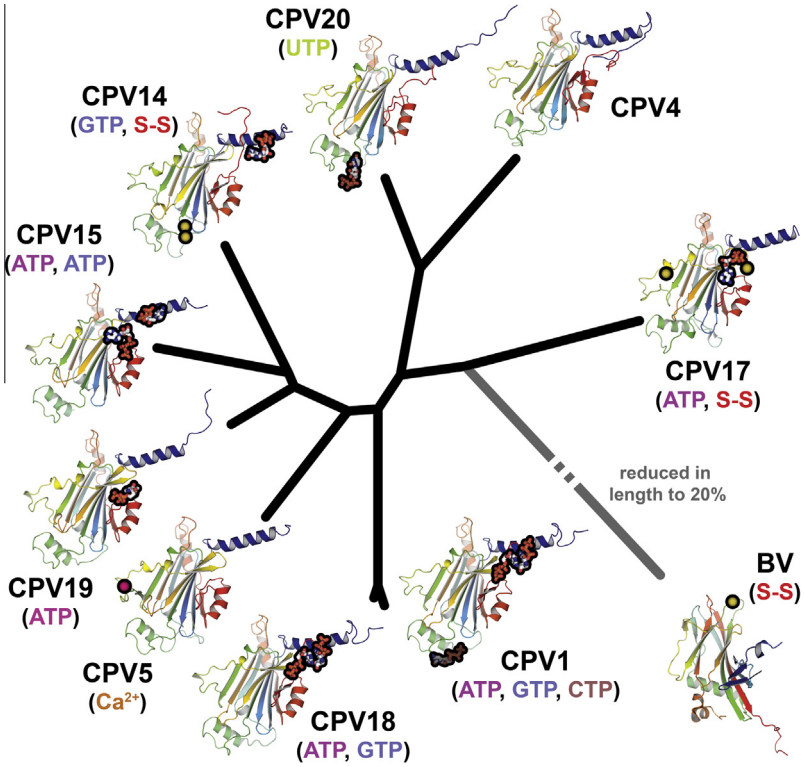


Fig. 6. Phylogenetic tree showing relationship between the polyhedrin proteins. Superpositions were performed using SHP (Stuart et al., 1979) and the phylogenetic tree calculated with PHYLIP (Felsenstein, 1989). The length of the branch to baculovirus polyhedrin (BV) has been reduced to 20% of its actual length. Structures of the proteins are shown coloured from N-terminus to C-terminus, blue to red. NTPs are represented in thickened stick style with a black outline to aid visualization. Cysteines involved in disulphide bridges are shown as black spheres, and the calcium ion in CPV5 as a pink sphere.

and histidine side chain NH groups hydrogen bond with ATP phosphates.

CPV4, 5 and 20 lack both ATP and GTP, whilst the other cypovirus types have varying combinations, in all cases the V1, V3, V4, V2c or V5 regions of the protein adopt alternate polypeptide conformations (Fig. 4). Thus in CPV4, the purines are replaced by greatly elongated V1 and V5 regions, whilst the shortened V4 loop would no longer be able to contact the ATP. In CPV20 the V4 loop is similarly shortened whilst V1, V2c, and V5 infiltrate the region. In CPV5 V3 (179–187) and V2c (135–140) form clamp-like protrusions which fill the GTP and ATP pockets, whilst in the SeMet form a calcium ion links four polyhedrin molecules (Fig. S3).

CPV15 is similar to CPV5 however V3 and V2c are shorter, leaving space for two ATPs in different conformations to CPV1: one binds V3 and V2c and the other stacks at the 2-fold axis with the backbone of V4 forming a further layer. In CPV14, a GTP is located ~5 Å from the CPV1 position, in a different conformation, binding to the backbone of V2c and embraced by V1 and bulky side chains of an elongated V5. A change in the conformation of the V4 hairpin switches it to bind another monomer. In CPV19, an extra V2c loop and elongated V1 eliminates the GTP space completely and displaces the ATP base ~13 Å *cf* CPV1, to a site surrounded by V4 and V5.

In summary the two purines are shrouded by amino acids, leaving almost no cavity (Fig. 4) and bridge at least four protein molecules to reinforce interactions between adjacent unit cells. Where one or both are missing, the polypeptide is altered to substitute for the missing moieties.

3.7.2. Pyridine region

NTP binding in this area is only observed in CPV1 and CPV20. In CPV1 a ball of 12 CTPs lie at the point of *23* symmetry at the heart of the dodecamer (Fig. 5). The pyridine in CPV20 is not well ordered, but can be identified as UTP based on the limited space and the hydrogen bonding environment provided by residue 101 main chain nitrogen and the side chain nitrogen of H79, whilst the triphosphates are supported by two lysines, two histidines and an arginine. The pyridines stabilise the dodecamer, although in CPV1 the CTPs also interact with each other, whilst the UTPs in CPV20 are more deeply buried in the trimer interface (Figs. 5 and S7).

In the cypovirus types which contain no pyridine the proteins take up a number of different configurations. In CPV14 and CPV19 the V2n loops move to occupy some of the central space. CPV4 has limited numbers of molecular interactions in this area, while in CPV5 a few strong trimer-trimer interactions are present. In CPV15, 13 disordered residues are situated around this region (indicated by dashed lines in Fig. 5) leaving the trimer contacts apparently weakened. Despite a strong overall similarity with CPV1 there is no pyridine bound in CPV18. Many of the key residues that bind the CTP in CPV1 (H76, N77, D78, S79, Y80, D81, D96 and R98) are substituted, including H76N and S79P, which interact with the phosphates and R98T which stacks with the base. Clearly these changes are sufficient to disrupt CTP binding (Fig. 5).

To summarise, different polyhedrins adopt different solutions to stabilising the crystal lattice, some use specific nucleotides, whilst others substitute protein chains, in all crystals however the lattice is conserved.

4. Discussion

Structures for a number of previously uncharacterised types of cypovirus polyhedrins have revealed a strongly conserved structural core. A phylogenetic tree constructed based on the structures (Fig. 6) shows, as expected, that CPV1 and CPV18 are closely related structurally, but otherwise the tree is rather star-like, with

CPV17 being perhaps rather an outlier and closest to the markedly different baculovirus polyhedrin. It is conceivable that CPV17 retains some features of an ancestral insect virus polyhedrin. Given the evolutionary explosion of insects some 400 million years ago the considerable divergence of the polyhedra suggests that much of the substantial evolution of the CPVs took place alongside these events, perhaps explaining why the build-up of variation appears almost saturated, with only two amino acids fully conserved across all nine types. Away from the core of the polyhedrin we identify five regions of structural variability and show that NTPs are used in the assembly of some but not all proteins, so that they impose little constraint on the polyhedrin structure. We also find that the whole base domain is dispensable for polyhedra formation, although the biochemical features are altered dramatically when it is removed. The base domains include the most variable region of the protein and nestle around a point of *23* symmetry. The bases likely attach to the virus to be encapsidated and are therefore either packed together inside the crystal or exposed on the outside, furthermore, since they do not drive assembly they will tend to remain accessible on the surface of the growing crystal, ready to engage a protein-only icosahedral shaped CPV particle. The outer faces of the base domains are positively charged across nearly all the types, at present there is no little structural information for charge status of the virion component that is likely to attach to the polyhedrin. It seems that in the final stage of assembly the base domains pull the lattice into a closer knit form, presumably allowing useful flexibility in the lattice in the vicinity of the viral inclusions (note that the presumed ancestral CPV17 does not fully lock down the lattice, retaining a larger unit cell). The variety of NTPs seem to act as small molecule 'fillers' in the absence of virus particles, allowing lattice completion and the avoidance of crystal voids whilst permitting appropriate specificity in the virus recognition motifs between CPV types. Overall this domain is likely a late addition to the fold (it is dispensable for protein folding) and probably was a feature acquired to aid viral recognition (it is distal to the point of attack by ions which initiate the dissolution process by modifying a point critical for assembly). There may be potential for exploiting the variation in the base domain for nanotechnology applications.

Structure based sequence alignments of the polyhedrins show that whilst approximately 80% of the C alphas are structurally equivalent (pairwise rmsd from 0.2 to 1.5 Å), the corresponding sequence identities are only 12–32% (Fig. S2). The exception is the polyhedrin of CPV18, which has >80% sequence identity with that of CPV1, consistent with their relatively recent divergence (Graham et al., 2007). Although the two molecules are very similar (rmsd of 247 structurally equivalent Cαs 0.23 Å (Fig. S1)) they behave dramatically differently in alkali buffers, CPV18 being far more sensitive to dissolution. Clearly functional differences have been locked into these molecules, probably arising from the structure around the CTP binding site. In fact the three CPVs (CPV1, 19 and 20) most resistant to alkali buffers all display more extensive interactions between the 3-fold symmetry-related subunits of the trimeric building block – strengthened by either pyridine binding (CPV1 and CPV20) or numerous protein–protein hydrogen bonds.

Overall polyhedra present a fascinating picture of the effect of 400 million years of evolution on a system where crystal lattice formation is conserved in the face of massive sequence variability. By spreading the crystal contacts over the majority of the surface of the molecule lattice formation remains robust in the face of many individual changes, in the same way that the tertiary structure is robust. Indeed whilst most proteins harbour the greatest fraction of conserved residues in the protein core (e.g. 63% for picornavirus 3C proteases) only a minority (43%) of the conserved residues in CPV polyhedrins are buried in the core.

Accession numbers

Coordinates and structures factors have been deposited in the PDB with accession codes: CPV1Δ 5A9B; CPV4 101 Å cell 5A8S; CPV4 104 Å cell 5A8T; CPV4Δ 5A9C; CPV5 native 5A8U; CPV5 SeMet 5A8V; CPV14 5A96; CPV15 5A98; CPV17 room temperature 4S1L; CPV17 100K 4S1K; CPV18 5A9P; CPV19 5A99; CPV20 5A9A.

Acknowledgements

We thank Jun Dong, Robert Esnouf and Jonathan Diprose for IT support and Diamond Light Source for access to beamline I24 (proposal number MX6387) that contributed to the results presented here. GS and DIS are supported by the MRC (Grant no G100099), HG is a Wellcome Trust Student. The Wellcome Trust core award 090532/Z/10/Z provided administrative support.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jsb.2015.08.009.

References

Abrahams, J.P., Leslie, A.G., 1996. Methods used in the structure determination of bovine mitochondrial F1 ATPase. Acta Crystallogr. D Biol. Crystallogr. 52, 30–42.

Adams, P.D., Afonine, P.V., Bunkoczi, 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. PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallogr. D 66, 213–221.

Anduleit, K., Sutton, G., Diprose, J.M., Mertens, P.P., Grimes, J.M., Stuart, D.I., 2005. Crystal lattice as biological phenotype for insect viruses. Protein Sci. 14, 2741–2743.

Attoui, H., Mertens, P.P.C., Becnel, J., Belaganahalli, S., Bergoin, M., Brussaard, C.P., Chappell, J.D., Ciarlet, M., del Vas, M., Dermody, T.S., Dormitzer, P.R., Duncan, R., Fang, Q., Graham, R., Guglielmi, K.M., Harding, R.M., Hillman, B., Makkay, A., Marzachi, A.C., Matthijnssens, J., Milne, R.G., Mohd Jaafar, F., Mori, H., Noordeloos, A.A., Omura, T., Patton, J.T., Rao, S., Maan, M., Stoltz, D., Suzuki, N., Upadhyaya, N.M., Wei, C., Zhou, H., 2012. Cypovirus, p. 572–580, in: A. M. Q. King, et al., Eds.), Virus Taxonomy: classification and nomenclature of viruses. Ninth Report of the International Committee on Taxonomy of Viruses, p, Academic Press, London.

Axford, D., Ji, X., Stuart, D.I., Sutton, G., 2014. In cellulo structure determination of a novel cypovirus polyhedrin. Acta Crystallogr. D Biol. Crystallogr. 70, 1435–1441.

Belloncik, S., Mori, H., 1998. Cypoviruses. In: Miller, L.K., Ball, L.A. (Eds.), The Insect Viruses. Plenum, New York, pp. 337–369.

Boudko, D.Y., Moroz, L.L., Harvey, W.R., Linser, P.J., 2001. Alkalinization by chloride/bicarbonate pathway in larval mosquito midgut. Proc. Natl. Acad. Sci. U.S.A. 98, 15354–15359.

Bricogne, G., Vonrhein, C., Flensburg, C., Schiltz, M., Paciorek, W., 2003. Generation, representation and flow of phase information in structure determination: recent developments in and around SHARP 2.0. Acta Crystallogr. D 59, 2023–2030.

Bricogne, G., Blanc, E., Brandl, M., Flensburg, C., Keller, P., Paciorek, W., Roversi, P., Sharff, A., Smart, O.S., Vonrhein, C., Womack, T.O., 2011. BUSTER version 2.9.2.

Chen, V.B., Arendall III, W.B., Headd, J.J., Keedy, D.A., Immormino, R.M., Kapral, G.J., Murray, L.W., Richardson, J.S., Richardson, D.C., 2010. MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr. D 66, 12–21.

Chiu, E., Coulibaly, F., Metcalf, P., 2012. Insect virus polyhedra, infectious protein crystals that contain virus particles. Curr. Opin. Struct. Biol. 22, 234–240.

Coulibaly, F., Chiu, E., Ikeda, K., Gutmann, S., Haebel, P.W., Schulze-Briese, C., Mori, H., Metcalf, P., 2007. The molecular organization of cypovirus polyhedra. Nature 446, 97–101.

Coulibaly, F., Chiu, E., Gutmann, S., Rajendran, C., Haebel, P.W., Ikeda, K., Mori, H., Ward, V.K., Schulze-Briese, C., Metcalf, P., 2009. The atomic structure of baculovirus polyhedra reveals the independent emergence of infectious crystals in DNA and RNA viruses. Proc. Natl. Acad. Sci. U.S.A. 106, 22205–22210.

Cowtan, K., 1994. 'dm': an automated procedure for phase improvement by density modification. Joint CCP4 and ESF-EACBM Newsletter on Protein Crystallography 31, 34–38.

Emsley, P., Cowtan, K., 2004. Coot: model-building tools for molecular graphics. Acta Crystallogr. D Biol. Crystallogr. 60, 2126–2132.

Evans, G., Axford, D., Owen, R.L., 2011. The design of macromolecular crystallography diffraction experiments. Acta Crystallogr. D 67, 261–270.

Felsenstein, J., 1989. PHYLIP - phylogeny inference package (version 3.2). Cladistics 5, 164–166.

Fortelle, E.d.L., Bricogne, G., 1997. Maximum-likelihood heavy-atom parameter refinement for multiple isomorphous replacement and multiwavelength anomalous diffraction methods, in: C.W. Carter, R.M. Sweet, (Eds.), Methods in Enzymology, Academic Press, San Diego, pp. 472–494.
Ginn, H.M., Brewster, A.S., Hattne, J., Evans, G., Wagner, A., Grimes, J.M., Sauter, N.K., Sutton, G., Stuart, D.I., 2015a. A revised partiality model and post-refinement algorithm for X-ray free-electron laser data. Acta crystallogr. D Biol. crystallogr. 71, 1400–1410.
Ginn, H.M., Messerschmidt, M., Ji, X., Zhang, H., Axford, D., Gildea, R.J., Winter, G., Brewster, A.S., Hattne, J., Wagner, A., Grimes, J.M., Evans, G., Sauter, N.K., Sutton, G., Stuart, D.I., 2015b. Structure of CPV17 polyhedrin determined by the improved analysis of serial femtosecond crystallographic data. Nat. Commun. 6, 6435.
Graham, R.I., Rao, S., Sait, S.M., Mertens, P.P., Hails, R.S., Possee, R.D., 2007. Characterisation and partial sequence analysis of two novel cypoviruses isolated from the winter moth *Operophtera brumata* (Lepidoptera: Geometridae). Virus Genes 35, 463–471.
Ji, X., Sutton, G., Evans, G., Axford, D., Owen, R., Stuart, D.I., 2010. How baculovirus polyhedra fit square pegs into round holes to robustly package viruses. EMBO J. 29, 505–514.
Kabsch, W., 1993. Automatic processing of rotation diffraction data from crystals of initially unknown symmetry and cell constants. J. Appl. Crystallogr. 26, 795–800.
Krissinel, E., Henrick, K., 2007. Protein interfaces, surfaces and assemblies. J. Mol. Biol. 372, 774–797.
Leslie, A.G.W., Powell, H.R., 2007. Processing diffraction data with mosflm, p. 41–51, in: R. R. J. and J. L. Sussman, Eds.), Evolving Methods for Macromolecular Crystallography.
McCoy, A.J., Grosse-Kunstleve, R.W., Adams, P.D., Winn, M.D., Storoni, L.C., Read, R.J., 2007. Phaser crystallographic software. J. Appl. Crystallogr. 40, 658–674.
O'Reilly, D.R., Miller, L.K., Luckow, V.A., 1994. Baculovirus expression vectors. A Laboratory Manual. Oxford University Press, New York.
Ohtsuka, Y., Nakai, D., Coulibaly, F., Chiu, E., Metcalf, P., Mori, H., 2010. Mutations of cypovirus polyhedrin and applications of polyhedra to protein nanocontainers. Nanotechnology 3, 292–295.
Otwinowski, Z., Minor, W., Charles W. Carter, J., 1997. Processing of X-ray diffraction data collected in oscillation mode, Methods in Enzymology, Academic Press, pp. 307–326.
Read, R.J., Schierbeek, A.J., 1988. A phased translation function. J. Appl. Crystallogr. 21, 490–495.
Rohrmann, G.F., 1986. Polyhedrin structure. J. Gen. Virol. 67 (Pt 8), 1499–1513.
Rothman, A.L., Myers, J., 1996. Debilitating effects of viral diseases on host Lepidoptera. J. Invertebr. Pathol. 67, 1–10.
Schrodinger, LLC. 2010. The PyMOL Molecular Graphics System, Version 1.3r1.
Strachan, T., Read, A.P., 1999. Human Molecular Genetics, New York.
Stuart, D.I., Levine, M., Muirhead, H., Stammers, D.K., 1979. Crystal structure of cat muscle pyruvate kinase at a resolution of 2.6 Å. J. Mol. Biol. 134, 109–142.
Vago, C., Croissant, O., 1959. Recherches sur la pathogenese des viroses d'insectes. La liberation des virus dans le tube digestif de l'insecte a partir des corps d'inclusion ingeres. Annales des épiphyties 10, 5–18.
Winn, M.D., Ballard, C.C., Cowtan, K.D., Dodson, E.J., Emsley, P., Evans, P.R., Keegan, R.M., Krissinel, E.B., Leslie, A.G.W., McCoy, A., McNicholas, S.J., Murshudov, G.N., Pannu, N.S., Potterton, E.A., Powell, H.R., Read, R.J., Vagin, A., Wilson, K.S., 2011. Overview of the CCP4 suite and current developments. Acta Crystallogr. D 67, 235–242.
Winter, G., 2010. Xia2: an expert system for macromolecular crystallography data reduction. J. Appl. Crystallogr. 43, 186–190.
Zhao, Y., Chapman, D.A., Jones, I.M., 2003. Improving baculovirus recombination. Nucleic Acids Res. 31, E6–6.

JOURNAL OF APPLIED CRYSTALLOGRAPHY

ISSN 1600-5767

On the release of *cppxfel* for processing X-ray free-electron laser images[1]

Helen Mary Ginn,[a] Gwyndaf Evans,[b] Nicholas K. Sauter[c] and David Ian Stuart[a,b]*

[a]Division of Structural Biology, Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford, Oxfordshire OX3 7BN, UK, [b]Diamond House, Harwell Science and Innovation Campus, Fermi Avenue, Didcot, Oxfordshire OX11 QX, UK, and [c]Molecular Biophysics and Integrated Bioimaging Division, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, California 94720, USA. *Correspondence e-mail: dave@strubi.ox.ac.uk

Received 18 December 2015
Accepted 25 April 2016

Edited by T. White, Center for Free-Electron Laser Science, Hamburg, Germany

[1] This article will form part of a virtual special issue of the journal on free-electron laser software.

Keywords: X-ray free-electron lasers; XFELS; serial femtosecond crystallography; data analysis; computer programs.

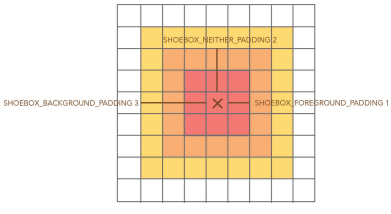
Supporting information: this article has supporting information at journals.iucr.org/j

As serial femtosecond crystallography expands towards a variety of delivery methods, including chip-based methods, and smaller collected data sets, the requirement to optimize the data analysis to produce maximum structure quality is becoming increasingly pressing. Here *cppxfel*, a software package primarily written in C++, which showcases several data analysis techniques, is released. This software package presently indexes images using DIALS (diffraction integration for advanced light sources) and performs an initial orientation matrix refinement, followed by post-refinement of individual images against a reference data set. *Cppxfel* is released with the hope that the unique and useful elements of this package can be repurposed for existing software packages. However, as released, it produces high-quality crystal structures and is therefore likely to be also useful to experienced users of X-ray free-electron laser (XFEL) software who wish to maximize the information extracted from a limited number of XFEL images.

1. Introduction

X-ray free-electron lasers (XFELs) provide a source of X-ray light in the form of short ~50 fs pulses, several orders of magnitude brighter than a synchrotron beam (Neutze *et al.*, 2000; Chapman *et al.*, 2011). Crystallographic methods can be extended to otherwise intractable micro- or nanocrystalline samples, where the majority of crystal destruction takes place after diffraction has occurred (Chapman *et al.*, 2014). This paves the way towards reconstructions of electron density from single-molecule diffraction patterns (Ekeberg *et al.*, 2015). In addition, time-resolved studies on the picosecond or femtosecond timescale are now possible owing to the precision of pulse timing, the very short pulse length and the considerable energy in each pulse (Neutze & Moffat, 2012). These studies in particular require a very large number of crystals in order to produce enough data per time point to reconstruct a movie of motion, and thus much effort has been devoted to reducing the number of images required to reconstruct high-quality electron density (Hattne *et al.*, 2014). There are now a number of software packages and their subsequent improvements employing data analysis methods that are capable of processing XFEL data sets, including *CrystFEL* (Kirian *et al.*, 2010; White *et al.*, 2012, 2013) and *cctbx.xfel* (Hattne *et al.*, 2014; Sauter *et al.*, 2013, 2014; Sauter, 2015). *CrystFEL* has recently benefited from additional algorithms for orientation matrix refinement and modelling partiality (White *et al.*, 2016).

The applications to soluble proteins include the determination of unknown structures by molecular replacement (notably for a type of cypovirus polyhedrin; Ginn, Messer-



OPEN ACCESS

schmidt *et al.*, 2015), proof of principle demonstrations of phasing by use of anomalous scattering (Barends *et al.*, 2013, 2014; Nakane *et al.*, 2015), and time-resolved studies on photoactive yellow protein (Tenboer *et al.*, 2014) and myoglobin (Barends *et al.*, 2015). Membrane proteins analysed include photosystem I (Chapman *et al.*, 2011), photosystem II (Kern *et al.*, 2012, 2013, 2014; Kupitz *et al.*, 2014; Suga *et al.*, 2015), rhodopsin bound to arrestin (Kang *et al.*, 2015) and a human serotonin receptor (Liu *et al.*, 2013).

Here an open-source software suite (*cppxfel*) is provided to showcase methods which have been previously published (Ginn, Messerschmidt *et al.*, 2015; Ginn, Brewster *et al.*, 2015). Implementation of these methods in other software packages is keenly encouraged. The DIALS (diffraction integration for advanced light sources) framework (Waterman *et al.*, 2013) encompasses the indexing and integration of images initially from synchrotron data using a wide range of optional methods with a clean command line interface. DIALS is also capable of indexing XFEL still images and is currently employed to produce indexing results which can be passed into the *cppxfel* integration and post-refinement pipeline to produce merged reflection lists. A summary of how to use the *cppxfel* software package to process XFEL stills is provided here, along with the tunable parameters of major interest for the user.

2. Description of the methods and implementation

2.1. Overview of *cppxfel*-specific features

Cppxfel currently creates merged reflection lists in MTZ format from image files which have been processed by the DIALS indexing algorithms (Waterman *et al.*, 2013). To improve the ease of use, we include with this *cppxfel* release scripts for indexing still images using DIALS.

There are two major innovations in the strategy employed by *cppxfel* which are likely to be of immediate value to other software: (*a*) the initial orientation matrix refinement as described by Ginn, Messerschmidt *et al.* (2015), and (*b*) the nature of the partiality model, which has been described by Ginn, Brewster *et al.* (2015).

The justification for the strategy employed during initial orientation matrix refinement, which has been extended to include refinement of unit-cell lengths, is that over-prediction can reveal spots that are in the correct position but too weak to be picked up by spot-finding algorithms. A small increase in photon counts at any detector pixel is likely to be noise but, if found at a position which is already known to correspond to a nearby Miller index, has a greater probability of being a true signal. By using this information, the number of spots which can be used to refine the orientation matrix is increased substantially, leading to a more accurate estimation of the orientation. As a result, this method is used in *cppxfel* to perform initial refinement of the orientation and unit-cell parameters.

The innovative partiality model, which is introduced during post-refinement, combines both bandwidth- and mosaicity-based models. This part of the pipeline fine-tunes the orientation of the unit cell, the wavelength of the XFEL pulse, and the crystal domain size and mosaicity against a reference data set, to achieve high-quality estimations of the full intensities of individual reflections. This model differs from others in the literature in explicitly modelling both bandwidth and mosaicity. The program can use a bandwidth-based model, a mosaicity-based model or a combination of the two. A bandwidth-based model would be best suited to low-mosaicity crystals where reflections are mainly illuminated by the spread of photon wavelengths. However, a mosaicity-based model may be better suited to experimental setups which are fitted with a monochromator. As an example of the differences between bandwidth and mosaicity models, the effects of the two models on a low-mosaicity crystal of CPV17 polyhedrin are shown in Fig. 1.

2.2. Programming environment and implementation

Cppxfel is almost entirely written in C++ but is executed at the highest level using Python. It is designed to be placed within the Computational Crystallography Toolbox (*cctbx*) project (Grosse-Kunstleve *et al.*, 2002) distribution, and although *cctbx.xfel* is not strictly necessary for compilation, many of the features of *cctbx.xfel* are useful for analysis and it is therefore recommended. *Cppxfel* supports threading within the C++ code on a single machine using the Boost thread library, which is now delivered along with the standard *cctbx* distribution. Machines require a C++ compiler with the c++0x standard available. The code was developed and tested on 64 bit CentOS 6.*x* systems (*i.e.* gcc 4.4.7), DIALS version 1.0-141-g20c8340, *cctbx* build dev-580, to create *cppxfel* version 1.1. The code should be compatible with versions as early as gcc version 4.3 if an earlier version is required for compilation. For the purposes of DIALS analysis, images should be in the same pickle format (Python's serialization format) used for *cctbx.xfel*. These can be generated directly from the raw streams using a PSDM-enabled *cctbx.xfel* installation, or converted from another supported file format (including `*.cbf` files which may be generated from synchrotron sources) using the command `cxi.image2pickle`, also contained within the *cctbx.xfel* distribution. To streamline the implementation of *cppxfel*, the next release will work with the HDF5 image format which is becoming a *de facto* standard for diffraction data (The HDF Group, 1997), such as that output from the new data processing pipeline at SACLA (Nakane *et al.*, 2016).

3. Optimal use of *cppxfel*

The software package includes a data set of 1000 Cypovirus polyhedra type 17 (CPV17) diffraction images for which structures have been previously published (Ginn, Messerschmidt *et al.*, 2015; Ginn, Brewster *et al.*, 2015) as an example. The data processing is broadly separated into four stages: (i) indexing with DIALS, (ii) initial refinement and integration, (iii) post-refinement, and (iv) final merge (Fig. 2). This section discusses the optimal use of *cppxfel*, which users may wish to consult for their particular study. The default parameters have

been chosen to process the majority of data sets well, but modifying certain parameters can often bring about improvements in the quality of the final data set.

3.1. Indexing

Images are read in and processed singly by DIALS. A Python script is provided to automate this and can be run with

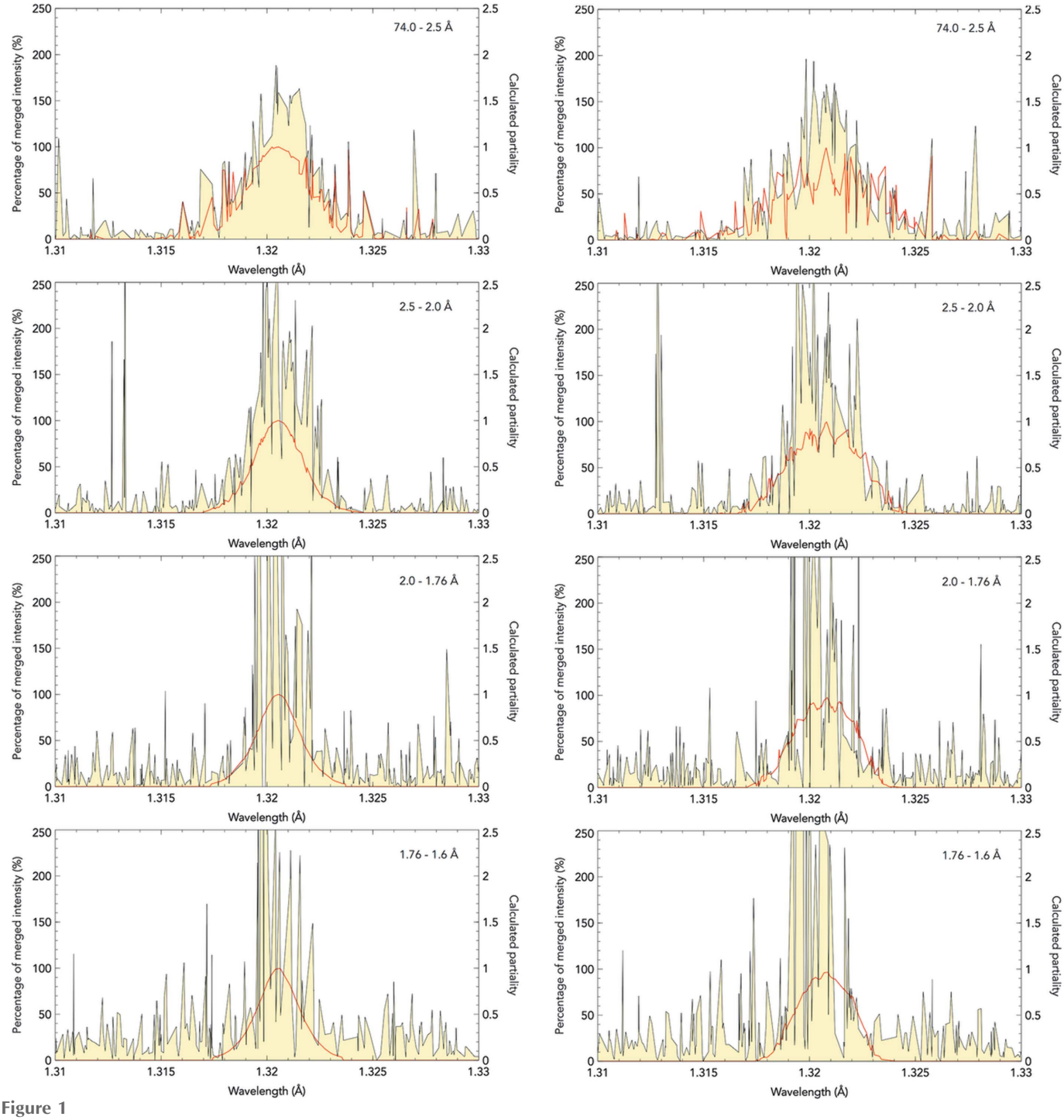


Figure 1
Plotting calculated partiality (red line) against estimate of partiality (yellow fill) by comparing against the reference data set, plotted against the Ewald sphere wavelength for the midpoint of each reflection. These are two models plotted vertically in four resolution shells to 1.6 Å resolution. The left and right models use two different bases: on the left, post-refinement has been carried out with a bandwidth of 0.18% and 0° mosaicity. On the right, the bandwidth was lowered to 0.07% and the mosaicity was increased to 0.03°. Both the bandwidth model and the mosaicity model were calculated with a super-Gaussian exponent of 1.5. The four panels on each side correspond to reflections at increasing resolution. The agreement is fairly close for both models at mid to high resolution, but the mosaicity model has a more erratic structure at low resolution. Overall, the correlation coefficient is reduced for the mosaicity model (89%) compared to the bandwidth model (96%), which suggests that a bandwidth model is preferred for the highly ordered CPV17 polyhedrin crystals from which these data were collected.

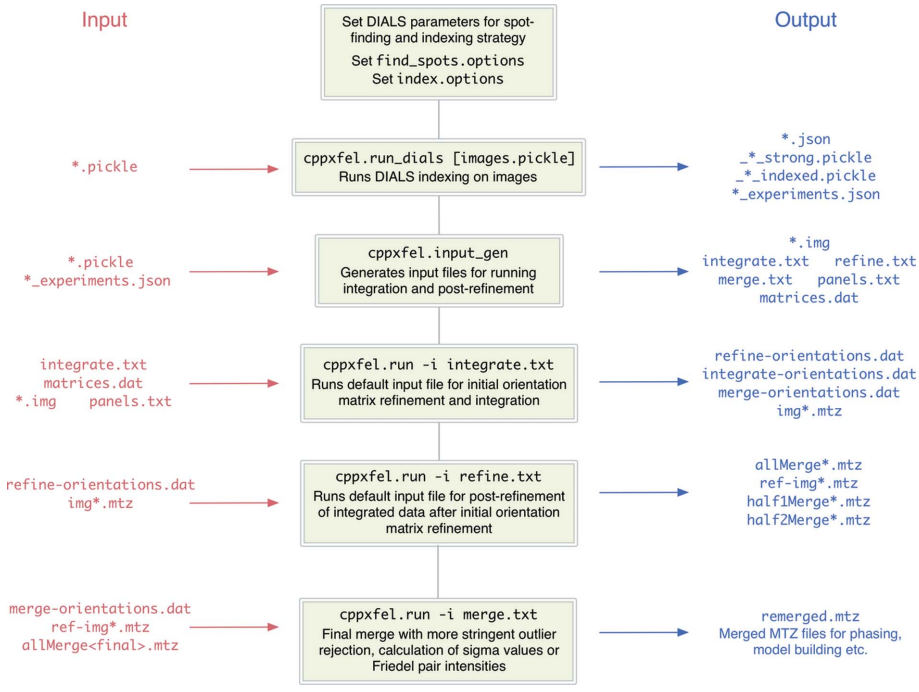


Figure 2
Workflow for executing indexing, initial orientation matrix refinement and post-refinement of individual images.

`cppxfel.run_dials *.pickle`. This will import single images, run spot finding and attempt to index each image. The parameters for these three steps can be customized by changing the option files `find_spots.options` and `index.options` which are appended to the command line call within the script. These can be used to change spot-finding parameters or provide target unit cells and space groups for indexing. Refer to the DIALS web site tutorials for more information (http://dials.github.io/documentation/tutorials/index.html). Depending on the data one may find it useful to enable the indexing of multiple crystal lattices (Gildea *et al.*, 2014).

After individual images have been indexed, the successfully indexed images must be prepared for entry into the *cppxfel* pipeline by running `cppxfel.input_gen`. This will search for successful images in the current directory and generate prototype input scripts which can be immediately fed into *cppxfel* or altered to improve the refinement strategy. These include the standard scripts `integrate.txt`, `refine.txt` and `merge.txt`.

3.2. Initial orientation matrix refinement and integration

Cppxfel is governed by input text files with a simple parameter and value entry format, followed by an optional list of commands. These are supplied to the program by running `cppxfel.run -i inputfile.txt`. The default input files are `integrate.txt`, `refine.txt` and `merge.txt`. The first carries out initial orientation matrix refinement and integration, generating a single MTZ file per image, the next runs post-refinement on those MTZ files, while the last recalculates sigma values in the final merge.

The main parameters of concern for integration are those defining the background and integration box, any required rough correction of detector geometry (which should not be required if detector geometry has already been corrected), the threshold at which a spot may correspond to a strong reflection, the extent of over-prediction, and parameters concerning the treatment of the unit cell and its refinement.

The parameter `METROLOGY_SEARCH_SIZE` can be used to roughly correct for detector geometry errors by searching for a nearby maximum pixel value according to the protocol defined previously (Ginn, Messerschmidt *et al.*, 2015). For example, a search size of 5 will recentre the integration box on the pixel with the maximum value up to 5 pixels away from the predicted spot position.

The simple integration box is primarily governed by three parameters: `SHOEBOX_FOREGROUND_PADDING`, `SHOEBOX_NEITHER_PADDING` and `SHOEBOX_BACKGROUND_PADDING`. These parameters define squares of integer distances from the central pixel, where pixels are treated as foreground signal, ignored pixels or background pixels as in Fig. 3. This is a very simple treatment of spot integration, but it has been found to be capable of generating high-quality structures. Spots are integrated using a simple summation of foreground pixels and removal of the estimation of background in this area. Integrated positions are considered to be potential reflections for the purposes of initial orientation matrix refinement if the measure of their signal exceeds a certain value defined by `INTENSITY_THRESHOLD`. The toggle `ABSOLUTE_INTENSITY`, which can be set to `ON` or `OFF`, determines whether this uses the raw intensity value or the $I/\sigma(I)$ value.

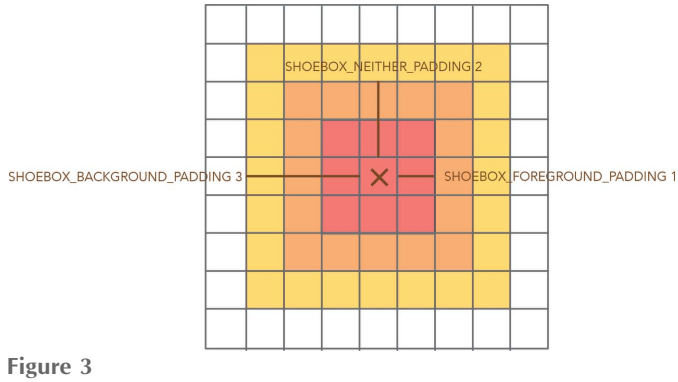


Figure 3
Simple integration box available in *cppxfel*, defined by the three parameters shown specifying the size of the boxes of foreground, neither and background flags.

The method used by initial orientation matrix refinement and the subsequent post-refinement relies on over-prediction of illuminated Miller indices. Some crystals may have a higher mosaicity than others, or have a poorer starting orientation matrix, and thus require a wider prediction of reflections. These are governed by two parameters. `OVER_PRED_BANDWIDTH` primarily increases over-prediction at high resolution by assuming a very wide bandwidth of the beam energy, and may need expanding from the default of 3% (representing the spread of two standard deviations of a super-Gaussian model for beam energy). In contrast `OVER_PRED_RLP_SIZE` increases over-prediction at low resolution by expanding the size of the reciprocal lattice points, thus increasing their likelihood of intersecting the illuminated Ewald spheres. `OVER_PRED_RLP_SIZE` may need to be increased from the default of 2×10^{-4} Å^{-1} radius (corresponding to a crystal domain size of 500 nm) if there are not enough predicted reflections at low resolution evident in later processing.

Finally, unit-cell dimensions from indexing can vary beyond the variation expected for a biological sample. This is due to uncertainty in other parameters and also because each image is treated independently and, to a first approximation, samples a central section of the reciprocal lattice which is often insufficient to properly define all unit-cell dimensions. Thus, there is often a benefit in resetting the unit-cell dimensions to their expected values, if they are known. This is governed by the toggle parameter `FIX_UNIT_CELL`, which, if set to ON, will reset the unit cell to the six parameters defined in `UNIT_CELL`. The unit-cell dimensions of cubic space groups are perfectly correlated with the beam energy, and hence further refinement is not needed. However, other space groups may require refinement of unit-cell dimensions. The unit-cell lengths can be refined by toggling the parameters `REFINE_UNIT_CELL_A`, `REFINE_UNIT_CELL_B` and `REFINE_UNIT_CELL_C`, and the unit-cell angles by `REFINE_UNIT_CELL_ALPHA`, `REFINE_UNIT_CELL_BETA` and `REFINE_UNIT_CELL_GAMMA`. *Cppxfel* ensures that unit-cell lengths are kept identical if strictly required by the chosen space group.

```
Wavelength histogram for shot-s01-20130316164914875.img
1.3131
1.32769
1.34228
1.35687
1.37146 ..
1.38605 ...
1.40064 ....
1.41523 .
1.42982 .........
1.44441 ..................................
1.459   ........................................
1.47359 ........................................
1.48818 .......
1.50277 .
1.51736 .....
1.53195 .
1.54654 .
1.56113
1.57572 .
1.59031 .
```

Figure 4
Example of an Ewald sphere wavelength histogram produced during initial orientation matrix refinement, which is indicative of a successfully refined image.

The output from integration will produce histograms showing the distribution of strong reflections over the corresponding Ewald sphere wavelengths at which they are illuminated (Fig. 4). This is primarily governed by the crystal orientation, and optionally by altering the unit-cell dimensions as well. One should modify these parameters to aim to produce histograms which are within a narrow range of Ewald sphere wavelengths with as few reflections as possible at the edges of the distribution. The results of this integration stage are a series of MTZ files, one per crystal. If DIALS was configured to index multiple crystals per lattice, this may result in more MTZ files than image files. The software also produces a list of updated matrices to send into the next stage of the pipeline, post-refinement. These are stored in a plain text file as a default set of files `*-orientations.dat` which can be fed into the various stages of data processing.

3.3. Post-refinement

Post-refinement is implemented as previously described (Ginn, Brewster *et al.*, 2015), with some improvements and alterations. There are certain parameters to carefully consider changing for optimal post-refinement. The most important parameters govern the choice of starting values for refining models for individual images and their step sizes during minimization. The following parameters in particular may require changing from their default values: (*a*) the starting value of the reciprocal lattice point size, (*b*) the parameters in the calculation of mean wavelength and (*c*) the step size of orientation matrix refinement. The reciprocal lattice point size determines the radius of the reciprocal lattice point in reciprocal ångströms at the origin of reciprocal space, before any additions due to mosaicity. The mean wavelength is the estimated wavelength of the X-ray pulse for that particular exposure, bearing in mind this scales directly with the overall unit-cell size.

Generally, the default parameters for post-refinement will provide a good quality set of intensities for good quality diffraction obtained at an XFEL. However, the user may wish to tweak parameters to achieve small improvements in the $CC_{1/2}$ or R_{split} values. For example, the step size of the orientation matrix refinement (`STEP_SIZE_ORIENTATION`) can specify how widely sampled the function landscape should be in the horizontal and vertical rotation parameters, and although this has a default setting of 0.06°, it may be increased or reduced in order to step over or avoid local minima in the function landscape. It may need to be reduced so as not to leave the over-predicted reflection zone, which would lead to extremely low numbers of reflections being accepted for a number of images. Conversely, it may need increasing in order to sufficiently explore the function landscape to find an appropriate minimum. It is also possible to further refine the unit-cell dimensions. The closer the initial estimate to the correct solution, the more likely it is that an image will refine to the global minimum and not a local minimum. The begin-

ning of post-refinement is also the point at which any indexing ambiguity is broken using an implementation of a cluster algorithm (Brehm & Diederichs, 2014), which is currently only supported for twofold indexing ambiguities.

