Quantifying Radiation Damage in X-ray Diffraction Experiments in Structural Biology



Jonathan Charles Brooks-Bartlett

Oriel College

University of Oxford

A thesis submitted in partial fulfilment of the degree of

Doctor of Philosophy

Trinity 2016

supervised by

Prof. Elspeth F. Garman

Abstract

Quantitative studies of global radiation damage are presented for two different types of experiments in structural biology: macromolecular crystallography (MX) and small angle X-ray scattering (SAXS)

MX is the most common technique to elucidate the atomic resolution structures of biological macromolecules. However, these molecules undergo radiation induced changes during the experiment that undesirably affect the data. Global radiation damage, which is characterised by an overall loss in the diffracted intensity of Bragg reflections, limits the amount of useful data that can be collected from a single crystal in an experiment. Furthermore, for experimental phasing experiments, the radiation induced intensity changes can be so significant that the phasing signal becomes undetectable, thereby hindering successful structure determination. This thesis investigates methods to track and correct the diffraction data that are affected as a result of global radiation damage. First, extensions to the diffraction weighted dose (DWD) metric are investigated for the ability of DWD to track the overall intensity decay of reflections. This metric then is combined with a new mathematical model of intensity decay to perform zero-dose extrapolation. An additional probabilistic extrapolation approach is incorporated into the traditional regression based approach to allow extrapolation of low multiplicity reflections. As an alternative approach, a new hidden Markov model representation of the data collection experiment is developed that allows the time-resolved calculation of structure factor amplitudes, with error estimates calculated explicitly. This method gives comparable refinement statistics to that obtained from data processed with the current data reduction pipeline, and improvements to the algorithm are proposed.

SAXS, on the other hand, is a complementary structural technique that results in low resolution information about macromolecules. However it still requires the probing of the macromolecules with ionising radiation, so radiation induced changes are still a problem. Unfortunately the tools for assessing radiation damage in SAXS experiments are not mature enough for them to be used routinely. This thesis presents extensions to RADDOSE-3D to perform dose calculations for SAXS samples. Additionally, a free, open source Python library has been developed to allow the exploration and visualisation of the results of a similarity analysis of frames within a dataset. These tools are then used to determine the efficacy of various radioprotectant compounds at different concentrations to mitigate radiation damage effects.

To Nana, for teaching me first hand the two most important lessons I've learned so far in life:

- against all odds, success can be achieved through hard work.
- the relationships built with friends and family are the most important things in life.

I know you would be proud to see how far we've come.

R.I.P.

Acknowledgements

First and foremost I thank my supervisor, Professor Elspeth Garman, for seeing my potential and giving me the opportunity to get to the position I am in now. The support she has given me over the last four years is second to none, and she has only ever put me in a position to succeed.

Markus Gerstel is ultimately responsible for the progress in my technical development. Without him I would be at least six months behind with my ninja programming skills (as Oli puts it). Markus is a great role model and I have largely looked up to him throughout my DPhil. The constant scientific discussions I have had with Charlie Bury have made my DPhil even more enjoyable. He knows what questions to ask that really rack my brain. Furthermore, he contributes to my work and development in many ways. Katharina has always been supportive and helps me with anything in the wet lab that I have little knowledge about (and that is a lot). I will never forget how stressful it is to extract crystals from glass LCP plates. Oli Zeldin has been a positive influence on me since day one. He and Markus first encouraged me to join the group and I have never looked back. To this day, Oli still believes in me and my ability and I am incredibly lucky to have him as a friend. Much of my crystallographic knowledge I owe to Edward Lowe. Is there anything he doesn't know? I can ask him anything and he will always take the time out to explain it to me, as well as help me with my data processing.

I also thank the many academics that have given their support outside of the lab: Professor Colin Please who has been a fantastic teacher and mentor ever since I began my maths degree (all those years ago), Garib Murshudov who is the major influence on the work I've carried out over the past year, and Harry Powell, Arwen Pearson, James Holton and Andrea Thorn for great discussions and encouragement over the years.

I owe thanks to my housemates Michael Barber and Nathalie Willhems for keeping me sane and ploughing me with some Dutch courage over the last three years. We have supported each other through tough times and made it to the other side. We are now Oxford Survivors.

None of this would be possible without the love and support from my family: My mother has given me everything I could ever need in life and has moulded me into the man I am today, and my sister for her ongoing love, support and encouragement that has never stopped. I could not ask any more from my godfather who has taken me under his wing and is the perfect father figure who I will look up to for the rest of my life. My partner, Beth Phipps has always been by my side through the good times and the tough times. She never stops believing in me. I am forever indebted to them all.

Finally, I am grateful to the Systems Biology program of the EPSRC funded Doctoral Training Centre for funding the studentship that made all of this work possible.

Abbreviations

ACT ..... Aimless CTruncate

CCP4 ..... Collaborative Computational Project No. 4

CMD ..... CorMap Derived

EKF ..... Extended Kalman Filter

FBA ..... Forward-Backward Algorithm

GPR ..... Gaussian Process Regression

HMM ..... Hidden Markov Model

MMSE ..... Minimum Mean Squared Error

MX ..... Macromolecular X-ray Crystallography

PCC ..... Pearson's Correlation Coefficient

RSA ..... Rotation Start Angle

REA ..... Rotation End Angle

SAXS ..... Small Angle X-ray Scattering

SGC ..... Structural Genomics Consortium

UKF ..... Unscented Kalman Filter

URTSS ..... Unscented Rauch-Tung-Striebel Smoother

UT ..... Unscented Transform

Contents

1 Introduction **25**

1.1 Macromolecular X-ray crystallography (MX) 26

1.1.1 Producing a protein crystal 26

1.1.2 Crystals 28

1.1.3 The diffraction experiment 30

1.1.4 Understanding diffraction from a crystal 31

1.1.5 From diffraction patterns to electron density - The theory 36

1.1.6 From diffraction patterns to electron density - In practice 37

1.2 Small Angle X-ray Scattering (SAXS) 39

1.3 Limitations of MX and SAXS and the alternatives 43

1.4 Radiation damage in MX 44

1.4.1 Types of X-ray interactions with atoms: primary damage 44

1.4.2 Secondary damage 47

1.4.3 Quantifying energy absorbance: dose 48

1.4.4 Manifestations of damage: global damage 51

1.4.5 Manifestations of damage: specific damage 54

1.4.6 Experimental methods for dealing with radiation damage 56

1.4.7 Modelling intensity decay 58

1.5 Radiation damage in SAXS 63

1.6 This thesis in context 64

2 Dose Decay Modelling **67**

2.1 Introduction 68

2.2 Experimental methods 68
2.2.1 Considerations 68
2.2.2 Crystallization 70
2.2.3 Data collection and dose calculation 72
2.2.4 Data processing 75
2.2.5 Calculating the relative intensity 76
2.3 Dose decay models 77
2.3.1 Validity test 82
2.3.2 Obtaining model parameter values 83
2.4 DDM comparison results 87
2.4.1 Uniform irradiation 87
2.4.2 Calculating the RDE 87
2.5 Further investigation of the RDE Leal model 94
2.6 Incorporating the RDE into DWD calculations 97
2.6.1 Predicting intensity loss 98
2.6.2 Offset simulations 101
2.6.3 Offset experiment 102
2.7 Discussion 108

3 Zero-Dose Extrapolation 113

3.1 Introduction 114
3.2 Extracting intensities and doses 115
3.3 Extrapolation routine 118

3.3.1 Regression extrapolation 118

3.3.2 Probabilistic extrapolation 125

3.4 Results 132

3.5 Discussion 138

4 A Markovian Data Reduction Framework 143

4.1 Introduction 144

4.2 Why Julia? 144

4.3 A hidden Markov model of the data collection experiment 146

4.3.1 Mathematical Notation 147

4.3.2 Bayesian optimal filtering 148

4.3.3 Bayesian smoothing 149

4.3.4 Process function and covariance 150

4.3.5 Observation function and covariance 152

4.3.6 Obtaining parameter values for the process and observation functions 153

4.3.7 Convergence of the forward-backward algorithm 154

4.3.8 Summary of hidden Markov model formulation 156

4.4 Extraction and treatment of reflection intensity data 156

4.4.1 Allocating observations to images 156

4.4.2 Treatment of intensity data 157

4.5 Simulation results 160

4.5.1 Weak data 165

4.6 Protein structure results 169

4.6.1 Bovine pancreatic insulin 169

4.6.2 Protein-DNA complex - C.Esp1396I 180

4.7 Discussion 189

4.7.1 Overview of the forward-backward algorthm 189

4.7.2 Improvements and extensions 193

4.7.3 Future work 195

5 X-ray Beam Analysis 197

5.1 Introduction 198

5.2 Processing aperture measurements 198

5.2.1 Experimental methods 199

5.2.2 Deconvoluting the X-ray beam measurements 199

5.2.3 Results 208

5.3 2D X-ray beam profile measurements 209

5.3.1 PGM file preprocessing 210

5.3.2 RADDOSE-3D simulation results 214

5.4 Processing multiple 1D aperture scan measurements 220

5.4.1 Acquiring the 1D aperture scans 220

5.4.2 Creating a 2D beam profile 220

5.5 Discussion 222

6 Methods to Assess Radiation Damage in SAXS 227

6.1 Introduction 228

6.2 Extending RADDOSE-3D for SAXS 228
6.2.1 RADDOSE-3D architecture 229
6.2.2 Cylindrical sample geometry 229
6.2.3 Determining the sample composition 232
6.2.4 Attenuation of X-ray flux due to capillary 233
6.2.5 Summary of SAXS extensions 236
6.2.6 Model considerations 236
6.3 Experimental methods 239
6.3.1 Sample preparation 239
6.3.2 SAXS data collection 240
6.3.3 Dose calculations 243
6.4 1D scatter curve similarity analysis 243
6.4.1 Data analysis - experiment 1 245
6.4.2 Data analysis - experiment 2 248
6.5 Results 253
6.5.1 Experiment 1 253
6.5.2 Experiment 2 264
6.6 Discussion 273

7 Conclusions **279**

7.1 Radiation damage in MX 280
7.1.1 Extending the DWD metric 280
7.1.2 Zero-dose extrapolation 281

7.1.3 Measuring X-ray beam profiles 282

7.2 Data reduction 283

7.3 Quantifying radiation damage in SAXS experiments 284

List of Figures

1.1 MX structure solution pipeline. 27

1.2 Lattice definition. 28

1.3 Two dimensional Bravais lattices. 29

1.4 Insulin diffraction image. 31

1.5 X-ray scattering from two electrons. 32

1.6 Centrosymmetric electron cloud and scattering factor from carbon 34

1.7 SAXS data collection schematic and resulting radially averaged intensity curve. 41

1.8 Example Kratky curves and distance distribution function 42

1.9 Primary X-ray interaction processes of X-ray photons with atoms 46

1.10 Relative contribution to the overall X-ray interaction cross section as a function of incident beam energy for chicken egg-white lysozyme 47

1.11 Dose distributions in crystals represented as a polyhedron 50

1.12 Global radiation damage metrics 53

1.13 Transition dynamics of the populations in the Blake and Phillips radiation damage model 59

2.1 Relative diffraction efficiency (RDE) schematic 69

2.2 Crystals used in the experiment at PETRA III, Hamburg 71

2.3 Beam profile at beamline P14, PETRA III, Hamburg. 74

2.4 Flow diagram of the crystal reorientation process prior to data collection at PETRA III. 74

2.5 Resolution analysis table from the log file of an AIMLESS job. 78

2.6 Relative intensity plotted against the average absorbed dose for one insulin crystal. 78

2.7 *BEST* intensity curve. 84

2.8 Scale and B factor values plotted against the average dose over the whole crystal. 85

2.9 Structure of the objective function to determine parameter values for dose decay models. 86

2.10 Crystal states as predicted by the Sygusch and Alliare model applied to insulin data. 88

2.11 Dose metrics and dose state of crystal showing uniform irradiation. 89

2.12 Comparison of each dose decay model's ability to predict intensity decay for all insulin data. 93

2.13 Leal *et al.* RDE model fitted to all insulin relative intensity data. 94

2.14 Effects of resolution cut offs on Leal *et al.* model. 96

2.15 Increasing and decreasing η functions 97

2.16 Beam profiles used in the varying dose contrast regime experiment in Zeldin *et al.* (2013). 98

2.17 Relative intensity plots comparing DWD with different η forms. 100

2.18 Offset simulations comparing the diffracted dose efficiency against DWD with the different η forms. 105

2.19 Beam profile used in the offset experiment described in Zeldin *et al.* (2013). . . 106

2.20 Dose isosurface map for the crystal used in the offset experiment described in Zeldin *et al.* (2013). 107

2.21 Dose metric comparison: DWD, Average dose (whole crystal) and Maximum dose 108

3.1 Relative diffraction efficiency as a function of the dose. 116

3.2 Scaled relative intensity decay. 117

3.3 Poor regression fit: overfitting. 120

3.4 Regression fits to reflections that result in a low correlation coefficient. 121

3.5 Poor regression fits. 122

3.6 Logarithm of the zero-dose mean intensities in 14 resolution shells. 127

3.7 Logarithm of the zero-dose mean intensities in 14 resolution shells with theoretical means for a later dataset. 128

3.8 Smoothing spline interpolation of estimated scale factors against dose in a single resolution shell. 129

3.9 Scale factor weighting against the residual of the observed and mean intensities. 131

3.10 Regression fits that satisfied the regression criteria. 134

3.11 Prior, likelihood and (unnormalised) posterior distributions for two centric reflections. 135

3.12 Reflection intensities of 5000 reflections before and after zero-dose extrapolation. 137

3.13 Reflection intensities of 5000 reflections before zero-dose extrapolation and the spherically averaged zero-dose data have been calculated without contribution from the abnormally high intensity values. 138

4.1 Benchmark times taken to run a given algorithm for various languages relative to C. 145

4.2 Hidden Markov model representation of the diffraction experiment. 147

4.3 Propagation of states through a non linear function for different transformation techniques. 155

4.4 Model of the spherical cap traversed in a data collection experiment. 159

4.5 Forward-backward algorithm results for simulated data for a strong reflection. 164

4.6 Log likelihood values calculated at the end of each smoothing pass for a strong reflection. 165

4.7 Forward-backward algorithm results for simulated data for a weak reflection. . 167

4.8 Log likelihood values calculated at the end of each smoothing pass for a strong reflection. 168

4.9 Calculated B factors for each image in the insulin dataset. 170

4.10 Normality of calculated B factor distribution for the insulin dataset. 172

4.11 Calculated scale factors for each image in the insulin dataset. 173

4.12 Histogram of scale factors for insulin dataset. 174

4.13 Amplitude estimates for four different reflections observed in the insulin dataset using the forward-backward algorithm. 177

4.14 $2F_o$ - F_c electron density maps for insulin dataset. 179

4.15 Difference electron density map for insulin dataset. 180

4.16 Structure of the C.Esp1396I protein-DNA complex. 181

4.17 Calculated B factors for each image in the C.Esp1396I dataset. 182

4.18 Normality of calculated B factor distribution for the C.Esp1396I dataset. 183

4.19 Calculated scale factors for each image in the C.Esp1396I dataset. 185

4.20 Histogram of scale factors for the C.Esp1396I dataset. 186

4.21 Amplitude estimates for two different reflections observed in the C.Esp1396I dataset using the forward-backward algorithm. 187

4.22 $2F_o$ - F_c electron density maps for the C.Esp1396I dataset 188

4.23 Difference electron density map for the C.Esp1396I dataset. 189

5.1 Flux measurements collected using 1D aperture scans at the Diamond Light Source synchrotron. 200

5.2 A schematic of the X-ray beam and aperture setup. 201

5.3 Aperture scan measurements and the Gaussian fits to the data. 204

5.4 2D X-ray beam profile reconstructions. 206

5.5 Gaussian model fitted to the deconvoluted X-ray beam profile. 207

5.6 Modelling of the point spread function. 211

5.7 Beam image segmentation and beam boundaries. 213

5.8 Background pixels of the X-ray beam image. 215

5.9 Processed beam profiles used for simulations in RADDOSE-3D. 217

5.10 $D_{1/2}$ values plotted against the percentage of zero pixel values in the processed pgm beam images. 219

5.11 Diode readings from aperture scans across the beam collected at beamline BM29, ESRF. 221

5.12 Interpolation of diode readings from the aperture scans. 223

5.13 Final averaged X-ray beam profile used for the SAXS experiments. 224

6.1 Flow diagram illustrating the structure of the RADDOSE-3D code. 230

6.2 Implementation of the cylindrical sample geometry in RADDOSE-3D. 231

6.3 X-ray mass attenuation coefficient data for carbon from NIST. 235

6.4 Updated flow diagram illustrating the structure of the RADDOSE-3D code with the SAXS extensions. 236

6.5 Flow diagram summarising the SAXS extensions to RADDOSE-3D. 237

6.6 Effect of diffusive turnover on the dose in SAXS. 238

6.7 RADDOSE-3D dose contour plot of a cylindrical SAXS sample. 239

6.8 Diode readings and flux estimates during the first SAXS repeat for the 1 mg/ml GI sample with no radioprotectant added. 242

6.9 A 2D reconstruction of beam used in the second SAXS experiment 243

6.10 RADDOSE-3D input file used for the dose calculations in Expt 1. 244

6.11 Increasing dissimilarity of 1D SAXS curves with increasing X-ray exposure. . . . 245

6.11 Fidelity value as a function of time and dose. 247

6.12 Similarity comparison with selected frames from the first experimental repeat with no radioprotectant added 251

6.13 Longest observed edge length against the frame number for pairwise comparisons with frame 1. 252

6.14 Dose at which significant radiation damage is determined to have occurred for different values of m, the number of consecutive dissimilar frames. 257

6.15 Dose at which significant radiation damage is determined to have occurred for different values of α, the threshold probability value to determine frame similarity. 261

6.16 Fidelity values as a function of dose for each radioprotectant 262

6.17 Radiation damage onset threshold dose values for both metrics used to assess the fidelity values. 263

6.18 Radioprotectant efficacy comparison boxplots for each concentration in the experiment. 266

6.19 Radiation damage threshold frame against reference frame (m = 3) for the first repeat with DTT as the added radioprotectant at a concentration of 10 mM. . . 267

6.20 Heat map of all possible pairwise frame comparisons for the first repeat with 10 mM concentration DTT added 268

6.21 Comparing dose and frame number as the metric by which to track radiation damage in SAXS. 271

6.22 RD onset ratio against concentration for the 8 radioprotectants 272

List of Tables

2.1 Data collection strategy for each protein crystal type at PETRA III. 75

2.2 Data processing statistics for the first data set collected from each of the processed insulin crystals. 76

2.3 Pearson Correlation Coefficient (PCC) values for linearly transformed intensity data 83

2.4 Best fit zero dose average intensity values for each resolution bin for the Holton dose decay model. 93

2.5 Parameter values for the dose decay models. 94

2.6 Parameter values for Leal *et al.* model for data scaled to different resolution limits. 97

2.7 Squared Euclidean norm values representing the scatter of the relative intensity data from the line of best fit. 100

2.8 Comparison of $D_{1/2}$ and s_{AD} values using DWD with different η forms 101

2.9 Offset experiment design. 106

3.1 Average DWD and scaled relative intensity values. 132

3.2 Regression fit quality indicators. 133

4.1 Final refinement statistics for data processed with the ACT and FBA pipelines. . 178

4.2 Final refinement statistics for data processed with the ACT and FBA pipelines. . 189

5.1 Comparison of calculated FWHM values with experimentally observed FWHM values. 208

CHAPTER 1

Introduction

Determination of the three-dimensional (3D) structure of biological macromolecules is essential in developing our understanding of vital processes that occur within the cells of living organisms. This is because the 3D structure of a macromolecule is one of the major factors that determine its function (Berg *et al.*, 2002; Hegyi and Gerstein, 1999). However, macromolecules are too small to be observed under a light microscope (the typical diameter for a soluble protein monomer is about 3-6 nm (Philips, 2015), whereas light microscopes can only resolve objects that are at least a few hundred nanometres in size (Starr *et al.*, 2010)) so alternative methods must be used to probe the molecular structure. Several techniques to determine 3D molecular structures have been developed since the beginning of the 20th century. The research presented in this thesis is concerned with methods development for the structural techniques of macromolecular X-ray crystallography (MX) and small angle X-ray scattering (SAXS) applied to protein molecules. In particular, quantitative methods for assessing and correcting radiation damage for these techniques are explored. Thus the following sections set the theoretical framework on which this work is based.

1.1 Macromolecular X-ray crystallography (MX)

Macromolecular X-ray crystallography (MX) is by far the most common technique used for solving the atomic structure of 3D macromolecules. As of 14th June 2016 the Protein Data Bank (PDB) contained 119,480 structures and just over 89% of those structures were solved by MX. Despite the dominance of MX, there are many stages in the structure solution pipeline that present their own challenges. The typical protein structure solution pipeline in X-ray crystallography is outlined in Figure 1.1 (Garman, 2014).

1.1.1 Producing a protein crystal

When a particular protein target has been identified, the first step is to produce it in an appropriate expression system. *E.coli* is the most common system since it is cheap, fast and produces higher quantities of protein compared to other systems (Rai and Padh, 2001). However other expression systems such as yeast, insect and mammalian cells are used when bacterial systems are not suitable e.g. bacterial cells lack the capacity to perform certain types of post-translational modification such as glycosylation. The target protein is

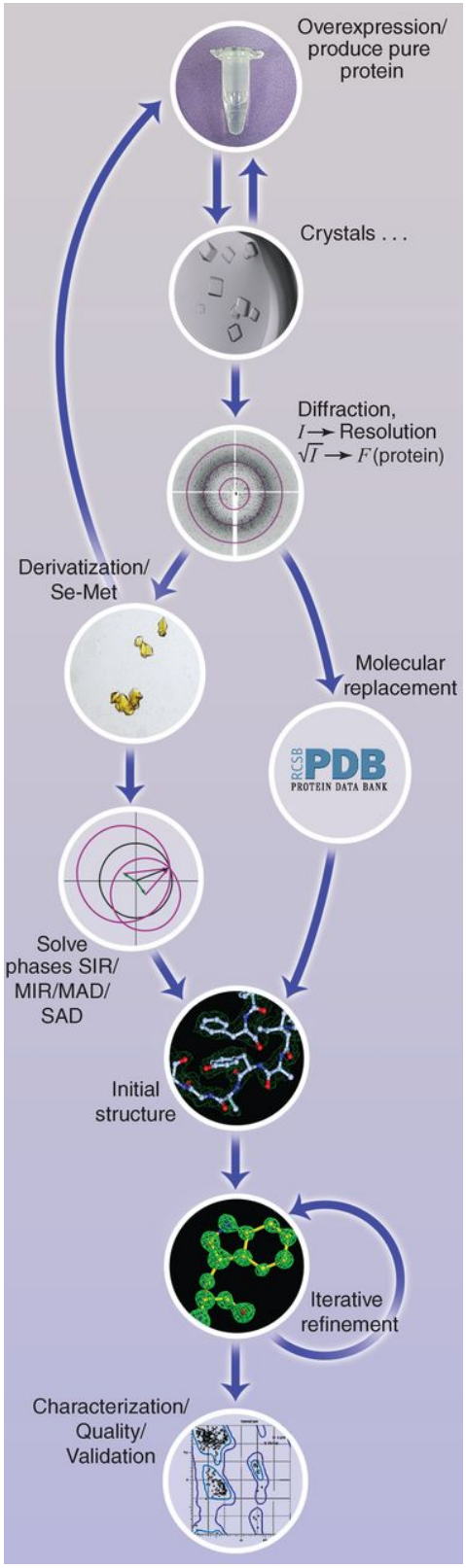
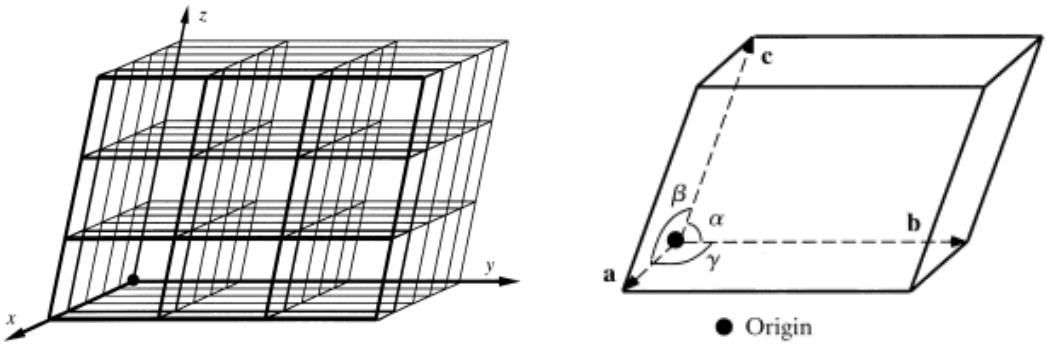


Figure 1.1: Typical protein structure solution pipeline in X-ray crystallography (Garman, 2014).

generally not the only product formed during the expression phase, so it has to be isolated. The process of isolating the target protein is known as purification, methods for which include filtration and chromatography (Gräslund *et al.*, 2008). The next step is crystallisation, which is generally regarded as the major bottleneck in crystallography (Garman, 2014). At this stage of the pipeline, a solution containing the purified target protein is mixed with a precipitant solution to achieve suitable chemical and environmental conditions for crystallogenesis. However, in general the exact conditions for crystallogenesis for a given protein are unknown, hence many laboratories use robots to screen multiple conditions in the hope of finding the correct one(s) (Luft *et al.*, 2007).

1.1.2 Crystals

A conventional crystal is essentially a 3D repeating array of identical subunits known as unit cells (Figure 1.2a). A unit cell is defined by six parameters - $a, b, c, \alpha, \beta, \gamma$ (Figure 1.2b). a, b, c represent the lengths of the edges of the unit cells and α, β, γ are the angles between them (Drenth, 2006).



(a) (b)

Figure 1.2: (a) Crystals are formed by repeating unit cells related by translations in all 3 dimensions. The intersection points of the lines are the lattice points. (b) A single unit cell with axes $\boldsymbol{a}, \boldsymbol{b}$ and $\boldsymbol{c}$, with angles between them α, β and γ. (Drenth, 2006)

If $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}$ are viewed as directional vectors (vectors will be generally represented by lowercase bold letters $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}$) instead of physical unit cell lengths, and only integer multiples of them are considered, then the resulting set of points is referred to as a (3D Bravais) lattice. Mathematically this is equivalent to:

$$\{n_1\boldsymbol{a} + n_2\boldsymbol{b} + n_3\boldsymbol{c} \mid n_1, n_2, n_3 \in \mathbb{N}\}, \tag{1.1.1}$$

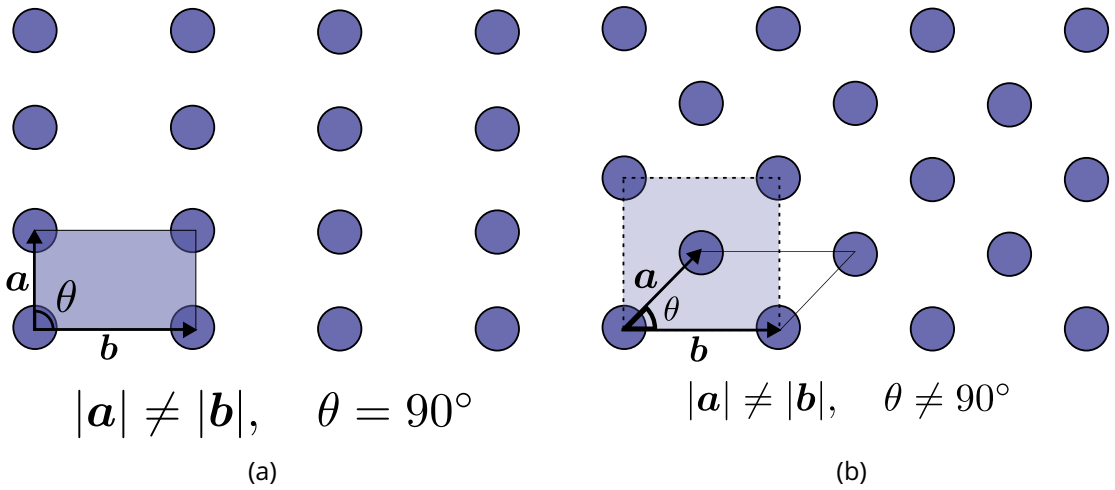


Figure 1.3: Two dimensional Bravais lattices. (a) Rectangular lattice system with no centering. (b) Centred rectangular lattice system.

where $\mathbb{N}$ denotes the set of natural numbers $\{$ 0, 1, 2, 3,...$\}$ (Note: sometimes the set of natural numbers is defined as excluding 0 but here it is included). The terms 'crystal' and 'lattice' are sometimes used interchangeably because they are very closely related. However, the distinction between the physical crystal and the hypothetical lattice is made clear by noticing that the set defined in equation 1.1.1 implies that the array of lattice points is infinite, whereas the physical crystal is not. Despite this distinction, the properties of lattices are applicable to crystals. In particular, the electron density, ρ, at a given point in the crystal, $\boldsymbol{r}$, is assumed to be invariant by a translation of a lattice vector, $\boldsymbol{t}$ (a lattice vector is any linear combination of unit lattice vectors $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}$). Explicitly this is

$$\rho(\boldsymbol{r}) = \rho(\boldsymbol{r} + \boldsymbol{t}), \quad \{\boldsymbol{t} = m_1\boldsymbol{a} + m_2\boldsymbol{b} + m_3\boldsymbol{c} \mid m_1, m_2, m_3 \in \mathbb{N}\}. \tag{1.1.2}$$

There are different types of lattices which are defined by their lattice (axial) system and their types (centerings) (examples given in Figure 1.3). In three dimensions there are 7 lattice systems (cubic, tetragonal, orthorhombic, monoclinic, triclinic, hexagonal rhombohedral), and 5 crystal centering types (primitive (P), base-centered (A, B, C), face-centered (F), Body-centered (I), Rhombohedrally-centered (R)), and these combine to make the set of 14 Bravais Lattices. (Note: a stricter way to define the 14 possible Bravais lattices is to realise that two lattices are equivalent if their symmetry groups are isomorphic. Bravais' original criterion only classifies 11 types of lattice (Pitteri and Zanzotto, 1996)).

Thus far, only translational symmetries have been considered for unit cells in a crystal.

These transformations move every point in space from their original position. However, Bravais lattices also possess symmetry operations that leave at least one point in space fixed. These are called the point group symmetries and consist of reflections and rotations. Protein molecules are chiral and hence symmetry operations that reverse the chirality of the molecule are not allowed for protein crystals (these correspond to the reflections - symmetry operations for which the determinant of the transformation matrix is equal to -1). Further to this, the crystallographic restriction theorem only allows for 2, 3, 4, and 6 fold rotations (Coxeter, 1973). This leaves only 11 enantiomorphic point groups allowed for protein crystals (Drenth, 2006).

The combination of fractional translations along a unit cell axis with rotations result in screw axis symmetry operations. When these operations are combined with the point group, the resulting set of operations is referred to as the space group. There are 65 enantiomorphic space groups allowed for protein crystals (Drenth, 2006).

1.1.3 The diffraction experiment

The final experimental stage of the structure solution pipeline is the data collection. Once a suitable crystal composed of the target protein has been grown, diffraction data are collected using a method that will also keep the crystal hydrated in the solution in which it was grown: the mother liquor. It is then mounted and irradiated with a beam of intense X-rays (energy usually between 6 and 18 keV), typically whilst being rotated, although other collection protocols are becoming increasingly common (e.g. helical scans). Over a small angular rotation range, the diffracted X-ray photons are collected on a position sensitive detector and produce a diffraction pattern (a diffraction image) that is unique to that crystal (Figure 1.4). The spots that are observed on the images are known as reflections. These observations are in fact the intensities of the reflections. The goal of the data collection experiment is to accurately measure the intensities of as many reflections (unique and multiplicitous) as possible. The space that allows indexing of these reflections is known as reciprocal space. Therefore to sample as much of reciprocal space as possible, multiple diffraction patterns are collected as the crystal rotates (often the number of images range in the hundreds). Understanding how the diffraction of the X-ray photons arise from their interaction with the atoms in the protein crystal allows crystallographers to interpret the pattern and

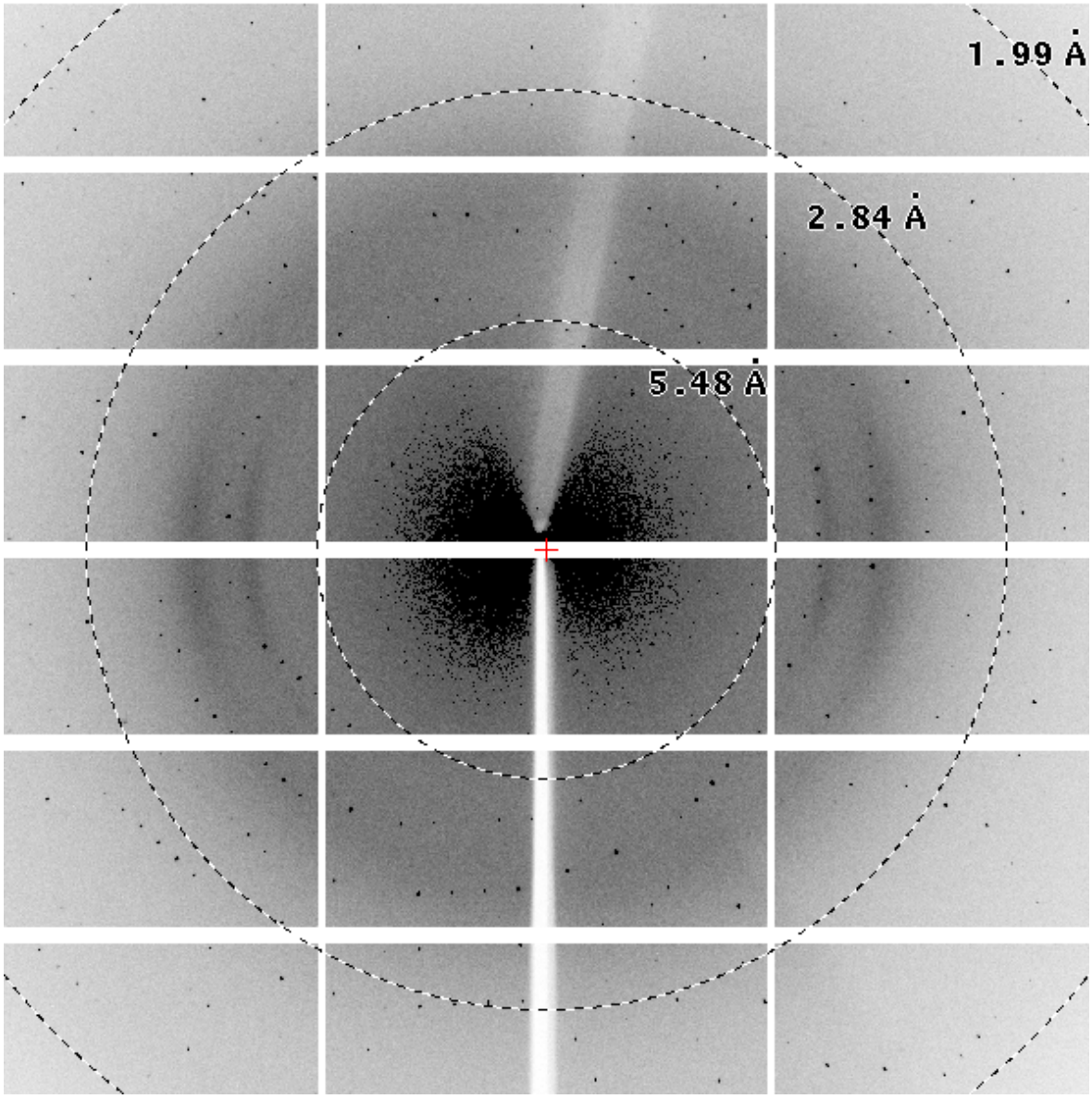


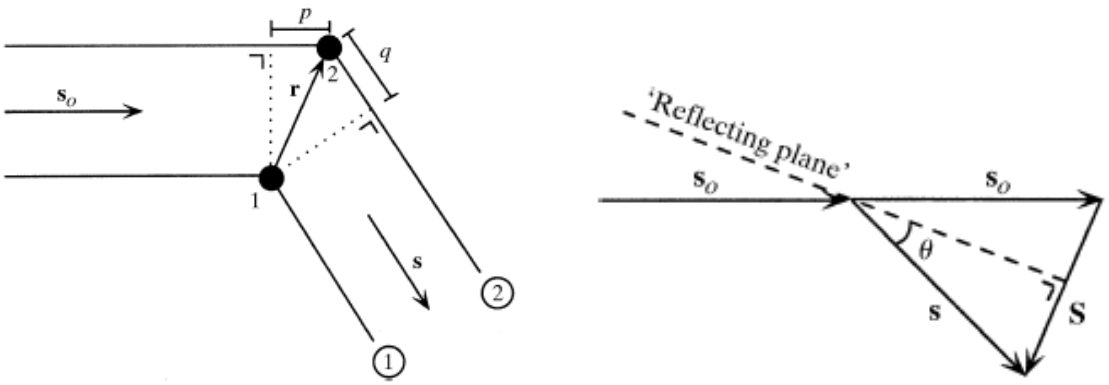
Figure 1.4: Section of a diffraction image from a crystal of bovine pancreatic insulin crystallised in space group $I2_13$ recorded in January 2014 on beamline P14 at the PETRA III synchrotron, Hamburg. The oscillation range for the image was 0.1° and the incident X-ray energy 12.7 keV. Individual spots known as reflections or Bragg peaks are clearly visible.

solve the atomic structure of the protein.

1.1.4 Understanding diffraction from a crystal

Scattering from a single electron

The diffraction of the X-ray beam by the crystal results from the interaction of the electric component of the X-ray beam with the electrons in the crystal. When an X-ray hits an electron, the electron begins to oscillate. When the scattering is elastic (the desired scattering



(a) (b)

Figure 1.5: (a) Scattering from a system of two electrons labelled 1 and 2. Incident beam direction is denoted $\boldsymbol{s}_0$ and the scattered beam direction is denoted $\boldsymbol{s}$. The origin is defined at the position of electron 1. Electron 2 is at a position $\boldsymbol{r}$ with respect to electron 1. The path length difference of the beam scattered by electron 1 and electron 2 is $p + q$. (b) The angle θ is the angle between the incident beam and the reflecting plane. $\boldsymbol{S} = \boldsymbol{s} - \boldsymbol{s_0}$ and is perpendicular to the reflecting plane. The lengths of the vectors $\boldsymbol{s}$ and $\boldsymbol{s_0}$ are arbitrary but are chosen to be $1/\lambda$ for convenience, where λ is the wavelength of the incident X-ray beam. This gives $|\boldsymbol{S}| = 2\sin(\theta)/\lambda$ (Drenth, 1999).

type for diffraction in MX), the original photon is absorbed by the electron which then emits an X-ray with the same wavelength as the incident X-ray photon, but with a phase shift of $180°$. The amplitude of the electric component, E_{el}, of the scattered X-ray photon at a distance r from a free (unbound) electron is given by

$$E_{el} = E_0 \frac{1}{r} \frac{e^2}{mc^2} \sin(\varphi), \tag{1.1.3}$$

where E_0 is the amplitude of the electric vector of the incident X-ray photon, e is the electron charge, m is the electron mass, c is the speed of light and $\sin(\varphi)$ is the proportion of the component of E_0 perpendicular to the direction of the scattering electron (Drenth, 2006).

Scattering from a two electron system

Suppose a beam of X-rays is incident on a system of two electrons where electron 1 is positioned at the origin and electron 2 is at position $\boldsymbol{r}$ relative to electron 1 as depicted in Figure 1.5a. The direction of the incident X-ray beam, $\boldsymbol{s}_0$, is altered after the scattering event, and the direction of the scattered beam is denoted $\boldsymbol{s}$. The lengths of these vectors can be arbitrarily chosen. The convenient choice is $|\boldsymbol{s}| = |\boldsymbol{s}_0| = 1/\lambda$ where λ is the wavelength of the incident beam. The path length difference of the beam scattered by electron 1 and electron 2 is $p + q = \lambda[\boldsymbol{r} \cdot (\boldsymbol{s}_0 - \boldsymbol{s})]$ (where p and q are defined as shown in Figure 1.5a). Thus the

second beam lags behind the first, and the resulting phase difference is

$$\frac{-2\pi}{\lambda} \times \lambda[\boldsymbol{r} \cdot (\boldsymbol{s}_0 - \boldsymbol{s})] = 2\pi \boldsymbol{r} \cdot \boldsymbol{S}, \tag{1.1.4}$$

where $\boldsymbol{S} = \boldsymbol{s} - \boldsymbol{s_0}$ (Drenth, 1999). Figure 1.5b is a graphical description showing

$$|\boldsymbol{S}| = 2\sin(\theta)/\lambda. \tag{1.1.5}$$

where θ is the angle of reflection of the incident beam by the reflecting plane. It is important to note here that the phase of a wave with respect to a different wave is dependent on the relative position of the electrons. This demonstrates the importance of the phases for generating accurate structural information (Taylor, 2003, 2010).

The amplitudes of the scattered waves from both electrons in the system are the same, they only differ in phase. If the amplitude of the scattered wave from an electron positioned at the origin (e.g. electron 1) is equal to 1, then the total scattered wave from the two-electron system shown in Figure 1.5a is $1+1\times\exp[2\pi i \boldsymbol{r}\cdot\boldsymbol{S}]$ where i is the imaginary number, $i = \sqrt{-1}$.

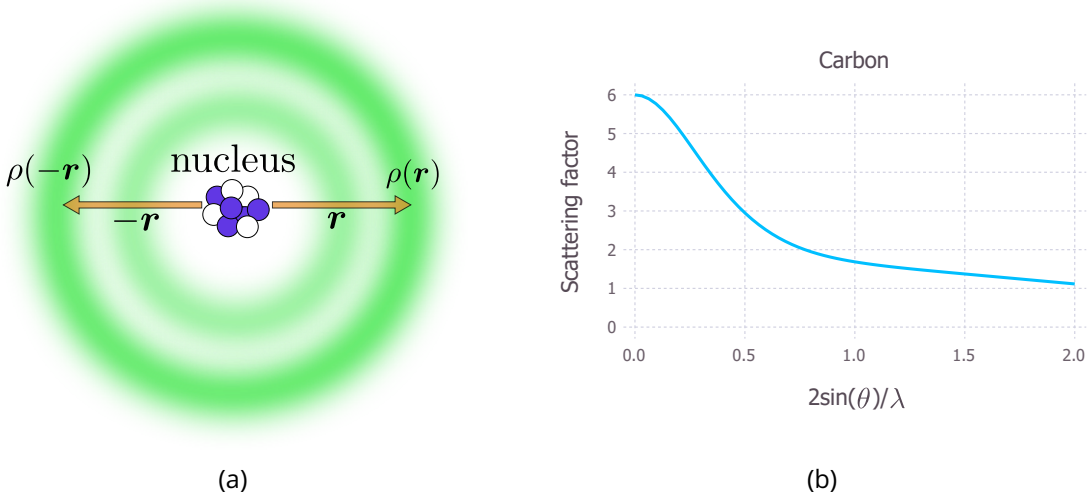
Scattering from an atom

The electron cloud of an atom scatters the incident X-ray beam. The strength of the scattered beam is dependent on the number of electrons and their positions in the electron cloud. The electrons in an atom are not free as was assumed in the original model of diffraction, but continuing to treat them as free electrons gives sufficient accuracy provided the wavelength of the incident beam is not too close to an absorption edge of the atom (Drenth, 2006). Setting the origin of the system at the centre of the atom, the total scattering from the atom, f, is calculated as

$$f = \int_{\boldsymbol{r}} \rho(\boldsymbol{r}) \exp[2\pi i \boldsymbol{r} \cdot \boldsymbol{S}] \, \mathrm{d}\boldsymbol{r}, \tag{1.1.6}$$

where $\rho(\boldsymbol{r})$ is the electron density at position $\boldsymbol{r}$ and the integral is over all space. f is known as the atomic scattering factor (Drenth, 1999).

The electron density of an atom is assumed to be perfectly centrosymmetric i.e. $\rho(\boldsymbol{r}) =$



(a) (b)

Figure 1.6: (a) An atom with a centrosymmetric electron density distribution (i.e. $\rho(\boldsymbol{r}) = \rho(-\boldsymbol{r})$) about the centre of the nucleus. (b) atomic scattering factor, f, of carbon as a function of $|\boldsymbol{S}| = 2\sin(\theta)/\lambda$. f is expressed in number of electrons. Notice an angle of $\theta = 0$, $f = 6$ corresponds to the atomic number of carbon. For this figure, $0^\circ \leq \theta \leq 90^\circ$ and λ = 1.0 Å.

$\rho(-\boldsymbol{r})$ (Figure 1.6a). Thus equation 1.1.6 can be simplified to

$$f = 2\int_{\text{half space}} \rho(\boldsymbol{r}) \cos[2\pi \boldsymbol{r} \cdot \boldsymbol{S}]\, \mathrm{d}\boldsymbol{r}. \tag{1.1.7}$$

The integral is only over half of the entire space and the quantity does not contain any imaginary terms, hence the atomic scattering factor is a real quantity.

The atomic scattering factor of an atomic is only dependent on the length of the vector $\boldsymbol{S}$ (equation 1.1.5). The dependence of the atomic scattering factor, f, on $|\boldsymbol{S}| = 2\sin(\theta)/\lambda$ is shown for carbon in Figure 1.6b.

Scattering from a unit cell

Imagine an atom placed in a unit cell where the origin of the system is placed at the origin of the unit cell. Now the atom has positional vector $\boldsymbol{r}$ to define its location. The scattering from this atom is now

$$\boldsymbol{f} = f \exp[2\pi i \boldsymbol{r} \cdot \boldsymbol{S}]. \tag{1.1.8}$$

Now suppose there are N atoms in the unit cell. The total scattering from the unit cell is now

$$\boldsymbol{F}(\boldsymbol{S}) = \sum_{j=1}^{N} f_j \exp[2\pi i \boldsymbol{r}_j \cdot \boldsymbol{S}]. \tag{1.1.9}$$

where $\boldsymbol{r}_j$ and f_j are the position and atomic scattering factor of the j-th atom respectively. $\boldsymbol{F}(\boldsymbol{S})$ is referred to as the structure factor because it depends on the structure composed of the various atoms within the unit cell (Drenth, 2006).

Scattering from a crystal

Suppose an origin is chosen to be at the corner of an arbitrary unit cell but the scattering from a different unit cell from the same crystal is to be calculated. Unit cells are related by translations along the unit cell basis vectors $\boldsymbol{a}, \boldsymbol{b}$ and $\boldsymbol{c}$ as described in section 1.1.2. So the unit cell of interest is located at position $t\boldsymbol{a}+u\boldsymbol{b}+v\boldsymbol{c}$, where $t, u, v \in \mathbb{Z}$ i.e. t, u, v are integers. Thus the scattering from this unit cell is

$$\boldsymbol{F}(\boldsymbol{S}) \times \exp[2\pi i t\boldsymbol{a} \cdot \boldsymbol{S}] \times \exp[2\pi i u\boldsymbol{b} \cdot \boldsymbol{S}] \times \exp[2\pi i v\boldsymbol{c} \cdot \boldsymbol{S}]. \tag{1.1.10}$$

If this is extended to include all unit cells then the total wave scattered by the crystal is

$$\boldsymbol{K}(\boldsymbol{S}) = \boldsymbol{F}(\boldsymbol{S}) \times \sum_{t=0}^{n_1} \exp[2\pi i t\boldsymbol{a} \cdot \boldsymbol{S}] \times \sum_{u=0}^{n_2} \exp[2\pi i u\boldsymbol{b} \cdot \boldsymbol{S}] \times \sum_{v=0}^{n_3} \exp[2\pi i v\boldsymbol{c} \cdot \boldsymbol{S}], \tag{1.1.11}$$

where n_1, n_2, n_3 are the number of unit cells along the direction of the basis vectors $\boldsymbol{a}, \boldsymbol{b}$ and $\boldsymbol{c}$ respectively. Since n_1, n_2 and n_3 are typically very large, the summations $\sum_{t=0}^{n_1} \exp[2\pi i t\boldsymbol{a} \cdot \boldsymbol{S}]$, $\sum_{u=0}^{n_2} \exp[2\pi i u\boldsymbol{b} \cdot \boldsymbol{S}]$ and $\sum_{v=0}^{n_3} \exp[2\pi i v\boldsymbol{c} \cdot \boldsymbol{S}]$ are usually equal to zero unless $\boldsymbol{a} \cdot \boldsymbol{S} = h$, $\boldsymbol{b} \cdot \boldsymbol{S} = k$ and $\boldsymbol{c} \cdot \boldsymbol{S} = l$ where $h, k, l \in \mathbb{Z}$. Therefore a crystal scatters X-rays if

$$\boldsymbol{a} \cdot \boldsymbol{S} = h, \tag{1.1.12a}$$

$$\boldsymbol{b} \cdot \boldsymbol{S} = k, \tag{1.1.12b}$$

$$\boldsymbol{c} \cdot \boldsymbol{S} = l. \tag{1.1.12c}$$

These are known as the Laue conditions (Drenth, 1999). The $h, k, l \in \mathbb{Z}$ are referred to as Miller indices and they define individual reflections (Figure 1.4). The result is that the amplitude of the total scattered wave from the crystal is proportional to the structure factor $\boldsymbol{F}(\boldsymbol{S})$ and the number of unit cells in the crystal (Drenth, 1999).

1.1.5 From diffraction patterns to electron density - The theory

The goal of structure determination is to obtain the atomic structure of the target molecule. Therefore given that it is understood how X-ray diffraction arises from the interaction of X-rays with a crystal, the inverse problem must be solved i.e. determining the atomic structure from the diffraction pattern.