Reciprocal lattice point sizes are proportional to the domain size of the crystal and expand with resolution according to mosaicity (Rossmann *et al.*, 1979). This model is also implemented in *cctbx.xfel* (Sauter *et al.*, 2014). The starting values of these parameters are defined by `INITIAL_RLP_SIZE` (default 1×10^{-4} Å^{-1}, which usually needs to be raised) and `INITIAL_MOSAICITY` (default 0). Changing the reciprocal lattice point size has the greatest effect at low resolution. Thus, errors in the reciprocal lattice point size can be judged by considering the merging statistics at low resolution and examining plots of partiality agreement by resolution (see §3.5). This can also help judge the requirement for including a mosaicity term: if there is under-prediction at high resolution which is not adequately addressed by the bandwidth parameter.

The illuminated wavelength is calculated by taking the mean Ewald sphere wavelength of all strong reflections. Thus, the threshold of a strong reflection should be set appropriately, using `REFINEMENT_INTENSITY_THRESHOLD` and specifying an absolute intensity cutoff above which reflections are included in the calculation. If outliers are causing problems in the determination of the beam wavelength, they can be excluded by specifying a wavelength beyond which reflections are excluded. This is determined by lower and upper wavelength boundaries following the keyword `WAVELENGTH_RANGE`. These parameters may need altering to produce the best merging statistics or can be judged by examining plots of partiality agreement by resolution (see §3.5).

Refinement of unit-cell parameters can be switched on and off by toggling the parameters `OPTIMISING_UNIT_CELL_A`, `OPTIMISING_UNIT_CELL_B` and `OPTIMISING_UNIT_CELL_C`, which can provide further improvements on the initial estimates from the integration stage for appropriate space groups.

Nelder–Mead minimization (Nelder & Mead, 1965) is used to refine to the local minimum, and the initial starting values for Nelder–Mead can have a great influence on its ability to fully minimize. Hence, the step sizes can be specified for each of the parameters according to the `STEP_SIZE_*` parameters as defined in the manual.

The primary parameter pertaining to partiality is `PARTIALITY_CUTOFF`, which determines the minimum partiality which would be accepted for merging and evaluation of the target function per image. The default is set to 0.2, but this may be raised to around 0.3 for poor crystals or lowered to 0.05 for well behaved crystals. The slicing of the numeric integration for partiality calculation, particularly at low resolution where the nest of Ewald spheres (representing different incident energies) is very tightly packed, may need to be finer for reciprocal lattice points of a larger size. The maximum partiality slicing is defined by `MAX_SLICES`, set to a default of 100, which is employed at resolutions below `CAREFUL_RESOLUTION`, where care must be taken to ensure adequate numerical integration. If the multiplicity of low-resolution reflections is particularly low, this parameter may need to be increased, as the multiplicity may be due to insufficient slicing to properly estimate how the bandwidth of the beam intersects the reciprocal lattice point. If `MAX_SLICES` is set too coarse this may result in suspiciously low $R_{\rm split}$ merging statistics and high $R_{\rm work}$ and $R_{\rm free}$ values.

Two methods of outlier rejection are employed, but depending on detector and data quality, one or both may need disabling or modifying. Rejection of reflections on individual images as described previously (Ginn, Brewster *et al.*, 2015) can be turned off using the parameter `CORRELATION_REJECTION`. Rejection at the merging stage by standard deviation from the mean can be turned off entirely using the parameter `OUTLIER_REJECTION`, or the standard deviation threshold can be modified from the default value of 1.8 using `OUTLIER_REJECTION_SIGMA`.

If a set of reference intensities is available, this can be provided as a path to an MTZ file using the command `INITIAL_MTZ`. This is used to provide an initial reference set of intensities, which are refined against but not included in the merges at the end of the macrocycles. This allows data sets to be processed with a more than twofold indexing ambiguity if the reference set provided is untwinned.

3.4. Final merge

Running `cppxfel.input_gen` creates a template input file `merge.txt` which should be run after `integrate.txt` and `refine.txt`. This merging stage can have outlier rejection or image rejection parameters different from earlier stages of refinement. This will create a final MTZ file named `remerged.mtz` which will have corrected sigma values for use in other crystallographic programs. These are calculated by a weighted standard deviation function divided by the square root of the number of observations, so are more akin to a term of precision. Setting the toggle `MERGE_ANOMALOUS` will output anomalous differences in Friedel pairs to `anomalous_diff.mtz` which can be used to assess anomalous data quality.

3.5. Analysing the data output

Merging statistics are a simple measure of data quality for a set of reflections and both $CC_{1/2}$ (Karplus & Diederichs, 2012) and $R_{\rm split}$ (White *et al.*, 2013) are output during each macrocycle of post-refinement. Errors in integration or data refinement strategies are likely if these values fail to tend towards 1.0 and 0.0, respectively. One should ensure wavelength histograms as in Fig. 4 are being obtained during the integration step, and check that the plots of partiality agreement over resolution shells do not show global errors in reciprocal lattice point size or calculation of mean wavelength. Data would be expected to decay with resolution.

Merged MTZ files are written out to files named `allMerge[n].mtz` where `[n]` is the cycle number of post-refinement. Individual halves of the data set used for $R_{\rm split}$ and $CC_{1/2}$ calculation, including a full reflection list, are written to `half1Merge[n].mtz` and `half2Merge[n].mtz` and can be

compared after refinement is complete using `cppxfel.run -cc half1Merge[n].mtz half2Merge[n].mtz`.

Other merging statistics are available within the program for unmerged MTZ files. A large MTZ file containing unmerged data is written to `unmerged[n].mtz` which can be used to calculate R_{merge}, R_{meas} and R_{pim} using the command `cppxfel.run -command unmerged[n].mtz`, where `command` is replaced by `rmerge`, `rmeas` or `rpim`.

It may be useful to analyse the quality of post-refinement of individual crystals against the reference data set to gain the best insight into errors in the post-refinement strategy. To do this, partiality agreement data in CSV format can be generated. This uses the command `cppxfel.run -partiality reference.mtz crystal.mtz [maxRes]`, where `[maxRes]` is the maximum resolution for analysis. This will generate CSV files for each of four resolution bins. Each file will contain a list of Ewald sphere wavelengths for each reflection in the crystal MTZ file, the percentage of the merged reflection intensity per reflection, the calculated partiality and the resolution of each reflection. The CSV files will be named `partiality_[m].csv`, where `[m]` is the bin number. These can then be imported into any graph drawing software to plot the data and identify pathologies as shown in Fig. 5.

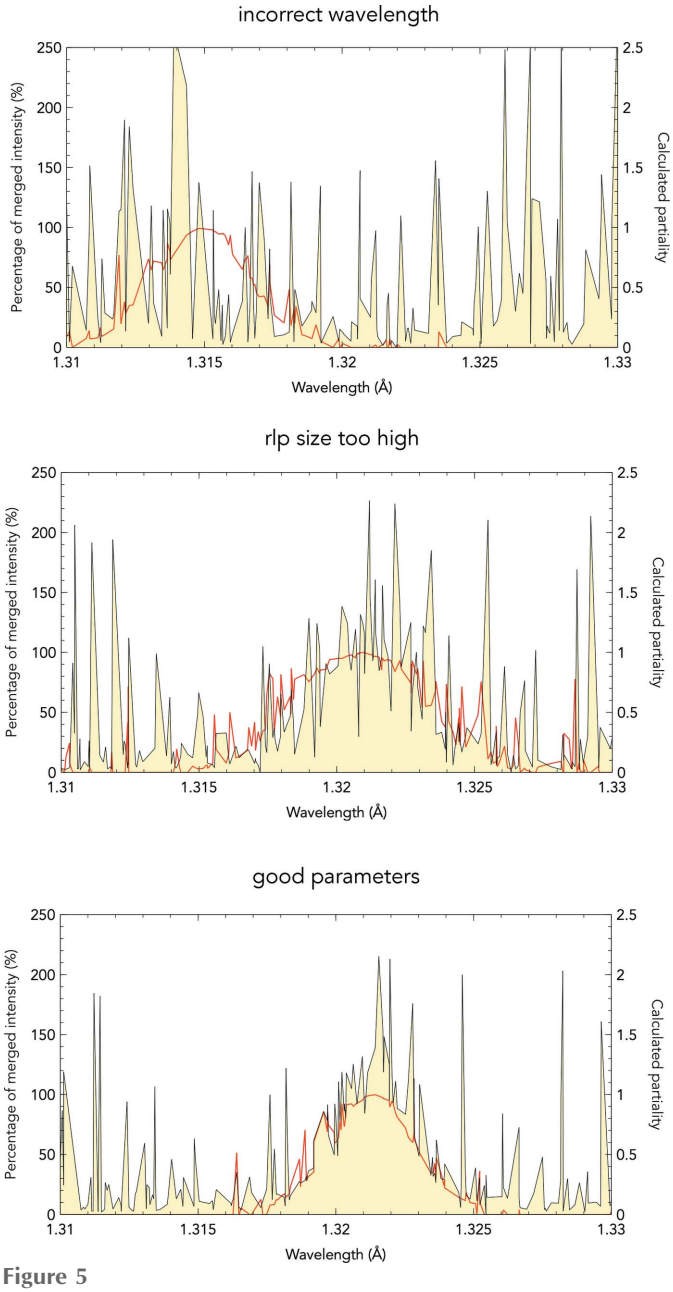


Figure 5
Plotting calculated partiality (red line) against estimate of partiality (yellow fill) by comparing against the reference data set, plotted against the Ewald sphere wavelength for the midpoint of every reflection under 2.5 Å resolution. These plots represent incorrect parameter choices. The mean wavelength of the X-ray pulse was manually reconfigured to be off by 0.005 Å and the effect of this error is shown in the top panel. The crystal is rotated by the minimization method far off the true solution in order to compensate for the incorrect wavelength, to the point that the true wavelength peak is no longer recognizable. The middle panel had the reciprocal lattice point (rlp) size inflated by 100%, also leading to an incorrect crystal rotation to compensate by broadening the distribution of wavelengths. Finally, the bottom panel shows a set of good parameter choices which lead to a good refinement solution. This image come from a data set collected on CPV17 crystals.

4. Conclusion

This paper provides an introduction to data processing using *cppxfel* to allow developers to understand the impact of the models used in this software package and to provide a guide to facilitate the effective use of the software by experienced XFEL users. The code is open source, and the associated Wiki entries may be accessed at http://viper.lbl.gov/cctbx.xfel/index.php/Cppxfel. This includes installation instructions and a step-by-step tutorial to process the provided CPV17 polyhedrin diffraction images.

Acknowledgements

DIS was supported by the Medical Research Council, grant G1000099. HMG was supported by the Wellcome Trust (studentship 075491/04). NKS was supported by US National Institutes of Health grants GM095887 and GM102520. We are grateful to Luc Bourhis for integrating the Boost thread libraries within the *cctbx* package. We are also grateful to Helen Duyvesteyn for providing invaluable comments for the manuscript and software. Portions of this research were carried out at the Linac Coherent Light Source (LCLS) at the SLAC National Accelerator Laboratory. LCLS is an Office of Science User Facility operated for the US Department of Energy Office of Science by Stanford University. Administrative support was received from the Wellcome Trust, grant 090532/Z/09/Z. This is a contribution from the Oxford Instruct Centre.

References

Barends, T. R. M., Foucar, L., Ardevol, A. *et al.* (2015). *Science*, 445–450.
Barends, T. R. M., Foucar, L., Botha, S. *et al.* (2014). *Nature*, **505**, 244–247.
Barends, T. R. M., Foucar, L., Shoeman, R. L. *et al.* (2013). *Acta Cryst.* D**69**, 838–842.
Brehm, W. & Diederichs, K. (2014). *Acta Cryst.* D**70**, 101–109.

Chapman, H. N., Caleman, C. & Timneanu, N. (2014). *Philos. Trans. R. Soc. London Ser. B*, **369**, 20130313.
Chapman, H. N., Fromme, P. *et al.* (2011). *Nature*, **470**, 73–77.
Ekeberg, T. *et al.* (2015). *Phys. Rev. Lett.* **114**, 098102.
Gildea, R. J., Waterman, D. G., Parkhurst, J. M., Axford, D., Sutton, G., Stuart, D. I., Sauter, N. K., Evans, G. & Winter, G. (2014). *Acta Cryst.* D**70**, 2652–2666.
Ginn, H. M., Brewster, A. S., Hattne, J., Evans, G., Wagner, A., Grimes, J. M., Sauter, N. K., Sutton, G. & Stuart, D. I. (2015). *Acta Cryst.* D**71**, 1400–1410.
Ginn, H. M., Messerschmidt, M., Ji, X., Zhang, H., Axford, D., Winter, G., Brewster, A. S., Hattne, J., Wagner, A., Grimes, J. M., Sauter, N. K., Sutton, G. & Stuart, D. I. (2015). *Nat. Commun.* **6**, 6435.
Grosse-Kunstleve, R. W., Sauter, N. K., Moriarty, N. W. & Adams, P. D. (2002). *J. Appl. Cryst.* **35**, 126–136.
Hattne, J. *et al.* (2014). *Nat. Methods*, **11**, 545–548.
Kang, Y. *et al.* (2015). *Nature*, **523**, 561–567.
Karplus, P. A. & Diederichs, K. (2012). *Science*, **336**, 1030–1033.
Kern, J., Alonso-Mori, R., Hellmich, J. *et al.* (2012). *Proc. Natl Acad. Sci. USA*, **109**, 9721–9726.
Kern, J., Alonso-Mori, R., Tran, R. *et al.* (2013). *Science*, **340**, 491–495.
Kern, J., Tran, R., Alonso-Mori, R. *et al.* (2014). *Nat. Commun.* **5**, 4371.
Kirian, R. A., Wang, X., Weierstall, U., Schmidt, K. E., Spence, J. C., Hunter, M., Fromme, P., White, T., Chapman, H. N. & Holton, J. (2010). *Optics Express*, **18**, 5713–5723.
Kupitz, C. *et al.* (2014). *Nature*, **513**, 261–265.
Liu, W. *et al.* (2013). *Science*, **342**, 1521–1524.
Nakane, T., Joti, Y., Tono, K., Yabashi, M., Nango, E., Iwata, S., Ishitani, R. & Nureki, O. (2016). *J. Appl. Cryst.* **49**, doi:10.1107/S1600576716005720.
Nakane, T., Song, C. *et al.* (2015). *Acta Cryst.* D**71**, 2519–2525.
Nelder, J. A. & Mead, R. (1965). *Comput. J.* **7**, 308–313.
Neutze, R. & Moffat, K. (2012). *Curr. Opin. Struct. Biol.* **22**, 651–659.
Neutze, R., Wouts, R., van der Spoel, D., Weckert, E. & Hajdu, J. (2000). *Nature*, **406**, 752–757.
Rossmann, M. G., Leslie, A. G. W., Abdel-Meguid, S. S. & Tsukihara, T. (1979). *J. Appl. Cryst.* **12**, 570–581.
Sauter, N. K. (2015). *J. Synchrotron Rad.* **22**, 239–248.
Sauter, N. K., Hattne, J., Brewster, A. S., Echols, N., Zwart, P. H. & Adams, P. D. (2014). *Acta Cryst.* D**70**, 3299–3309.
Sauter, N. K., Hattne, J., Grosse-Kunstleve, R. W. & Echols, N. (2013). *Acta Cryst.* D**69**, 1274–1282.
Suga, M., Akita, F., Hirata, K., Ueno, G., Murakami, H., Nakajima, Y., Shimizu, T., Yamashita, K., Yamamoto, M. & Ago, H. (2015). *Nature*, **517**, 99–103.
Tenboer, J., Basu, S., Zatsepin, N., Pande, K., Milathianaki, D., Frank, M., Hunter, M., Boutet, S., Williams, G. J. & Koglin, J. E. (2014). *Science*, **346**, 1242–1246.
The HDF Group (1997). Hierarchical Data Format, Version 5, http://www.hdfgroup.org/HDF5.
Waterman, D. G., Winter, G., Parkhurst, J. M., Fuentes-Montero, L., Hattne, J., Brewster, A., Sauter, N. K. & Evans, G. (2013). *CCP4 Newsl. Protein Crystallogr.* **49**, 16–19.
White, T. A., Barty, A., Stellato, F., Holton, J. M., Kirian, R. A., Zatsepin, N. A. & Chapman, H. N. (2013). *Acta Cryst.* D**69**, 1231–1240.
White, T. A., Kirian, R. A., Martin, A. V., Aquila, A., Nass, K., Barty, A. & Chapman, H. N. (2012). *J. Appl. Cryst.* **45**, 335–341.
White, T. A., Mariani, V., Brehm, W., Yefanov, O., Barty, A., Beyerlein, K. R., Chervinskii, F., Galli, L., Gati, C., Nakane, T., Tolstikova, A., Yamashit, K., Yoon, C. H., Diederichs, K. & Chapman, H. N. (2016). *J. Appl. Cryst.* **49**, 680–689.

STRUCTURAL BIOLOGY
ISSN 2059-7983

Fixed target combined with spectral mapping: approaching 100% hit rates for serial crystallography

Saeed Oghbaey,[a]‡ Antoine Sarracini,[a]‡ Helen M. Ginn,[b,c]‡ Olivier Pare-Labrosse,[a] Anling Kuo,[d] Alexander Marx,[e] Sascha W. Epp,[e] Darren A. Sherrell,[c] Bryan T. Eger,[d] Yinpeng Zhong,[e] Rolf Loch,[e] Valerio Mariani,[f] Roberto Alonso-Mori,[g] Silke Nelson,[g] Henrik T. Lemke,[g] Robin L. Owen,[c] Arwen R. Pearson,[h] David I. Stuart,[b,c] Oliver P. Ernst,[d,i] Henrike M. Mueller-Werkmeister[a,d,e] and R. J. Dwayne Miller[a,e,h]*

[a]Department of Physics and Chemistry, University of Toronto, 60 St George Street, Toronto, ON M5S 1A7, Canada, [b]Division of Structural Biology, The Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford OX3 7BN, England, [c]Diamond Light Source, Harwell Science and Innovation Campus, Didcot OX11 ODE, England, [d]Department of Biochemistry, University of Toronto, 1 King's College Circle, Toronto, Ontario, Canada, [e]Atomically Resolved Dynamics, Max-Planck-Institute for the Structure and Dynamics of Matter, Luruper Chaussee 149, Hamburg, Germany, [f]Center for Free-Electron Laser Science, DESY, 22607 Hamburg, Germany, [g]SLAC National Accelerator Laboratory, 2575 Sand Hill Road, Menlo Park, CA 94025, USA, [h]Hamburg Centre for Ultrafast Imaging, University of Hamburg, Hamburg, Germany, and [i]Department of Molecular Genetics, University of Toronto, 1 King's College Circle, Toronto, ON M5S 1A8, Canada. *Correspondence e-mail: dwayne.miller@mpsd.mpg.de

Received 11 February 2016
Accepted 4 July 2016

Edited by E. F. Garman, University of Oxford, England

‡ These authors contributed equally to this work.

Keywords: femtosecond time-resolved crystallography; high-throughput serial crystallography; fixed-target sample delivery; room-temperature crystallography.

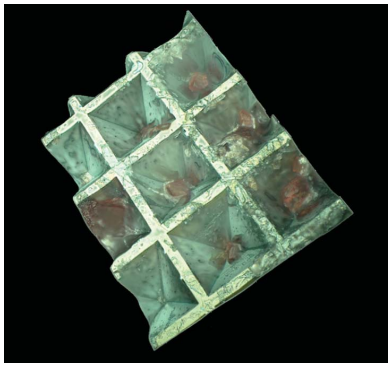
Supporting information: this article has supporting information at journals.iucr.org/d

The advent of ultrafast highly brilliant coherent X-ray free-electron laser sources has driven the development of novel structure-determination approaches for proteins, and promises visualization of protein dynamics on sub-picosecond timescales with full atomic resolution. Significant efforts are being applied to the development of sample-delivery systems that allow these unique sources to be most efficiently exploited for high-throughput serial femtosecond crystallography. Here, the next iteration of a fixed-target crystallography chip designed for rapid and reliable delivery of up to 11 259 protein crystals with high spatial precision is presented. An experimental scheme for predetermining the positions of crystals in the chip by means of *in situ* spectroscopy using a fiducial system for rapid, precise alignment and registration of the crystal positions is presented. This delivers unprecedented performance in serial crystallography experiments at room temperature under atmospheric pressure, giving a raw hit rate approaching 100% with an effective indexing rate of approximately 50%, increasing the efficiency of beam usage and allowing the method to be applied to systems where the number of crystals is limited.

1. Introduction

X-ray diffraction is currently the standard approach for high-resolution structure determination of proteins, and is conventionally carried out at synchrotron light sources. Recent developments in X-ray sources have led to the realisation of hard X-ray free-electron lasers (XFELs). The advantage of these fourth-generation X-ray light sources is their high peak brilliance and improved beam coherence in contrast to synchrotron-generated X-ray radiation. Furthermore, X-rays are generated as short pulses in the 10–100 fs regime, enabling experiments to probe ultrafast chemical and biochemical dynamics (Kern *et al.*, 2014; Tenboer *et al.*, 2015; Barends *et al.*, 2015).

Protein crystallography experiments at XFELs are based on the approach now known as serial femtosecond crystallography (SFX), where the above beam properties allow single



© 2016 International Union of Crystallography

diffraction patterns to be recorded from micrometre-sized protein crystals in a few tens of femtoseconds (Chapman *et al.*, 2011). Using this approach, it is possible to collect a large number of diffracted photons from each protein crystal before the onset of X-ray radiation damage ('diffraction before destruction'; Boutet *et al.*, 2012; Barty *et al.*, 2011). However, a peak brilliance 10^8 times higher than that of a synchrotron source and the short duration of the pulse at an XFEL do result in severe radiation damage to each crystal and make a sample change between each X-ray shot necessary (Barty *et al.*, 2014). This requires either a very large number of small protein crystals or the translation and exposure of multiple regions of larger crystals in order to obtain a sufficient sampling of reciprocal space to allow protein structure determination. Samples must be delivered into the X-ray beam with high spatial precision and, ideally, a sufficient speed/repetition rate to fully use the X-ray beam flux and to avoid misalignment or mistiming in presenting the target to the X-ray beam. Currently, SFX experiments predominantly use variations of a liquid jet (DePonte *et al.*, 2008; Boutet *et al.*, 2012; Sierra *et al.*, 2012), LCP injectors (Weierstall *et al.*, 2014) in vacuum, a grease matrix (Sugahara *et al.*, 2015) or an acoustic injector (Roessler *et al.*, 2016) for sample delivery. In the liquid-jet method, the crystal solution is ejected at high velocity (albeit with a small net volume flow) in a stream onto which, for time-resolved experiments, the pump (laser) and probe (X-ray) pulses are incident. These approaches can support structure determination, but suffer from low hit rates since the random nature of the position of crystals within the jet means that while all X-ray pulses may intersect with the jet, few hit a crystal, wasting photon flux. Even worse for many experiments where crystal production is difficult or expensive, the vast majority of the crystals are wasted, since to maintain the thin flow of crystals the jet velocity needs to be high, increasing the number of crystals which do not intersect an X-ray pulse. The hit rate has been estimated to be as low as one hit per 1600 crystals in some cases (Weierstall *et al.*, 2014). Furthermore, the maximum pressure that the crystals can tolerate before flow-induced damage occurs limits the crystal flow. The flowing stream of sample also adds to the difficulty (compounded by the destructive nature of the X-ray pulse) of taking diffraction reference shots, which would simplify pump–probe experiment data analysis. It is therefore, highly desirable to develop sample-delivery approaches with hit rates that approach 100% and which provide the possibility of recording reference data. For time-resolved studies, the requirement for a high hit rate is particularly acute since sufficient data for structure analysis must be obtained for each and every time point. This significantly increases the volume of sample needed, bearing in mind the high signal to noise required to resolve the very small changes in structure that direct chemical and biological processes or other structural transitions.

Fixed targets provide an attractive solution to the problems described above, and in principle allow the convenient harvesting of crystals and their precise presentation to laser and X-ray beams. Several fixed-target systems have been developed. These include polydimethylsiloxane (PDMS)-based chips (Guha *et al.*, 2012; Lyubimov *et al.*, 2015), in which the crystals sit on the substrate, resulting in noticeable background scatter from the substrate material. An alternate approach utilizes a thin film as a substrate for crystal sample deposition (Hunter *et al.*, 2014; Murray *et al.*, 2015) or patterned silicon substrates (Roedig *et al.*, 2015, 2016) for rapid sample positioning and polycarbonate grids coupled with a goniometer (Baxter *et al.*, 2016; Cohen *et al.*, 2014). Most of these thin-film approaches contribute less background but eliminate the possibility of prior mapping out of crystal positions. The use of a large single crystal as a fixed target has also been implemented (Cohen *et al.*, 2014; Chreifi *et al.*, 2016; Hirata *et al.*, 2014; Keedy *et al.*, 2015; Zhou *et al.*, 2015), but large crystals with the appropriate shape and large accessible surface area and thickness are not easily available for many samples of interest.

We have previously reported a crystallography chip, made by silicon nanofabrication methods (Zarrine-Afsar *et al.*, 2012), for use at microfocus synchrotron beamlines and for time-resolved studies at XFELs that addresses all of the above problems (Mueller *et al.*, 2015). The chip represents a significant improvement over injector schemes, as it requires a smaller amount of sample and effectively minimizes background scatter compared with other fixed-target approaches owing to the removal of excess mother liquor/solvent, a feature that was specifically incorporated into the design of the chip-loading process. The chip contains precisely defined 'features' (pit or well shapes) in which the crystals are trapped. The feature sizes and chip-support layer match the crystal size distribution of the sample under study. The chip structure allows a large number of microcrystals to be loaded, using turbulent fluid flow to introduce random crystal orientations to ensure sufficient sampling of reciprocal space (Mueller *et al.*, 2015). The loaded chip can be mounted on a high-speed translation stage for crystal positioning in the X-ray beam (Sherrell *et al.*, 2015). The accuracy of the nanofabrication process allows precise positioning of the crystals, increasing the hit rate tremendously while keeping sample waste to a minimum, owing to both the relatively high occupancy and the possibility of recovering crystals that were not trapped during the loading process, since all eluent can be collected. The chip also enables the recording of reference diffraction patterns and thus, radiation damage and signal to noise permitting, use of the ratio method for differential detection of the very small early structural changes typically involved in time-resolved studies (Coppens *et al.*, 2009; Mueller *et al.*, 2015). This effectively reduces the volume of data required for structural dynamics studies and further increases the overall efficiency with respect to sample consumption.

In our initial reports the chip could not be operated with the maximum possible hit rate as all features on the chip were probed sequentially, irrespective of the presence or absence of a crystal (Zarrine-Afsar *et al.*, 2012; Mueller *et al.*, 2015). In the present design, we have further improved the chip to reduce multiple crystal instances per loading site, and added a system of fiducial marks for rapid alignment with well defined

reference points and quick identification of chip compartments. The fiducial system enables error-free position determination and identification of each crystal in the array. Finally, we have developed a high-speed single-crystal spectrometer to map the crystal positions prior to the diffraction experiment. High-speed mapping is important for crystals with finite lifetimes once mounted, and increases hit rates substantially as only features containing crystals are sampled at the beamline.

As a proof of principle, we demonstrate here that this chip concept, with a fiducial system and mapping protocol, can achieve a raw hit rate of up to 85%. When inherent variations in crystal quality as well as intensity fluctuations owing to the SASE profile of the XFEL beam are taken into account, this approaches the maximum achievable sampling efficiency. Furthermore, the method facilitates a data-collection rate that comes close to the fully optimized use of the XFEL X-ray flux with low sample consumption, thus opening up the study of precious protein systems. Specifically, we demonstrate the efficacy of the chip concept using carboxymyoglobin, as a prelude to femtosecond time-resolved studies of ligand dissociation from the haem binding site.

2. Materials and methods

2.1. Nanofabrication

The chip is made out of a single-side polished silicon wafer (WRS material) with a thickness of 550 µm coated with 50 nm low-pressure chemical vapour-deposition (LPCVD) silicon nitride, applied as an etching mask for KOH etching. Conventional lithography is utilized to transfer the photomask pattern onto an organic photoresist (Shiply S1818), which is spin-coated onto the silicon wafer. After exposure to UV light, the pattern is developed and cleaned. In the next step, the silicon nitride under the photoresist is etched away using SF_6 plasma reactive-ion etching. The photoresist is used in the plasma-etching stage as an etch mask and must be stripped away afterwards. The wafer is then placed in a supersaturated aqueous KOH solution [40%(*w*/*w*)] at 80°C to etch away bulk silicon until the desired feature depth is reached. A similar process is carried out for developing the back-side features until the desired hole size is reached. The fabrication process is shown in Fig. 1. The wafer is then cleaned using a standardized cleaning process with a solution consisting of H_2O_2:NH_4OH:H_2O in a ratio of 1:1:5 by volume at 70°C, which leaves the chip surface hydrophilic. Finally, the wafer is separated into individual chips by dicing.

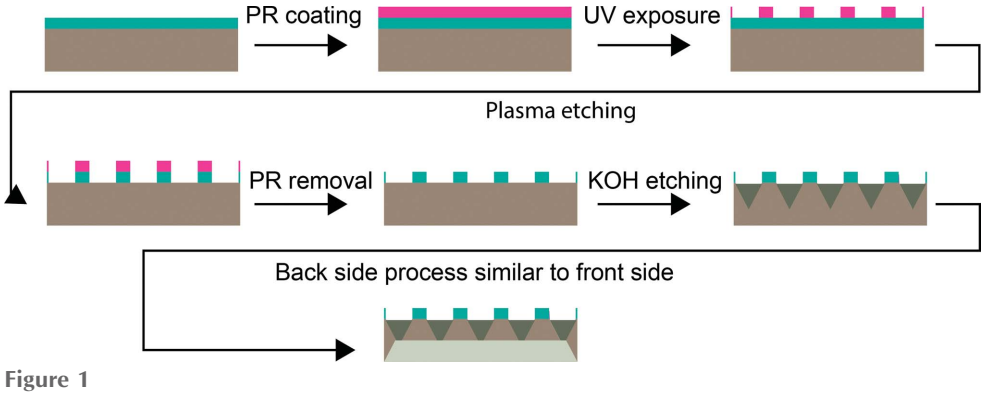


Figure 1
Fabrication scheme for the crystallography chip. The process starts with conventional UV lithography to transfer the pattern of loading sites to a silicon wafer [grey, with a layer of silicon nitride (SiN) in cyan] coated with photoresist (pink). The pattern is then transferred to the SiN layer using reactive-ion etching and the photoresist is stripped. The SiN layer plays the role of the mask for wet etching in the subsequent steps. KOH etching results in a trapezoidal shape that enhances loading efficiency. The same procedure is repeated for the back side of the chip using a different photomask to make it thin enough to fit a specific size of crystals.

2.2. Protein expression, purification and crystallization

2.2.1. Sperm whale myoglobin expression and purification. The sperm whale (*Physeter catodon*) myoglobin (SWMb) expression plasmid (pMb413a) with a D122N mutation was kindly provided by Professor John S. Olson (Rice University, USA) and the protein was expressed constitutively in *Escherichia coli* BL21(DE3) cells at 37°C for 3 days. SWMb with bound carbon monoxide (SWMb–CO) was purified as described previously, with some modifications (Springer & Sligar, 1987; Mueller *et al.*, 2015). Briefly, (i) all buffers used in purification were saturated with CO gas at 1 atm; (ii) after ammonium sulfate precipitation, the collected precipitant was dissolved in a small volume of 20 m*M* sodium phosphate pH 6.0 and dialyzed overnight against this buffer; (iii) the dialyzed protein was concentrated and applied onto a CM Sepharose (GE Healthcare Life Sciences) fast flow column which had been pre-equilibrated with the same buffer and a linear pH gradient of 20 m*M* sodium phosphate pH 6.0 to 50 m*M* sodium phosphate pH 9.0 was used for elution; (iv) eluted protein fractions were concentrated and applied onto a Superdex 75 gel-filtration column (GE Healthcare Life Sciences) in 20 m*M* Tris–HCl, 150 m*M* NaCl buffer pH 8.0 to remove aggregates; (v) the peak fractions were combined, solid sodium dithionite was added (20 m*M* final concentration) and, after 10 min incubation, the protein was rapidly applied onto a Sephadex G-25 column equilibrated with 10 m*M* Tris–HCl buffer pH 9.0 under 1 atm of CO (Li *et al.*, 1994). The eluate containing SWMb–CO was concentrated for crystallization.

2.2.2. SWMb–CO crystallization. To ensure that the SWMb crystals were grown in the reduced CO-bound form, the crystallization buffers were saturated with CO gas. The SWMb–CO crystals were grown in 3 ml Monoject blood collection tubes (Covidien, Mansfield, Massachusetts, USA) prefilled with CO gas at 1 atm. 180–200 µl crystallization solution containing 50–60 mg ml^{-1} protein in 10 m*M* Tris–HCl pH 9.0, 2.5–2.6 *M* ammonium sulfate was added to the Monoject tubes. Seeding was used to promote the nucleation of

large numbers of small crystals. The SWMb–CO crystals were crushed in 10 m*M* Tris–HCl pH 9.0, 3.2 *M* ammonium sulfate. The seeds were diluted 1:100 under the same buffer conditions and 10 µl was added to each tube. The size distribution of the final crystals was controlled by adding 180 µl 10 m*M* Tris–HCl pH 9.0, 3.2 m*M* ammonium sulfate pre-saturated with CO gas to each tube after crystals appeared. This method produces crystals of appropriate sizes, corresponding to the chip parameters, as described previously (Mueller *et al.*, 2015). Crystals were loaded onto the chip as shown in Fig. 2. To prevent salt-crystal formation, all experiments for chip loading were performed in a portable glove bag (Sigma–Aldrich) at a temperature of 24°C and a humidity of 60%.

2.3. *In situ* microspectrometer

A fibre-coupled microfocus absorption spectroscopy setup for sample characterization and chip mapping was developed (adapted from Kirchner *et al.*, 2014). The setup utilizes a fibre-coupled broadband light source, off-axis parabolic mirrors, a fibre-coupled USB spectrometer and a three-axis translation stage (SmarAct) for accurate sample positioning (similar to the beamline end station described in Sherrell *et al.*, 2015).

A deuterium–halogen lamp (Ocean Optics; 215–2500 nm) or an LED lamp (Ocean Optics; 425–650 nm) was used as a fibre-coupled light source. The light source is coupled to a 200 µm core diameter fibre. The first off-axis parabolic mirror (OPAM) collimates and reflects the light by 90° onto a second OPAM with half the effective focal length of the first. This mirror focuses the light onto the sample. These two mirrors constitute an all-reflective 2:1 microscope, resulting in a focal spot size of nominally 100 µm at the sample. A third OPAM collimates the transmitted light onto a fourth OPAM that focuses the light into a fibre-coupled USB spectrometer sensitive to light in the wavelength range 200–850 nm (Ocean Optics; USB2000-UV-VIS). Absorption spectra were generated by taking the logarithm of the background-subtracted transmission spectra divided by the light-source spectrum.

The parabolic mirror pairs are placed in a rigid four-bar optical cage system (Thorlabs; Fig. 3*a*) using a custom-designed mount, which fixes the rotation of the mirrors. This ensures very easy and precise alignment, which does not need to be adjusted. A CMOS camera (Thorlabs) and a close-focus zoom lens (Edmond Optics), with a minimum field of view of approximately 6 mm and a large working distance, is used as an off-axis viewing system for chip alignment.

2.4. SFX data collection

X-ray data collection was performed at the XPP instrument at LCLS at room temperature with an X-ray beam size of 20 × 20 µm and 10^{10} X-ray photons per pulse with a photon energy of 9.58 keV and a nominal sample-to-detector distance of 67.8 mm under proposal LF70. A diamond (111) monochromator was used, resulting in a narrow bandwidth: $(\Delta E/E) = 5 \times 10^{-5}$. The diffraction data-collection rate was 10 Hz, which was limited by the detector-readout speed. Data were recorded using a Rayonix MX170-HS CCD detector.

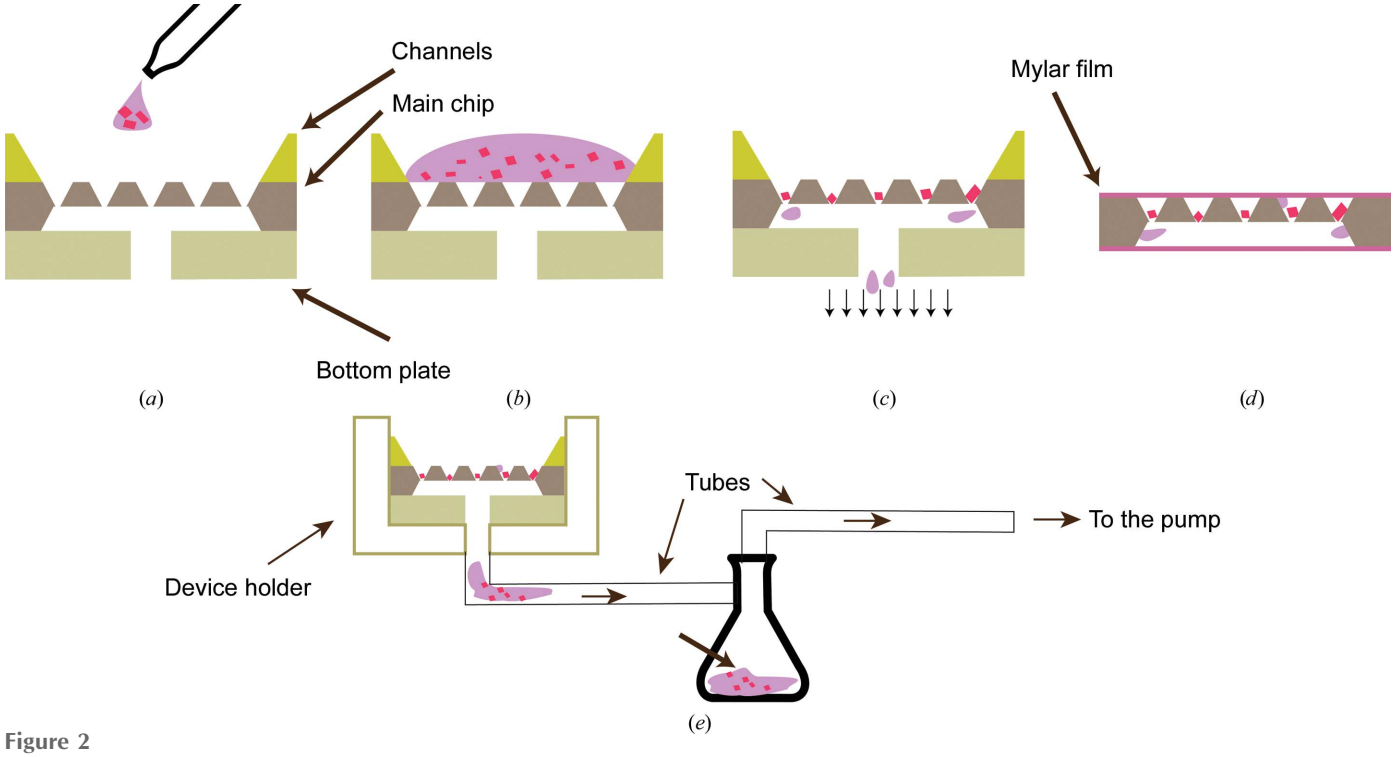
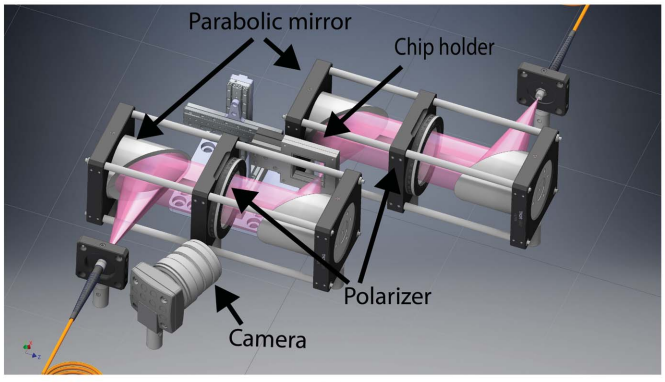


Figure 2
(*a*, *b*, *c*) Chip-loading scheme with chip components designed for crystal solution loading. The silicon piece with channels in it is placed on top of the chip to guide crystal solution into the compartment rows. The bottom plate helps to apply effective negative pressure to trap the crystals in features (*c*). (*d*) The loaded chip is wrapped in a thin film of Mylar to avoid crystal dehydration. The crystal solution removed from the chip can be collected in an additional reservoir for use with a different chip size (*e*). All figures are schematics and are not to scale.

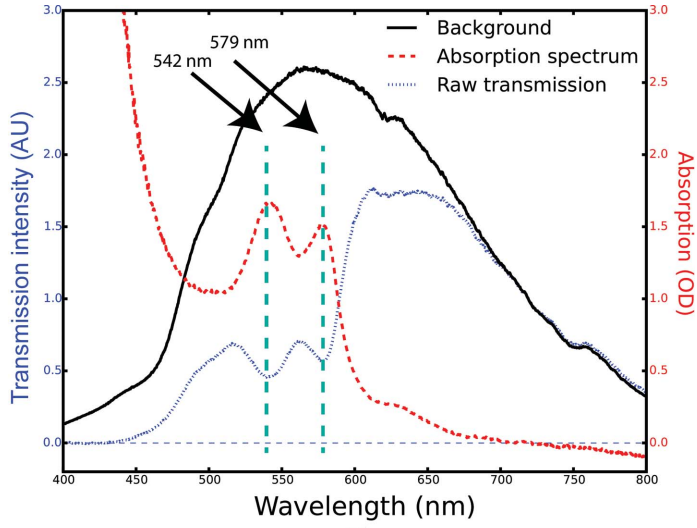
In the manner described previously (Mueller *et al.*, 2015), data collection was organized such that ten shots of the XFEL were integrated into a single frame of the detector, providing the reference or 'before' image, and in a second step a single shot of the same crystal was recorded. In a pump–probe experiment, this second image would be the only one with a possible pump interaction. We note that the first image integrated from ten individual XFEL shots was recorded with a beam attenuated roughly 30 times after the monochromator, ensuring that radiation damage is below the detection threshold. The total absorbed X-ray dose per crystal was 5.7 kGy for the sum of ten attenuated pulses plus 16.8 kGy without the attenuator, using a 480 µm sapphire attenuator. These values were calculated using the *RADDOSE-3D* software suite (Zeldin *et al.*, 2013), ignoring the effect of photoelectron escape. These are well below the damage threshold (Garman limit) at which noticeable spectroscopic changes are observed as determined in previous studies (Owen *et al.*, 2011). A broadly accepted rule of thumb in macromolecular crystallography is that a structural change of less than 20% does not contribute significantly to the final electron density (Pearson *et al.*, 2007; see the Supporting Information for safe-dose calculation).

2.5. Data analysis

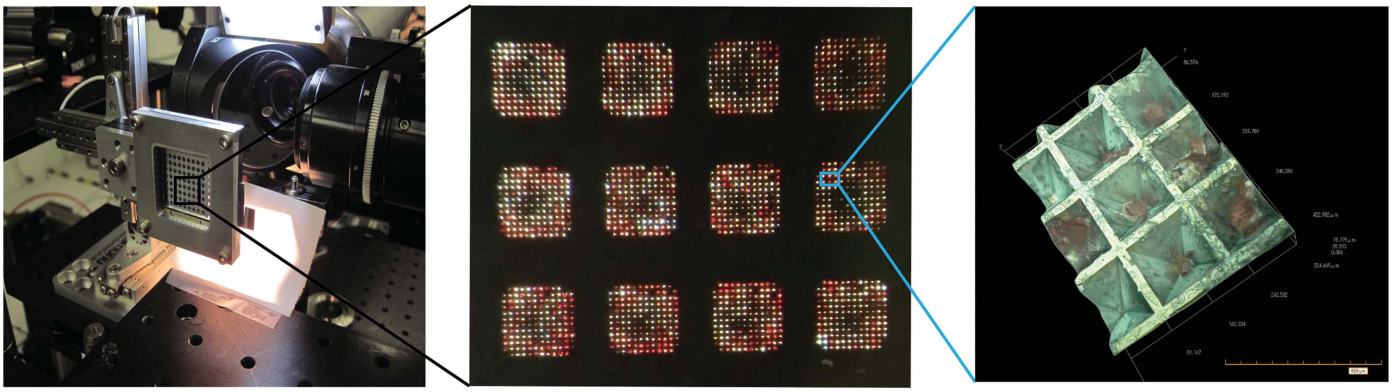
Individual images were extracted from the XTC streams and stored in HDF5 format for indexing with *Cheetah* and *indexamajig* (*CrystFEL*; White *et al.*, 2012). From the output of *indexamajig*, information was extracted to calculate both raw hit rates ($h_{\rm raw}$) and fractions of indexed images ($h_{\rm idx}$). A spot search was carried out on these extracted images using the *DIALS* spot-finding routine (Waterman *et al.*, 2016), which is based on the *XDS* algorithm (Kabsch, 2010). For integration, data merging and structure determination, with an improved method based on post-refinement and partiality correction (Ginn, Brewster *et al.*, 2015), images were indexed using the *TakeTwo* (Ginn, Roedig *et al.*, 2016) algorithm in *cppxfel* (Ginn, Evans *et al.*, 2016) and integrated using simple summation, with an integration box with a 5 × 5 foreground,



(*a*)



(*b*)



(*c*) (*d*) (*e*)

Figure 3
(*a*) Schematic representation of the high-speed spectroscopy setup for mapping the loaded chip. (*b*) Example spectrum of a carboxymyoglobin crystal recorded from a crystal-containing feature. The characteristic absorption bands for carboxymyoglobin (542 and 579 nm; Springer & Sligar, 1987) are indicated. (*c*) The chip mounted in the beam at the XPP instrument at LCLS. (*d*) A magnified picture of chip compartments loaded with crystals. The carboxymyoglobin crystals can be easily identified owing to their red colour. (*e*) Three-dimensional image of crystals trapped inside a feature (image taken using a Hirox RH2000 microscope as part of user facilities on site at SACLA, RIKEN, Japan).

Table 1
Refinement statistics.

Resolution range (Å)	42.26–1.90
Completeness (%)	100
R_{work}/R_{free} (%)	21.1/26.0
No. of non-H atoms	
Protein	1225
Sulfate	1
Water	20
R.m.s. deviations	
Bonds (Å)	0.016
Angles (°)	1.8
Ramachandran plot	
Most favoured (%)	96.7
Allowed (%)	3.3

one-pixel buffer and two layers of additional surrounding pixels for background subtraction. The detector distance was refined to 68.6 mm and the wavelength to 1.293 Å, in contrast to the values of 67.8 mm and 1.312 Å, respectively, which were recorded during beam time. Crystal orientations were initially refined according to a previous protocol to minimize the spread of Ewald spheres on which the reflections are observed (Ginn, Messerschmidt *et al.*, 2015), with additional refinement of the three unit-cell dimensions for the orthorhombic space group. These positions were integrated, according to the chosen shoebox, and the potential reflection was classed as a 'strong' reflection if it had an integrated intensity of 32 counts or greater. Post-refinement was carried out against a reference data set according to a previous protocol (Ginn, Brewster *et al.*, 2015) using a binary partiality of 0 or 1. A partiality of 1 is assigned to reflections for which the conventionally calculated partiality is 0.2 or greater and 0 if this is not the case. Only one cycle of post-refinement was performed after an initial Monte Carlo merge in order to refine which reflections were included and excluded in the merge, but partiality correction on these data was found to be unstable, likely owing to the monochromatic beam. The reference data set was generated by an initial merge of the images and was used to post-refine individual images to improve the merging statistics R_{split} and $CC_{1/2}$. The bandwidth for refinement was set to 0.01% of the pulse wavelength and the mosaicity to 0.04°, leading to around 166 accepted reflections per image. Images were rejected if they were below a correlation threshold of 0.8 and individual reflections were rejected beyond 1.8 standard deviations from the mean of all observation intensities for a given reflection (parameters as defined in Ginn, Brewster *et al.*, 2015). Linear scale factors were applied to each image during merging to best match the reference density. Intensities were converted to amplitudes using *TRUNCATE* (Winn *et al.*, 2011). The protein and haem group from the sperm whale myoglobin structure (PDB entry 1vxa; Yang & Phillips, 1996) were used as a starting model for molecular replacement using *Phaser* (McCoy *et al.*, 2007). Building and refinement were performed using *Coot* (Emsley *et al.*, 2010) and *REFMAC*5 (Murshudov *et al.*, 2011). Following the placement of waters and a single sulfate atom, carbon monoxide was built into the density and refined (Fig. 4*b*). Refinement statistics are presented in Table 1. Coordinates and structure factors for the *P. macrocephalus* myoglobin with carbon monoxide covalently bound have been deposited in the Protein Data Bank with accession code 5jom.

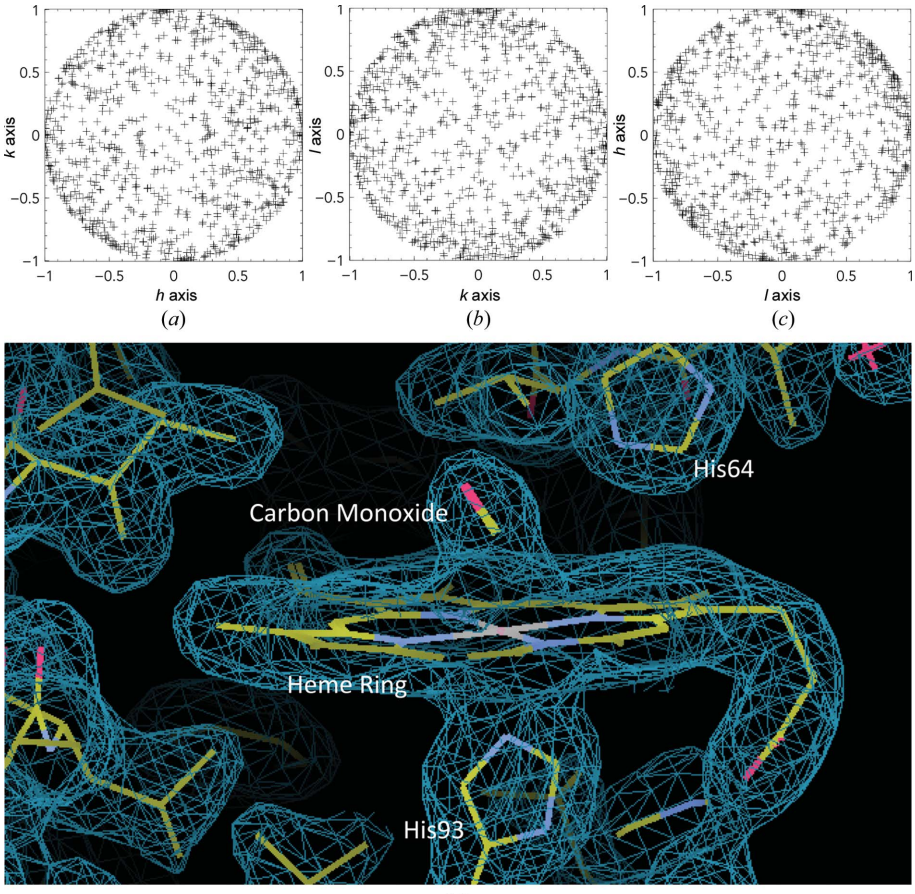


Figure 4
(*a*) A (1, 0, 0) vector was multiplied by each crystal orientation matrix and its associated symmetry operations. The data for each pair of axes are plotted against each other, showing a uniform sampling of crystal orientation. (*b*) Representative electron density at 1.9 Å resolution, contoured at 1σ, for carbon monoxide (CO) bound to the haem iron of sperm whale (*P. macrocephalus*) myoglobin generated from data collected using the crystallography chip.

3. Results

3.1. Chip design

The crystallography chip is a microfabricated 28 × 28 mm silicon substrate consisting of compartments with features designed specifically for trapping micrometre-sized crystals. The current design is comprised of 9 × 9 compartments, where each

compartment contains an array of 12 × 12 features (Fig. 5), resulting in 11 664 positions on the chip, 11 259 of which are for trapping crystals. Each feature is square-shaped with a 100 µm wide top aperture and a bottom aperture size ranging in width from 5 to 30 µm.

A key characteristic of the etched pattern on the silicon wafer is the trapezoidal geometry resulting from anisotropic etching of the different silicon crystal planes (Seidel *et al.*, 1990). The aperture and depth of the resulting features can be optimized to match the average crystal size of the system of interest. The opening at the bottom of each feature allows the removal of excess solution to recover any untrapped smaller crystals and to remove as much of the mother liquor remaining from the loading process as possible, dramatically reducing unwanted background scatter.

The thin regions of the chip designed for crystal trapping are easily deformable and, if not supported, can flex during loading and translation, resulting in a loss of positioning accuracy. The chip therefore includes thick silicon spacer bars between compartments to increase rigidity (Figs. 5*a* and 5*b*).

It should be noted that the thickness of the raw silicon wafer varies by 10 µm, which affects the final feature sizes and the etching rate at the extremities of the individual compartments and the chip itself. The etching is slightly slower at the edge of the chip than in the centre. This leads to a radial distribution of feature sizes of about 5 µm over each individual compartment and over the whole chip. This non-uniformity means that a small number of features on the outermost part of the chip might not be fully open at the bottom and thus cannot hold a crystal. Nonetheless, this geometry and method of fabrication result in a chip capable of trapping single crystals. The method of fabrication can be improved and easily scaled and thus allows the adjustment of the number of compartments and features to further increase the chip capacity if desired.

Each compartment has five fiducials, three of which are at the corners for alignment and two additional ones serving as unique addresses, allowing the precise and unambiguous positioning of every feature on the chip. The fiducial marks are designed to enable alignment of the chip using an inline imaging system with minimum aberrations to locate the centroid of the fiducial mark. The fiducial marks have a length of 100 µm and width of 10 µm and are in the form of a double cross or '#' symbol to facilitate automated image analysis (Figs. 5*c* and 5*d*). The '#' reduces edge confusion in the image-processing/alignment algorithm, hence reducing the number of false-positive matches. In principle, the corner features of every compartment could be used as intrinsic fiducial marks, but as the feature dimensions have some irregularity, owing to the silicon processing, this is less reliable.

3.2. Chip loading

Two auxiliary parts have been fabricated for use in the loading process to improve its ease and efficiency (Fig. 2). The first is an extremely flat silicon 'bottom plate' (<5 nm surface roughness over the whole contact area) with a hole through it, which lies underneath the crystallography chip. This very smooth surface provides a good seal for a closed-flow system to direct the crystals through the features to trap the crystals. The second is a silicon chip with nine linear channels etched across it, which span the length of the chip and with widths matching the size of one compartment. This part sits on top of the crystallography chip to direct the flow of liquid during sample loading, reducing sample waste by minimizing flow over dead space on the chip and maximizing the flow rate over the features. This avoids the adsorption of crystals on to the chip at locations where they would not be sampled in the diffraction experiment. The whole assembly is stacked and held together accurately using a custom holder.

To minimize stress on the crystals, the loading process is performed at a controlled temperature to match that of crystallization, and at controlled humidity to prevent salt-crystal formation from the mother liquor (see §2 for details). First the crystals, in their mother liquor, are dispensed onto the chip using a pipette, using siliconated tips to minimize sticking of crystals to the pipette tip. Depending on the crystal density, volumes of 60–80 µl crystal slurry are sufficient to load a chip with approximately 40% occupancy. Negative pressure is applied to the underside of the chip and the solution is sucked through the features, leaving the crystals trapped

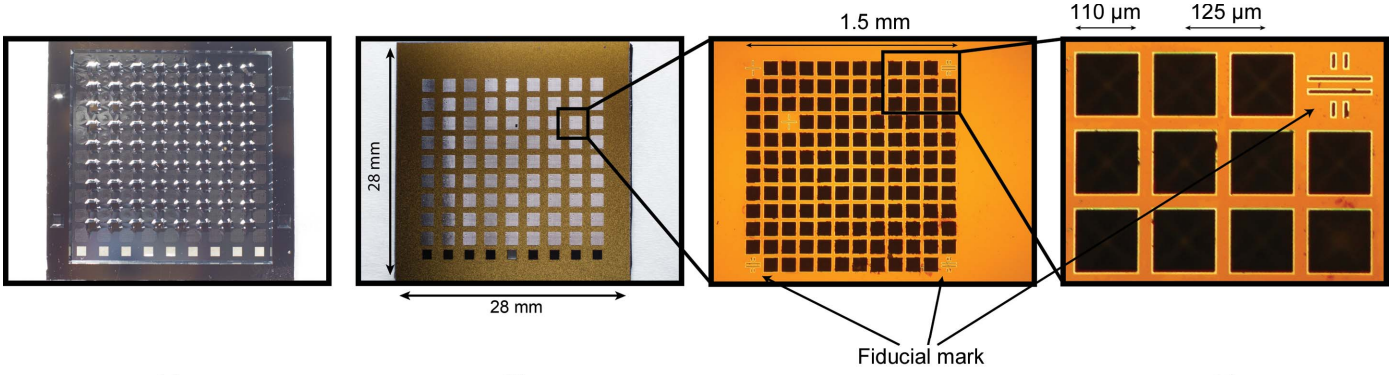


Figure 5
Back (*a*) and front (*b*) of the chip and a close-up view of an individual compartment (*c*). Each chip comprises 9 × 9 compartments with 12 × 12 features per compartment. As each feature is designed to hold a single crystal, the maximum capacity of the chip is 11 259 (139 × 81) crystals (based on 144 features minus five fiducials per compartment). (*d*) Magnified portion of a compartment showing the crystal-trapping features and a fiducial mark.

inside a feature with a minimal volume of mother liquor. The etched away area on the back side of the chip creates a large evacuated volume for efficient removal of the mother liquor (Fig. 2*c*). The amount of negative pressure applied during the loading process needs to be carefully adjusted. Too much suction dehydrates the protein crystals too quickly, while too little suction leaves excess mother liquor on the surface of the chip, resulting in increased background scatter. To avoid the formation of salt crystals, adjusting the suction and humidity is important when high salt concentrations are used as precipitant. Next, the chip is removed from the closed-flow system and rapidly sealed with 4 µm thick Mylar film on each side. Finally, the chip is placed in a custom holder, which is specially designed to accommodate the two 4 µm films of stretched Mylar and the chip.

The permeability of the Mylar film is low enough to minimize protein dehydration and maintain the crystals in a humid environment long enough for data collection. The Mylar introduces some diffuse background to the diffraction pattern (see Fig. 6), which is subtracted during the standard data-processing procedures. The bottom hole of the feature acts as an aperture which helps to remove background scatter (see Fig. 6). In the future other, lower-scatter, polymer films or 30–50 nm SiN films could be substituted to significantly further reduce the background scatter. In addition, the silicon nitride and etched silicon surfaces, which are usually covered with their native oxides, are very hydrophilic, and therefore it may be possible to improve the loading efficiencies and removal of excess mother liquor by reducing the hydrophilicity of the chip using simple surface treatments.

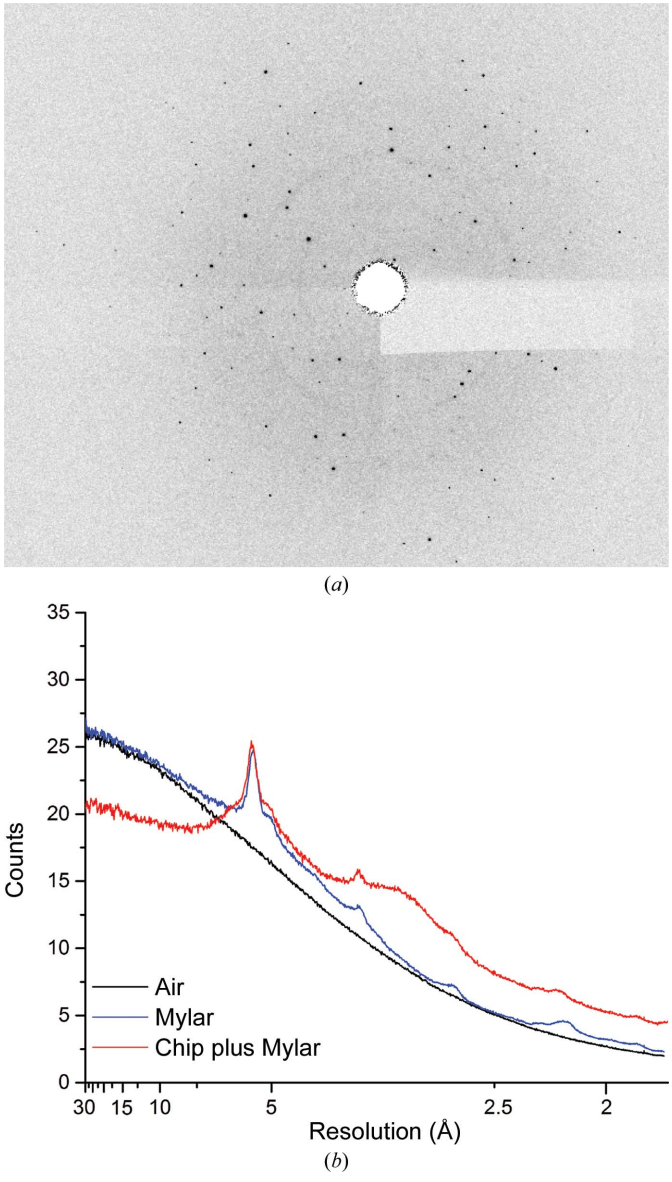


Figure 6
(*a*) Example of diffraction image with Mylar background (image obtained at the XPP endstation at LCLS), (*b*) comparison of diffuse scatter of air, air plus Mylar and a loaded chip (empty feature) as a function of resolution. The feature acts as an aperture, reducing the background at low resolution (data obtained on beamline I24 at Diamond Light Source).

3.3. Optical spectroscopy setup

To increase the hit rate, we have developed a high-speed UV–Vis microspectrometer (adapted from Kirchner *et al.*, 2014) with a focus spot size of 100 µm and that is sensitive to a wavelength range of 215–850 nm to map crystal positions in the chip. Transmission spectra are recorded from each feature to determine whether or not it harbours a crystal. This information can subsequently be passed to the beamline data-acquisition software so that empty features are not addressed during the SFX experiment. This approach also allows differentiation of protein and salt crystals, as salt crystals strongly scatter the light and can therefore be easily identified. The rapid filling of the chip relative to the rate of salt-crystal formation greatly reduces the growth of salt crystals during chip loading (§3.2). However, for salt-based crystallization conditions, some features can still be occupied with salt crystals and, in addition, salt crystals can grow on protein crystal surfaces.

To ensure complete compatibility between the chip mapping and positioning of the chip at the X-ray beamline, we use identical precision translation stages for the spectroscopy/mapping and for the X-ray beamline and the same coordinate-system transform to convert feature positions into translation-stage motions as described previously (Sherrell *et al.*, 2015). Once mounted, the chip is translated along the spectroscopy probe beam axis until it is at the beam focus and the optical spot size is minimized. Compensation is made for chip tilt in the coordinate-system transform based on the position of the three outermost corner fiducial marks on the chip. The chip is then translated successively to each feature, where a spectrum is recorded and named according to the feature address as defined by the coordinate system. The experimental setup is shown schematically in Fig. 3(*a*), together with representative spectra of carboxymyoglobin crystals recorded from the chip (Fig. 3*b*). The time needed to acquire a spectrum with acceptable signal to noise limits the speed at which a chip can be scanned. In the experiments reported here the integration time for recording each feature spectrum was 250 ms, resulting

Table 2
Hit-rate statistics for mapped chips.

Chip No.	1	2	3	4	5	6	Average (%)
Positive spectra	6983	2237	9174	4325	8618	3782	52† (fill factor)
Maximum No. of diffraction images possible‡	4467	1485	5993	3100	6064	2609	57
Diffraction images with >10 peaks	4093	1432	5298	2790	4161	2148	84 (h_{raw})
Indexed images	1802	825	2479	1611	2863	1171	46 (h_{idx})

† The number of features with a verified spectrum of carboxymyoglobin. ‡ The maximum expected number of diffraction images taking empty (negligible amplitude) X-ray pulses into account.

in a 40 min acquisition time for a whole chip map. However, improvements in integration time and the use of an external trigger have since been implemented, leading to an acquisition time of approximately 8 min per chip.

After the chip has been fully scanned, the background-corrected transmission spectra are processed to determine the loaded feature positions. This is performed by taking the ratio of the integral of a part of the transmission spectrum which is expected to show significant absorption (500–700 nm for carboxymyoglobin) in comparison to the integral of a part of the spectrum where there is little absorption (700–850 nm). A feature is marked as containing a crystal if the resulting number is below the threshold ratio of an empty feature. A list is then generated containing each feature address and a binary value corresponding to whether or not the feature contains a crystal. This scheme is very fast and is generally effective, but suffers from a few drawbacks. Firstly, the effectiveness of the asymmetric filter described above depends on the specific absorption spectrum of the sample. Secondly, it is impossible, based on the spectroscopic data alone, to distinguish crystalline sample from amorphous sample. Subsequent diffraction data showed that 15% of features marked as containing a crystal did not yield Bragg peaks, indicating the presence of amorphous material in our crystal preparation. Nonetheless, we were able to consistently achieve 85% raw hit rates using the XPP instrument at LCLS (Chollet *et al.*, 2015; §3.4). It is possible that a more robust way of correlating the spectral signal with that of crystalline protein can be found to increase the hit rate; however, there is likely to be some residual crystal-to-crystal variation in diffraction quality that will preclude reaching hit rates of exactly 100%. Nevertheless, the present hit rate already allows the highly efficient use of the current XFEL pulse stream.

3.4. Data acquisition and analysis

For a proof of principle, we investigated the use of the crystallography chip with spectroscopic mapping during an experiment using the XPP instrument at the Linac Coherent Light Source (LCLS), determining the structure of carboxymyoglobin using the SFX approach. Spectroscopic mapping was carried out on six chips, which then underwent X-ray data collection. The use of a monochromator resulted in $\sim 10^{10}$ photons per image, which was considerably fewer than the 10^{12} photons per image typical in an SFX experiment. For this reason, in a typical 'full-throttle' experiment, we would expect a useful data set to be collected from substantially fewer images, with the expected attendant improvements in signal to noise with more diffracted X-ray photons. The average number of X-ray hits recorded for a scanned chip was around 5000 (Table 2). Here, an X-ray hit is defined as an image containing >10 intensity maxima, as determined by a fast peak search using *OnDA* (Mariani *et al.*, 2016), which we report here as raw hit rates (h_{raw}). This indicates that the crystal loading efficiency on the chip was approximately 50%. The use of a shot list based on the spectroscopic chip mapping resulted in an actual overall hit rate of greater than 85%.

A quantity that describes the success rate of diffraction events more reliably is the fraction of shots that lead to indexable diffraction patterns, h_{idx}, which is available after data reduction and indexing of the entire data set. The 'indexable hit rate' h_{idx} can never be higher than h_{raw}, and will usually be considerably lower, as not every crystal diffraction pattern can be indexed. Obviously h_{idx} depends not only on the efficiency of sample delivery, but also on the type and quality of the crystals and on the performance of the indexing software. Utilizing *CrystFEL* to index the diffraction data gives an h_{idx} of 43%, which represents on average 2000 successfully indexed images per chip. Finally, because of the stochastic nature of X-ray pulse generation, there is a shot-to-shot variation in both pulse intensity and spectral distribution. In our experiments an X-ray monochromator was used. As a consequence, the natural variation in the X-ray pulse intensity leads to a significant fraction of 'empty' pulses. As a result, even features containing a 'good' crystal may not result in a 'hit'. To take this into account, corrected hit rates, h'_{raw} and h'_{idx}, were calculated on the basis of the number of X-ray pulses observed with intensity greater than a certain threshold. This correction leads to an increase in the h_{idx} rates to 46%, *i.e.* close to 50% indexable diffraction patterns. The total number of diffraction images (>10 peaks, h_{raw}) collected from the six chips was of the order of 20 000, of which half resulted in successfully indexed diffraction patterns (Table 2, the exact number is 10 751). Not all of these images were used in final structure determination (see Table 1 and §2).

The dimensions of the crystals that were successfully indexed were 20 × 20 × 40–70 µm. Analysis of the orientation matrices (Fig. 4*a*) confirms that the crystals are reasonably randomly oriented, with no evidence of preferred alignment within the chip features. The quality of the intensities derived by the application of *cppxfel* (Ginn, Brewster *et al.*, 2015) was reasonable (Table 3), despite the heavy attenuation of the XFEL pulses (see §2). The *TakeTwo* algorithm within *cppxfel* increased the indexing rates to approximately 100% of images which registered more than 20 spots using *DIALS* spot finding (Waterman *et al.*, 2016). These data, when merged (with *cppxfel*), yielded a complete data set and supported model building and refinement to obtain a high-resolution model capable of supporting the analysis of small conformational changes. Table 1 summarizes the refinement statistics and

Table 3
Data collection and processing.

Values in parentheses are for the highest resolution shell.

Diffraction source	LCLS XFEL source, XPP endstation
Wavelength (Å)	1.29
Temperature (K)	293
Detector	Rayonix MX170-HS CCD
Crystal-to-detector distance (mm)	340
Exposure time per image (s)	0.1
Space group	$P2_12_12_1$
a, b, c (Å)	38.13, 46.48, 84.52
α, β, γ (°)	90, 90, 90
Resolution range (Å)	33.43–1.90 (1.93–1.90)
Images merged	8957
No. of unique reflections	12354
Completeness (%)	100
$\langle I/\sigma(I)\rangle$	6.0
Multiplicity	51.6
$R_{\rm split}$ (%)	10.1 (46.1)
Matthews coefficient $V_{\rm M}$ (Å^3 Da^{-1})	2.18
Solvent content $V_{\rm s}$ (%)	43.5
Observations per image to edge of detector	166 ± 28
$CC_{1/2}$ (%)	99.6 (42.6)

Fig. 4(*b*) shows a representative portion of the electron-density map.

4. Discussion

The batch process for chip fabrication makes it possible to produce many chips at once. The tunability of the fabrication process enables the development of different-sized features to accommodate a range of crystal sizes up to hundreds of micrometres. This makes the crystallography chip well suited for a large range of crystalline protein samples down to dimensions of a few micrometres. This aspect of the chip is especially important for SFX experiments and the study of proteins for which large crystals are difficult to obtain. In addition, owing to the well shaped features, the loading leads to centred crystals, ensuring that the X-ray beam hits the centre of the crystal rather than an edge.

The minimal amount of mother liquor around each crystal and the thin Mylar 'windows' result in very low background, making the chip an attractive alternative compared with other sample-delivery methods such as the liquid jet and LCP injectors, where the mother liquor or LCP carrier introduces significant background scatter to the diffraction data. An additional advantage of the silicon-based chip is that it can be cleaned and reused. Finally, we have not observed any evidence of radiation damage to the chip itself at either microfocus synchrotron beamlines or when exposed to an XFEL beam.

One of the greatest advantages of the fixed-target chip, in combination with spectroscopic mapping, is the ability to define a spatial address for each protein crystal located in the chip and thus extract the corresponding data selectively for a particular sample. This provides the possibility of performing additional diagnostics with the potential to improve SFX analysis, for example post-sorting of diffraction data based on optical absorption. In general, any measurable characteristic of an individual crystal can be recorded and referenced later for data analysis.

Another application of the optical spectroscopy setup is *in situ* sample characterization. Unstable samples that require sealing from the atmosphere or specific ambient conditions can be tested for longevity, for example in the case of carboxymyoglobin the ligand-binding state can be monitored for each individual crystal. In addition, chips can be scanned before and after X-ray exposure to observe the extent of X-ray-induced and laser-induced damage (in case of time-resolved SFX experiments). This is important as femtosecond time-resolved experiments generally involve very small structural changes and thus there is a need to go to the maximum photoexcitation possible to ensure the majority of molecules in the crystal are reacting. This requires laser powers that are typically right at the limits of multiphoton processes unrelated to the photophysics and photochemistry of interest. This capability allows spectroscopic analysis to ensure proper excitation conditions below this limit and the onset of X-ray damage.

The improved jet-based sample-delivery system utilizing LCP media has reported hit rates around 8% with an indexing success rate within this subset of images of 22%, resulting in an overall successful $h_{\rm idx}$ rate of about 1.6% (Weierstall *et al.*, 2014). Even allowing for possible differences in crystal quality, the hit rates achieved with this chip constitute an improvement of at least an order of magnitude over competing methods. The improved percentage of successfully indexed shots is owing to a combination of reduced background scatter with the chip (see unmapped hit rates; Fig. 7) with a further gain in efficiency owing to spectroscopic mapping (Table 2 and Fig. 7). Differential polarization analysis, using a simple polarizer between the sample and detector in the microspectrometer, or second-harmonic generation-based imaging to distinguish amorphous and crystalline protein will further improve this correlation to close to the 100% $h_{\rm raw}$ limit.

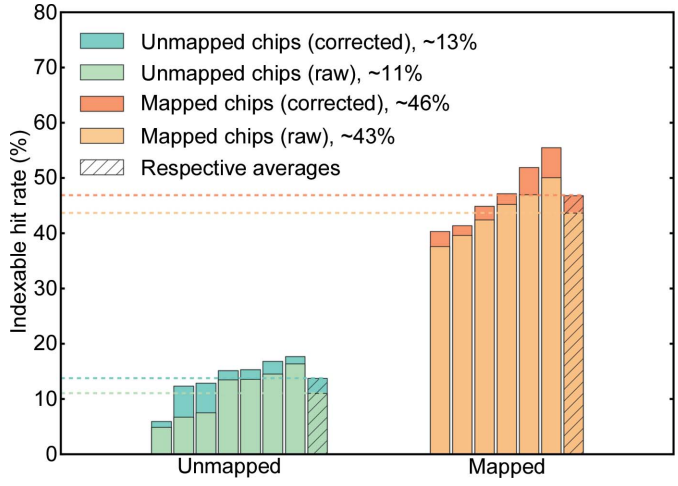


Figure 7
Percentage of indexed diffraction patterns for unmapped (left) and mapped (right) chips: light colours, without correction for missing X-ray pulses; dark colours, after correction. Mapping the chip improves the success rate approximately threefold.

With improvements in post-refinement (using a modified approach based on Ginn, Brewster *et al.*, 2015), we find 8957 indexable diffraction patterns to be sufficient to obtain a structure from monochromatic XFEL data. This means that in principle, with improved spectral mapping (including the differentiation between crystalline and amorphous material) *and* if optimal loading is achieved (>50%), one or two chips could provide sufficient diffraction data for structure determination (ultimately depending on the diffraction quality and the parameters of the crystal such as the space group). Data collection at XFELs is currently limited in principle by the repetition rate of the FEL (120 Hz for LCLS in the US, 30 Hz for SACLA in Japan), but practically by the repetition rate of the available detectors. At the XPP instrument, the Rayonix detector used in these experiments runs at 10 Hz. Taking into account the mounting and alignment time, this would mean sufficient data for *de novo* structure determination could be collected within less than 20 min. For the chips described (feature spacing of 125 µm), the present translation-stage system is capable of scanning at 107 Hz using the current approach of pausing at each feature for laser/X-ray exposure.

Larger data sets will, of course, improve the overall signal-to-noise ratio, and 20 000 useful diffraction patterns could, in principle, be acquired at 100 Hz from an optimally loaded and mapped chip containing a larger number of features in less than 4 min. For time-resolved studies, this would mean that a full time series with up to 100 time points could be collected within a few hours. This would enable multiple systems to be studied for comparative studies of structural dynamics or studies over many decades in time in a single experimental visit to capture more fully relevant dynamics.

5. Conclusion

We present an improvement of a chip-based approach to sample delivery for SFX (an extension of Mueller *et al.*, 2015; Zarrine-Afsar *et al.*, 2012), which combines spectroscopic mapping of crystal positions with an innovative low-background fixed target to provide a versatile and highly efficient system with significant advantages over other sample-delivery methods. In its current version, the chip is designed for time-resolved SFX experiments, but it can be used at virtually any synchrotron or XFEL beamline in combination with the dedicated mini endstation (Sherrell *et al.*, 2015). The efficient process of loading and handling samples, in combination with spectral mapping, significantly increases the productive data-acquisition time. Further improvements in pre-mapping of the chip, in combination with fast read-out X-ray detectors, will greatly facilitate high-throughput serial crystallography at XFELs, potentially requiring only minutes of data collection for static structure determination. This chip concept is therefore an ideal approach for precious crystalline protein samples and oversubscribed XFEL beamlines. The most significant future improvement will be the identification of spectral features or other probes of single-crystal character that are better correlated with X-ray diffraction quality so that the mapping procedure can lead to the maximum possible indexable hit rates for a given system.

6. Related literature

The following reference is cited in the Supporting Information for this article: von Stetten *et al.* (2015).

Acknowledgements

This research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) (RJDM), the Max Planck Society (RJDM), the Canadian Institute for Advanced Research (RJDM, OPE and DIS) and the Canada Excellence Research Chair program (OPE). OPE is the Anne and Max Tanenbaum chair in Neuroscience at the University of Toronto. The research leading to these results has received funding from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme (FP7/2007–2013) under REA grant agreement No. 623994 (HM-W). We would like to thank the staff at Toronto Nanofabrication Centre (TNFC) for their support. Use of the Linac Coherent Light Source (LCLS), SLAC National Accelerator Laboratory is supported by the US Department of Energy, Office of Science and Office of Basic Energy Sciences under Contract No. DE-AC02-76SF00515. This work made use of user facilities at SACLA, RIKEN, Japan with the approval of the Japan Synchrotron Radiation Research Institute (JASRI) (Proposal No. 2016A8036).

References

Barends, T. R. M. *et al.* (2015). *Science*, **350**, 445–450.
Barty, A. *et al.* (2011). *Nature Photonics*, **6**, 35–40.
Barty, A., Kirian, R. A., Maia, F. R. N. C., Hantke, M., Yoon, C. H., White, T. A. & Chapman, H. (2014). *J. Appl. Cryst.* **47**, 1118–1131.
Baxter, E. L. *et al.* (2016). *Acta Cryst.* D**72**, 2–11.
Boutet, S. *et al.* (2012). *Science*, **337**, 362–364.
Chapman, H. N. *et al.* (2011). *Nature (London)*, **470**, 73–77.
Chollet, M. *et al.* (2015). *J. Synchrotron Rad.* **22**, 503–507.
Chreifi, G., Baxter, E. L., Doukov, T., Cohen, A. E., McPhillips, S. E., Song, J., Meharenna, Y. T., Soltis, S. M. & Poulos, T. L. (2016). *Proc. Natl Acad. Sci. USA*, **113**, 1226–1231.
Cohen, A. E. *et al.* (2014). *Proc. Natl Acad. Sci. USA*, **111**, 17122–17127.
Coppens, P., Pitak, M., Gembicky, M., Messerschmidt, M., Scheins, S., Benedict, J., Adachi, S., Sato, T., Nozawa, S., Ichiyanagi, K., Chollet, M. & Koshihara, S. (2009). *J. Synchrotron Rad.* **16**, 226–230.
DePonte, D. P., Weierstall, U., Schmidt, K., Warner, J., Starodub, D., Spence, J. C. H. & Doak, R. B. (2008). *J. Phys. D Appl. Phys.* **41**, 195505.
Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. (2010). *Acta Cryst.* D**66**, 486–501.
Ginn, H. M., Brewster, A. S., Hattne, J., Evans, G., Wagner, A., Grimes, J. M., Sauter, N. K., Sutton, G. & Stuart, D. I. (2015). *Acta Cryst.* D**71**, 1400–1410.
Ginn, H. M., Evans, G., Sauter, N. K. & Stuart, D. I. (2016). *J. Appl. Cryst.* **49**, 1065–1072.
Ginn, H. M., Messerschmidt, M., Ji, X., Zhang, H., Axford, D., Gildea, R. J., Winter, G., Brewster, A. S., Hattne, J., Wagner, A., Grimes, J. M., Evans, G., Sauter, N. K., Sutton, G. & Stuart, D. I. (2015). *Nature Commun.* **6**, 6435.

Ginn, H. M., Roedig, P., Kuo, A., Evans, G., Sauter, N. K., Ernst, O., Meents, A., Müller-Werkmeister, H., Miller, R. J. D. & Stuart, D. I. (2016). *Acta Cryst.* D**72**, 956–965.
Guha, S., Perry, S. L., Pawate, A. S. & Kenis, P. J. A. (2012). *Sens. Actuators B Chem.* **174**, 1–9.
Hirata, K. *et al.* (2014). *Nature Methods*, **11**, 734–736.
Hunter, M. S. *et al.* (2014). *Sci. Rep.* **4**, 6026.
Kabsch, W. (2010). *Acta Cryst.* D**66**, 125–132.
Keedy, D. A. *et al.* (2015). *Elife*, **4**, e07574.
Kern, J. *et al.* (2014). *Nature Commun.* **5**, 4371.
Kirchner, F. O., Lahme, S., Riedle, E. & Baum, P. (2014). *AIP Adv.* **4**, 077134.
Li, T., Quillin, M. L., Phillips, G. N. Jr & Olson, J. S. (1994). *Biochemistry*, **33**, 1433–1446.
Lyubimov, A. Y., Murray, T. D., Koehl, A., Araci, I. E., Uervirojnangkoorn, M., Zeldin, O. B., Cohen, A. E., Soltis, S. M., Baxter, E. L., Brewster, A. S., Sauter, N. K., Brunger, A. T. & Berger, J. M. (2015). *Acta Cryst.* D**71**, 928–940.
Mariani, V., Morgan, A., Yoon, C. H., Lane, T. J., White, T. A., O'Grady, C., Kuhn, M., Aplin, S., Koglin, J., Barty, A. & Chapman, H. N. (2016). *J. Appl. Cryst.* **49**, 1073–1080.
McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C. & Read, R. J. (2007). *J. Appl. Cryst.* **40**, 658–674.
Mueller, C. *et al.* (2015). *Struct. Dyn.* **2**, 054302.
Murray, T. D., Lyubimov, A. Y., Ogata, C. M., Vo, H., Uervirojnangkoorn, M., Brunger, A. T. & Berger, J. M. (2015). *Acta Cryst.* D**71**, 1987–1997.
Murshudov, G. N., Skubák, P., Lebedev, A. A., Pannu, N. S., Steiner, R. A., Nicholls, R. A., Winn, M. D., Long, F. & Vagin, A. A. (2011). *Acta Cryst.* D**67**, 355–367.
Owen, R. L., Yorke, B. A., Gowdy, J. A. & Pearson, A. R. (2011). *J. Synchrotron Rad.* **18**, 367–373.
Pearson, A. R., Pahl, R., Kovaleva, E. G., Davidson, V. L. & Wilmot, C. M. (2007). *J. Synchrotron Rad.* **14**, 92–98.
Roedig, P., Duman, R., Sanchez-Weatherby, J., Vartiainen, I., Burkhardt, A., Warmer, M., David, C., Wagner, A. & Meents, A. (2016). *J. Appl. Cryst.* **49**, 968–975.
Roedig, P., Vartiainen, I., Duman, R., Panneerselvam, S., Stübe, N., Lorbeer, O., Warmer, M., Sutton, G., Stuart, D. I., Weckert, E., David, C., Wagner, A. & Meents, A. (2015). *Sci. Rep.* **5**, 10451.
Roessler, C. G. *et al.* (2016). *Structure*, **24**, 631–640.
Seidel, H. (1990). *J. Electrochem. Soc.* **137**, 3612–3626.
Sherrell, D. A., Foster, A. J., Hudson, L., Nutter, B., O'Hea, J., Nelson, S., Paré-Labrosse, O., Oghbaey, S., Miller, R. J. D. & Owen, R. L. (2015). *J. Synchrotron Rad.* **22**, 1372–1378.
Sierra, R. G. *et al.* (2012). *Acta Cryst.* D**68**, 1584–1587.
Springer, B. A. & Sligar, S. G. (1987). *Proc. Natl Acad. Sci. USA*, **84**, 8961–8965.
Stetten, D. von, Giraud, T., Carpentier, P., Sever, F., Terrien, M., Dobias, F., Juers, D. H., Flot, D., Mueller-Dieckmann, C., Leonard, G. A., de Sanctis, D. & Royant, A. (2015). *Acta Cryst.* D**71**, 15–26.
Sugahara, M. *et al.* (2015). *Nature Methods*, **12**, 61–63.
Tenboer, J. *et al.* (2014). *Science*, **346**, 1242–1246.
Waterman, D. G., Winter, G., Gildea, R. J., Parkhurst, J. M., Brewster, A. S., Sauter, N. K. & Evans, G. (2016). *Acta Cryst.* D**72**, 558–575.
Weierstall, U. *et al.* (2014). *Nature Commun.* **5**, 3309.
White, T. A., Kirian, R. A., Martin, A. V., Aquila, A., Nass, K., Barty, A. & Chapman, H. N. (2012). *J. Appl. Cryst.* **45**, 335–341.
Winn, M. D. *et al.* (2011). *Acta Cryst.* D**67**, 235–242.
Yang, F. & Phillips, G. N. Jr (1996). *J. Mol. Biol.* **256**, 762–774.
Zarrine-Afsar, A., Barends, T. R. M., Müller, C., Fuchs, M. R., Lomb, L., Schlichting, I. & Miller, R. J. D. (2012). *Acta Cryst.* D**68**, 321–323.
Zeldin, O. B., Gerstel, M. & Garman, E. F. (2013). *J. Appl. Cryst.* **46**, 1225–1230.
Zhou, Q. *et al.* (2015). *Nature (London)*, **525**, 62–67.

STRUCTURAL BIOLOGY

ISSN 2059-7983

TakeTwo: an indexing algorithm suited to still images with known crystal parameters

Helen Mary Ginn,[a,b] Philip Roedig,[c] Anling Kuo,[d] Gwyndaf Evans,[b] Nicholas K. Sauter,[e] Oliver Ernst,[d,f] Alke Meents,[c] Henrike Mueller-Werkmeister,[d,i,g] R. J. Dwayne Miller[g,h,i] and David Ian Stuart[a,b]*

[a]Division of Structural Biology, Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford OX3 7BN, England, [b]Diamond House, Harwell Science and Innovation Campus, Fermi Avenue, Didcot OX11 0QX, England, [c]Deutsches Elektronen-Synchrotron, Notkestrasse 85, 22607 Hamburg, Germany, [d]Department of Biochemistry, University of Toronto, King's College Circle, Toronto, ON M5S 1A8, Canada, [e]Molecular Biophysics and Integrated Bioimaging Division, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, USA, [f]Department of Molecular Genetics, University of Toronto, King's College Circle, Toronto, ON M5S 1A8, Canada, [g]Atomically Resolved Dynamics, Max-Planck-Institute for Structure and Dynamics of Matter, Luruper Chaussee 149, Hamburg, Germany, [h]Hamburg Centre for Ultrafast Imaging, University of Hamburg, Hamburg, Germany, and [i]Departments of Physics and Chemistry, University of Toronto, 80 St George Street, Toronto, ON M5S 1H6, Canada. *Correspondence e-mail: dave@strubi.ox.ac.uk

Received 19 May 2016
Accepted 1 July 2016

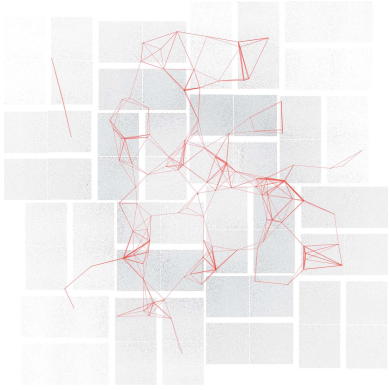
Edited by R. J. Read, University of Cambridge, England

Keywords: *TakeTwo*; data processing; serial crystallography; XFELs; X-ray free-electron lasers.