Recalling equation 1.1.9 which allows the calculation of the structure factor as a summation over all atoms in the unit cell, it is possible to write this as an integral over the electron density in the cell instead, giving

$$\boldsymbol{F}(\boldsymbol{S}) = \int_{cell} \rho(\boldsymbol{r}) \exp[2\pi i \boldsymbol{r} \cdot \boldsymbol{S}] \, \mathrm{d}v. \tag{1.1.13}$$

Introducing fractional coordinates x, y, z (i.e. $0 \leq x < 1$ and similarly for y and z) and given the unit cell has volume V, the volume element, $\mathrm{d}v$, can be rewritten as

$$\mathrm{d}v = V \times \mathrm{d}x \, \mathrm{d}y \, \mathrm{d}z. \tag{1.1.14}$$

The position $\boldsymbol{r}$ can also be rewritten as $\boldsymbol{r} = \boldsymbol{a}x + \boldsymbol{b}y + \boldsymbol{c}z$, so this implies that

$$\boldsymbol{r} \cdot \boldsymbol{S} = \boldsymbol{a} \cdot \boldsymbol{S} \, x + \boldsymbol{b} \cdot \boldsymbol{S} \, y + \boldsymbol{c} \cdot \boldsymbol{S} \, z \tag{1.1.15}$$

$$= hx + ky + lz, \tag{1.1.16}$$

using the Laue conditions (equation 1.1.12). Thus $\boldsymbol{F}(\boldsymbol{S})$ can be written as a function of the Miller indices $\boldsymbol{F}(h, k, l)$ giving

$$\boldsymbol{F}(h, k, l) = V \int_{x=0}^{1} \int_{y=0}^{1} \int_{z=0}^{1} \rho(x, y, z) \exp[2\pi i(hx + ky + lz)] \, \mathrm{d}x \, \mathrm{d}y \, \mathrm{d}z. \tag{1.1.17}$$

Equation 1.1.17 shows explicitly that the structure factor $\boldsymbol{F}(h, k, l)$ is the Fourier transform of the electron density $\rho(x, y, z)$. The electron density in the unit cell can be obtained by taking the inverse Fourier transform of the structure factor

$$\rho(x, y, z) = \frac{1}{V} \sum_{h} \sum_{k} \sum_{l} |\boldsymbol{F}(h, k, l)| \exp -2\pi i(hx + ky + lz) + i\alpha(h, k, l), \tag{1.1.18}$$

where $|\boldsymbol{F}(h, k, l)|$ is the structure factor amplitude and $\alpha(h, k, l)$ is the phase. Equation

1.1.18 is known as the electron density equation and accurately calculating this is the ultimate goal of structure determination in MX. This is because it gives the density of electrons at every point in space in the unit cell. However, this equation can only be calculated once the amplitudes, $|\boldsymbol{F}(h,k,l)|$, and phases, $\alpha(h,k,l)$, are known. Notice that the summation is over reflections $(hkl\text{s})$ which explains why it is desirable to accurately obtain as many reflections as possible in the data collection experiment. The more (accurate) hkl terms that are used in the electron density equation, the better the Fourier series representation of the electron density in the unit cell (i.e. the better the electron density map). The amplitudes can be derived from the experimentally observed intensities in the data collection experiment but the phases are lost. This is known as the phase problem (Taylor, 2010).

1.1.6 From diffraction patterns to electron density - In practice

In practice, much of the theory is abstracted from the crystallographer and many software programs perform the necessary calculations 'behind the scenes'.

A set of diffraction images from a data collection experiment is known as a dataset. Extracting the amplitudes of each reflection in the dataset is the ultimate aim of the first series of programs. The first step is to calculate the intensities of the observations from the images. Programs such as MOSFLM (Leslie and Powell, 2007), XDS (Kabsch, 2010b) and more recently DIALS (Waterman *et al.*, 2013, 2016), are used to find the spots, index (which includes determining the unit cell dimensions and its orientation with respect to the beam and assigning Miller indices to each reflection) and finally integrate the intensity of the observations on the images. Commonly it is the case that a single observation is observed on multiple images because the rotation range of the crystal that results in a diffraction image (the oscillation angle) does not sample the entirety of a reflection. The fraction of the reflection that is recorded on a given image, the partiality of a reflection, has to be calculated and the experimental parameters refined in a process called post-refinement (Rossmann *et al.*, 1979; Rossmann, 1979).

Although the space group of the crystal is determined during the indexing stage, it is still only an informed prediction. An improvement of the predicted space group can be made after the integration stage with a CCP4 program called POINTLESS (Evans, 2011).

The intensities of reflections are affected not only by the number of electrons in the unit cell, but also by the variation of other systematic factors such as the rotation rate, incident beam intensity, the path length of the X-ray beam through the crystal, and the secondary absorption etc. (Evans, 2006). These factors must be taken into account so that the intensities of all reflections are on the same scale. To achieve this, the difference between intensities of reflections that should be identical according to the symmetry of the crystal are used to estimate the necessary parameters in a process called scaling. This is carried out by programs such as AIMLESS (Evans and Murshudov, 2013) and XSCALE (Kabsch, 2010b).

Once the intensity estimates have been put on an internally consistent scale, it is possible to derive the amplitudes. In theory the amplitudes are equal to the square root of the intensities. However the conversion is not as straightforward as this, because subtraction of background noise in the diffraction images during the integration step leads to weakly measured intensity observations having negative values. This is a problem because it has been established that the amplitude should be a real quantity. One way to deal with negative intensity reflections is to set the intensity values to zero. However French and Wilson developed a treatment that uses Bayesian analysis (French and Wilson, 1978) and Wilson statistics (Wilson, 1949) to calculate a better estimate of the amplitude given that a negative intensity has been observed. Programs such as CTRUNCATE (Evans, 2011) perform this analysis and calculate further data assessment statistics to check for crystal pathologies such as twinning.

Although phases are not observed directly in the diffraction experiment, they can be derived in multiple ways. If the sequence of the target protein is similar to another structure that has already been solved ($\approx$ 25-35% (Taylor, 2010; Abergel, 2013)) or the structure is thought to be similar to another structure, then initial phases can be taken from the homologous structure. The homologue structure is transformed via a series of rotations and translations within the asymmetric unit of the target crystal so that the calculated Patterson map best matches the one that has been obtained experimentally (McCoy, 2007). Programs such as PHASER perform these operations (McCoy *et al.*, 2007).

If no structural homologue exists, then the phases must be derived experimentally. Isomorphous replacement is a method in which experimental data are collected on the native crystal and then again with a crystal in which a heavier element has been incorporated without changing the arrangement of the other atoms in the unit cells (Perutz, 1956). The phases

can then be derived by analysing the difference in the intensity observations (Taylor, 2010). Another method known as anomalous dispersion can also be used, in which a heavy element is incorporated into the structure (sometimes this is not necessary if the structure already contains enough elements such as sulphur or heavier) and multiple datasets are collected at different X-ray photon energies. If possible, some of the datasets are collected at close to the absorption edge of the heavy element so appreciable differences between the Bijvoet pair can be measured (Bijvoet, 1954). If only 1 dataset is collected, this is known as single wavelength anomalous dispersion (SAD) and differences in intensities are analysed between Friedel pairs ([h, k, l] and [-h, -k. -l]) of reflections. If datasets are collected at different incident X-ray energies, the method is known as multiple wavelength anomalous diffraction/dispersion (MAD) (Hendrickson, 1991; Taylor, 2010).

With both amplitudes and phases obtained, it is possible to perform the Fourier transform defined in equation 1.1.18 to obtain an electron density map. The atomic structure can then be built into the resulting map. In general the initial map is unlikely to provide a satisfactory structure, so a process of refinement is carried out to improve the agreement of the amplitudes calculated from the model that has been built, with those that were derived from the experimental intensities. Programs such as REFMAC (Murshudov *et al.*, 2011) and PHENIX.REFINE (Adams *et al.*, 2010) perform these operations until a satisfactory structure has been obtained. An atomic model is determined to be satisfactory when the statistical quantities, R_{work} and R_{free}, are low enough. To ensure that the final structure is reliable, software programs such as MOLPROBITY (Chen *et al.*, 2010) and PROCHECK (Laskowski *et al.*, 1993) are used to check the stereochemical quality of the structure.

1.2 Small Angle X-ray Scattering (SAXS)

Small angle X-ray scattering (SAXS) is a technique used to determine the overall shape and size of a macromolecule, which like X-ray crystallography, requires a sample containing the molecule to be irradiated with a beam of X-rays. However, unlike X-ray crystallography, SAXS does not require the growth of crystals. Instead, the protein solution is irradiated and the scattered X-rays are collected on a position sensitive detector. In solution, the movement of molecules is not restricted and they can adopt random orientations with respect to one another (Blanchet and Svergun, 2013). Therefore the data observed on the detector are not

individual Bragg peaks. Although SAXS does not provide atomic level detail of a structure, it does give information about the interatomic distances between atoms, as well as an overall envelope of the structure and the molecular weight of the molecule (Pollack, 2011).

The X-ray scattering intensity is circularly symmetric (the intensity varies with radius) and hence it is determined by taking a radial average of the intensities recorded on the detector image (Franke *et al.*, 2015). The radially averaged intensity is usually written as a function of the momentum transfer, q, which itself is defined in terms of the scattering angle:

$$q = \frac{4\pi \sin(\theta)}{\lambda} \tag{1.2.1}$$

The final scattering pattern is the result of subtracting the radially averaged intensity of the solution containing the protein from the radially averaged intensity of the solution containing the buffer without protein (Figure 1.7).

Several graphs are produced during the analysis of SAXS data which can give much information about the state of the protein. For example, Kratky plots give indications of whether or not a protein is in a folded state (Figure 1.8a). This is because Debye's equation, (equation 1.2.2), which describes the scattering intensities, I_{Gauss}, for molecules behaving as Gaussian coils, plateaus in a $q^2 \times I(q)$ vs q plot within a limited range of data.

$$I_{Gauss}(q) = 2I_0 \frac{\exp(-u) + u - 1}{u^2}, \tag{1.2.2}$$

where I_0 is incident X-ray intensity and

$$u = q^2 R_g^2, \tag{1.2.3}$$

where R_g is the radius of gyration.

The Fourier transform of the SAXS pattern gives the pair-distance distribution function, a function that provides information about the interatomic distances in the protein:

$$p(r) = \frac{r^2}{2\pi^2} \int_0^\infty q^2 I(q) \frac{\sin(qr)}{qr} \, \mathrm{d}q. \tag{1.2.4}$$

where r is the distance between electrons in the molecule. Examples of distance distribution functions for various geometrical shapes are shown in Figure 1.8b.

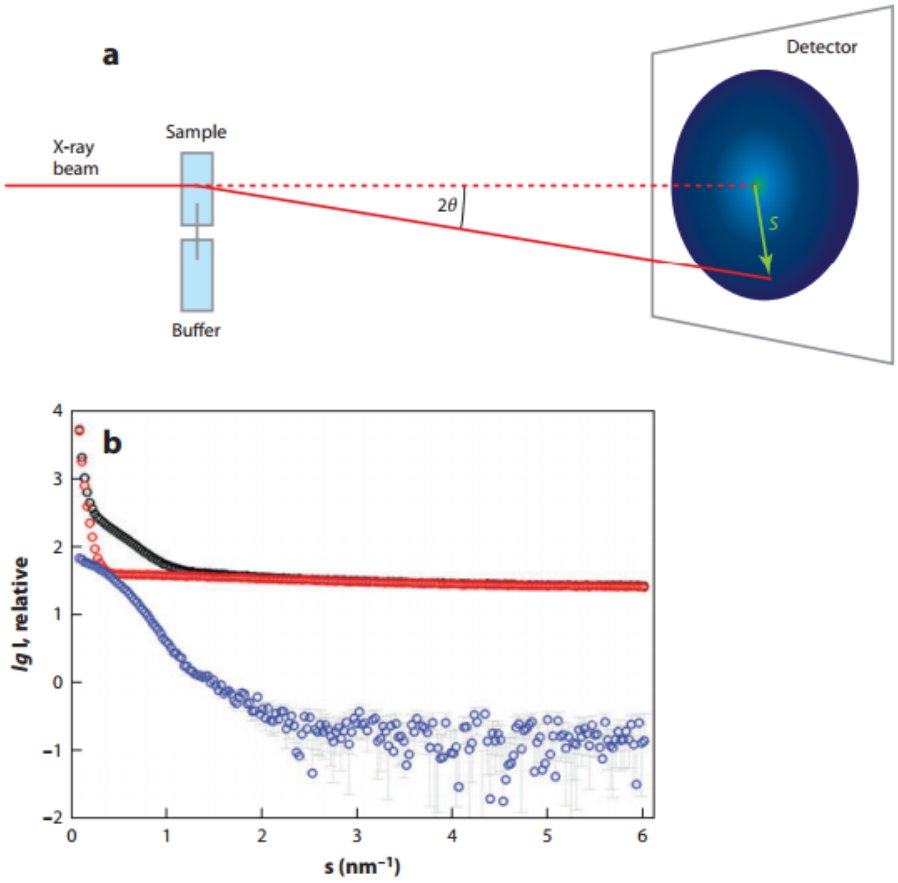
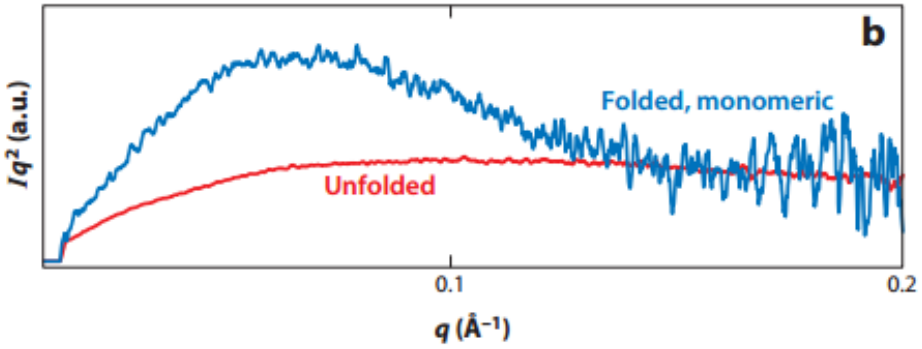
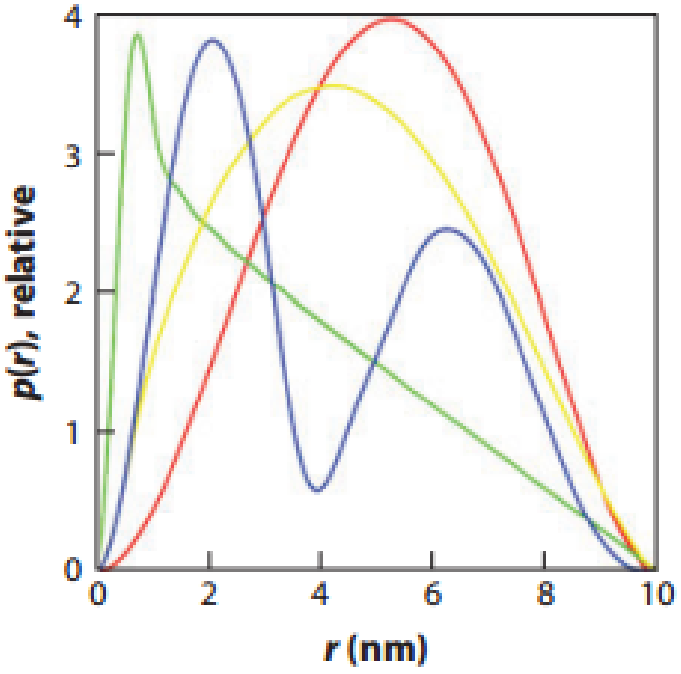


Figure 1.7: (a) X-ray beam (typical energies between 7 and 12.5 keV (Hopkins and Thorne, 2016)) is incident on a SAXS sample and the scattered radiation is collected on a detector. The symbol s in the figure is equivalent to the momentum transfer denoted q in equation 1.2.1. (b) Radially averaged intensity curves from a solution of bovine serum albumin (BSA). The radially averaged intensity of the solution containing only the buffer (red curve) is subtracted from the radially averaged intensity of the solution containing BSA and buffer (black curve) to obtain the resulting intensity curve for BSA (blue curve) (Blanchet and Svergun, 2013).



(a)



(b)

Figure 1.8: (a) Kratky plots. The red curve shows the unfolded RNA domain P4-P6 from the *Tetrahymena* ribozyme where the salt concentration was low. The blue curve implies the domain was folded and monomeric (Pollack, 2011). (b) Distance distribution functions for several geometrical volumes: sphere (red), dumbbell (blue), cylinder (green) disk (yellow) (Blanchet and Svergun, 2013).

1.3 Limitations of MX and SAXS and the alternatives

Despite the dominance of MX as a method for structural analysis and the recent emergence of SAXS as a reliable method to provide complementary structural and functional information, both methods have drawbacks. It has already been mentioned that producing protein crystals is a bottleneck in the pipeline in MX (section 1.1.1). Even if a crystal can be grown it must be of a sufficient size and quality to give good quality diffraction. Often it is the case that there is a limit to the size and quality of the crystal that can be grown. Hence the hardware and software for data collection is constantly being further developed to handle these situations. Theoretical work on the minimum crystal size for which sufficient data can be collected from a single crystal in a typical synchrotron radiation experiment (Holton and Frankel, 2010) has inspired work at various synchrotron beamlines to achieve this limit. Another drawback of MX is the fact that it is essentially a method that only provides a temporal and spatial average of the molecules in the crystal, and hence no dynamical information is readily obtained from the experiment. However a relatively new technique making use of the Hadamaard transform is being developed that promises to give dynamic structural information about the molecule (Yorke *et al.*, 2014).

With regards to SAXS, the major disadvantage is that it does not provide atomic resolution information. Thus relatively subtle structural dynamics of molecules will not be uncovered by using this technique.

Several alternatives to MX and SAXS exist. These include the use of X-ray free electron lasers (XFELs) and single particle cryo-electron microscopy (cryo-EM). In XFEL experiments, very intense beams of X-rays irradiate crystals over a very short time period (about 10^{12} photons in a single pulse in about 10 femtoseconds (fs) (Chapman *et al.*, 2011)). XFELs can obtain high resolution diffraction information from nanocrystals. However the technique is still relatively immature. The data processing is not trivial (the crystals not being rotated during exposure and unknown distributions of X-ray beam wavelengths per pulse are just a couple of the factors complicating data processing), and access to XFELs is very limited. Cryo-EM, like SAXS, does not require the production of crystals but can additionally provide atomic resolution detail due to recent hardware and software advances (Bai *et al.*, 2015). In cryo-EM a set of projection images of single molecules is taken with an electron microscope (af-

ter some sample preparation). The images are then computationally combined to obtain a density distribution of the molecule (Milne *et al.*, 2013). Again, cryo-EM comes with its drawbacks. Obtaining high resolution structures from small (< 200-300 kDa), unstable or flexible molecules is not trivial and the technique is not yet accessible enough (Bai *et al.*, 2015).

1.4 Radiation damage in MX

A common theme amongst all of the structural techniques mentioned thus far is that the sample is probed with ionising radiation. This results in the sample suffering radiation damage. Even samples studied at XFEL sources have also shown signs of radiation damage when a pulse longer than 30 fs is used (Nass *et al.*, 2015). Being the most mature of these structural techniques however, a greater volume of literature exists about radiation damage for synchrotron based MX. Radiation damage is a problem because it limits the amount of useful data that can be collected from a single crystal during an MX experiment (Garman, 2010). It is therefore the major cause of unsuccessful data collection at synchrotron sources given a sufficiently well diffracting crystal (Zeldin *et al.*, 2013a).

1.4.1 Types of X-ray interactions with atoms: primary damage

In a fairly typical MX experimental set-up, only a small fraction of the incident X-ray photons will interact in any way with the atoms in the crystal. Garman reported that for a 100 μm thick protein crystal only 2% of the incident photons of a 12.4 keV (1 Å) incident beam will interact with it (Garman, 2010). This section will explain the possible interactions that occur and how they can give rise to the data or result in radiation damage, a phenomenon which is the focus of this thesis.

Elastic scattering (Figure 1.9a)

Elastic scattering (also referred to as Rayleigh, Thompson or coherent scattering) is a type of interaction in which no energy from the incident X-ray photon is deposited in the sample (Nave, 1995). The resulting scattered waves interfere to give rise to the observed diffraction pattern. It is this type of interaction that the experimenter would like to maximise. However,

of the 2% of 12.4 keV incident X-ray photons that interact with the sample, only about 8% of these result in elastic scattering (Ravelli and Garman, 2006).

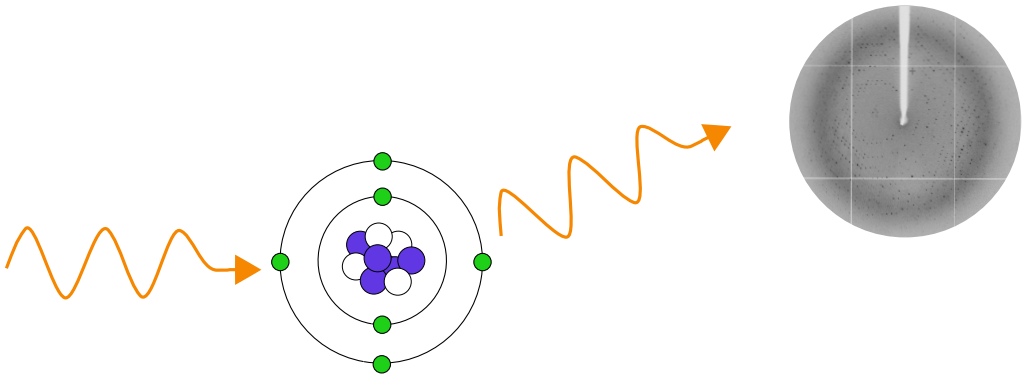
Compton Scattering (Figure 1.9b)

Another 8% of the interaction is due to Compton scattering. This occurs when the X-ray photon scatters incoherently from the crystal, thereby transferring some of its energy to an electron. The resulting scattered X-ray photon leaves with less energy (higher wavelength). It is possible for the recoil electron (the name given to the electron in which the energy is deposited) also to be ejected from the atom (Nave, 1995).

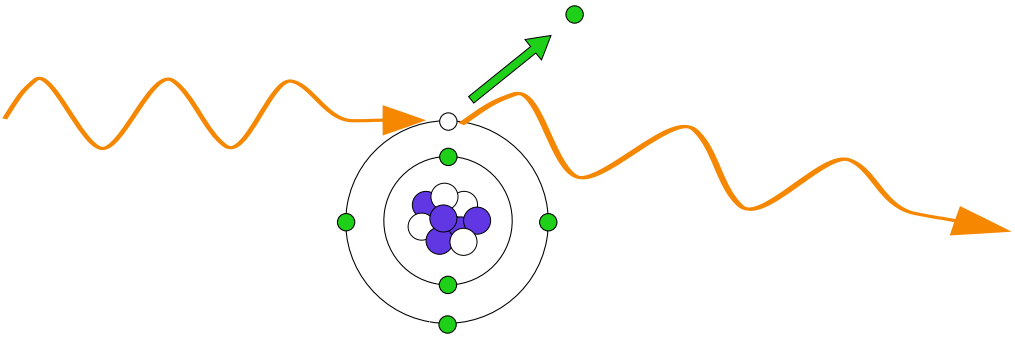
Photoelectric effect (Figure 1.9c)

By far the most prevalent interaction is the photoelectric effect which is responsible for the other 84% of the X-ray-electron interactions. Here the incident X-ray photon is completely absorbed and an inner shell electron is ejected (Garman, 2010). The ejected electron is called a photoelectron. The vacancy left by the ejection of the photoelectron is filled by another electron. This transition can lead to two different outcomes: either the production of a characteristic X-ray, known as fluorescence, or the ejection of an (Auger) electron from an outer shell (Nave, 1995). The relative probability of either fluorescence or Auger emission occurring is dependent on the atomic species. With lighter elements that constitute the vast majority of protein molecules, by far the more likely process is Auger emission.

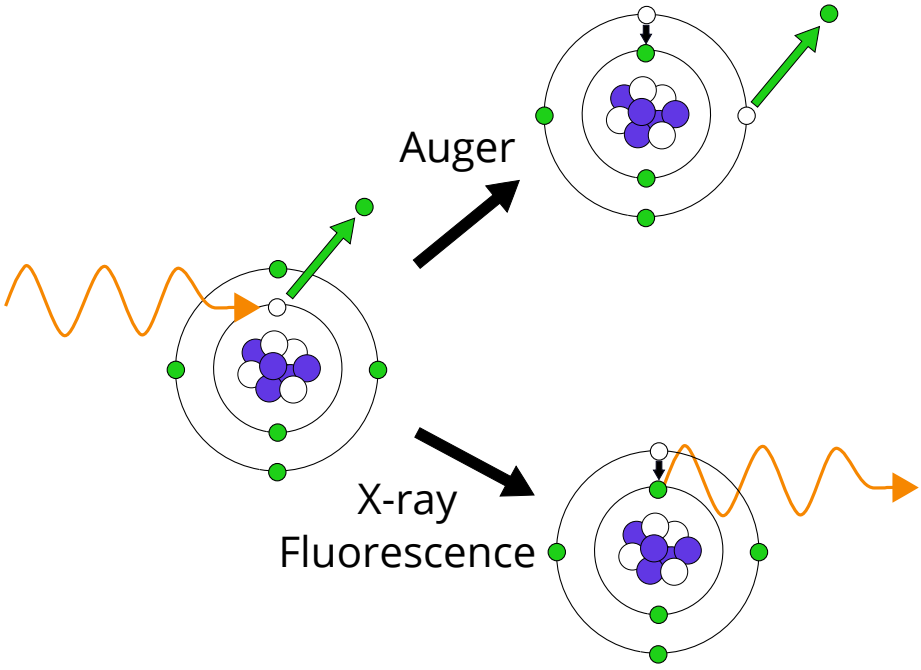
An initial ionisation event caused by the photoelectric effect or Compton scattering is referred to as primary damage, although the precise use of this term varies in the literature (Garman, 2010). The relative probability of each of the three interactions (cross section) described above varies with incident X-ray beam energy (Figure 1.10). The atomic number also affects the amount of energy absorption in the sample, with heavier elements contributing more to the overall absorption per atom. To maximise the amount of elastic scattering it would initially seem advisable to increase the energy of the incident X-ray beam from the range of typical values used in X-ray crystallography, usually around 12 keV (Paithankar and Garman, 2010). However, the diffracted intensity per incident photon is lower for incident beams with higher energy. In practice, the optimal energy for an MX experiment is highly de-



(a) Elastic Scattering



(b) Compton Scattering



(c) Photoelectric effect

Figure 1.9: Primary X-ray interaction processes of the X-ray photons with the atoms of the crystal and solvent. (a) Elastic scattering. There is no energy deposited in the crystals and these photons are scattered, resulting in the observed diffraction pattern. (b) Compton scattering. Some photon energy is transferred in the sample which contributes to the absorbed dose and results in emission of a photon with a longer wavelength. This can sometimes cause the emission of a recoil electron. (c) Photoelectric effect. All of the energy is absorbed by the atom in the sample resulting in the emission of an inner shell electron (the photoelectron). An outer shell electron then occupies the vacancy which can either result in Auger electron emission or emission of a fluorescent X-ray.

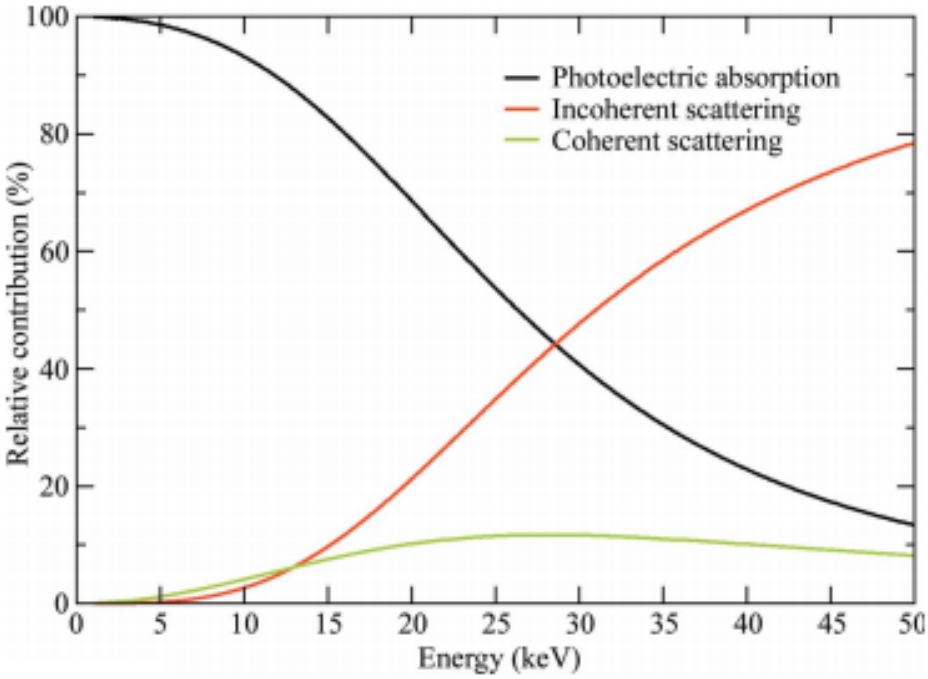


Figure 1.10: Relative contribution to the overall X-ray interaction cross section as a function of incident beam energy for chicken egg-white lysozyme (Paithankar and Garman, 2010). The photoelectric effect dominates at typical wavelengths used in MX (about 1 Å$\approx$ 12 keV). However, at higher energies the elastic contribution increases but the Compton effect increases faster and begins to dominate.

pendent on the sample composition and the aims of the experiment (i.e. the design may be different if the experiment aims to collect suitable data for phasing) and no general consensus on the best incident X-ray energy has been established (Paithankar and Garman, 2010).

1.4.2 Secondary damage

In contrast to primary damage (ionisation due to the incident photon), secondary damage is the damage propagated by the energetic photoelectrons or fluorescent X-rays produced as a result of the former. In fact, it is secondary damage that is responsible for the majority of the damage that is observed in MX experiments. A single photoelectron has enough energy to cause $\approx$ 500 ionisation events (O'Neill *et al.*, 2002). The ionisation event is referred to as direct damage if the event occurs within the protein molecule. If the event occurs within the surrounding mother liquor then it is referred to as indirect damage. The radical species produced as a result of these events can diffuse through the crystal disrupting the structure of the proteins by breaking chemical bonds, causing redox processes and producing more radical species (Meents *et al.*, 2010). Some of the radical species produced are hydroxyl radicals, electrons and hydrogen ions (Garman, 2010). At temperatures below $\approx$ 180 K (the

glass transition) it is thought that the only mobile radical species are electrons (Jones *et al.*, 1987) due to the solvent possessing a quasi-infinite viscosity (Weik *et al.*, 2001). Electrons are able to overcome energy barriers by quantum mechanical tunnelling mechanisms (Garman and Nave, 2009) and cause the majority of the damage observed in 100 K MX experiments (Garman, 2010).

1.4.3 Quantifying energy absorbance: dose

The first systematic study of radiation damage in protein crystals was carried out on sperm whale myoglobin in 1962 by Blake and Phillips (Blake and Phillips, 1962). They determined that radiation damage progression is proportional to the absorbed dose. Therefore being able to quantify the amount of energy absorbed by the crystal is vital to track the progression of radiation damage in an MX experiment (Blake and Phillips, 1962; Holton, 2009). Dose is the metric by which the energy absorbed in the crystal is quantified. It is defined as the energy absorbed per unit mass and the SI unit used in MX is the gray (Gy), where 1 Gy = 1 J kg^{-1}. In typical cryo-MX experiments it is common for a crystal to receive a dose of the order of millions of gray (MGy) (Garman, 2010) due to the brilliance of third generation synchrotron sources (Mitchell *et al.*, 1999). However, due to the complex processes (e.g. the stochastic nature of the absorption events, electron cascades, etc.), the dose cannot be measured, so it must be calculated using the parameters of the experiment instead. The formula used to calculate the dose for a small volume of sample is

$$Dose = \frac{E_{incident}}{M_{vol}} \left(1 - e^{-\mu_{abs} z}\right), \tag{1.4.1}$$

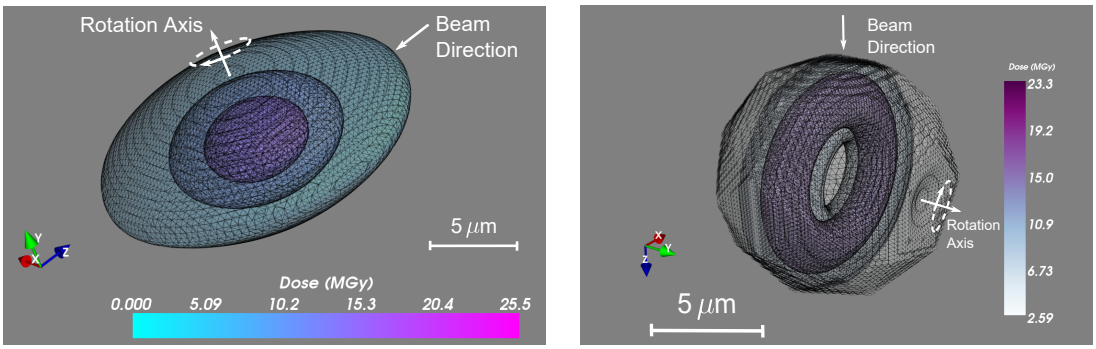
Where $E_{incident}$ is the energy incident on the sample, M_{vol} is the mass of the irradiated volume and $1 - e^{-\mu_{abs} z}$ is the fraction of the incident beam absorbed through the volume of thickness z, μ_{abs} is the absorption coefficient of the sample (a value dependent on the atomic cross sections that determines the likelihood of absorbance of the X-ray beam) and e = 2.718... is Euler's number (Zeldin, 2013).

RADDOSE is a software program that was developed to improve the ease of calculating the dose absorbed by a protein crystal (Murray *et al.*, 2004). It increases the accuracy by implementing a 2D model of the experiment as opposed to the 1D model described in equation

1.4.1. Information about the crystal (size, unit cell parameters, number of molecules per unit cell, number of residues, solvent content), the beam (energy (keV), flux (photons per second), size (μm^2) and intensity profile), exposure time per image and the total number of images, are required for the calculation. RADDOSE uses these parameters to determine the maximum flux within the beam profile and simulates an experiment in which the crystal is exposed homogeneously to the maximum beam intensity to provide a worst case scenario dose estimate: the maximum dose. Several improvements to the model were incorporated in successive versions of RADDOSE. Version two (Paithankar *et al.*, 2009) included calculation of the fluorescent emission probability of the constituent atoms and the probability that these fluorescent photons could escape the crystal. Version three (Paithankar and Garman, 2010) took into account the energy loss in the crystal due to the Compton effect, which had been neglected in previous versions, but only significantly affects the dose values for incident energies above 25 keV

Despite these improvements, RADDOSE still has several limitations. One of the major problems with the dose calculation is that some variables in the data collection experiment are known with more certainty than others. Parameters such as the beam flux and exposure time are known to a relatively high degree of accuracy. However, the extent to which parameters such as the crystal volume (Holton, 2009) and beam profile (Krojer and von Delft, 2011) are known is often significantly less. This can lead to large errors in the calculated dose values and inferences about the radiation susceptibility of protein crystals (Krojer and von Delft, 2011). The uncertainty can be accounted for by assigning a “factor of 2” decision threshold (Holton, 2009). This means that a particular radiation damage effect described by quantitative values (e.g. intensity loss or dose) can be considered the same if the values are within a factor of 2 of one another. Methods are being developed to better parametrise the data collection experiment for the crystal shape (Svensson *et al.*, 2015; Khan *et al.*, 2012; Brockhauser *et al.*, 2008) and the beam (Bowler *et al.*, 2015) but these methods are not yet routine on the majority of beamlines.

Another inaccuracy in the calculation of dose in RADDOSE is the two-dimensional nature of the model. It lacks the three-dimensional information about the damage state of the crystal. This is important for optimising the use of crystal volume to collect better quality data (Zeldin *et al.*, 2012, 2013a). Furthermore, RADDOSE assumes that the crystal is always totally immersed within the beam and does not take into account crystal rotation. With



(a) (b)

Figure 1.11: Dose distributions in crystals represented as a polyhedron. (a) An oblate ellipsoid (x = z = 10 μm, y = 6 μm) crystal that has been irradiated with a Gaussian profile beam with full width half maxima (FWHM) of 5 $\mu m \times 5\,\mu m$, flux of 10^{10} photons per second and an incident photon energy of 12.7 keV for a rotation of 360°. The collimation was set to 20 $\mu m \times 20\,\mu m$ and hence the entire volume of the crystal was irradiated. The 3 visible contours represent dose values of 20 MGy, 15 MGy and 5 MGy from inner to outer contour levels. (b) An icosahedral crystal with maximum dimensions of 10 μm in each of the x, y and z dimensions, which has been irradiated with a Gaussian profile beam with FWHM 4 μm in both the z and y directions, flux 2×10^{10} photons per second and an incident photon energy of 12.7 keV. The collimation was again set to 20 $\mu m \times 20\,\mu m$. The rotation axis was offset 5 μm from the centre of the crystal in the x-direction and the crystal was rotated 360° for a total exposure time of 180 seconds. Contouring levels correspond to 2.59 MGy, 12.5 MGy and 18.75 MGy from inner to outer contour levels. It is evident that different crystal, beam and experiment design parameters lead to very different distributions of absorbed dose within a protein crystal.

the size of X-ray beams becoming smaller over the last 10 years, it is commonly the case that crystals are no longer completely immersed in the beam, meaning that new, previously unirradiated parts of the crystal could rotate into the beam during the experiment. Thus it is now more critical to model crystal rotation.

This geometric issue is addressed in the RADDOSE-3D (Zeldin *et al.*, 2013b) software program, a successor to RADDOSE. RADDOSE-3D takes into account the three-dimensional geometry of the MX experiment to provide a temporally and spatially resolved dose distribution within a protein crystal. Initially RADDOSE-3D was written to model only cuboid or spherical crystals, but now it has been extended to be able to capture any polyhedral shape (Figure 1.11).

The raw output of RADDOSE-3D is a 3D scalar field which assigns a dose to every voxel (3D volume element) in the crystal. This output is not very useful for the experimenter to assess the damage state of the crystal, especially when the dose distributions are highly inhomogeneous (Figure 1.11). Thus several metrics were proposed to provide useful summaries of the data. These include the maximum dose (the same as the dose value as output by the original versions of RADDOSE), average dose for the whole crystal (AD-WC), and the dose inefficiency (i.e. the maximum dose divided by the total absorbed energy), amongst others (Zeldin *et al.*, 2012). The metric that was found to be most promising in faithfully represent-

ing the damage state of the crystal was the diffraction weighted dose (DWD) (Zeldin *et al.*, 2013a). The DWD is a weighted average, where the weight at each voxel position in the crystal is given by the X-ray fluence through that voxel. It considers the effective dose absorbed by the crystal and its impact on the diffraction pattern for any given image of the dataset. Mathematically the DWD is defined for each image as

$$DWD^i = \frac{\int_{t_{i-1}}^{t_i} \int_{crystal} D(\boldsymbol{x}, t) \times F(\boldsymbol{x}, t) \, \mathrm{d}\boldsymbol{x}\mathrm{d}t}{\int_{t_{i-1}}^{t_i} \int_{crystal} F(\boldsymbol{x}, t) \, \mathrm{d}\boldsymbol{x}\mathrm{d}t} \tag{1.4.2}$$

where i is the image number, t is the time, $\boldsymbol{x}$ is the position in the crystal, D is the total cumulative absorbed dose (Gy) at that position, and F is the fluence (photons per unit area).

The DWD metric only takes into account the fluence incident on a unit volume during the collection of data for a given image. However as the dose increases, a particular volume of the crystal will only contribute to the background (Blake and Phillips, 1962). The definition of DWD in equation 1.4.2 does not take into account this loss of diffraction efficiency of the crystal volume due to global radiation damage.

1.4.4 Manifestations of damage: global damage

The damage processes described in sections 1.4.1 and 1.4.2 are observed in a variety of ways in MX, and are classified accordingly. Global radiation damage is observed in reciprocal space and results in changes in mosaicity, increase in the scaling and Wilson B factors, unit cell volume expansion, increases in data quality R values ($R_{merge}, R_{meas}, R_{pim}$) and decreasing $CC_{1/2}$ and CC^* (Garman, 2010). Perhaps the most iconic symptom of global radiation damage is the loss of intensity of reflections in the diffraction pattern, with the reflections corresponding to higher resolution information fading the quickest. Global radiation damage is particularly problematic in data collection for experimental phasing where multiple datasets are collected for comparison of reflection amplitudes. For experimental determination of phases in isomorphous replacement or anomalous scattering experiments, it is necessary to distinguish intensity changes of around 4% (Taylor, 2010). However it was calculated that for a 0.5% change in all unit cell dimensions or a 0.5° rotation of the molecule about a single axis through the centre of gravity of the molecule both perpendicular or parallel to a crystallographic axis of a 100 Å^3 unit cell, there would be a change in the intensity of

a general 3 Å reflection by 15% and 16% respectively (Crick and Magdoff, 1956). Changes in the intensity due to non-isomorphism caused by global radiation damage can exceed these limits, thereby swamping the phasing signal and ultimately hindering structure solution.

To observe appreciable changes in reciprocal space, radiation damage must be affecting the long range crystalline order of the crystal (Meents *et al.*, 2010). At room temperature, many more radical species are mobile and damage occurs much more quickly than at cryo-temperatures (100 K) (Henderson, 1990; Nave and Garman, 2005; Weik and Colletier, 2010). However, global radiation damage still occurs at cryo-temperatures and it has been proposed that this is due to hydrogen gas build up. This is because diffusion rates are smaller at lower temperatures. This build up of hydrogen is thought to exert a disruptive force on the crystalline structure leading to the reduction in structural integrity (Meents *et al.*, 2010).

The parameters affected by the various symptoms, or variants of them, have been proposed as metrics to assess the extent of radiation damage as a function of the dose (Figure 1.12). An ideal metric would change reproducibly and monotonically with increase in dose. This rules out using the mosaicity as a suitable metric as it violates both of these criteria (Garman, 2010). Unit cell volume expansion was once thought to be a suitable metric (Ravelli *et al.*, 2002), however it was determined later that it was not reliable because the expansion was not consistent amongst crystals of the same protein and the same size irradiated under identical conditions (Murray and Garman, 2002). Three promising metrics that seem to adhere to the criteria (within a factor of 2) are: the relative intensity (Owen *et al.*, 2006), the relative B factor (Kmetko *et al.*, 2006) and the decay R factor (Diederichs, 2006).

Relative intensity

The relative intensity is defined as I_n/I_1, where I_n is the summed mean intensity of a complete data set n (or equivalent sections of data) after a dose D, and I_1 is the mean intensity of the first data set (Garman, 2010). This metric is desirable because it directly analyses the decay of the experimental data: the intensities. Furthermore it exhibits a linear dependence on the dose for a suitably low dose (Owen *et al.*, 2006; Zeldin *et al.*, 2013a). An experimental dose limit of 30 MGy was established based on the observation that crystals of proteins with different compositions decay at the same rate at cryo-temperatures (Owen *et al.*, 2006). At this dose the relative intensity of the crystals had decayed to 70% of their initial value, at

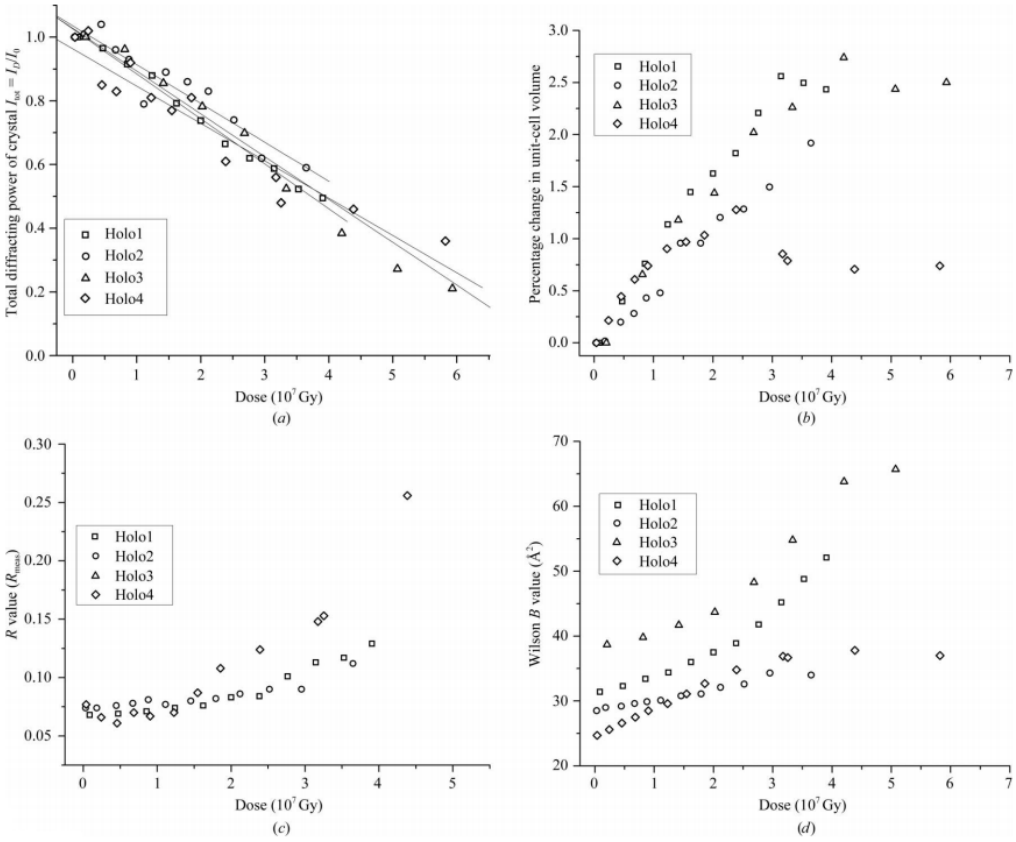


Figure 1.12: Various global radiation damage metrics as a function of dose for four holoferritin crystals. (a) Relative intensity, I_n/I_1. (b) % change in unit cell volume. (c) R value (R_{meas}). (d) Wilson B value. Work performed by Owen *et al.* (2006), reproduced from Garman (2010).

which point it was observed that the structural information that can be obtained from the sample has been significantly compromised (Owen *et al.*, 2006; Blundell and Johnson, 1976).

Relative B factor

The relative B factor, $B_{rel} = B_n - B_1$, is a metric that calculates the difference between the isotropic B factor for the current dataset, n, and the initial dataset collected on the same crystal (Kmetko *et al.*, 2006). Various studies have shown that B_{rel} increases linearly with dose (Kmetko *et al.*, 2006; Borek *et al.*, 2007; Bourenkov and Popov, 2010; Leal *et al.*, 2012). Another closely related metric which is derived from B_{rel} is the coefficient of sensitivity to absorbed dose, $s_{AD} = \Delta B_{rel}/\Delta D 8\pi^2$ where ΔB_{rel} and ΔD are the change in relative B factor and dose respectively. The advantage of s_{AD} is that it is robust (to within a factor of 2) for cryo-cooled protein crystals of differing molecular weights and solvent contents (Kmetko *et al.*, 2006).

Decay R factor

The decay R factor, R_d is a decay metric that assesses the pairwise difference between symmetry related reflections on different images (Diederichs, 2006). This results in a relatively smooth function against the difference in image number between two measurements (used as a proxy for dose) that can be interpreted as follows: if R_d stays constant for each image then no significant radiation damage has occurred, whereas if the R_d increases then radiation damage is progressing throughout the dataset. A desirable feature of this metric is that it can be used with a single dataset. Therefore radiation damage analysis can be performed on highly sensitive crystals that may only withstand enough dose for a partial or single dataset. The problem is that it requires high multiplicity which is not always achievable. This metric is less commonly used in radiation damage studies than the other two.

1.4.5 Manifestations of damage: specific damage

In contrast to global radiation damage, specific radiation damage is classified as damage that is observed in real space in the electron density maps and occurs up to $\approx$ 60 times faster (Holton, 2009). It is characterised by what appear to be specific chemical changes to

the molecule and occurs in a reproducible order at particular sites (Ravelli and McSweeney, 2000; Weik *et al.*, 2000; Gerstel *et al.*, 2015):

1. Metallo-centres are reduced at doses as low as 45 kGy (Owen *et al.*, 2011).
2. Disulphide bonds are elongated and broken (Burmeister, 2000; Ravelli and McSweeney, 2000).
3. Glutamate and aspartate residues are decarboxylated (Burmeister, 2000; Weik *et al.*, 2000; Ravelli and McSweeney, 2000).
4. The $S^{\delta} - C^{\epsilon}$ bond of methionine residues is cleaved (Burmeister, 2000).
5. Covalent bonds to metal atoms are cleaved (Ramagopal *et al.*, 2005).

Dehydroxylation of tyrosine residues has also been observed (Burmeister, 2000), but the physical basis and the evidence of this being a general symptom of specific radiation damage is contentious (Charlie Bury and Elspeth Garman, personal communication).

Due to the fact that specific damage changes the structural information derived from the model, it can lead to incorrect biological interpretation from structures. Metrics to assess the progression of specific damage also exist but are not always described as a function of dose.

Peaks in Fourier difference maps

Fourier difference maps, $F_n - F_1$, are electron density maps where the amplitudes of the first dataset F_1 are subtracted from the amplitudes of a later dataset F_n. Negative values correspond to areas where electron density has been lost as the experiment progressed. Sometimes this can be due to noise in the data, but significantly high values (difference peaks) that are coincident with the structural model from the first dataset suggest specific structural damage has occurred. In most studies, this analysis is carried out via manual inspection of the electron density map (Burmeister, 2000; Weik *et al.*, 2000; Ravelli and McSweeney, 2000). However work towards automating this analysis is progressing well, and new metrics are being developed to better characterise the level of specific damage (Bury *et al.*, 2015). For performing radiation damage analysis with Fourier difference maps, it is usual to not refine

the phases for the later datasets. Instead the phases from the first dataset are used for the later datasets.

Atomic B-factors

Atomic B-factors (also referred to as atomic displacement parameters) describe the level of dynamic disorder of atoms, with each atom having its own B-factor. Generally this value increases as radiation damage progresses and the level of disorder of the structure increases. In practice the B-factor is also substantially influenced by other factors such as static disorder (occupancy), errors in model building and the packing density of the atom (Gerstel *et al.*, 2015). This means that atoms with higher isotropic and anisotropic B-factors may not necessarily be more damaged than atoms with a smaller B-factor.

B damage

B damage, B_{damage}, is a metric that aims to deconvolute the dependence of the atomic B factor on its packing density. It is defined as the ratio of an atom's B-factor to the average B-factor of atoms that have a similar packing density environment (Gerstel *et al.*, 2015). This is a normalised metric, therefore atoms that show damage will have B_{damage} values significantly greater than 1.

Analysis of 2 PDB structures with B_{damage} showed a positive correlation between radiation damage and solvent accessibility (Gerstel *et al.*, 2015), in agreement with statements made in a previous study (Sygusch and Allaire, 1988). However other work has claimed otherwise (Coquelle *et al.*, 2007; Homer *et al.*, 2011), demonstrating the lack of consensus on metrics and on conclusions in radiation damage studies.

1.4.6 Experimental methods for dealing with radiation damage

Due to the challenges presented by radiation damage in MX to achieve successful structure determination, various parameters of the experiment have been varied in an attempt to mitigate radiation damage effects.