The indexing methods currently used for serial femtosecond crystallography were originally developed for experiments in which crystals are rotated in the X-ray beam, providing significant three-dimensional information. On the other hand, shots from both X-ray free-electron lasers and serial synchrotron crystallography experiments are still images, in which the few three-dimensional data available arise only from the curvature of the Ewald sphere. Traditional synchrotron crystallography methods are thus less well suited to still image data processing. Here, a new indexing method is presented with the aim of maximizing information use from a still image given the known unit-cell dimensions and space group. Efficacy for cubic, hexagonal and orthorhombic space groups is shown, and for those showing some evidence of diffraction the indexing rate ranged from 90% (hexagonal space group) to 151% (cubic space group). Here, the indexing rate refers to the number of lattices indexed per image.

1. Introduction

Indexing, or deducing the specimen orientation from crystalline diffraction patterns, can potentially be performed with high accuracy and precision owing to the integral nature of the Miller indices at which Bragg reflections are located. Indexing algorithms implemented in programs such as *XDS* (Kabsch, 1993), *iMosflm* (Powell *et al.*, 2013), *DENZO* (Otwinowski & Minor, 2006), *LABELIT* (Sauter *et al.*, 2004) and *DIALS* (Gildea *et al.*, 2014) are well established for data-collection strategies that involve crystal rotation, which are typically employed at synchrotron sources. For data collected at X-ray free-electron laser (XFEL) sources, where each image represents diffraction from a separate nonrotating specimen, the measure of success is less well defined. Data-analysis pipelines sometimes include a preprocessing 'hitfinder' step, which distinguishes images without diffraction (blanks) from hits that exceed a certain threshold number of candidate Bragg spots, which is typically set to around 20. The indexing rate is therefore defined as the percentage of hits for which crystal orientations can be determined. Reported indexing rates are



OPEN ACCESS

Table 1
Spot-finding parameters used for *DIALS* spot finding prior to indexing.

Full parameter definitions are available at the *DIALS* website (http://dials.lbl.gov/). A summary follows. Gain: the ratio of the detector output to the input number of photons per pixel. Global threshold: all pixels less than this value are considered to be background. Minimum spot size: minimum number of pixels required to perform the thresholding operation. Sigma background: number of standard deviations of the coefficient of variation in the local area below which a pixel is classified as background. Sigma strong: number of standard deviations above the mean in the local area above which a pixel will be classified as strong. Low/high-resolution cutoff: lowest/highest resolution considered in spot finding.

Parameter	CPV	BEV	Thermolysin	Myoglobin
Gain	14	14	14	0.1
Global threshold	0	2000	800	0
Minimum spot size	2	2	2	1
Sigma background	6	30	6	6
Sigma strong	3	6	3	3
Low-resolution cutoff (Å)	N/A	27	N/A	15
High-resolution cutoff (Å)	N/A	N/A	N/A	2

often quite low (Barends *et al.*, 2013; Liu *et al.*, 2013; Johansson *et al.*, 2013; Ginn, Brewster *et al.*, 2015; Chapman *et al.*, 2011), although the rate depends on a number of factors, including the strength of diffraction on each image and the diffraction resolution. Our own two analyses of diffraction data from Cypovirus type 17 polyhedrin (CPV17) yielded mediocre indexing rates of 53% (Ginn, Messerschidt *et al.*, 2015) and 36% (Ginn, Brewster *et al.*, 2015). Here, we investigate whether alternate algorithms can help improve the indexing rate, with the overall goal of producing good-quality structures while consuming a minimal amount of crystalline sample and beam time.

All indexing methods, regardless of the data source, begin by transforming candidate Bragg spot coordinates measured on the detector into corresponding three-dimensional coordinates in reciprocal space. The periodic arrangement of reciprocal-lattice points (rlps) is then detected using one of several methods. For rotation data generally, equivalent determinations of the three-dimensional periodic repeat can be deduced either by a single three-dimensional fast Fourier transformation of the entire reciprocal-space pattern (Campbell, 1998) or by separately computing one-dimensional FFTs of the pattern projected onto individual directional axes in reciprocal space and considering all possible directional axes, finely sampled on a spherical grid (Steller *et al.*, 1997). A key success factor is that the specimen rotation provides sufficient sampling in three dimensions to dramatically overdetermine the three-dimensional lattice repeat. In contrast, still shots produced at XFEL sources contain only the limited three-dimensional information that is afforded by the curvature of the Ewald sphere, which becomes essentially non-existent in the low-resolution limit. Another condition making it difficult to detect periodicity from still shots is the minimal number of Bragg spots that meet the reflection conditions, especially from crystals with smaller unit-cell dimensions.

A number of approaches have been adopted to mitigate these inherent difficulties in resolving the lattice for still-shot data. The *cctbx.xfel* software suite (Hattne *et al.*, 2014) uses the one-dimensional FFT method to identify candidate basis vectors that can potentially span the lattice, but then uses prior knowledge of the unit-cell parameters to choose a basis set (three vectors combined) that best agrees with the known unit-cell lengths and angles. Even if the unit cell is initially unknown, a good target cell may be derived from an initial first pass of data reduction (Zeldin *et al.*, 2015). Secondly, a recent method implemented within the *DIALS* toolbox (Waterman *et al.*, 2013) avoids the one-dimensional FFT search altogether (Gildea *et al.*, 2014). The basic idea here is that the FFT is primarily useful to identify the periodic repeat spacing if the unit-cell length is completely unknown. However, since the typical unit-cell length can be treated as prior knowledge, it is sufficient to perform an exhaustive search for the already known periodic spacings over a grid of directional axes. This method has a high success rate for finding the basis vectors from XFEL still shots, and has also been used to index shots where multiple lattices are present, including CPV polyhedrin diffraction data from Diamond Light Source beamline I24, which eventually resulted in structure solution (Ginn, Messerschmidt *et al.*, 2015). A final adaptation concerns the rocking curve that is normally observed in rotation data sets, as the reciprocal-lattice point is first rotated into and then out of the exact reflecting condition. At synchrotrons it is easy to collect individual image frames with a fine enough rotational slicing to determine the spot centroid position with high accuracy, but for still data the degree to which the rlps are offset from the Ewald sphere is initially unknown. To compensate, parameter refinement targets have been developed (Sauter *et al.*, 2014; Kabsch, 2014) that restrain the lattice orientation (after the basis vectors are chosen) so that the rlps are positioned as close as possible to the Ewald sphere.

In our efforts to analyse a number of new challenging data sets, including that of *Bovine enterovirus* (BEV), we revisited the indexing stage of the data-processing pipeline. At first, the BEV data appeared to be intractable owing to the low signal to noise and small separation between neighbouring spots. Initial attempts at indexing, specifying the known unit cell and space group, produced indexing solutions with refinement statistics that appeared to be of extremely good quality. Finally, however, we realised that the solutions were fundamentally incorrect owing to a large (20 mm) error in the detector distance that initially went undetected.

In the process of developing automated approaches to fix the distance problem, we also sought new ways to detect the basis vectors, considering the sparsity of Bragg spots that potentially precludes the use of FFT methods. Older indexing methods developed around 1990 (Kabsch, 1988; Higashi, 1990; Kabsch, 1993) succeeded without any explicit grid search or FFT method by considering the difference vectors that connect rlps that are close in reciprocal space. Kabsch (1988) was able to index rotation data by creating a three-dimensional histogram of such difference vectors and identifying clusters which correspond to candidate basis vectors for the lattice.

Inspired by these early experiences with indexing from a limited set of difference vectors, we devised a similar

algorithm that allows us to tackle difficult diffraction patterns and correct for errors in the measurements of the experimental geometry. We have named this algorithm *TakeTwo*, reflecting the underlying idea of taking pairs of spots to form vectors and pairs of vectors to generate indexing solutions. *TakeTwo* attempts to make maximal use of the information contained within a single still image by considering all inter-spot vectors that could match to a vector between reflections in reciprocal space. This indexing algorithm provides high indexing rates for cubic and hexagonal space groups, and a high degree of success in indexing multiple lattices. *TakeTwo* is applicable to both serial synchrotron and serial femtosecond crystallography. We expect the *TakeTwo* algorithm to markedly improve the indexing rates of many of the available XFEL data sets.

2. Materials and methods

2.1. Data acquisition for various protein crystals

A summary of the experimental parameters and crystal symmetry for the various data sets is provided in Table 1. The diffraction patterns used for CPV17 had previously been used for structure determination (Ginn, Brewster *et al.*, 2015; Ginn, Messerschmidt *et al.*, 2015). Data were collected at the XPP endstation at the Linac Coherent Light Source (LCLS) under proposal LH90 from BEV crystals on a silicon chip (Roedig *et al.*, 2015; cubic space group $F23$, unit-cell dimension 437 Å). Reflections were observed to approximately 2.0 Å resolution and the minimum separation of individual reflections was three pixels. The data from thermolysin crystals, collected in 2011, have previously been processed and optimal unit-cell dimensions established (Uervirojnangkoorn *et al.*, 2015). All of the above data were recorded using the CSPAD detector. Diffraction patterns from myoglobin crystals crystallized in space group $P2_12_12_1$ provided an example of an orthorhombic space group and were recorded using a Rayonix MX170-HS detector at the XPP endstation using an alternative silicon-chip delivery system (Sherrell *et al.*, 2015; Zarrine-Afsar *et al.*, 2012; Mueller *et al.*, 2015), which has been shown to support structure solution (Oghbaey *et al.*, 2016).

2.2. Hit finding and spot finding using *DIALS*

Hits for the CPV17, BEV and thermolysin data sets were determined with *cctbx.xfel* using the default minimum spot count of 20 to distinguish a hit from a non-useful image. Images for the myoglobin data set were identified as hits if they had at least 30 spots from *DIALS* (Waterman *et al.*, 2013) spot-finding analysis, which uses *XDS* algorithms (Kabsch, 1977). For our indexing algorithm, spots were identified using *DIALS*, manually finding an optimal combination of spot-finding parameters (Table 2). *DIALS* produced spot-centroid coordinates on the surface of the detector.

Table 2
Parameters for each cycle of BEV indexing to achieve the final indexing rate of 97.6%.

The indexing method used was either the network or the cluster method. Runs were attempted aiming to index a maximum of one lattice per image. Runs were carried out with a constant reciprocal distance tolerance between vectors. Subsequent 'network' method runs were run with altered reciprocal distance tolerance and network threshold number values: the reciprocal distance tolerance was tightened with subsequent runs and the network threshold number was relaxed. The requirement for adding inter-spot vectors which included a spot common with the existing vectors was toggled.

Cycle	Indexing method	Reciprocal tolerance	Threshold number (vectors within network)	Common spots required	Indexed images (cumulative %)
1	Network	4×10^{-4}	22	False	84.0
2	Network	3×10^{-4}	15	False	93.0
3	Network	3×10^{-4}	12	True	97.0
4	Cluster	2.5×10^{-4}	N/A	N/A	97.6

2.3. Generating inter-spot vectors

Spots were back-projected onto the Ewald sphere using the known experimental parameters (incident energy and detector geometry). Three-dimensional vectors were calculated between spots in reciprocal space. The coordinates and vectors generated from back-projection onto the Ewald sphere are referred to as 'observed space', as they come directly from the spots and parameters of the experiment.

2.4. Optional filtering of inter-spot vectors

In cases where substantial amounts of noise are picked up by the spot-finding algorithms or multiple lattices are found on the same image, one can perform an optional filtering of inter-spot vectors to help prevent spot-to-noise, noise-to-noise or inter-lattice vectors from being included in pseudo-powder pattern generation or indexing. This was performed by tabularing all inter-spot vectors and removing those which had the fewest neighbours, as explained below.

An inter-spot vector was considered to be a neighbour of another vector if they were within a certain tolerance of each other. The tolerance was either calculated for a given resolution based on the energy bandwidth and the rlp size, or alternatively a constant tolerance was assumed across the entire image. The variable distance tolerance was chosen as the maximum possible separation between rlps of finite size on opposite sides of the nest of Ewald spheres. This assures that the vector is of a similar length and direction, and is more likely to be a repeated vector from the same lattice. Each vector was assigned a score which was equal to the total number of neighbouring vectors. The vectors were ordered in terms of their score in descending order, and a particular fraction of the vectors were removed if they did not reach a threshold p_t, which was based on the number of expected lattices for a given image, as follows. The number of expected lattices (n) on the image was calculated based on the number of spots on the image (s) divided by the number expected per lattice (l). The value of l was manually estimated after obtaining a few indexing results as the average number of

spots picked by *DIALS* which were predicted (and removed) by one lattice solution. n was rounded to the nearest integer:

$$n = \lfloor s/l + 0.5 \rfloor. \quad (1)$$

The total number of inter-spot vectors for a single lattice is proportional to the square of the total number of spots. However, if there were two lattices on an image, each spot could generate a vector either with a member of its own lattice or the other lattice, with roughly equal probability. Thus, one would only expect half of the spot vectors to originate between member vectors of the same lattice. The proportion of vectors (p_t) which were likely to originate from a single lattice can therefore be determined:

$$p_t = 1/n. \quad (2)$$

The threshold t was set so that only the proportion p_t of vectors would remain. Since the list is ordered so that those with the most neighbours are at the top, taking this proportion from the top of the list aims to enrich those that originate from the same lattice.

2.5. Theoretical distances for crystal lattices and space groups

Theoretical distances between individual reflections according to the space group and unit-cell dimension were calculated by considering all Miller indices and their relation to the origin of reciprocal space. Care was taken over systematic absences: those owing to the centring type of the lattice of the crystal prevent some inter-spot distances from ever appearing, as opposed to axial systematic absences which do not affect the occurrence of general inter-spot distances. The reciprocal unit cell generated by applying the transformation matrix onto Cartesian coordinates is hereby referred to as 'theoretical space'. The indexing solution then defines the orientation matrix which was applied to the transformation matrix to give rise to the spots observed on the detector.

2.6. One-dimensional pseudo-powder patterns

One-dimensional pseudo-powder patterns were generated from a small number of diffraction patterns (Fig. 1) by generating a histogram of inter-spot distances. These were compared with the predicted reflections given the unit-cell dimensions and space-group systematic absences, and the wavelength and detector distance were adjusted to ensure a good fit between the theoretical and observed powder diffraction patterns. The beam centre correctness, metrology, detector distance and wavelength parameters all affected the sharpness of the powder rings.

2.7. Inter-spot vector-distance matching

In order to use the observed space vectors for indexing, they were filtered according to length. Observed space vectors are considered for indexing if they match one or more of the theoretical space vectors in length, within a certain tolerance, calculated as in §2.4.

2.8. First indexing method: rotation-matrix clusters

Two related indexing algorithms based on inter-spot vectors have been created which are suited to different situations. The *TakeTwo* algorithm therefore has two branches: the rotation-matrix cluster method and the inter-spot vector-network method.

An indexing solution cannot be determined by one inter-spot vector alone as this only anchors reciprocal space in one axis. An indexing solution must be generated by a minimum of two vectors which are not linearly dependent. This mostly used four spots, but can use three if one spot is shared between the two inter-spot vectors, as explained in §2.10. Inter-spot vectors were considered a candidate for a solution if the angle between them in the observed space matched the angle between matching vectors in the theoretical space within 1°. For each of these pairs of inter-spot vectors, a matrix was generated which aligns the observed space onto the theoretical space (the opposite transformation was achieved using the inverse). This was combined with the transformation matrix to generate a potential indexing solution. Owing to the presence of noise and multiple lattices on a single image, the map of orientation matrices produced using this method had a low signal-to-noise ratio. Indexing solutions were chosen by selecting matrices which were closely surrounded by neighbouring solutions, determined by a reimplementation of the matrix similarity metric contained within *X-PLOR* (Brünger, 1990, 1992), which is reproduced here for clarity. For two rotation matrices P and Q, and n symmetry operators where O_s is the rotational component of symmetry operator s, the metric $m(P, Q)$ is defined as

$$m(P, Q) = \min_{s=1,n}\{\mathrm{Tr}[(P - O_sQ)(P - O_sQ)^t]\}^{1/2}. \quad (3)$$

If this metric has a value below the threshold of 0.25, this corresponds to a rotation of 10° or less and the two are therefore treated as duplicate solutions.

To visualize the indexing solutions generated using this method, a standard unit vector, for example (1, 0, 0), was rotated by an orientation matrix to generate a new vector. This was decomposed into polar coordinates and the two angles (θ, φ) were plotted. Of course, this did not contain all of the information contained within the orientation matrices, but provides a two-dimensional representation which was easily plotted to visually identify clusters of orientation matrices (Fig. 3). Clusters were identified by having the highest number of neighbours (calculated using the *X-PLOR* method) within an 8° radius between the central spot and any potential neighbour. Indexing solutions were filtered such that symmetrically identical solutions did not occur more than once, allowing the selection of multiple lattices. Indexing solutions were then selected for initial orientation-matrix refinement according to a previously outlined protocol (Ginn, Messerschmidt *et al.*, 2015).

Indexing solutions were accepted on the basis of the success of the initial orientation-matrix refinement. These initial refinements generate histograms of Ewald sphere wavelengths, which are defined as the inverse of the Ewald sphere

radius on which the centre of the modelled rlp lies. Images were accepted if their standard deviation of rlp midpoint Ewald sphere radii were below a certain threshold σ_t or if the histogram of wavelength peaks had a sufficiently high peak. These thresholds were manually chosen on a case-by-case basis for each crystal form and are highly dependent on the other parameters of the initial orientation-matrix refinement.

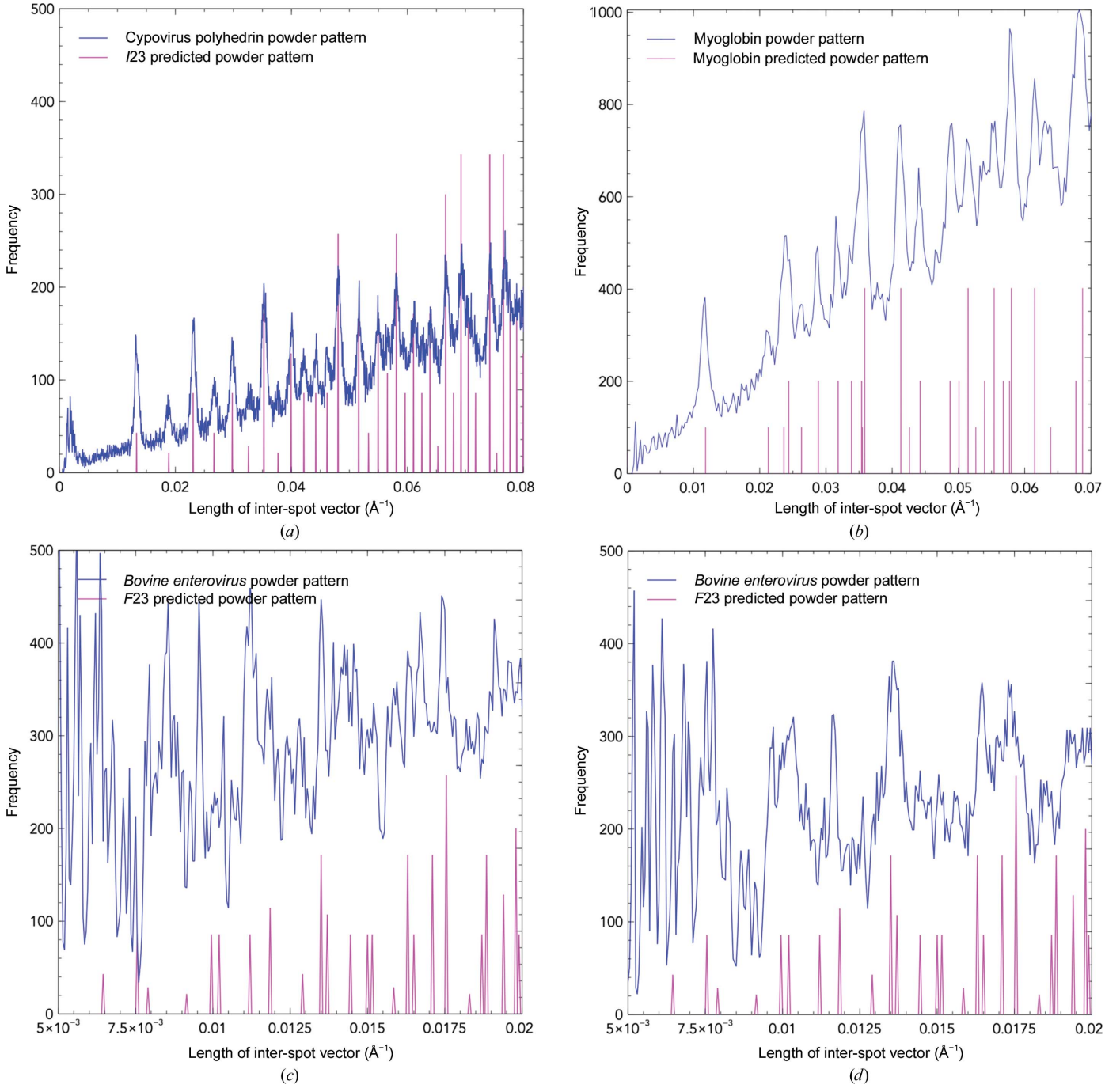


Figure 1
CPV17 at 101.2 mm detector distance (*a*), myoglobin (*b*) and BEV pseudo-powder patterns (*c*, *d*) generated using optimized experimental parameters without spot-vector filtering. The blue line is generated from experimental data, whereas the pink line is the expected pseudo-powder pattern given the unit-cell and space-group information. The height of the magenta line denotes the relative number of Miller indices which give rise to the particular inter-spot vector length; the absolute values are not shown for clarity. The CPV17 powder pattern was generated from 262 images. Owing to the higher mosaicity of the myoglobin crystals (464 images), the experimental peaks for each powder ring in (*b*) are broadened compared with the CPV17 powder pattern (*a*). For the BEV pseudo-powder patterns (304 images), the pattern in (*c*) was generated from the recorded detector distance of 105 mm. However, the pattern in (*d*), which shows a better fit to the pseudo-powder pattern, was obtained at a detector distance of 85 mm. The peaks are less well separated owing to the high error associated with measuring spot vectors between extremely close spots.

2.9. Second indexing method: building up an inter-spot vector network

The second indexing method built up an interconnected network of vectors which were all self-consistent for a single indexing solution. The network was considered as an indexing solution if the number of vectors reached a certain threshold, to distinguish it from 'noise'. This threshold was set to 20 by default. An indexing solution was built up recursively from a starting vector by adding further vectors which belong to the same indexing solution. In most cases, a new vector was not added unless it shared at least one spot with the existing network to increase the likelihood of picking spots from only a single lattice. However, this could be toggled in cases where only single lattices were suspected to be present. In order to increase the speed of calculation, vectors were 'pre-screened' to ensure that the angles between the pre-screened vector and the current network of existing vectors were consistent. The resulting orientation matrices between the pre-screened vector and existing vectors were then calculated according to §2.10 and the vector was accepted if it matched within a 8° tolerance of the existing indexing solutions as judged by the *X-PLOR* metric.

2.10. Generating an orientation matrix from two identified inter-spot vectors

A rotation matrix was generated from two identified inter-spot vectors i and j. Reciprocal space was rotated to line up the first observed vector with the corresponding theoretical vector. The cross product and angle between the observed vector $i_{\rm obs}$ and the theoretical vector $i_{\rm thr}$ were calculated:

$$C = i_{\rm obs} \times i_{\rm thr}, \tag{4}$$

$$\cos(\alpha) = \frac{i_{\rm obs} \cdot i_{\rm thr}}{|i_{\rm obs}||i_{\rm thr}|}. \tag{5}$$

We then generated a rotation matrix $\mathbf{Q}_a$, rotating by the angle α around the axis C, which aligns one axis in reciprocal space. The second axis must be aligned using the second vector $j_{\rm obs}$. The rotation matrix $\mathbf{Q}_b$ rotates $j_{\rm obs}$ by an angle β around the axis $i_{\rm thr}$ such that the angle between $\mathbf{Q}_b \cdot j_{\rm obs}$ and $j_{\rm thr}$ is minimized. Let $\mathbf{Q}_c$ be $\mathbf{Q}_b\mathbf{Q}_a$. The inverse of $\mathbf{Q}_c$, $\mathbf{Q}_c^{-1}$, equals $\mathbf{Q}_c^{\rm T}$, and is now equivalent to the $\mathbf{U}$ matrix defined by Busing & Levy (1967). The unit-cell transformation matrix is generated from the unit-cell parameters, which is equivalent to the Busing–Levy $\mathbf{B}$ matrix. The entire orientation matrix $\mathbf{R}$ is calculated as

$$\mathbf{R} = \mathbf{UB}. \tag{6}$$

This matrix R can be applied to integer (h, k, l) values to map them onto rotated Cartesian coordinates in reciprocal space.

2.11. Indexing additional lattices using the inter-spot vector network

The indexing of additional lattices using the inter-spot vector network is made possible by a previous method (Gildea *et al.*, 2014) in which spots which are predicted by an existing solution are removed from the data set and indexing is resumed with the remaining spots. Inter-spot vector networks which lead to a previously determined orientation matrix, or a geometrically equivalent matrix, are rejected. Solutions are scored for their equivalence as described above.

3. Results

3.1. One-dimensional pseudo-powder patterns

We find that well defined pseudo-powder patterns (generated as described in §2) can be obtained even when true powder patterns cannot be usefully calculated. For instance, when few patterns are available pseudo-powder patterns can still be produced. These are not only useful for diffraction patterns with small unit-cell dimensions (Fig. 1*a*), but are still interpretable for diffraction patterns from crystals with large unit cells and very poor reflection separation, as for BEV crystals (Fig. 1*c*).

An example of the vectors used in a single image to contribute to the overall pseudo-powder pattern is shown for CPV17 crystals in Fig. 2. The conventional powder patterns generated by overlaying multiple images may show fewer peaks than the pseudo-powder patterns owing to axial systematic absences. The corresponding distances will not be present in a conventional powder pattern, but will appear in the pseudo-powder pattern as such vectors occur off-axis.

As the number of BEV crystals was very small (304 images), a conventional powder pattern could not be generated by superimposition of the images, and if it had been possible, the

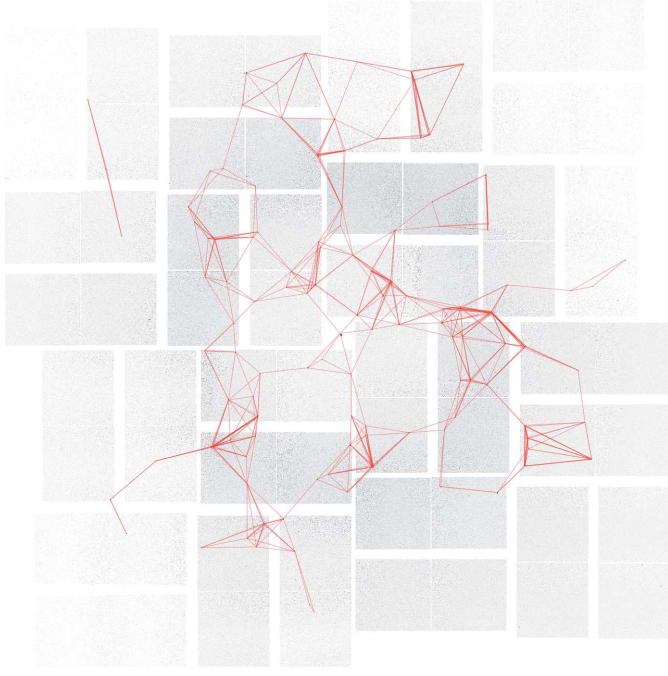


Figure 2
An example CPV17 crystal diffraction image at 101.2 mm detector distance with vectors picked between spots which lie within 0.1 Å^{-1} of each other. This will contribute a small portion of the pseudo-powder pattern as seen in Fig. 1. The Miller index (0, 0, 0) is also included in the analysis to boost the number of inter-spot vectors identified within the image.

rings would be impossible to distinguish: the unit-cell dimension is 437 Å and even the longest reciprocal-space distance in the pseudo-powder pattern (50 Å) only spanned 8 mm across the detector and was therefore lost in the unrecorded low-angle region. Only after a one-dimensional pseudo-powder pattern had been generated was it revealed that the original detector-distance reading was misaligned by 20 mm. This value could be refined manually by observing the effect on the pseudo-powder pattern (Figs. 1*c* and 1*d*). This enabled indexing of these diffraction patterns in a situation where the presence of this anomaly was otherwise difficult to establish. The BEV images had very poor separation of individual reflections, with the Miller index vector (1, 1, 1) corresponding to only three pixels. Assuming an error in a spot position of ∼0.5 pixels, the error in such a vector will be about 23%, having a major impact on the pseudo-powder pattern especially at small reciprocal distance values.

Owing to the orthorhombic nature of myoglobin, powder rings will overlap in ways which are not observed in cubic space groups. This will lead to a degree of uncertainty during the indexing stage when assigning theoretical vectors to observed vectors according to their distances. The pseudo-powder pattern for the myoglobin data set is shown in Fig. 1(*b*).

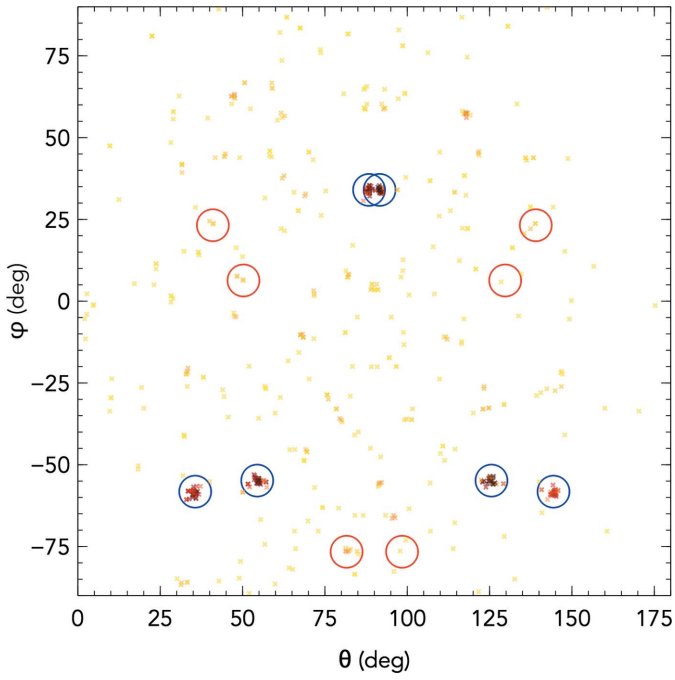


Figure 3
Matrix-cluster solutions decomposed into two dimensions, showing one strong (denoted by blue circles) and one weak solution (denoted by red circles) on the same image. Darker areas show matrices which have a high number of neighbouring solutions within an 8° angle, whereas yellow solutions have the lowest number of neighbours within the 8° angle and are likely to be noise. Both circles correspond to solutions which lead to a successful, unique indexing solution. Each solution has six symmetry-related solutions generated by indexing relative to geometrically equivalent axes. The image is from a CPV17 sample at 101.2 mm.

3.2. Rotation-matrix clusters

The rotation-matrix cluster method, as described in §2, has a poor signal-to-noise ratio and can only reliably reach above the level of noise in cubic space groups, where only geometrically equivalent powder rings overlap. Projecting the orientation matrices onto two dimensions shows clusters of orientation matrices which have a high number of neighbours, and these clusters lead to indexing solutions (Fig. 3). However, this method is not well suited to the hexagonal space group of

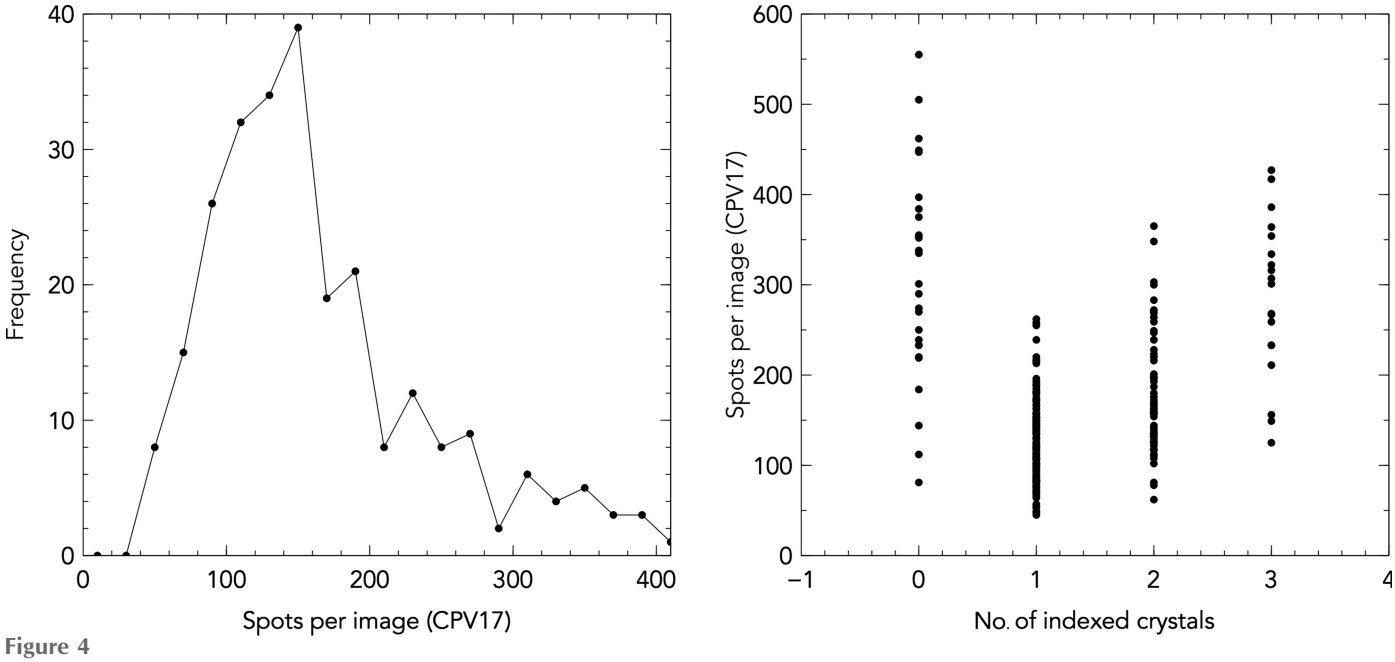


Figure 4
Left: a histogram of frequency of images with a given number of spots for the CPV17 sample at 101.2 mm detector distance. Images with higher numbers of spots typically have larger numbers of lattices. The graph is cropped at a maximum of 400 spots for clarity. Right: the frequency of spots on the same images grouped by the number of successfully indexed lattices. Images which fail to index typically have significantly larger spot counts.

Table 3
Crystal symmetry and wavelength/detector distance for the data sets.

Data set	Space group	Unit cell (Å)	Wavelength (Å)	Distance (mm)	Total images	Best indexing rate (%)
CPV17	$I23$	106.1 × 106.1 × 106.1	1.46	101.2	262	151.1
CPV17	$I23$	106.1 × 106.1 × 106.1	1.46	91.0	1380	99.3 (single lattice)
BEV	$F23$	437 × 437 × 437	1.31	85.0	304	97.6
Thermolysin	$P6_122$	92.9 × 92.9 × 130.4	1.27	175.0	1866	90.1
Myoglobin	$P2_12_12_1$	38.0 × 46.9 × 84.5	1.29	68.6	3767	112.2

thermolysin, perhaps because of the less distinct separation of powder rings. A sample of 262 CPV17 images demonstrated a 130% indexing rate (*i.e.* more than one lattice per hit). Of all the images, 87% provided at least one lattice, with data extending to 1.98 Å at the edge of the detector. These images often contained multiple lattices, which were identifiable by the abnormally high numbers of spots (Fig. 4). The indexing rates obtained for these images in the previously published structure (Ginn, Messerschmidt *et al.*, 2015) was 52%.

3.3. Inter-spot vector networks

Inter-spot vector networks were successfully applied to all of the test systems. For CPV17, inter-spot vector distances were considered up to a reciprocal distance of 0.19 Å^{-1}, which was manually chosen to give the optimal indexing rate. When the number of inter-spot vectors within the network exceeded 20, the corresponding indexing solution was confirmed by initial orientation-matrix refinement (Ginn, Messerschmidt *et al.*, 2015). At this point the majority of inter-spot vector networks provided a correct indexing solution, as judged by orientation-matrix refinement. When indexing was limited to one lattice solution per image, a lattice was found for 92% of images, a higher proportion of indexed images than the result for rotation-matrix clusters. Enabling multiple lattice indexing (up to three distinct lattices on one image) increased the indexing rate to 151%. Of the 29 images which failed to produce any orientations after indexing, 18 were weak or of low resolution, two showed no evidence of diffraction, one had over 1000 spots (at the extreme of the distribution shown in Fig. 4) and eight showed no obvious reasons for indexing failure. Another sample of 1380 CPV17 images which extended to 1.84 Å at the edge of the detector demonstrated an indexing rate of 99.3% when only considering single lattices.

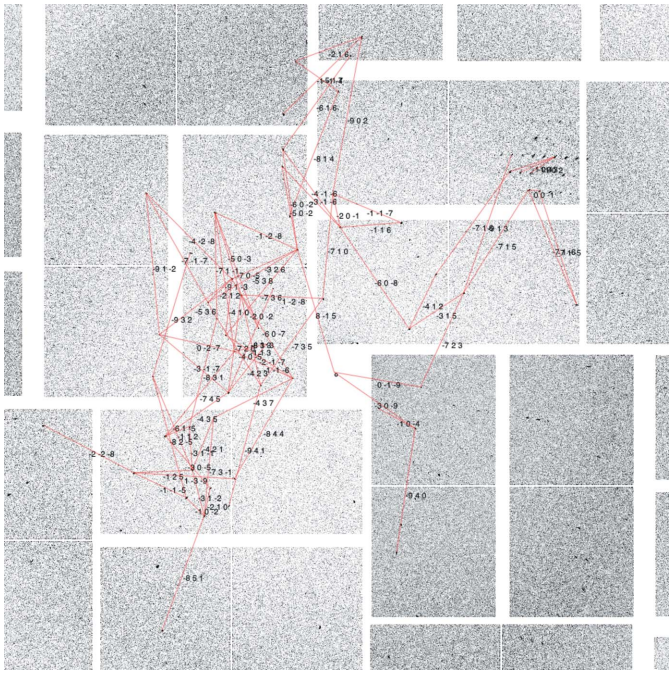


Figure 5
Map of inter-spot vectors within an image of a thermolysin crystal diffraction pattern which contribute to a correct indexing solution. The Miller index translation between the spots is marked beside each vector. The total number of vectors was limited to 80.

Thermolysin crystals were indexed, selecting vectors with lengths of up to 0.1 Å^{-1} also manually chosen to give the best indexing rates. There was a high incidence of multiple lattices in this data set, which made finding inter-spot vector networks more difficult (example in Fig. 5). However, when searching for multiple lattices an indexing rate of 90% was achieved for this data set.

BEV crystals performed similarly to CPV17 crystals but had to be processed several times with differing combinations of parameters (Table 3) to achieve the highest indexing rate of 97.6% from a pool of 304 images, despite the spots being much closer together. The greatest distance considered in reciprocal space is 0.06 Å^{-1}. The inter-spot vector network is likely to comprise more vectors which span an appreciable distance across the detector, including longer vectors where the errors are proportionately lower.

With respect to myoglobin, 69% of the crystals, which were crystallized in space group $P2_12_12_1$ with distinct unit-cell axes, generated an indexing solution using the inter-spot vector-network method using a vector distance tolerance of 5 × 10^{-4} Å^{-1}. When searching for multiple lattices, the indexing rate increased to 112%. Inter-spot vectors were considered within a reciprocal-lattice distance of 0.15 Å^{-1}. The failed crystals were generally those of lower resolution and had insufficient spots to support indexing with the chosen parameters. Overall, all the images had an average of 148 ± 106 reflections and those which failed to index had a lower average of 80 ± 22 reflections, illustrating the poorer crystal quality of these images.

3.4. Tolerance to errors in experimental parameters

To test the robustness of the algorithm, the experimental parameters were altered and the effect on the indexing success rate was observed using CPV17 data. As the detector distance and beam centre are the most likely to be incorrectly measured at an XFEL beamline, these parameters were varied. The detector distance was varied from its true value of 101.2 mm by 1.0 mm in either direction, and similarly for the

beam centre X position. This algorithm had a good tolerance to experimental errors, and although the indexing rates decreased when the model parameters deviated from their true values, this was in the form of a gradual decay. Indexing success only rapidly decreased after moving approximately 0.5 mm from the optimal value for both detector distance and beam centre (Fig. 6). For comparison, the shortest vector for this space group would have spanned a minimum of 1.77 mm across the detector if it had been located near the beam centre.

4. Discussion and conclusions

For the reliable indexing of diffraction patterns, experimental parameters such as the direct beam position and crystal-to-detector position need to be known reasonably accurately. Powder patterns aggregated from a large number of diffraction patterns are useful in establishing these parameters. We have found that in cases where only a small number of images are available, or the unit cell is large enough to obscure the low-resolution rings or prevent adequate separation of rings, a pseudo-powder pattern can still be generated by considering the projected three-dimensional vectors between spots on the Ewald sphere. We find that such pseudo-powder patterns are generally useful tools for detecting and correcting significant errors in experimental parameters such as the crystal-to-detector distance, which can still be poorly estimated at XFEL beamlines.

We describe two new approaches to indexing still images such as those derived from serial crystallography. The first approach, termed the rotation-cluster method, is well suited to low-resolution data with a small number of reflections, whereas the second method, using inter-spot vector networks, was found to be more generally robust and particularly suited to the test case in a hexagonal space group. Taken together, the methods presented here have been shown to be highly effective for cubic (CPV17 and BEV), hexagonal (thermolysin) and orthorhombic (myoglobin) space groups, and are likely to be of general utility. In the test cases chosen, the indexing rate was high and the number of lattices indexed ranged from 80 to 150% of the number of images classified as harbouring diffraction.

The large majority of the images which failed to be indexed by the methods described here appeared to originate from weak crystals resulting in diffraction to only low resolution, low XFEL pulse intensity or a combination of these, as judged from the 262-image sample of CPV17. This provides reassurance that the images which failed to index would have provided the least amount of useful information for the data-reduction stage if they had been successfully indexed.

The *TakeTwo* algorithm could be incorporated into other XFEL crystallography software suites as an alternative method. Combining the results from multiple algorithms and removing duplicate solutions should lead to a larger indexed percentage than that obtained using any one algorithm alone. We include the *TakeTwo* algorithm in the *cppxfel* software suite (Ginn *et al.*, 2016). Potential users may download this software and to do so should visit http://viper.lbl.gov/cctbx.xfel/index.php/Cppxfel.

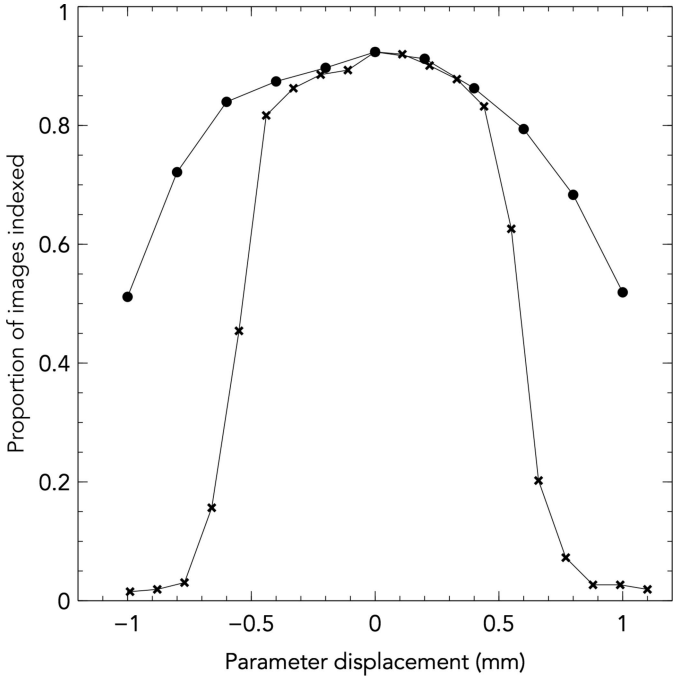


Figure 6
Tolerance of the indexing algorithm to the detector-distance parameter (filled circles) and beam-centre X parameter (crosses) are shown. Indexing success decreases significantly after 0.5 mm displacement from the optimal detector distance. The CSPAD detector pixel size is 0.11 mm. Calculated for the CPV17 sample at 101.2 mm detector distance.

Acknowledgements

DIS was supported by the Medical Research Council, grant MR/N00065X/1 and previously G1000099. HMG was supported by the Wellcome Trust (studentship 075491/04). NKS was supported by US National Institutes of Health grant R01-GM102520. The Canadian Institute for Advanced Research supported RJDM, OPE and DIS. Portions of this research were carried out at the Linac Coherent Light Source (LCLS) at the SLAC National Accelerator Laboratory. The research leading to these results has received funding from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme (FP7/20072013) under REA grant agreement No. 623994 (HM-W). We are grateful to Helen Duyvesteyn for testing and using the software. LCLS is an Office of Science User Facility operated for the US Department of Energy Office of Science by Stanford University. Administrative support was received from the Wellcome Trust (grant 090532/Z/09/Z). This is a contribution from the Oxford Instruct Centre.

References

Barends, T. R. *et al.* (2013). *Acta Cryst.* D**69**, 838–842.
Brünger, A. T. (1990). *Acta Cryst.* A**46**, 46–57.
Brünger, A. T. (1992). *X-PLOR* v.3.1. *A System for X-ray Crystallography and NMR*. Yale University, New Haven, USA.
Busing, W. R. & Levy, H. A. (1967). *Acta Cryst.* **22**, 457–464.
Campbell, J. W. (1998). *J. Appl. Cryst.* **31**, 407–413.

Chapman, H. N. *et al.* (2011). *Nature (London)*, **470**, 73–77.
Gildea, R. J., Waterman, D. G., Parkhurst, J. M., Axford, D., Sutton, G., Stuart, D. I., Sauter, N. K., Evans, G. & Winter, G. (2014). *Acta Cryst.* D**70**, 2652–2666.
Ginn, H. M., Brewster, A. S., Hattne, J., Evans, G., Wagner, A., Grimes, J. M., Sauter, N. K., Sutton, G. & Stuart, D. I. (2015). *Acta Cryst.* D**71**, 1400–1410.
Ginn, H. M., Evans, G., Sauter, N. K. & Stuart, D. I. (2016). *J. Appl. Cryst.* **49**, 1065–1072.
Ginn, H. M., Messerschmidt, M., Ji, X., Zhang, H., Axford, D., Winter, G., Brewster, A. S., Hattne, J., Wagner, A., Grimes, J. M., Sauter, N. K., Sutton, G. & Stuart, D. I. (2015). *Nature Commun.* **6**, 6435.
Hattne, J. *et al.* (2014). *Nature Methods*, **11**, 545–548.
Higashi, T. (1990). *J. Appl. Cryst.* **23**, 253–257.
Johansson, L. C. *et al.* (2013). *Nature Commun.* **4**, 2911
Kabsch, W. (1977). *J. Appl. Cryst.* **10**, 426–429.
Kabsch, W. (1988). *J. Appl. Cryst.* **21**, 67–72.
Kabsch, W. (1993). *J. Appl. Cryst.* **26**, 795–800.
Kabsch, W. (2014). *Acta Cryst.* D**70**, 2204–2216.
Liu, W. *et al.* (2013). *Science*, **342**, 1521–1524.
Mueller, C. *et al.* (2015). *Struct. Dyn.* **2**, 054302.
Oghbaey, S., Sarracini, A., Ginn, H. M., Pare-Labrosse, O., Kuo, A., Marx, A., Epp, S. W., Sherrell, D. A., Eger, B. T., Zhong, Y., Loch, R., Mariani, V., Alonso-Mori, R., Nelson, S., Lemke, H. T., Owen, R. L., Pearson, A. R., Stuart, D. I., Ernst, O. P., Mueller-Werkmeister, H. M. & Miller, R. J. D. (2016). *Acta Cryst.* D**72**, 944–955.
Otwinowski, Z. & Minor, W. (2006). *International Tables for Crystallography*, Vol. *F*, 1st online ed., edited by M. G. Rossmann & E. Arnold, pp. 226–235. Chester: International Union of Crystallography.
Powell, H. R., Johnson, O. & Leslie, A. G. W. (2013). *Acta Cryst.* D**69**, 1195–1203.
Roedig, P., Vartiainen, I., Duman, R., Panneerselvam, S., Stübe, N., Lorbeer, O., Warmer, M., Sutton, G., Stuart, D. I., Weckert, E., David, C., Wagner, A. & Meents, A. (2015). *Sci. Rep.* **5**, 10451.
Sauter, N. K., Grosse-Kunstleve, R. W. & Adams, P. D. (2004). *J. Appl. Cryst.* **37**, 399–409.
Sauter, N. K., Hattne, J., Brewster, A. S., Echols, N., Zwart, P. H. & Adams, P. D. (2014). *Acta Cryst.* D**70**, 3299–3309.
Sherrell, D. A., Foster, A. J., Hudson, L., Nutter, B., O'Hea, J., Nelson, S., Paré-Labrosse, O., Oghbaey, S., Miller, R. J. D. & Owen, R. L. (2015). *J. Synchrotron Rad.* **22**, 1372–1378.
Steller, I., Bolotovsky, R. & Rossmann, M. G. (1997). *J. Appl. Cryst.* **30**, 1036–1040.
Uervirojnangkoorn, M., Zeldin, O. B., Lyubimov, A. Y., Hattne, J., Brewster, A. S., Sauter, N. K., Brunger, A. T. & Weis, W. I. (2015). *Elife*, **4**, e05421.
Waterman, D. G., Winter, G., Parkhurst, J. M., Fuentes-Montero, L., Hattne, J., Brewster, A., Sauter, N. K. & Evans, G. (2013). *CCP4 Newsl. Protein Crystallogr.* **49**, 16–19.
Zarrine-Afsar, A., Barends, T. R. M., Müller, C., Fuchs, M. R., Lomb, L., Schlichting, I. & Miller, R. J. D. (2012). *Acta Cryst.* D**68**, 321–323.
Zeldin, O. B., Brewster, A. S., Hattne, J., Uervirojnangkoorn, M., Lyubimov, A. Y., Zhou, Q., Zhao, M., Weis, W. I., Sauter, N. K. & Brunger, A. T. (2015). *Acta Cryst.* D**71**, 352–356.

JOURNAL OF SYNCHROTRON RADIATION

ISSN 1600-5775

Received 18 July 2016
Accepted 23 September 2016

Edited by M. Weik, Institut de Biologie Structurale, Univ. Grenoble Alpes, CEA, CNRS, France

Keywords: SAXS; radiation damage; *RADDOSE-3D*; radioprotectants; CorMap visualization.

Supporting information: this article has supporting information at journals.iucr.org/s

Development of tools to automate quantitative analysis of radiation damage in SAXS experiments

Jonathan C. Brooks-Bartlett,[a] Rebecca A. Batters,[a] Charles S. Bury,[a] Edward D. Lowe,[a] Helen Mary Ginn,[b] Adam Round[c,d,e] and Elspeth F. Garman[a]*

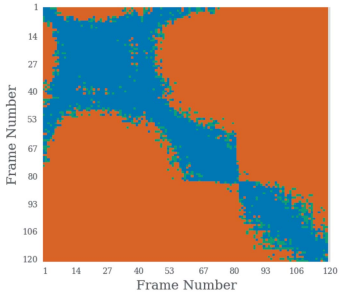
[a]Department of Biochemistry, University of Oxford, Oxford OX1 3QU, UK, [b]Division of Structural Biology, Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford OX3 7BN, UK, [c]European Molecular Biology Laboratory, Grenoble Outstation, 71 avenue des Martyrs, CS 90181, 38042 Grenoble, France, [d]SPB/SFX European XFEL, Holzkoppel 4, 22869 Schenefeld, Germany, and [e]Faculty of Natural Sciences, Keele University, Staffordshire ST5 5BG, UK. *Correspondence e-mail: elspeth.garman@bioch.ox.ac.uk

Biological small-angle X-ray scattering (SAXS) is an increasingly popular technique used to obtain nanoscale structural information on macromolecules in solution. However, radiation damage to the samples limits the amount of useful data that can be collected from a single sample. In contrast to the extensive analytical resources available for macromolecular crystallography (MX), there are relatively few tools to quantitate radiation damage for SAXS, some of which require a significant level of manual characterization, with the potential of leading to conflicting results from different studies. Here, computational tools have been developed to automate and standardize radiation damage analysis for SAXS data. *RADDOSE-3D*, a dose calculation software utility originally written for MX experiments, has been extended to account for the cylindrical geometry of the capillary tube, the liquid composition of the sample and the attenuation of the beam by the capillary material to allow doses to be calculated for many SAXS experiments. Furthermore, a library has been written to visualize and explore the pairwise similarity of frames. The calculated dose for the frame at which three subsequent frames are determined to be dissimilar is defined as the radiation damage onset threshold (RDOT). Analysis of RDOTs has been used to compare the efficacy of radioprotectant compounds to extend the useful lifetime of SAXS samples. Comparison of the RDOTs shows that, for radioprotectant compounds at 5 and 10 m*M* concentration, glycerol is the most effective compound. However, at 1 and 2 m*M* concentrations, dithiothreitol (DTT) appears to be most effective. Our newly developed visualization library contains methods that highlight the unusual radiation damage results given by SAXS data collected using higher concentrations of DTT: these observations should pave the way to the development of more sophisticated frame merging strategies.

1. Introduction

Biological small-angle X-ray scattering (SAXS) is an experimental technique that provides low-resolution structural information on macromolecules. The surge in popularity of the technique is a result of recent improvements in both software and hardware, allowing for high-throughput data collection and analysis (Bizien *et al.*, 2016; Graewert & Svergun, 2013). This is reflected in the increasing number of dedicated SAXS beamlines such as BM29 at the ESRF, P12 at the EMBL Hamburg and B21 at Diamond Light Source (Blanchet *et al.*, 2015; Brennich *et al.*, 2016; Materlik *et al.*, 2015; Pernot *et al.*, 2013).

However, as for most other macromolecular structural techniques, radiation damage is still a major factor hindering the success of experiments. The high solvent proportion of



OPEN ACCESS

biological SAXS samples means that hydroxyl, hydroperoxyl radicals and hydrated electrons are produced in abundance by the radiolysis of water when it is irradiated with X-rays (Allan *et al.*, 2013; Garman, 2010; Kuwamoto *et al.*, 2004). These radicals can then interact with the protein molecules, ultimately leading to protein aggregation, fragmentation or unfolding (Garrison, 1987; Hopkins & Thorne, 2016; Houée-Lévin *et al.*, 2015; Kuwamoto *et al.*, 2004). The categories of these radiation-induced processes can generally be distinguished by their various effects on the SAXS data. Aggregation manifests as an increase in scattering intensity at low angles, whereas fragmentation has the opposite effect, resulting in a decrease in scattering at low angles (Jeffries *et al.*, 2015; Kuwamoto *et al.*, 2004). Furthermore, molecular repulsion due to protein charging can also decrease the scattering at low angles. Unfolding, on the other hand, results in an increase in the radius of gyration and a decrease in the Porod exponent (Hopkins & Thorne, 2016).

Common methods used to reduce radiation damage to biological SAXS samples are generally concerned with limiting the X-ray exposure to any given volume of sample. These include: flowing or oscillating the sample in the container (usually a quartz or glass capillary, but could also be a flat cell with mica windows or a printed solution holder), reducing the exposure time, and attenuation or defocusing of the X-ray beam (Fischetti *et al.*, 2003; Jeffries *et al.*, 2015; Martel *et al.*, 2012; Pernot *et al.*, 2010). In an analogous manner to cryo-cooling in macromolecular crystallography (MX) (Garman, 1999), cryo-cooling samples down to 100 K for SAXS (cryoSAXS) has been reported to increase the dose tolerance of SAXS samples by at least two orders of magnitude (Meisburger *et al.*, 2013). Despite developments to improve the experimental apparatus for cryoSAXS (Hopkins *et al.*, 2015), it still requires specialized equipment and a certain level of technical expertise, which prevents it from currently being a commonly accessible technique (Jeffries *et al.*, 2015). Additives such as glycerol, ethylene glycol, sucrose and sodium ascorbate can be added to the SAXS sample to increase the dose tolerance of the sample (Grishaev, 2012; Kuwamoto *et al.*, 2004). However, additives are also known to reduce the overall scattering signal (Jeffries *et al.*, 2015).

Application of the above radiation damage mitigation approaches are unable to completely circumvent its detrimental effects, in particular the change of the scattering profile throughout the experiment. It is necessary to determine whether any two scattering profiles are similar so that noise can be reduced by averaging over similar curves. The χ^2 statistic, which describes the global goodness-of-fit of a model (Franke *et al.*, 2012), and the SAXS Merge method, which employs Gaussian process regression (Spill *et al.*, 2014), have previously been used to determine the similarity of scattering profiles. The CorMap test (Franke *et al.*, 2015) is the most recent method developed to assess the similarity of frames. Its advantage is that it does not use the errors on the experimental intensities, which are generally incorrectly estimated due to inaccurate propagation through the data processing pipeline.

For experiments by different researchers to be intercomparable, the progression of radiation damage is most usefully tracked as a function of the dose absorbed by the sample. *RADDOSE-3D* is a free and open source software program used to calculate the time- and space-resolved three-dimensional distribution of the dose absorbed by a protein crystal in an MX experiment (Zeldin *et al.*, 2013*a*). However, there is no equivalent software available for SAXS. Radiation damage studies in SAXS thus currently require the experimenters to correctly parameterize the experiment and manually calculate a single estimate of the dose within the sample (Hopkins & Thorne, 2016; Jeffries *et al.*, 2015; Meisburger *et al.*, 2013).

Here, extensions to *RADDOSE-3D* are presented, which enable the convenient calculation of doses for SAXS experiments, reducing the burden of manually performing the calculation. The three-dimensional geometry of the experiment is also taken into account. Furthermore, a new Python library has been developed to provide visual analysis of the results of scattering curve similarity tests from applying the CorMap test. Using this, it was possible to define a simple metric for significant radiation damage onset. These tools are used to compare the efficacy of eight different radioprotectant additives tested at various concentrations for their capability to increase the dose tolerance of glucose isomerase in a SAXS experiment. Additionally, the reduction in the scattered intensity signal is also analysed and the trade-off between dose tolerance and the signal reduction is briefly discussed.

2. Materials and methods

2.1. Sample preparation

Glucose isomerase (GI) from *Streptomyces rubiginosus* was chosen as the model protein for these experiments because it is a stable and soluble globular protein that has well defined SAXS behaviour, and it is sufficiently large to scatter well at modest concentrations (Kozak, 2005). It was purchased in tetrameric form (1552 residues, 172 kDa) from Hampton Research (GI mutant E186Q, now no longer commercially available) and was dissolved and dialysed for 24 h at 277 K against 100 m*M* HEPES and 10 m*M* $MgCl_2$ buffer at pH 7.0. The final GI concentration, 1 mg ml^{-1}, was used for all data collection runs and was determined using the known extinction coefficient, 45660 M^{-1} cm^{-1} at 280 nm.

Eight soluble additives were tested over a range of concentrations for their radiation damage protection capabilities: dithiothreitol (DTT), ethylene glycol, glycerol, sodium ascorbate, sodium nitrate, sucrose, (2,2,6,6-tetramethylpiperidin-1-yl)oxyl (TEMPO) and trehalose.

The additives were added to the buffer solution (100 m*M* HEPES and 10 m*M* $MgCl_2$ buffer at pH 7.0) without protein at four different concentrations: 10 m*M*, 5 m*M*, 2 m*M* and 1 m*M*, except for glycerol and ethylene glycol, which were both prepared at 10% *v*/*v*, 5% *v*/*v*, 2% *v*/*v* and 1% *v*/*v* immediately prior to data collection.

These additives were also prepared to the same final concentration in the solution containing both the buffer and protein.

2.2. Data collection

Data collection was performed at the ESRF BioSAXS beamline BM29 (Blanchet *et al.*, 2015; Brennich *et al.*, 2016; Materlik *et al.*, 2015; Pernot *et al.*, 2013). The photon energy used throughout was 12.5 keV and the photon flux was estimated from the beamline diode readings which were recorded for every frame using the conversion formula

$$\text{flux} = 5.72293 + 2.72295 \times 10^{15} \times d_{\text{c}}, \quad (1)$$

where d_{c} is the reading of the diode mounted within the backstop. The flux obtained using this formula was calibrated as described by Owen *et al.* (2009), here using an OSD1-0 photodiode purchased from Optoelectronics, which was a 500 µm-thick silicon diode with a 1 mm^2 active area.

The data were recorded using a Pilatus 1M detector from Dectris. 15 µl of each sample was loaded into a 1.8 mm external diameter quartz capillary (1.7 mm internal diameter, thus the wall thickness is 50 µm) held at 20°C using the BioSAXS Sample Changer described by Round *et al.* (2015). For every additive, data were collected at each of the concentrations stated in §2.1, and each of these individual data collection runs was repeated three times (*i.e.* three datasets per radioprotectant per concentration). The exposure time for each frame was 1 s, and a total of 120 frames were collected for each repeat (*i.e.* 120 s total exposure for each dataset) with the sample kept static. After data collection on a sample, the capillary was washed with cleaning solution (2% Hellmanex, 10% ethanol and 88% distilled water), rinsed with distilled water and dried with dry air, a procedure also described by Round *et al.* (2015). For each radioprotecant concentration, a single dataset was collected with only the buffer (no protein) and the radioprotectant, so that a suitable buffer correction (subtraction) could be applied during data analysis. To obtain scattering curves on an absolute scale, a water calibration measurement was used.

2.3. Data processing

Azimuthal integration of diffraction frames was performed as described in the corresponding section (§3.1) of Brennich *et al.* (2016). A custom script was written in Python to average the frames from the datasets collected with only the buffer with each radioprotectant added. These averaged frames were then subtracted from the frames collected with both the buffer and protein sample. Finally, the frames were cropped at both the lowest and highest scattering angles using the same Python script. The cropped scattering angles were the same for all datasets, and the choice of angles was determined by visual inspection to remove the regions with a higher level of noise.

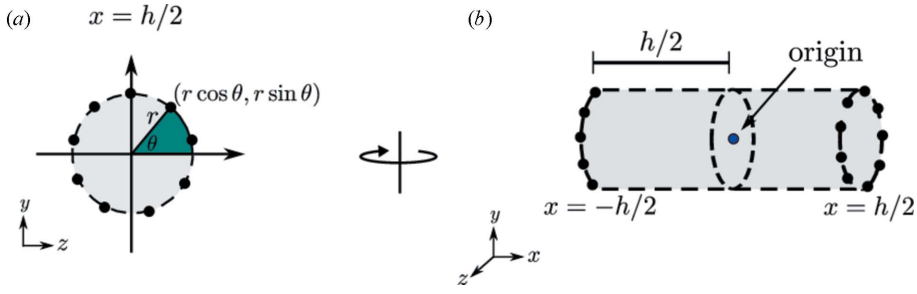


Figure 1
Implementation of the SAXS cylindrical sample geometry in *RADDOSE-3D* given a defined diameter, d, and height, h. (*a*) Evenly spaced points around a circle are generated given the radius, $r = d/2$, of the circular cross section. *RADDOSE-3D* defaults to 32 points. (*b*) In three dimensions the points represent the circles at each end of the cylinder at a distance of $h/2$ from the origin located at the centre of the cylinder.

2.4. Extending *RADDOSE-3D* for SAXS

2.4.1. Cylindrical sample geometry. In many SAXS experiments, liquid samples are contained in, or flowed through, a cylindrical capillary during the X-ray exposure, so it is necessary to model cylindrical sample shapes. To extend the capability of *RADDOSE-3D* to be used for the calculation of dose in SAXS experiments, a cylindrical shape geometry was implemented. It specifies the geometry of the sample alone, but not the capillary in which it is contained, and is graphically depicted in Fig. 1. (The effect of the capillary is dealt with separately in §2.4.3.) First, the points around a circle are generated using the diameter of the circular cross section. (*RADDOSE-3D* uses 32 points around the circle by default). The points are evenly spaced around the circle with y, z coordinates $(r\cos\theta, r\sin\theta)$. The angle (in radians) between any two consecutive points is $2\pi/32$. A cylinder can be defined by the circles at either end of the shape, so this is achieved using the final coordinate x. Depending at which end a particular point lies, it will have coordinates $(x, y, z) = (h/2, r\cos\theta, r\sin\theta)$ or $(x, y, z) = (-h/2, r\cos\theta, r\sin\theta)$. Note that *RADDOSE-3D* assumes that the origin of the system is located at the centre of the cylinder by default but that this can be changed by the user.

2.4.2. Determining the sample composition. The overall absorption coefficient of the sample, μ_{abs}, is calculated from the individual atomic absorption coefficients, σ_j, as

$$\mu_{\text{abs}} = (1/V_{\text{c}}) \sum_{j=1}^{N} \sigma_j, \quad (2)$$

where V_{c} is the volume of the unit cell, N is the number of atoms in the unit cell and $\sigma_j = \sigma_j^{\text{Thompson}} + \sigma_j^{\text{Compton}} + \sigma_j^{\text{Photoelectric}}$ (Murray *et al.*, 2004). To determine the atomic composition of the sample, a volume of liquid is defined and its contents estimated, given its protein concentration and buffer composition.

First the molarity of the solution is calculated using the formula

$$\text{Molarity (mol L}^{-1}) = \frac{\text{sample concentration (g L}^{-1})}{\text{molecular mass (g mol}^{-1})}. \quad (3)$$

The sample concentration is provided in units of grams per litre ($\equiv$ mg ml^{-1}). The molecular mass of the molecule is calculated from other parameters provided. If the sequence file is given for the protein (the sample can also contain DNA and RNA), then the molecular mass can be determined by summing the molecular mass of each residue in the file. Otherwise, an average molecular weight is used for each residue (110.0 Da for protein residues, 339.5 Da for RNA nucleotides and 327.0 Da for DNA nucleotides).

The number of molecules in the volume can then be calculated by multiplying the molarity, volume and Avogadro's number ($N = 6.022 \times 10^{23}$ mol^{-1}), and the result is then rounded to the nearest integer.

2.4.3. Beam attenuation due to the capillary. In a typical MX experiment at 100 K, a crystal is exposed directly to the X-ray beam. In contrast, samples from SAXS experiments are held inside a quartz capillary. Thus the X-ray flux is attenuated due to the capillary, and account must be taken of this effect before calculating the dose absorbed by the sample. The transmission fraction of an X-ray beam due to a material with mass thickness x and density ρ is given by

$$I/I_0 = \exp\left[-(\mu/\rho)x\right], \quad (4)$$

where I is the emergent intensity of the beam after penetrating the material, I_0 is the incident intensity and μ/ρ is defined as the mass attenuation coefficient (Hubbell & Seltzer, 1995). The mass thickness, x, is defined as the mass per unit area and is given by $x = \rho t$ where t is the thickness of the material. The attenuation fraction caused by the capillary can hence be calculated as $1 - I/I_0$.

The mass attenuation coefficients for each element are tabulated in the National Institute of Standards and Technology (NIST) tables. For mixtures, the total attenuation coefficient is given by

$$\mu/\rho = \sum_i w_i(\mu/\rho)_i, \quad (5)$$

where w_i and $(\mu/\rho)_i$ are the fraction by weight and the mass attenuation coefficient of the ith atomic constituent, respectively.

2.5. Dose calculation

Doses were calculated using *RADDOSE-3D* (Zeldin *et al.*, 2013*a*) and all doses referred to here are diffraction weighted dose (DWD) values (Zeldin *et al.*, 2013*b*). The flux was calculated for each frame because the diode readings continuously changed between frames due to the decay of the electron current in the storage ring (see Fig. S1 of the supporting information). However, the overall change in diode current was only 0.54% during the course of a single run. Despite the small percentage change, this effect was still taken into account in the analysis.

A 100 µm-diameter circular aperture was scanned across the X-ray beam to obtain measurements of the beam profile.

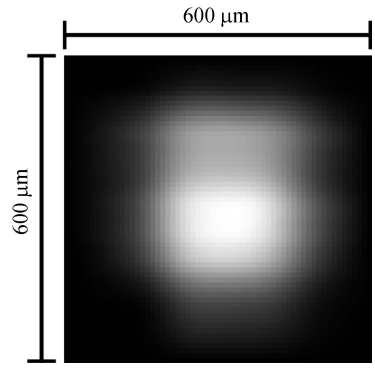


Figure 2
A two-dimensional reconstruction of the beam used in the experiment shown as a greyscale image. The intensity scales linearly between pixels.

The readings were taken at 10 µm intervals with an OSD1-0 photodiode purchased from Optoelectronics. The scanning was performed six times with the collection of three horizontal and three vertical scans (Fig. S2 of the supporting information).

To calculate a full two-dimensional beam profile from these aperture scans, a computational rectangular grid was constructed with the edges of the measurement positions used as the boundaries of the grid. The flux at and beyond the grid boundaries was assumed to be zero. The diode measurements from the vertical aperture scans were placed in their corresponding positions on the grid and interpolation between these values was performed using the 'RectBivariateSpline' function in the *SciPy* package in the Python programming language (Jones *et al.*, 2001). The same procedure was performed for the data in the horizontal direction. The interpolated two-dimensional beams were then averaged to obtain the final two-dimensional beam profile which was used in the *RADDOSE-3D* simulation (Fig. 2).

2.6. Radiation damage onset using the CorMap test

The program *DATCMP*, distributed as part of the *ATSAS* suite of programs for processing SAXS data (Petoukhov *et al.*, 2012), was used to perform the one-dimensional scatter curve similarity analysis. *DATCMP* implements the CorMap test for assessing frame similarity (Franke *et al.*, 2015). This test performs a pairwise correlation between one-dimensional scattering curves derived from the diffraction frames, which involves taking the difference between the scattering curves. If the two curves are similar, and hence come from the same distribution, then the chance of observing a positive or negative value is 50%, which is the same chance as observing a head or tail when an unbiased coin is tossed. The Schilling distribution quantifies the likelihood of observing C number of heads/tails in a row, and this is extended to observing a given number of positive or negative values in the pairwise correlation map. The CorMap test calculates the probability, p, of observing the longest stretch of positive or negative values under the null hypothesis that the curves are similar. If p is

above a given significance threshold, then the frames can be considered similar.

For the purposes of the current investigation, the first three frames of each experiment were compared in a pairwise manner using the CorMap test to ensure that they were similar (p value > 0.01). Then all subsequent frames were compared with frame 1. Radiation damage was assumed to have become significant at the point when three consecutive frames (in order to exclude outliers, for example bubbles or particles, passing through the beam) were found to be dissimilar as determined by the CorMap test at the p = 0.01 significance level. The dose absorbed in the sample for the first of the three consecutive dissimilar frames was then denoted the threshold dose, $D_{\rm Thresh}$.

2.7. Signal reduction

As well as providing protection from radiation damage, the addition of a radioprotectant compound to the sample decreases the scattered intensity signal that constitutes the experimental data.

Fig. 3 shows the reduction in the scattered intensity signal when glycerol and trehalose are added as radioprotectants to the sample. To quantify this reduction, the ratio of the first 20 points (lowest q angles) of the samples containing the radioprotectant to the sample containing no radioprotectant was calculated for each of the eight compounds at each concentration. The mean of these values was then calculated and subtracted from 1 to give the average fractional reduction in the scattered signal. The lowest q values were chosen because that is where the signal reduction is most prominent.

3. Results

3.1. CorMap visualization tools

A library for the visualization of the results of the CorMap test was developed, largely due to the fact that no open source alternative existed. One of the core visualizations is the correlation matrix between pairwise frames (Fig. 4). These correlation maps give an indication of the similarity of two frames. When two frames are similar, the correlation map resembles a randomized lattice (Figs. 4*a* and 4*b*), whereas, when there are systematic differences between any two frames, there are large continuous regions of either black or white (Figs. 4*c* and 4*d*).

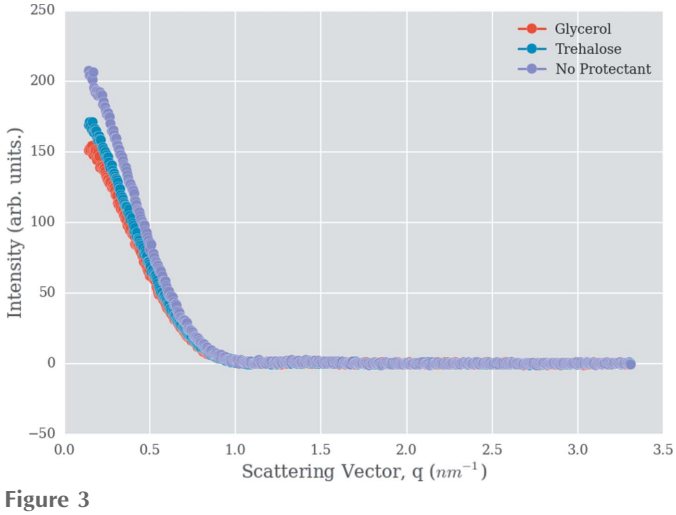


Figure 3
Scattered intensity signal reduction due to addition of radioprotectants. A reduction in the scattered intensity signal is seen when glycerol and trehalose are added as radioprotectants to the sample.

Another visual representation developed for the current investigation which gives insight into the damage progression is the scatter plot showing the similarity of all frames with any chosen reference frame. Fig. 5 shows such a scatter plot where frames 2–120 are compared with frame 1. The variable C represents the length of the longest continuous patch of white or black in the corresponding pairwise correlation map. Each of the points is coloured by the p value of the CorMap test. If the p value is equal to 1, suggesting that the frames are similar, then the point is coloured blue, whereas if the p value is below a specified significance level (default level is 0.01) suggesting that the frames are different, then it is coloured orange. Points that have p values between a specified significance level and 1 are coloured green.

3.2. Concentration dependence of radioprotectant efficacy

$D_{\rm Thresh}$ values for each experimental run for each additive radioprotectant compound were calculated and are plotted in Fig. 6. These enabled the order of increased dose tolerance to be determined at each radioprotectant concentration; however, to quantify the improvement, a new metric called the radiation damage onset threshold (RDOT) was developed. This metric is defined as the ratio of the median $D_{\rm Thresh}$ value with added radioprotectant to the median $D_{\rm Thresh}$ value for the same sample with no protection. Values below 1 correspond to a reduction in radiation tolerance whereas values above 1 show improved radiation tolerance. This metric was calculated for each compound at each concentration and the results are plotted in Fig. 7.

Significant concentration dependence can be observed for several of the radioprotectants. In particular, the efficacy of ascorbate, glycerol and sodium nitrate all exhibit a strong positive concentration dependence, *i.e.* the higher the concentration, the better the protection ability. However, DTT exhibits the opposite behaviour: at low concentrations it has the highest ratio, but this decreases at the higher concentrations. Sucrose, TEMPO and trehalose show a very small positive dependence but even at the highest concentration (10 mM) the RDOT is less than 2. This suggests that these radioprotectants are not very efficient at increasing the dose tolerance of the sample. The RDOT for ethylene glycol decreases as the concentration increases from 1 to 5 mM, but then there is a large increase at 10 mM. Thus the protection ability of ethylene glycol is not a simple monotonic function of the concentration.

The most effective radioprotectant varies depending on the concentration of the compound used in the sample. At low concentrations (1 and 2 mM) DTT is the most effective

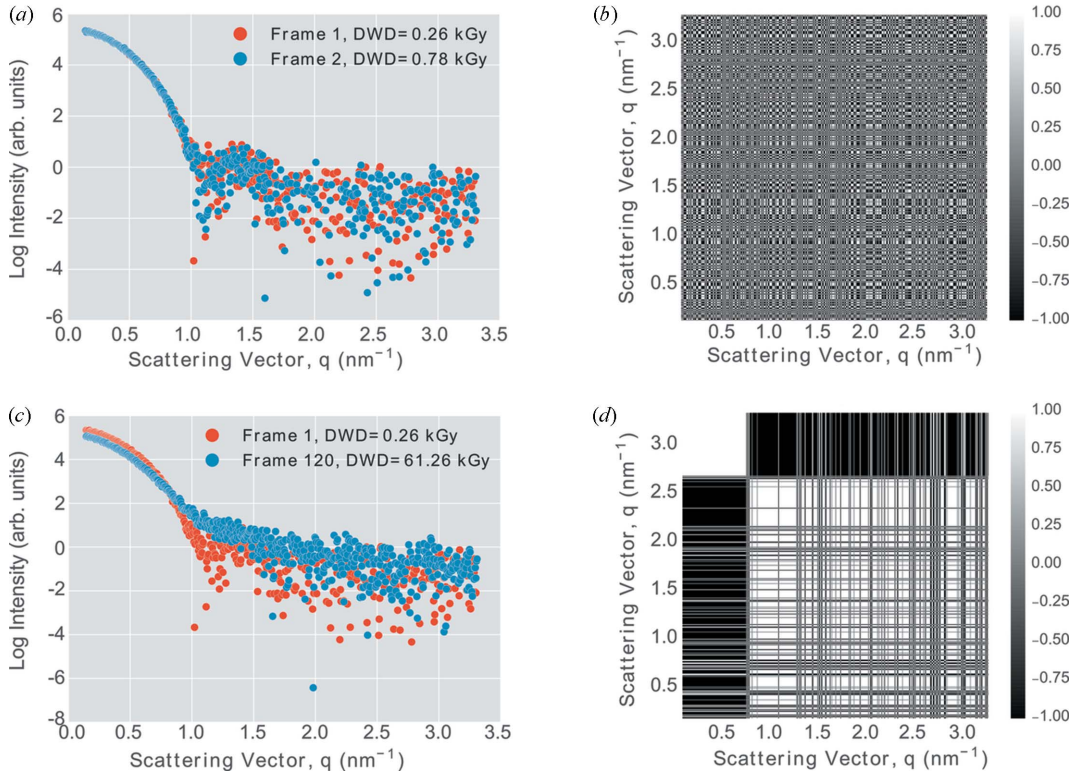


Figure 4
Similarity comparison with selected frames from the first experimental repeat for the GI protein and buffer but with no radioprotectant added. (*a*) One-dimensional scatter curves for frames 1 and 2. These two curves overlap well and are classed as similar. (*b*) Pairwise CorMap between frames 1 and 2. The ostensibly randomized lattice pattern suggests that the one-dimensional curves are similar. (*c*) One-dimensional scatter curves for frames 1 and 120. It is clear that these frames do not overlap. (*d*) Pairwise CorMap between frames 1 and 120. The dissimilarity between the two frames is represented by the large black and white regions.

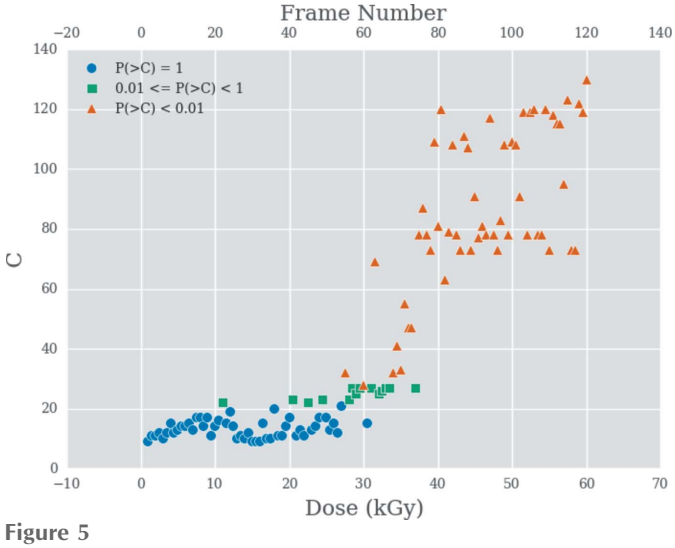


Figure 5
Longest observed black/white edge length in the correlation map, *C*, against the frame number and dose (kGy) for pairwise comparisons with frame 1. For similar frames to frame 1, the pairwise CorMaps resemble randomized lattices (Figs. 4*a* and 4*b*) and hence *C* is fairly small. Therefore, the chance of observing a longer edge length than *C* is high [$P(>C) = 1$; blue circles]. As frames start becoming more dissimilar, the *C* values increase and the $P(>C)$ values fall. These are represented by green squares. When frames are very dissimilar, *C* becomes very large (Figs. 4*c* and 4*d*) and $P(>C) < 0.01$. The symbols representing the comparison with these frames are orange triangles.

radioprotectant; however, at the higher concentrations (5 and 10 m*M*) glycerol becomes the most effective additive. Furthermore, the results also suggest that DTT is the least effective radioprotectant at the higher concentrations.

3.3. All pairwise frame comparisons should be made

To understand why the efficacy of DTT changed so drastically as the concentration was increased, further investigation on frame combinations was undertaken. Rather than just taking pairwise frame comparisons with the first frame alone (which implicitly assumes that the first frame is undamaged), all possible pairwise comparisons were performed and plotted in a heat map for the data where DTT was added at 10 m*M* concentration (Fig. 8). The colour scheme is the same as that used for the scatter plot, namely blue, orange and green suggest, respectively, that the frames are similar, dissimilar or may be considered similar but that the *p* value is not equal to 1. The first row of the heat map gives the results of the CorMap test for all pairwise frame comparisons with the first frame. It can be seen that a region of orange begins from frame 8 (DWD = 4.32 kGy), demonstrating why $D_{\rm Thresh}$ was so low for 10 m*M* DTT. The large blue square region in the top left of Fig. 8 suggests that these frames are all similar, and so much more useful data can be obtained if frames 7 (DWD = 3.74 kGy) to 57 (DWD = 32.50 kGy) are used for merging.

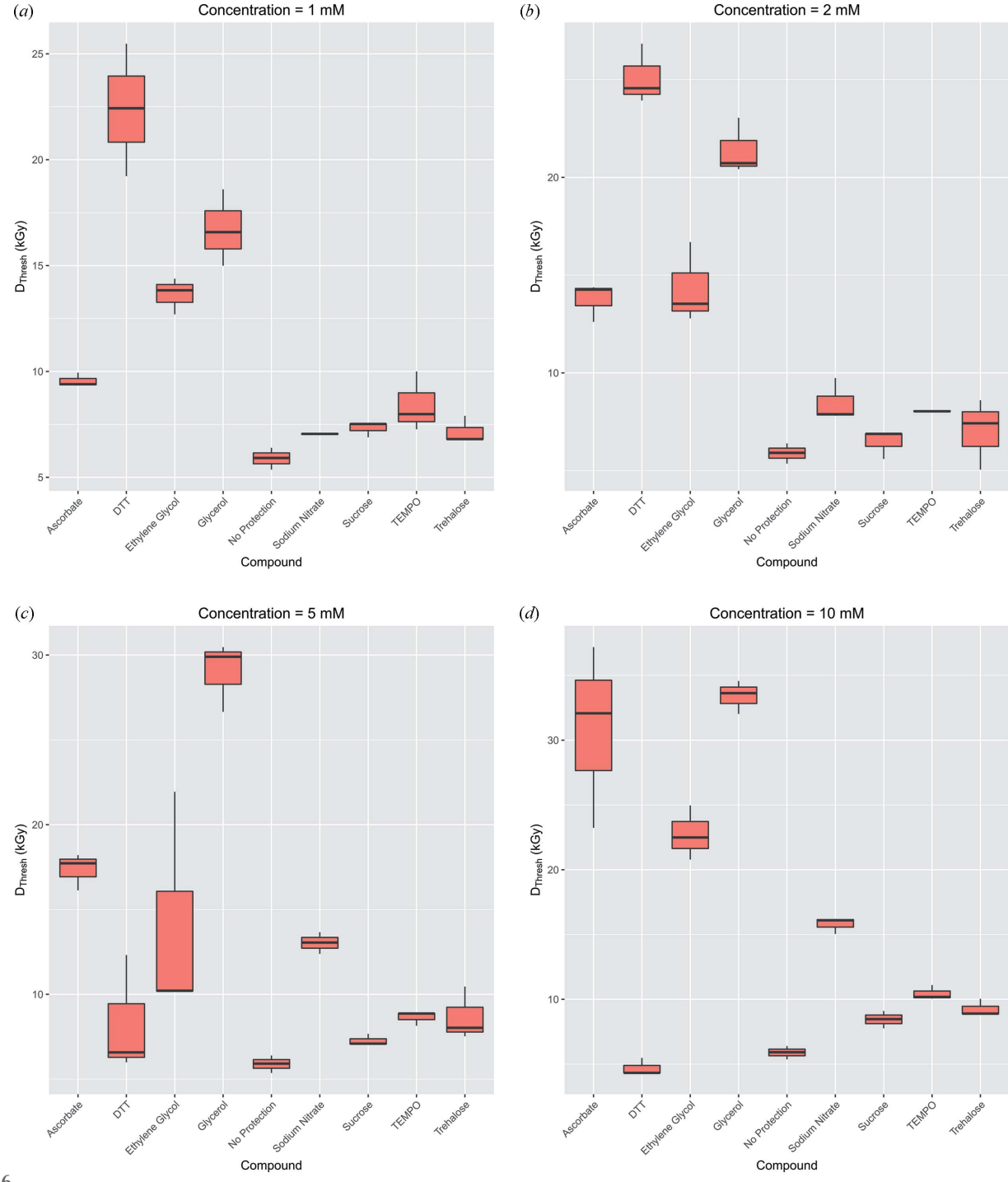


Figure 6
D_{Thresh} values as box plots for each of the three experimental repeats for each additive radioprotectant compound, including no additive at all. (*a*) 1 m*M*, (*b*) 2 m*M*, (*c*) 5 m*M*, (*d*) 10 m*M*.

This means that assessing all pairwise comparisons, as opposed to pairwise comparisons with the first frame only, may result in better quality data and improve the consistency of the radiation damage analysis.