Temperature

It has been known for decades that diffraction experiments carried out at cryo-temperatures (down to 100 K) improves the lifetime of biological samples (Henderson, 1990; Brooks-Bartlett and Garman, 2015) by around $\sim$ 25 - 110 times over room temperature (RT) experiments (Southworth-Davies *et al.*, 2007). Cryo-cooling requires careful protection of the sample to prevent freezing, because ice also exhibits a crystalline structure that will diffract coherently in the MX experiment. Furthermore, ice has 7% greater volume than liquid water so the formation of ice can increase the disorder of the molecules in the crystal. Cryo-protection techniques have been developed since the late 1980s (Garman and Schneider, 1997; Hope, 1988; Teng, 1990) and cryo-crystallography is now more routinely used than RT (Garman, 2014). The main problem with cryo-cooling a crystal is that it can decrease the crystalline order (Nave and Garman, 2005) although this is not always the case (Garman, 1999).

There is evidence to suggest that using temperatures below 100 K using an open flow helium cryostat may further improve the lifetime of crystals and the data quality (Meents *et al.*, 2010; Teng and Moffat, 2002). However, whether the benefits of cooling to temperatures of 40 K or below are significant enough to warrant the costs required to achieve those temperatures routinely is still not obvious (Weik and Colletier, 2010).

Dose rate

Damage processes occur over a finite time period and this has been exploited by increasing the rate at which the dose is deposited in the crystal (dose rate) in an attempt to outrun radiation damage. At cryo-temperatures increased dose rates have been shown to decrease crystal lifetimes by up to 10% (Owen *et al.*, 2006) although some report that there is no effect at all (Sliz *et al.*, 2003). At RT there is growing evidence that a higher dose rate increases crystal lifetimes. (Southworth-Davies *et al.*, 2007; Owen *et al.*, 2012, 2014). With the increasing popularity of *in situ* data collection of crystals in their growth trays (Axford *et al.*, 2015, 2012), particularly for protein crystals that are not amenable to cryo-cooling, high dose rate experiments are likely to become more routine.

Serial femtosecond crystallography (SFX) at XFELs use extremely fast X-ray pulses (as short as 5 fs (Boutet *et al.*, 2012)) to probe protein crystals to obtain “diffraction before destruc-

tion" (Chapman *et al.*, 2014). However, experimental data have now shown indications of radiation damage at XFELs if the pulse length is too long (Nass *et al.*, 2015). SFX at XFELs is still in its infancy but progress, is continuously being made and it is a promising field for the development of structural biology (Garman, 2014; Brooks-Bartlett and Garman, 2015).

Scavengers

Scavengers are small molecules introduced into the crystal to mitigate the effects of secondary damage by intercepting or interacting with radicals to make them less reactive. At present, multiple studies have reported conflicting evidence regarding the efficacy of scavengers (Barker *et al.*, 2009; Kmetko *et al.*, 2011). It is worth noting that these studies used different global radiation damage metrics (I_n/I_1 improved with use of scavengers, whereas B_{rel} did not) to come to their conclusions. The effect of scavengers to protect against specific radiation damage is more promising (Southworth-Davies and Garman, 2007). A comprehensive summary of all scavenger literature to date is provided in Allan *et al.* (2012).

1.4.7 Modelling intensity decay

Structure factor amplitudes are derived from the integrated intensity measurements obtained from the experimental data. Therefore accurate intensity values are required to obtain reliable structural information. The problem is that the intensity measurements are affected by several systematic factors (Evans, 2006). In particular, global radiation damage to the sample can be a significant effect contributing to the overall intensity variation. Thus attempts to correct for it have been proposed and implemented during data processing.

In the radiation damage study carried out by Blake and Phillips in 1962 (Blake and Phillips, 1962), repeated intensity measurements of a subset of reflections along the 010 zone were taken with increasing X-ray exposure. It was noted that there was a general decrease in the intensity values, with reflections corresponding to higher resolution decaying quickest. If the variation was due solely to thermal disorder then the intensity decay was expected to obey the form

$$I = I_0 \exp\left(-\frac{B\sin^2(\theta)}{\lambda^2}\right), \tag{1.4.3}$$

where I is the experimentally measured intensity, I_0 is the initial intensity, B is a measure

of the disorder, θ is the diffraction angle and λ is the wavelength of the incident beam. However this form did not explain the observation that the intensity decay flattened out at high θ angles. Therefore a compartmental model of the crystal was proposed. The three states that were used to describe the evolution of the crystal damage were

1. A_1 - an undamaged fraction of the crystal which contributes to diffraction at all angles.
2. A_2 - a highly disordered fraction that only contributes to diffraction at low angles.
3. $A_3 = 1 - (A_1 + A_2)$ - an amorphous fraction that no longer diffracts coherently.

These three populations were suggested to contribute to the diffracted intensity, I, at time t as

$$\frac{I(t)}{I(0)} = A_1(t) + A_2(t) \exp\left(-\frac{B \sin^2(\theta)}{\lambda^2}\right). \tag{1.4.4}$$

To explain the crystal population dynamics (i.e. how the crystal transitions between damaged, highly disordered and amorphous states) it was assumed that radiation damage effects are irreversible. The transitions are described according to Figure 1.13 where k_1, k_2 and k_3 are rate constants to be determined.

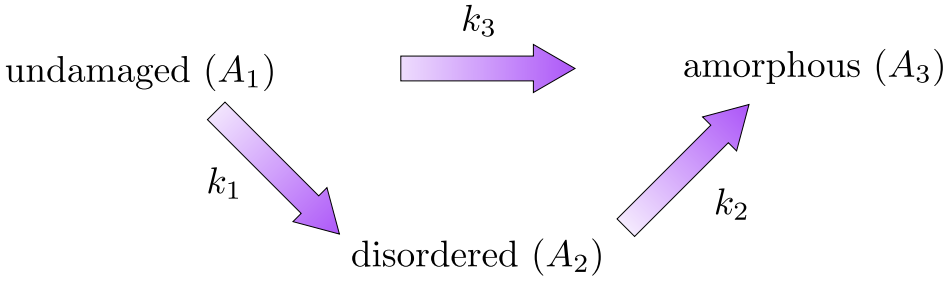


Figure 1.13: Transition dynamics of the populations in the Blake and Phillips model. Note here that the states are irreversible and it is possible to transition directly from the undamaged fraction to the amorphous fraction.

Intensity measurements corrected using equation 1.4.4 generally agreed well with the measurements from the relatively undamaged crystal.

The Blake and Phillips model was also used to correct for radiation damage in a study on crystals of lamprey haemoglobin (Hendrickson *et al.*, 1973). 150 reflections were observed in order to determine the change in intensity as a function of dose. The data agreed well with the model of Blake and Phillips.

In 1976, Fletterick *et al.* proposed a modification to the Blake and Phillips model for their

study on glycogen phosphorylase *a* crystals (Fletterick *et al.*, 1976). The modification disallowed a direct transition from the undamaged state to the amorphous state (i.e. $k_3 = 0$ in Figure 1.13). This gives the sequential transition model

$$A_1 \xrightarrow{k_1} A_2 \xrightarrow{k_2} A_3. \tag{1.4.5}$$

In 1976, Hendrickson studied all possible transitions from the Blake and Phillips model (Figure 1.13) assuming irreversible transition states (Hendrickson, 1976). This included $k_2 = k_3$ (Hendrickson *et al.*, 1973), $k_1 = k_2$ and $k_3 = 0$, $k_3 = 0$ (Fletterick *et al.*, 1976), $k_1 + k_3 = k_2$ and the arbitrary time-dependence model ($k_1 \neq k_2 \neq k_3$). These models were fitted to the data from the Blake and Phillips myoglobin study (Blake and Phillips, 1962). It was concluded that all models explain the data well for moderate dose values ($\approx$ 26% decrease in relative intensity), but none explain radiation damage at high doses very well ($\approx$ 72% decrease in relative intensity). The general model resulted in refined k_3 values that were very close to zero, and the model with $k_3 = 0$ consistently resulted in the best fit to the data at higher dose values. This suggests that the radiation damage process is always sequential.

In 1988 Sygusch and Allaire proposed an extension to the Fletterick *et al.* model. They suggested that a fourth fraction, A_1', that has only suffered small perturbations of surface residues and disulphide bonds, still has the capability of contributing to diffraction that occurs at all angles (Sygusch and Allaire, 1988). This is because these surface perturbations are caused before there are any significant changes in molecular conformation. The resulting sequential model is expressed as

$$A_1 \xrightarrow{k_0} A_1' \xrightarrow{k_1} A_2 \xrightarrow{k_2} A_3. \tag{1.4.6}$$

This model was fitted with data collected on crystals of rabbit skeletal muscle aldolase. The authors show that the proposed model provides a very good fit for radiation induced intensity decay even at high doses.

In 2009 James Holton proposed a resolution dependent model of intensity decay for a reflection (Holton, 2009). This model considers the dose absorbed by the crystal and describes

the intensity decay of a reflection, I, as

$$I = I_0 \exp\left(-\ln(2)\frac{D}{Hd}\right) \tag{1.4.7}$$

where I_0 is the intensity at zero dose, D is the absorbed dose in MGy, d is the resolution of the reflection in Å corresponding to the distance between successive Bragg planes, and H is the Howells *et al.* criterion which proposed a resolution dependent dose limit of 10 MGy per Å of resolution (Howells *et al.*, 2009). As an example, if a resolution of 3 Å is required, the Howells criterion suggests that the dose limit of the crystal would be 30 MGy. This model was shown to produce results in agreement with those from Owen *et al.* and Kmetko *et al.*.

In 2012 Leal *et al.* modified a traditional scaling model to account for radiation damage (Leal *et al.*, 2012). The scaling model is defined as

$$J(D, \boldsymbol{h}) = J(h_d) \times K(D) \times \exp\left(-B(D)h_d^2/2\right) \tag{1.4.8}$$

where $J(D, \boldsymbol{h})$ is the expected intensity of a given reflection with Miller indices $\boldsymbol{h} = hkl$ after the crystal has absorbed a dose D, and $J(h_d)$ is the expected reflection intensity at reciprocal distance $h_d = |\boldsymbol{h}| = 1/d$ from the origin in the absence of any radiation damage. $K(D)$ and $B(D)$ are the scale and B-factors respectively, and both are dependent on the absorbed dose. The authors assumed a linear model of B-factor increase given by

$$B(D) = B_0 + D\beta, \tag{1.4.9}$$

where B_0 and β are empirical constants to be determined. The functional form of the scale factor, K, was empirically deduced as

$$K(D) = C \exp\left(-\gamma^2 D^2\right), \tag{1.4.10}$$

where C and γ are empirical constants to be determined. The D^2 in equation 1.4.10 means that the model proposed by Leal *et al.* exhibits Gaussian decay behaviour. This model was shown to agree well with the overall diffraction intensity decay for data collected on 15 different crystals at RT to a relative intensity down to 60% of the original value.

Despite the differences between these models, they each rely on an accurate calculation of

the dose absorbed by a crystal. Of particular note, the models proposed by Sygusch and Allaire and Leal *et al.* allow for a delayed intensity decay. For the Leal *et al.* model, this behaviour follows from the Gaussian form of the equation. The Sygusch and Allaire model was shown to exhibit this behaviour by fitting the model to experimental data (Owen *et al.*, 2014).

Current scaling methods employ a scaling model in the same form as equation 1.4.8 to place multiple datasets on the same scale (Kabsch, 2010a). However the functional forms of the scale and B-factors are not necessarily assumed to take those given by equations 1.4.9 and 1.4.10. The scaling of intensities on images within a dataset also exhibits a relationship similar to equation 1.4.8. Take, for example the scaling model used in AIMLESS:

$$g = g_1(\varphi) \times g_2(\boldsymbol{s}_2) \times \exp\left(\frac{-2B(\varphi)\sin^2(\theta)}{\lambda^2}\right), \tag{1.4.11}$$

where g is the factor that places the intensity on an internally consistent scale with the other images in the dataset, g_1 and g_2 are scaling factors, φ is the crystal rotation angle which is used as a proxy for the primary beam direction and the dose, and $\boldsymbol{s}_2$ is the secondary beam direction (Evans and Murshudov, 2013). The crystal rotation angle, φ, is used as a proxy for dose because the dose is not routinely calculated in MX experiments. It is noted by Evans and Murshudov that the B-factor correction in this model is largely an average correction.

Many studies, including the first by Blake and Phillips, found that the decay of individual reflection intensities is not necessarily monotonic (Blake and Phillips, 1962; Hendrickson *et al.*, 1973; Hendrickson, 1976). As well as diffraction being anisotropic (Abrahams and Marsh, 1987), some reflections also increase in intensity with increasing dose (Abrahams, 1973).

Therefore the average radiation damage correction will not account for the specific changes that are known to occur. Diederichs *et al.* used a linear function to correct individual reflection intensities up to a dose of 10 MGy and showed that this reflection specific correction could improve results in SAD phasing (Diederichs *et al.*, 2003). Diederichs later used quadratic and exponential functions to correct individual reflection intensities (Diederichs, 2006). It was concluded that the quadratic model suffers from the problem that it requires two parameters to be fitted for each reflection and it is not suitable when the unit cell ex-

pands because the Fourier transform of the molecule is no longer consistent between observations of the same reflection. The linear model suffers from the problem that the fitted parameter does not have a physical interpretation and the correction can result in negative extrapolated intensities. The exponential model has the advantage that it only results in positive values but again it lacks a physical interpretation.

Despite some of the drawbacks, these specific correction models have shown some promise, but they are not yet standard. This is because they are not necessarily straightforward to implement and they do not always improve the data quality (Phil Evans, personal communication).

1.5 Radiation damage in SAXS

The radiation damage literature for SAXS experiments is far less extensive than that for MX, and the phenomenon is less well understood. What is known is that the absorption events that occur in MX (primary damage) will occur in SAXS, since the physics of the interactions will not change. SAXS experiments are generally carried out at RT which means that in addition to electron radicals, hydroxyl (OH) and hydroperoxyl (HO_2) radicals created by the interaction of X-rays with water will also be mobile in the buffer solution (Jeffries *et al.*, 2015; Garrison, 1987). These radicals attach to the protein backbone and/or sidechains which in turn induce protein aggregation through covalent or non-covalent bonding (Kuwamoto *et al.*, 2004), fragmentation, conformational changes and unfolding (Hopkins and Thorne, 2016). Radiation damage is observed in the experimental data as a lack of overlap between 1D diffraction intensity curves for successive frames.

There are several methods to mitigate the radiation damage caused in SAXS experiments. These include:

- The sample can be flowed through a capillary to limit the X-ray exposure for a given volume of sample. An alternative approach to this is to translate the capillary in the beam (Jeffries *et al.*, 2015).
- In a similar manner to MX, the beam can be attenuated or defocussed (Jeffries *et al.*, 2015).

- Radioprotectants such as DTT, TCEP, glycerol, ethylene glycol, ascorbate or sucrose can be added. These are thought to be able to capture hydroxyl radicals (Grishaev, 2012).
- The SAXS experiment can be carried out at cryo-temperatures which reduces the damage per unit dose by up to five orders of magnitude compared with experiments at RT (Meisburger *et al.*, 2013). However performing cryo-SAXS is not routine because the experimental set-up is far from trivial (Jeffries *et al.*, 2015). It should be noted that dose estimates in SAXS experiments are calculated using equation 1.4.1, which is a one dimensional representation of the experiment.

Programs for primary data analysis do exist to assess the quality of the data (Petoukhov *et al.*, 2012). In particular, DATCMP has recently been improved to provide correlation maps (CorMaps) to assess the similarity of the 1D scattering curves produced in a SAXS experiment (Franke *et al.*, 2015). However, there is still room to improve the assessment of radiation damage progression online during the experiment. A recent radiation damage study attempts to go further than just assessing the similarity of frames, and tries to determine the nature and rate of radiation damage (Hopkins and Thorne, 2016). However, this type of analysis is not yet mature and the quantitative methods are too involved for routine use by inexperienced experimenters.

1.6 This thesis in context

The work presented in this thesis extends previous work on radiation damage correction models. A new data reduction algorithm to process MX data has been developed that explicitly tracks the changes in the amplitude values throughout the diffraction experiment. The several benefits of this approach include:

- Explicit tracking of the amplitude uncertainties, even for negative and weak positive reflections.
- Estimates of intensity and amplitude values at every point in the diffraction experiment.
- The possibility to explicitly correct the amplitude values for radiation damage, as opposed to correcting the intensities.

Furthermore, RADDOSE-3D has been extended to calculate doses for SAXS experiments. This has been used to assess the efficacy of several radioprotectants for SAXS experiments. As a result of this work, a new metric has been developed for defining the threshold at which 1D SAXS curves are significantly damaged.

- *Chapter 2* describes the experiment that was performed to collect data at 100 K from several crystals of different proteins with a specially designed top-hat profile X-ray beam. These data are then used to test the validity of the decay models presented in section 1.4.7. Finally the chosen model is used to improve the DWD metric proposed by Zeldin *et al.* (2013).

- *Chapter 3* describes the regression model used to make corrections to intensities of individual reflections using RADDOSE-3D and DWD. It also proposes an extension to the regression model using Bayesian inference methods to extrapolate intensity values for low multiplicity reflections.

- *Chapter 4* reports the main results of this thesis. In this chapter, the new data reduction algorithm is presented. The results are also presented and compared to those using the current data reduction pipeline (AIMLESS & CTRUNCATE). Considerations of the algorithm and extensions are also discussed.

- *Chapter 5* describes work on processing experimentally measured X-ray beam profiles for use in RADDOSE-3D. Methods for processing 1D beam profile data and 2D beam images with significant background contribution are presented.

- *Chapter 6* describes the extension of RADDOSE-3D to calculate dose values for SAXS experiments. This extension is used to assess the dose suffered by samples in an experiment to determine the efficacy of radioprotectants in SAXS experiments. A new metric to define the threshold at which 1D SAXS curves are significantly damaged is also presented.

- *Chapter 7* Summarises the work and the conclusions drawn in the previous chapters and suggests future directions as a result of the research presented.

CHAPTER 2

Dose Decay Modelling

2.1 Introduction

The diffraction weighted dose (DWD) is a dose metric that exhibits a consistent and reproducible relationship with the relative decay in total diffraction intensity from a protein crystal down to a relative intensity of 40% (Zeldin *et al.*, 2013a) (relative intensity is defined as I_n/I_1, where I_n is the summed mean intensity of a complete data set n (or equivalent sections of data) after a dose D, and I_1 is the mean intensity of the first data set). This is because DWD spatially resolves the exposure of the crystal to the X-ray beam in addition to accounting for the absorbed dose. Mathematically this spatial resolution of the exposure is incorporated via the fluence weighting in the DWD calculation (equation 1.4.2). However, the DWD currently does not take into account the loss of diffraction power due to the current damage state of the crystal. The *relative diffraction efficiency* (RDE), η, is introduced as a function of the absorbed dose that describes the diffracting power of any given volume of the crystal. The RDE is defined as the ratio of the proportion of incident photons that are currently being diffracted to the proportion that initially diffracted. This is graphically depicted in Figure 2.1.

The RDE is incorporated into the DWD equation as an additional weighting term. Explicitly this is

$$DWD^i = \frac{\int_{t_{i-1}}^{t_i} \iiint_{crystal} D(\boldsymbol{x},t) \times F(\boldsymbol{x},t) \times \eta(D(\boldsymbol{x},t)) \quad \mathrm{d}\boldsymbol{x}\mathrm{d}t}{\int_{t_{i-1}}^{t_i} \iiint_{crystal} F(\boldsymbol{x},t) \times \eta(D(\boldsymbol{x},t)) \quad \mathrm{d}\boldsymbol{x}\mathrm{d}t}. \tag{2.1.1}$$

The work presented in this chapter describes the experiment and analysis carried out to determine a suitable functional form for the RDE. The RDE is then incorporated into the DWD using equation 2.1.1 and compared with the simple DWD (equation 1.4.2) and applied to the data used in the previous DWD study (Zeldin *et al.*, 2013a).

2.2 Experimental methods

2.2.1 Considerations

To explore the behaviour of η as a function of absorbed dose it is necessary to understand the role of η in the calculation of the DWD (equation 2.1.1). The crystal is represented as a

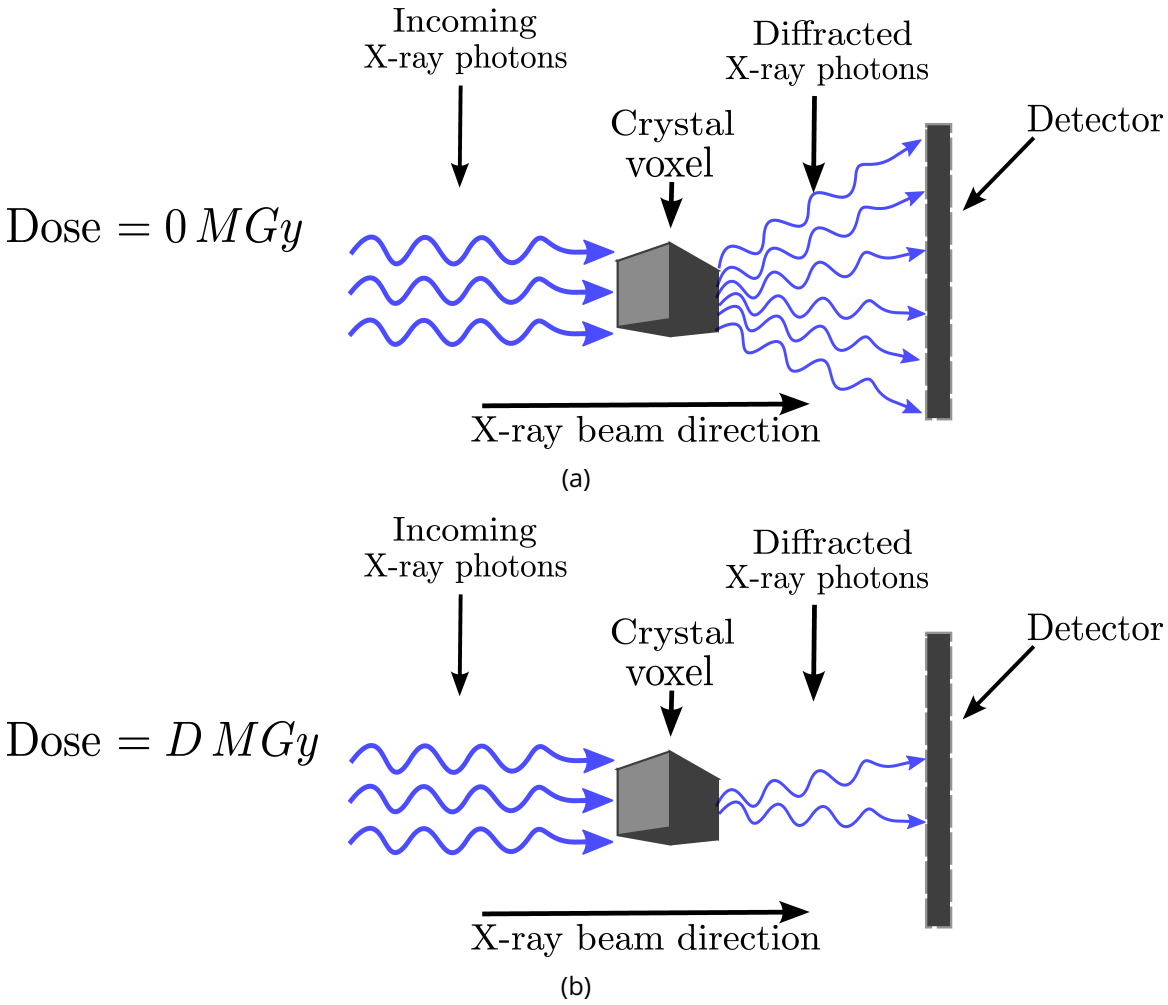


Figure 2.1: A visual example of the relative diffraction efficiency (RDE). (a) Initially at dose = $0\,MGy$ the crystal voxel elastically scatters 6 photons for a given number of incident photons. (b) Later in the experiment when the crystal voxel has absorbed a dose $= D\,MGy$, it elastically scatters only 2 photons for the same number of incident photons. The ratio of the current number of diffracted photons with the initial number is $2/6 = 1/3$, therefore in this example the RDE is 1/3. The components in the figure are not drawn to scale.

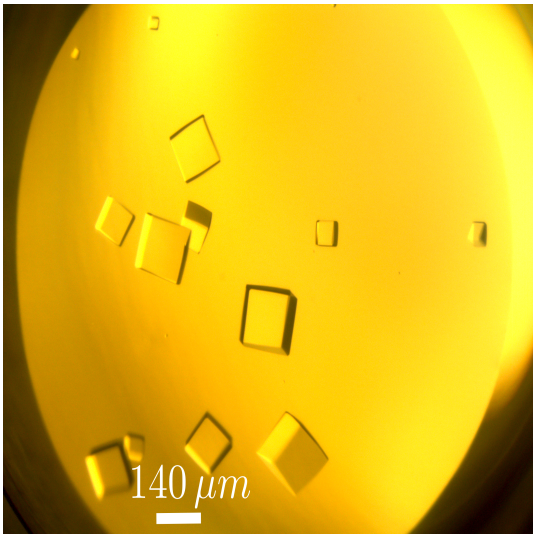
collection of voxels* in RADDOSE-3D. η is calculated for each voxel within the crystal and the absorbed dose within each voxel is assumed to be homogeneous. Therefore to experimentally determine the behaviour of η within a voxel as a function of absorbed dose, the crystals in the diffraction experiment must be irradiated uniformly to produce a homogeneous dose distribution. To achieve this it is necessary to use a flat (top-hat) beam profile with the entire crystal volume completely immersed within the X-ray beam throughout the rotation. At the time of the experiment (January 2014) RADDOSE-3D was only able to model cuboid or spherical crystal shapes. Therefore the crystals used in the experiment were grown to be as close to cuboid in shape as possible.

2.2.2 Crystallization

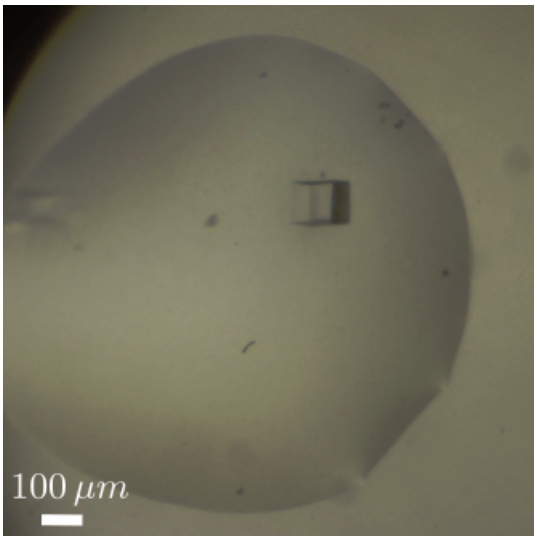
Crystals of bovine pancreatic insulin purchased from Sigma-Aldrich (Lot # SLBJ0654V) were grown by the sitting-drop vapour diffusion method. The well solution consisted of 0.243 M Na_2HPO_4, 0.007 M Na_3PO_4 at pH 10, and 0.01 M Na_3EDTA. 2 μl of the well solution was added to an equal volume of the protein solution which consisted of 20 mg/ml insulin protein, 0.0195 M Na_2HPO_4, 0.0005 M Na_3PO_4 at pH 10, and 0.01 M Na_3EDTA. The crystals were stored at room temperature ($\approx$ 293 K) and grew in a morphologically cuboid shape within 48 hours (Figure 2.2a). Cuboid shaped crystals less than 140 μm in each dimension were selected and soaked for 30 - 60 seconds in a cryoproctectant solution with an identical composition to that of the well solution except with 30% v/v glycerol substituted for water, before being flash cooled into liquid nitrogen (77 K).

Human Haspin and myelocytomatosis (MYC) induced nuclear antigen (MINA) protein crystals that were cuboid in shape were kindly provided by the Structural Genomics Consortium (SGC) (Figures 2.2b and 2.2c). These crystals were selected for their cubic shape out of the many human proteins crystallised at the SGC. They were cryoprotected in their native well solution with 25% v/v glycerol substituted for water.

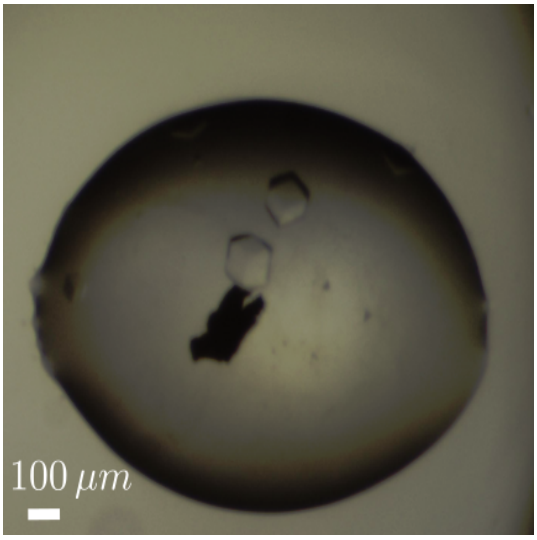
*A voxel is the smallest distinguishable volume element in a three-dimensional representation of a computationally modelled object.



(a) Insulin



(b) Haspin



(c) MINA

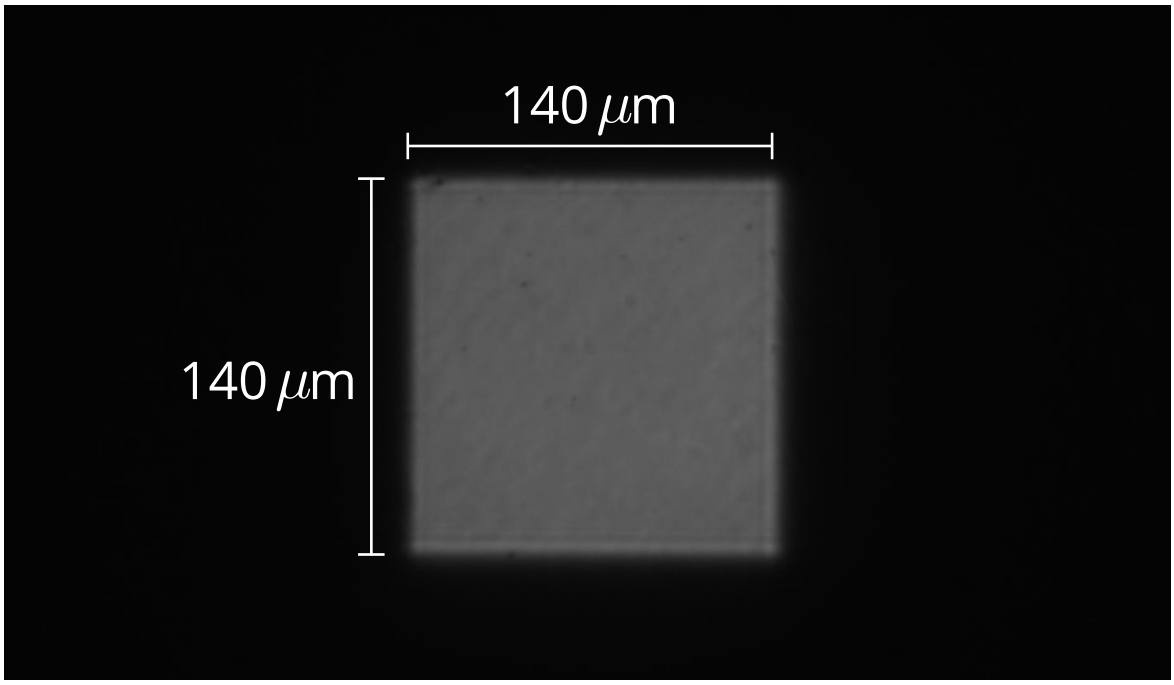
Figure 2.2: Crystals used in the experiment.

2.2.3 Data collection and dose calculation

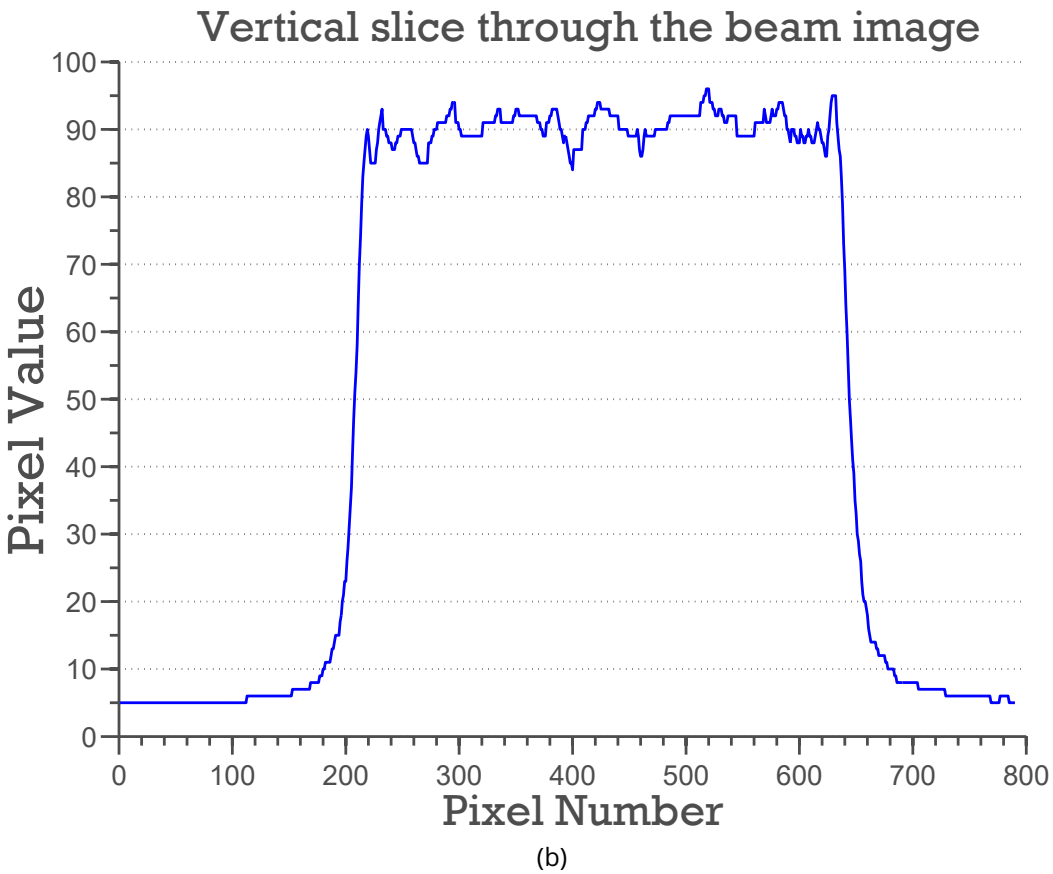
All data were collected at 100 K on the PETRA III Hamburg beamline P14 using an X-ray energy of 12.7 keV ($\lambda = 0.9764$ Å), in collaboration with beamline scientist Dr. Gleb Bourenkov and EMBL beamline director Dr. Thomas Schneider. The experimentally measured beam profile was determined by placing a scintillator combined with an Allied Vision GC1350C CCD camera directly in the beam path. This produced a quantitative map of the beam profile in a portable graymap (pgm) file. The flat profile of the beam (coefficient of variation[†] of the beam is 2.09% vertically and 2.24% horizontally) was achieved by removing the focusing mirrors, and the slits were adjusted to achieve an aperture of 140 $\mu m \times$140 μm (Figure 2.3). The beam current was measured using a 500 μm thick silicon PIN diode placed in the sample position from which the photon flux could be calculated (Owen *et al.*, 2009). Before the data were collected from each crystal, an indexing set (100 frames of 0.1° rotation and 0.1 s exposure time per frame) was acquired. These frames were then indexed to provide the information necessary to reorient the crystal. One of the crystal faces was then aligned perpendicular to the beam direction such that the plane containing the beam direction vector was perpendicular to two of the edges of the aligned face (Figure 2.4). Alignment was performed using an Arinax MD3 mini kappa goniometer with the spindle axis mounted in a vertical and downward configuration. The crystal was centred on the beam position to make sure the entire crystal volume was completely immersed in the beam during the experiment. The dimensions of the crystals were measured on the screen prior to data collection. Table 2.1 contains the details of the data collection strategies for each crystal type.

Dose values were calculated using RADDOSE-3D. The photon flux was determined to be 1.9 $\times$ 10^{11} ph/s, the composition of the crystal was obtained from Dr. Oliver B. Zeldin's thesis (Zeldin, 2013) and from the constituents of the crystallisation solution as in Section 2.2.2. Although the crystal composition in Zeldin's thesis is incorrect (he specifies that there is a zinc atom for every insulin monomer in the unit cell whereas the true composition has only two zinc atoms per insulin hexamer), using the same composition allows direct comparison with the results obtained in (Zeldin *et al.*, 2013a). Functionality to handle the experimentally measured beam profile was added to RADDOSE-3D to further improve the simulation of the absorbed dose.

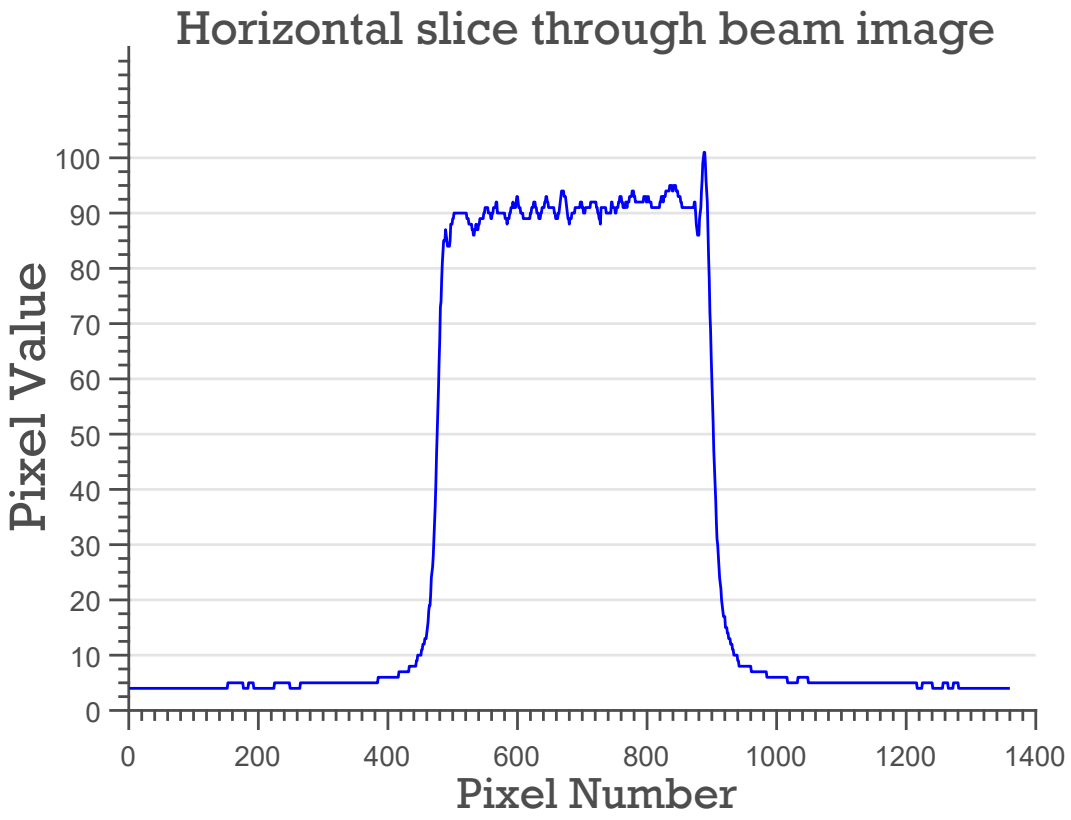
[†]The coefficient of variation is defined at the ratio of the standard deviation to the mean of a set of values which can also be expressed as a percentage by multiplying by 100%.


140 μm
140 μm

(a)


Vertical slice through the beam image
Pixel Value
Pixel Number
0
10
20
30
40
50
60
70
80
90
100
0
100
200
300
400
500
600
700
800

(b)



(c)

Figure 2.3: (a) Image (790 x 1360 pixels) of the experimentally determined beam profile. (b) Vertical slice through the beam image showing the flat profile of the beam. (c) Horizontal slice through the beam image showing the flat profile of the beam.

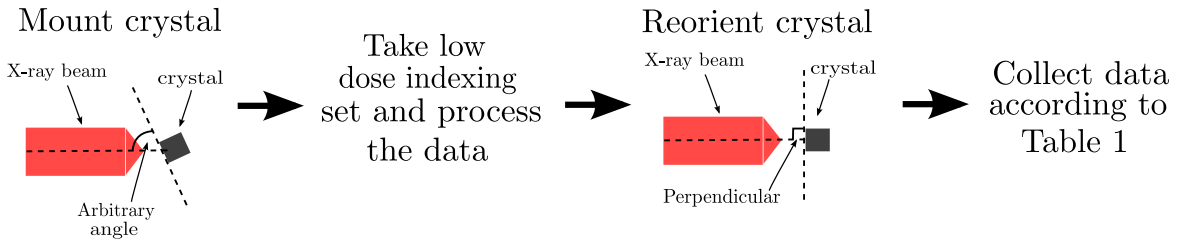


Figure 2.4: Flow diagram of the crystal reorientation process prior to data collection.

Table 2.1: Data collection strategy for each protein crystal type

Protein Crystal Type	Number of Crystals	Total Number of Frames per crystal	Rotation per image (°)	Total rotation (°)	Exposure Time per image (seconds)	Approx resolution
Insulin	8	14400	0.1	1440	0.5	1.38 Å
Haspin	5	7200	1	7200	1	2.9 Å
MINA	3	7200/3600	0.1	720/360	1	3.0 Å

2.2.4 Data processing

The data collected for the SGC MINA and haspin crystals were non-trivial to analyse and hence the data processing procedures described here are relevant only for the insulin crystals. Data were processed using the Collaborative Computational Project No. 4 (CCP4) suite (Winn *et al.*, 2011) with a standardised script used to call each program from within the suite to ensure identical treatment of all crystals. MOSFLM (Leslie and Powell, 2007) was run manually with the space group set to $I2_13$. All images collected from a crystal were integrated together, fixing the unit cell angles and the detector distance but allowing the unit cell dimensions and mosaicity to be refined during integration. This was done because it is well documented that the unit cell expands and the mosaicity increases as radiation damage progresses (Garman, 2010).

The data were then scaled with AIMLESS (Evans and Murshudov, 2013) in batches of 900 frames (equivalent to 90° rotations) separated 50 frames apart (equivalent to 5°). This resulted in several overlapping datasets being extracted which allowed the progression of radiation damage to be tracked in much more detail than would be the case if datasets did not overlap. It is important to note however that these datasets will have sampled different regions of reciprocal space and so will therefore have different numbers of total observations, however, the differences are not expected to be significant. 14400 images were collected from each insulin crystal, so processing the data in this way produced 271 datasets per crystal. Despite the fact that the insulin crystals diffracted to 1.38 Å, the data were scaled to a resolution limit of 1.8 Å. This was to ensure that the processing for the highly damaged datasets were less likely to fail and to allow more direct comparison of the intensity values between datasets. Only five of the eight insulin datasets produced data that were processed straightforwardly in MOSFLM. Data processing statistics for those five insulin crystals are shown in Table 2.2.

Table 2.2: Overall data processing statistics for the first data set collected from each of the processed insulin crystals (the completeness and multiplicity values are similar for the other datasets because the angular range is the same and the highest observable resolution data, out to 1.4 Å, has been discarded for this part of the analysis).
Values in parentheses are for the outer shell (1.83-1.79Å). Unit cell and mosaicity are average values for all 14400 images

Crystal	0259	128	172	137	180
Space group	$I2_13$	$I2_13$	$I2_13$	$I2_13$	$I2_13$
Unit-cell parameters					
$a = b = c$ (Å)	78.28	78.28	78.35	78.36	78.40
$\alpha = \beta = \gamma$ (°)	90	90	90	90	90
Total No. of reflections	71955 (4528)	70860 (4208)	70757 (4423)	71968 (4554)	71580 (4463)
No. of unique reflections	7446 (474)	7436 (445)	7478 (471)	7478 (473)	7468 (471)
Completeness (%)	99.8 (100)	99.9 (100)	100 (100)	99.9 (100)	99.9 (100)
Multiplicity	9.7 (9.6)	9.5 (9.5)	9.5 (9.4)	9.6 (9.6)	9.6 (9.5)
$I/\sigma(I)$	33.3 (12.3)	37.9 (20.4)	32.3 (11.5)	43.6 (22.7)	33.9 (13.2)
R_{merge}	0.040 (0.128)	0.041 (0.097)	0.041 (0.158)	0.036 (0.086)	0.041 (0.128)
$CC_{1/2}$	0.999 (0.991)	0.998 (0.995)	0.994 (0.986)	0.988 (0.997)	0.999 (0.989)
Mosaicity (°)	0.42	0.30	0.29	0.34	0.28

2.2.5 Calculating the relative intensity

The relative intensity of a uniformly irradiated crystal is effectively the same as the RDE. Recall that the relative intensity is defined as I_n/I_1, where I_n is the summed mean intensity of a complete data set n (or equivalent sections of data) after a dose D, and I_1 is the mean intensity of the first data set. Implicit in this definition is that the total intensity is integrated for the same reflections for each dataset. Note: the relative intensity in this thesis has been calculated incorrectly and is described below.

Figure 2.5 shows a table from the log file of an AIMLESS job. The overall mean intensity given in the table (highlighted red in Figure 2.5) is calculated from intensities of the reflections within that particular dataset. The relative intensity can be calculated by dividing this value by the corresponding overall mean intensity from the first dataset, referred to in Figure 2.6 as *Aimless average intensity*.

Another way to calculate the relative intensity is to calculate the total summed intensity of a dataset, then divide that by the total summed intensity of the first dataset, referred to in Figure 2.6 as *total intensity*. This was achieved by multiplying the number of measured

reflections in each resolution bin (*Nmeas* - highlighted orange in Figure 2.5) by the average intensity in that resolution bin (*AvI* - highlighted blue in Figure 2.5), then summing them. The latter method of calculating the relative intensity is was preferred over the previous method because it is more representative of the overall intensity loss, as opposed to the mean intensity loss. Therefore the relative intensity was always calculated using the total intensity method. Figure 2.6 shows the results of calculating the relative intensity using both methods and shows that the Aimless average intensity method decays more slowly than the total intensity method as the relative intensity drops below 0.6.

It is important to note that both of the methods described above for calculating the relative intensity are incorrect for two main reasons:

1. each dataset was scaled separately in Aimless. However Aimless assigns a different scale factor for each run and therefore the scaled values are not necessarily comparable between datasets.

2. the summed intensity values are derived from all reflections observed in a particular dataset. However these datasets sample different sections of reciprocal space and hence the same reflections are not necessarily observed in all datasets.

To calculate the relative intensity correctly, all datasets should be scaled together and only the reflections that are observed in all datasets should be used for the total summed intensity calculation. This method of calculating the relative intensity has not been used for the data in this thesis because this was discovered during the viva and it was determined that it is likely that the conclusions of the work will not change using the new method.

2.3 Dose decay models

With a calculated dose and a relative intensity value for each dataset, a relationship between radiation damage progression and absorbed dose can be determined via dose decay models (DDMs). A DDM is a function that describes the change of reflection intensity as a function of the absorbed dose. Several DDMs have been proposed over the last few decades as enumerated in section 1.4.7. In particular, three different DDMs have been successful at

$TABLE: Analysis against resolution, New:
$GRAPHS:I/sigma, Mean Mn(I)/sd(Mn(I)):0|0.305791x0|66.3004:2,13,14:
:Rmerge, Rfull, Rmeas, Rpim v Resolution:0|0.305791x0|0.135111:2,4,5,6,7:
:Average I, RMSdeviation and Sd:0|0.305791x0|12681:2,10,11,12:
:Fractional bias:0|0.305791x-0.100293|0.011556:2,15:
$$

N	1/d^2	Dmid	Rmrg	Rfull	Rcum	Rmeas	Rpim	Nmeas	AvI	RMSdev	sd	I/RMS	Mn(I/sd)	FrcBias
1	0.0064	12.53	0.027	-	0.027	0.029	0.010	600	12681	811	498	15.6	60.5	-0.002
2	0.0191	7.23	0.030	-	0.028	0.032	0.010	1172	4961	247	226	20.1	53.4	-0.008
3	0.0319	5.60	0.029	-	0.029	0.031	0.011	1323	7403	367	316	20.2	54.1	-0.002
4	0.0446	4.74	0.030	-	0.029	0.032	0.010	1591	10684	540	434	19.8	61.3	0.008
5	0.0573	4.18	0.031	-	0.030	0.033	0.010	2017	12067	613	485	19.7	66.3	0.012
6	0.0701	3.78	0.033	-	0.031	0.035	0.011	2182	10410	599	429	17.4	61.8	0.003
7	0.0828	3.47	0.034	-	0.031	0.036	0.011	2478	7527	441	331	17.1	58.2	0.008
8	0.0956	3.23	0.033	-	0.031	0.035	0.011	2439	5525	293	262	18.8	53.8	0.002
9	0.1083	3.04	0.037	-	0.032	0.039	0.012	2739	3962	231	207	17.2	48.1	-0.005
10	0.1210	2.87	0.035	-	0.032	0.036	0.012	2901	3047	163	174	18.7	44.3	0.000
11	0.1338	2.73	0.035	-	0.032	0.037	0.012	3099	2753	158	165	17.5	42.2	-0.015
12	0.1465	2.61	0.036	-	0.032	0.038	0.012	3190	2201	125	145	17.5	38.5	-0.020
13	0.1593	2.51	0.039	-	0.033	0.041	0.013	3270	2005	123	139	16.3	36.7	-0.021
14	0.1720	2.41	0.041	-	0.033	0.044	0.014	3410	1717	113	130	15.2	34.2	-0.023
15	0.1847	2.33	0.048	-	0.033	0.051	0.016	3685	1482	114	124	13.0	30.9	-0.031
16	0.1975	2.25	0.051	-	0.034	0.054	0.017	3558	1370	110	121	12.5	29.2	-0.041
17	0.2102	2.18	0.054	-	0.035	0.057	0.018	3967	1249	104	118	12.0	27.3	-0.031
18	0.2230	2.12	0.064	-	0.035	0.067	0.021	3667	1115	113	114	9.9	25.4	-0.047
19	0.2357	2.06	0.075	-	0.036	0.079	0.025	3906	859	102	105	8.4	21.2	-0.049
20	0.2485	2.01	0.086	-	0.037	0.092	0.030	3722	780	108	101	7.2	19.9	-0.042
21	0.2612	1.96	0.092	-	0.037	0.097	0.031	4115	638	94	95	6.8	17.6	-0.055
22	0.2739	1.91	0.099	-	0.038	0.105	0.034	4070	571	91	93	6.3	16.6	-0.067
23	0.2867	1.87	0.113	-	0.039	0.119	0.038	4315	463	85	88	5.5	14.3	-0.088
24	0.2994	1.83	0.128	-	0.040	0.135	0.043	4524	375	78	85	4.8	12.3	-0.100
$$														
Overall:			0.040	0.000	0.040	0.042	0.013	71940	2798	241	168	11.6	33.3	-0.008
N	1/d^2	Dmid	Rmrg	Rfull	Rcum	Rmeas	Rpim	Nmeas	AvI	RMSdev	sd	I/RMS	Mn(I/sd)	FrcBias

Figure 2.5: A table given in the log file of an AIMLESS job from an insulin crystal. The table gives data for 24 resolution bins including the number of measured reflections in each resolution bin, Nmeas (highlighted orange), and the Average Intensity in each resolution bin, AvI (highlighted blue). The value highlighted in red is the overall average intensity for a dataset calculated using reflections that are measurable in that data set.

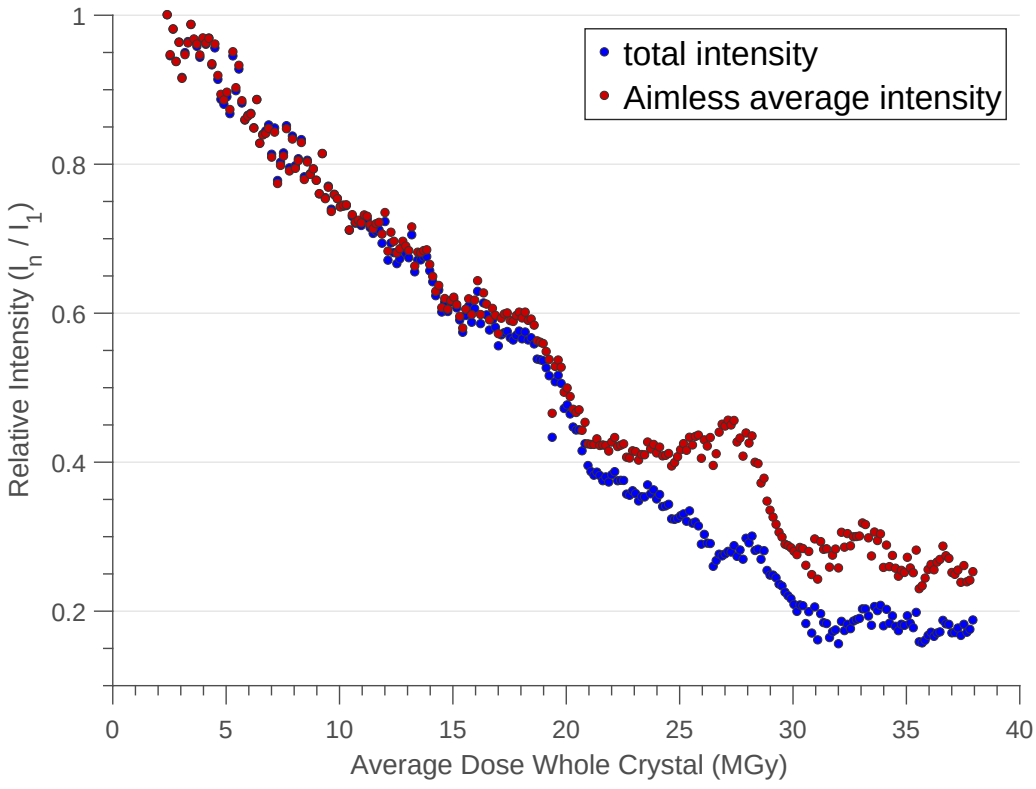


Figure 2.6: Relative intensity plotted against the average absorbed dose for one insulin crystal using the two methods described in section 2.2.5.

describing radiation damage progression. These models were analysed and compared, to determine which one best described the data collected in this experiment.

As mentioned in chapter 1, the first of the three models was proposed by Sygusch & Alliare (1988) (Sygusch and Allaire, 1988), which was developed from the original model proposed by Blake & Phillips (1962) (Blake and Phillips, 1962) and subsequently altered by Hendrickson, Fletterick and others. Sygusch & Allaire's model assumes that the intensity of a reflection from a protein crystal is a linear combination of scattering contributions from 4 states; an undamaged fraction, A_1, that contributes to diffraction at all angles, a fraction that has undergone 'surface modification' but still conformationally resembles the undamaged protein, A_1', and hence also contributes to diffraction at all angles, a disordered fraction, A_2, that mainly contributes to diffraction at low angles, and finally an amorphous fraction, A_3, that is no longer capable of coherent scattering. The model also assumes that radiation damage is a sequential and irreversible process, so the crystal transitions between states as follows:

$$A_1 \xrightarrow{k_0} A_1' \xrightarrow{k_1} A_2 \xrightarrow{k_2} A_3. \tag{2.3.1}$$

Mathematically the intensity of an individual reflection, I, as a function of absorbed dose, D is given by:

$$I(D)/I(0) = A_1(D) + A_1'(D) + A_2(D)\exp\left[-\frac{B}{2}h^2\right], \tag{2.3.2}$$

assuming isotropic atomic vibrations, where B is the thermal parameter related to the mean square displacement of atomic vibration (Drenth, 1999), and $h = 1/d$, where d is the distance between successive Bragg planes.

The crystal fractions are assumed to evolve according to the following system of coupled first order linear ordinary differential equations (ODEs)

$$\frac{\mathrm{d}A_1}{\mathrm{d}t} = -k_0 I_0, \tag{2.3.3a}$$
$$\frac{\mathrm{d}A_1'}{\mathrm{d}t} = k_0 I_0 - k_1 A_1', \tag{2.3.3b}$$
$$\frac{\mathrm{d}A_2}{\mathrm{d}t} = k_1 A_1' - k_2 A_2, \tag{2.3.3c}$$
$$\frac{\mathrm{d}A_3}{\mathrm{d}t} = k_2 A_2, \tag{2.3.3d}$$

subject to the following constraints:

$$A_1(D) + A_1'(D) + A_2(D) + A_3(D) = A_0, \tag{2.3.4a}$$

$$A_1'(0) = A_2(0) = A_3(0) = 0, \tag{2.3.4b}$$

$$A_1(0) = A_0, \tag{2.3.4c}$$

where I_0 represents the intensity of the incident irradiation and A_0 is the quantity of protein in the irradiated sample.

The solution of the set of ODEs (2.3.3) subject to the constraints (2.3.4) is presented here:

$$A_1 = A_0 - k_0 I_0 D, \tag{2.3.5a}$$

$$A_1' = \frac{k_0 I_0}{k_1}\left(1 - e^{-k_1 D}\right), \tag{2.3.5b}$$

$$A_2 = \frac{k_0 I_0}{k_2}\left(1 - e^{-k_2 D}\right) + \frac{k_0 I_0}{k_2 - k_1}\left(e^{-k_2 D} - e^{-k_1 D}\right), \tag{2.3.5c}$$

$$A_3 = k_0 I_0 D + \left(\frac{k_0 I_0}{k_2 - k_1} - \frac{k_0 I_0}{k_2}\right)\left(1 - e^{-k_2 D}\right) - \frac{k_2 k_0 I_0}{k_1(k_2 - k_1)}\left(1 - e^{-k_1 D}\right). \tag{2.3.5d}$$

Equations (2.3.5a), (2.3.5b) and (2.3.5c) are explicitly given in (Sygusch and Allaire, 1988), whereas equation (2.3.5d) is my own work for this thesis. Since the transition $A_1 \rightarrow A_1'$ is zero-order, eventually all of the undamaged crystal fraction, A_1, will be converted into A_1'. Above this dose, $D = D_L = A_0/k_0'$, where $k_0' = k_0 I_0$ (set $A_1 = 0$ and rearrange equation 2.3.5a), the solutions given by equation 2.3.5 will no longer be valid since A_1 will become negative. To obtain a solution for $D > D_L$ I observed that $k_0 \equiv 0$ and hence solved the set of ODEs (2.3.3) subject to the constraints (2.3.4) with the additional constraint on k_0 yields:

$$A_1 = 0, \tag{2.3.6a}$$

$$A_1' = A_{10}' e^{-k_1(D - D_L)}, \tag{2.3.6b}$$

$$A_2 = \frac{k_1 A_{10}'}{k_2 - k_1} e^{-k_1(D - D_L)} + \left(A_{20} - \frac{k_1 A_{10}'}{k_2 - k_1}\right) e^{-k_2(D - D_L)}, \tag{2.3.6c}$$

$$A_3 = A_{30} + \left(A_{20} - \frac{k_1 A_{10}'}{k_2 - k_1}\right)\left(1 - e^{-k_2(D - D_L)}\right) + \frac{k_2 A_{10}'}{k_2 - k_1}\left(1 - e^{-k_1(D - D_L)}\right), \tag{2.3.6d}$$

where A_{10}', A_{20} and A_{30} are the values of the crystal fractions A_1', A_2 and A_3 at dose $D = D_L$ respectively. Although a solution for the model is undefined for $k_1 = k_2$, Table 2.5 shows that the best fit values for the two parameters are different enough to neglect this case.

The strength of this model lies in the fact it explicitly predicts the proportions of damaged

states of the crystal as a function of the absorbed dose via the system of ODEs, and uses that solution to determine the intensity of a reflection. The drawbacks of this model are that it requires 3 parameters to be determined and also that its analytical form is not easy to intuitively interpret.

The second model was proposed by Dr. James Holton and assumes that the average intensity of a reflection $I(D)$ decays exponentially (Holton and Frankel, 2010). Mathematically this is expressed as:

$$I(D) = I_{ND} \exp\left[-\ln(2)\frac{Dh}{H}\right], \tag{2.3.7}$$

where I_{ND} is the average reflection intensity expected in the absence of radiation damage, $\ln(2)$ is the natural logarithm of 2 ($\approx$ 0.693), D is the dose, $h = 1/d$, where d is the distance between successive Bragg planes and H is Howell's criterion (10 MGy Å^{-1}). The advantages of this model are that it is very simple and only has a maximum of two parameters to be determined, I_{ND} and H. To simplify the parameter value extraction, the value of H can be assumed to be 10 MGy Å^{-1} and I_{ND} can be approximated to the intensity values from the first data set. Therefore this model is relatively easy to apply and does not require any further simplifying assumptions.

However the simplicity of this model is also the cause of its main disadvantage: it does not predict a lag phase for intensity decay. At cryotemperatures this prediction may be correct but recent results suggest otherwise for room temperature MX (Owen *et al.*, 2014). Therefore this model already has apparent limitations in the applicability of its predictive power.

The third DDM was proposed by Leal *et al.* (Leal *et al.*, 2012) and uses a similar radiation damage model to those used in many scaling programs (Evans and Murshudov, 2013; Kabsch, 2010b). It is:

$$J(D) = J(h) \times scale(D) \times \exp\left[-\frac{B(D)h}{2}\right], \tag{2.3.8}$$

with

$$scale(D) = K \exp\left[-\gamma^2 D^2\right] \tag{2.3.9}$$

$$B(D) = B_0 + \beta D \tag{2.3.10}$$

where $J(D)$ is the expected intensity after the crystal has absorbed a dose D, $J(h)$ is the

expected reflection intensity at reciprocal distance h from the origin in the absence of any radiation damage and B_0, β, K and γ are parameters to be determined by fitting the model to the data, and are completely empirical.

This model essentially describes a Gaussian decay of the intensity with dose. This gives it the potential to predict lag phases, which will be dependent on the parameter values obtained from the data. It has already been shown to successfully predict relative intensity decay at room temperature (Leal *et al.*, 2012) at dose rates below those used in (Owen *et al.*, 2014) (0.05 - 300 kGy s^{-1}). The main disadvantage of this model is that the scale function $K \exp\left[-\gamma^2 D^2\right]$ is completely empirical and has no obvious physical interpretation.

2.3.1 Validity test

Since the parameters for the models are fitted, judging the models based on how well they fit the data would not necessarily validate any particular model. A more robust test would be to check whether the data transforms in the way that the model predicts they should. In particular, models given by equations 2.3.7 and 2.3.8 can be transformed into linear forms

$$\ln(I) = \ln\left(I_{ND}\right) - \frac{h\ln(2)}{H} \times D, \tag{2.3.11}$$

$$\ln\left(\frac{J(D)}{J(h)}\right) = \ln\left(scale(D)\right) - \frac{B(D)}{2} \times h^2. \tag{2.3.12}$$

$$\tag{2.3.13}$$

These equations are of the linear form *y = mx + c*. Note that the Sygusch & Allaire model does not transform easily into a linear form. If the data follow the relationships described by the models, then transforming the data according to equations 2.3.11 and 2.3.12 means that plots of $\ln(I)$ against D (equation 2.3.11) and $\ln\left(\frac{J(D)}{J(h)}\right)$ against h^2 (equation 2.3.12) should resemble straight lines. A Pearson Correlation Coefficient (PCC) value[‡] can then be used to determine the strength of the linear relationship. PCC values were found for all data in each resolution shell where the relative intensity values (I_n/I_1) were above the experimental dose limit, 0.7 (beyond this point the data are likely to be biologically compromised (Owen *et al.*, 2006)). Table 2.3 shows the results of this test. The Leal *et al.* model seems to explain

[‡]The Pearson Correlation Coefficient is a measure of linear correlation between two variables and gives values between -1 and 1 inclusive. A value of 1 represents a positive correlation, -1 represents a negative correlation and 0 represents no correlation.

the data better. However the correlation coefficients determined for Holton's model are generally strong enough not to reject the possibility that this model may still sufficiently describe the data.

Table 2.3: Pearson Correlation Coefficient (PCC) values for linearly transformed intensity data. The values are negative, showing the negative correlation between intensity and dose i.e. as the dose increases, the intensity decreases.

	Holton Model	Leal *et al.* Model
Mean PCC	-0.9240	-0.9915
Max PCC	-0.2599[§]	-0.9896
Min PCC	-0.9819	-0.9935

2.3.2 Obtaining model parameter values

The transformations given by equations (2.3.11) and (2.3.12) produce suitably linear plots (Table 2.3), which allow the parameters for the Holton and Leal *et al.* models to be determined by calculating the gradient and intercepts for each plot respectively. Data collected before the experimental dose limit was reached were used to obtain fitted parameters for each of the models. I_{ND} was obtained by exponentiating (with Euler's number $e = 2.718...$) the intercept of the plot of $\ln(I)$ against dose, where I is the average reflection intensity in a given resolution bin. Note that I_{ND} will be different for each resolution bin (Table 2.4). H was obtained by calculating the gradient from the straight line plot for each resolution bin and rearranging according to equation 2.3.11. However H should be a constant value across all resolution bins, so the 'best' estimate of H was taken from the resolution bin that gave the best straight line fit.

The $B(D)$ and $scale(D)$ values for the Leal *et al.* model were found for each dose by calculating the gradient and intercept from a plot of $\ln\left(\frac{J(D)}{J(h)}\right)$ against h^2 for each dose. The $J(h)$ values are given by the BEST intensity curve (Popov and Bourenkov, 2003), derived from data collected from 72 different protein crystals at the DESY protein crystallography beamlines (Figure 2.7), and the $J(D)$ values are the measured mean intensities in resolution bins. The values of B_0, β, K and γ are found by fitting the functions given by equations (2.3.9) and (2.3.10) to the data values of $scale(D)$ and $B(D)$ using the Cauchy M-estimator (Figure 2.8). The Cauchy M-estimator[¶] was used instead of the commonly used least-squares fitting

[§]This value is an outlier from the lowest resolution shell. The PCC value for the next shell is -0.7578.

[¶]M-estimation is a fitting technique designed to be insensitive to outliers. The idea is to try to minimise a function of the residual values that increases less than the square, because the square of the residual of an

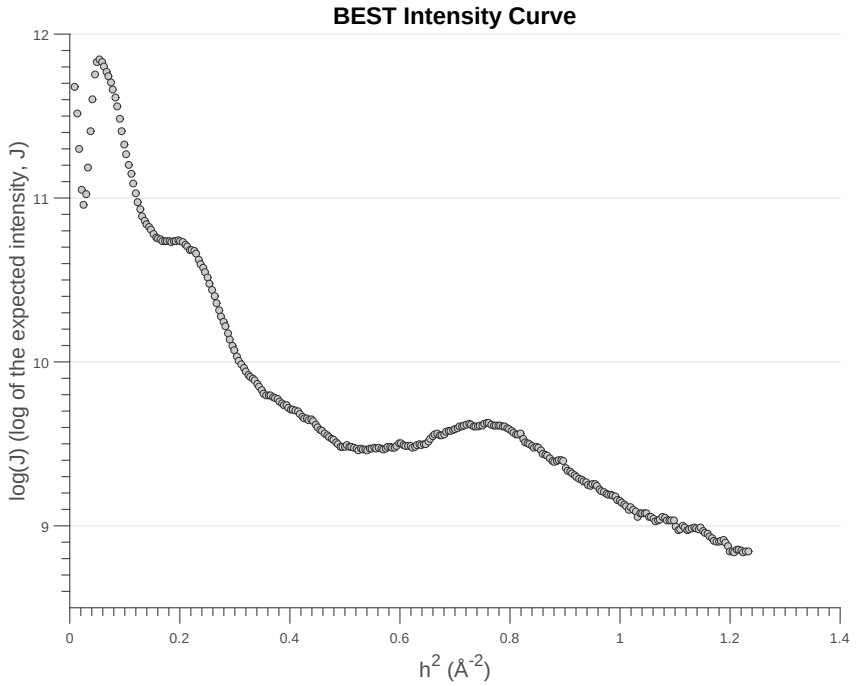
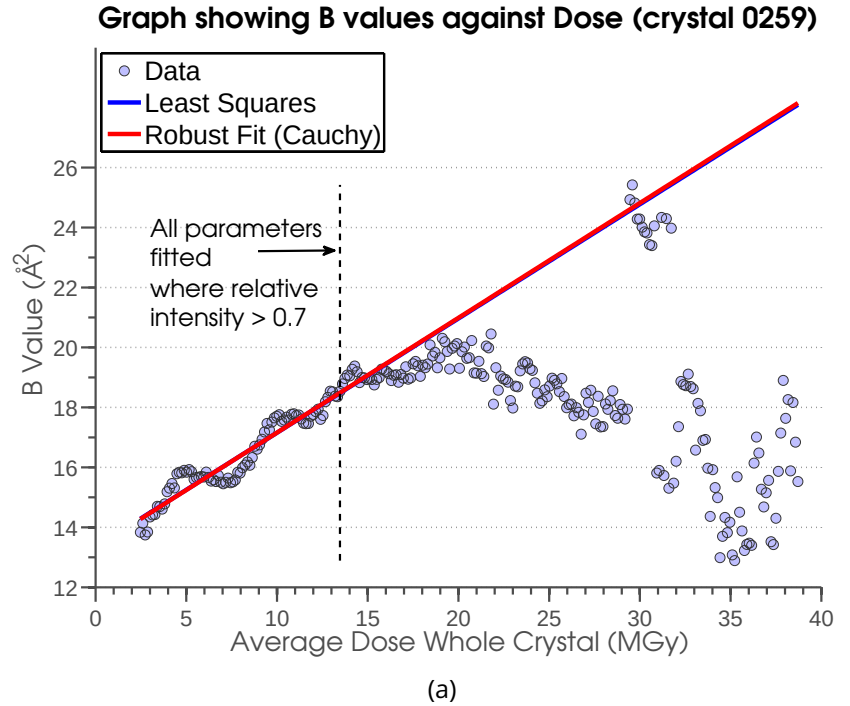


Figure 2.7: Logarithm of the expected intensity values against resolution for 72 different protein crystals which were scaled together (Popov and Bourenkov, 2003). The proteins had different folds, molecular masses, space groups, data collection resolutions and were collected at different temperatures (both RT and cryo). This curve is reproduced from data kindly provided by Dr. Gleb Bourenkov.

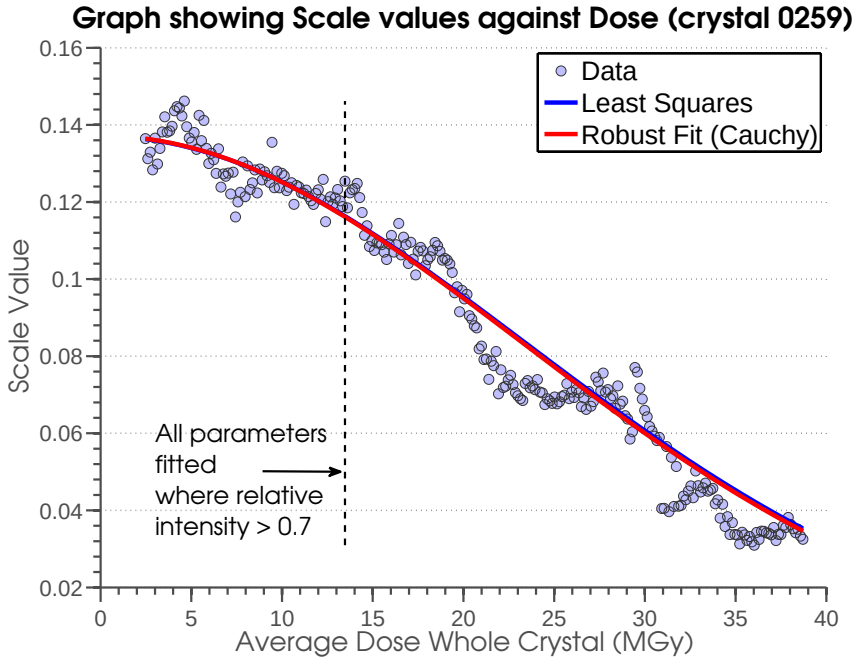
procedure to reduce the influence of errors on the fit.

The parameter values for the Sygusch & Alliare model (k_0', k_1 and k_2) were found using an iterative numerical minimisation procedure for which an objective function (i.e. a function that returns a value to be minimised) was created. The objective function was constructed according to the flowchart in figure 2.9. This function was minimised using the Matlab `lsqnonlin` procedure which in turn uses the trust-region-reflective algorithm (Coleman and Li, 1996). The results are shown in Figure 2.10. The fit to the relative intensity data is very good and the crystal proportions change in a similar manner to that found in Owen *et al.* (2014). It is worth noting that numerical minimisation procedures can return values that correspond to local minima in the parameter space, and therefore the initial choice of values for k_0', k_1 and k_2 will affect the resulting values returned by the procedure. Initial values in this case were determined by inspection to make sure that the calculated relative intensity values were a close match to the measured intensities as possible. A full parameter analysis has not been performed and therefore the values for k_0', k_1 and k_2 presented in Table 2.5 may not be the combination that provide the best overall fit to the data. Thus more analysis

outlier is large. Thus using a function that increases less than the square decreases the influence of outliers in the fitting.



(a)



(b)

Figure 2.8: (a) B values plotted against average dose (whole crystal). Initially the B values increase linearly as reported in the literature (Kmetko *et al.*, 2006; Warkentin and Thorne, 2010; Leal *et al.*, 2012). However the linearity begins to break down as radiation damage progresses. The group of data points in the top right are possibly an artefact of unstable data processing for highly damaged datasets. This was a major factor in the decision to fit the parameters in Tables 2.5 and 2.4 using only data where the relative intensity was above a given threshold value. The threshold relative intensity value of 0.7 was chosen since other data quality indicators suggest that data beyond this point become significantly biologically compromised (Owen *et al.*, 2006). (b) Scale values plotted against the average dose (whole crystal). The blue and red curves are fits to the data using equation 2.3.9 and are almost coincident. The same function was also used successfully to fit data from crystals irradiated at room temperature (Leal *et al.*, 2012), suggesting that this function is suitable for parameterising both cryo and room temperature data.

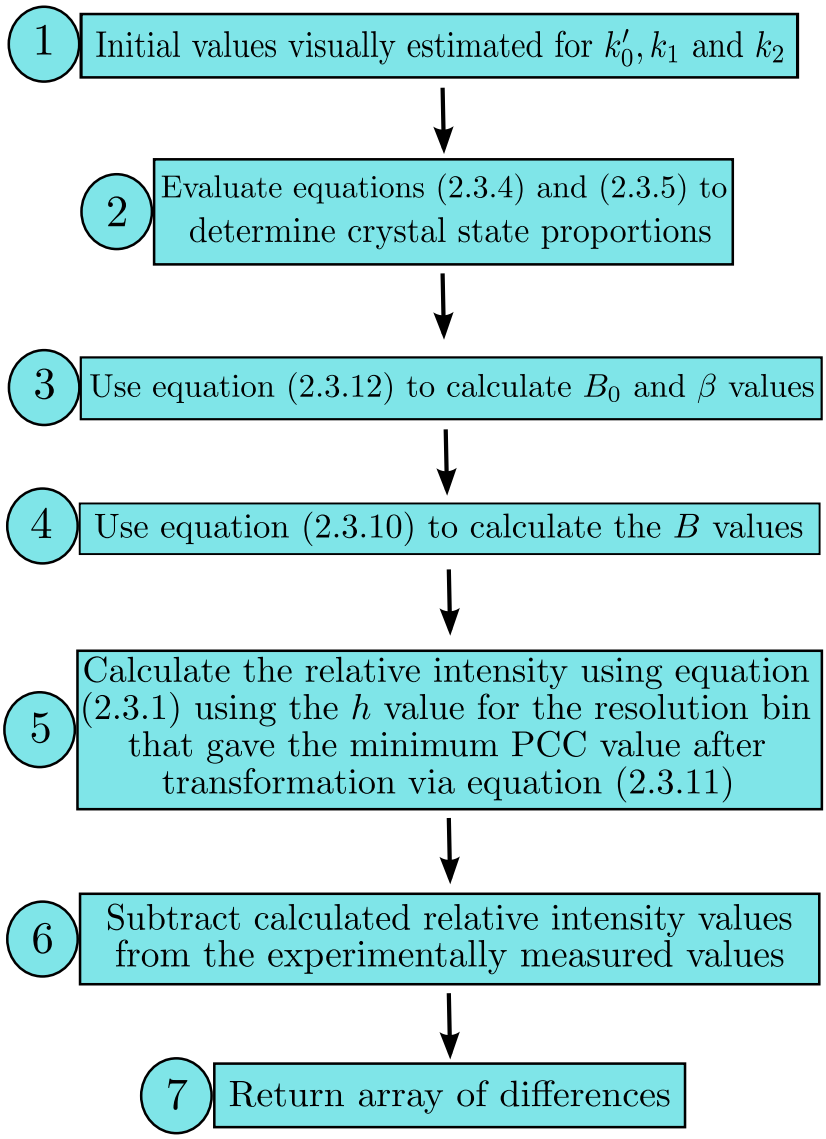


Figure 2.9: Flow chart outlining the structure of the objective function used to find best fit parameters for k_0', k_1 and k_2.

is required before the model can be fully evaluated, but these results show the model to be very promising.

2.4 DDM comparison results

2.4.1 Uniform irradiation

To validate the subsequent analysis performed to find a suitable DDM, it was essential that the crystals were uniformly irradiated. A homogeneous dose distribution resulting from uniform irradiation would mean that the various dose metrics: average dose, maximum dose and DWD, should all give the same value. Simulations performed in RADDOSE-3D show that all crystals were completely immersed in the X-ray beam and the predictions displayed showed a very homogeneous dose distribution (Figure 2.11).

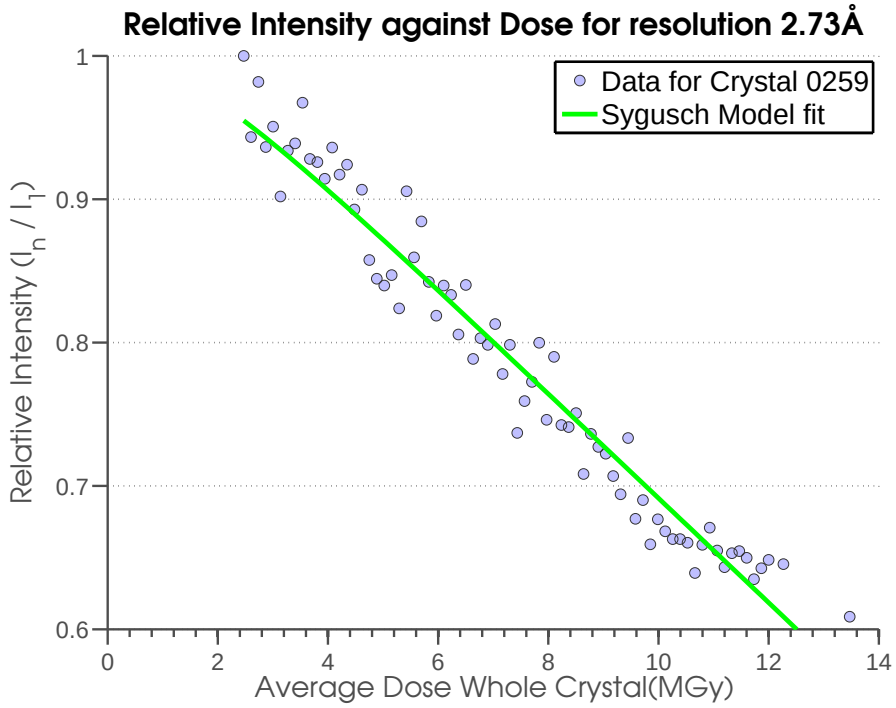
2.4.2 Calculating the RDE

To compare the theoretical RDEs calculated from the DDMs with the experimentally measured relative intensities, a spherically symmetric volume integral is performed on each of equations (2.3.2), (2.3.7) and (2.3.8) up to the resolution limit of the data (1.8 Å) at each of the recorded dose values. These values are then divided by the same integral at dose $D =$ 0 MGy to give an estimate of the relative intensities. The resulting equations are given below and plotted in Figure 2.12 along with the experimentally measured relative intensity values I_n/I_1 for the insulin crystals using the parameter values given in tables 2.5 and 2.4.

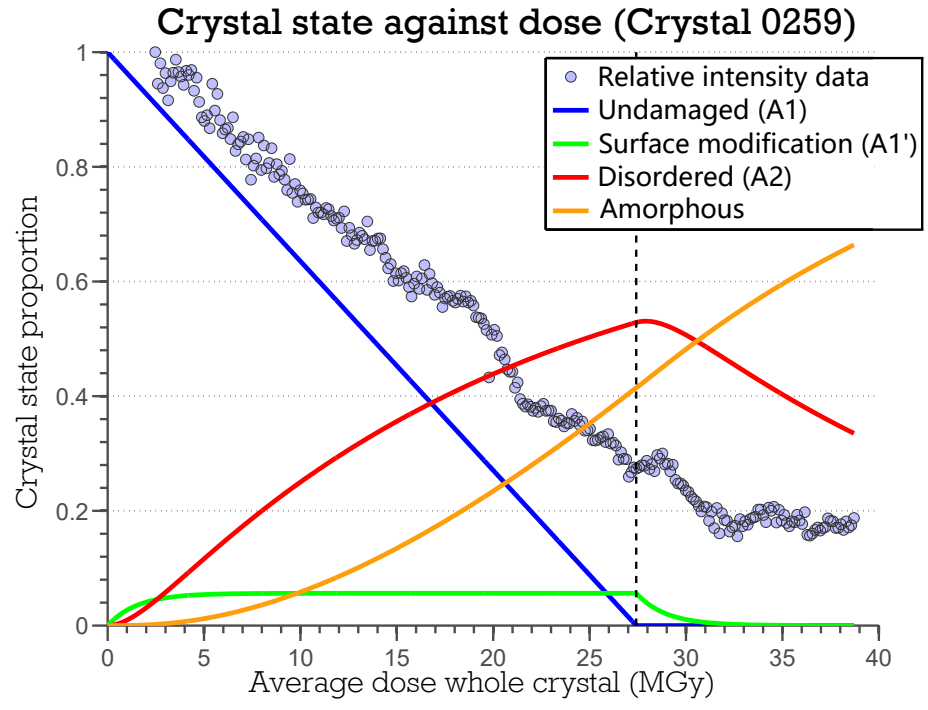
$$\text{RDE Holton} = \frac{\int_{h_{min}}^{h_{max}} h^2 I_{ND}(h) \exp\left[\frac{-\ln(2)\times D\times h}{H}\right] \mathrm{d}h}{\int_{h_{min}}^{h_{max}} h^2 I_{ND}(h)\, \mathrm{d}h} \tag{2.4.1}$$

$$\text{RDE Leal} = \frac{\exp\left[-\gamma^2 D^2\right] \int_{h_{min}}^{h_{max}} h^2 J(h) \exp\left[-\frac{1}{2}(B_0 + \beta D)h^2\right] \mathrm{d}h}{\int_{h_{min}}^{h_{max}} h^2 J(h) \exp\left[-\frac{1}{2}B_0 h^2\right] \mathrm{d}h} \tag{2.4.2}$$

$$\text{RDE Sygusch} = \frac{\int_{h_{min}}^{h_{max}} h^2 I(D=0,h) \left[A_1(D) + A_1'(D) + A_2(D) \exp\left(\frac{-Bh^2}{2}\right)\right] \mathrm{d}h}{\int_{h_{min}}^{h_{max}} h^2 I(D=0,h) \left[A_1(0) + A_1'(0) + A_2(0) \exp\left(\frac{-Bh^2}{2}\right)\right] \mathrm{d}h} \tag{2.4.3}$$

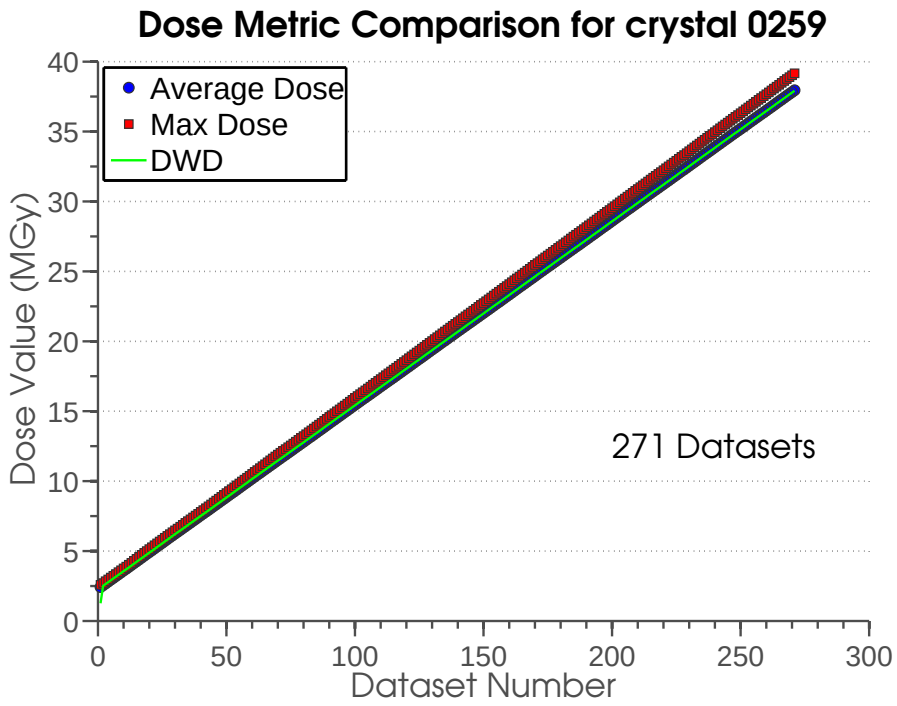


(a)

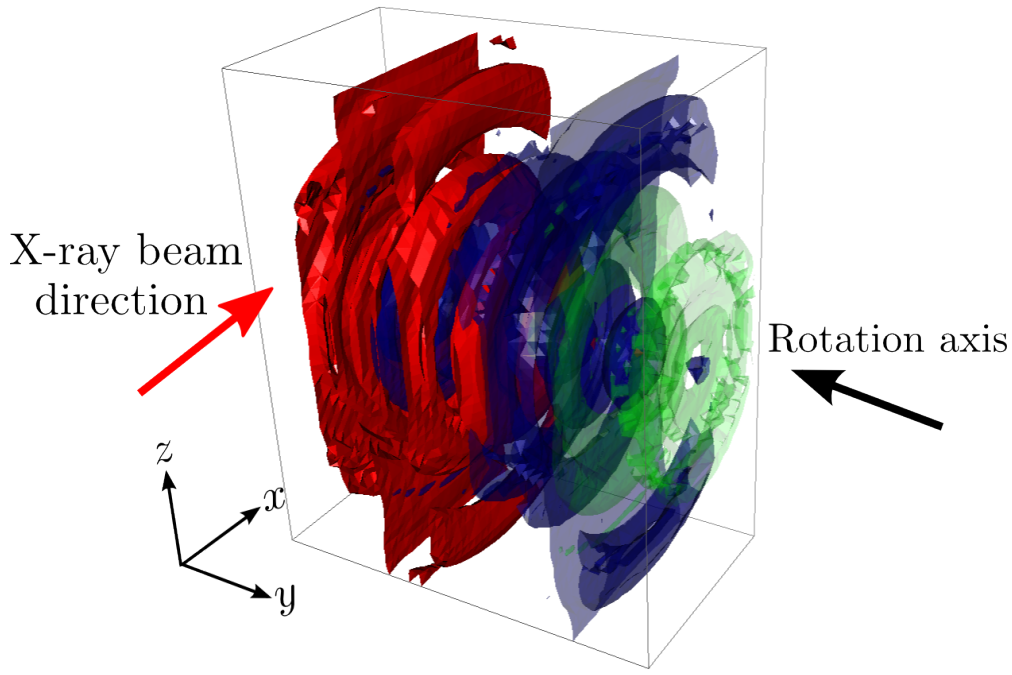


(b)

Figure 2.10: (a) Relative intensity against dose (whole crystal) for the 2.73 Å resolution bin. The Sygusch & Allaire model was fitted to these data to obtain parameter values for k_0', k_1 and k_2 using a non-linear least squares procedure. This resolution bin was chosen for this crystal because the data in the resolution bin gave the best PCC value after transformation via equation (2.3.11). (b) Crystal state proportions as a function of the average absorbed dose (whole crystal) according to the Sygusch & Allaire model. The dose at which all of the A_1 proportion has been converted, $D_L = 27.42\,MGy$, is marked as a black vertical dashed line. The overall relative intensity data, I_n/I_1, for the crystal are also overlaid on the graph.



(a)

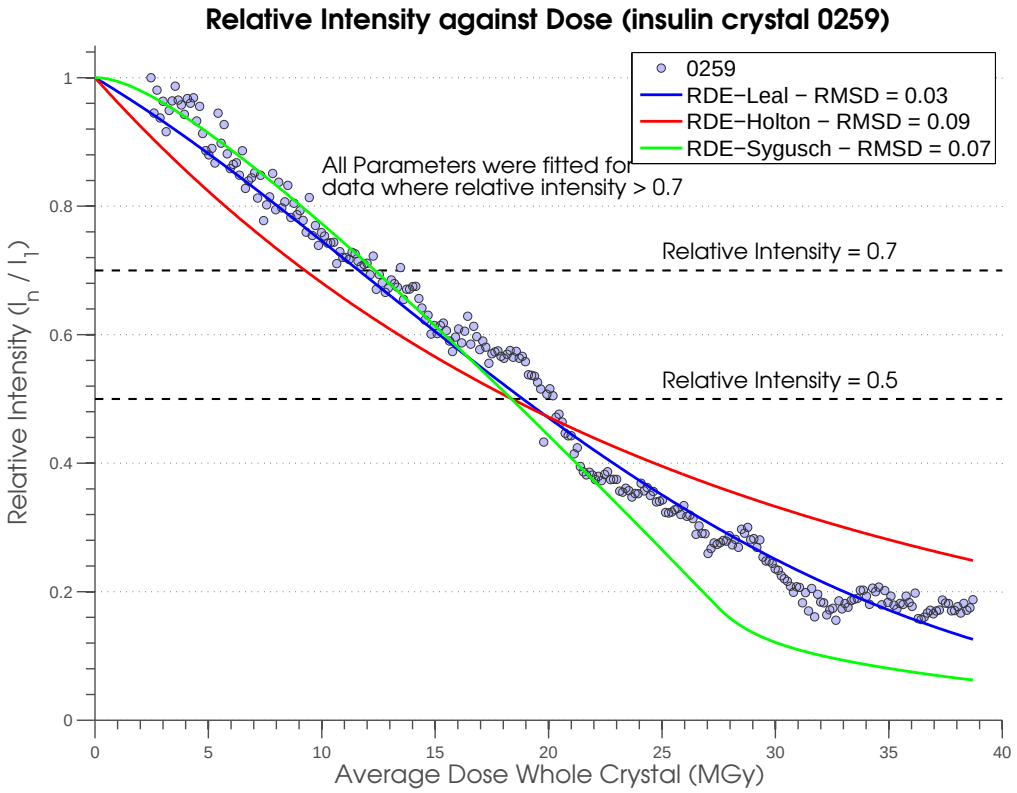


(b)

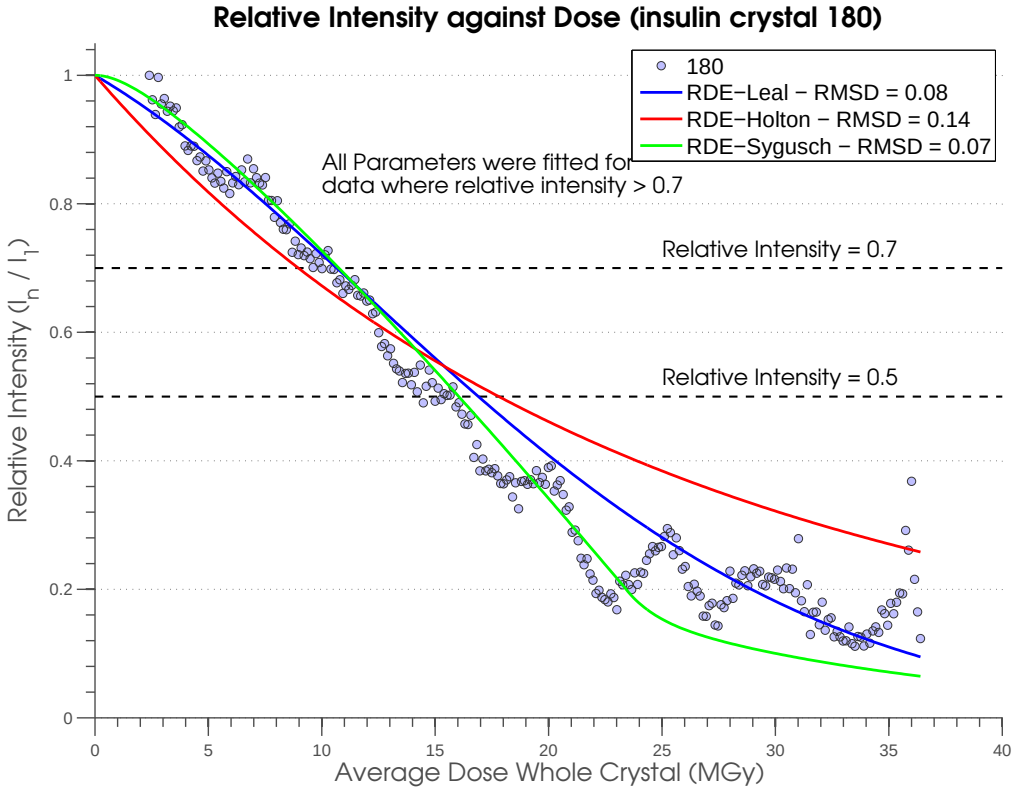
Figure 2.11: (a) The three dose metric values are predicted to be very similar throughout the experiment. Here the DWD values are calculated using equation 1.4.2. (b) The calculated dose map from an MX simulation in RADDOSE-3D of an insulin crystal ($89\,\mu m \times 74\,\mu m \times 40\,\mu m$) exposed to the beam shown in Figure 2.3a over a 1440° rotation. The dose isosurfaces are: green $= 38\,MGy$, dark blue $= 38.5\,MGy$ and red $= 39\,MGy$. The dose distribution is very homogeneous.

The root mean squared deviation (RMSD) was used to assess the fit of the RDE models to the data, with a lower RMSD suggesting a superior fit. Although the parameter values were determined using only the data down to 70% of the initial intensity, the RMSD values are found using the entire range of data for each crystal. Figure 2.12 shows that the RDE using the Leal *et al.* intensity decay model best describes the crystal intensity decay for three out of the five crystals (crystal IDs: 0259, 137 and 128) according to the RMSD. For the other two crystals (crystal IDs: 180 and 172), the RDE Sygusch model gives the lowest RMSD, suggesting that this model is also adequate at describing the relative intensity decay. An important feature of the RDE Sygusch model is that it displays at least two phases during the relative intensity decay. Both phases appear to decay linearly, with the first linear phase decaying faster than the second phase. The RDE Holton model consistently gives the highest RMSD value for all five crystals and the decay curve does not correlate well visually with the decay shown in the data. It should be noted that the RMSD does not account for the different degrees of freedom of the models. The Holton model has fewer degrees of freedom than the other two models and so it would be expected to perform worse. A more representative metric to assess the quality of the fit would be something like the Pearson's chi-squared test, however this has not been performed in this analysis.

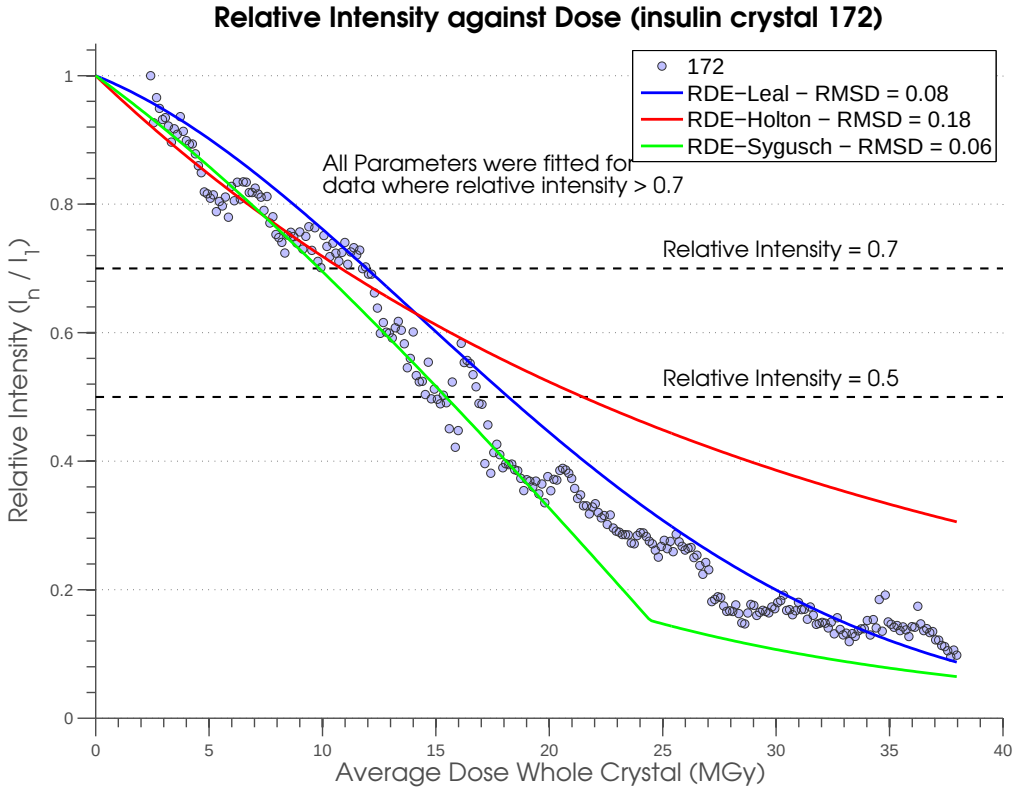
The RDE Leal model was chosen as the one to investigate with the DWD due to its simplicity and superior predictive ability. Figure 2.13 shows the relative intensity values predicted by the RDE Leal model with parameters obtained as described in section 2.3.2 and averaged for each of the 5 insulin crystals processed. However, this time the data were processed to a resolution limit of 1.4 Å (as opposed to the 1.8 Å limit used previously) to ensure that the difference in resolution would not affect the predictive ability of the model.