3.4. Reduction of scattered intensity signal

Fig. 9 shows the reduction in the scattered intensity signal for each radioprotectant. The majority of the data show a reduction between 16 and 26% (for q values from 0.14 to 0.23 nm^{-1}). The compounds for which the signal reduction exceeds 26% are glycerol, sodium ascorbate and ethylene glycol, all of which were the best radiation tolerant compounds at concentrations of 5 and 10 m*M*, according to the RDOT metric. Notably, at 5 m*M* concentration, sodium ascorbate appears to reduce the signal the least, despite being the second most radiation tolerant additive of those used in our experiment. However, counterintuitively, at the lower

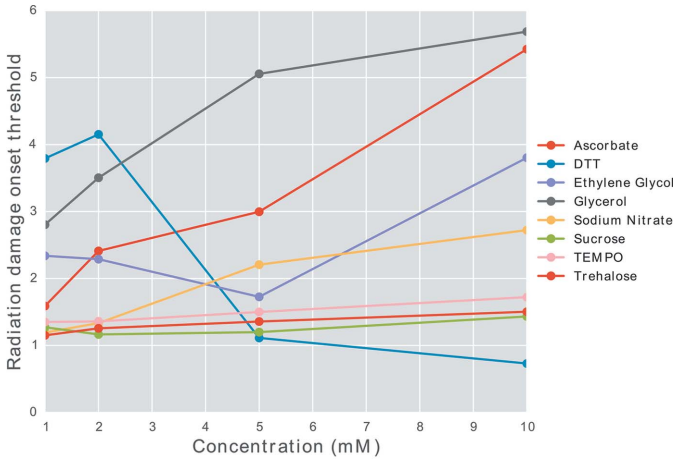


Figure 7
RDOT against concentration for the eight radioprotectants.

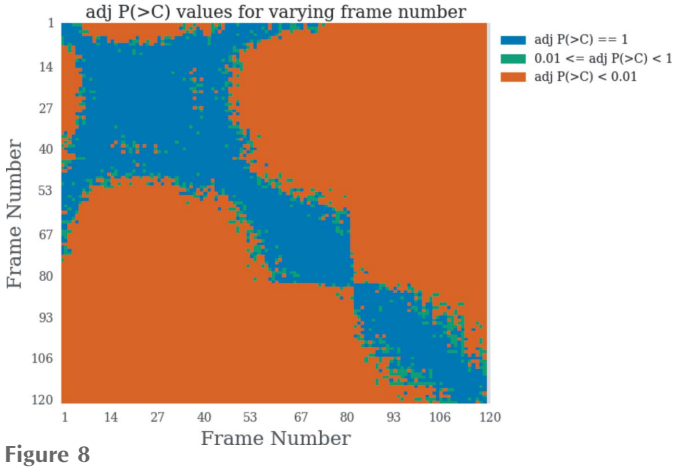


Figure 8
Heat map of all possible pairwise frame comparisons for the first repeat with 10 m*M* concentration DTT added to the GI sample. The *y*-axis represents the reference frame to which all other frames on the *x*-axis are compared. Blue: $P(>C) = 1$. Green: $0.01 \leq P(>C) < 1$. Orange: $P(>C) < 0.01$. The dose range is 0.29 kGy (frame 1) to 68.74 kGy (frame 120).

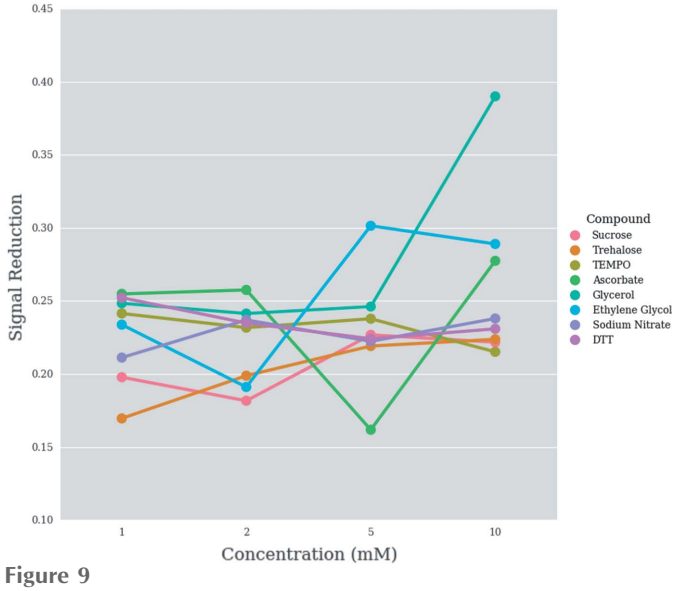


Figure 9
Scattered intensity signal reduction for each radioprotectant (see text for method of calculation).

concentrations (1 and 2 m*M*), it decreases the intensity signal more than any of the other compounds. This suggests that there may be a non-linear relationship between signal reduction, radiation tolerance and concentration. DTT, sodium nitrate and TEMPO exhibit a slightly negative correlation between the signal reduction and the concentration, *i.e.* as the concentration increases, the reduction in scattered intensity signal decreases.

4. Discussion

The work presented here describes how extensions to *RADDOSE-3D* to simulate SAXS experiments, along with additional analysis of results from the CorMap test, can be used to perform quantitative radiation damage analysis of SAXS experiments. In particular, the three major additions to *RADDOSE-3D* were:

(i) Implementation of a cylindrical sample geometry.

(ii) Determination of the sample composition given a mg ml^{-1} protein concentration.

(iii) Attenuation of the X-ray beam due to a surrounding capillary.

In isolation, the CorMap test performs pairwise comparisons of frames. This means that an individual frame can be detected as being dissimilar to another frame regardless of the similarity assessment of frames immediately before and after the frame in question. This dissimilarity of an individual frame relative to its neighbour therefore may not indicate a true systematic change in the molecules of the sample; rather, it may just be an outlier. Thus, a more robust indicator for radiation damage in this work was defined as the point whereby three consecutive frames were assessed as being dissimilar to the first frame.

The additions described above allowed the convenient visualization of the results from the original CorMap test, and thus enabled the comparison of the efficacy of eight different additive radioprotectant compounds for their ability to improve the dose tolerance of a protein sample (glucose isomerase, GI). It was established that some radioprotectant compounds exhibit a stronger concentration dependence than others in their ability to increase the dose tolerance of GI. Explicitly, at the lower concentrations, DTT was the most effective radioprotectant whereas, at the higher concentrations, glycerol was the most effective, resulting in more than a fivefold improvement in dose tolerance at 10 m*M* concentration.

Furthermore, the visualization library developed for this work was able to highlight why DTT performed poorly, and show that performing all pairwise comparisons could result in more consistent frame-merging methods.

Despite the extensions described above, the SAXS data collection model in *RADDOSE-3D* makes some implicit assumptions. For instance, with regard to the capillary, the atomic composition is assumed to be uniform throughout, and the thickness to be constant around the entire sample volume. The advantage of these assumptions is that the attenuation by

the capillary only needs to be calculated once, regardless of any movement or rotation of the capillary. This is valid for a cylindrical capillary since the thickness penetrated by the X-ray beam is the same regardless of any rotations or translations.

Additionally, the sample itself is assumed to be static, moving as a rigid body when rotated or translated, and also to be completely filling the capillary. This greatly reduces the computational cost when compared with the possibility of modelling more realistic fluid dynamics. The assumption that the capillary volume is completely filled is generally valid since this is usually the case during an experiment. The static assumption, on the other hand, is not always appropriate, especially when the sample is flowed through the capillary. Hopkins & Thorne (2016) calculated that for typical experimental parameters the velocity profile is expected to exhibit the quadratic Poiseuille flow profile, which arises from the axial symmetry and no-slip boundary assumptions (the velocity at the centre of the tube moves the fastest while the velocity at the boundary is equal to zero provided the capillary is also stationary). Hopkins & Thorne also calculated that the residence times of the sample in the beam were too short for any appreciable radial diffusive mixing, so that the flow profile results in radius-dependent residence times in the X-ray beam. Therefore, the static assumption made in *RADDOSE-3D* will give misleading dose values if calculated for experiments where the sample is flowed through the beam position.

Diffusive turnover is another phenomenon that will affect the dose calculation. Molecules have the ability to diffuse into and out of the illuminated volume, with the additional complexity that a non-uniform beam profile will cause differential diffusion across the beam due to higher sample heating at the peak of the beam profile. In the current work, no account was taken for molecular diffusion. However, beam sizes in SAXS experiments are typically quite large and exposure times are not long enough for the diffusive turnover to cause a significant effect. For example, Jeffries *et al.* (2015) used a 500 µm × 250 µm sized beam with a maximum exposure time of 141 s per sample while irradiating at 283 K (Jeffries *et al.*, 2015).

An important consideration regarding the concentration dependence of the efficacy of the additives is that they can alter the preferred environments of the sample. For example, it is known that DTT reduces disulfide bonds and undergoes oxidation. Glycerol at higher concentrations, on the other hand, is known to increase the noise and hence reduce the observable signal obtained from the sample (Jeffries *et al.*, 2015). In our study the reduction in the scattered intensity signal was examined for the different radioprotectants. Generally, the best performing radioprotectants, in terms of increasing the radiation tolerance of the sample, also reduced the scattered signal most. However, this order is not necessarily consistent, which suggests that the relationship is likely to be non-linear. Therefore, striking the optimal balance between increasing radiation tolerance and the level of signal reduction will still require some level of experience and input from the experimenter.

The definition of the signal reduction used in this study only considered the differences of the scattered intensity curve at the lowest q values where the difference was most prominent. However, the shape of the one-dimensional scattering curve can vary dramatically for different protein samples. Therefore, a metric for assessing the reduction that faithfully accounts for the difference across the entire q range may be more desirable.

Another important factor to consider is that the efficacy of radioprotectants may alter for different protein samples. This will likely be due to the various interaction processes that will occur for different macromolecules. For example, radioprotectants that performed best in this study could be involved in the prevention of the GI tetramer oligomerization breakdown. This process may not be as important for other samples, and hence the order of efficacy found in this study may not be applicable to a different sample.

During the data processing stage, the data were cropped by visual inspection to remove noisy sections. This is a subjective choice and therefore reduces the reproducibility of this work. To overcome this problem, quantitative criteria could be applied to find the optimum cut-off values for the data. Investigation of these possible criteria should be considered for further investigation of this method.

The $D_{\rm Thresh}$ values in our study range between 2.37 and 51.24 kGy. These differ significantly from the various threshold values in other types of diffraction experiments: room-temperature MX ∼150–500 kGy (Roedig *et al.*, 2015; Southworth-Davies *et al.*, 2007), cryo-crystallography ∼30 MGy (Owen *et al.*, 2006) and cryo-SAXS >3.7 MGy for µL samples and between 100–300 kGy for nL samples (Meisburger *et al.*, 2013). These differences are likely to be attributable to the differences in the experimental factors, *e.g.* temperature, sample type *etc.* However, there is still significant variation in the critical doses calculated for other room-temperature SAXS experiments [∼400 Gy (Kuwamoto *et al.*, 2004) and 284–7700 Gy (Jeffries *et al.*)] when compared with the $D_{\rm Thresh}$ values calculated in our study. Some of the variation can be attributed to the different protein samples and concentrations used in the experiments for the studies. However, the most likely cause of the apparent discrepancy is the various definitions of the critical dose. Hopkins & Thorne (2016) show that applying the critical dose definition from Jeffries *et al.* (2015) to their own GI data results in a critical dose of ∼66000 kGy, whereas the range in the Jeffries *et al.* (2015) study for GI was 5964–7056 Gy. Furthermore, they also state that 'the molecular weight shows a significant (13%) change after just ∼75 kGy'. Hence the critical dose value is highly dependent on the definition chosen to calculate it. Additionally, differences can also be expected due to the fact that the critical dose estimates from the previous studies are based upon (pseudo) radius of gyration values, whereas in our study the $D_{\rm Thresh}$ values are determined by analysis of frame similarity.

Although these issues should be taken into account for any given experiment to ensure optimum and successful data collection, the ability to predict the dose expected in a BioSAXS experiment and to rationally choose a suitable

additive to improve the starting conditions offers considerable benefits for data quality and efficiency of BioSAXS experiments. The methods and tools presented in this paper can be used in a complementary manner to other metrics created for assessing the quality of SAXS data (Grant *et al.*, 2015; Hopkins & Thorne, 2016).

The CorMap analysis visualization source code is freely available on Github: https://github.com/GarmanGroup/CorMapAnalysis.

Acknowledgements

We thank the ESRF for beam time under the auspices of the Radiation Damage BAG and the staff of beamline BM29, Martha Brennich and Petra Pernot, for their support. We gratefully acknowledge the UK Engineering and Physical Sciences Research Council for studentship funding in the Systems Biology Programme of the University of Oxford Doctoral Training Centre (JBB and CSB, grant number EP/G03706X/1), and studentship funding from the Wellcome Trust (HMG, studentship number 075491/04). We also thank the referees for their insightful comments.

References

Allan, E. G., Kander, M. C., Carmichael, I. & Garman, E. F. (2013). *J. Synchrotron Rad.* **20**, 23–36.
Bizien, T., Durand, D., Roblina, P., Thureau, A., Vachette, P. & Pérez, J. (2016). *Protein Pept. Lett.* **23**, 217–231.
Blanchet, C. E., Spilotros, A., Schwemmer, F., Graewert, M. A., Kikhney, A., Jeffries, C. M., Franke, D., Mark, D., Zengerle, R., Cipriani, F., Fiedler, S., Roessle, M. & Svergun, D. I. (2015). *J. Appl. Cryst.* **48**, 431–443.
Brennich, M. E., Kieffer, J., Bonamis, G., De Maria Antolinos, A., Hutin, S., Pernot, P. & Round, A. (2016). *J. Appl. Cryst.* **49**, 203–212.
Fischetti, R. F., Rodi, D. J., Mirza, A., Irving, T. C., Kondrashkina, E. & Makowski, L. (2003). *J. Synchrotron Rad.* **10**, 398–404.
Franke, D., Jeffries, C. M. & Svergun, D. I. (2015). *Nat. Methods*, **12**, 419–422.
Franke, D., Kikhney, A. G. & Svergun, D. I. (2012). *Nucl. Instrum. Methods Phys. Res. A*, **689**, 52–59.
Garman, E. (1999). *Acta Cryst.* D**55**, 1641–1653.
Garman, E. F. (2010). *Acta Cryst.* D**66**, 339–351.
Garrison, W. M. (1987). *Chem. Rev.* **87**, 381–398.
Graewert, M. A. & Svergun, D. I. (2013). *Curr. Opin. Struct. Biol.* **23**, 748–754.
Grant, T. D., Luft, J. R., Carter, L. G., Matsui, T., Weiss, T. M., Martel, A. & Snell, E. H. (2015). *Acta Cryst.* D**71**, 45–56.
Grishaev, A. (2012). *Curr. Protoc. Protein. Sci.* **70**, 17.14.1–17.14.18.
Hopkins, J. B., Katz, A. M., Meisburger, S. P., Warkentin, M. A., Thorne, R. E. & Pollack, L. (2015). *J. Appl. Cryst.* **48**, 227–237.
Hopkins, J. B. & Thorne, R. E. (2016). *J. Appl. Cryst.* **49**, 880–890.
Houée-Lévin, C., Bobrowski, K., Horakova, L., Karademir, B., Schöneich, C., Davies, M. J. & Spickett, C. M. (2015). *Free Radic. Res.* **49**, 347–373.
Hubbell, J. & Seltzer, S. (1995). Report NISTIR 5632. National Institute of Standards and Technology, Gaithersburg, MD, USA.
Jeffries, C. M., Graewert, M. A., Svergun, D. I. & Blanchet, C. E. (2015). *J. Synchrotron Rad.* **22**, 273–279.
Jones, E., Oliphant, T., Peterson, P. & others (2001). *SciPy: Open source scientific tools for Python*, http://www.scipy.org/.
Kozak, M. (2005). *J. Appl. Cryst.* **38**, 555–558.
Kuwamoto, S., Akiyama, S. & Fujisawa, T. (2004). *J. Synchrotron Rad.* **11**, 462–468.
Martel, A., Liu, P., Weiss, T. M., Niebuhr, M. & Tsuruta, H. (2012). *J. Synchrotron Rad.* **19**, 431–434.
Materlik, G., Rayment, T. & Stuart, D. I. (2015). *Philos. Trans. R. Soc. A*, **373**, 20130161.
Meisburger, S. P., Warkentin, M., Chen, H., Hopkins, J. B., Gillilan, R. E., Pollack, L. & Thorne, R. E. (2013). *Biophys. J.* **104**, 227–236.
Murray, J. W., Garman, E. F. & Ravelli, R. B. G. (2004). *J. Appl. Cryst.* **37**, 513–522.
Owen, R. L., Holton, J. M., Schulze-Briese, C. & Garman, E. F. (2009). *J. Synchrotron Rad.* **16**, 143–151.
Owen, R. L., Rudiño-Piñera, E. & Garman, E. F. (2006). *Proc. Natl Acad. Sci.* **103**, 4912–4917.
Pernot, P., Round, A., Barrett, R., De Maria Antolinos, A., Gobbo, A., Gordon, E., Huet, J., Kieffer, J., Lentini, M., Mattenet, M., Morawe, C., Mueller-Dieckmann, C., Ohlsson, S., Schmid, W., Surr, J., Theveneau, P., Zerrad, L. & McSweeney, S. (2013). *J. Synchrotron Rad.* **20**, 660–664.
Pernot, P., Theveneau, P., Giraud, T., Fernandes, R. N., Nurizzo, D., Spruce, D., Surr, J., McSweeney, S., Round, A., Felisaz, F., Foedinger, L., Gobbo, A., Huet, J., Villard, C. & Cipriani, F. (2010). *J. Phys. Conf. Ser.* **247**, 012009.
Petoukhov, M. V., Franke, D., Shkumatov, A. V., Tria, G., Kikhney, A. G., Gajda, M., Gorba, C., Mertens, H. D. T., Konarev, P. V. & Svergun, D. I. (2012). *J. Appl. Cryst.* **45**, 342–350.
Roedig, P., Vartiainen, I., Duman, R., Panneerselvam, S., Stübe, N., Lorbeer, O., Warmer, M., Sutton, G., Stuart, D. I., Weckert, E., David, C., Wagner, A. & Meents, A. (2015). *Sci. Rep.* **5**, 10451.
Round, A., Felisaz, F., Fodinger, L., Gobbo, A., Huet, J., Villard, C., Blanchet, C. E., Pernot, P., McSweeney, S., Roessle, M., Svergun, D. I. & Cipriani, F. (2015). *Acta Cryst.* D**71**, 67–75.
Southworth-Davies, R. J., Medina, M. A., Carmichael, I. & Garman, E. F. (2007). *Structure*, **15**, 1531–1541.
Spill, Y. G., Kim, S. J., Schneidman-Duhovny, D., Russel, D., Webb, B., Sali, A. & Nilges, M. (2014). *J. Synchrotron Rad.* **21**, 203–208.
Zeldin, O. B., Brockhauser, S., Bremridge, J., Holton, J. M. & Garman, E. F. (2013*b*). *Proc. Natl Acad. Sci.* **110**, 20551–20556.
Zeldin, O. B., Gerstel, M. & Garman, E. F. (2013*a*). *J. Appl. Cryst.* **46**, 1225–1230.

High-speed fixed-target serial virus crystallography

Philip Roedig[1,11], Helen M Ginn[2,3,11], Tim Pakendorf[1], Geoff Sutton[2], Karl Harlos[2], Thomas S Walter[2], Jan Meyer[1], Pontus Fischer[1], Ramona Duman[3], Ismo Vartiainen[4], Bernd Reime[1], Martin Warmer[1], Aaron S Brewster[5], Iris D Young[5], Tara Michels-Clark[5], Nicholas K Sauter[5], Abhay Kotecha[2], James Kelly[6,7], David J Rowlands[6], Marcin Sikorsky[8], Silke Nelson[8], Daniel S Damiani[8], Roberto Alonso-Mori[8], Jingshan Ren[2], Elizabeth E Fry[2], Christian David[9], David I Stuart[2,3], Armin Wagner[3] & Alke Meents[1,10]

© 2017 Nature America, Inc., part of Springer Nature. All rights reserved.

We report a method for serial X-ray crystallography at X-ray free-electron lasers (XFELs), which allows for full use of the current 120-Hz repetition rate of the Linear Coherent Light Source (LCLS). Using a micropatterned silicon chip in combination with the high-speed Roadrunner goniometer for sample delivery, we were able to determine the crystal structures of the picornavirus bovine enterovirus 2 (BEV2) and the cytoplasmic polyhedrosis virus type 18 polyhedrin, with total data collection times of less than 14 and 10 min, respectively. Our method requires only micrograms of sample and should therefore broaden the applicability of serial femtosecond crystallography to challenging projects for which only limited sample amounts are available. By synchronizing the sample exchange to the XFEL repetition rate, our method allows for most efficient use of the limited beam time available at XFELs and should enable a substantial increase in sample throughput at these facilities.

X-ray crystallography has been the dominant method for the determination of high-resolution virus structures in the past 30 years, although electron microscopy plays an increasing role. For nonenveloped viruses, numerous X-ray crystallographic structures have been solved, at resolutions of up to 1.4 Å (ref. 1). Owing to the large unit cell and limited sizes of the crystals, Bragg reflections from virus crystals are typically weak[2,3]. Thus, X-ray structure determination of virus crystals is ideally carried out at highly brilliant X-ray sources, thus allowing a large number of photons to be focused into a small spot ideally matching the size of the virus crystals[4]. A further challenge is radiation damage. Structure determination from biological macromolecules is usually carried out at cryogenic temperatures to decrease the detrimental effect of ionizing radiation on crystal diffraction[5–7]. Although finding appropriate conditions for cryoprotection is often straightforward for many protein crystals, doing so has remained a challenge for virus crystals, because they possess only weak crystal contacts, and a small increase in crystal mosaicity often results in overlapping reflections and a decrease in the measured resolution[2,3]. Thus, to date, relatively few virus structures have been determined at cryogenic temperatures[1], and most work is performed at room temperature[4].

A promising method for biological structure determination from virus crystals is serial femtosecond X-ray crystallography (SFX)[8–11] at XFELs, a technique that is well suited to room-temperature experiments and can overcome classical radiation dose limits by several orders of magnitude[12,13]. By taking up to hundreds of thousands of snapshots of nano- to micrometer-sized crystals, SFX has been used to solve more than 80 biomolecular structures to date. The method has also been successfully extended to the time domain to investigate protein kinetics and enzyme reactions[14,15]. SFX therefore not only has the potential to yield high-resolution information about the virus structure itself but also allows for investigation of dynamics, because many viral proteins undergo structural changes during their life cycles, for example as a result of protein interactions with cellular receptors and pH changes during entry through the endosome[16–18].

A continuing limiting factor for virus SFX in particular is efficient sample delivery, because typically only microgram amounts of virus crystals are available. Most SFX experiments today are carried out with gas dynamic virtual nozzles and high-viscosity gel matrices typically requiring more than 500 µg (~370 nL) of sample for structure determination[19–26].

In a first attempt of virus crystallography at XFELs[27], 10^4 microcrystals of the Sindbis virus with an ~700 Å unit-cell dimension were suspended in a slow-moving stream of agarose and exposed to XFEL pulses. Through this approach, diffraction patterns were obtained with up to ~40-Å resolution and a 0.8% hit rate, defined as the percentage of images containing a certain number of Bragg spots. Owing to the technological challenges, especially

[1]Photon Science, Deutsches Elektronen-Synchrotron DESY, Hamburg, Germany. [2]Division of Structural Biology, Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK. [3]Diamond Light Source Limited, Didcot, UK. [4]Institute of Photonics, University of Eastern Finland, Joensuu, Finland. [5]Molecular Biophysics and Integrated Bioimaging Division, Lawrence Berkeley National Laboratory, Berkeley, California, USA. [6]School of Molecular and Cellular Biology, Faculty of Biological Sciences, University of Leeds, Leeds, UK. [7]Pirbright Institute, Pirbright, UK. [8]Linac Coherent Light Source, SLAC National Accelerator Laboratory, Menlo Park, California, USA. [9]Paul Scherrer Institut, Villigen-PSI, Switzerland. [10]Center for Free Electron Laser Science (CFEL), Hamburg, Germany. [11]These authors contributed equally to this work. Correspondence should be addressed to A.M. (alke.meents@desy.de).

RECEIVED 8 DECEMBER 2016; ACCEPTED 15 MAY 2017; PUBLISHED ONLINE 19 JUNE 2017 ; DOI:10.1038/NMETH.4335

the amount of sample required, to our knowledge, no previous virus structure has been determined at an XFEL.

A promising sample-delivery approach for SFX is the use of solid sample supports, also referred to as fixed targets[28–37], especially if the sample is limited. Here, several tens to thousands of crystals are loaded onto a structured solid support and are automatically raster-scanned during X-ray exposure. Major challenges for fixed-target experiments are the fast and precise scanning of these supports and the synchronization of predefined sample positions with the arrival of XFEL pulses.

Such experiments have been performed in step-scanning mode, which is less efficient than other sample-delivery methods, owing to the relatively long data collection times[34,38]. Another obstacle for fixed-target experiments is the relatively high background scattering level, which is mainly caused by air scattering from the direct beam. Further contributions originate from scattering by nonsample material such as the surrounding mother liquor or the Kapton or Mylar foil typically used as a sealing material to prevent crystal dehydration.

Here, we report a method for fixed-target serial crystallography at low background levels and a sample exchange rate of 120 Hz, which requires only micrograms of sample. We demonstrate the applicability of our method for virus structure determination, shown for an intact virus, BEV2, and cytoplasmic polyhedrosis virus type 18 polyhedrin (CPV18).

BEV2 belongs to the virus family Picornaviridae, genus *Enterovirus*, and is a nonenveloped, positive-stranded RNA virus of 30 nm in diameter, which is endemic in cattle environments[39–41]. Unlike some other picornaviruses, it is not a severe economic or animal-health threat and is therefore a suitable model system for the investigation of biological processes such as virus uncoating, which are universal to all enteroviruses. Enteroviruses (including BEV2) are stabilized by lipid cofactors such as sphingosine, which bind to a hydrophobic pocket of the VP1 capsid protein[42]. Potent binders to this pocket have potential as antivirals (by preventing uncoating); hence, we investigated the specificity of binding natural and synthetic moieties in this pocket. BEV2 crystals represent a challenging system for current SFX experiments, possessing a large unit-cell constant of 436.6 Å; moreover, typically only microgram quantities of small crystals are available.

Cytoplasmic polyhedrosis viruses (CPVs) are found as parasites in many insects and cause substantial losses in silkworm cocoon harvests[43]. CPV strains have polyhedral crystals varying in size from hundreds of nanometers to several micrometers and contain several thousands of CPV particles[43–46], thus providing protection against environmental insults. CPV polyhedra were chosen as a well-established and robust model system for SFX data collection at cryogenic temperatures[47].

RESULTS

High-speed fixed-target structure-determination approach

Microcrystals of BEV2 and CPV18 were measured by fixed-target SFX with the Roadrunner goniometer installed at the XPP instrument at the LCLS (**Supplementary Fig. 1**). The Roadrunner setup consists of high-precision *x* and *y* piezomotor-driven scanning stages mounted on a horizontal translation stage and a vertical rotation axis (**Supplementary Fig. 2**; details in Online Methods). A high-magnification inline microscope allows for visualization of samples and their support structure (**Supplementary Fig. 3**).

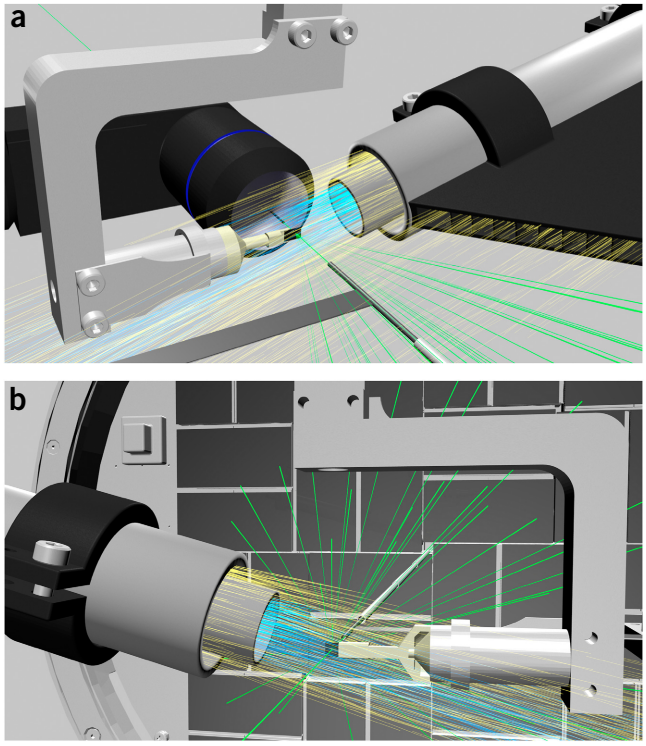


Figure 1 | Low-background experimental setup for fast fixed-target SFX experiments using the Roadrunner goniometer. (**a**) Front view. The silicon chip is raster-scanned through the X-ray beam (green) while being maintained under a continuous stream of humidified air (blue). A helium sheath flow (yellow) is used to confine the humidity stream and to decrease air scattering. Air scattering is further decreased by helium injection along the beam path. An inline microscope is used for proper chip alignment and definition of the scanning grid. (**b**) Back view. X-ray diffraction caused by the sample crystals is recorded with a Cornell-SLAC hybrid pixel array detector (CSPAD). After hitting the sample, the primary beam is enclosed by a molybdenum tubule and additional steel tubules, which further absorb air-scattered photons. In **b**, the inline microscope is omitted for clarity.

The Roadrunner setup is capable of data collection at both room temperature and cryogenic temperatures.

To decrease air scattering of X-rays, most of the path of the direct beam in air is enclosed in capillary shields, both upstream and downstream of the sample, thus decreasing the free path of the direct beam in air to 20 mm (**Supplementary Fig. 4**). By streaming helium gas across the remaining unenclosed direct beam behind the sample, the number of photons scattered by air is further decreased (**Fig. 1**). By combining these two approaches, air scattering can be decreased by a factor of ~8.

By use of micropatterned chips made of single-crystalline silicon as a substrate material, the background scattering signal caused by the support can be further decreased[31]. Dehydration of the crystals is prevented either by keeping them constantly under a stream of humidified gas[32,48] or by flash freezing them and collecting data at cryogenic temperatures[31] (**Supplementary Fig. 5**). With this approach, sealing of the sample holders, for example with Mylar foil, which would result in an increased background scattering level, is not required.

Our silicon chip (**Fig. 2a**) provides 22,500 pores for crystals[31]. For sample loading, 2–3 µL of sample suspension is pipetted onto

© 2017 Nature America, Inc., part of Springer Nature. All rights reserved.

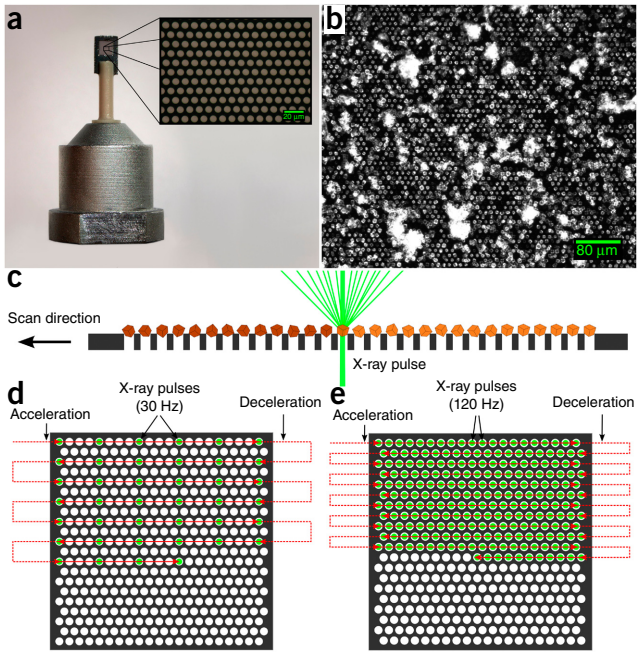


Figure 2 | Design of the micropatterned silicon chip and data collection strategy. (**a**) The chip is attached to a plastic rod for the purpose of thermal isolation. The membrane part within the outer frame consists of micropores with diameters of typically 4–8 µm, which are arranged in a triangular grid (inset). (**b**) The chip acts as a sample holder for more than 20,000 microcrystals, which largely organize themselves according to the pore pattern. (**c**) After loading, the microcrystals are scanned through the X-ray beam. By shooting the beam through the micropores in the chips, the interaction of the X-rays with any support material is further minimized. (**d**,**e**) Scanning strategies for measurements performed at room temperature (**d**) and cryogenic temperatures (**e**) (details in text).

the chip, so that the amount of sample material used is typically in the range of several micrograms, depending on the crystal sizes, the desired coverage of the chip membrane, and the amount of sample available[31,32]. The resulting arrangement of the crystals is due to the pore pattern and allows for a highly efficient measurement strategy by shooting the free-electron laser (FEL) pulses through all pores in a fully automated procedure (**Fig. 2b**,**c**). With the Roadrunner control software, the scan points are defined by drawing a grid, which is graphically overlaid on the inline-microscope image (**Supplementary Fig. 6**). For data collection, the coordinates of scan points are downloaded to the motion controller, and the entire chip is scanned in a meander-scan manner (**Fig. 2d**,**e**; details in Online Methods). With this approach, most of the crystalline material is used for the diffraction experiment and is not wasted.

To achieve fixed-target data collection rates of 120 Hz, we developed an improved version of the so-called fly scan. In a conventional fly scan, the sample is accelerated and then moved along a predefined trajectory at a constant velocity. This approach would in principle allow for data collection at 120 Hz, but in most cases it results in lower hit rates than those obtained with our method, because the X-ray pulses do not always hit through the pores where the crystals are located. To achieve higher hit rates—and thereby require substantially less sample—more precise motion control is required. Such an approach demands, in addition to velocity control, phase control of the movement of the stages with respect to the arrival of the X-ray pulses. The synchronized movement assures that every X-ray pulse hits through a pore (**Fig. 2c**–**e**, Online Methods and **Supplementary Fig. 7**).

BEV2 and CPV18 data collection

In the case of BEV2 samples, which were measured at room temperature, diffraction images were collected at an FEL pulse rate of 30 Hz per line, thus resulting in an effective frame rate of up to 12.2 images/s for each chip, when taking the time for line switching into account (**Table 1** and **Supplementary Table 1**). A maximum hit rate of more than 9% for one chip and 5% on average was achieved for BEV2. The relatively low hit rates in this case were because the density of the crystals on the chip was low. Because only limited amounts of sample were available, we aimed at making the most efficient use of the available sample instead of optimizing hit rates. For room-temperature data collection, it was not possible to run at the full LCLS frame rate of 120 Hz, because crystals in the neighboring compartments were already predamaged by the wings of the X-ray beam. When shooting every fourth hole (40-µm separation) in the horizontal direction and every second row in the vertical direction (20-µm separation), we observed no effects of predamage (**Fig. 2d**). With a maximum speed of the horizontal scanning stage of 2.5 mm/s, data collection at 60 Hz with 40 µm separation or at 120 Hz with 20 µm separation would also have been possible.

For CPV18, data collection was performed at cryogenic temperatures with the full LCLS repetition rate of 120 Hz, thus resulting in an average data collection rate of 33.6 images/s (**Fig. 2e**). From these images, more than 70% were classified as hits. In other runs, we were able to achieve hit rates of more than 90%. No predamage of the neighboring crystals by the wings of the X-ray beam was observed for CPV18, probably because of decreased diffusion rates of free radicals and the resulting higher radiation tolerance of macromolecular crystals at cryogenic temperatures[5–7].

Table 1 | Data collection parameters and hit rates for individual runs

Run number	Sample	Temperature Conditions	Number of images	Acquisition time (s)	Effective acquisition rate (images/s)	Number of hits[a]	Hit rate (%)	Indexed
294	BEV2	293 K	2,280	230.7	9.9	159	7.0	124
296	BEV2	293 K	180	14.8	12.2	17	9.4	16
298	BEV2	293 K	1,950	178.0	11.0	76	3.9	69
301	BEV2	293 K	2,015	180.7	11.2	150	7.4	113
303	BEV2	293 K	2,387	214.4	11.1	44	1.8	30
47	CPV18	100 K	19,028	566.1	33.6	13,424	70.5	16,739

[a]Images containing ≥50 (BEV2) and 20 (CPV18) strong spots were considered hits. Spot-finding parameters can be found in **Supplementary Table 1**.

© 2017 Nature America, Inc., part of Springer Nature. All rights reserved.

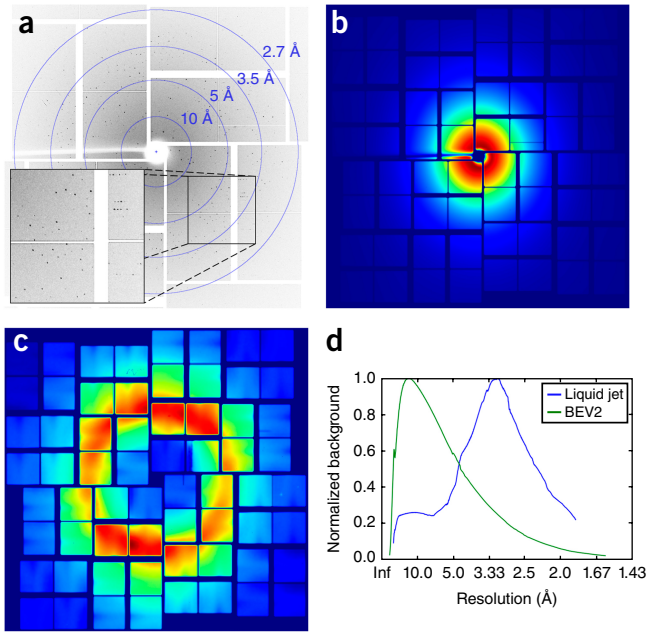


Figure 3 | Exemplary BEV2 diffraction pattern and comparison of background scattering levels achievable with different sample-delivery methods. (**a**) Diffraction image of BEV2 microcrystals obtained at the XPP instrument at LCLS by using the micropatterned silicon chip as a sample holder. (**b**) Owing to the efficient removal of successive mother liquor during sample loading, no water ring is observed in the averaged background image of the chip. (**c**) Averaged background image from a typical SFX liquid jet experiment with CPV17 crystals[47], shown for comparison. Inf, infinite. (**d**) The azimuthally averaged radial distribution of both images is plotted as a function of resolution. Both curves are normalized because measurements were performed under different experimental conditions, and therefore a direct comparison was not possible.

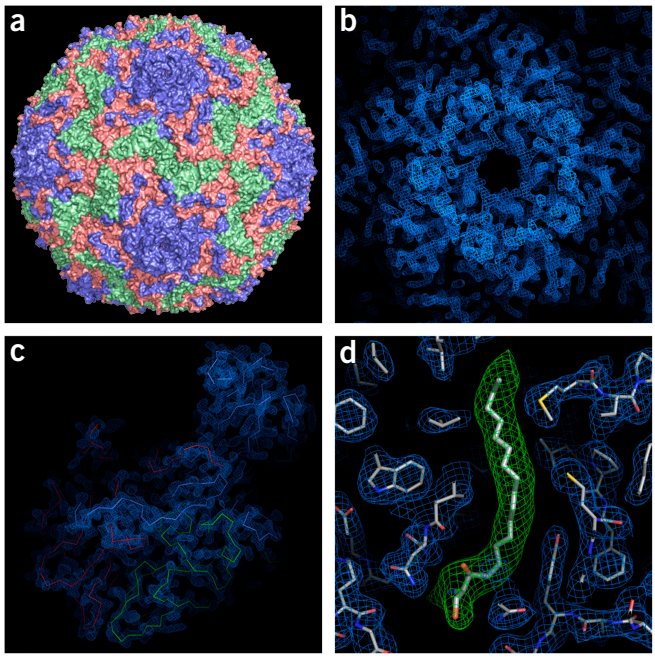


Figure 4 | Overall structure of BEV2 and corresponding high-resolution electron density maps. (**a**) Surface representation of the BEV2 particle, as viewed toward an icosahedral two-fold axis. VP1, VP2, and VP3 are shown in blue, green, and red, respectively. (**b**,**c**) Electron density maps after one cycle of five-fold real-space averaging by using the phases calculated from the current refined model showing the electron density around the five-fold axis in **b** and for a biological protomer in **c**. (**c**) Cα traces of VP1–3, colored as in **a**. (**d**) A close-up view of the electron density for protein residues around the pocket-factor-binding site of VP1 (blue mesh and thinner sticks) and density for the pocket factor (thicker sticks show a sphingosine fitted to the density, and the green density is for a simulated annealing omit map).

© 2017 Nature America, Inc., part of Springer Nature. All rights reserved.

Image quality and background analysis

The diffraction patterns extended to a resolution of 2.3 Å for BEV2 and 2.4 Å for CPV18. An example diffraction image (**Fig. 3a**) obtained from a BEV2 crystal illustrates the high quality of the diffraction patterns obtained with the Roadrunner goniometer. The low background contribution of the chip resulted in high-resolution diffraction data with high signal-to-noise ratios. We analyzed the averaged background signal of the measured diffraction images (**Fig. 3b**) and compared it with that from an SFX experiment in which a liquid jet was used for sample delivery (**Fig. 3c**). The azimuthally averaged background signal as a function of resolution clearly showed that in the diffraction images measured with our fixed-target setup, the background was dominated by air scattering, which was most prominent at resolutions below 10 Å (**Fig. 3d**). The chip itself consisted of single-crystalline silicon and therefore did not contribute to any background signal. The absence of a water ring for room-temperature data collection revealed the efficient removal of mother liquor during sample loading. In typical liquid jet experiments, the averaged background signal showed a strong water ring around 3.1 Å.

Structure determination

Summary information regarding data collection and structure refinement for both samples is given in **Supplementary Table 2** (PDB 5MQU for BEV2 and PDB 5MQW for CPV18). The structure of BEV2 was solved by using diffraction data obtained from five chips, from a total of ~446 crystal hits in less than 14 min of scanning time. For CPV18, the complete structure was able to be solved with the data obtained from only a single chip, collected in less than 10 min. Unfortunately, the resolution was limited by the dimensions of the detector for the given detector distance in that case. Structure refinement (on the basis of prior models) yielded high-quality electron density maps (**Fig. 4**) with model R values of R_{work}/R_{free} = 23.3/25.7% for BEV2 (to 2.3-Å resolution; data were measureable to 2.0-Å resolution, but the resulting map was only marginally improved, and the statistics were significantly worse) and 11.3/14.5% for CPV18 (to 2.4-Å resolution).

The CPV18 data were measured with high redundancy (>100-fold) and were of very high quality (R_{split} 9.2%, $CC_{1/2}$ 0.993), thus supporting the excellent refinement. In contrast, the BEV2 data were derived from only 324 crystals and had a multiplicity of only 2, and the merging statistics were correspondingly poor (R_{split} 48.6%, $CC_{1/2}$ 0.746). We therefore performed two tests to determine whether the amplitudes contained sufficient information to determine the high-resolution structure in the absence of accurate phase information. First, we extended the phases by using five-fold noncrystallographic symmetry (NCS) averaging and solvent flattening from 5-Å to 2.5-Å resolution. The result was an excellent map with largely successful phase

recovery (details in Online Methods and **Supplementary Fig. 8**). Second, we performed molecular replacement starting from a distantly related virus (FMDV type A22, sequence identity 19.5%). As expected, the initial map showed substantial bias; however, this bias was eliminated by cyclic averaging, thus again resulting in a high-quality electron density map (Online Methods and **Supplementary Figs. 9** and **10**).

The BEV2 structure was determined as part of an investigation of the specificity of the hydrophobic pocket in VP1 for different fatty acids. The structure revealed that despite the cocrystallization of the virus with lauric acid, the pocket factor present in the particles was indistinguishable from that observed in native particles and was well modeled as sphingosine (**Fig. 4d**), thus explaining the biophysical observations that lauric acid has essentially no effect on the stability of the virus particles.

The CPV18 structure is similar to recently published structures of isolated crystals of CPV18 (refs. 31,49). A part of the electron density map is shown in **Supplementary Figures 11** and **12**.