Relative Intensity against Dose (insulin crystal 0259)
0259
RDE−Leal − RMSD = 0.03
RDE−Holton − RMSD = 0.09
RDE−Sygusch − RMSD = 0.07
All Parameters were fitted for
data where relative intensity > 0.7
Relative Intensity = 0.7
Relative Intensity = 0.5
Relative Intensity (I_n / I_1)
Average Dose Whole Crystal (MGy)
0
0.2
0.4
0.6
0.8
1
0
5
10
15
20
25
30
35
40

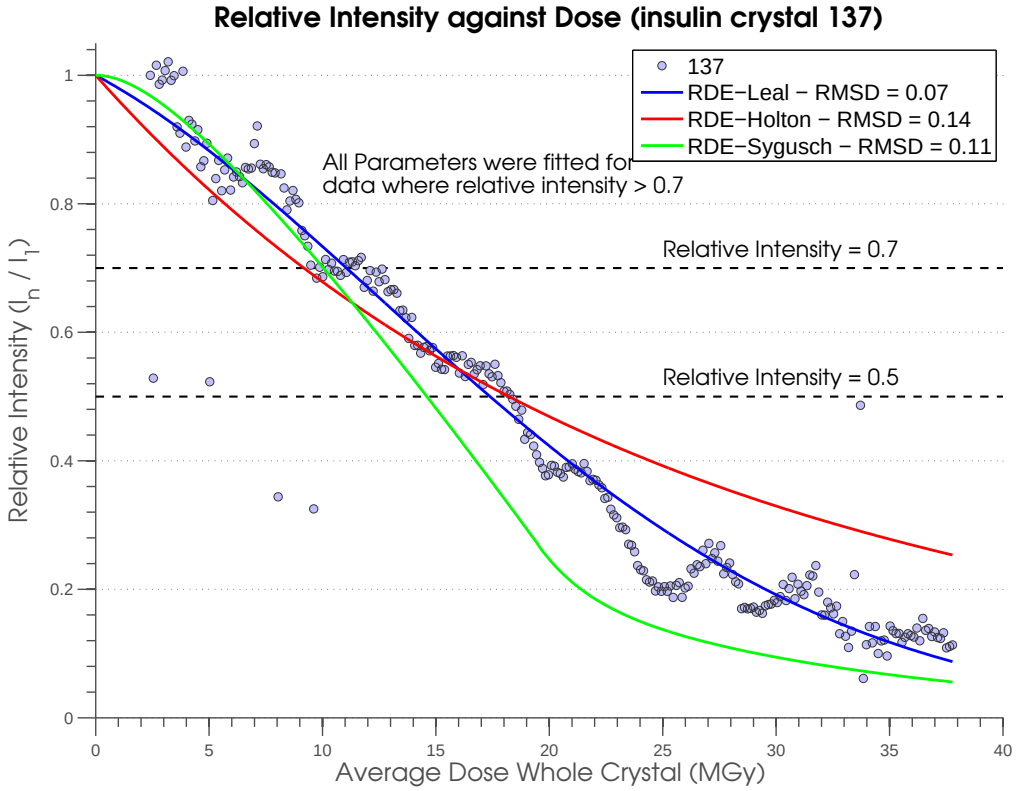
(a)


Relative Intensity against Dose (insulin crystal 180)
180
RDE−Leal − RMSD = 0.08
RDE−Holton − RMSD = 0.14
RDE−Sygusch − RMSD = 0.07
All Parameters were fitted for
data where relative intensity > 0.7
Relative Intensity = 0.7
Relative Intensity = 0.5
Relative Intensity (I_n / I_1)
Average Dose Whole Crystal (MGy)
0
0.2
0.4
0.6
0.8
1
0
5
10
15
20
25
30
35
40

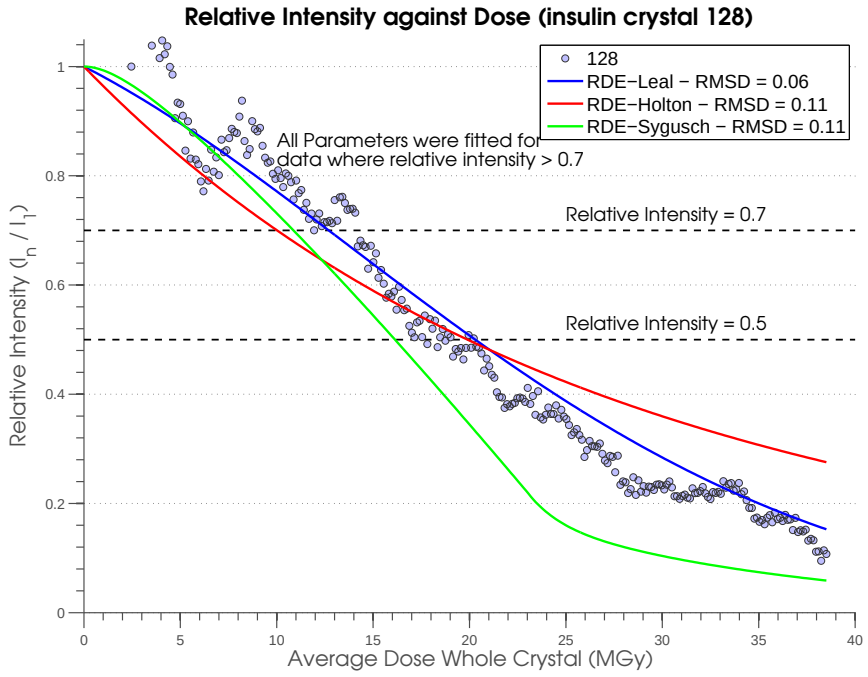
(b)


Relative Intensity against Dose (insulin crystal 172)
172
RDE−Leal − RMSD = 0.08
RDE−Holton − RMSD = 0.18
RDE−Sygusch − RMSD = 0.06
All Parameters were fitted for
data where relative intensity > 0.7
Relative Intensity = 0.7
Relative Intensity = 0.5
Relative Intensity (I_n / I_1)
Average Dose Whole Crystal (MGy)

(c)


Relative Intensity against Dose (insulin crystal 137)
137
RDE−Leal − RMSD = 0.07
RDE−Holton − RMSD = 0.14
RDE−Sygusch − RMSD = 0.11
All Parameters were fitted for
data where relative intensity > 0.7
Relative Intensity = 0.7
Relative Intensity = 0.5
Relative Intensity (I_n / I_1)
Average Dose Whole Crystal (MGy)

(d)



(e)

Figure 2.12: (a)-(e) Relative intensity plotted against the average dose over the whole crystal for each of the insulin crystals. Grey circles represent the experimental data processed to 1.8 Å. Blue, red and green solid lines represent the RDE-Leal, RDE-Holton and RDE-Sygusch models respectively. The RMSDs for each model are given in the figure legend on each plot.

Table 2.4: Best fit zero dose average intensity values ($I_{ND} = I(D = 0, h)$ (arb. units)) for each resolution bin. Note that $1/h_{mid} = d_{mid}$ where d_{mid} is the mid-point resolution in each resolution bin

	Crystal ID				
resolution (Å) ($1/h_{mid}$)	0259	180	172	137	128
12.53	12984	8271	20087	29840	37536
7.23	5161	3342	7857	12277	14732
5.60	7764	4946	11258	17550	21769
4.74	11244	7043	15363	25479	31497
4.18	12772	8000	18242	28230	35589
3.78	11325	7050	16401	25625	31156
3.47	8234	5197	11827	19044	22606
3.23	6035	3787	9008	13885	16543
3.04	4293	2673	6447	9662	11656
2.87	3328	2104	5075	7592	9092
2.73	3021	1904	4426	6903	8122
2.61	2417	1522	3569	5509	6473
2.51	2232	1369	3130	4907	5909
2.41	1913	1181	2697	4261	5101
2.33	1659	1010	2332	3668	4406
2.25	1525	959	2170	3479	3990
2.18	1376	853	1982	3041	3619
2.12	1240	746	1697	2656	3237
2.06	958	607	1341	2159	2451
2.01	871	540	1217	1898	2172
1.96	719	429	922	1483	1783
1.91	645	398	835	1367	1594
1.87	524	324	674	1091	1257
1.83	422	263	534	891	1041

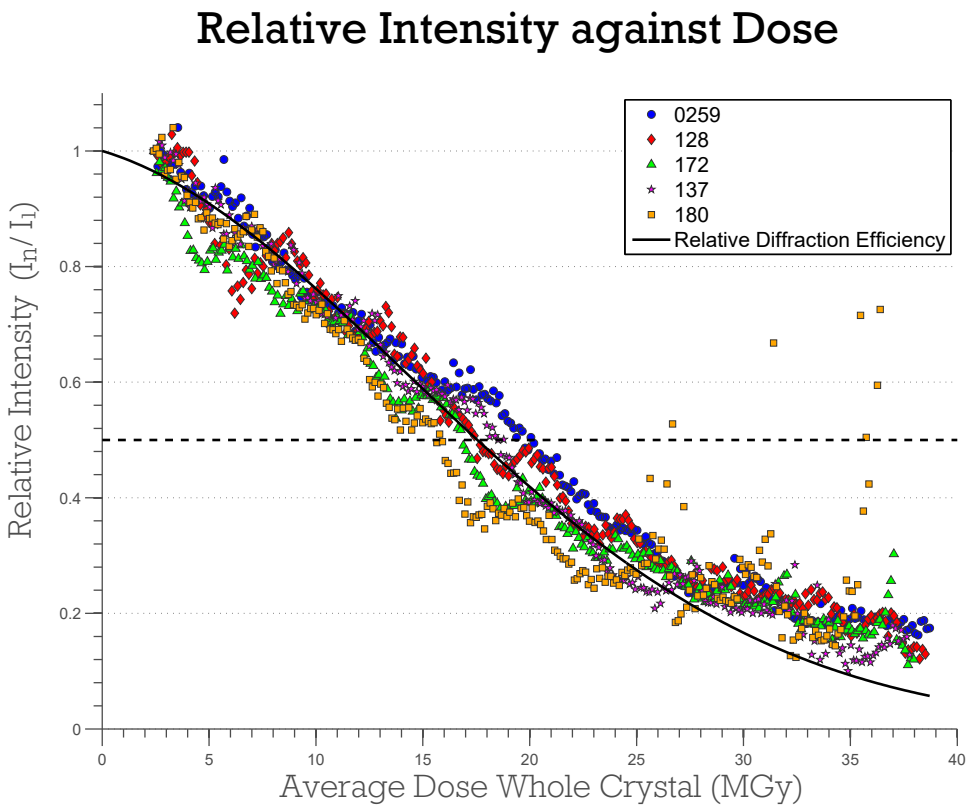


Figure 2.13: Relative intensity plotted against the average dose over the whole crystal for all of the insulin crystals. Discrete data points represent the experimental data for each crystal processed to 1.4 Å. The black solid line is the RDE Leal model with parameters obtained as described in section 2.3.2 and with averaging of the parameter values for each crystal.

Table 2.5: Parameter values for the dose decay models. Note: K does not explicitly appear in equation (2.4.2) and hence is not required for the analysis. The large k_1 parameter value for crystal 172 seems highly unphysical when compared with the other values. This large value implies that any site specific damaged crystal proportion in the crystal immediately becomes disordered, which suggests that the site specific state is present in negligible quantities for this crystal. Accounting for the intrinsic variability of crystals and the fact that the surface modification fraction of crystal 0259 was never higher than 5% of the total crystal fraction, it is not too surprising that at least one of the crystals may show a negligible surface modification fraction.
units: $k_0',\ k_1,\ k_2 \equiv MGy^{-1}, \quad H \equiv MGy\,\text{Å}^{-1}, \quad B_0 \equiv \text{Å}^2 \quad \beta \equiv \text{Å}^2 MGy^{-1} \quad \gamma \equiv MGy^{-1}$

Crystal	k_0'	k_1	k_2	H	B_0	β	γ
0259	3.6×10^{-2}	0.648	5.0×10^{-2}	5.466	13.329	0.383	0.030
180	4.3×10^{-2}	0.772	5.0×10^{-2}	5.274	13.408	0.377	0.036
172	4.1×10^{-2}	2310	5.0×10^{-2}	6.345	14.818	0.266	0.037
137	5.1×10^{-2}	0.473	5.0×10^{-2}	5.391	14.136	0.353	0.036
128	4.3×10^{-2}	0.624	5.0×10^{-2}	5.842	14.538	0.338	0.029

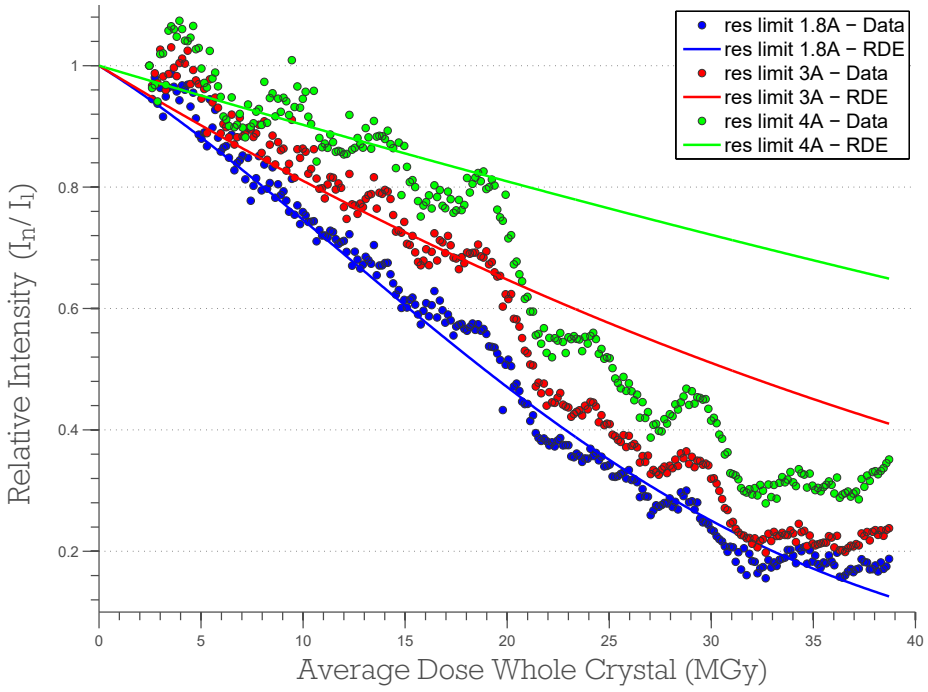
2.5 Further investigation of the RDE Leal model

One of the factors that will affect the parameter values that are obtained from the analysis is the resolution of the data collected (Leal *et al.*, 2012). It is well known that the intensity of higher resolution reflections decay more quickly than for lower resolution reflections.

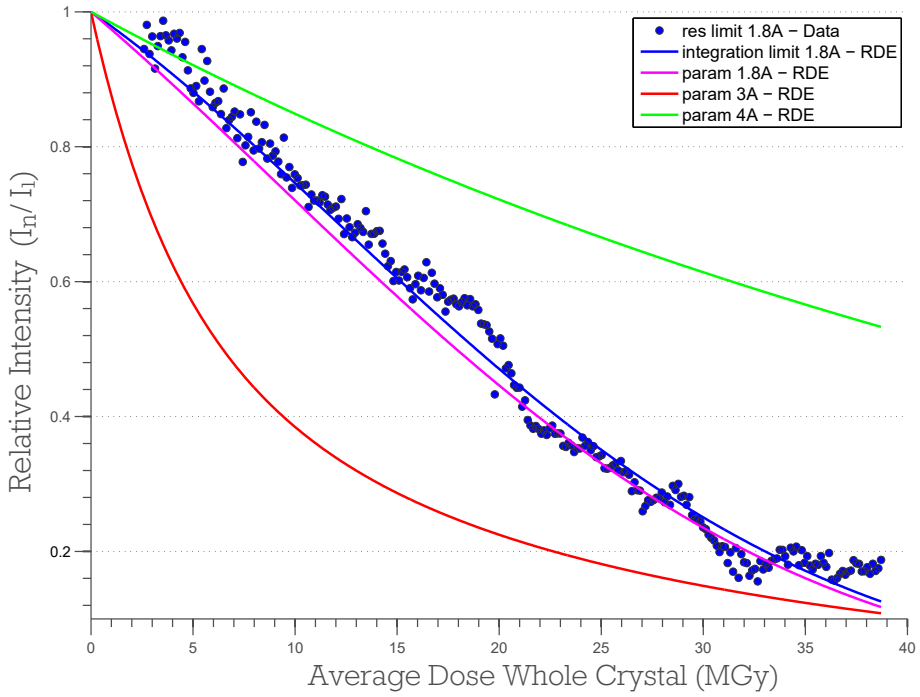
Therefore if the data were collected at a lower resolution, the relative intensity throughout the experiment may be expected to be systematically higher than data collected from the same crystal at a higher resolution. To determine the effect that the resolution dependence would have on the parameter values, the data were scaled to three different resolutions: 1.8 Å, 3 Å, and 4 Å. The parameter values found for each resolution are given in Table 2.6 and the resulting RDE models plotted with the experimental data are shown in Figure 2.14a. It is clear that the calculated relative intensity predictions become worse as the resolution of the data decreases.

The other factor that will affect the results of the RDE model are the resolution limits used to perform the RDE Leal integrals (equation 2.4.2). The resolution of the *BEST* data extend from 10.54 Å to 0.90 Å (Figure 2.7), thus the integrals can be performed between these limits provided the parameter values have been determined. What is unclear is whether performing the integral to the high resolution *BEST* limit, 0.66 Å, would allow the model to converge for the high resolution relative intensity data despite obtaining the parameter values using lower resolution data. Figure 2.14b shows the RDE models derived from processing the data to the three different resolutions (1.8 Å, 3 Å, and 4 Å) but with the integration performed over the resolution limits of the *BEST* data. The magenta, red and green curves correspond to the RDE models using the parameters determined from the data processed to 1.8 Å, 3 Å, and 4 Å respectively. The green and red curves clearly do not converge towards the expected values of intensity decay (blue circles). The green curve suggests that the crystal is much more resilient than the measured 1.8 Å data. The red curve displays an exponential decay shape that is not seen in relative intensity decay data of crystals at cryotemperatures. On the other hand the magenta curve looks much more reasonable for the decay of cryotemperature data.

These results suggest that to obtain the best predictions of the RDE from the data, it is necessary to collect the data to the highest resolution possible. Furthermore, if the data are collected to a high enough resolution, then calculating the RDE using the *BEST* limits should give a more representative model of the true RDE of the crystal. This is because the integral is performed over a large resolution range regardless of the diffraction resolution, which will affect the apparent radiation sensitivity.



(a)

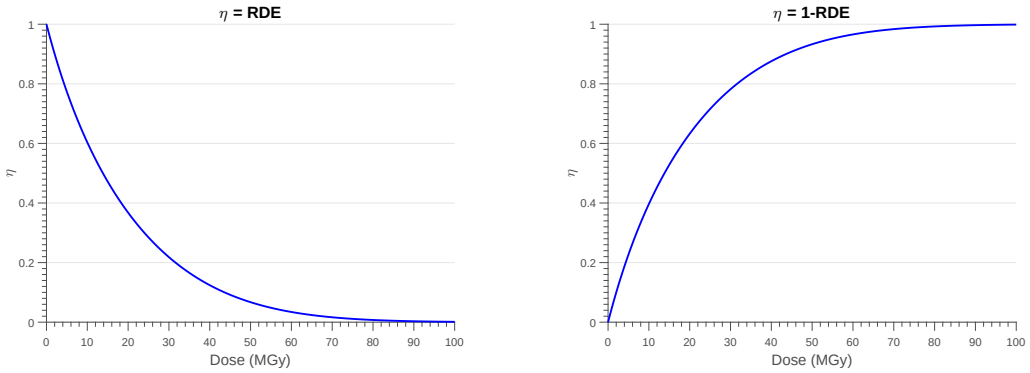


(b)

Figure 2.14: (a) Relative intensity data (circles) plotted with the calculated relative intensity using equation 2.4.2, with the corresponding parameter values for each resolution limit from Table 2.6. (b) Relative intensities for the data processed to 1.8 Å (circles) plotted with the calculated relative intensity using equation 2.4.2 with the corresponding parameter values for each resolution limit from Table 2.6, but the high resolution integral limit used was the limit of the *BEST* intensity data, 0.66 Å. The blue solid line corresponds to the calculated relative intensity value with the high resolution integral limit set at 1.8 Å.

Table 2.6: Parameter values for Leal *et al.* model determined by the method described in section 2.3.2 with data scaled to different resolution limits.

	Parameter Values		
Scaled resolution limit (Å)	B_0 (Å^2)	β (Å$^2 MGy^{-1}$)	γ (MGy^{-1})
1.8	13.329	0.383	0.030
3.0	1.190	0.621	0.010
4.0	46.277	0.536	0.006



(a) (b)

Figure 2.15: Two different forms of η used in equation 2.1.1. (a) $\eta =$ RDE. (b) $\eta = 1 -$ RDE.

2.6 Incorporating the RDE into DWD calculations

The RDE is a monotonically decreasing function of the dose which is bounded between the values of 0 and 1 (equation 2.4.2). If $\eta =$ RDE then the decreasing nature of the RDE means that as the dose increases, the DWD defined in equation 2.1.1 will decrease at some point. This may be a desirable characteristic if only diffraction is to be considered. However if the DWD is used to characterise the damage in a crystal, then it would be expected to always increase with increasing radiation exposure. For this reason two forms of η were investigated along with the simple DWD where η had no dose dependence:

$$\eta = 1 \qquad \text{simple } \eta \text{ form}, \tag{2.6.1}$$

$$\eta = \text{RDE} \qquad \text{decreasing } \eta \text{ form}, \tag{2.6.2}$$

$$\eta = 1 - \text{RDE} \qquad \text{increasing } \eta \text{ form} \tag{2.6.3}$$

Both dose dependent forms are bounded between 0 and 1. The difference between them is that equation 2.6.2 decreases from 1 to 0 whereas equation 2.6.3 increases from 0 to 1 as shown in Figures 2.15a and 2.15b respectively.

The remainder of this section presents the results of the analysis carried out with the data

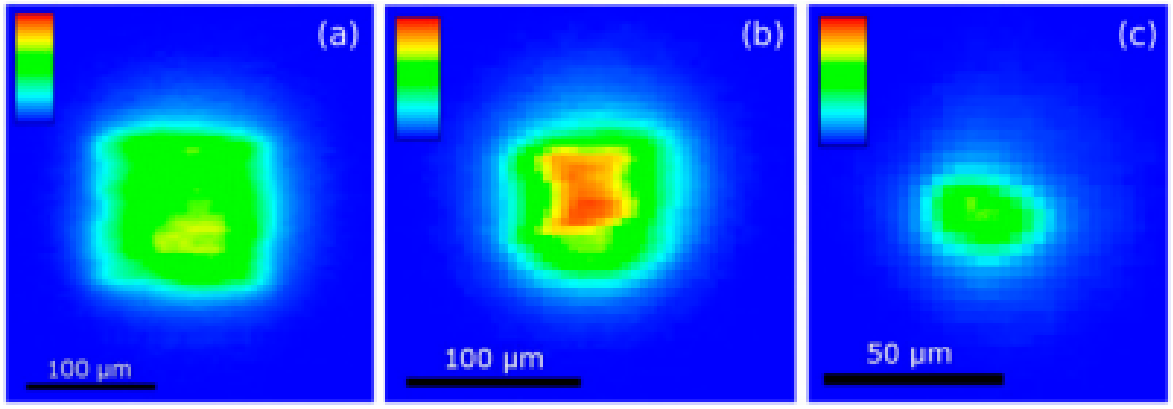


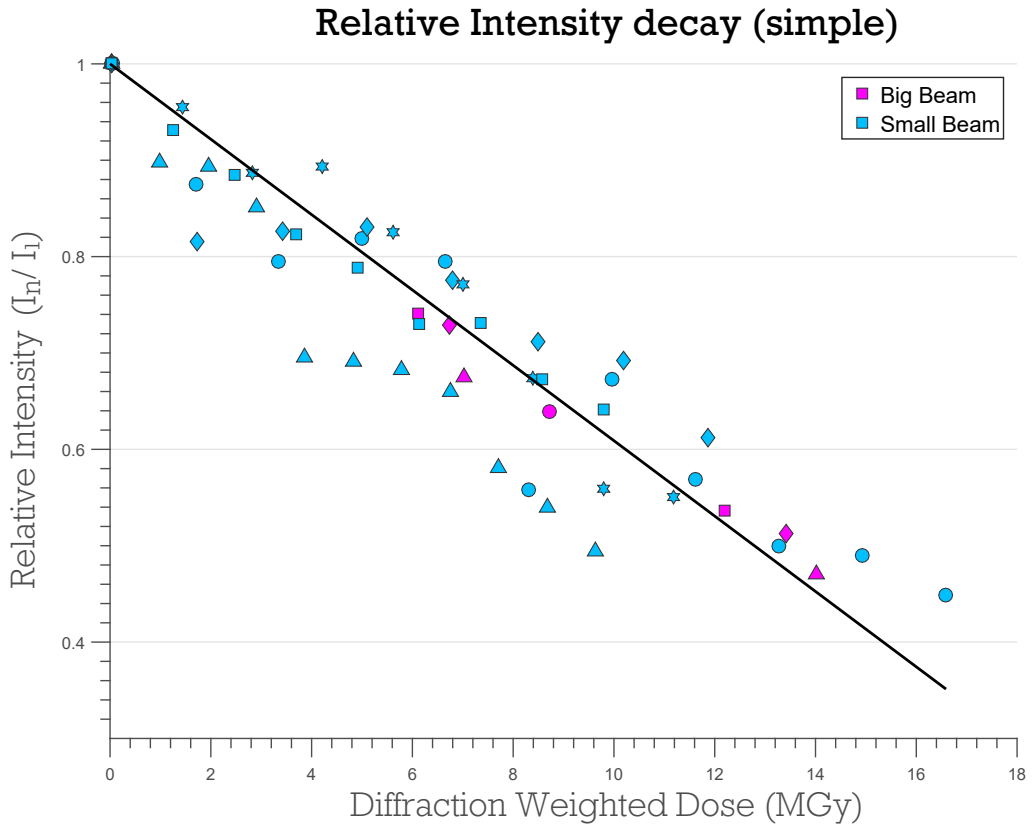
Figure 2.16: False color images of the beam profiles used for the experiments in Zeldin *et al.* (2013). The pixel size is $5 \times 5\ \mu m^2$ in all cases: (a) big beam, (b) medium beam and (c) small beam.

from Zeldin *et al.* (2013) using the forms of η given above, comparing their performance with the simple DWD (equation 1.4.2).

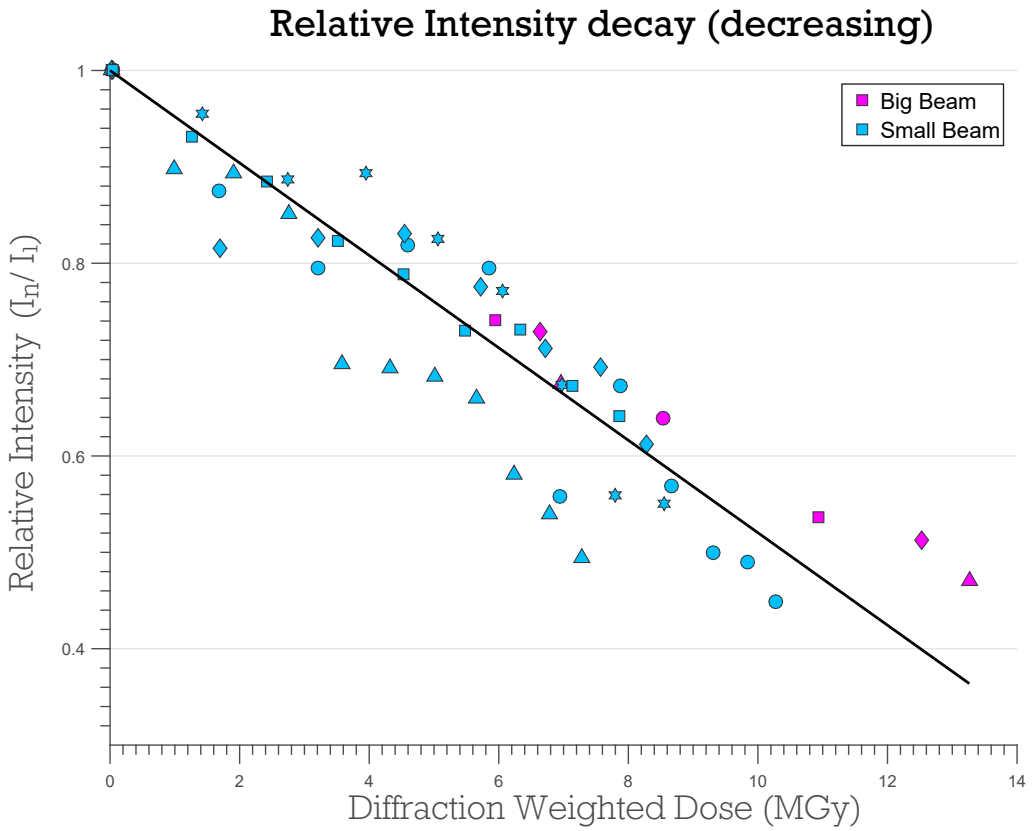
2.6.1 Predicting intensity loss

In the study carried out by Zeldin *et al.* (Zeldin *et al.*, 2013a) cubic crystals of bovine pancreatic insulin were irradiated under different dose contrast conditions. The three beams used in the study: big, medium and small, are shown in Figure 2.16. It was shown that the DWD is a significantly better metric for assessing the extent of radiation damage compared to the average dose for the whole crystal (AD-WC) or the maximum dose, because the spread of relative intensity values around the line of best fit of the data was greatly reduced (plots A-C in Figure 2.21 reproduced from Zeldin *et al.* (2013.)). The data for the big and small beams were available so the analysis performed by Zeldin *et al.* (2013) was repeated for each of the η forms. The results are shown in Figure 2.17. The main difference that can be seen in the plots is the DWD range. The increasing η function results in the largest range of DWD values. Conversely, the decreasing η function reduces the range of DWD values when compared to the simple DWD. Another difference is that introducing a dose dependent η function shifts the big beam data relative to the small beam data. The decreasing η function shifts the big beam data towards higher dose values relative to the small beam data, whereas the increasing function shifts the big beam data towards lower dose values. This suggests that there is a signature from the different beams. However this is an undesirable feature because DWD should already account for the different beam conditions via the flux weighting.

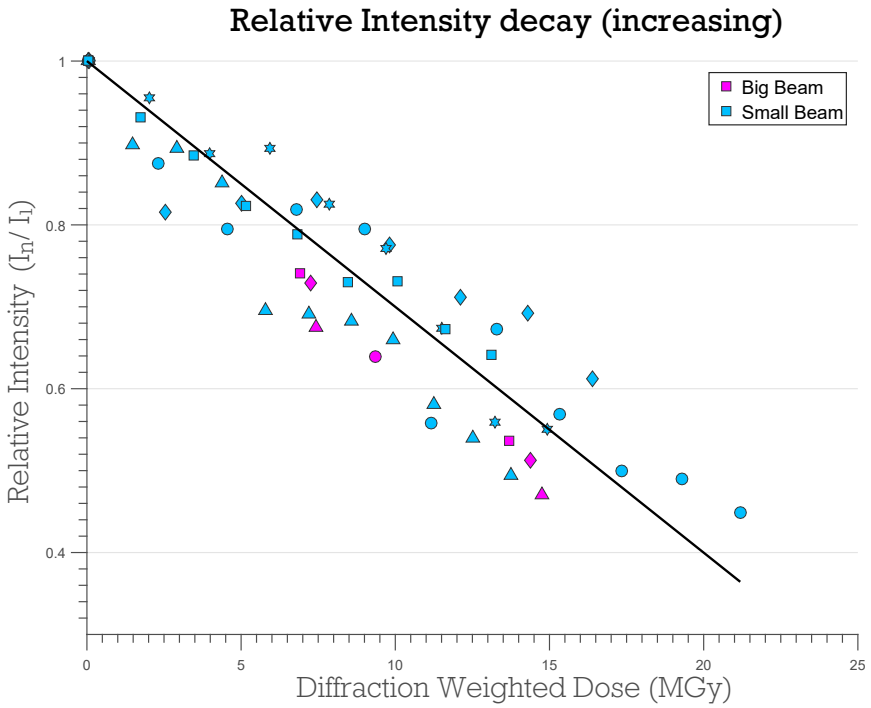
A line of best fit was calculated using a least squares fitting procedure and plotted (solid


Relative Intensity decay (simple)
Big Beam
Small Beam
Relative Intensity (In/ I1)
Diffraction Weighted Dose (MGy)
1
0.8
0.6
0.4
0
2
4
6
8
10
12
14
16
18

(a)


Relative Intensity decay (decreasing)
Big Beam
Small Beam
Relative Intensity (In/ I1)
Diffraction Weighted Dose (MGy)
1
0.8
0.6
0.4
0
2
4
6
8
10
12
14

(b)



(c)

Figure 2.17: Relative intensity decay against the DWD using the different forms of η given by equations 2.6.1, 2.6.2 and 2.6.3. (a) Simple η. (b) Decreasing η. (c) Increasing η.

black line Figure 2.17). A measure of the overall deviation of the data from the line was obtained by calculating the squared Euclidean norm[‖] for each DWD form (Table 2.7). The DWD form that gave the lowest value for the squared Euclidean norm was the simple DWD form. This suggests that the data are less spread overall without adding a dose dependent form for η.

Table 2.7: Squared Euclidean norm values of the line of best fit with the data in Figure 2.17.

η form	Squared Euclidean norm
$\eta = 1$ (simple)	0.203
$\eta = RDE$ (decreasing)	0.211
$\eta = 1 - RDE$ (increasing)	0.206

$D_{1/2}$ is a metric of the radiation sensitivity of a protein crystal that is defined as the dose at which the relative intensity falls to 50%. $D_{1/2}$ values were calculated from the line of best fit for each of the DWD forms. Furthermore s_{AD} values (as defined in section 1.4.4) were also calculated and both sets of values are given in Table 2.8. The spread of the $D_{1/2}$ values for the simple DWD, 0.38 MGy, is much smaller than the spread for the decreasing η form, 2.27 MGy, and the increasing η form, 3.87 MGy. This confirms the result that the spread of the data is increased by incorporating the dose dependent η forms. The ranges of s_{AD} values

[‖] The squared Euclidean norm is defined as $\sum_i (f(x_i) - y_i)^2$.

Table 2.8: $D_{1/2}$ and s_{AD} values calculated using DWD with different η forms. The DWD with the simple η form results in the most similar $D_{1/2}$ values, whereas using the increasing or decreasing η forms gives significantly different $D_{1/2}$ values. The differences in the ranges of s_{AD} are relatively large and are not significantly altered by including the various η forms.

	$D_{1/2}$ (MGy)			s_{AD} (Å^2/MGy)		
Beam size	Simple	Decreasing	Increasing	Simple	Decreasing	Increasing
Big beam	12.94	12.18	13.97	0.0125	0.0133	0.0116
Small beam	13.32	9.91	17.84	0.0068	0.0092	0.0050

are not greatly improved by adding the dose dependent forms of η to the DWD equation. However, given the fact that DWD does not significantly reduce the data spread for the B_{rel} metric (Zeldin *et al.*, 2013a), a reduction in the spread of s_{AD} values was not expected.

2.6.2 Offset simulations

To quantify the efficiency of a given data collection strategy, Zeldin *et al.* introduced a metric called the diffracted dose efficiency (DDE) defined as the ratio of elastically scattered photons to DWD. The DDE states the number of elastically scattered photons that are diffracted per unit dose and it is this quantity that should be maximised for a given experiment. To explore this metric, the authors simulated experiments where the rotation axis was offset from the beam axis by various distances to determine how spreading the dose affected the DDE. These simulations were repeated for this work with each η form incorporated into the DWD. In all simulations a cuboid crystal of various sizes and "average" crystal absorption (absorption coefficient = 0.237 mm^{-1} (Zeldin *et al.*, 2013b)) was exposed to a Gaussian profile beam with a 20 $\times$ 20 μm^2 FWHM, a flux of 5 $\times$ 10^{11} ph/s and incoming photon energy of 12.4 keV (1 Å). The collimation was rectangular and set to 40 $\mu m\times$ 40 μm and the total exposure time was 60 seconds. The results of these simulations with each of the forms of DWD are presented in Figure 2.18 for two cubic crystals with edge lengths of 400 μm and 60 μm.

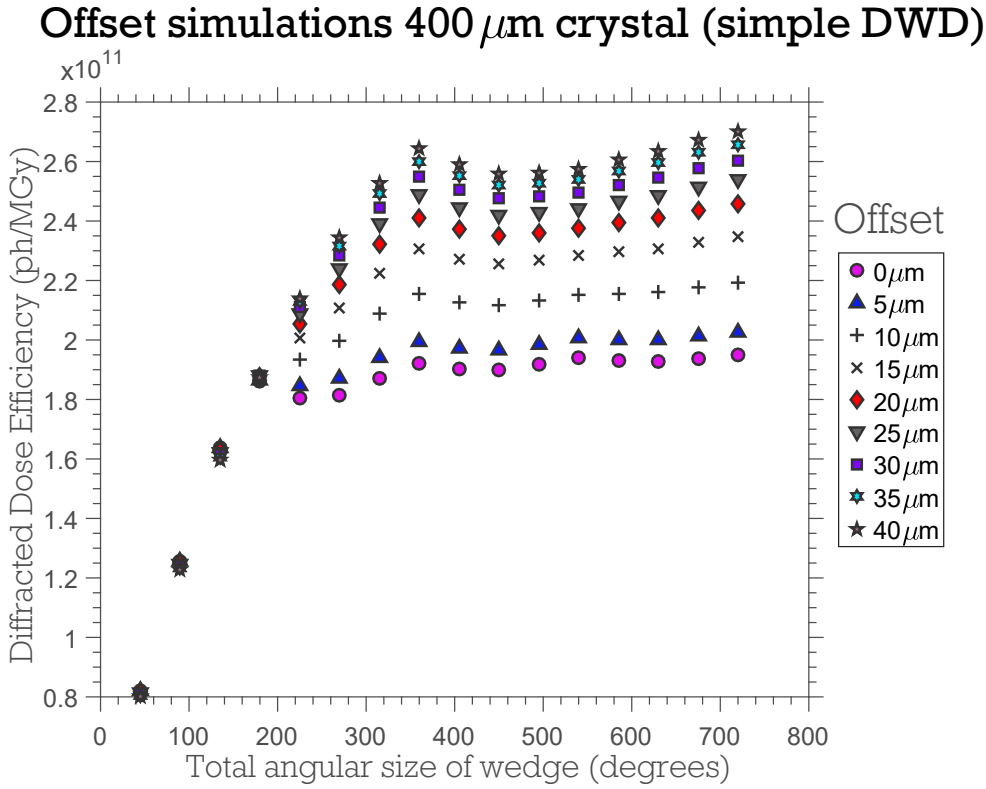
For the 400 μm crystal, Figures 2.18a, 2.18b and 2.18c show that the DDE increases with a greater rotation range up to a 360° rotation. However the decreasing η form of the DWD suggests that the DDE values are generally higher than for the simple form, whereas the opposite is true for the increasing η form. This means that the relative improvements in the diffraction efficiency are different for the different η forms. Even more important is that the inferences that would be made from the simulations are different depending on which form of η is used. The simple η form suggests that there is an almost undetectable difference

in the DDE values for the different offsets until the rotation angle is larger than 180°. On the other hand, using the decreasing η form, suggests that offsetting the rotation axis is detrimental for a rotation range below 180°, but for angles larger than 180° the offsetting becomes beneficial. The situation is very different for the increasing η form which suggests that there is negligible difference in offsetting strategy for angles up to 90°, but for rotations ranges larger than 90° it is beneficial to offset the rotation axis.

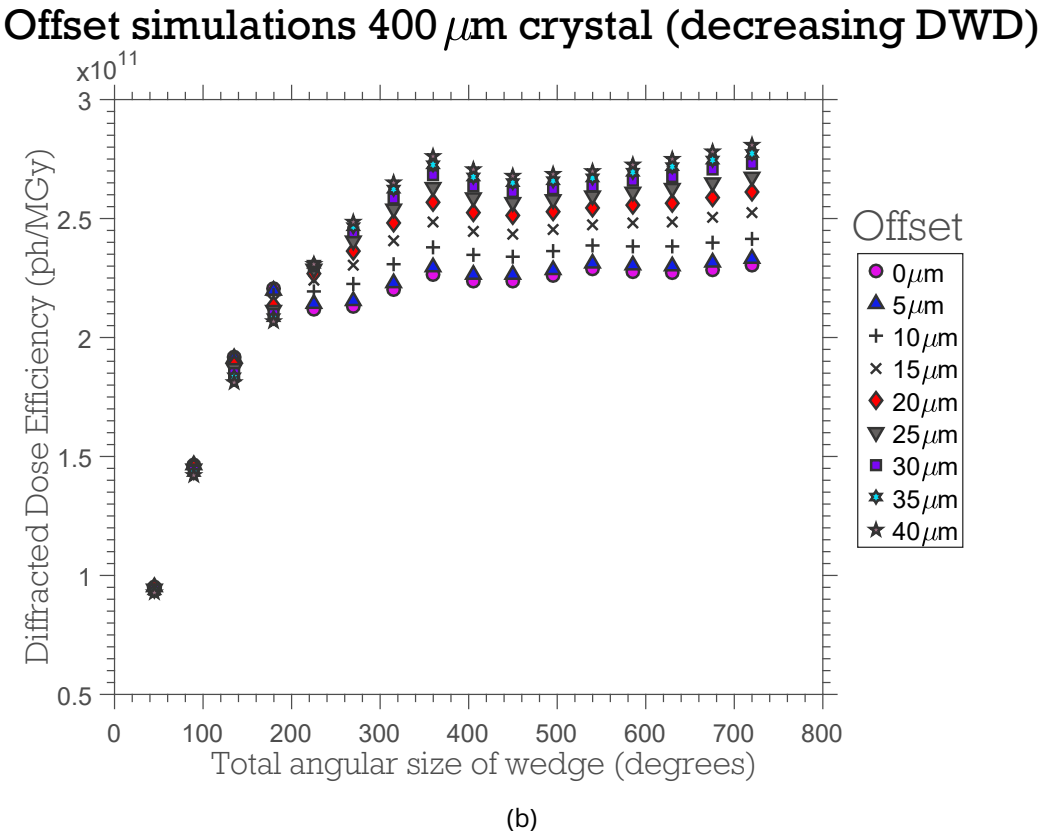
For the 60 μm crystal (Figures 2.18d, 2.18e and 2.18f) the differences are less pronounced. Of course the DDE values differ, again the decreasing η form suggests an increase in DDE values whereas a decrease in DDE is predicted with the increasing η form. All η forms suggest that no improvement in DDE is gained by offsetting with rotation angles up to 180°, but from 270° upwards it is best to offset the rotation axis by 15 μm from the beam axis. The main differences result from offsetting the crystal by either 5 μm or 25 μm with a rotation range larger than 270°. The simple η form suggests that offsetting by either distance gives similar DDE values but the better distance to offset fluctuates with the overall rotation angle. In contrast, the decreasing η form suggests that offsetting by 5 μm produces superior DDE values than offsetting by 25 μm. However the increasing η form suggests the opposite i.e. for a rotation range larger than 360° it is always better to offset by 25 μm.

2.6.3 Offset experiment

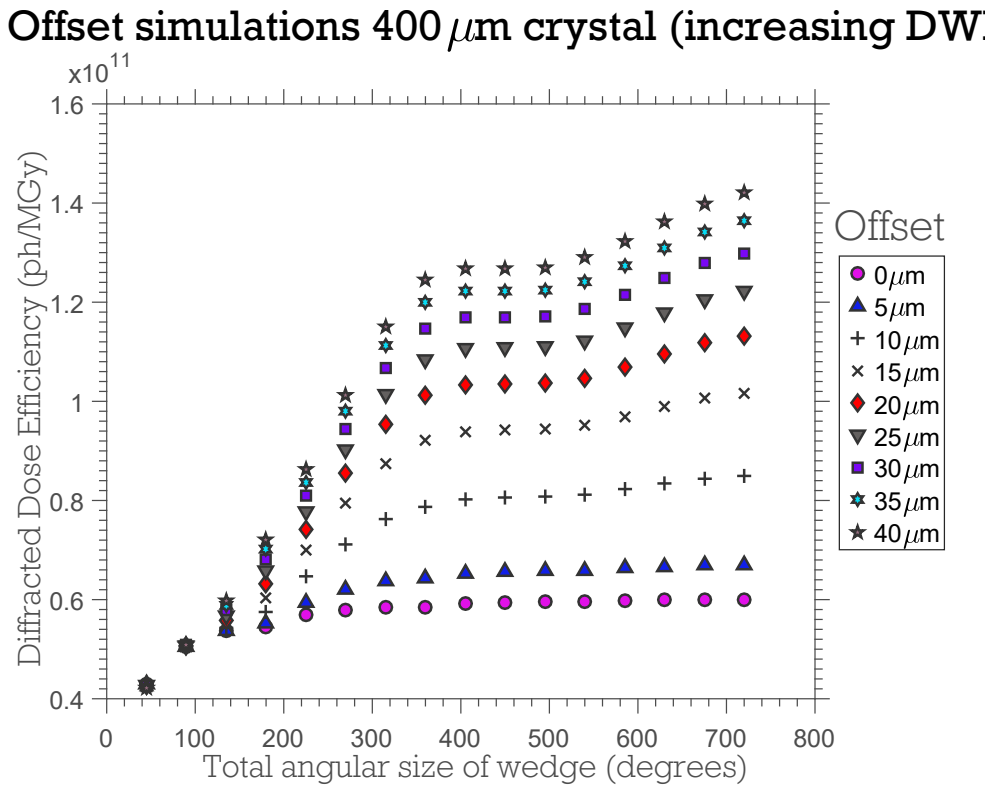
An experiment to validate the offset simulations was performed by Zeldin *et al.* and reported in (Zeldin *et al.*, 2013a). In the experiment, a cuboid crystal of bovine pancreatic insulin (460 $\times$ 550 $\times$ 260 μm^3) was irradiated in two regions with a beam of approximate size 40 $\times$ 70 μm^2 (Figure 2.19). The first position corresponded to the rotation axis aligned with the beam axis (standard strategy). The second position corresponded to the rotation axis offset from the beam axis by 1.25 times the beam FWHM (50 μm - offset strategy). At each of the two positions three datasets were collected: first a 180° low dose *probe* dataset, then a high-dose burn dataset, and then finally another low dose *probe* dataset to evaluate the damage state of the crystal after it had been subjected to a high dose X-ray exposure. To achieve the same DWD value (for the simple η = 1 form) for each of the high dose datasets, the standard strategy exposed the crystal for 126 seconds whereas for the offset strategy the exposure was 162 seconds. The final dose state of the crystal as calculated in RADDOSE-3D is shown


Offset simulations 400 μm crystal (simple DWD)
x10^11
Diffracted Dose Efficiency (ph/MGy)
Total angular size of wedge (degrees)
Offset
0μm
5μm
10μm
15μm
20μm
25μm
30μm
35μm
40μm

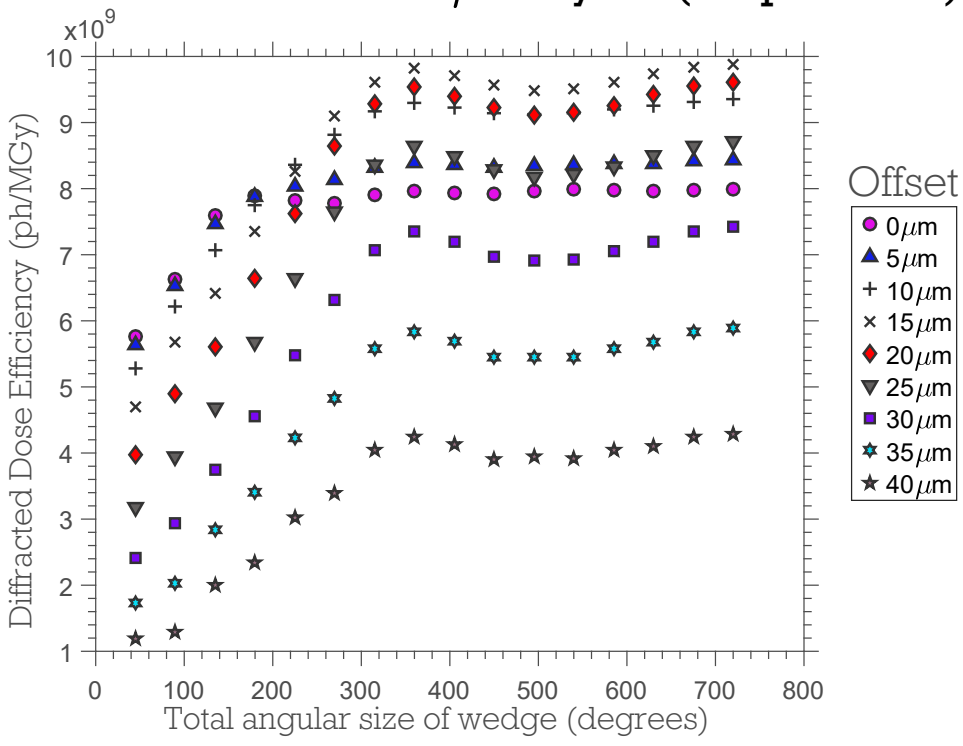
(a)


Offset simulations 400 μm crystal (decreasing DWD)
x10^11
Diffracted Dose Efficiency (ph/MGy)
Total angular size of wedge (degrees)
Offset
0μm
5μm
10μm
15μm
20μm
25μm
30μm
35μm
40μm

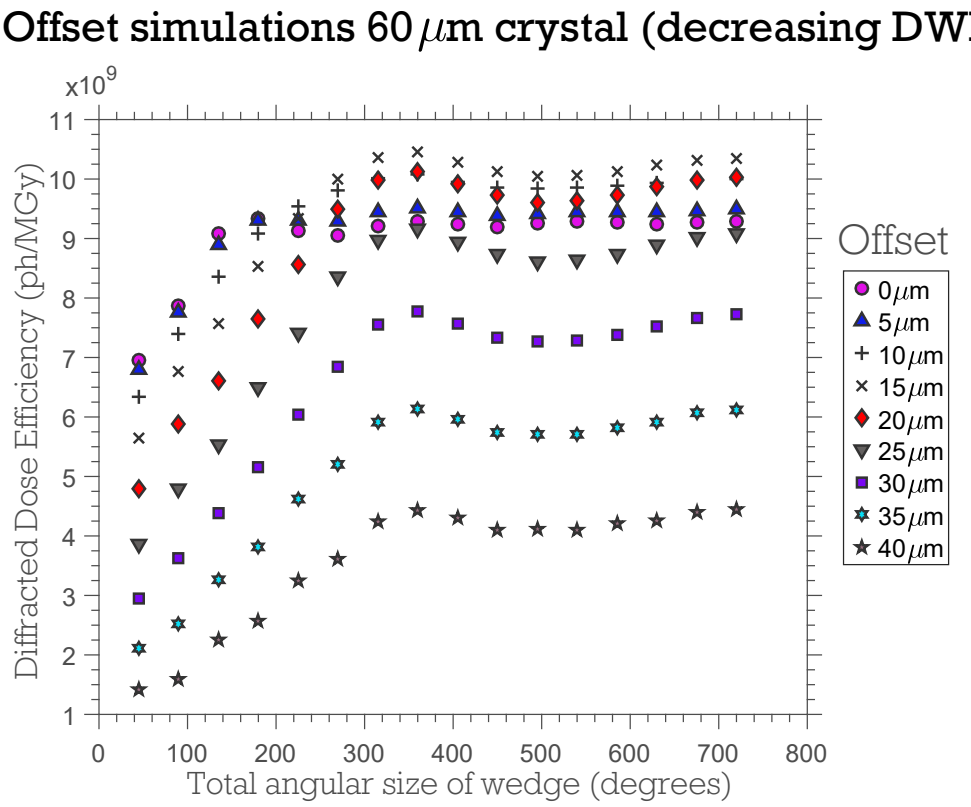
(b)


Offset simulations 400 μm crystal (increasing DWD)
x10^11
Diffracted Dose Efficiency (ph/MGy)
Total angular size of wedge (degrees)
Offset
0μm
5μm
10μm
15μm
20μm
25μm
30μm
35μm
40μm

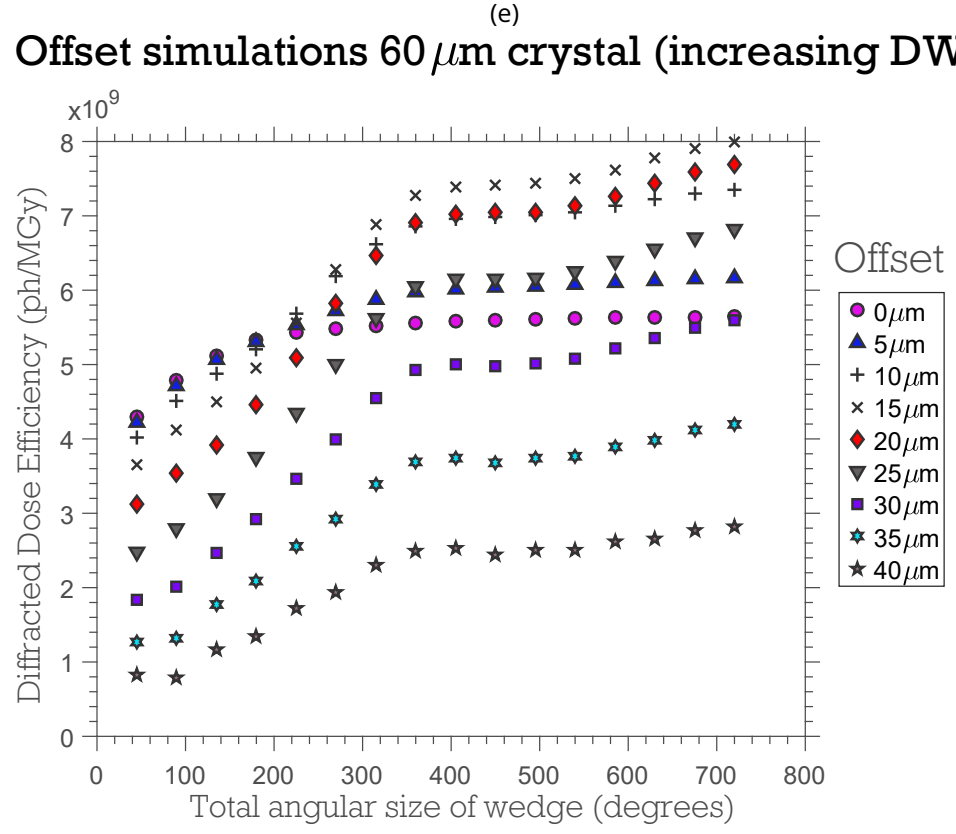
(c)


Offset simulations 60 μm crystal (simple DWD)
x10^9
Diffracted Dose Efficiency (ph/MGy)
Total angular size of wedge (degrees)
Offset
0μm
5μm
10μm
15μm
20μm
25μm
30μm
35μm
40μm

(d)



(e)



(f)

Figure 2.18: Results from the offset simulations showing the diffracted dose efficiency (DDE) plotted against the total rotation range for two different sized cubic crystals: 400 μm and 60 μm edge lengths. The DDE values are calculated from the DWD using the different forms of η given by equations 2.6.1, 2.6.2 and 2.6.3

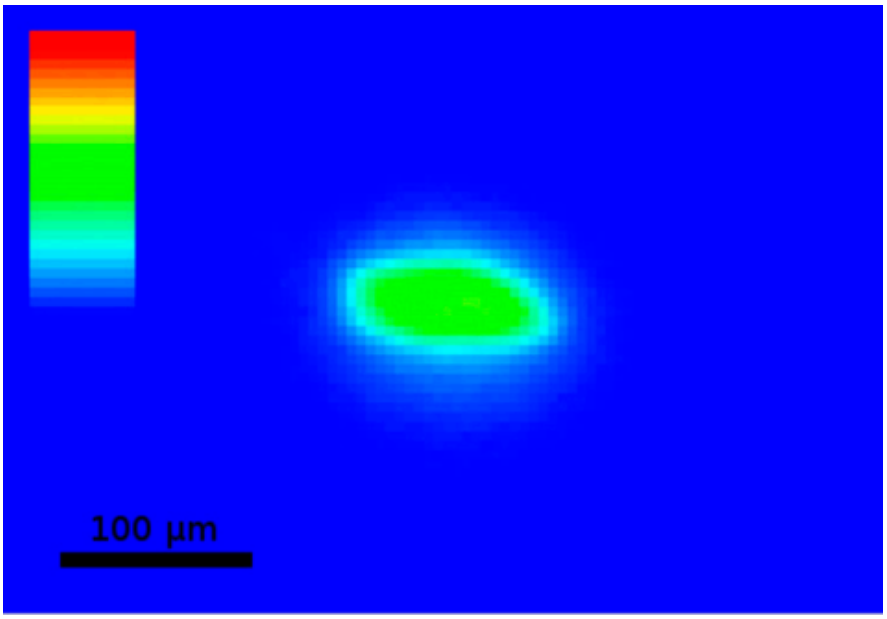


Figure 2.19: False colour images of the beam profile used in the offset experiment by Zeldin *et al.* (2013). The colour bar represents 0-255 intensity units.

in Figure 2.20. The results from the data processing of this experiment with the different forms of DWD can be seen in Table 2.9. The resulting DWD values for this experiment tell

Table 2.9: The rows are presented in chronological order of the experiment. Syntax of the first column is: P1: 1st probe dataset, P2: 2nd probe dataset, HD: high dose dataset, -S: standard strategy, -O: offset strategy.

Wedge	Total time**(s)	Elastic yield (ph)	DWD simple η (MGy)	DWD decreasing η (MGy)	DWD increasing η (MGy)
P1-S	5.4	5.5×10^{10}	0.08	0.08	0.16
P1-O	5.4	5.3×10^{10}	0.08	0.08	0.14
HD-O	162.0	1.6×10^{12}	1.86	1.77	2.85
HD-S	126.0	1.3×10^{12}	2.00	1.80	3.73
P2-S	5.4	5.5×10^{10}	3.81	1.77	7.03
P2-O	5.4	5.3×10^{10}	3.56	3.34	4.83

very different stories about the states of the crystal. For the first probe datasets (P1-S and P1-O), the DWD values obtained using the decreasing η form agrees perfectly (to 2 d.p.) with the resulting DWD values using the simple η form. The DWD values for the increasing η form are practically double that value. For the high dose datasets (HD-S and HD-O) that were designed so that the simple η form DWD was similar for both strategies, the decreasing η form DWD values for the two experiments are the most similar. On the other hand, the increasing form η DWD values differ by almost 1 MGy. This result suggests either that the increasing η form may not actually be representative of the real DWD (i.e. this is the incorrect form), or that the estimated DWD values with the other η forms are misleading. Finally there are several differences in the results for the second probe datasets (P2-S and P2-O). The simple and

**Equivalent at 100% transmission (1.4×10^{12} ph/s)

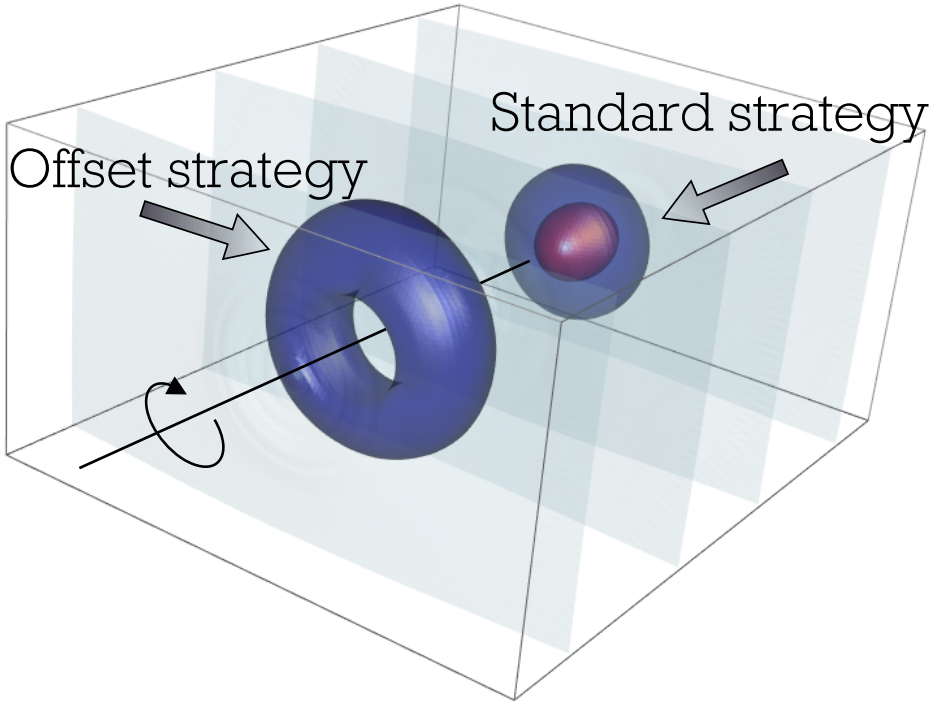


Figure 2.20: Dose isosurface map for the crystal used in the offset experiment. Isosurfaces are at 0.1 MGy (light blue), 5 MGy (dark blue) and 10 MGy (red). Note that the dose at any point in the crystal for the offset experiment is lower than the maximum dose for the standard strategy.

increasing η form DWD values suggest that the standard strategy results in a higher DWD value than the offset experiment. This is the expected result. The main difference however, is the large range in DWD values for the increasing η DWD (2.2 MGy) compared to the range for the simple η DWD (0.25). Given that the relative intensity (I_n/I_1) for the standard and offset strategies for the second probe datasets are 0.79 and 0.85 respectively, the higher DWD difference could account for the relative intensity difference. The DWD values using the decreasing η form for the second probe datasets could be considered counterintuitive at first sight. The result states that the DWD for the standard strategy is lower than the DWD for the high dose dataset which was taken before the second probe dataset. If radiation damage is considered progressive then this result rules out using the decreasing form of η. However it may be that the diffraction quality is better for the probe dataset than the high dose dataset, in which case this form of η does not represent the level of damage, but instead, it is a dose value that describes the quality of the diffraction. This result suggests that the relationship between the relative intensity and the DWD using the decreasing η form is not a one-to-one function. There are many possible relative intensity values for a given DWD value because the DWD can increase and then decrease again. The DWD values are similar for the simple and decreasing η forms for the offset strategy in the second probe experiment. This is due to the fact that the dose for the offset experiment is smaller and hence the reduction of η

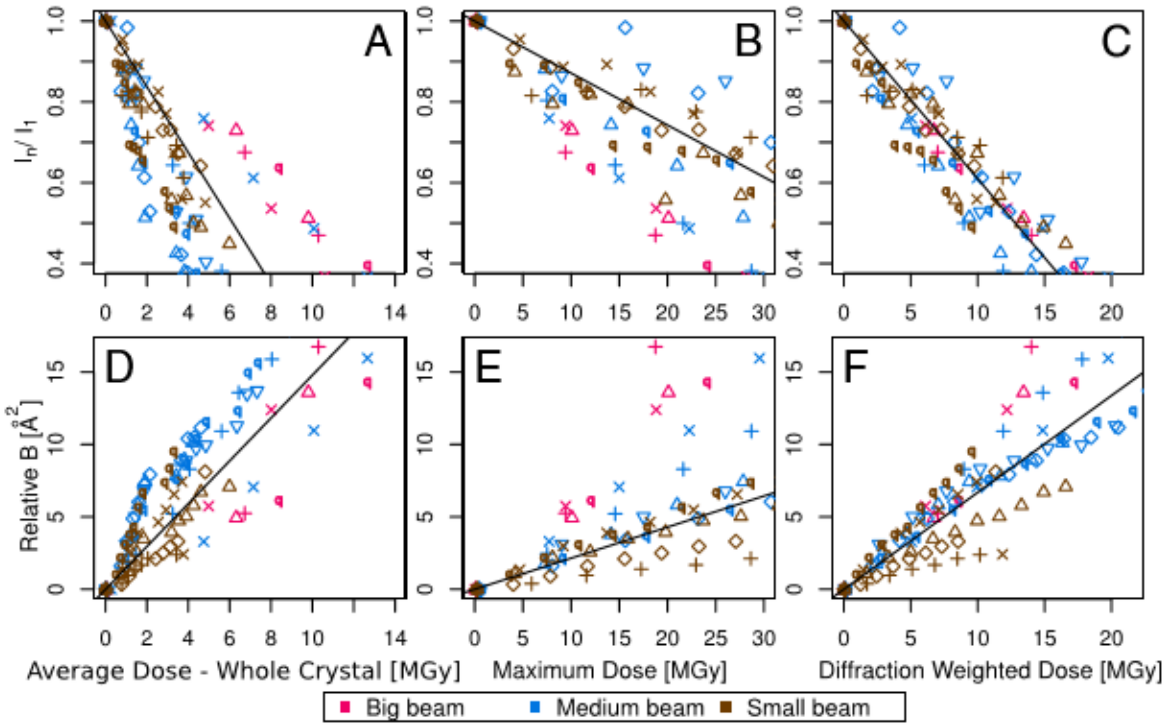


Figure 2.21: (A-C) I_n/I_1 against dose. The reduced scatter along the line of best fit in (C) is evidence that DWD is invariant to the dose distribution in the crystal. (D-F) B_{rel} against dose. There is no obvious reduction in scatter around the line of best fit for these metrics. This suggests that DWD offers no significant improvement over other dose metrics when using B_{rel} as a measure of radiation damage progression. The symbols represent individual crystals. Reproduced from (Zeldin *et al.*, 2013a).

for the decreasing form was not significant enough to lower the DWD, as occurred for the standard experiment.

2.7 Discussion

The DWD is a dose metric that was introduced by Zeldin *et al.* which represents "the average dose in the diffracting crystal volume from which the photons making up a given image have scattered" (Zeldin *et al.*, 2013a). In essence, the DWD is a weighted average of the dose where the weights are given by the fluence at each point in the crystal (equation 1.4.2). Zeldin *et al.* demonstrated that DWD was superior to other metrics (average dose - whole crystal and the maximum dose) in predicting the relative intensity loss of protein crystals whilst accounting for different beam profiles. The improvement in relative intensity decay prediction makes DWD a promising metric for comparing radiation damage studies and also for zero dose extrapolation. However, data points are still scattered around the line of best fit in Figure 2.21 (C). The scatter is thought to be caused by two factors: the first is the

intrinsic variation in crystal quality, and the second is the fact that the DWD does not account for the effects of inhomogeneous dose distribution, such as an uneven loss of diffraction efficiency throughout the crystal (Zeldin *et al.*, 2013a).

The incorporation of an additional weighting term in the definition of the DWD, which accounts for the loss of diffracting efficiency, is hoped to reduce the remaining scatter (equation 2.1.1). The RDE was introduced as a function of the absorbed dose that would represent the loss of diffraction efficiency. Note that equations 1.4.2 and 2.1.1 are equivalent when $\eta = 1$ (the simple η form). To obtain experimental estimates of the RDE of a crystal it was necessary to perform an experiment where the crystals are irradiated uniformly. This would eliminate any dependence on F and η in equation 2.1.1, and thus the result of any intensity changes being described solely by the absorbed dose. At the time of the experiment, RADDOSE-3D could only model cuboid and spherical crystals, so only cuboid shaped crystals were used in the experiment. Diffraction data were processed for five insulin crystals and the results data were used to study the intensity decay as a function of the dose.

Three dose decay models that describe the relationship of intensity decay of a reflection with absorbed dose were investigated for their suitability as models for the RDE: the Sygusch & Allaire model, the Holton model and the Leal *et al.* model. The Sygusch & Allaire model required the solution of a system of coupled first order linear ordinary differential equations for two domains. The interesting feature of this model is that it predicts two different dynamic regions (Figure 2.10b). The Holton and Leal *et al.* models are relatively simple in comparison and they can be converted into linear forms to readily obtain parameter values. To study these functions as RDEs and compare them to the relative intensity decay, these functions had to be integrated over a sphere in reciprocal space and then be normalised with respect to the zero dose integral. The RDE forms of these models were then assessed for their ability to explain the observed data for the five insulin crystals using the RMSD. The Leal *et al.* model narrowly outperformed the Sygusch & Allaire model overall. However, the behaviour of the relative intensity data above around 27 MGy, particularly for crystals 0259 (Figure 2.12a) and 172 (Figure 2.12c), exhibits a slower decay than the decay below that dose. This suggests that the Sygusch & Allaire model prediction of two behavioural domains may be valid and perhaps the model is a more accurate description of the true dynamics.

The resulting RDE using the Leal *et al.* model (equation 2.4.2) was investigated to determine

how resolution changes would affect the calculated parameter values and its ability to explain the observed data. It was found that collecting and processing data to the highest resolution possible, as well as performing the spherical integration to the maximum resolution limits of the *BEST* data is the best approach to faithfully representing the observed data.

The RDE model was incorporated into the DWD using equation 2.1.1 and tested with the data from the Zeldin *et al.* study. On its own, the RDE is a monotonically decreasing function of the dose. This is sensible when describing intensity decay but not necessarily when being used as a dose metric to represent damage in the crystal. Hence by defining $\eta = 1 - RDE$, an increasing function could be used to describe the progression of damage as the dose increases. The main difference in the resulting DWD values using the new forms of η is that the dose values are increased when using the increasing function of η, whereas the dose values are decreased for the decreasing η function. One of the important results from the analysis is that the observed scatter in the data is increased when using the dose dependent forms of η, a result that contradicts the hypothesis that adding the terms would decrease the scatter. Another interesting result from including the decreasing η form to the DWD is that it not only decreases the range of the resulting DWD values, but the DWD values decrease once an upper threshold dose limit is reached.

To determine which η form is correct, the offset simulations should be carried out as actual experiments. This is because the different η forms suggest slightly different results regarding the expected benefits obtained by offsetting the rotation axis from the beam axis with crystals of different sizes. In particular, the differences in the benefits of offsetting the rotation axis by $0\mu m$ and $40\mu m$ for the 400 μm are more pronounced than for the 60 μm crystal. What would need to be established first is how the DDE correlates with the observed intensities i.e. if the difference in DDE between two offsets is x ph/MGy, then what is the expected difference in the observed (relative) intensities? Furthermore to make the results statistically significant, at least three repeat experiments would need to be carried out to give increased confidence in the conclusions that could be drawn.

Given that the simple DWD (equation 1.4.2) already gives a measure of the absorbed dose in the crystal accounting for the crystal composition, it is reasonable to ask what would be added by introducing another dose dependent term into the DWD equation. It may help to

think about two crystals of the same protein that crystallise in the same space group and are irradiated under the same conditions by the same source. The quality of the crystals will differ due to the intrinsic crystal variation, resulting in different data quality and possibly different rates of intensity decay. This could be due to different mosaicity or unit cell volumes of the two crystals, which may also alter differentially during the experiment. The absorbed dose alone does not account for these factors, and hence a function that takes them into account, in theory, should account for the additional variation between crystals. In the analysis, the scatter of the data was increased by incorporating a term that was supposed to reduce it. This result could be due to incorrect parameter values for the RDE model. The method used to determine the parameter values (transforming the data to a linear form before performing a linear fit) may not give the optimum results. The values may be improved by fitting the function to the original data without any transformation. It is unknown how sensitive the dose values are to small deviations from the true parameter values, hence any perturbations of the parameter values could lead to misleading results.

Another question is whether the η function should be an increasing or decreasing function. From the results presented in this chapter, the answer depends on whether DWD should describe the extent of damage or the quality of diffraction. Radiation damage is generally a progressive process which means that reducing DWD values do not make sense when thinking about the metric in terms of describing damage. Hence to describe damage, η should be an increasing function. The results presented here provide some unconvincing evidence for using the increasing function of η over the simple form. Hence an experiment, such as the offsetting one described above, would need to be carried out to determine which form to use.

When framing the DWD as a metric to describe the quality of diffraction, it helps to think about an experiment where a crystal is irradiated by a Gaussian beam. In this case, the Gaussian beam has tails that are being diffracted from relatively undamaged regions of the crystal, even at late stages of the experiment. So when "highly damaged" data are processed, the structure factors that are obtained are more similar to the zero-dose case than they were in the middle of the dataset when the crystal in the bright part of the beam was contributing significantly (James Holton, personal communication). In this case it makes sense for the DWD to decrease because it tells us that the quality of the diffraction has improved, despite the overall damage state of the crystal being worse.

This presents a case for using two metrics concerned with different aspects of radiation damage. One metric that assesses the *damage* caused by the X-rays, which is generally the interpretation of the dose that has been used until now. A second metric could be used to assess the *diffraction quality*. A first step towards this would be to use the RDE as defined in this chapter.

CHAPTER 3

Zero-Dose Extrapolation

3.1 Introduction

Mathematical correction of reflection intensities to account for radiation damage has been a common procedure for decades (Hendrickson *et al.*, 1973; Abrahams and Marsh, 1987). Before the development of RADDOSE and RADDOSE-3D, the progression of radiation damage was tracked with time as the independent variable. Functions to describe the intensity loss took the form of the Blake and Phillips model detailed in equation 1.4.4, or isotropic polynomial expressions (Abrahams and Marsh, 1987).

Radiation damage correction is still an integral part of current scaling methods (Otwinowski *et al.*, 2003; Evans, 2006; Kabsch, 2010a). However, these correction models only compensate for the average decay of reflection intensities. In the correction algorithms, account is not taken of specific changes. Diederichs *et al.* introduced a linear model to perform zero-dose extrapolation, which was tested on data collected on bovine brain tubulin at 15 K (Diederichs *et al.*, 2003), and later introduced exponential and quadratic models (Diederichs, 2006), to account for these specific changes. These authors showed that zero-dose extrapolation can improve phasing statistics and SAD electron density maps. The problem is that reflections exhibit a diverse range of behaviours that are yet to be accounted for by a single Mathematical relationship (Blake and Phillips, 1962; Abrahams, 1973).

Over the last decade there have been several developments in the methods and understanding of radiation damage in MX. Some of the key advances in this field have been:

- RADDOSE-3D (Zeldin *et al.*, 2013b), which allows accurate 3D modelling of the diffraction experiment.
- the introduction of DWD (Zeldin *et al.*, 2013a), which is a useful metric for assessing the damage state of a crystal that resulted in the diffraction for a given image.
- the Leal *et al.* dose decay model (Leal *et al.*, 2012). This relatively simple model was shown to accurately describe the intensity decay of reflections at RT.

The work described in this chapter is concerned with using the developments presented above to perform zero-dose extrapolation on data collected from one of the cubic insulin crystals (crystal ID 0259) described in chapter 2.

3.2 Extracting intensities and doses

To perform zero-dose extrapolation, it is necessary to know both the intensities of each observation of a reflection and the dose absorbed by the crystal at the time of the observation. To extract the reflection intensity data, the insulin diffraction images were integrated using MOSFLM as described in section 2.2.4. The data were then scaled using AIMLESS, but the B factor correction was turned off to prevent correcting for overall radiation damage. The reflection intensities were then output to an unmerged MTZ file and the data were read by running MTZDUMP and parsing the log file. Importantly, each reflection intensity observation was stored along with its corresponding *batch* (image) number, which would ultimately allow it to be assigned the correct dose.

To extract the correct DWD values, RADDOSE-3D was run with the corresponding experimental parameters as described in section 2.2.3. The difference was that rather than outputting one DWD value to represent the average DWD for all images in the dataset, the DWD values were obtained for each image. To use the DWD with the dose dependent η functions developed in chapter 2, the parameter values B_0, β and γ must be found. However the method used to obtain the parameter values in section 2.3.2 requires the collection of multiple datasets, which may not be possible if the crystal is very radiation sensitive. Thus a new method to obtain the parameter values was developed.

The idea is to exploit the fact that the volume element (voxel) of the crystal has its own relative diffraction efficiency. The weighted average of the individual relative diffraction efficiencies will result in the observed relative intensity of the dataset. Mathematically this is

$$I_n/I_0 = \frac{\int_{\boldsymbol{x}} RDE(D(\boldsymbol{x}); B_0, \beta, \gamma) \times F(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}}{\int_{\boldsymbol{x}} F(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}}, \tag{3.2.1}$$

where I_n/I_0 is the theoretical relative intensity value (note that the denominator is I_0 rather than I_1 as the extrapolation back to zero dose is the aim), RDE is the RDE function given by equation 2.4.2, D is the absorbed dose, F is the fluence and the vector $\boldsymbol{x}$ is the position in the crystal. The absorbed dose and fluence for each voxel is returned in a file output by RADDOSE-3D representing the dose state of the crystal. The parameters, B_0, β and γ can then be found using an optimisation routine to minimise the squared residual between I_n/I_0 and I_n/I_1, where I_n/I_1 represent the measured relative intensities. Mat-

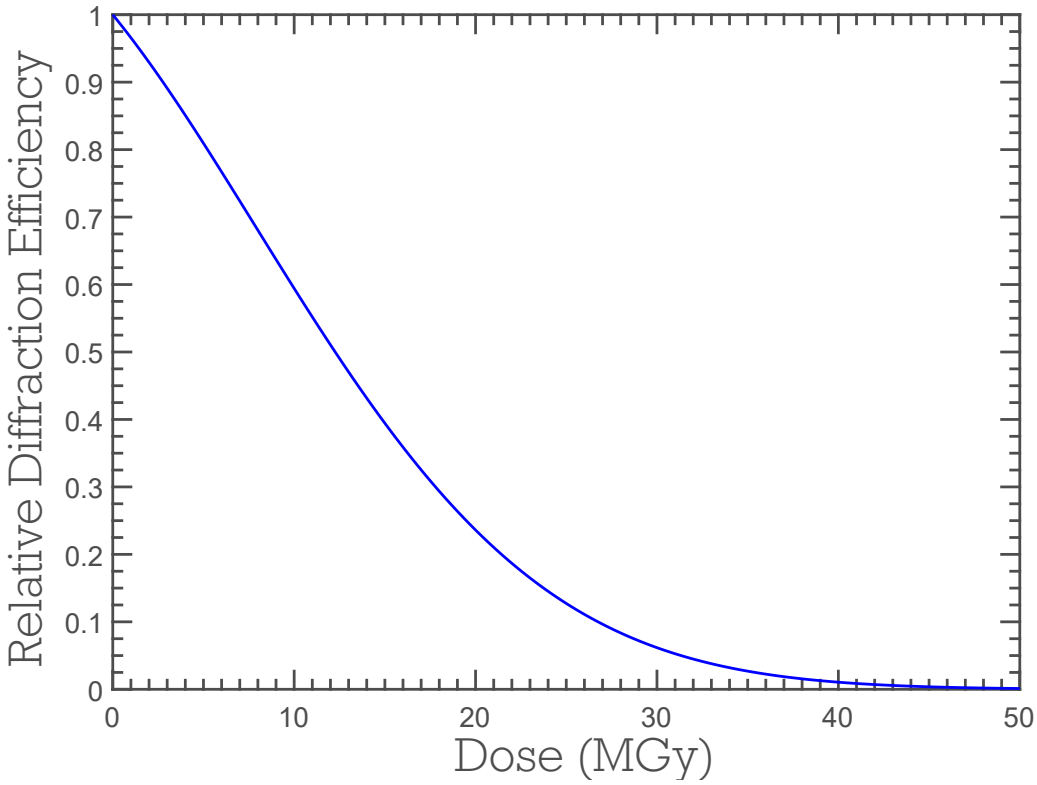


Figure 3.1: Relative diffraction efficiency plot with parameter values $B_0 = 14.28\,\text{Å}^2, \beta = 0.510\,\text{Å}^2\text{MGy}^{-1}$ and $\gamma = 0.047\,\text{MGy}^{-1}$ for insulin crystal 0259.

lab's `lsqnonlin` function was used for the optimisation, which in turn uses the trust-region-reflective algorithm (Coleman and Li, 1996) or the Levenberg-Marquardt algorithm (Moré, 1978) if the number of relative intensity measurements is less than the number of parameters. Further to being able to fit the decay parameters for a single dataset, this method is able to deal with fitting parameters from inhomogeneous dose distributions. An example case where ten 90° wedge datasets from crystal 0259 were used, the parameter values obtained are: $B_0 = 14.28\,\text{Å}^2, \beta = 0.510\,\text{Å}^2\text{MGy}^{-1}$ and $\gamma = 0.047\,\text{MGy}^{-1}$, which are comparable to the values given in Table 2.5. The resulting RDE is plotted in Figure 3.1.

One obvious issue with only using a single dataset for this optimisation is that $I_n/I_1 = 1$ for the first dataset by definition, whereas $RDE(D)$ is less than 1 for any dose greater than zero. To circumvent this issue, the observed relative intensity data are scaled to more accurately represent the relative intensity with respect to the theoretical zero-dose, I_0, prior to the parameter optimisation. To scale the data, the following function was first fitted to the data

$$I_n/I_0(D) = ke^{aD^2+bD}, \tag{3.2.2}$$

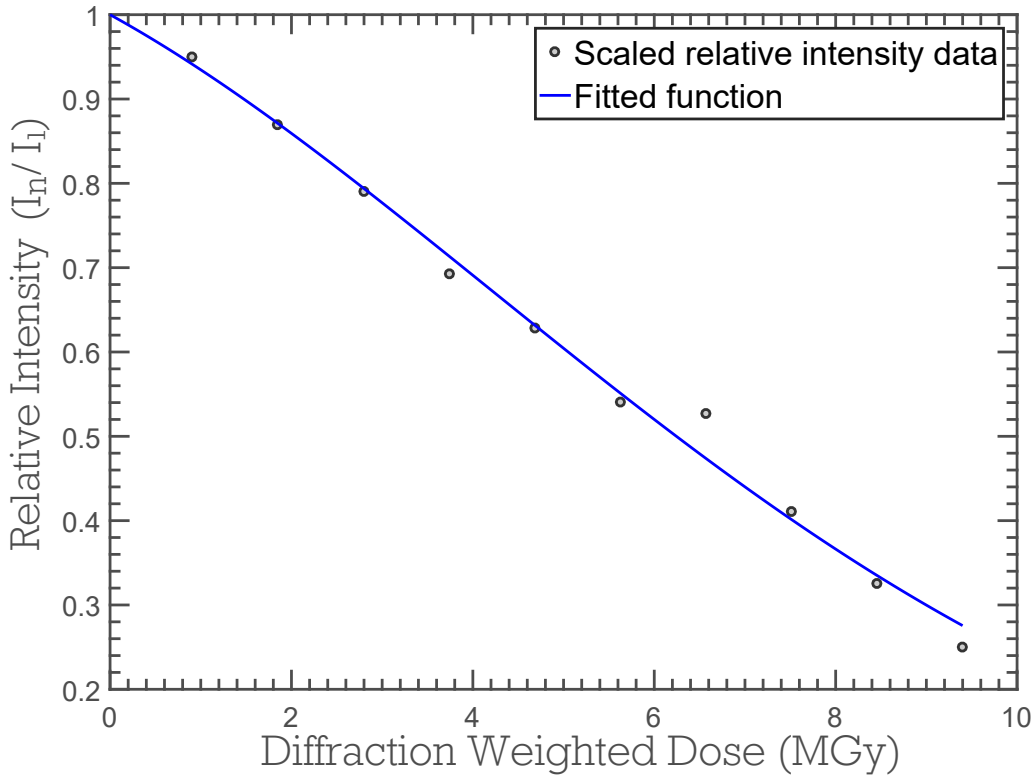


Figure 3.2: Scaled relative intensity data (grey circles) for cubic insulin crystal (crystal ID 0259) plotted with the fitted function (equation 3.2.2) overlaid.

where k, a and b are parameters to be determined. A closer inspection shows that this equation is a more general form of the Leal *et al.* model where the resolution dependence has been omitted. The dose metric used here is the simple DWD (equation 1.4.2), because at this point there is no information about the parameter values required to calculate η. For small dose ranges, the simple DWD is a suitable approximation for the DWD with dose dependent η terms, because for small doses, $\eta(D) \approx 1$ (assuming the decreasing η form). A least squares curve fitting procedure, Matlab's `lsqcurvefit` routine, was used to fit the relative intensity data to the function defined in equation 3.2.2. The relative intensity data were then scaled by k. The result of the relative intensity scaling for the example above (ten 90° wedge datasets from crystal 0259) is shown in Figure 3.2.

3.3 Extrapolation routine

3.3.1 Regression extrapolation

With both the reflection intensities and the doses extracted, extrapolation can be performed using standard regression techniques. In particular, parametric regression in which a particular functional form describing the relationship between the two variables was implemented. The function that was used to describe the relationship between the reflection intensity, I, and the dose, D, for each individual reflection was:

$$I(D) = K_{\boldsymbol{h}} \exp\left[-A_{\boldsymbol{h}}^2 D^2 - B_{\boldsymbol{h}} D h^2/2\right], \tag{3.3.1}$$

where $\boldsymbol{h}$ represents the Miller indices of a reflection, $h = |\boldsymbol{h}| = 1/d$, d is the spacing between adjacent Bragg planes, and $K_{\boldsymbol{h}}, A_{\boldsymbol{h}}$ and $B_{\boldsymbol{h}}$ are reflection specific parameters to be determined. This function was chosen because it is identical to the Leal *et al.* model (equations 2.3.8, 2.3.9 and 2.3.10) where

$$A_{\boldsymbol{h}}^2 = \gamma^2, \tag{3.3.2}$$

$$B_{\boldsymbol{h}} = \beta, \tag{3.3.3}$$

$$K_{\boldsymbol{h}} = K \exp\left[-B_0 h^2/2\right]. \tag{3.3.4}$$

Essentially equation 3.3.1 is a Gaussian function. Hence the Leal *et al.* model describes a Gaussian decay of reflection intensity with the dose. The values of the parameters $K_{\boldsymbol{h}}, A_{\boldsymbol{h}}$, and $B_{\boldsymbol{h}}$ will determine which section of the Gaussian will best describe the change in intensity for a given reflection.

To determine the parameter values (and hence the decay of intensity of a reflection), all observations of equivalent* reflections were grouped together to form the set of observed intensities. The parameter values were then found as the values that minimised the function:

$$\sum_{obs} w_{obs} \left(I_{obs}(D_{obs}) - I_{calc}(D_{obs})\right)^2, \tag{3.3.5}$$

where $w_{obs} = 1/\sigma_{obs}^2$ is the weighting term and σ_{obs} is the standard deviation of the ob-

*Equivalent in this sense means all symmetry equivalents and Friedel pairs.

servation, $I_{obs}(D_{obs})$ is the observed intensity value at dose D_{obs} and I_{calc} is the calculated intensity as given by equation 3.3.1. The optimization routine is performed using the MATLAB `fminsearch` routine which uses the simplex search method of Lagarias et al. (Lagarias *et al.*, 1998). This is a direct search method that does not use numerical or analytic gradients.

To ensure that the `fminsearch` routine converges to sensible final parameter values, it has to be seeded with suitable initial estimates of the values. To create the initial estimates, 3 observations (first, last and an observation from the middle of a dataset - provided they meet the criteria mentioned later in this section) are used to analytically find the parameter values such that the function passes through all three points.

One of the main problems with obtaining accurate data in crystallography is the limited multiplicity of reflection observations. This can also be problematic for performing accurate regression. Equation 3.3.1 requires three parameter values, therefore there should be at least three observations in order to perform extrapolation. However with only three observations, the regression is likely to lead to overfitting of the model to the data (Figure 3.3), so to avoid this, more observations are required.

An additional procedure that was implemented to prevent overfitting was to ensure that the fit correlated with the data. Therefore after the regression was performed, a correlation coefficient was calculated between the model fit and the data. If the resulting correlation was below a specified value then the extrapolation procedure was abandoned. A low correlation suggests that there were problems with the fit, as illustrated in Figure 3.4a where the correlation coefficient was 0.209. However, despite abandoning the procedure, a low correlation value did not always suggest a bad fit as shown in Figure 3.4b, where the correlation was low (0.231) but the fit looks reasonably good when accounting for the noise level.

Other checks were also performed to make sure that the resulting fits and zero-dose extrapolated intensities were reasonable:

1. that the first M observations are significantly above zero, where M is a user defined integer value. This value was set to the same value as the minimum number of observations required to perform a regression fit. The exact procedure was to check that for the first M observations $I_{obs}(D_{obs}) - n_0\sigma_{obs} > 0$, where $n_0 = 0.5$. This was carried out to prevent overfitting that could result from trying to fit to very small or negative

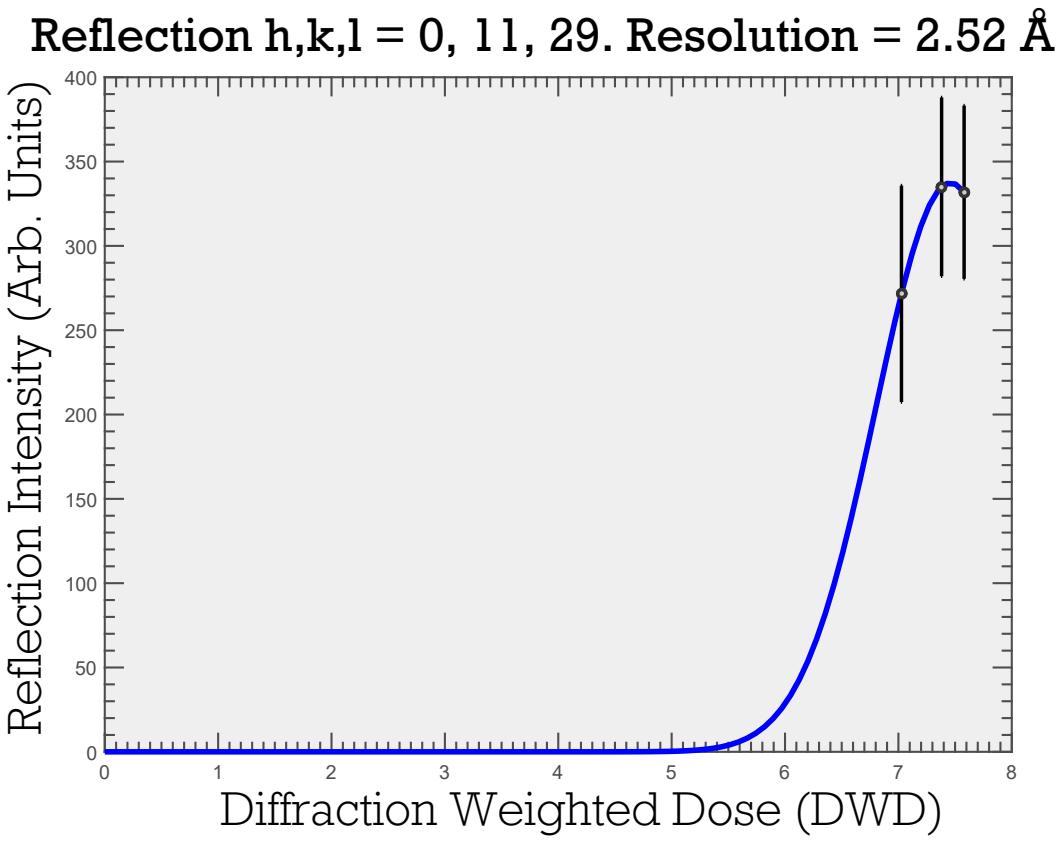


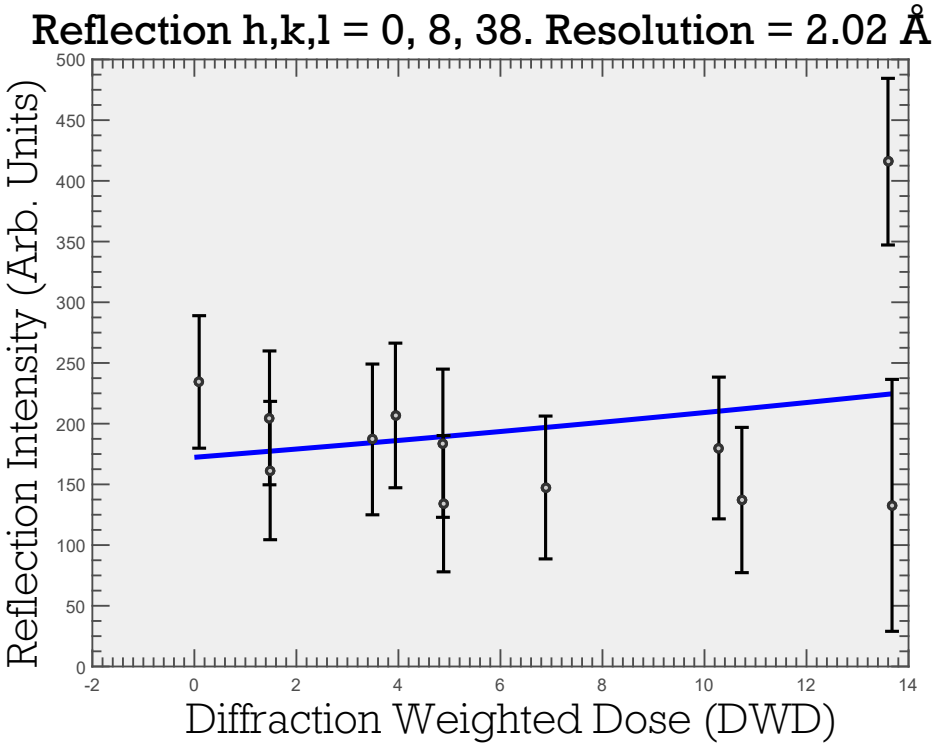
Figure 3.3: Regression performed where the model (solid blue line) has been overfitted to the data (grey circles). The function passes perfectly through all three data points but predicts a zero-dose intensity of 0. The solid black vertical lines are the standard deviations, σ_{obs}, of the reflection intensity observations.

intensity values as exemplified in Figure 3.5a. This fit is poor because the Leal *et al.* model is always positive for any positive real-valued dose. Hence fitting the function to small or negative values is likely to result in an untrustworthy fit.

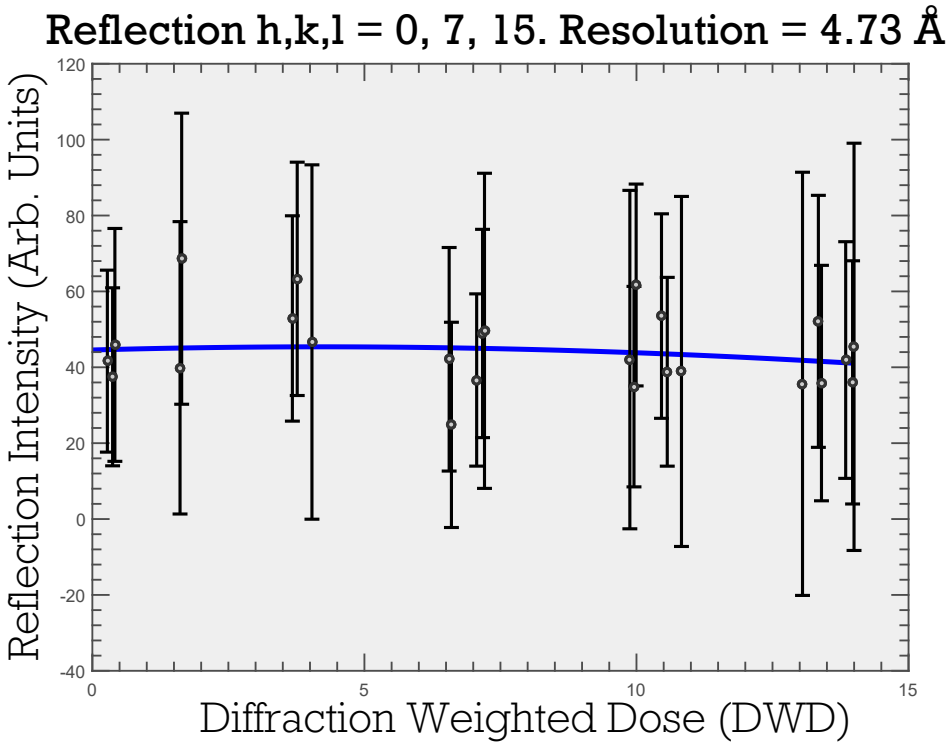
2. that the calculated zero-dose intensity using the fitted parameter values gives a finite value and it is not significantly larger than the intensity values from the rest of the dataset. This is to prevent unreasonable fits as shown in Figure 3.5b.

All the above checks are made to ensure that the model fit is representative of the data. However, the zero-dose values are not obtained from these fits (i.e. the zero dose value of the blue line in Figures 3.3, 3.4 and 3.5 is not used). Instead, the zero dose value is obtained separately for each observation to preserve the spread of the data for rescaling. This is exactly the same method used by Diederichs *et al.* (Diederichs *et al.*, 2003). To obtain the zero-dose estimate for each observation, $K_{\boldsymbol{h}}^{obs} = I_{obs}(0)$, equation 3.3.1 is rearranged to get

$$K_{\boldsymbol{h}}^{obs} = \exp\left[\log(I_{obs}(D_{obs})) + A_{\boldsymbol{h}}^2 D_{obs}^2 + B_{\boldsymbol{h}} D_{obs} h/2\right], \tag{3.3.6}$$

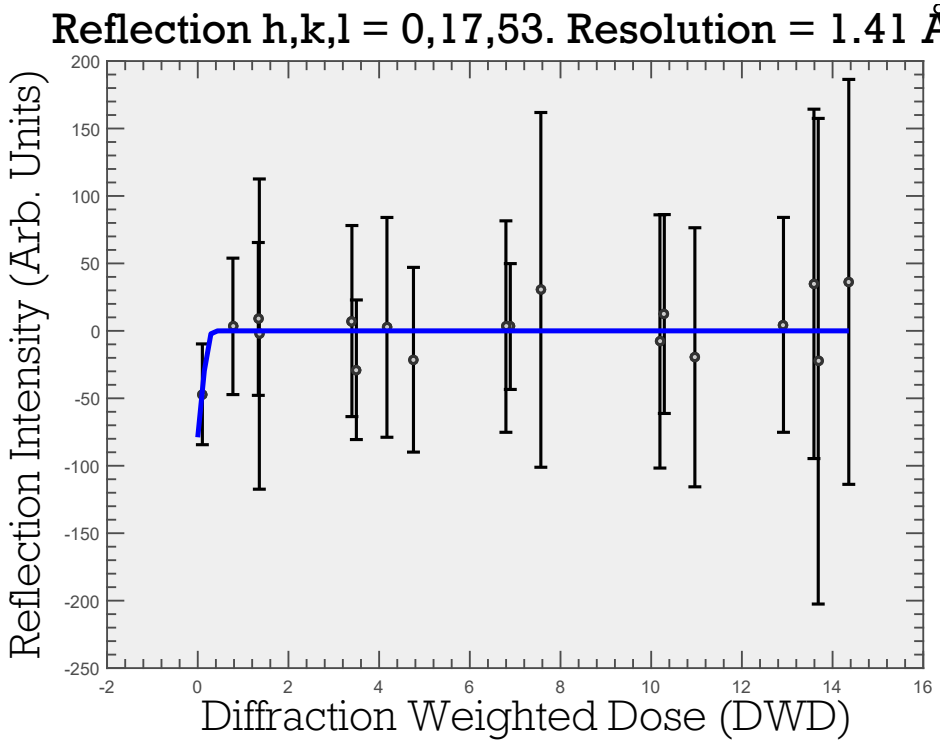


(a)

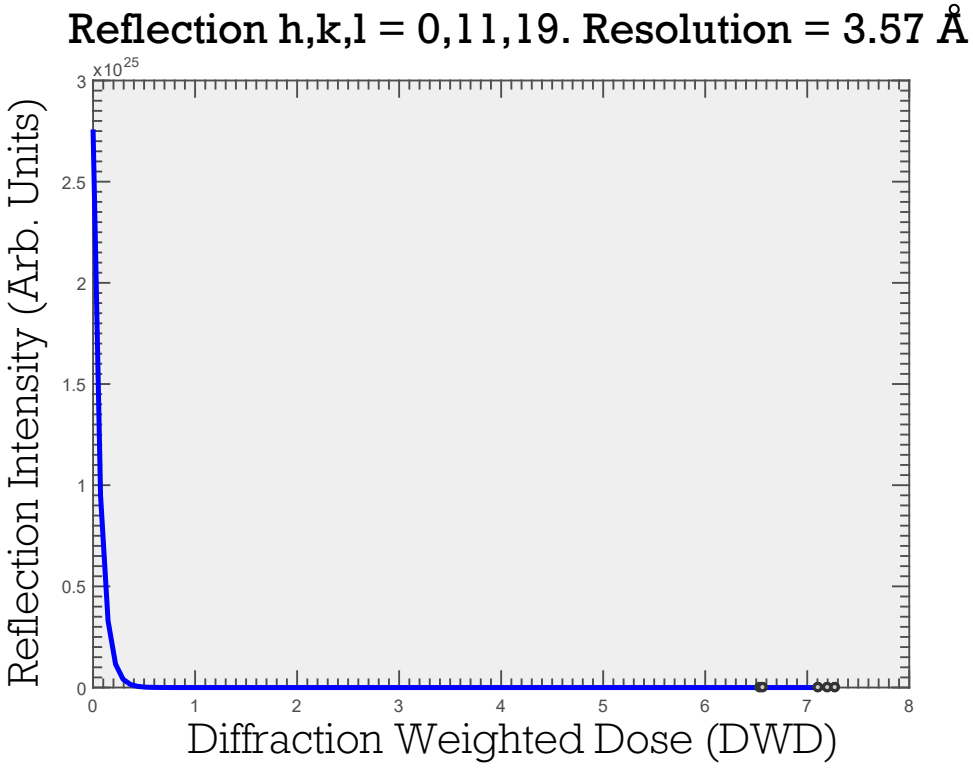


(b)

Figure 3.4: (a) Regression performed for a reflection where the correlation coefficient was determined to be 0.209. The low correlation coefficient is consistent with the bad fit of the model to the data. (b) Regression performed for a reflection where the correlation coefficient was determined to be 0.231. The low correlation coefficient in this case is clearly due to the noisy measurements but the fit looks reasonable given that the intensity predictions are consistently within 1 standard deviation of every observation.



(a)



(b)

Figure 3.5: (a) Poor model fit as a result of using data where the values are too small or negative. Equation 3.3.1 is always positive for all positive real-valued dose values so the fit does not exhibit a physically reasonable relationship when fitting to negative data. (b) A zero-dose estimate that is unphysically high despite a good fit to 5 data points. This often happens when the only observations of a reflection occur relatively close to each other but far away from the zero-dose state and decrease monotonically as the dose increases.

where the values for $A_{\boldsymbol{h}}$ and $B_{\boldsymbol{h}}$ are as determined from the regression fit.

Although the zero-dose observations are estimated using regression analysis, the standard deviations are not dealt with by this method. The standard deviation is computed as

$$\sigma_{\boldsymbol{h}}^{corr} = \sqrt{\sigma_{obs}^2(1 + D_{min}^2)}, \tag{3.3.7}$$

where $\sigma_{\boldsymbol{h}}^{corr}$ is the corrected standard deviation, σ_{obs} is the observed standard deviation and D_{min} is the dose value of the first observation. Equation 3.3.7 takes the same form as the corrected standard deviation in Diederichs *et al.* (2003), with a couple of differences:

1. in the study by Diederichs *et al.*, the standard deviation is inflated as

$$\sigma_{ij}^{corr} = \left[\sigma_{ij}^2 + (\sigma_\beta x_{ij})^2\right]^{1/2}, \tag{3.3.8}$$

   where σ_{ij} is the standard deviation of the i^{th} observation in the j^{th} dataset of a reflection, σ_β is the standard deviation of the damage factor β (essentially the gradient of the assumed linear intensity decay) which is common to all observations of a reflection in all datasets, and x_{ij} is the dose at which the observation occurred. In the current work, a different decay form is used for the extrapolation, so there is no single term that is directly analogous to the decay factor.