DISCUSSION

To our knowledge, this report describes the first structure of a virus particle determined at an XFEL by means of serial crystallography. The electron density maps obtained for BEV2 provide a high level of detail sufficient to demonstrate that lauric acid cannot displace sphingosine from the VP1 pocket, which is the major target for the design of antienteroviral compounds. We further showed that the method is also applicable to data collection at cryogenic temperatures, at which hit rates of more than 70% were achieved, and the structure of CPV18 crystals was solved from the measurement of one chip loaded with ~4 µg of protein.

The periodic arrangement of the crystals on our chip in combination with the Roadrunner goniometer allows for very effective use of beam time. With data collection rates of 120 Hz during a line scan, combined with hit rates of more than 70%, we were able to obtain up to 29.6 indexable diffraction patterns per second. Sample loading onto the chip is very efficient, and no precious crystalline material is lost. The method is probably more reliable than liquid jet experiments, which often have drawbacks such as nozzle clogging and crystal settling, which lead to substantial downtime during the experiments.

A further benefit of our method is its ultralow sample consumption, which requires orders of magnitude less sample than that required in current liquid jet methods at XFELs and room-temperature experiments at synchrotrons. A synchrotron structure[4] of the apo form of the BEV2 capsid has previously been determined at 2.1-Å resolution, on the basis of the measurement of 28 crystals of a cubic edge length of ~50 µm, thus amounting to a total crystal volume of 3.5 nL. Our work was based on data collected from 446 much smaller crystals with a cubic edge length of only ~8 µm, corresponding to a total volume of 228 pL. Cocrystallization of BEV2 with lauric acid limits the achievable crystal size and renders these crystals far too small for conventional synchrotron structure determination. The use of the XFEL therefore not only provided the high intensity X-ray pulses required to generate sufficiently strong diffraction to solve the ligand-bound virus structure but also decreased the total sample amount used for structure-factor calculation by 15-fold. Notably, high-quality phases and hence electron density maps were able to be derived from amplitudes assembled from such little material and such a low-multiplicity XFEL data set assembled from only 324 crystals. We attribute this ability to the high quality of the data obtained from this experimental setup, to advances in data processing methods, and in part to the five-fold noncrystallographic symmetry.

In the current setup, the X-ray-scattering background is dominated by air scattering from the short remaining beam path in humidified air or cold nitrogen gas. By further decreasing the path of the primary beam in air and by replacing air or nitrogen with helium, we aim to significantly decrease the background level in future experiments to achieve higher-resolution data from even smaller crystals. A larger chip design with up to 200,000 micropores in combination with faster scanning stages would allow for longer data collection runs at frame rates of up to 1 kHz, thereby resulting in even more efficient use of beam time at both XFELs and synchrotron facilities.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank the LCLS staff for their support during experiment LH90. Use of the LCLS, SLAC National Accelerator Laboratory, is supported by the US Department of Energy, Office of Science, Office of Basic Energy Sciences under contract no. DE-AC02-76SF00515. This work was further supported by the European Cluster of Advanced Laser Light Sources (EUCALL) and the Virtual Institute VH-VI-403 of the Helmholtz Association. H.M.G. was supported by the Wellcome Trust (studentship 075491/04). D.I.S. was supported by the Medical Research Council, grant MR/N00065X/1 and previously by G1000099. N.K.S. was supported by the US National Institutes of Health grants GM102520 and GM117126 for data processing methods. We further thank J. Bergtholdt from DESY for preparation of the technical drawings of the Roadrunner setup.

AUTHOR CONTRIBUTIONS

P.R., T.P., P.F., J.M., A.W., and A.M. designed the experiment. P.R., G.S., K.H., T.S.W., R.D., I.V., M.W., A.K., J.K., and D.J.R. were involved in sample preparation. P.R., T.P., G.S., K.H., J.M., P.F., R.D., B.R., M.S., S.N., D.S.D., R.A.-M., C.D., A.W., and A.M. participated in data collection. P.R., H.M.G., A.S.B., I.D.Y., T.M.-C., N.K.S., J.R., E.E.F., and D.I.S. analyzed the data. P.R., H.M.G., D.I.S., A.W., and A.M. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Zocher, G. *et al.* A sialic acid binding site in a human picornavirus. *PLoS Pathog.* **10**, e1004401 (2014).
2. Fry, E.E., Abrescia, N.G.A. & Stuart, D.I. in *Macromolecular Crystallography: Conventional and High-Throughput Methods* (eds. Sanderson, M.R. & Skelly, J.V.), 245–264 (Oxford, 2007).
3. Fry, E.E., Grimes, J. & Stuart, D.I. Virus crystallography. *Mol. Biotechnol.* **12**, 13–23 (1999).
4. Axford, D. *et al.* In situ macromolecular crystallography using microbeams. *Acta Crystallogr. D Biol. Crystallogr.* **68**, 592–600 (2012).
5. Hope, H. Cryocrystallography of biological macromolecules: a generally applicable method. *Acta Crystallogr. B* **44**, 22–26 (1988).
6. Owen, R.L., Rudiño-Piñera, E. & Garman, E.F. Experimental determination of the radiation dose limit for cryocooled protein crystals. *Proc. Natl. Acad. Sci. USA* **103**, 4912–4917 (2006).

© 2017 Nature America, Inc., part of Springer Nature. All rights reserved.

7. Meents, A., Gutmann, S., Wagner, A. & Schulze-Briese, C. Origin and temperature dependence of radiation damage in biological samples at cryogenic temperatures. *Proc. Natl. Acad. Sci. USA* **107**, 1094–1099 (2010).
8. Chapman, H.N. *et al.* Femtosecond X-ray protein nanocrystallography. *Nature* **470**, 73–77 (2011).
9. Boutet, S. *et al.* High-resolution protein structure determination by serial femtosecond crystallography. *Science* **337**, 362–364 (2012).
10. Redecke, L. *et al.* Natively inhibited *Trypanosoma brucei* cathepsin B structure determined by using an X-ray laser. *Science* **339**, 227–230 (2013).
11. Ayyer, K. *et al.* Macromolecular diffractive imaging using imperfect crystals. *Nature* **530**, 202–206 (2016).
12. Neutze, R., Wouts, R., van der Spoel, D., Weckert, E. & Hajdu, J. Potential for biomolecular imaging with femtosecond X-ray pulses. *Nature* **406**, 752–757 (2000).
13. Barty, A. *et al.* Self-terminating diffraction gates femtosecond X-ray nanocrystallography measurements. *Nat. Photonics* **6**, 35–40 (2012).
14. Kupitz, C. *et al.* Serial time-resolved crystallography of photosystem II using a femtosecond X-ray laser. *Nature* **513**, 261–265 (2014).
15. Tenboer, J. *et al.* Time-resolved serial crystallography captures high-resolution intermediates of photoactive yellow protein. *Science* **346**, 1242–1246 (2014).
16. Perera, R., Khaliq, M. & Kuhn, R.J. Closing the door on flaviviruses: entry as a target for antiviral drug design. *Antiviral Res.* **80**, 11–22 (2008).
17. Connolly, S.A., Jackson, J.O., Jardetzky, T.S. & Longnecker, R. Fusing structure and function: a structural view of the herpesvirus entry machinery. *Nat. Rev. Microbiol.* **9**, 369–381 (2011).
18. Harrison, J.S. *et al.* Role of electrostatic repulsion in controlling pH-dependent conformational changes of viral fusion proteins. *Structure* **21**, 1085–1096 (2013).
19. Weierstall, U. *et al.* Lipidic cubic phase injector facilitates membrane protein serial femtosecond crystallography. *Nat. Commun.* **5**, 3309 (2014).
20. DePonte, D.P. *et al.* Gas dynamic virtual nozzle for generation of microscopic droplet streams. *J. Phys. D Appl. Phys.* **41**, 195505 (2008).
21. Caffrey, M. & Cherezov, V. Crystallizing membrane proteins using lipidic mesophases. *Nat. Protoc.* **4**, 706–731 (2009).
22. Liu, W. *et al.* Serial femtosecond crystallography of G protein-coupled receptors. *Science* **342**, 1521–1524 (2013).
23. Fromme, R. *et al.* Serial femtosecond crystallography of soluble proteins in lipidic cubic phase. *IUCrJ* **2**, 545–551 (2015).
24. Sugahara, M. *et al.* Grease matrix as a versatile carrier of proteins for serial crystallography. *Nat. Methods* **12**, 61–63 (2015).
25. Sugahara, M. *et al.* Oil-free hyaluronic acid matrix for serial femtosecond crystallography. *Sci. Rep.* **6**, 24484 (2016).
26. Conrad, C.E. *et al.* A novel inert crystal delivery medium for serial femtosecond crystallography. *IUCrJ* **2**, 421–430 (2015).
27. Lawrence, R.M. *et al.* Serial femtosecond X-ray diffraction of enveloped virus microcrystals. *Struct. Dyn.* **2**, 041720 (2015).
28. Zarrine-Afsar, A. *et al.* Crystallography on a chip. *Acta Crystallogr. D Biol. Crystallogr.* **68**, 321–323 (2012).
29. Hunter, M.S. *et al.* Fixed-target protein serial microcrystallography with an x-ray free electron laser. *Sci. Rep.* **4**, 6026 (2014).
30. Mueller, C. *et al.* Fixed target matrix for femtosecond time-resolved and in situ serial micro-crystallography. *Struct. Dyn.* **2**, 054302 (2015).
31. Roedig, P. *et al.* A micro-patterned silicon chip as sample holder for macromolecular crystallography experiments with minimal background scattering. *Sci. Rep.* **5**, 10451 (2015).
32. Roedig, P. *et al.* Room-temperature macromolecular crystallography using a micro-patterned silicon chip with minimal background scattering. *J. Appl. Cryst.* **49**, 968–975 (2016).
33. Oghbaey, S. *et al.* Fixed target combined with spectral mapping: approaching 100% hit rates for serial crystallography. *Acta Crystallogr. D Struct. Biol.* **72**, 944–955 (2016).
34. Cohen, A.E. *et al.* Goniometer-based femtosecond crystallography with X-ray free electron lasers. *Proc. Natl. Acad. Sci. USA* **111**, 17122–17127 (2014).
35. Murray, T.D. *et al.* A high-transparency, micro-patternable chip for X-ray diffraction analysis of microcrystals under native growth conditions. *Acta Crystallogr. D Biol. Crystallogr.* **71**, 1987–1997 (2015).
36. Baxter, E.L. *et al.* High-density grids for efficient data collection from multiple crystals. *Acta Crystallogr. D Struct. Biol.* **72**, 2–11 (2016).
37. Owen, R.L. *et al.* Low-dose fixed-target serial synchrotron crystallography. *Acta Crystallogr. D Struct. Biol.* **73**, 373–378 (2017).
38. Sherrell, D.A. *et al.* A modular and compact portable mini-endstation for high-precision, high-speed fixed target serial crystallography at FEL and synchrotron sources. *J. Synchrotron Radiat.* **22**, 1372–1378 (2015).
39. Smyth, M. *et al.* Preliminary crystallographic analysis of bovine enterovirus. *J. Mol. Biol.* **231**, 930–932 (1993).
40. Smyth, M. *et al.* Implications for viral uncoating from the structure of bovine enterovirus. *Nat. Struct. Biol.* **2**, 224–231 (1995).
41. Goens, S.D., Botero, S., Zemla, A., Zhou, C.E. & Perdue, M.L. Bovine enterovirus 2: complete genomic sequence and molecular modelling of a reference strain and a wild-type isolate from endemically infected US cattle. *J. Gen. Virol.* **85**, 3195–3203 (2004).
42. Filman, D.J. *et al.* Structural factors that control conformational transitions and serotype specificity in type 3 poliovirus. *EMBO J.* **8**, 1567–1579 (1989).
43. Coulibaly, F. *et al.* The molecular organization of cypovirus polyhedra. *Nature* **446**, 97–101 (2007).
44. Ji, X. *et al.* How baculovirus polyhedra fit square pegs into round holes to robustly package viruses. *EMBO J.* **29**, 505–514 (2010).
45. Anduleit, K. *et al.* Crystal lattice as biological phenotype for insect viruses. *Protein Sci.* **14**, 2741–2743 (2005).
46. Ji, X. *et al.* Polyhedra structures and the evolution of the insect viruses. *J. Struct. Biol.* **192**, 88–99 (2015).
47. Ginn, H.M. *et al.* Structure of CPV17 polyhedrin determined by the improved analysis of serial femtosecond crystallographic data. *Nat. Commun.* **6**, 6435 (2015).
48. Sanchez-Weatherby, J. *et al.* Improving diffraction by humidity control: a novel device compatible with X-ray beamlines. *Acta Crystallogr. D Biol. Crystallogr.* **65**, 1237–1246 (2009).
49. Axford, D., Ji, X., Stuart, D.I. & Sutton, G. In cellulo structure determination of a novel cypovirus polyhedrin. *Acta Crystallogr. D Biol. Crystallogr.* **70**, 1435–1441 (2014).

© 2017 Nature America, Inc., part of Springer Nature. All rights reserved.

ONLINE METHODS

A detailed step-by-step protocol of our method is provided as a **Supplementary Protocol** and is also available on the Protocol Exchange from Nature Protocols[50].

Sample preparation. Bovine enterovirus 2 light fraction (BEV2) was prepared at the Department of Microbiology, University of Leeds, UK, by the group of D.J.R. Particles were purified by ultracentrifugation over a 30% sucrose cushion followed by a 15–45% sucrose gradient. Fractions containing BEV2 particles were pooled, incubated with lauric acid, and pelleted by ultracentrifugation by J.K. To remove excess lauric acid, particles were solubilized in 20 mM HEPES, pH 8.0, 200 mM NaCl, and 0.5% NP-40 at 4 °C, and pelleted again by ultracentrifugation with a 30% sucrose cushion. The pellet was solubilized in 20 mM HEPES, pH 8.0, and 200 mM NaCl. Crystals were grown from the purified concentrated particles by vapor diffusion in Greiner 'CrystalQuick X' plates at 293 K. Drops of 100 nL of BEV2 particles (A_{280} = 5) were mixed with 100-nL drops[51] of the precipitant (1.5 M ammonium sulfate and 0.1 M Tris, pH 8) and equilibrated against 35 µL of precipitant. More than 100 small crystals, with typical dimensions of $8 \times 8 \times 8$ µm^3, appeared in each drop within hours. An image of the BEV2 crystals is shown in **Supplementary Figure 13**. CPV18 polyhedrin crystals were prepared as described previously[31].

Chip design and fabrication. The chip design is illustrated in **Figure 2**. The chips were made from single-crystalline silicon by UV lithography and had overall dimensions of 2.5×4 mm^2 with a thickness of 0.1 mm. The inner-membrane part with an area of 1.5×1.5 mm^2 was thinned down to a thickness of 10 µm and provided a hexagonal dense pattern of pores with diameters between 4 µm and 8 µm and a 10-µm periodicity (**Fig. 2a**). The chips were glued to plastic pins, which could be mounted on conventional magnetic caps routinely used in macromolecular crystallography. More details about chip design, fabrication, loading, and handling can be found in ref. 31.

Preorientation of the chips. With the extremely high X-ray intensity per FEL pulse, Bragg reflections arising from the silicon chip material can easily damage the detector. Hence, it is essential to know the exact angular orientation of the chips with respect to the incident X-ray beam to avoid these Bragg reflections. As a reference mark, the magnetic caps carrying the chips were modified by removing some material at the lower rim of the caps, as shown in **Figure 2a**. All chips glued to the plastic pins were then oriented and fixed such that the chip surface was always parallel to the face of the magnetic caps.

Sample loading. Sample loading was performed by applying 2–3 µL of crystal suspension to the upper side of the chip. Additional mother liquor was then removed by soaking with a wedge of filter paper attached to the lower side of the chip (details in ref. 31). The chip allowed for data collection at both room temperature (BEV2) and cryogenic temperatures (CPV18). For room-temperature data collection, the samples were loaded onto the chips directly at the experimental setup. Similarly to recently performed synchrotron experiments[32], a humidified gas stream with adjustable relative humidity was used to prevent the crystals from drying out during loading and data collection[48,52]. For BEV2 data collection, the relative humidity was set to 96%. The experimental setup could be also used for data collection at cryogenic temperatures. The major difference was simply the replacement of the humidity stream used for room-temperature data collection by a cold nitrogen-gas stream. In that case, prefrozen samples on the chips were directly mounted on the goniometer.

Roadrunner goniometer. For the experiment, we designed a special goniometer whose main part was a fast piezomotor-driven *x*,*y* translation stage for fast raster-scanning of the chips carrying the samples. A technical overview drawing of the Roadrunner goniometer is provided in **Supplementary Figure 1**. The setup consisted of three major components—a high-resolution inline sample-viewing microscope, the high-precision goniometer itself, and a post sample beam pipe unit—all mounted on a common support frame structure. With outer dimensions of 250 mm along the beam direction, a width of 400 mm and a height of 515 mm, the entire setup is compact and could be therefore easily installed at different experimental endstations such as XPP (as presented here), the new MFX endstation at LCLS, or other X-ray sources.

The first element in the X-ray beam path, the inline sample-viewing microscope, is shown and described in more detail in **Supplementary Figure 3**. The microscope provides a high-resolution image of the samples mounted on the goniometer and is used for precise alignment of the chips with respect to the X-ray beam. The X-ray beam passes through a molybdenum collimator tube inserted into the hole of the objective lens with an inner diameter of 0.35 mm. The capillary is used to prevent X-ray damage to the microscope lenses, and it extends to only 3 mm from the sample position to decrease air scattering along the beam path.

The micropatterned silicon chip carrying the samples is mounted on a high-precision goniometer axis. A technical drawing of the goniometer and a detailed description of its functionality are provided in **Supplementary Figure 2**. The main element is the high *x*,*y* precision stage for scanning of the chips synchronized to the time structure of the X-ray pulses. The *x*,*y* scanning stage was controlled by a DMC-4080 motion controller from Galil. The motion controller was capable of synchronizing the two axes of the scanning stage to the repetition rate of the LCLS beam at 120 Hz. The synchronization was done on a line-by-line basis to ensure that each X-ray pulse hit the center of the holes in the silicon chip. For each line, the motion controller was provided with the starting point, the number of scan points (number of holes), the angular orientation of the line, and the repetition rate of the LCLS beam. The synchronization scheme is illustrated and described in more detail in **Supplementary Figure 7**. After the start, the controller moves the scanning stages to a defined position before the first scan point and sends a trigger signal to the LCLS control system. This trigger signal induces a defined sequence of TTL signals to be sent from the LCLS control system to the motion controller, to allow the scanning stage to reach a constant speed and the position of the chip pores to be in phase with the arrival of the X-ray pulses at the predefined beam position. After the starting point of the grid is reached, the pulse picker opens, and the X-rays hit the crystals located in the pores of the chip. At each scan point, the current position is read out by the controller, and any error is instantaneously injected into the control loop and compensated for to prevent accumulation of errors. After a predefined number of pulses (equal to the number

© 2017 Nature America, Inc., part of Springer Nature. All rights reserved.

of pores selected) has been reached, the pulse picker closes, and the scanner decelerates before switching to the next line of the chip. This sequence is repeated for all rows of the chip in a meander-scan-like manner.

After interacting with the sample, the direct undiffracted beam is guided into a beam pipe, which is shown and described in **Supplementary Figure 4**. By enclosing the beam in a beam pipe, all X-rays scattered by air are absorbed in the walls of the tube and thereby do not contribute to background scattering at the detector.

Roadrunner control system and software. Alignment of the goniometer setup, control of individual motors, prealignment of the individual chips, definition of the scan grid, and data collection are controlled by custom software written in the Python programming language with the underlying control system TANGO. The software provides a graphical user interface (GUI) for easy and efficient operation. A screenshot of the Roadrunner GUI together with a more detailed functionality of the software is provided in **Supplementary Figure 6**. The style and functionality of the GUI were adapted from GUIs typically used at protein crystallography beamlines. The software is available for free download at http://doi.org/10.5281/zenodo.571598.

Data collection. Measurements were conducted on the XPP instrument at the LCLS at SLAC under experiment number LH90. An X-ray energy of 9.5 keV was chosen for the experiment because it provides a good compromise between detector efficiency and pulse intensity on the one hand and X-ray absorption by the silicon chips on the other hand. The X-ray beam size at the sample was $3 \times 3\ \mu m^2$. X-ray pulse energies were attenuated to 40% of the full flux. A photograph of the Roadrunner setup installed at the XPP instrument at LCLS is shown in **Supplementary Figure 5**.

For measurements performed at room temperature, it was observed that, with the aforementioned procedure, only the first shot of each line yielded useful diffraction data, probably because of predamage of the subsequent crystals by the wings of the X-ray beam. It was therefore necessary to increase the displacement of subsequent pulses to prevent damage of the crystals caused by the previous X-ray pulse. For this reason, the beam shutter was used to chop the repetition rate of the laser to 30 Hz, shooting only every fourth hole in the chip (**Fig. 2d**). In addition, during the scan, only every second line of the chip membrane was scanned. In this way, an effective acquisition rate of up to 12.2 images/s was achieved for room-temperature measurements.

At cryogenic temperatures in each line, the chip was translated with a speed of 1.2 mm/s, so that the displacement of subsequent pulses matched the distance between two neighboring holes in the chip (**Fig. 2e**). Thus, a maximum data acquisition rate of 120 Hz could be achieved within a line. After the end of a line was reached, the chip moved to the next line and scanned in the reverse direction. This process allowed for scanning of the entire chip membrane with ~19,000 collected detector frames in less than 10 min, thus resulting in an effective data acquisition rate of 33.6 images/s (**Table 1**).

Data evaluation and structure refinement. Diffraction images considered hits were isolated from the XTC streams with *cctbx.xfel* and were passed into the data processing pipeline *cppxfel*[53]. Technical difficulties with indexing the BEV2 diffraction patterns drove the development of a new indexing algorithm called TakeTwo[54]. Owing to the increased indexing rate from this algorithm, the TakeTwo algorithm was used to index both the BEV2 and CPV18 samples. Integration, initial orientation matrix refinement, and postrefinement were carried out similarly to previously described procedures[47,53]. Geometry was refined initially with *cctbx.xfel* and then was further refined with the geometry algorithm in *cppxfel* by using the spot predictions from the indexing solutions and the nearest peak pixel value. For BEV2, a 2×2 foreground integration window was used to match the spot size, and care was taken to ensure that the background subtraction region did not overlap with a neighboring spot. After geometry refinement, the accuracy of spot prediction allowed interpolation between pixels to be used. After postrefinement, the BEV2 data were reintegrated with the updated orientation matrix to more accurately predict the spot positions. For CPV18, the integration window was 5×5, owing to the larger spot size.

For the BEV2 samples, 446 detector frames out of 8,812 collected images from five different chips were classified as possible hits (**Table 1**). Among these, 352 indexed diffraction patterns were obtained, of which 324 diffraction patterns were included in the final data set and used for structure refinement. To test whether the amplitudes were sufficiently accurate and complete to support phase determination, an initial map was calculated from phase information by using the known BEV coordinates truncated to 5 Å. Density modification, NCS averaging and gradual phase extension was performed from 5 Å to 2.5 Å, thus providing a good interpretable map with clear side chain density (**Supplementary Fig. 8**, which also shows the relationship between the phases derived from phase extension and those derived from an averaged map obtained from the synchrotron data). The atomic coordinates were rebuilt into the map derived from phase extension to remove bias from the model reported by Axford *et al.*[4], by using CNS[55] with strict NCS constraints. The resulting BEV2 electron densities are shown in **Figure 4**. In a second test, phases derived from the suitably placed capsid of FMDV A22 were combined with amplitudes from BEV2 (to a resolution of 2.3 Å). The level of sequence identity between the capsid protein of these two viruses was only 19.5%. This map was then refined by cyclic density modification and NCS averaging (**Supplementary Figs. 9** and **10**).

For the CPV18 sample, 13,424 diffraction images out of 19,028 collected images were regarded as hits, all from one single chip. Images were indexed with the multiple lattice version of the TakeTwo algorithm, thus producing 16,739 indexing solutions. Up to five diffraction patterns could be indexed on a single image, owing to multiple hits (**Supplementary Fig. 14**). Finally, 9,293 patterns were included in the final data set. For structure refinement of CPV18, phases were introduced from PDB 4OTS as a template file, and the structure was refined with Phenix[56]. Further data evaluation details are summarized in **Supplementary Table 2**. Part of the electron density map for CPV18 is shown in **Supplementary Figures 10** and **11**.

Statistics. Statistical measures addressing the quality of the two refined structures are provided in **Supplementary Table 2**.

© 2017 Nature America, Inc., part of Springer Nature. All rights reserved.

doi:10.1038/nmeth.4335

Code availability. The Roadrunner control software is available under doi 10.5281/zenodo.571598 (http://doi.org/10.5281/zenodo.571598).

Data availability. Solved structures have been deposited in the Protein Data Bank under PDB 5MQU and PDB 5MQW for BEV2 and CPV18, respectively. Source data for **Figure 3d** are provided as a separate data file.

50. Roedig, P. *et al.* Sample preparation and data collection for high-speed fixed-target serial femtosecond crystallography. *Protocol Exchange* http://dx.doi/org/10.1038/protex.2017.059 (2017).
51. Walter, T.S. *et al.* A procedure for setting up high-throughput nanolitre crystallization experiments: crystallization workflow for initial screening, automated storage, imaging and optimization. *Acta Crystallogr. D Biol. Crystallogr.* **61**, 651–657 (2005).
52. Wheeler, M.J., Russi, S., Bowler, M.G. & Bowler, M.W. Measurement of the equilibrium relative humidity for common precipitant concentrations: facilitating controlled dehydration experiments. *Acta Crystallogr. F Struct. Biol. Cryst. Commun.* **68**, 111–114 (2012).
53. Ginn, H.M., Evans, G., Sauter, N.K. & Stuart, D.I. On the release of *cppxfel* for processing X-ray free-electron laser images. *J. Appl. Crystallogr.* **49**, 1065–1072 (2016).
54. Ginn, H.M. *et al.* TakeTwo: an indexing algorithm suited to still images with known crystal parameters. *Acta Crystallogr. D Struct. Biol.* **72**, 956–965 (2016).
55. Brunger, A.T. Version 1.2 of the Crystallography and NMR system. *Nat. Protoc.* **2**, 2728–2733 (2007).
56. Adams, P.D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 213–221 (2010).

© 2017 Nature America, Inc., part of Springer Nature. All rights reserved.

JOURNAL OF SYNCHROTRON RADIATION

ISSN 1600-5775

The slip-and-slide algorithm: a refinement protocol for detector geometry

Helen Mary Ginn[a,b]* and David Ian Stuart[a,b]

[a]Division of Structural Biology, Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford OX3 7BN, UK, and [b]Diamond House, Harwell Science and Innovation Campus, Fermi Avenue, Didcot OX11 0QX, UK. *Correspondence e-mail: helen@strubi.ox.ac.uk

Received 1 June 2017
Accepted 18 September 2017

Edited by G. Grübel, HASYLAB at DESY, Germany

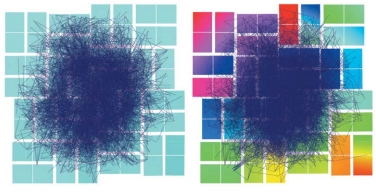
Keywords: serial crystallography; free-electron laser; SFX; stills; XFEL; detector; geometry.

Geometry correction is traditionally plagued by mis-fitting of correlated parameters, leading to local minima which prevent further improvements. Segmented detectors pose an enhanced risk of mis-fitting: even a minor confusion of detector distance and panel separation can prevent improvement in data quality. The slip-and-slide algorithm breaks down effects of the correlated parameters and their associated target functions in a fundamental shift in the approach to the problem. Parameters are never refined against the components of the data to which they are insensitive, providing a dramatic boost in the exploitation of information from a very small number of diffraction patterns. This algorithm can be applied to exploit the adherence of the spot-finding results prior to indexing to a given lattice using unit-cell dimensions as a restraint. Alternatively, it can be applied to the predicted spot locations and the observed reflection positions after indexing from a smaller number of images. Thus, the indexing rate can be boosted by 5.8% using geometry refinement from only 125 indexed patterns or 500 unindexed patterns. In one example of cypovirus type 17 polyhedrin diffraction at the Linac Coherent Light Source, this geometry refinement reveals a detector tilt of 0.3° (resulting in a maximal Z-axis error of $\sim$0.5 mm from an average detector distance of $\sim$90 mm) whilst treating all panels independently. Re-indexing and integrating with updated detector geometry reduces systematic errors providing a boost in anomalous signal of sulfur atoms by 20%. Due to the refinement of decoupled parameters, this geometry method also reaches convergence.

1. Introduction

A key part of data processing for any diffraction experiment, including X-ray diffraction, is determining the parameters of the experiment: those of the sample (*e.g.* unit-cell dimensions, crystal orientation, size, mosaicity), those of the probe (*e.g.* the beam focus, divergence, energy) and the recording device (*e.g.* detector gain curve, angle relative to beam, placement relative to sample, segment placements relative to each other). Due to well characterized detectors present in most synchrotron beamlines, researchers benefit from very good calibration of these detectors, including gain and precise knowledge of the placement of the sensors with respect to each other.

The machines with the most brilliant hard X-ray sources are now X-ray free-electron lasers (XFELs): the Linac Coherent Light Source (LCLS) and the SPring-8 Angstrom Compact Free Electron Laser (SACLA) are already in use and the European XFEL (EuXFEL) and SwissFEL have begun lasing. These are equipped with custom, in-house detectors, which embody technological advances necessary to cope with the demands of recording XFEL data, including: the Cornell-SLAC hybrid Pixel Array Detector (CSPAD) (Hart *et al.*, 2012) at LCLS, the MultiPort Charged Coupled Device



OPEN ACCESS

(MPCCD) (Kameshima *et al.*, 2014) detector at SACLA, and the upcoming Adaptive Gain Integrating Pixel Detector (AGIPD) (Henrich *et al.*, 2011) and the Jungfrau detector (Mozzanica *et al.*, 2014) at the EuXFEL and SwissFEL, respectively. The former two detectors are widely used at their respective light sources. These detectors are all segmented, and are being upgraded frequently. The CSPAD in particular is a 64-segmented detector of application specific integrated circuit (ASIC) modules which are bump-bonded into pairs. This requires regular dismantling and rebuilding, which inevitably changes the relative positions of the detector segments. This is an important part of the technological development of these detectors. However, the frequent (and unknown) rearrangement of the detector panels has a pronounced effect on the ability to process the data. The resulting panel shifts are largely unknown, and have to be back-calculated from data collected at the XFEL. For the regular crystallographer, the careful calibration of commercial detectors at synchrotron beamlines may have allowed this aspect of detector calibration to drift to the back of the mind.

XFEL data provide the biggest challenge for back-calculation of detector geometry. At an XFEL, crystals enter the beam with an unknown orientation, and a portion of X-rays entering a crystal will satisfy the Bragg condition and diffract to produce a measurable spot on the detector. A crucial stage of the data analysis workflow is working out the relative positions of the putative reflections which gave rise to these spots. This dictates how X-rays are traced from spot to sample and determines the intersection with the Ewald sphere, a critical component of determining crystal orientation. The geometry is also essential for predicting the positions of the diffracted X-rays in preparation for integrating reflection positions. Even a sub-pixel offset in the assigned position of a panel from its true position will disturb the integrated value for multiple reflections.

Geometry refinement calculations have been carried out by a number of groups. Work in early phases of XFEL data processing noted the requirement for further detector geometry refinement from the initial parameters, being incorporated into *geoptimiser* (Yefanov *et al.*, 2015), a stand-alone software package associated with *CrystFEL* (White *et al.*, 2016), and a geometry (also termed metrology) refinement module in *cctbx.xfel* (Hattne *et al.*, 2014). Many of the authors of the latter went on to incorporate a different form of detector geometry refinement in the DIALS framework for goniometer-based experiments and this has been adapted to exploit XFEL data (Waterman *et al.*, 2016; Brewster *et al.*, 2017).

The first two XFEL-specific methods both concentrate on the challenging arrangement of the CSPAD detector. The methods of geometry refinement by Hattne *et al.* (2014) were carried out using an initial GUI-based manual adjustment of the summed-image powder rings to obtain an initial starting solution, which is performed automatically using *geoptimiser*. Both of these reported methods then refine the positions and rotations of all ASIC pairs on the CSPAD, the orientation and unit-cell dimensions for each crystal, and the beam translation and sample-to-detector distance by least-squares refinement to minimize the discrepancy between observed and predicted spot positions from known indexed lattices. On the other hand, Yefanov *et al.* (2015) refined three parameters per individual panel: translation in X and Y (axes perpendicular to the beam) and a rotation around the centre of the panel, all kept in plane with no variation in the Z axis. Although this appears to be a less ambitious approach, because certain combinations of parameters can look remarkably similar, refining fewer parameters prevents one parameter from compensating for errors in another correlated parameter. As a result, this method may be advantageous against the chosen target function.

Detector geometry refinement in DIALS aims to refine geometry for both single-shot experiments as per XFEL experiments and goniometer-based rotation experiments at synchrotrons. This is a more ambitious project which includes fully three-dimensional detector models with an optional hierarchical structure, potentially able to model the panels arranged on a cylinder of the Pilatus 12M detector at Diamond, beamline I23 (Wagner *et al.*, 2016). In the DIALS framework, during refinement, panels are rotated with respect to their basis vectors, and also translated along these three vectors (which are not necessarily parallel to laboratory framework axes). However, several combinations of parameters defined in this model are still correlated, which significantly complicates the process of refinement.

In this paper, we present a new method of parameterization of a detector in three dimensions which aims to deconvolute the vast majority of the correlated parameters. These can be split into two major modes of movement: Euclidean movements which affect the relative arrangement of putative reflection coordinates in reciprocal space, and non-Euclidean movements of spherical geometry which rotate a panel around the sample but do not affect the relative arrangement of putative reflection coordinates (Fig. 1). This has been termed the slip-and-slide algorithm, as one first 'slips' into the correct orientation relative to the sample (Euclidean movements) and then 'slides' around the sample (non-Euclidean movements). These can be refined against different target functions. Refining these movements separately prevents a one-step movement in the X axis, and this must be decomposed into its component modes (Fig. 2). In addition, we present two types of target function, one of which uses the traditional approach of comparing observed to predicted pixel positions of diffracted rays through illuminated Bragg peaks, and another function which takes advantage of the 'unit cell fingerprint' of the lattice-derived pattern of spots on the detector, with no concern given to the orientation of the lattices.

The effect of this method of geometry refinement on even a data set of high initial quality is demonstrated by observing the effect on an independent parameter (anomalous signal from sulfur), which should indicate the relative importance of geometry calibration. This corroborates the message from a study on sulfur SAD phasing of thaumatin crystals (Nass *et al.*, 2016) which found that applying further geometry correction to gadolinium-phasing of lysozyme (Barends *et al.*, 2014) could

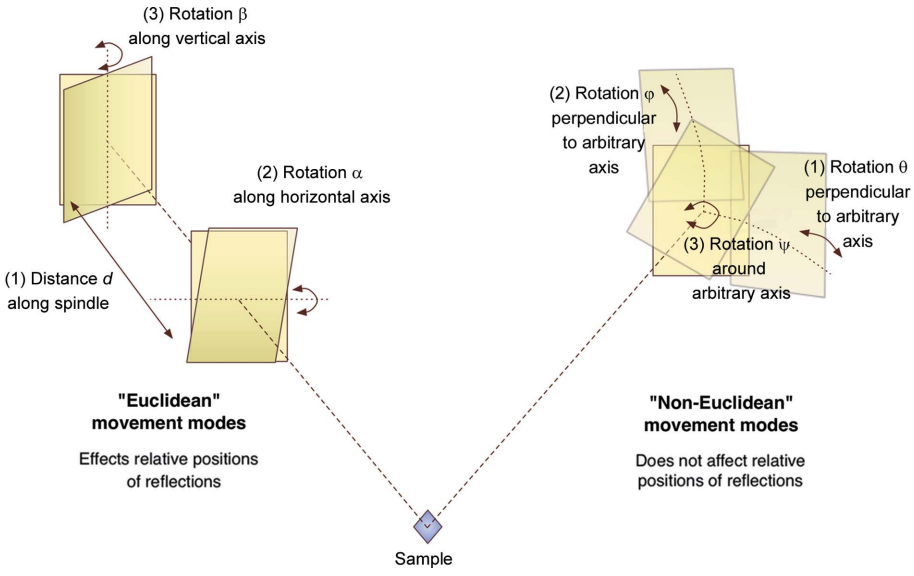


Figure 1
Types of movement: Euclidean (left) and non-Euclidean (right). Left: the detector panel can move backwards and forwards on the spindle connecting its midpoint to the sample by distance d, and rotate by two angles on two axes orthogonal to the spindle axis (α and β). This will affect the predicted spot locations (or, conversely, it will affect the relative arrangement of back-projected rays). Right: the detector panel can move around polar coordinates θ, φ and ψ around any arbitrary axis passing through the sample. This will not affect the relative arrangement of ray projections back onto the Ewald sphere, only their absolute position relative to the beam centre.

reduce the required patterns from 60000 in the original study to 7000.

2. Materials and methods

2.1. Data input

The initial geometry was loaded from the *CrystFEL* geometry file for the CSPAD (at LCLS) into the *cppxfel* suite. Files in the CXI format (using the HDF5 specification) were created using *Cheetah* (Barty *et al.*, 2014) and can now be directly loaded into *cppxfel* (Ginn *et al.*, 2016*a*). Cypovirus type 17 polyhedrin (CPV17) diffraction patterns from a prior study (Ginn *et al.*, 2015*b*), selected using *Cheetah* hit-finding [data deposited in CXIDB (Maia, 2012), code 29], were used, from LCLS proposal LS06. Note that, for the CSPAD, the bump-bonded ASIC pairs were treated as a single unit.

2.2. Detector geometry model

The underlying model of the detector geometry is no more complex than that implemented in DIALS. The laboratory basis vectors were chosen such that the Y axis pointed to the ceiling, the Z axis along the beam from sample to detector, and the X axis in the appropriate direction to complete a right-handed orthogonal coordinate system. The origin was defined as the sample position. Each segment of the detector stored three properties: its midpoint coordinate, and the direction vectors which specify the fast-scan and slow-scan axes (from which the orthogonal vector may be derived). The 32 individual ASICs of the detector were paired together to form 16 ASIC pairs. Individual ASICs first inherited the properties from these higher-order ASIC pairs, before applying further corrections. These 16 ASIC pairs were paired recursively, ending with the definition of a single 'master' detector object (Fig. 3).

During the course of refinement, small offsets are applied to the detectors in order to minimize the target function, which are applied on-the-fly. At the end of each cycle of refinement of a detector panel, these small offsets (rotation angles and translations as defined in §2.3) are absorbed into the direction vectors which specify the fast- and slow-scan axes. The small

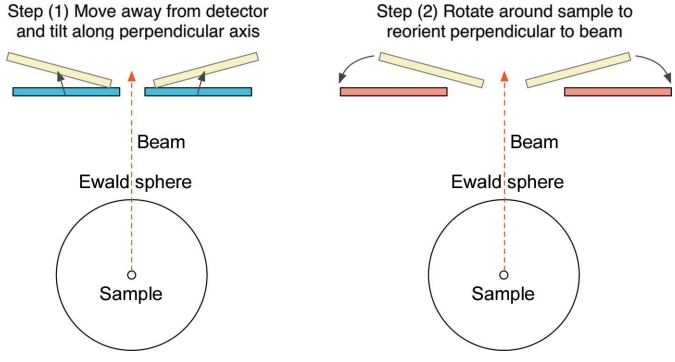


Figure 2
A lateral movement in panels must be split into the Euclidean elements (step 1, left) and the non-Euclidean elements (step 2, right). The former aligns the spot projections or predicted reflection positions within a single panel. The latter aligns the panels relative to the beam centre or each other.

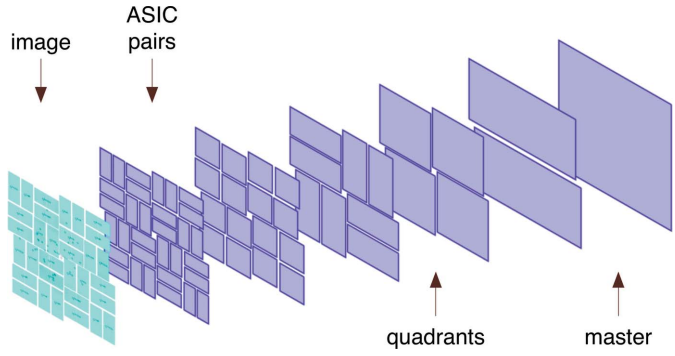


Figure 3
Detector hierarchy for a CSPAD image (left). Subsequent images show the pairs of detector groups joined in a higher-level group from the previous image, finishing with the final 'master' group. Each group will inherit the properties of the parent group recursively. This enables blanket changes to be made to the entire detector or specific sections of it, and also aids the process of refining geometry from unindexed diffraction patterns.

offsets are then set to zero before the next round of refinement. This is merely a detail of the implementation.

2.3. Panel movement modes

Movement modes were divided into Euclidean and non-Euclidean movements. The three Euclidean movements are: (*a*) movement back and forth along a spindle connecting the centre of the panel to the sample, (*b*) a rotation around an axis normal to the spindle, and (*c*) a rotation around the other orthogonal axis normal to the spindle. There were three types of non-Euclidean movement, polar coordinates θ, φ and ψ around any axis passing through the sample (Fig. 1).

2.4. Refinement of unit-cell dimensions

To refine unit-cell dimensions independently from detector geometry, a combination of over-prediction and a local search for the highest peak, followed by integration, could 'catch' reflections which would not satisfy the Bragg condition using the parameters of the initial model. This identified illuminated reflections which were further afield from the initial prediction, and was followed by initial orientation refinement. Reflections were registered as either strong or not strong, and thus provided data which were independent of the diffraction geometry. Global unit-cell refinement was then performed to bring the unit-cell dimensions in line with what was required to focus the reflections around the mean wavelength. Iterations of orientation refinement followed by unit cell refinement were performed until convergence.

2.5. Geometry refinement using indexing solutions

Diffraction patterns of CPV17 run 4 were indexed using an initial estimate of the detector geometry. The initial offsets between the predicted and observed locations for the vast majority of ASIC pairs were within 0 to 2 pixels, which was also affected by the unit cell and distance correction described above. Indexing was performed using the TakeTwo algorithm (Ginn *et al.*, 2016*b*), including solution validation. During indexing, the surrounding area up to five pixels away from the predicted peak was searched, and the highest pixel value stored as an offset. The offset was stored in both pixels and in reciprocal coordinates when back-projected onto the Ewald sphere. This local search was also used during the error checking process of indexing solutions for the initial geometry only, which relaxes the restraints. For the validation of indexing rates determined with refined geometry, the local search was not performed. Only the strongest peaks were recorded using an equivalent of an I/SIGI cutoff of 0.67 (using a detector gain of around 9.0 ADU per photons for the 1.46 Å wavelength pulses used).

The Euclidean movement modes were refined first using the Nelder–Mead algorithm (Nelder & Mead, 1965) against a measure of the spread of the pixel offsets. As there were many incorrectly located peaks, a least-squares offset would put undue emphasis on outliers where the majority of the errors are not similar to those derived from a Gaussian error model. Instead, an empirical 'reward-based' target function was developed (rather than penalizing for being far from the true value).