2. D_{min} is used instead of the actual dose that has been absorbed by the crystal during that particular observation because I have concluded that a big factor in the reliability of the extrapolation is how early the initial measurement was made. A thought experiment to illustrate this is to suppose a reflection is observed on every frame in an MX experiment where the dose range is 1-10 MGy, and another reflection is observed on every frame from a dose of 5.5 MGy onwards. Assume that observations of these reflections on the same image have the same sigma values and also that the regression fits are theoretically perfect for both reflections. In the case where the dose value at which the observation was made is used to inflate the standard deviations, the inflated values for the extrapolation of the observations will have the same value because they have the same dose and same sigma value. This is despite the fact that one of the reflections was only initially observed half way through the experiment. In the latter case, it is clear that the zero-dose value should have a higher uncertainty than the

zero-dose value for the reflection that was observed on the first frame regardless of the dose at which an observation is made.

Assessing the quality of the regression fit

Even though checks are performed to make sure that the zero-dose extrapolation fit is sensible, it is useful to assess the quality of all regression fits. A relatively obvious criterion for deciding whether a fit is adequate is to determine how close the calculated intensity of an observation is to the measured intensity value for the observation. The R_{fit} metric is given by

$$R_{fit} = \frac{\sum_{hkl} R_{err}^{hkl}}{N_{ref}}, \tag{3.3.9}$$

where N_{ref} represents the number of extrapolated reflections and the numerator is a sum over the number of extrapolated reflections where

$$R_{err}^{hkl} = \frac{\sum_{i}^{n_{obs}} w_i |I_{obs}^{i}(D_{obs}) - I_{calc}^{i}(D_{obs})|}{\sum_{i}^{n_{obs}} w_i I_{obs}^{i}(D_{obs})}, \tag{3.3.10}$$

and n_{obs} is the number of observations of a particular reflection.

R_{fit} (equation 3.3.9) is equivalent to the mean absolute percentage error.

The R_{fit} metric represents a global assessment of the quality of the overall regression fits to the data. However, examining the quality of the zero-dose extrapolated values is also desirable. Although the actual zero-dose values are never observed, the R_{fit} metric determined for only the **first** observation of a reflection gives an idea of how well the first observation is predicted by the regression fits. This metric is termed R_{zero}.

R_{fit} and R_{zero} are global metrics that quantify the quality of the overall regression fits and the ability of the method to predict the first observation of reflections respectively. The lower the value, the better the quality of the model fits. However, what constitutes a "good" quality fit is not immediately obvious. One criterion to determine whether a fit is good is that the calculated intensity value is within one standard deviation of the observed intensity value. Therefore R_{fit} and R_{zero} can be calculated such that $|I_{obs}(D_{obs}) - I_{calc}(D_{obs})| = \sigma_{obs}$ for every observation. These values are termed R_{fit}^{thresh} and R_{zero}^{thresh} respectively. A fit can then be considered adequate if $R_{fit} < R_{fit}^{thresh}$ and $R_{zero} < R_{zero}^{thresh}$ because this would mean

that the calculated intensity values are generally within 1 standard deviation of the observed intensity values.

3.3.2 Probabilistic extrapolation

The regression analysis performed as described above requires several checks to be made to ensure that the fitted model is reasonable. This filters out reflections that do not behave according to the relationship expected by equation 3.3.1, so some reflections are not extrapolated. Therefore Bayesian inference is used to extrapolate these remaining reflections. Bayes theorem in model form states

$$P(model|data) \propto P(data|model) \times P(model), \tag{3.3.11}$$

where $P(model|data)$ is the posterior distribution - the updated probability of the model given the observed data, $P(data|model)$ is the likelihood function - the probability of observing the data given the current model and $P(model)$ is the prior distribution - the probability of the model in the absence of any data. In this case, the $model \equiv J$, where J is the expected zero-dose intensity, and $data \equiv I, D$ where I represents the observed intensity and D is the dose at which the reflection is observed.

$P(J|I, D)$ is the (posterior) distribution of interest because this will give the probability of an observation having a particular zero-dose intensity value, J, given the intensity observed, I, at a particular dose, D. The expected value of this distribution will be the extrapolated intensity value. The expected value, $E(J|I, D)$, is given by

$$E(J|I, D) = \int_0^\infty J \times P(J|I, D)\,\mathrm{d}J. \tag{3.3.12}$$

Furthermore the variance, $var(J|I, D)$, of the observation can be obtained from the posterior distribution as

$$var(J|I, D) = \int_0^\infty [J - E(J|I, D)]^2 \times P(J|I, D)\,\mathrm{d}J \tag{3.3.13}$$

Therefore the likelihood $\equiv P(I|J, D)$ and prior $\equiv P(J)$ distributions are required to calculate the posterior distribution.

The prior, $P(J)$, is given by the Wilson distribution for intensities (Wilson, 1949). For acentric reflections the Wilson distribution is

$$P_a(J) = \begin{cases} (\varepsilon_{\boldsymbol{h}}\Sigma)^{-1}\exp(-J/\varepsilon_{\boldsymbol{h}}\Sigma) & \text{if } J \geq 0, \\ 0 & \text{if } J < 0, \end{cases} \tag{3.3.14}$$

whereas for centric reflections it is

$$P_c(J) = \begin{cases} (2\pi\varepsilon_{\boldsymbol{h}}\Sigma J)^{-1/2}\exp(-J/2\varepsilon_{\boldsymbol{h}}\Sigma) & \text{if } J \geq 0, \\ 0 & \text{if } J < 0, \end{cases} \tag{3.3.15}$$

where Σ is the distribution parameter, which here represents the mean intensity in the corresponding resolution shell of reciprocal space and $\varepsilon_{\boldsymbol{h}}$ is the multiplicity of the reflection with reciprocal lattice vector $\boldsymbol{h}$. $\varepsilon_{\boldsymbol{h}}$ is an integer value and the multiplication by Σ is to account for the increase in the expected intensity due to the space group symmetry (Blessing *et al.*, 1998). A rule for calculating ε is proposed by Stewart and Karle (1976) (Stewart and Karle, 1976): "ε is the number of times the transformed vector, $\boldsymbol{h}_t = (h,k,l)_t$, is identical to a given reflection, $\boldsymbol{h} = (h,k,l)$, under all distinct pure rotational symmetry operations $\boldsymbol{R}$ of the space group; $\boldsymbol{h}_t = \boldsymbol{h}\boldsymbol{R}_t$." The mean intensity of a resolution bin, Σ, must be corrected for ε. Therefore, the mean intensity of a resolution bin is given by

$$\Sigma = \langle I/\varepsilon \rangle = \frac{\sum_i^{n_{bin}} I_{\boldsymbol{h}}^i/\varepsilon_{\boldsymbol{h}}}{n_{bin}}, \tag{3.3.16}$$

where n_{bin} is the number of reflection observations in a resolution shell and $I_{\boldsymbol{h}}^i$ is the intensity of the i^{th} reflection observation with reciprocal lattice vector $\boldsymbol{h}$.

It should be noted here that since the aim is to extract the zero dose intensity of observations, the mean intensity value in the resolution bins, Σ, is not calculated from the intensity measurements of the data. Instead they are calculated from the zero dose values that were calculated from the regression analysis (Figure 3.6). This method does not account for anisotropic diffraction, therefore the errors are likely to be great if applied to anisotropic data.

Now that the prior distribution has been specified, it is left to define the likelihood distribution, $P(I|J,D)$. Using the same approach as the French and Wilson truncation algorithm

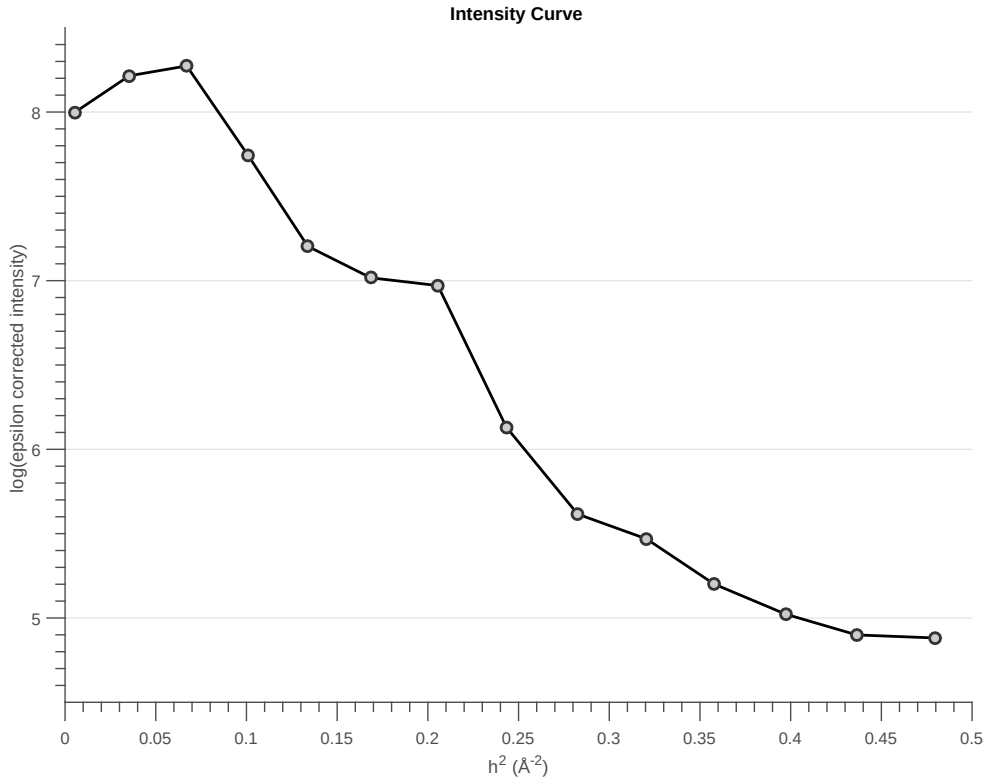


Figure 3.6: Logarithm of the mean intensities in 14 resolution shells. The mean intensities are calculated from the zero-dose extrapolated intensities that were found using the regression analysis.

(French and Wilson, 1978), the likelihood is assumed to be a normal distribution. The main difference is that the intensity data are assumed to have been changed by a scale factor, s, which is a function of the dose, D. Therefore the intensity of an observation is normally distributed:

$$I_{obs} \quad \sim \quad \mathcal{N}(J \times s(D_{obs}), \sigma^2_{obs}). \tag{3.3.17}$$

Or more explicitly

$$P(I_{obs}|J, D_{obs}) = \frac{1}{\sigma_{obs}\sqrt{2\pi}} \exp\left[-\frac{(I_{obs} - Js(D_{obs}))^2}{2\sigma^2_{obs}}\right]. \tag{3.3.18}$$

So all that remains is to find $s(D)$.

An estimate of $s(D)$ can be found for resolution shells by calculating the mean intensity inside resolution bins of observations that were collected within a *dose shell*. For example, if data were collected up to a dose of 10 MGy, then 5 dose shells can be chosen as: 0 - 2 MGy, 2 - 4 MGy, 4 - 6 MGy, 6 - 8 MGy, 8 - 10 MGy. The average dose in these dose shells represents the dose at which the mean intensity was observed. The mean intensities will be

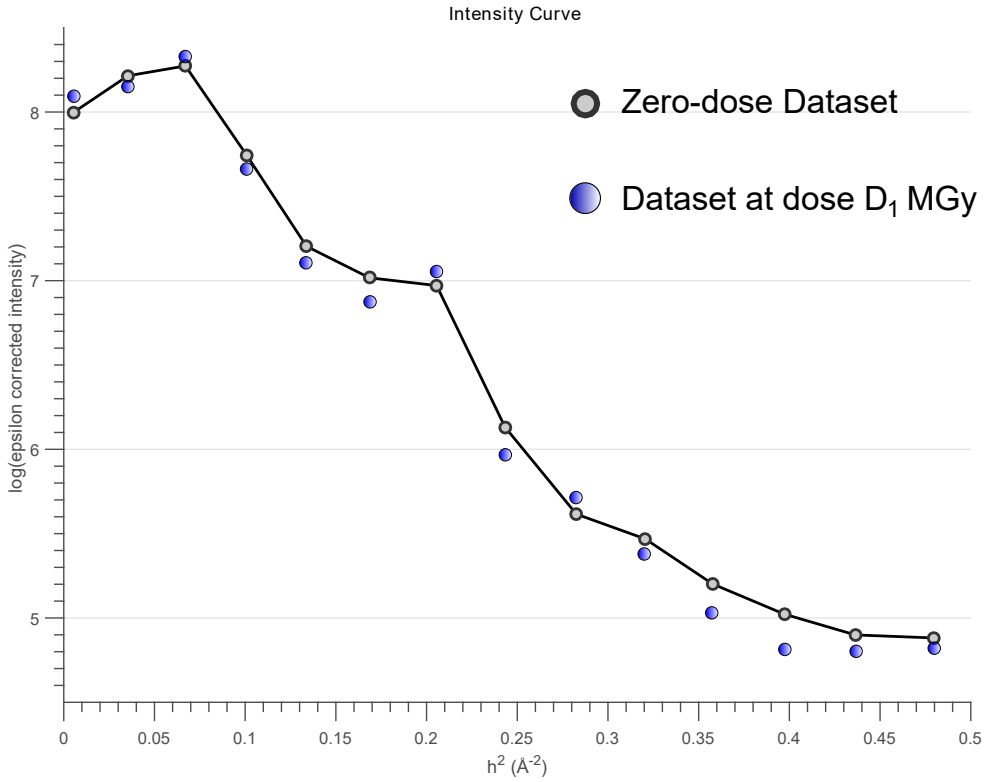
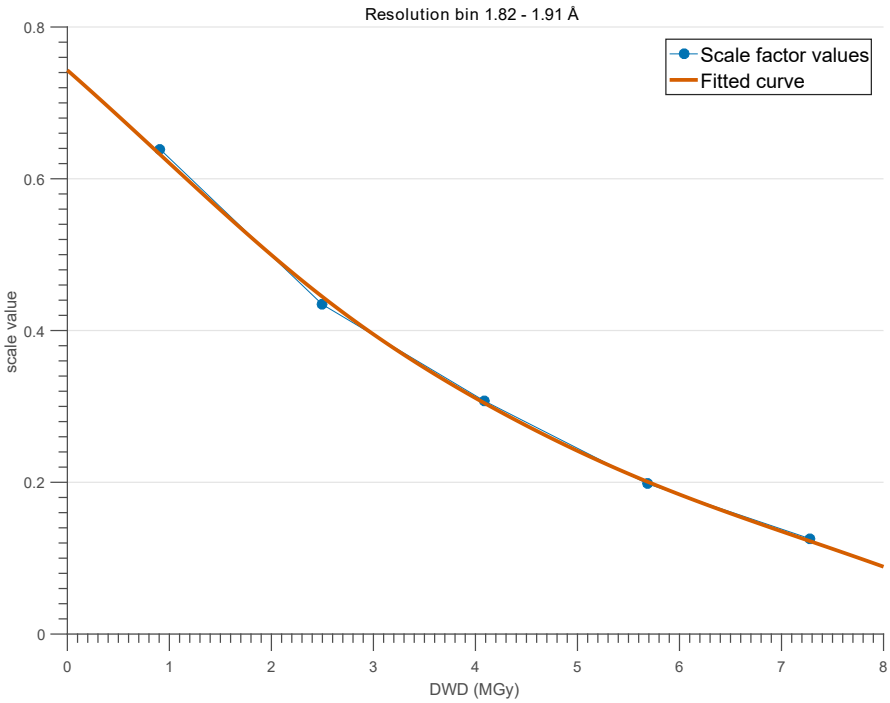


Figure 3.7: Logarithm of the mean intensities in 14 resolution shells for the extrapolated zero-dose dataset (obtained using the regression based extrapolation described in 3.3.1) and theoretical mean intensity values at dose D_1. $s(D_1)$ for each resolution shell can be estimated as the value s_f such that $\langle I(0)\rangle = s_f\langle I(D_1)\rangle$.

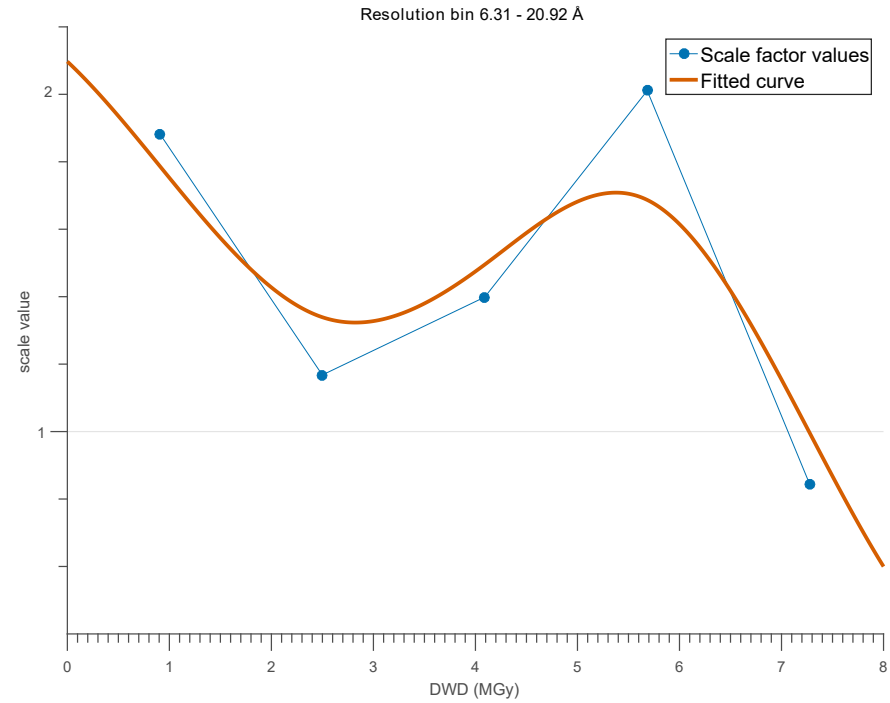
different from the zero-dose mean intensity as depicted in Figure 3.7. If $\langle I(D_1)\rangle$ represents the mean intensity of a reflection observed in a particular resolution shell in a dose shell with average dose D_1, then the scale factor, $s(D_1)$, can be estimated as the value s_f such that $\langle I(0)\rangle = s_f\langle I(D_1)\rangle$, where $\langle I(0)\rangle$ is the average zero-dose intensity in that resolution shell. This can be performed over several resolution shells to build a set of points of $s(D)$ at $D_1, D_2, \ldots, D_m$, as long as there are a suitable number of reflections in the dose shell and resolution shell. This allows interpolation of scale factors for doses between the values used in the dose shells. A smoothing spline was used to interpolate scale factors in the present work, as shown in Figure 3.8 .

When performing this analysis it is important to take into account the fact that there may not be enough reflections in a resolution shell to obtain reliable estimates of the average intensity. This would correspond to missing points in Figure 3.6. Two ways to resolve this issue are:

1. choose larger resolution shells, hence the number of reflections in a given resolution



(a)



(b)

Figure 3.8: Smoothing spline interpolation of estimated scale factors against dose in a single resolution shell. The blue circles represent the calculated $s(D_{obs})$ values whereas the orange line represents the smoothing spline interpolation. (a) At high resolution the scale factor exhibits a monotonic decrease in scale factor suggesting an overall decrease in average intensity with dose. (b) At lower resolutions the scale factor dose not necessarily display a monotonic relationship with dose.

shell increases.

2. interpolate/extrapolate with the data points that are available.

The minimum number of reflections in a resolution bin is a parameter that can be manually set in the current implementation. The interpolation/extrapolation method has also been implemented. This uses a simple line of best fit through the existing points, however a more sophisticated implementation would involve using information from the BEST curve (Figure 2.7).

Another point that must be considered is that not all of the reflections in a resolution shell will follow the behaviour pattern of the "*average*" reflection according to the scale factor, $s(D)$. This is important because using an incorrect model for the radiation decay is more detrimental than not correcting the intensities at all. Therefore only a fraction of the scale factor for that shell value may be applied when considering each reflection. A measure to determine whether a reflection may behave like the "*average*" is the absolute difference in intensity between the observed intensity and the mean intensity multiplied by the corresponding reflection multiplicity $\varepsilon_{\boldsymbol{h}}$, which will be denoted r. If the difference is large then the reflection may not behave like the "*average*" and hence only a small fraction of $s(D_{obs})$ should be applied when performing the extrapolation. The logistic function that is bounded between $s(D_{obs})$ and 1 can be used to quantify the proportion of $s(D_{obs})$ that should be applied to the expected zero dose intensity J, and is a function of the absolute residual r. Effectively it represents the extent to which the data are corrected according to the correction model. Mathematically f is given by

$$f(r;D) = \frac{s(D) - 1}{1 + (r^2/A)^B} + 1; \tag{3.3.19}$$

where f is the amount of $s(D)$ that actually multiplies J, and A and B are user specified parameters that determine respectively where the inflection point occurs and the steepness of the logistic function slope. The relationship between f and r can be seen in Figure 3.9 for $s(D_{obs})$ = 0.5, A = 10 and B = 5. It can be seen that $f(r)$ can only take values between $s(D_{obs})$ and 1. When the residual is small ($r \approx 0$), i.e. the difference between the observed intensity and the mean intensity is small, $f(r) \approx 0.5$. As the residual increases, $f(r) \to 1$. How fast this change occurs depends on the values of A and B, which in turn will be crystal dependent.

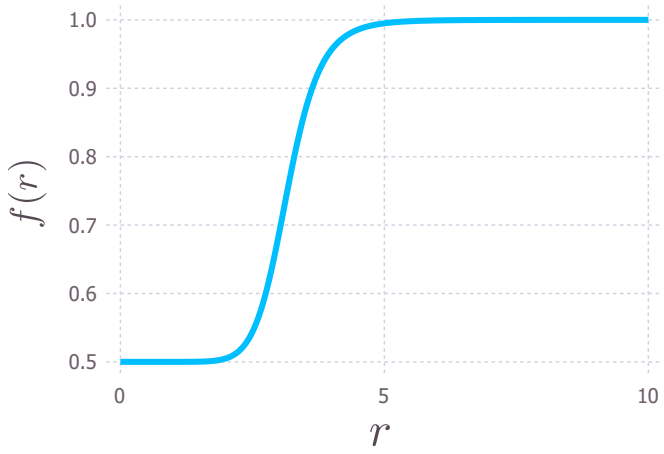


Figure 3.9: Scale factor weighting against the residual of the observed and mean intensities. The function always gives values between $s(D_{obs})$ and 1.0 and can therefore be viewed as the amount of the $s(D_{obs})$ correction that is applied.

Diederichs *et al.* use a similar principal of down weighting their decay factors and standard deviations (Diederichs *et al.*, 2003).

If $f(r)$ is preferred over using the calculated scale factor, $s(D)$, then the likelihood distribution takes the form

$$P(I_{obs}|J, D_{obs}) = \frac{1}{\sigma_{obs}\sqrt{2\pi}} \exp\left[-\frac{(I_{obs} - Jf(r))^2}{2\sigma_{obs}^2}\right]. \tag{3.3.20}$$

A robust method for determining suitable values of the parameters A and B is yet to be devised.

Now that the prior and likelihood distributions have been specified, the posterior distribution can be found by multiplying both distributions and dividing through by a normalising constant (the marginal distribution of the data). This can be written as

$$P(J|I_{obs}, D_{obs}) = \frac{P(I_{obs}|J, D_{obs}) \times P(J)}{P(I_{obs})}. \tag{3.3.21}$$

Finding the exact value of the marginal distribution of the data (the denominator in equation 3.3.21) is not practical analytically. Instead, the marginal distribution is found by multiplying the likelihood and prior distributions and integrating out the model parameter, J, so explicitly the posterior distribution is given by

$$P(J|I_{obs}, D_{obs}) = \frac{P(I_{obs}|J, D_{obs}) \times P(J)}{\int_0^\infty P(I_{obs}|J, D_{obs}) \times P(J)\,\mathrm{d}J}, \tag{3.3.22}$$

which is calculated numerically. The zero dose intensity is then determined by calculating the expected value and the variance according to equations 3.3.12 and 3.3.13.

3.4 Results

The extrapolation algorithm presented in section 3.3 has been written in the Matlab programming language. Intensity data from a single insulin crystal (crystal ID 0259) were collected as described in section 2.2.3 and processed as outlined in section 3.2. The first 5000 reflections from five 90° degree datasets were used to perform zero-dose extrapolation. The average DWD and scaled relative intensity values for each dataset are given in Table 3.1. The doses here represent the average doses in the dose shells used in the probabilistic extrapolation. The 25,000 rows of data from the *HKL* list in the MTZ files resulted in 1255

Table 3.1: Average DWD and scaled relative intensity values. The scaling value, k, was 1.0888

Dataset Number	Average DWD (MGy)	Relative Intensity I_n/I_0
1	0.91	0.9184
2	2.50	0.7812
3	4.09	0.6533
4	5.68	0.5252
5	7.28	0.4286

unique reflections on which the extrapolation as described in section 3.3 was performed. The minimum number of observations to perform the extrapolation was set at 8, and the threshold correlation coefficient was set at 0.5. These values were chosen by inspecting a subset of extrapolated curves to determine which values generally resulted in curves that gave visually sufficient fits to the data. They are quite conservative so it is expected that values which would result in a higher number of reflections extrapolated via the regression method (i.e. lower number of observations and lower correlation coefficient) might lead to better overall extrapolated values. If the natural logarithm of any intensity value was over 14, then regression based extrapolation was not performed on the reflection. This value was chosen because the plot of the data (Figure 3.12a) visually show that values beyond 14 are highly unlikely. However, it should be noted that these values have not been statistically determined as outliers and may actually represent true values of very strong reflections. With the above criteria, the conventional regression analysis was only performed on 748 (59.60%) of the 1255 unique reflections. The rejection list was as follows:

- 43 (3.43%) reflections were rejected because there were less than 8 observations.
- 375 (29.88%) reflections were rejected because the first 8 intensity observations were too weak.
- 89 (7.09%) reflections were rejected because the correlation coefficient was below 0.5.

The assessment criteria for the regression fits are shown in Table 3.2. Both $R_{fit} < R_{fit}^{thresh}$ and $R_{zero} < R_{zero}^{thresh}$, suggesting that the regression fits and zero-dose extrapolated intensity values, are sensible. Fits from two of the successfully extrapolated reflections are shown in Figure 3.10. It is apparent from these two reflections that the behaviour of the intensities is very different. Despite the general decrease in the average reflection intensities throughout the experiment, some reflection intensities increase before they start to decay (Figure 3.10b).

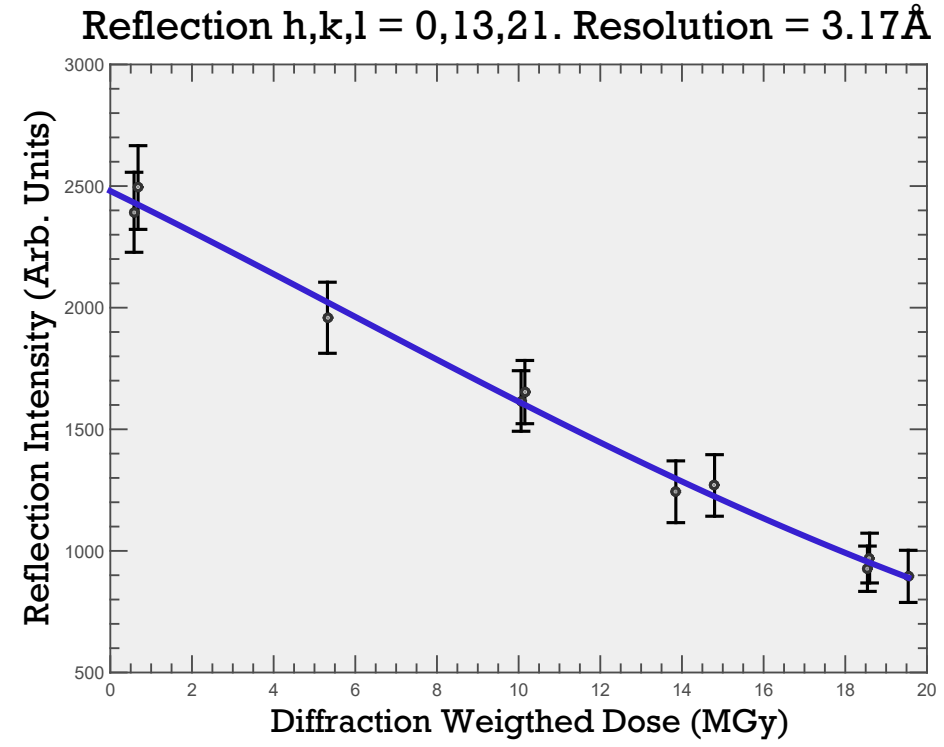
Table 3.2: Regression fit quality indicators.

R_{fit}	R_{fit}^{thresh}	R_{zero}	R_{zero}^{thresh}
0.136	3.756	0.098	0.180

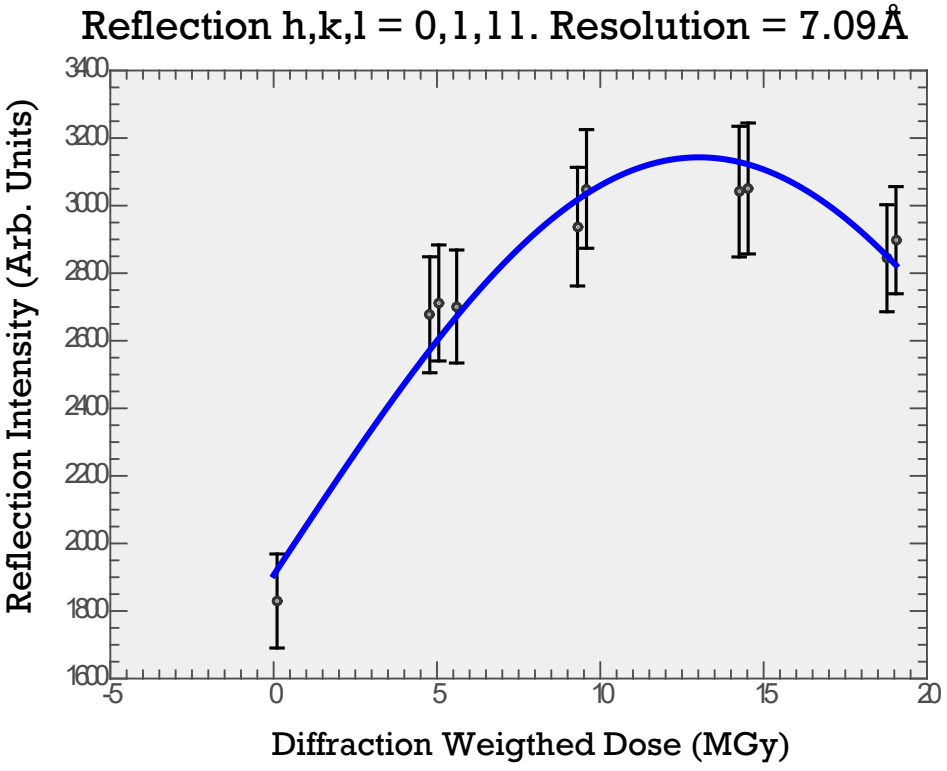
A total of 507 (40.40%) reflections were rejected for the regression fits and were thus extrapolated using the probabilistic approach. The calculated distributions for 2 centric reflections are shown in Figure 3.11. The extrapolated intensities for negative reflections are made positive because the Wilson distribution has a zero probability value for negative intensity values. The intensity values for weak reflections are significantly affected by the prior distribution. This is evidenced by the fact that the posterior distribution resembles the prior distribution for weak reflections (Figure 3.11a). In contrast, strong reflections are almost unaffected by the prior distribution (Figure 3.11b). As a result, the scaling factor $f(r)$ dominates the resulting zero-dose extrapolated intensity value for strong reflections.

Figure 3.12 shows the intensities of all 5000 reflections before and after the extrapolation. The overall take away message is that there is a general increase in the intensity values after performing the extrapolation. This is expected if the true zero-dose intensity values were obtained, because the overall relative intensity decreased during the experiment (Table 3.1).

The red and blue solid lines in Figure 3.12 represent the spherically averaged intensity observations in the resolution shells before and after the extrapolation respectively. The relative

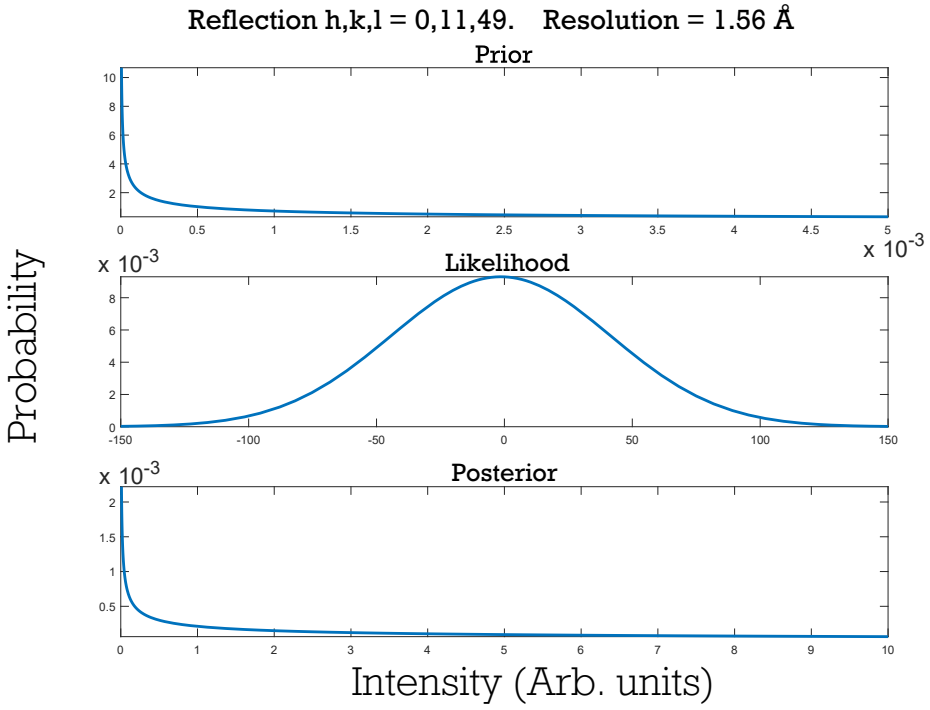


(a)

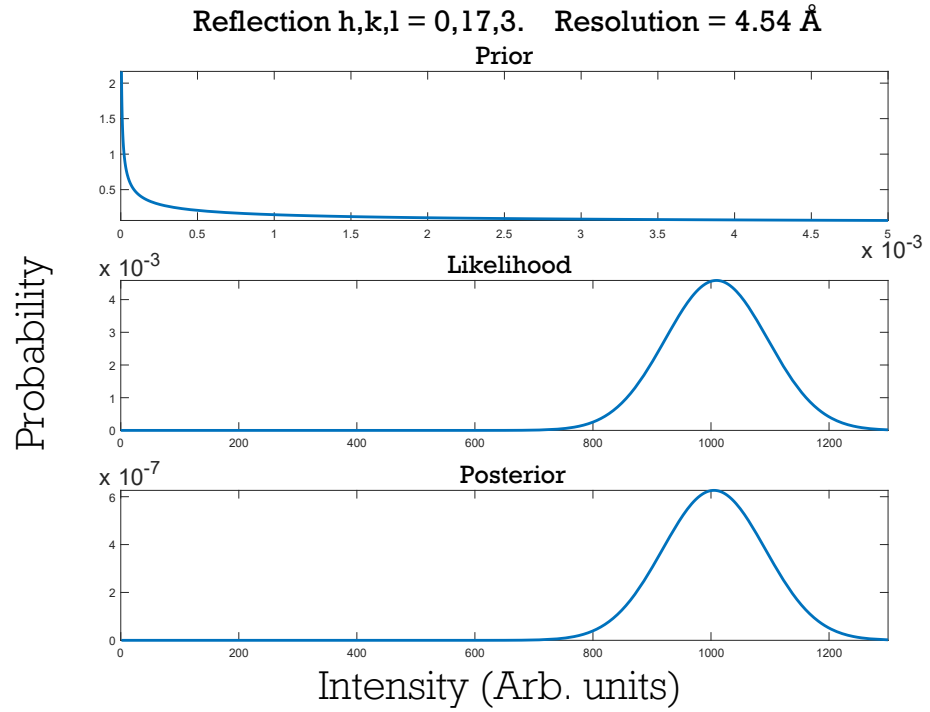


(b)

Figure 3.10: Regression fits that satisfied the regression criteria outlined in section 3.3.1 for two centric reflections. (a) The intensity of this reflection decreases linearly and monotonically. (b) The intensity increases before it starts to decrease. These are examples of the variety and non-linearity in the behaviour of individual reflection intensities.



(a)



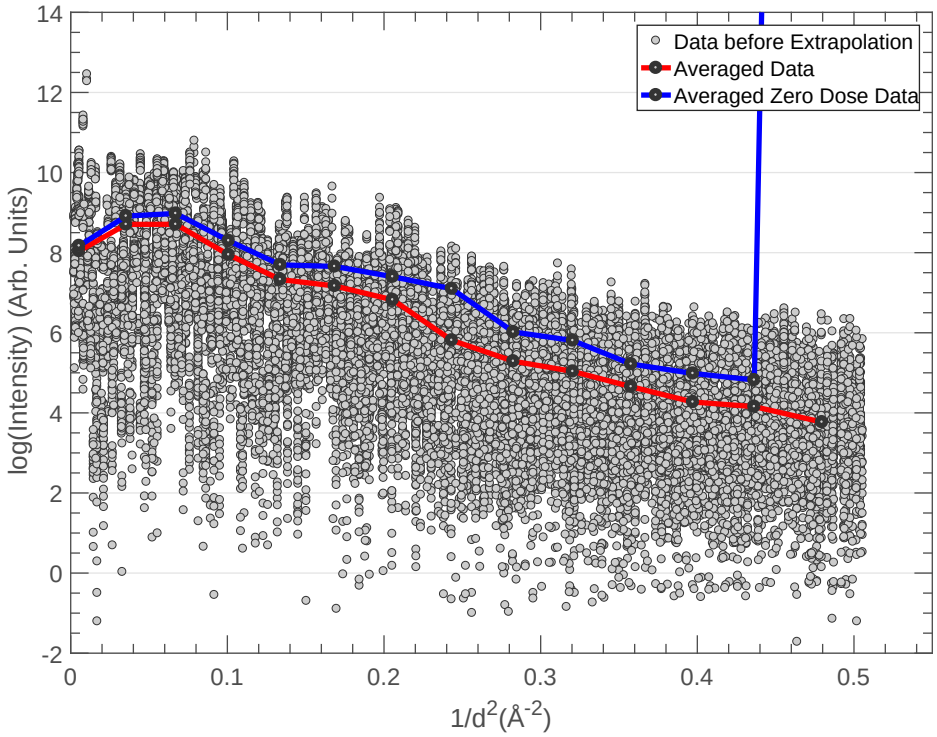
(b)

Figure 3.11: Prior, likelihood and (unnormalised) posterior distributions for two centric reflections. (a) Reflection 0,11,49. The observed intensity is negative, I_{obs} = -1.4, with a relatively large standard deviation (σ_{obs} = 42.9) in comparison. The posterior distribution resembles the prior distribution which is a consequence of the fact that the reflection intensity is weak. The resulting zero-dose value calculated as the expected value from the posterior distribution is 19.35. (b) Reflection 0,17,3. I_{obs} = 1009.33 and σ_{obs} = 86.99). This reflection is very strong (I_{obs}/σ_{obs} = 11.60), hence the posterior distribution resembles the likelihood distribution. The calculated zero-dose value is 1005.02.

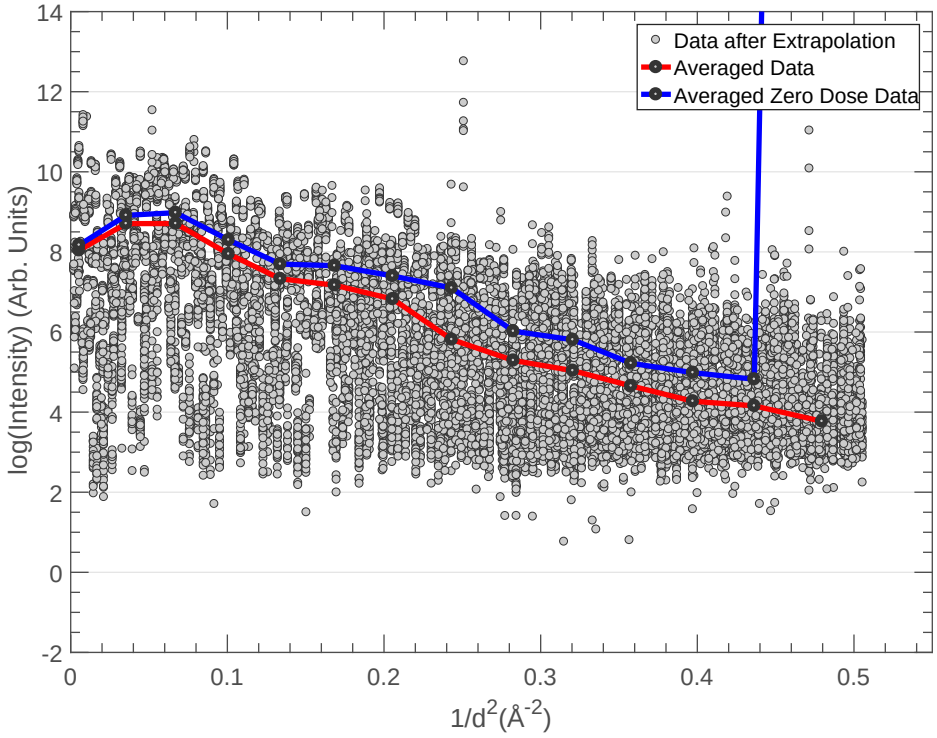
increase in the extrapolated intensities is greater on average for higher resolution reflections, as evidenced by the increase in the gap between the red and blue solid lines with increasing resolution. This is expected because the intensity of higher resolution reflections are known to decay more quickly than low resolution reflections.

There are some reflection observations that have abnormally high extrapolated intensity values. These act to highly skew the spherically averaged mean intensity of resolution shells, as can be seen by the blue line sharply increasing for the highest resolution shells. If these reflections are removed by only using zero dose intensities that are below six standard deviations above the mean value of the original uncorrected intensities then the sharp increase in the average intensity is removed (Figure 3.13). For this resolution bin (1.48Å - 1.41Å) there are 59 reflections (2.57% of the reflections in that bin) that exceed this threshold, the maximum of which has an intensity of 1.75×10^{42}. These excessively high values explain the sharp increase in the spherically averaged intensity at the highest resolution shell in Figure 3.12. These high intensity values are not a result of the regression based extrapolation because the averaged zero-dose intensities after the regression analysis do not exhibit this feature (Figure 3.6). Therefore these are the result of the probabilistic extrapolation and are most likely due to applying $f(r) \approx s(D_{obs})$ to a reflection intensity where the reflection does not behave like the "average" reflection in the resolution shell. These outliers could simply be rejected using suitable outlier criteria. An obvious choice would be to reject observations where the intensity is above a given threshold. A more sophisticated outlier rejection method using Wilson statistics could be implemented (Read, 1999).

A feature of Figure 3.12b that may not be expected is that many of the data points have shifted away from zero. This is unexpected because the prior for the reflections is centred on zero (Figure 3.11). It should be noted that Figure 3.12b only includes positive intensity observations and therefore does not include the predicted zero dose intensities of 215 observations. However, the bias in the figure is unexplained and could be due to a bug in the numerical implementation of the Bayesian calculation.



(a)



(b)

Figure 3.12: Reflection intensities of 5000 reflections before and after zero-dose extrapolation. The points on the red and blue solid lines represent the mean intensity of the resolution shells before and after the extrapolation respectively and are shown on both (a) and (b) for ease of viewing. As expected, there is a general increase in the reflection intensities after extrapolation.

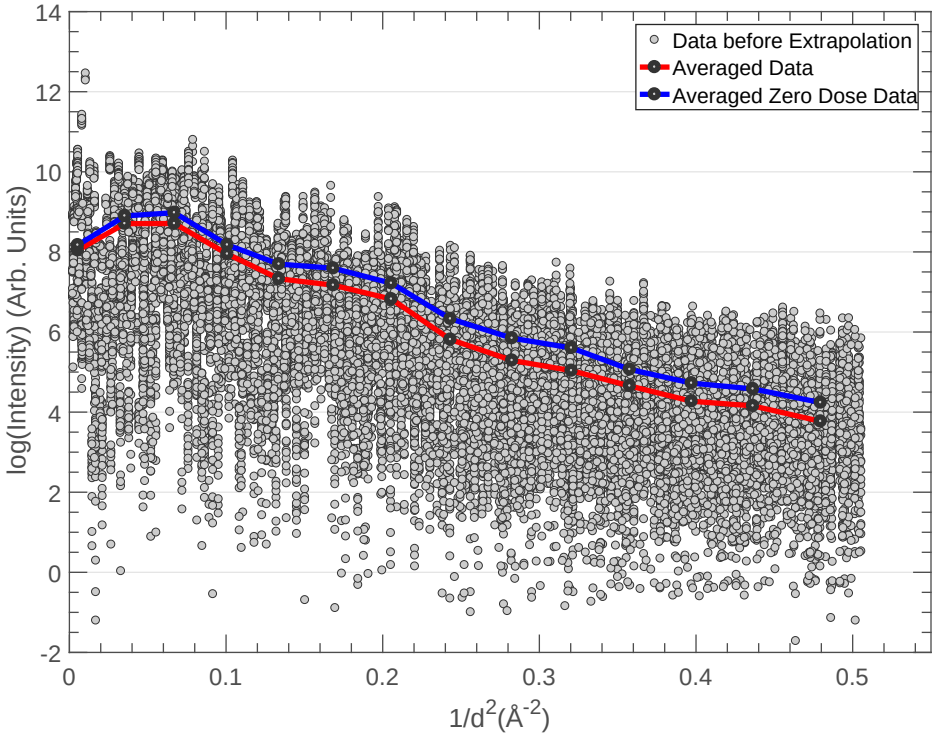


Figure 3.13: Reflection intensities of 5000 reflections before zero-dose extrapolation, however, this time the spherically averaged zero-dose data have been calculated without contribution from the abnormally high intensity values. It can be seen that the spherically averaged data no longer diverge as they did in Figure 3.12.

3.5 Discussion

As outlined in the introduction, global radiation damage in MX results in changes in the intensities of reflections throughout the experiment. These intensity changes ultimately compromise the quality of the data and can affect the biological interpretations. Crystallographers attempt to correct for these intensity changes during scaling by applying some form of overall damage correction. These corrections do not account for specific intensity changes and can therefore introduce systematic errors (Diederichs *et al.*, 2003).

Zero-dose extrapolation is a method that attempts to correct for the specific intensity changes to improve data quality. It has already been shown to improve phasing statistics and SAD electron density maps (Diederichs *et al.*, 2003). However the linear model used in the 2003 study by Diederichs *et al.* was only valid for a limited dose range, and fails to capture the variable behaviour of reflection intensity changes. Quadratic and exponential models have also been explored with some success (Diederichs, 2006), but ultimately the extrapolation is rarely used/effective ((Borek *et al.*, 2007), Ed Lowe and Phil Evans, personal communication).

The work presented in this chapter takes advantage of a decade of additional radiation damage research to revisit zero-dose extrapolation. The advances of particular interest to the research are the development of RADDOSE-3D, the DWD metric and the Leal *et al.* dose decay model.

Data from a uniformly irradiated insulin crystal (section 2.2.3) were used to perform the extrapolation. Dose calculations were carried out using RADDOSE-3D and the DWD metric was used as the independent variable in which to track the changes in reflection intensities. In a similar manner to the approach used by Diederichs *et al.*, some of the reflection intensities were extrapolated to zero-dose using regression analysis. However, in the method presented here, multiple criteria had to be satisfied by a reflection before the extrapolation was carried out on it. This was done to reduce the likelihood of overfitting the model to the data. These included checking that the fits correlated well with the data, there were a minimum number of observations, the intensity observations were not too small and that the extrapolated intensity values did not give intensity values that were significantly different from the observed distribution.

To check the quality of the model fits and the zero-dose extrapolated intensity values, the R_{fit} and R_{zero} metrics were created. Furthermore, the R_{fit}^{thresh} and R_{zero}^{thresh} metrics were used as thresholds to compare with the R_{fit} and R_{zero} and ensure that the overall regression extrapolations were reasonable.

Further to the regression based analysis, a probabilistic extrapolation using Bayesian inference was applied to the subset of data that was unsuitable for regression. Using the data from the already extrapolated reflection observations, it was possible to estimate distribution parameters for the prior and likelihood distributions, which ultimately enabled estimates of the zero-dose intensity values for other observations. This approach is new for zero-dose extrapolation and means that only a subset of the data are required to have a high enough multiplicity. It is an important development because the low multiplicity of diffraction data is generally regarded as a major problem with the traditional regression based approach.

The extrapolation method presented here gave results consistent with that which was expected. The Leal *et al.* model was able to capture some of the non-linear and non-monotonic behaviour of the reflection intensities, and for the majority of reflections, the probabilistic

method seemed to give reasonable zero-dose estimates. In general the extrapolated intensities of reflections showed an increase from the observed values.

However, the method presented still requires some work before it can be used reliably for general crystallographic data. The most obvious improvement is to find a method to prevent the probabilistic extrapolation method giving abnormally high intensity values. As mentioned earlier, this issue could be corrected by using an outlier rejection method. There is also currently no check for the quality of the extrapolation from the probabilistic approach, which would aid in highlighting potential problems. A possible validation criteria may be that the probabilistic extrapolation gives values that are similar to the regression based extrapolation values for the reflections that have already been extrapolated. This basically checks the consistency of the two methods. The reason that the consistency check is a good way to validate that the probabilistic method works is that the regression based method is already validated with several checks and its own validation metrics. The drawback with this is that the distribution parameters are calculated from the intensities extrapolated using the regression method. This means that it is more likely that the two methods will seem consistent, but there is no guarantee that it will be reliable for the reflections that were not extrapolated using the regression approach. Another approach to assessing the quality of the extrapolation is the using cross validation as is typically used in refinement in MX (Brünger, 1997). This involves keeping a relatively small subset of reflections aside and performing the extrapolation on the remaining data, then finally assessing the quality of the fits using all of the reflections at the end. The potential problem with this approach is that there is generally not enough data to start with due to low multiplicity and the method requires a subset of reflections to have a sufficient multiplicity to perform the regression based extrapolation. Therefore the "free" observation set may not be able to be picked completely randomly in all cases.