The equation E which was to be maximized using the Euclidean movement modes only is described in equation (1) below. This calculated and summed a value for every strong reflection *i*, which was the sum of an exponential function of the axis offsets *x* and *y* of every other reflection *j* from reflection *i*. Hence this was independent of the origin of the pixel offsets and would be best maximized by reducing the spread of neighbouring offsets. The non-Euclidean movements were refined against the deviation of the reciprocal offsets (in reciprocal coordinates) from the origin. When X-rays diffract through a finite reciprocal lattice point at a wide angle and hit a detector far from the perpendicular, the rays will be distributed with an increasing spread of incident angles. The choice to calculate in reciprocal offsets guarded against the over-compensation which would occur by trying to centre on the midpoint of the pixel offset distribution. This was calculated using a related equation (2) using non-Euclidean movements to rotate the panels until the offset is zero. Equation (2) is less computationally expensive than equation (1). The constant k in both equations is a scale factor to roughly scale the offsets to the appropriate scale, and was set to 5.0 for pixel offsets and 0.002 Å^{-1} for reciprocal coordinate offsets. The flow of operations are described in Fig. 4.

$$\mathrm{E} = \sum_i \sum_j \exp\left\{-k^2\left[\left(x_i - x_j\right)^2 + \left(y_i - y_j\right)^2\right]\right\}, \qquad (1)$$

$$\mathrm{N} = \sum_i \exp\left[-k^2\left(x_i^2 + y_i^2\right)\right]. \qquad (2)$$

Each mode of movement was refined independently for each panel. This ensures that any global trends observed in the detector have not been biased by using a hierarchical detector. This means any trends in overall tilt or detector distance can be determined independently.

2.6. Geometry refinement using unit-cell information only

Spot-to-spot vectors were determined in the same manner as previously described (Ginn *et al.*, 2016*b*). 'Intra-panel' vectors are vectors which span two spots on the same detector panel, whereas 'inter-panel' vectors are those which have two spots each from a different panel.

For a given panel, all pairs of intra-panel vectors sharing a common spot were stored, and on each iteration of refinement the two reciprocal lengths of the vectors and the angle between them (restricted to $0 \leq \text{angle} < 90$) were recalculated. This produces a three-dimensional histogram of two lengths, and angle and a frequency. When summed over multiple still shots from several angles, a sampling of all potential vector pairs is expected. For each panel, this histogram was refined against the ideal histogram calculated from the unit-cell dimensions, using the Euclidean movement parameters. This is refined using the Nelder–Mead algorithm to convergence.

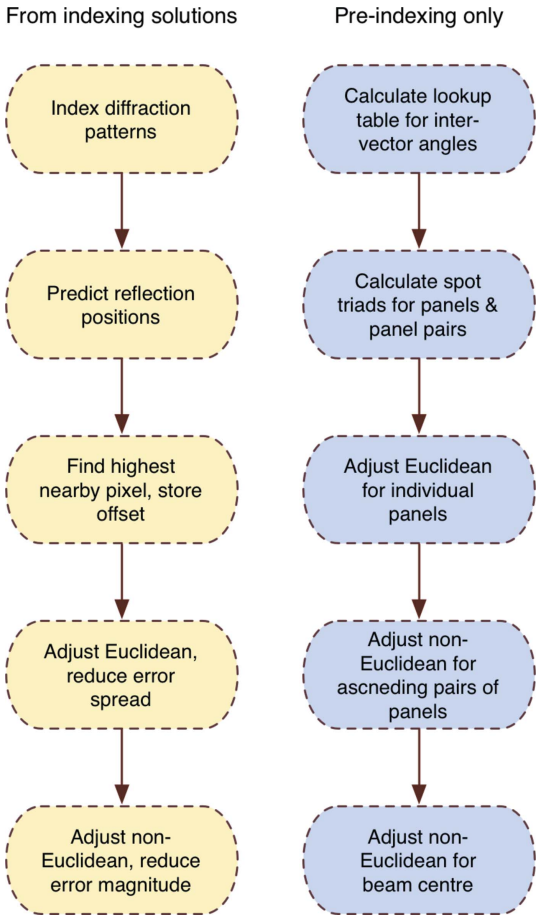


Figure 4
The order of events is shown for geometry refinement from indexing solutions for individual crystals (left) and for using the pre-indexing 'lattice-like' information only (right).

Inter-panel vectors between two chosen panels were paired either with other inter-panel vectors or intra-panel vectors on one panel or the other, as long as they shared a common spot. The three-dimensional histogram was generated and refined against the non-Euclidean movements of the two panels next to each other. The movement is defined by the polar coordinates θ, φ and ψ, the axis for which was defined as the midpoint between the centres of the two chosen panels. The panels are restricted to perform equal and opposite movements (*e.g.* rotation by θ of one panel must be opposed with a $-\theta$ rotation of the other), to prevent unnecessary drift of panels around the experimental hutch. A hierarchy of panels was formed for both the CSPAD and MPCCD detectors, with ascending pairs of panels up to the entire detector. For the CSPAD, adjacent ASIC pairs are paired, followed by paired ASIC foursomes, up to quadrants, halves and then the master detector. Each level of the hierarchy was refined using the Nelder–Mead algorithm to convergence.

The target function for the unit-cell fingerprint was calculated from the ideal unit-cell dimensions and a reciprocal-lattice-point radius (akin to the profile radius in *CrystFEL* or the crystal domain size for *cctbx.xfel*). These acted as tolerances for the lengths of vectors to be accepted as resulting from the chosen lattice parameters. Vectors were rejected if the magnitude of the vector exceeded 0.15 Å^{-1} (reciprocal coordinate units). The angles were calculated from the tolerances on the lengths of the vectors such that shorter vectors had wider tolerances. These are stored in memory with 240 voxels sampling each dimension, producing 240^3 single-precision floating points (occupying 55.3 MB in memory). Each voxel contained a 'score' for a given combination of two vector lengths and the angle. This was a linear ascent from 0 at the edge of the acceptable tolerance to 1 at the most ideal calculated value. Sets of three spots were taken for each target function, and the three vectors between them were stored in memory. For the target function, a sum of the individual scores for each spot triad were calculated. Each pair of the three vectors in a spot triad were given a score between 0 and 1 according to the lookup table, and multiplied together, which can also be calculated in advance.

2.7. Calculation of anomalous signal

Diffraction patterns were indexed using a given geometry file and orientation matrices were refined according to a previous protocol (Ginn *et al.*, 2015*b*). The peak locations for these orientations were then predicted, and could be treated as exact or approximate. For the former, images were integrated at the exact peak locations; for the latter, after a three-pixel local search for the highest pixel value, on which the integration window was centred. Integration and post-refinement of these diffraction patterns were performed as previously described (Ginn *et al.*, 2015*a*) and anomalous differences were calculated per cycle.

3. Results

Geometry movements were categorized into either 'non-Euclidean' or 'Euclidean' movements. The non-Euclidean geometry movements are those which do not affect the relative positions of diffracted rays on the surface of the panel, *i.e.* any movement on a sphere around the sample of a constant radius. The other movements are those which only affect the relative positions of the diffracted rays and are classed as Euclidean movements. These two modes of movements are refined against separate target functions, involving no component to which the parameters are insensitive. This means that, apart from the exception of moving an entire single-panel detector forwards or backwards, any single movement along the traditional $X/Y/Z$ axes is disallowed. These movements must be decomposed into the Euclidean and non-Euclidean elements (Fig. 2) and refined against separate target functions.

Two methods can be used to refine the geometry, depending on whether indexing solutions have been obtained or not. If the geometry is far from the truth, orientation matrices may be extremely difficult to determine. However, the images contain a great deal of information as the spots still obey a 'lattice-like nature' defined by the unit-cell parameters and space group centring. Hence the Euclidean and non-Euclidean movements

can be refined against related target functions, depending on whether the geometry is close to the true positions or needs to be refined from a far starting position.

3.1. Corrections to unit-cell parameters

The unit-cell dimensions and detector geometry parameters must be carefully separated. Using the same target functions to refine the unit-cell dimension as to refine the detector geometry will lead to incorrectly determined parameters. Hence, a method to abstract the information necessary for unit-cell refinement from the detector geometry is performed instead (see §2.4). This clusters the prediction of the strong reflections around the mean wavelength recorded by integration of the spectrometer output at the XFEL for both data sets. This method was used to update the unit-cell dimensions used for CPV17 from 106.1 Å to 105.5 Å (deflation of approximately 0.6%). An approximate correction was applied to the initial detector distance (inflation by 0.6%) to compensate for the change in unit cell.

3.2. Indexing-driven geometry refinement

If the detector geometry is sufficient to index and produce pixel offsets not diverging too far from zero, the Euclidean and non-Euclidean movements can be refined against the appropriately sensitive target functions. The Euclidean movements are refined against the spread of pixel offsets (but need not be equal to zero), and the non-Euclidean movements are refined against the spread of reciprocal coordinate offsets being centred around zero. The result of applying the two-stage protocol on individual panels is presented in Fig. 5, and the overall effect on the pixel offsets on each quadrant is shown in Fig. 6.

The TakeTwo algorithm (Ginn *et al.*, 2016*b*) may be able to index with very poor geometry, in the case where the independent panels are well oriented in their Euclidean movements but are poorly oriented with respect to either each other or the beam centre. If there are enough panels oriented correctly with respect to the sample, and enough interconnected panels to find enough vectors consistent with a single indexing solution, identifying the correct basis vectors is very likely to be successful. However, the pixel offsets may be extremely divergent. If the indexing rate is very low to begin with, pre-indexing geometry refinement can be performed to help bootstrap indexing.

3.3. Pre-indexing geometry refinement

By analysing pairs of spot-to-spot vectors, one can essentially produce an orientation-less 'unit-cell fingerprint' of the space group and centring (disregarding axial systematic absences). This unit-cell fingerprint will be perturbed if the non-Euclidean movement parameters of a single panel are incorrect, and so can be used as a target function to lock this information. This is demonstrated graphically in Fig. 7.

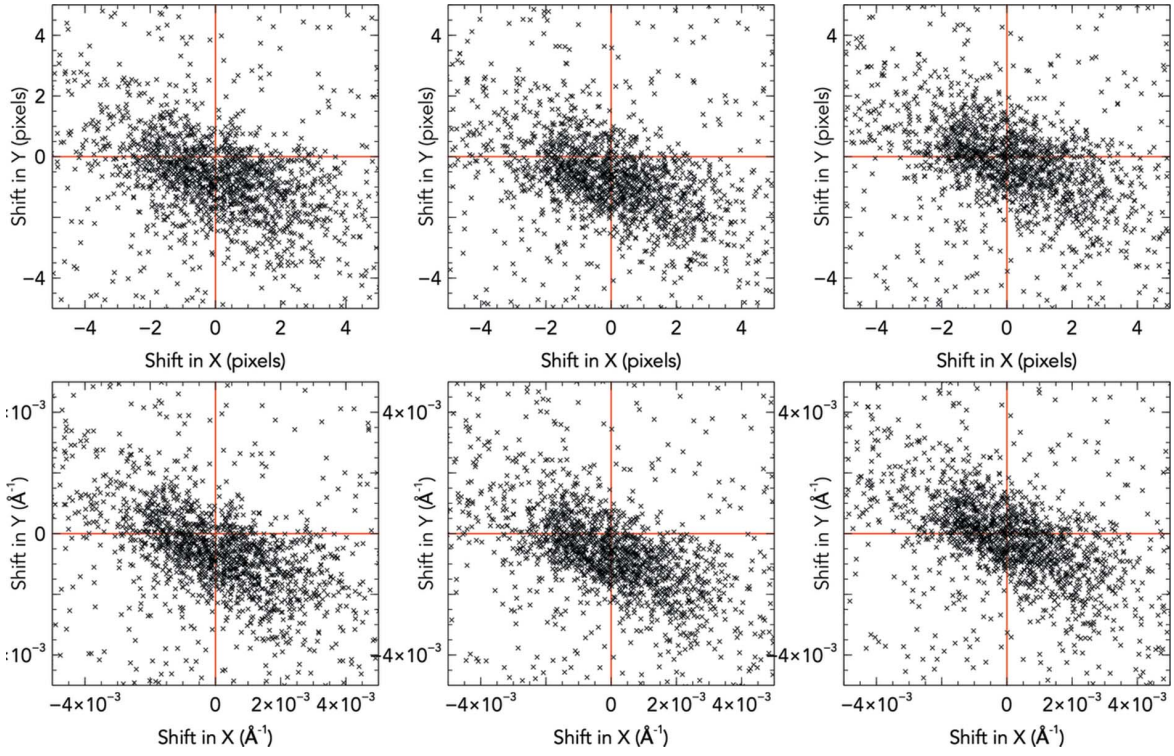


Figure 5
The shifts in predicted reflection positions on a mid-resolution panel at the start (left), after Euclidean movements (centre), and after non-Euclidean movements (right). The former is refined against the spread of pixels (top row), resulting in a reduction of the standard deviation of reciprocal offsets by 0.8% (including outliers), whereas the latter is refined against the centre-point of the reciprocal offsets (bottom row), resulting in a reduction of the mean reciprocal offset by 77.1%.

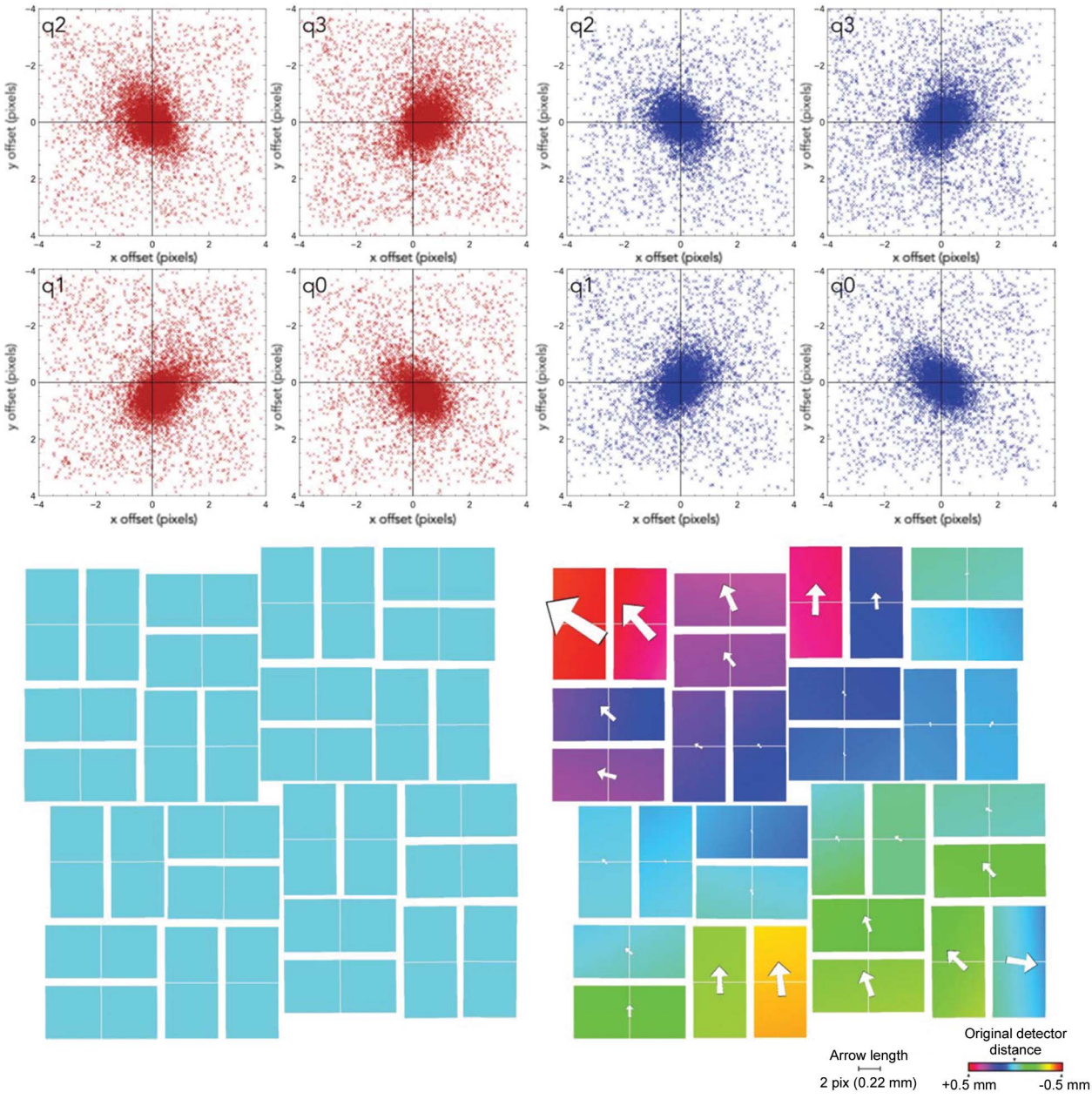


Figure 6
The shifts of the pixel offsets from the predicted and observed peak locations in each quadrant of the CSPAD from the original offsets in the geometry file (top, red), and after refinement (top, blue). On the bottom is the before and after CSPAD corresponding to the graphs above, with the Z axis spanning a 1 mm rainbow around the original detector distance. Arrows shown are proportional to the offset in X and Y of each panel from the original position and independent of the Z axis. Geometry refinement performed using 170 indexing solutions.

Due to the lack of exact orientation information, it is harder to lock panels in an absolute position relative to the beam centre. However, it is possible to lock panels to one another, starting with pair-wise interactions between independent panels, and then pair-wise interactions between pairs, in an ascending hierarchy, until the whole of the detector has been aligned with itself. The inter-panel vectors contain the information which defines the relative positions of the two participating panels and this forms the basis of this part of the calculation.

Finally, the entire detector must be aligned with the beam centre. However, this is unlikely to diverge far from the original centre, since locking onto an alternative Miller index is likely if only the entire tilt of the detector has exceeded the 2θ scattering angle of the lowest resolution Bragg peak (in the case of the CPV17 unit cell, this is 1.1°).

3.4. Geometry refinement reveals a tilt in the CSPAD

Geometry refinement using the indexing solutions produces a 5.8% boost in indexing rate using information from only 125 images of CPV17 data used for a previous structure solution (Ginn *et al.*, 2015*b*). The boost in the indexing rate is comparable with that using 500 unindexed

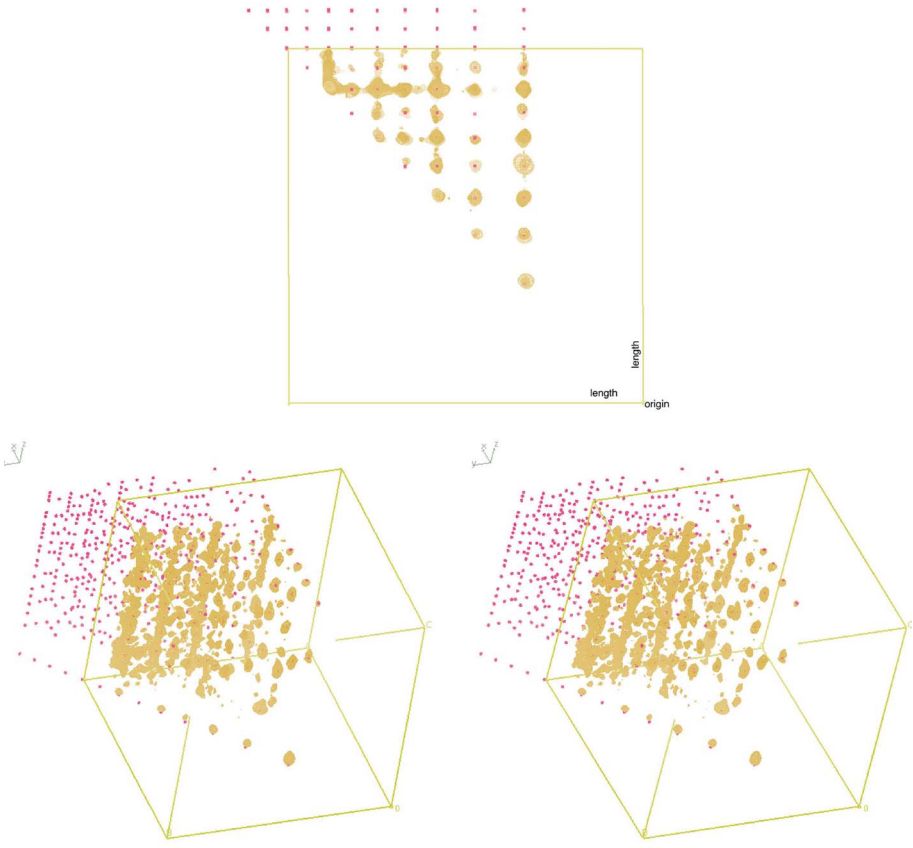


Figure 7
Representation of the three-dimensional histogram of inter-spot vectors using *Coot*. Atoms denote the positions of the expected positions, calculated with a buffer region outside of the observed vector positions (within the bounds of the unit cell). The map density reflects the set of observed vectors, contoured at 6.5σ. Reciprocal distances cover the range 0–0.04 Å^{-1}. The top diagram is shown with all angles projected down perpendicular to the page. The bottom diagram is a stereo image showing two lengths (x, y axes) and angle (z axis) illustrating all dimensions.

patterns (Fig. 8). The former method of refinement can exploit more information and therefore reaches convergence faster than using pre-indexing information only. However, the calculation time scales proportionally to the square of the images from indexing solutions [due to equation (1)], but is directly proportional to the number of images from the pre-indexing information. So including more data has a reduced impact on the calculation time using pre-indexing information. Furthermore, the latter method cuts out the time required to index beforehand. Therefore, under time-constrained conditions, one can approach an approximate solution from a larger number of unindexed diffraction patterns, and fine-tune the geometry using a smaller number of indexing solutions.

Because the geometry is run separately on individual panels using indexing solutions (and for the Euclidean stages of pre-indexing information), any global trends which are observed throughout the detector can be considered to have been reached in an unbiased manner. In this case, both forms of geometry refinement reveal a tilt in the detector (Fig. 9) of approximately 0.3°, which produces a 1.1% error in the Z axis (a detector distance in the centre of 90.3 mm). Integration and post-refinement confirms that the nominal beam centre passes through the (0, 0, 0) Miller index.

3.5. Improvement in random error reduction and indexing rates

In order to judge the reduction in random error, images in the data sets were split into odd and even chunks. Indexing solutions from each half were used to direct geometry refinement using both indexing solutions and pre-indexing information only. The changes in the X, Y and Z coordinates of the four panel corners (all of which are a net effect of both movement modes) were stored for each half after refinement, and the correlation between them measured for increasing numbers of images. This increases with use of additional images and rises faster for indexing information than pre-indexing information (Fig. 8).

The indexing rates show an immediate improvement after geometry refinement even with the minimum number of images (125 images), despite reducing the local search area to zero pixels, which produce more stringent indexing solution checks. The vectors used for indexing *via* the TakeTwo algorithm can also be plotted on a single image and this may reveal portions of the detector (or vectors between parts of the detector) which are not used during indexing. This 'iron filing' plot can highlight poor geometry between two areas or with an area of the detector (Fig. 10).

3.6. Data set quality improvement

In order to judge the reduction of systematic error, the anomalous signal was chosen as an independent metric, using *ANODE* (Thorn & Sheldrick, 2011) to calculate the average sulfur signal from Met and Cys residues after refinement. This was performed for (*a*) the initial geometry without a local maximum search, (*b*) the initial geometry with a three-pixel maximum search, and (*c*) the best diffraction geometry without a local maximum search. Post-refinement was performed on indexing solutions derived and initially refined with the best geometry, and then post-refined with either initial or refined geometry definitions. Indexing solutions were initially obtained and refined using the best geometry, resulting in 7613 indexed lattices (from 7000 images).

All data sets respond to post-refinement, increasing the anomalous signal per cycle (Fig. 11). Integrating the 7613-lattice data set using the initial geometry without performing

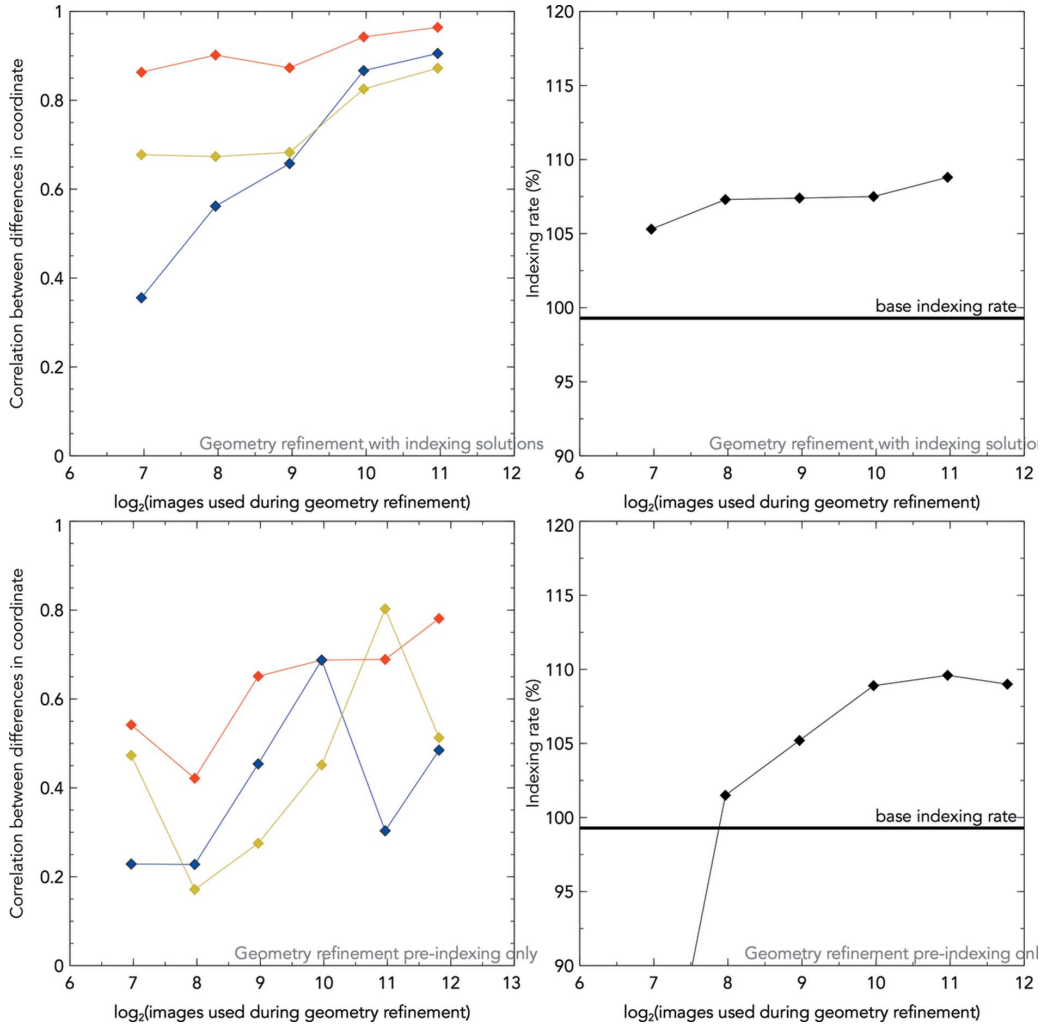


Figure 8
From the CPV17 data set: the correlation coefficients of the *X* (blue), *Y* (yellow) and *Z* (red) coordinate offsets of the four panel corners from the original positions are plotted (left), from 125 images to 2000 images (top, with indexing solutions) and from 125 to 3500 images (bottom, pre-indexing only). At the top right, the base indexing rate is shown in black for the original, partially refined *CrystFEL* format geometry file for CPV17, with a local search size of 5 pixels for the maximum peak (used for indexing solution verification). After refinement from indexing solutions, with a local search size reduced to 0 in all cases, the improvement in the indexing rates are shown. A noticeable improvement in the indexing rate can be seen from 125 images using indexing solutions. At the bottom right, 125 images fail to produce verifiable indexing solutions, but 250 images produce a small improvement in the indexing rate.

any local search produced the worst anomalous signal, reaching a maximum of 3.21σ across all sulfur atoms (an increase of 43% from cycle 0). Integrating using initial geometry combined with a local maximum search stabilizes the post-refinement process, with a maximum anomalous signal of 3.45σ (an increase of 41% from cycle 0). However, this will be held back by poorly estimated integration windows for weak reflections and would be unsuitable for careful structural analysis. This causes a systematic inflation of weak reflections which perturbs the twinning H-test by skewing the intensity distribution [*Truncate* (Winn *et al.*, 2011) reports a twinning fraction of 9.1% for the reflection intensities after a local search, *versus* 0% for reflection intensities without]. The systematic inflation of weak reflections could leave a structure prone to model bias, as phases of inflated weak reflections can be manipulated to add Fourier waves supporting any correct or incorrect input model in the electron density map. Using the best geometry to predict exact peak locations for the 7613-lattice data set, the anomalous signal reaches a maximum of 3.89 (an increase of 49% from cycle 0), showing a significant improvement derived from a more accurate prediction of the spot locations, and does not systematically inflate weak reflections.

In summary, a local search for strong reflections can mitigate the effect of poor geometry definition, improving the anomalous signal over predicting at exact locations. However, it is no substitute for accurate geometry determination.

4. Discussion

Using non-indexed diffraction information relies on good spot-finding information and consistent unit-cell parameters from crystal to crystal, and computation time scales with the number of images.

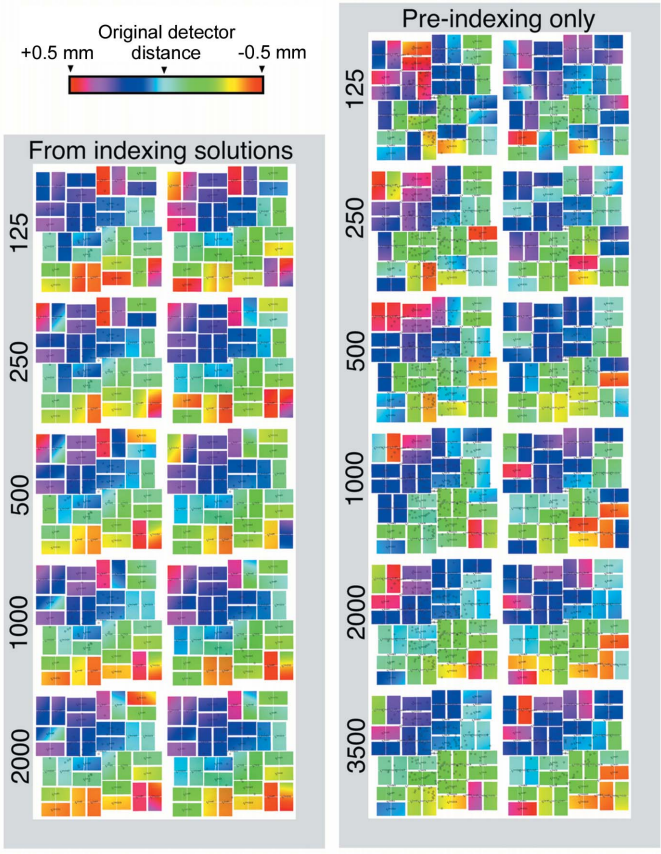


Figure 9
Side-by-side comparison of geometry refinement with indexed and unindexed diffraction patterns. Images of the CSPAD after geometry refinement from oriented indexing solutions (left), and using the pre-indexing information contained within the diffraction patterns only (right). Each column is divided into two half-data sets which were used to run the geometry calculation independently. The number of images which were given to each half are displayed on the left-hand side.

Using indexed diffraction information relies heavily on the ability to index, and computation time scales with the square of the number of images. However, it will be more robust for weak data where the knowledge of the predicted peak location can be used to select high value pixels that are more likely to be from a Bragg peak-diffracted X-ray. The geometry refinement described here allowed the structure solution of bovine enterovirus type 2 by using only 352 indexed diffraction patterns (Roedig *et al.*, 2017). In that case geometry refinement revealed a 0.5 mm tilt across the detector, correction for which enabled significantly more accurate prediction of Bragg spot locations and hence more reliable intensity estimates.

This algorithm leaves one type of parameter correlation not particularly well separated: the radial distance often has quite a similar effect to tilting along both of the perpendicular axes simultaneously. These effects can only be separated if there is a sufficient distortion in the spread of spots due to a tilt: the spots at one extreme end of a panel would be clustered together compared with those at the other extreme end. If the spread of angles between diffracted rays allows sufficient sensitivity to separate these effects, the geometry algorithm is capable of doing so. However, this becomes harder to separate

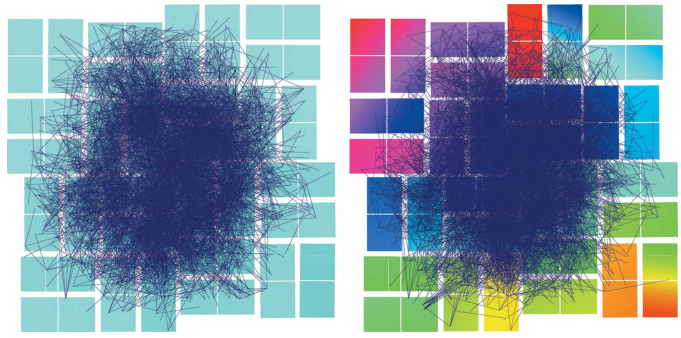


Figure 10
All inter-spot vectors used for indexing displayed on a single 'iron filing' plot, before (left) and after (right) geometry refinement with indexing solutions from 2000 images. Notice that the quadrant in the top left of the image has a poor share of the total used vectors, and, after being shifted significantly further back after refinement, has regained a more equal share of the used vectors.

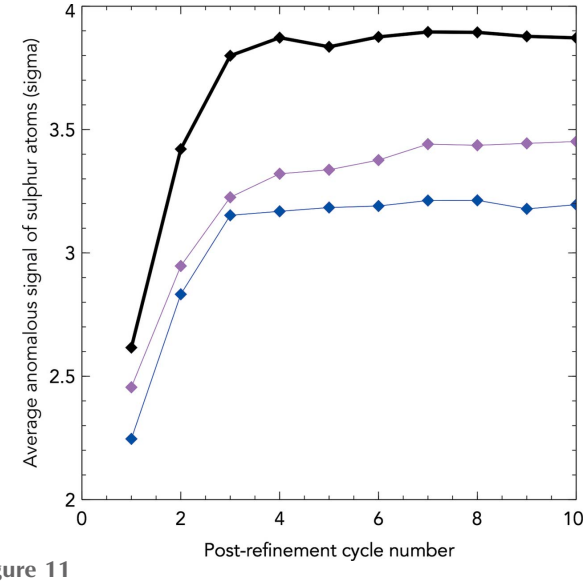


Figure 11
Graph showing the average sulfur atom anomalous signal per post-refinement cycle number for various treatments. 7613 orientations were derived from indexing and initial refinement using the best geometry file (2000 images and their indexing solutions). Crystal orientations were used to re-predict peak locations using the initial geometry integrating at the exact location (blue), or with a three-pixel local maximum search (magenta), or using the best geometry file and integrating at the exact location (black).

as the wavelength becomes shorter (leading to a flatter Ewald sphere) and when the detector distance is large. This is a loss of information which is inherent to all geometry refinement experiments, and therefore experiments to acquire good geometry should be performed with a long wavelength and short sample-to-detector distance.

Because the DIALS framework adjusts the position and rotation of the detector panels with respect to the vectors which define the panel edges and the cross product of the two, the refinement of the detector at the I23 beamline would heavily emulate the non-Euclidean movement modes described here, as the panels are deliberately arranged in a circle to match the origin of the Ewald sphere (sample posi-

tion). However, in the vast majority of experiments the detector panels lie on a largely flat surface.

Assuming correct unit-cell information, one can evaluate the success of the detector geometry refinement by (*a*) starting from different detector distances and seeing its ability to converge on a solution from different initial positions, (*b*) comparing the similarity of two half-data sets and their convergence and (*c*) checking that the spread in pixel or reciprocal offsets has reduced from indexing solutions. If the unit-cell dimensions are inflated or contracted from their true value, this may cause the detector to become concave or convex to try to correct for the error in the Ewald sphere curvature. Hence, the unit-cell dimensions may need correction beforehand.

This algorithm provides a method to quickly uncover previously unnoticed tilts in the detector using a very small number of images. This is achieved by taking the set of observations and dividing the target function into parts which are sensitive to different types of motions of the detector. These two parts of the target function can converge independently. Using indexing solutions, tilts revealed across the entire detector are unbiased because the hierarchy of the detector is not considered, and each panel is refined independently. The effect of geometry refinement has a positive effect on indexing solutions and anomalous peak heights. The code has been released for use in *cppxfel* (Ginn *et al.*, 2016*a*).

Acknowledgements

We thank Thomas A. White for providing an initial geometry estimate and initial processing with *Cheetah* and we thank Helen Duyvesteyn for hit-finding with *Cheetah* for additional CPV17 runs. DIS was supported by the Medical Research Council, HMG by the Wellcome Trust and administration support also came from the Wellcome Trust (see grants and funding details below). Portions of this research were carried out at the Linac Coherent Light Source (LCLS) at the SLAC National Accelerator Laboratory. LCLS is an Office of Science User Facility operated for the US Department of Energy Office of Science by Stanford University. This is a contribution from the Oxford Instruct-ERIC Centre.

Funding information

Funding for this research was provided by: Wellcome Trust (award Nos. ALR00040, 090532/Z/09/Z); Medical Research Council (award No. MR/N00065X/1 and previously G1000099).

References

Barends, T. R., Foucar, L., Botha, S., Doak, R. B., Shoeman, R. L., Nass, K., Koglin, J. E., Williams, G. J., Boutet, S., Messerschmidt, M. & Schlichting, I. (2014). *Nature (London)*, **505**, 244–247.

Barty, A., Kirian, R. A., Maia, F. R. N. C., Hantke, M., Yoon, C. H., White, T. A. & Chapman, H. (2014). *J. Appl. Cryst.* **47**, 1118–1131.

Brewster, A. S. *et al.* (2017). In preparation.

Ginn, H. M., Brewster, A. S., Hattne, J., Evans, G., Wagner, A., Grimes, J. M., Sauter, N. K., Sutton, G. & Stuart, D. I. (2015*a*). *Acta Cryst.* D**71**, 1400–1410.

Ginn, H. M., Evans, G., Sauter, N. K. & Stuart, D. I. (2016*a*). *J. Appl. Cryst.* **49**, 1065–1072.

Ginn, H. M., Messerschmidt, M., Ji, X., Zhang, H., Axford, D., Winter, G., Brewster, A. S., Hattne, J., Wagner, A., Grimes, J. M., Sauter, N. K., Sutton, G. & Stuart, D. I. (2015*b*). *Nat. Commun.* **6**, 6435.

Ginn, H. M., Roedig, P., Kuo, A., Evans, G., Sauter, N. K., Ernst, O. P., Meents, A., Mueller-Werkmeister, H., Miller, R. J. D. & Stuart, D. I. (2016*b*). *Acta Cryst.* D**72**, 956–965.

Hart, P., Boutet, S., Carini, G., Dubrovin, M., Duda, B., Fritz, D., Haller, G., Herbst, R., Herrmann, S., Kenney, C., Kurita, N., Lemke, H., Messerschmidt, M., Nordby, M., Pines, J., Schafer, D., Swift, M., Weaver, M., Williams, G., Zhu, D., Van Bakel, N. & Morse, J. (2012). *Proc. SPIE*, **8504**, 85040C.

Hattne, J., Echols, N., Tran, R., Kern, J., Gildea, R. J., Brewster, A. S., Alonso-Mori, R., Glöckner, C., Hellmich, J., Laksmono, H., Sierra, R. G., Lassalle-Kaiser, B., Lampe, A., Han, G., Gul, S., DiFiore, D., Milathianaki, D., Fry, A. R., Miahnahri, A., White, W. E., Schafer, D. W., Seibert, M. M., Koglin, J. E., Sokaras, D., Weng, T. C., Sellberg, J., Latimer, M. J., Glatzel, P., Zwart, P. H., Grosse-Kunstleve, R. W., Bogan, M. J., Messerschmidt, M., Williams, G. J., Boutet, S., Messinger, J., Zouni, A., Yano, J., Bergmann, U., Yachandra, V. K., Adams, P. D. & Sauter, N. K. (2014). *Nat. Methods*, **11**, 545–548.

Henrich, B., Becker, J., Dinapoli, R., Goettlicher, P., Graafsma, H., Hirsemann, H., Klanner, R., Krueger, H., Mazzocco, R., Mozzanica, A., Perrey, H., Potdevin, G., Schmitt, B., Shi, X., Srivastava, A. K., Trunk, U. & Youngman, C. (2011). *Nucl. Instrum. Methods Phys. Res. A*, **633**, S11–S14.

Kameshima, T., Ono, S., Kudo, T., Ozaki, K., Kirihara, Y., Kobayashi, K., Inubushi, Y., Yabashi, M., Horigome, T., Holland, A., Holland, K., Burt, D., Murao, H. & Hatsui, T. (2014). *Rev. Sci. Instrum.* **85**, 033110.

Maia, F. R. (2012). *Nat. Methods*, **9**, 854–855.

Mozzanica, A., Bergamaschi, A., Cartier, S., Dinapoli, R., Greiffenberg, D., Johnson, I., Jungmann, J., Maliakal, D., Mezza, D., Ruder, C., Schaedler, L., Schmitt, B., Shi, X. & Tinti, G. (2014). *J. Instrum.* **9**, C05010.

Nass, K., Meinhart, A., Barends, T. R. M., Foucar, L., Gorel, A., Aquila, A., Botha, S., Doak, R. B., Koglin, J., Liang, M., Shoeman, R. L., Williams, G., Boutet, S. & Schlichting, I. (2016). *IUCrJ*, **3**, 180–191.

Nelder, J. A. & Mead, R. (1965). *Comput. J.* **7**, 308–313.

Roedig, P., Ginn, H. M., Pakendorf, T., Sutton, G., Harlos, K., Walter, T. S., Meyer, J., Fischer, P., Duman, R., Vartiainen, I., Reime, B., Warmer, M., Brewster, A. S., Young, I. D., Michels-Clark, T., Sauter, N. K., Kotecha, A., Kelly, J., Rowlands, D. J., Sikorsky, M., Nelson, S., Damiani, D. S., Alonso-Mori, R., Ren, J., Fry, E. E., David, C., Stuart, D. I., Wagner, A. & Meents, A. (2017). *Nat. Methods*, **14**, 805–810.

Thorn, A. & Sheldrick, G. M. (2011). *J. Appl. Cryst.* **44**, 1285–1287.

Wagner, A., Duman, R., Henderson, K. & Mykhaylyk, V. (2016). *Acta Cryst.* D**72**, 430–439.

Waterman, D. G., Winter, G., Gildea, R. J., Parkhurst, J. M., Brewster, A. S., Sauter, N. K. & Evans, G. (2016). *Acta Cryst.* D**72**, 558–575.

White, T. A., Mariani, V., Brehm, W., Yefanov, O., Barty, A., Beyerlein, K. R., Chervinskii, F., Galli, L., Gati, C., Nakane, T., Tolstikova, A., Yamashita, K., Yoon, C. H., Diederichs, K. & Chapman, H. N. (2016). *J. Appl. Cryst.* **49**, 680–689.

Winn, M. D., Ballard, C. C., Cowtan, K. D., Dodson, E. J., Emsley, P., Evans, P. R., Keegan, R. M., Krissinel, E. B., Leslie, A. G. W., McCoy, A., McNicholas, S. J., Murshudov, G. N., Pannu, N. S., Potterton, E. A., Powell, H. R., Read, R. J., Vagin, A. & Wilson, K. S. (2011). *Acta Cryst.* D**67**, 235–242.

Yefanov, O., Mariani, V., Gati, C., White, T. A., Chapman, H. N. & Barty, A. (2015). *Opt. Express*, **23**, 28459–28470.