It is likely that the abnormally high extrapolated intensities are caused by applying the scale factor, $s(D)$, to reflections that do not obey the behaviour of the scale function. $f(r)$ is a factor that was introduced to dampen the proportion of $s(D)$ that is applied to reflection observations. But there are two problems. Firstly, it is assumed that the similarity of the behaviour of a reflection to that of the scale factor is a function of the difference between the observed intensity value and the mean intensity value, r, in the appropriate resolution shell. However, there is no apparent evidence to support this assumption, so it may not

hold. Secondly, the parameter values A and B in the equation 3.3.19 are not guaranteed to be optimal. They were chosen by visual inspection of the behaviour of a limited set of observations and therefore were not rigorously determined.

The current method calculates the distribution parameter for the prior distribution using information obtained by spherically averaging the zero-dose extrapolated values. This should work fine when the data are isotropic, but could result in large errors when the data are anisotropic. A method that accounts for any potential anisotropy in the data is necessary to reduce these errors.

Furthermore, there are improvements that could be made to the regression based extrapolation. It was observed that a "bad" correlation between the fitted model and the data was simply due to noise. This implies that the correlation coefficient check may be too stringent, and in fact some reflections may be rejected despite the possibility that the model would have given a reasonable extrapolated zero-dose intensity. Perhaps removing the correlation requirement in favour of a different criterion could improve the data quality by not rejecting "well behaved" reflections due to systematic noise. The reason this was not done in the present analysis is because it was thought to be more important to have fewer reliable zero-dose estimates than a huge number where some estimates were questionable.

The corrected standard deviation equation (3.3.7) is based on the form of the analogous equation in (Diederichs *et al.*, 2003). Other forms of the correction function should be explored with a good physical/statistical justification.

A limitation of using the Leal *et al.* model for the regression fits is that the model assumes positive intensity values for all (real) dose values. As a physical model, this assumption is completely valid; however actual intensity data can be both positive and negative. In the regression analysis, any reflection which had too many negative/small intensity values was rejected because this could lead to unreliable extrapolated intensity values. Further work could include suggesting alternative physical dose decay models or amendments of the Leal *et al.* model to satisfactorily account for negative intensity values. However, it is currently reasonable to suggest that perhaps a single parametric dose decay model is not capable of describing the relationship between the change in intensity and the dose. With this in mind, it could be the case that non-parametric regression techniques, such as Gaussian process regression, may prevail over parametric ones. The problem is that non-parametric

regression techniques often require more data for a reliable fit. Furthermore, in my own experience, they are good for interpolation but not necessarily extrapolation.

The current implementation of the extrapolation algorithm was written in the Matlab programming language. This was a suitable language to perform the data exploration necessary to do the analysis. However the bottleneck with the program is reading the intensity data from the MTZDUMP output. Matlab is over 1000 times slower than C for parsing integers (Bezanson *et al.*, 2014). To make the program useable to the general crystallographic community, the program would need to be written in an open source and free programming language. Python and C/C++ are currently the standard languages in the crystallographic community and would thus make good candidates.

CHAPTER 4

A Markovian Data Reduction Framework

4.1 Introduction

The radiation damage correction models implemented thus far have all focussed on correcting reflection intensities. Scaling methods generally employ an average resolution dependent correction to all reflection intensities (Evans, 2006; Evans and Murshudov, 2013; Kabsch, 2010a), whereas specific correction methods employ regression analysis on individual reflections (Diederichs *et al.*, 2003; Diederichs, 2006). However correcting the reflection intensities presents many problems because each independent reflection exhibits its own behaviour, which is not necessarily linear, nor monotonic (Abrahams and Marsh, 1987). Therefore, rather than addressing the problem of damage at the level of the reflection intensities, an alternative approach is to track the changes at the level of the sample undergoing the radiation damage.

This chapter presents a (parametric) time series model used to describe crystal changes during the X-ray crystallography experiment as a Markovian process. Within this framework existing algorithms - the Unscented Kalman filter and the Unscented Rauch-Tung-Striebel smoother - are used to determine the "optimal" values of the underlying crystal state, defined as the set of structure factor amplitudes, at each point in time during the MX experiment.

4.2 Why Julia?

The code for the algorithm presented in this chapter was all written in a recently released programming language called Julia. Given that this language is still in its beta version (it has yet to reach version 1 release) and is relatively unknown in the crystallography community, this choice may seem very unorthodox, so it is worth discussing why this language was chosen.

The total time of the implementation and run time of any piece of code can be crudely given as

$$\text{Total time} = \text{Computation time} + \text{Development time}$$

A compromise between the development time and the computation time has to be made,

with the additional constraint that the developer's time is more important than the time of a computer. Given that the development of a new algorithm requires a lot of data and parameter exploration, writing prototype code in a low-level language such as C/C++ or Fortran would not necessarily be the optimum choice, particularly when the developer is unfamiliar with these languages. A popular alternative is to use a high-level dynamic language such as Python, R or Matlab, which are usually designed, or have packages to support technical computing. The productivity that can be achieved with these languages is counteracted by the relatively slow computation time when compared to low-level languages.

Julia is a dynamic language designed for technical computing (Bezanson *et al.*, 2014, 2012) that is also very fast. Generally the computational performance of algorithms written in Julia executes within a factor of 2 of the speed of C (Figure 4.1). Therefore despite having to learn

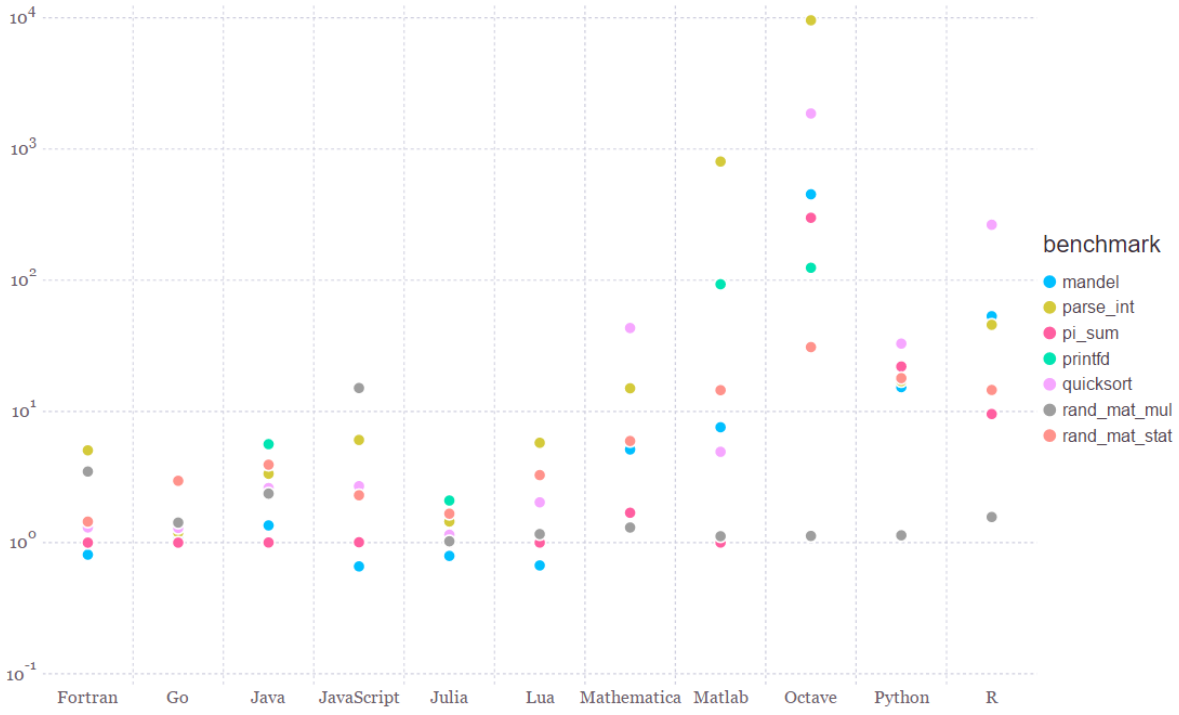


Figure 4.1: Benchmark times taken to run a given algorithm for various languages relative to C (smaller is better, C performance = 1.0). The Python implementations of rand_mat_stat and rand_mat_mul use NumPy (v1.9.2) functions, the rest are pure Python implementations. The table of data for this plot can be found on the main Julia programming language webpage, `http://julialang.org/`, along with a link to the plot.

the Julia language, it seemed to be the best trade-off for coding the algorithm described here.

As a dynamic language that employs type inference, Julia code can be written in such a way that it compiles to non-optimal machine code.. This means that Julia has to be written in a particular manner to achieve the performance claimed by the language authors. A further

argument for using Python or C/C++ over Julia is the large library of existing crystallographic packages that have been written in those languages. However Julia has built-in support for calling methods from C/Fortran libraries with a single line of code. Additionally, the `PyCall` package was written in Julia to allow Python libraries to be accessed directly from Julia with a single line of code. Thus the crystallographic libraries were still easily accessible.

4.3 A hidden Markov model of the data collection experiment

The data collection experiment can be viewed as a time series: a sequence of diffraction data generated by a time-dependent process. Time series analysis seeks to understand the underlying processes that produced the data and allow forecasting or monitoring of the process. A more accurate analogy for the data collection experiment would be to describe it as a dose series, because the changes in the crystal state (and hence the observed data) are generally attributed to a dose dependent process.

Figure 4.2 shows a schematic of the dose series model of the experiment. At time t = i - 1 the crystal is in its initial state where the atoms have relatively well defined positions. After an initial X-ray exposure the crystal state has changed at time t = i. The atomic positions have changed and they have an effective smearing of their position due to their increased atomic B-factors. It is this state of the crystal that gives rise to the diffraction pattern at time t = i. The process repeats itself at time t = i + 1 and beyond. Thus each diffraction image can be regarded as being generated from a different but related crystal. In the model, the only components that are observed are the diffraction patterns, despite the fact that the crystal states are the desired quantities. The crystal states are effectively 'hidden'. Furthermore, the changes in crystal state as a result of X-ray exposure are assumed a Markovian process i.e. the state of the crystal at time i is completely determined by the state of the crystal at time i - 1. The model described here is known as a hidden Markov model.

The problem can now be stated as: **what is the most likely sequence of crystal states that generated the sequence of observed diffraction patterns?**

Before addressing the question, it is necessary to understand what is meant by the crystal state. The state of the crystal is defined by its constituent atoms and their positions. Equivalently the state of the crystal can be described by the set of structure factors in reciprocal

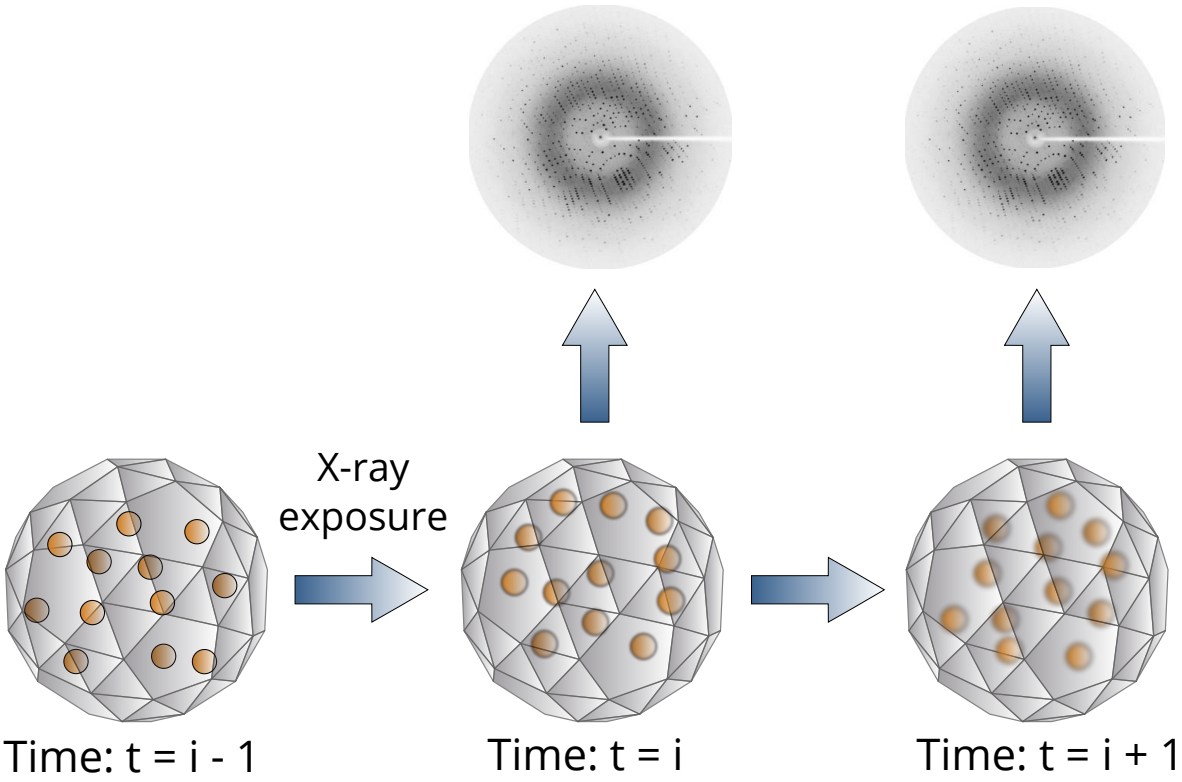


Figure 4.2: Hidden Markov model representation of the diffraction experiment. At time t = i - 1 the crystal is in an undamaged state and the constituent atoms have fairly well defined positions. After an X-ray exposure the crystal changes state at time t = i. Some of the constituent atoms change positions and their atomic B factors increase, which is represented by a slight blurring of the atoms. However the state of the crystal is not observed by the experimenter, instead the experimenter observes the diffraction image that is generated as a result of the exposure. This process repeats itself, typically until enough images have been collected to solve the structure.

space: amplitudes and their corresponding phases. However, at the data reduction stage of the crystallographic structure solution pipeline, the phases are completely unknown. Therefore the state of the crystal is represented solely by the set of structure factor amplitudes.

4.3.1 Mathematical Notation

This chapter contains many formulas and symbols so a glossary of terms that define the notations for this chapter is provided below.

Symbol	Definition
$\boldsymbol{F}_i$	Structure factor of a reflection at discrete time point i.
$\lvert\boldsymbol{F}_i\rvert = F_i$	Structure factor amplitude of a reflection at discrete time point i
F_c	Structure factor amplitude to be calculated using Bayesian inference.
F_0	Initial structure factor amplitude calculated from a cycle of the forward-backward algorithm (FBA).
$\{\lvert\boldsymbol{F}\rvert\}_i$	Set of structure factor amplitudes at discrete time point i
$\{O\}_i$	Set of intensity observations on an image collected at discete time point i
$P_a(\boldsymbol{F}_{i+1};\boldsymbol{F}_i)$	Probability of a structure factor, $\boldsymbol{F}_{i+1}$, of an acentric reflection at time point $i+1$ conditional on the structure factor value, $\boldsymbol{F}_i$, at time i.
$P_c(\boldsymbol{F}_{i+1};\boldsymbol{F}_i)$	Probability of a structure factor, $\boldsymbol{F}_{i+1}$, of a centric reflection at time point $i+1$ conditional on the structure factor value, $\boldsymbol{F}_i$, at time i.
I_i	Intensity of a reflection observed on the image generated at time point i.
K	Scale factor which is assumed constant throughout the experiment.
B_i	Scaling B factor calculated from the image generated at time point i.
$\Delta\boldsymbol{r}$	Average coordinate error.
$\boldsymbol{s}$	Reciprocal space vector.
ΔB	Change in B between consecutive time points i and $i+1$.
Σ_N	Expected intensity of a reflection.
σ^2	= $(1-\lvert\boldsymbol{D}\rvert^2)\Sigma_N$ is the variance.
ε	Multiplicity of a reflection.
$\boldsymbol{D}$	= $\exp\left(-2\Delta B\frac{\sin^2(\theta)}{\lambda^2}\right)\cos\left(2\pi\boldsymbol{s}\cdot\Delta\boldsymbol{r}\right)$. (Although in the general formulation (see Read (1990)) $\boldsymbol{D}$ is a complex number is hence given a bold symbol).
$\boldsymbol{D}\boldsymbol{F}_i$	The (complex) product of the structure factor, $\boldsymbol{F}_i$, with the multiplier, $\boldsymbol{D}$.

4.3.2 Bayesian optimal filtering

Methods used to estimate the (hidden) state, $\boldsymbol{x}_j$, of a time-varying system observed indirectly via noisy measurements, $\boldsymbol{y}_j$, at a given point in time, j, are known as *optimal filtering* methods. The filtering distribution can be defined mathematically as

$$P(\boldsymbol{x}_j;\boldsymbol{y}_1,\ldots,\boldsymbol{y}_j), \tag{4.3.1}$$

i.e. the probability distribution of the state given all of the previous observations up to time point j. Filtering is sometimes regarded as a *forwards pass* through the data. There are many ways to define what is meant by *optimal* (Chen, 2003): here the optimality criterion is the minimum mean-squared error (MMSE) estimate of the state of the system. Bayesian filtering refers to the formulation of an optimal filter within a Bayesian framework. A Bayesian optimal filter can therefore be used to solve the problem of determining the crystal states given a set of (noisy) diffraction images. Due to the non-linear relationship ($I \propto |\boldsymbol{F}|^2$) between the reflection intensity, I, and the structure factor amplitudes, $|\boldsymbol{F}|$, it is necessary to use a non-linear Bayesian optimal filter for the crystallographic diffraction experiment problem. The filter chosen is the Unscented Kalman filter (UKF) because it propagates the probability density function in an effective manner (using the unscented transform) to achieve up to third order accuracy in the posterior mean and covariance estimates (Wan and Van Der Merwe, 2000). The UKF algorithm is outlined in Wan and van der Merwe (2002).

4.3.3 Bayesian smoothing

Bayesian Smoothing can be considered to be a class of methods within the field of Bayesian filtering. Whereas filters generally compute estimates of the system state based on the observation history, smoothers use all of the available information and thus they can estimate states that happened before the current time (Särkkä, 2013). Mathematically the smoothing distribution is

$$P(\boldsymbol{x}_j; \boldsymbol{y}_1, \ldots, \boldsymbol{y}_j, \ldots, \boldsymbol{y}_\tau) \tag{4.3.2}$$

where $j < \tau$. Smoothers are regarded as *backwards passes* because they can be used to estimate states prior to the current time.

The goal of the data reduction stage in the macromolecular structure solution pipeline is to reduce the intensity values to accurate estimates of the structure factor amplitude for each reflection. If this problem is phrased in a probabilistic manner, then the distribution of interest is

$$P(\{|\boldsymbol{F}|\}_i\,;\, \{O\}_1, \ldots, \{O\}_i, \ldots, \{O\}_\tau), \tag{4.3.3}$$

where $\{|\boldsymbol{F}|\}_i$ is the set of structure factor amplitudes and $\{O\}_i$ is the set of intensity measurements, I_{hkl} at time point i. Equation 4.3.3 is the probability distribution that describes

the values of the set of structure factor amplitudes given all of the observed data on the diffraction images. Comparison of equations 4.3.3 and 4.3.2 show that they are identical and hence using a Bayesian smoother (after application of the UKF) should provide the desired solution to the data reduction problem. The application of both the forwards and backwards passes is known as the *forward-backward algorithm*. The important difference of this formulation compared to the current data reduction methods is that the set of structure factor amplitudes is found for every time point, i, as opposed to just producing a single set of amplitudes from the data.

The Bayesian smoother chosen for this problem was the unscented Rauch-Tung-Striebel smoother (URTSS) (Särkkä, 2008).

4.3.4 Process function and covariance

In order to apply the UKF it is necessary to define the function that relates the crystal state at time point i to the crystal state at time point i + 1. This function is known as the process function, and is typically taken as the expected value of a conditional probability distribution (transition probability) of the probability of crystal state at i + 1 given the crystal state at i. The process covariance is also an important quantity because it effectively describes the level of uncertainty of the process.

The crystal state changes as a result of the X-ray exposure (the duration of which will differ for a single diffraction image depending on the goal of the experiment) and the magnitude of these changes will vary depending on the radiation sensitivity of the irradiated crystal. It is assumed that X-ray irradiation on the timescale of image collection is short enough such that the change in crystal state is fairly small. Thus the crystal state at one image is closely related to the crystal state for the subsequent image. This assumption along with the assumptions that changes in structure factors are independent* and identically distributed, gives the Luzzati distributions (Luzzati, 1952; Read, 1990; Pannu and Read, 1996)

$$P_a(\boldsymbol{F}_{i+1};\boldsymbol{F}_i) = \frac{1}{\pi\varepsilon\sigma^2}\exp\left(-\frac{|\boldsymbol{F}_{i+1}-\boldsymbol{D}\boldsymbol{F}_i|^2}{\varepsilon\sigma^2}\right), \tag{4.3.4}$$

*In reality structure factors are not independent, however this assumption simplifies the equations significantly and still provides useful information (Pannu and Read, 1996).

for acentric reflections and

$$P_c(\boldsymbol{F}_{i+1};\boldsymbol{F}_i) = \frac{1}{\left[2\pi\varepsilon\sigma^2\right]^{1/2}}\exp\left(-\frac{|\boldsymbol{F}_{i+1}-\boldsymbol{D}\boldsymbol{F}_i|^2}{2\varepsilon\sigma^2}\right), \tag{4.3.5}$$

for centric reflections where $P(\boldsymbol{F}_{i+1};\boldsymbol{F}_i)$ denotes the probability of structure factor, $\boldsymbol{F}_{i+1}$, at time i + 1 given the structure factor, $\boldsymbol{F}_i$, at time i, ε is the multiplicity of the reflection, $\sigma^2 = (1-|\boldsymbol{D}|^2)\,\Sigma_N$, where Σ_N is the expected intensity of the reflection. In this work the expected intensity value was calculated as the sum of the squared atomic scattering factors, and $\boldsymbol{D}$ is a (complex) multiplier, which quantifies the effects of crystal perturbations on the structure factor and is explicitly defined in section 4.3.6.

As discussed previously, the phases are unknown during the data reduction stage and hence the only quantity that can be inferred is the amplitude of the reflection. The unknown phase can be integrated over equations 4.3.4 and 4.3.5 (marginalisation) to obtain the probability of the structure factor amplitudes (transition probability):

$$P_a(|\boldsymbol{F}_{i+1}|;|\boldsymbol{F}_i|) = \frac{2|\boldsymbol{F}_{i+1}|}{\varepsilon\sigma^2}\exp\left(-\frac{|\boldsymbol{F}_{i+1}|^2+\boldsymbol{D}^2|\boldsymbol{F}_i|^2}{\varepsilon\sigma^2}\right)I_0\left(\frac{2|\boldsymbol{F}_{i+1}|\boldsymbol{D}|\boldsymbol{F}_i|}{\varepsilon\sigma^2}\right), \tag{4.3.6}$$

$$P_c(|\boldsymbol{F}_{i+1}|;|\boldsymbol{F}_i|) = \left[\frac{2}{\pi\varepsilon\sigma^2}\right]^{1/2}\exp\left(-\frac{|\boldsymbol{F}_{i+1}|^2+\boldsymbol{D}^2|\boldsymbol{F}_i|^2}{2\varepsilon\sigma^2}\right)\cosh\left(\frac{|\boldsymbol{F}_{i+1}|\boldsymbol{D}|\boldsymbol{F}_i|}{\varepsilon\sigma^2}\right) \tag{4.3.7}$$

where $I_0(\cdot)$ is the zero order modified Bessel function of the first kind and $\cosh(\cdot)$ is the hyperbolic cosine function. Note equation 4.3.6 is known as the Rice distribution and equation 4.3.7 is known as the Woolfson distribution (Woolfson, 1956; McCoy, 2004).

The mean (process function), μ_{Rice}, and variance (process variance), σ^2_{Rice}, for acentric structure factor amplitudes can are calculated by integrating equation 4.3.6 to give

$$\mu_{Rice} = \sigma\sqrt{\frac{\pi}{2}}L_{1/2}\left(-\frac{\boldsymbol{D}^2|\boldsymbol{F}_i|^2}{2\sigma^2}\right), \tag{4.3.8}$$

$$\sigma^2_{Rice} = 2\sigma^2+\boldsymbol{D}^2|\boldsymbol{F}_i|^2-\frac{\pi\sigma^2}{2}L^2_{1/2}\left(-\frac{\boldsymbol{D}^2|\boldsymbol{F}_i|^2}{2\sigma^2}\right), \tag{4.3.9}$$

where

$$L_{1/2}(x) = \exp(x/2)\left[(1-x)I_0\left(\frac{-x}{2}\right)-xI_1\left(\frac{-x}{2}\right)\right] \tag{4.3.10}$$

is the Laguerre polynomial and $L^2_{1/2}$ denotes the square of the Laguerre polynomial $L_{1/2}$. $I_1(\cdot)$ is the first order modified Bessel function of the first kind (den Dekker and Sijbers,

2014). For strong reflections the corresponding Rice distribution for the amplitude can be approximated well with a Gaussian function. In this case the process function and covariance become

$$\mu_{Gauss} = \boldsymbol{D}\boldsymbol{F}_i \tag{4.3.11}$$

$$\sigma^2_{Gauss} = \left(1 - |\boldsymbol{D}|^2\right)\Sigma_N. \tag{4.3.12}$$

For centric reflections the covariance is simply twice the variance for acentric reflections (Terwilliger and Berendzen, 1996). The process function for centric reflections is assumed identical to the process function for acentric reflections for the sake of simplicity. For strong reflections this assumption is valid, since the Gaussian distributions (equations 4.3.4 and 4.3.5) differ only in the variance. For weak reflections this assumption breaks down and the expected value of the Woolfson distribution should be explicitly calculated.

4.3.5 Observation function and covariance

In addition to the process function, it is necessary to define the process by which diffraction images are generated from the crystal state. This is known as the observation function. In an analogous manner to the process function, the observation function is taken as the expected value of a conditional probability distribution (emission probability) describing the probability of the intensity of a reflection I_i, given the structure factor amplitude $|\boldsymbol{F}_i|$ at time i. The observation model is given by (Otwinowski *et al.*, 2003)

$$I = K|\boldsymbol{F}|^2, \tag{4.3.13}$$

where K is the scale factor. However, due to the measurement process being inherently noisy, the process is better approximated as a probability distribution. Assuming a normally distributed measurement error, the emission probability is given by

$$P(I_i; |\boldsymbol{F}_i|) = \frac{1}{\sigma_m\sqrt{2\pi}} \exp\left(-\frac{(I_i - K|\boldsymbol{F}_i|^2)^2}{2\sigma_m^2}\right), \tag{4.3.14}$$

where σ_m^2 is the variance of the measurement process, which is given as a result of the integration process. Not all reflections are observed on a diffraction image, so these obser-

vations are not given a variance value by the integration software. The variance for these missing reflections is therefore given a large value (e.g. 10^{10}) to effectively represent an infinite uncertainty on the observation for the image.

4.3.6 Obtaining parameter values for the process and observation functions

The process and observation functions are parameterised by $\boldsymbol{D}$ and K respectively and hence the values of these parameters must be determined. The multiplier $\boldsymbol{D}$ is given by (Murshudov *et al.*, 1997; Leal *et al.*, 2012)

$$\boldsymbol{D} = \exp\left(-2\Delta B\frac{\sin^2(\theta)}{\lambda^2}\right)\cos\left(2\pi\boldsymbol{s}\cdot\Delta\boldsymbol{r}\right). \tag{4.3.15}$$

where ΔB is the change in scaling B factor and $\Delta\boldsymbol{r}$ is the average coordinate error from time point i to time point i + 1, θ is the Bragg angle and λ is the wavelength of the incident X-ray. (All references to the B factor in the rest of the chapter refer to the scaling B factor unless otherwise stated). Implicit to the form of $\boldsymbol{D}$ in equation 4.3.15 is the fact that the B-factor is assumed to be isotropic. Furthermore, it is assumed that the irradiation process only changes the B factor (not the coordinate error of the atoms i.e. $\Delta\boldsymbol{r} = 0$) and the change in B-factor is the same for every atom in the structure (the Wilson B-factor). This reduces equation 4.3.15 to

$$\boldsymbol{D} = \exp\left(-2\Delta B\frac{\sin^2(\theta)}{\lambda^2}\right). \tag{4.3.16}$$

Thus $\boldsymbol{D}$ is ultimately determined by the change in B factor, ΔB.

The B and scale factor can be determined using the scaling equation

$$I_{obs} = K\sum_j f_j^2\exp\left(-2B\frac{\sin^2(\theta)}{\lambda^2}\right), \tag{4.3.17}$$

where I_{obs} is the observed intensity and f_j is the atomic scattering factor for an atomic species within the unit cell, which can be calculated for a given reflection using the Cromer-Mann coefficients (Cromer and Mann, 1968). Taking the natural logarithm of both sides and rearranging yields.

$$\ln\left(\frac{I_{obs}}{\sum_j f_j^2}\right) = \ln(K) - 2B\frac{\sin^2(\theta)}{\lambda^2}. \tag{4.3.18}$$

Plotting $\ln\left(\frac{I_{obs}}{\sum_j f_j^2}\right)$ against $\frac{\sin^2(\theta)}{\lambda^2}$ gives a straight line where $\ln(K)$ is the intercept and $-2B$ is the gradient. This procedure can be performed for each image to obtain a sequence of K and B values. These values will be noisy because no single image contains the entirety of reciprocal space (i.e. no image contains all reflections). Assuming the scale and B functions to be smooth, continuous functions can be fitted through the values obtained from the data to get estimates of the true scale and B factors for each image. For simplicity the B factor is assumed linear and the scale factor is assumed constant. The assumption of linearity for the B factor is valid for low dose ranges (Kmetko *et al.*, 2006; Borek *et al.*, 2007; Leal *et al.*, 2012) and hence ΔB can be given as the difference in B factor between images. On the other hand the constant scale factor assumption is not valid for general MX experiments, but may be a suitable approximation for the case where a crystal is completely immersed in a top-hat profile X-ray beam, as for instance was the case for the data collection described in Chapter 2.

4.3.7 Convergence of the forward-backward algorithm

The initial estimate of the structure factor amplitude required to start the forward-backward algorithm may not be close to the true value. The algorithm will thus propagate the incorrect amplitude value through the filter until the time point when the first actual observation is made. From then on the estimates should be closer to the true values of the amplitude in the experiment. In particular, the backwards pass should result in a better estimate of the true initial amplitude. The improved initial amplitude estimate can then be provided to the forward-backward algorithm again to obtain better estimates of the amplitude evolution of a reflection, including a further improved initial structure factor amplitude value. Hence the procedure is iterative.

The question then becomes *"When has the solution reached convergence?"*

Log Likelihood

One way to determine when the solution has converged is to determine the point at which the increase in the log likelihood is smaller than a given tolerance. The likelihood is a function describing how likely the data observed are to occur given the current model. For the

Kalman filter it is defined as (Cressie and Wikle, 2015)

$$L = P(F_0) \prod_i P(I_i|F_i) \times P(F_i|F_{i-1}), \tag{4.3.19}$$

where $F_i = |\boldsymbol{F}_i|$ is the structure factor amplitude of a reflection and I_i is the intensity of the reflection at time point i. Initially it seems that $P(I_i|F_i)$ and $P(F_i|F_{i-1})$ should represent the emission and transition probabilities respectively. However this is not exactly the case because the UKF propagates Gaussian models. Thus the mean and covariance of the states calculated at each time point in the HMM are used as parameters for the Gaussian distribution that is used as an approximation of the true crystal state (Figure 4.3). Hence $P(I_i|F_i)$ and $P(F_i|F_{i-1})$ are Gaussian distributions. The log likelihood is computationally more convenient to deal with (and often analytically too) than the likelihood, and hence it is the log likelihood that is calculated instead of the likelihood.

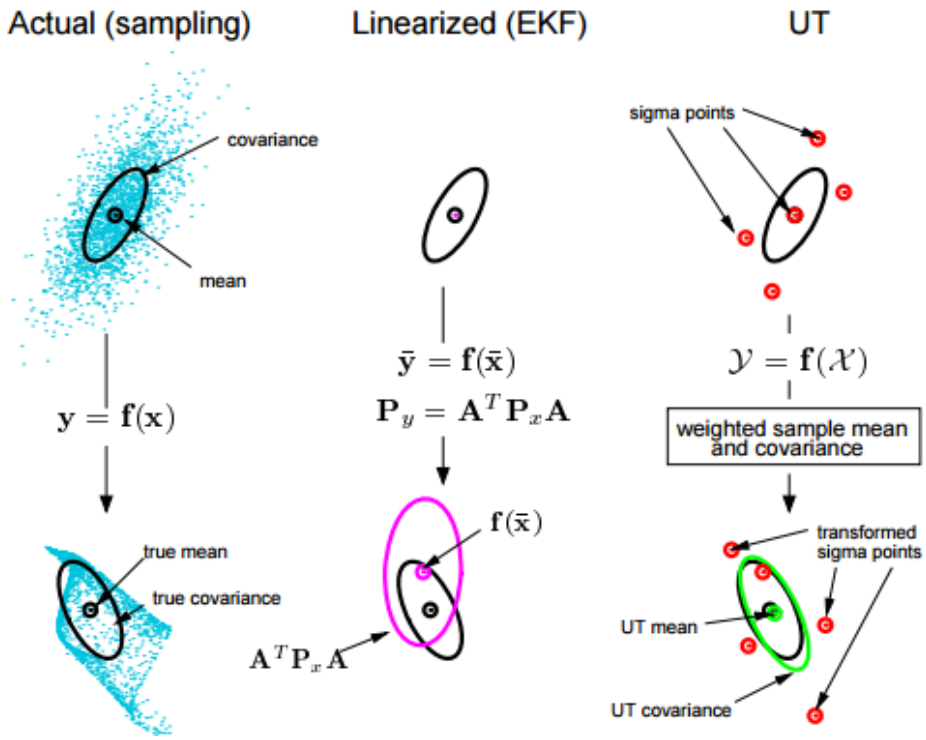


Figure 4.3: Propagation of states $\mathbf{x}$ through a non linear function $\mathbf{f}$ for different transformation techniques. The left plot shows the true mean and covariance propagation which uses Monte-Carlo sampling. The middle shows the linearisation performed by the extended Kalman filter (EKF), another non-linear Kalman filter. The right plot shows the propagation performed with the unscented transform (UT) used by the UKF. The performace of the UKF is superior to that of the EKF and uses many fewer sampling points (called sigma points) than Monte-Carlo sampling.

4.3.8 Summary of hidden Markov model formulation

In summary, in the current work the diffraction experiment has been represented as a hidden Markov model where a process function describes how the (hidden) state of the crystal at a particular time point i generates the consecutive crystal state due to X-ray irradiation. An observation model also describes how the crystal, which is in a particular state, generates the observed intensities. The UKF and the URTSS algorithms together give the forward-backward algorithm, which is designed to give the optimal sequence of crystal states that best describe the observed data, consistent with the defined process and observation functions. The process and observation (co)variances quantify the level of uncertainty of the corresponding processes. The forward-backward algorithm is applied iteratively to obtain the improved estimates of the structure factor amplitude of a reflection with each iteration. The log likelihood is calculated for each cycle, and when the improvement of the log likelihood is smaller than a given tolerance value, the solution is considered to have converged. The final result of the algorithm is a sequence of the set of structure factor amplitudes for each time point in the data collection experiment.

4.4 Extraction and treatment of reflection intensity data

4.4.1 Allocating observations to images

The algorithm presented was applied to the data that were collected on a crystal of bovine pancreatic insulin (Crystal ID 0259) as described in Chapter 2. However, before the forward-backward algorithm can be applied, the data for each full observation have to be extracted and allocated to a single diffraction image. An MTZ file containing the integrated data from the set of diffraction images was produced by processing the images with MOSFLM (Leslie and Powell, 2007). MTZDUMP, a program from the CCP4 software suite (Winn *et al.*, 2011), was used to extract the data from the integrated MTZ file, with additional commands to obtain the space group symmetry and image (batch) information. A custom script was written to parse the MTZDUMP output and extract the observation information. Each image is given a *rotation start angle* (RSA) and a *rotation end angle* (REA) and each observation has a *rotation centroid* value. An observation is allocated to an image if its centroid value lies between

the image's RSA and REA. For fully recorded reflections this is straight forward because the entire observation is recorded on a single image. This is not the case for partially observed reflections i.e. when a single reflection observation is partially measured on multiple images. Each detection of a partially measured reflection is given an estimate of the rotation centroid of the full observation. In theory this centroid value should be the same for each measurement of the same reflection, but in practice this is rarely the case. To determine the actual centroid value, the mean average of all centroid values of each measurement is calculated, and this value is used to allocate a partial reflection observation to a given image.

At the beginning and end of a data collection experiment, some reflections have not been completely traversed, resulting in observations where the calculated rotation centroid lies outside the oscillation range of the data collection. In these cases the reflection is allocated to either the first or last image if the calculated centroid is smaller than the RSA of the first image or the REA of the last image respectively.

4.4.2 Treatment of intensity data

Extracting intensity values for fully traversed reflections

When reflection observations are integrated with MOSFLM, two intensity estimates are calculated: a profile fitted intensity and a summed intensity. The profile fitted intensity value is generally a better estimate and this is especially true for weak data. On the other hand, the summed estimate should be more accurate for the strongest reflections (`http://www.ccp4.ac.uk/html/aimless.html`). In exactly the same manner as utilised in AIMLESS (Evans and Murshudov, 2013), the custom written parser can use either of the two intensity estimates but defaults to using a combination of the two. The approach of combining the two is to calculate a weighted average of the two intensity estimates such that the profile fitted estimate is weighted higher for weak reflections and vice versa for strong reflections. The equation used to extract the combined intensity, I_{com} is

$$I_{com} = wI_{pr} + (1 - w)I_{sum}, \tag{4.4.1}$$

where I_{pr} is the profile fitted intensity estimate, I_{sum} is the summed intensity estimate and w is the weight defined as

$$w = \frac{1}{1 + \left(\frac{I_{raw}}{I_{mid}}\right)^{I_{pow}}}. \tag{4.4.2}$$

In AIMLESS, I_{pow} defaults to 3, I_{raw} is the intensity value before Lorentz-Polarisation (LP) correction and I_{mid} is optimised to give the best overall R_{meas} value. The custom parser uses $I_{mid} = (I_{pr} + I_{sum})/2$, $I_{raw} = I_{mid} \times$LP and $I_{pow} = 3$.

The intensity value (either I_{com}, I_{pr} or I_{sum}) is calculated for each measurement of a reflection observation. For full reflections the resulting intensity value is used as the full observation intensity. For partial reflections this value has to be summed for each measurement of the same reflection observation to obtain the full estimate.

Estimating the intensity of non-fully traversed reflections

Some reflection observations are not fully traversed and hence the intensity values for these reflections are incomplete. Estimates of the true intensity of these reflections can be made by assuming a standard uniform shape for the reflection. If the reflection is assumed spherical in shape then the measured fraction of the reflection is a spherical cap, shown in Figure 4.4. The ratio of the volume of the spherical cap to the volume of the sphere, p is given by

$$p = \frac{h^2}{d^3}(3d - 2h), \tag{4.4.3}$$

where d is the diameter of the sphere and h is the height of the spherical cap. If the height of the spherical cap is given as a fraction of the diameter, denoted q, and the diameter is set to 1, then equation 4.4.3 becomes

$$p = 3q^2 - 2q^3, \tag{4.4.4}$$

which is the same formula as given in Rossman *et al.* (1979). Angles are used as a proxy for the actual lengths h and d because the lengths are not given. For reflections where the calculated centroid is before the RSA of the first image, h is approximated as the absolute difference between the RSA of the first image and the mid point of the RSA and REA of the last image on which the reflection was observed. The spherical diameter of the reflection,

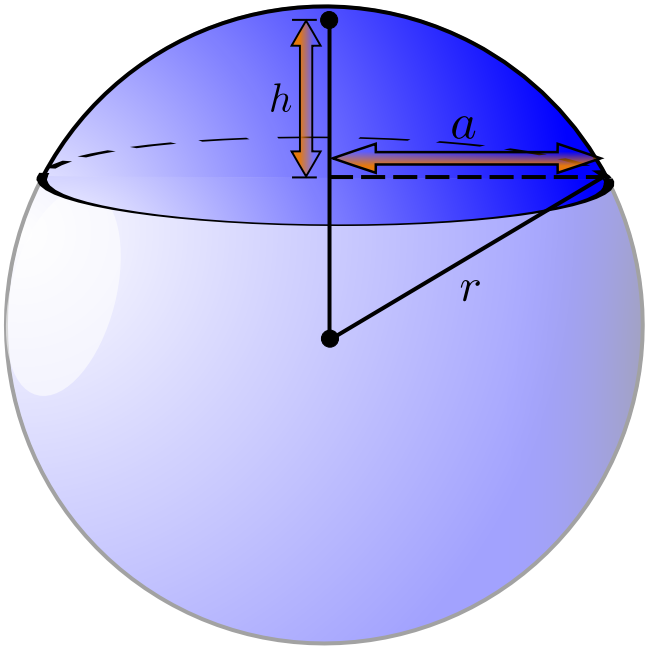


Figure 4.4: Model of the spherical cap traversed in a data collection experiment. The translucent volume is the volume that was not recorded. h represents the height of the spherical cap, a represents the radius of the circular base of the cap and r is the radius of the sphere.

d is approximated as twice the absolute difference between the mid point of the RSA and REA of the last image on which the reflection was observed and the rotation centroid. The analogous values are used for reflections where the reflection centroids were calculated beyond the REA of the last image.

Quantifying additional uncertainty in the observation variance

The standard deviation for each measurement is also calculated and provided in the output by MOSFLM. Again, for full reflections this value can be used as the final standard deviation for each measurement. However, for partial reflections the standard deviations for each partial measurement have to be combined. This is achieved by summing the variances for each partial measurement giving a total variance denoted σ^2_{sum}.

However, two more factors complicate the variance calculation. Firstly, it is acknowledged that the crystal is in a slightly different dose state at each image: hence in combining the variances it is also necessary to increase the uncertainty due to the change in crystal state between images. This additive uncertainty factor is calculated as

$$\sigma^2_{im} = \sum_{i}^{\text{images}} (1 - (\boldsymbol{D}_i^{im})^2) I_i^{im}, \tag{4.4.5}$$

where the sum is over all images on which the measurements of an observation are recorded, I_i^{im} is the measured intensity of the observation a image i and $\boldsymbol{D}_i^{im}$ is defined as

$$\boldsymbol{D}_i^{im} = \exp(-2|\Delta B_i^{diff}|\sin^2(\theta)/\lambda^2)), \tag{4.4.6}$$

where ΔB_i^{diff} is the difference in B factor between image i and the centroid image. The explicit calculation and results of the B factor analysis is given in sections 4.6.1 and 4.6.2.

Secondly, the total fraction of each measurement is calculated for each reflection (denoted `FRACTIONCALC` in the MTZ column from MOSFLM). The sum of these values for partial measurements of an individual observation should be equal to 1. This is rarely the case because these values are not accurate. In AIMLESS the criteria for an observation to be regarded as fully recorded over its partial measurements is if the sum of the `FRACTIONCALC` is bewteen 0.95 and 1.05. The criterion used by the custom parser script only flags observations where the sum is less than 0.95. If this is the case then the observations can either be rejected, or the variance of the intensity value can be further inflated by a value proportional to the inverse of the total calculated fraction. Explicitly the additive factor is calculated as

$$\sigma_{fr}^2 = \varepsilon \times \Sigma \times (1 - fr_{tot}), \tag{4.4.7}$$

where ε is multiplicity of the reflection, Σ is the expected intensity value and fr_{tot} is the sum of the individual calculated fractions for each partial measurement.

The final variance for a single observation is then given by

$$\sigma^2 = \sigma_{sum}^2 + \sigma_{im}^2 + \sigma_{fr}^2. \tag{4.4.8}$$

4.5 Simulation results

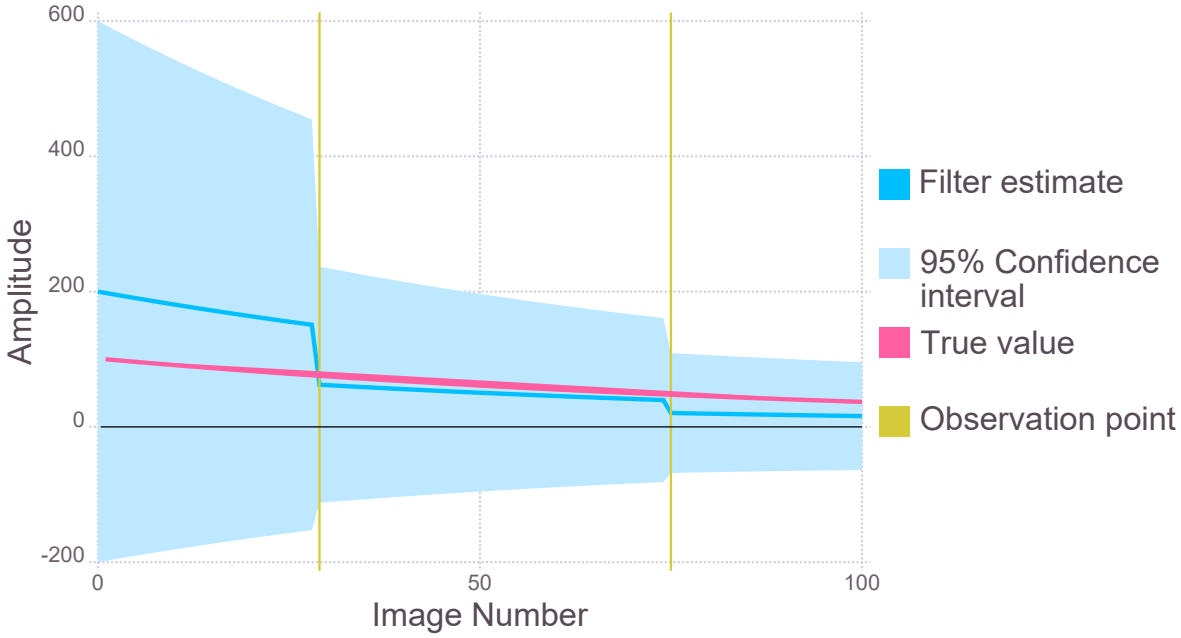
A simulation of the behaviour of a reflection in a diffraction experiment was performed to test the performance of the forward-backward algorithm. In the simulation, 100 images were recorded in a diffraction experiment where the intensity of a particular reflection was observed twice, once on the 27th image and again on the 76th image with simulated Gaussian noise. The observed intensities were 71.43 and 40.13 on images 27 and 76 respectively.

The true structure factor amplitude of the reflection is initially 100 and it decays by 1% after each image is collected (whether it is observed or not). The forward-backward algorithm is applied where the process function is defined such that the amplitude decays by 1% for each image and the observation function is defined as in equation 4.3.13. The estimate supplied to the forward-backward algorithm for the initial amplitude is 200, double the value of the true value of 100. The results of the forward-backward algorithm for cycles 1, 2 and 10 are shown in Figure 4.5.

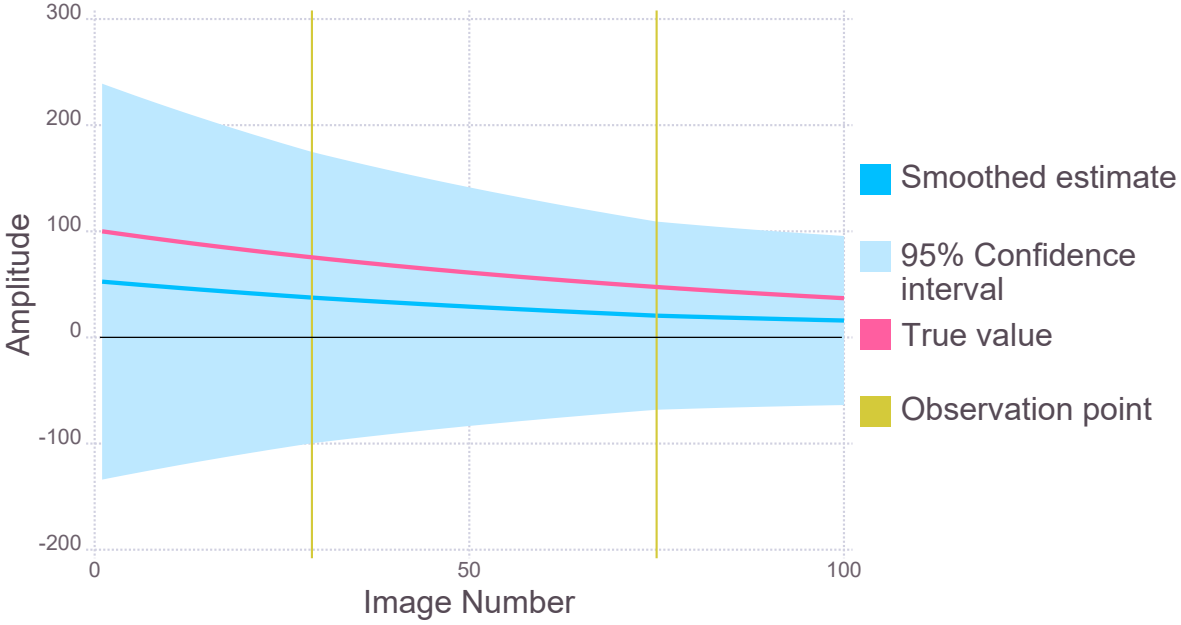
At the very beginning of the algorithm, the filtering estimates (solid blue line Figure 4.5a) do not predict the true value (solid pink line) very well. This is because no observations are made until the 27th image, therefore the estimates propagate according to the defined process function. At image 27 the first observation is made, represented by the vertical solid gold line, which is when the filtering estimate approaches the amplitude value required to produce the observed intensity according to the observation equation 4.3.13. The forward pass continues to propagate the estimates according to the process function until it reaches the second observation and sharply changes value to what it deems optimum. The smoothing algorithm attempts to consolidate the optimal values found during the forwards pass whilst maximising the probability of the estimate of the state at time i producing the state at time i + 1. This results in smoother state predictions, as evidenced by the smoother blue solid line in Figure 4.5b, and a reduced overall uncertainty (the light blue shaded region represents the 95% confidence interval). The initial value found during the smoothing is also closer to the true value. The second forward-backward pass shows the same characteristics, whereby the smoother gives a smaller uncertainty estimate and smoothes the values from the filtering pass (Figures 4.5c and 4.5d). After 10 forward-backward cycles the smoothed estimates are very close to the true values (Figure 4.5f). Importantly the uncertainty values are smallest at the points where the observations are made, as expected. Conversely, the uncertainty is largest at the points furthest away from the observations. This is because the uncertainty in the amplitude values propagates when the data are not observed.

The log likelihoods for the 10 cycles, calculated using the natural logarithm of equation 4.3.19, were also computed and are shown in Figure 4.6.

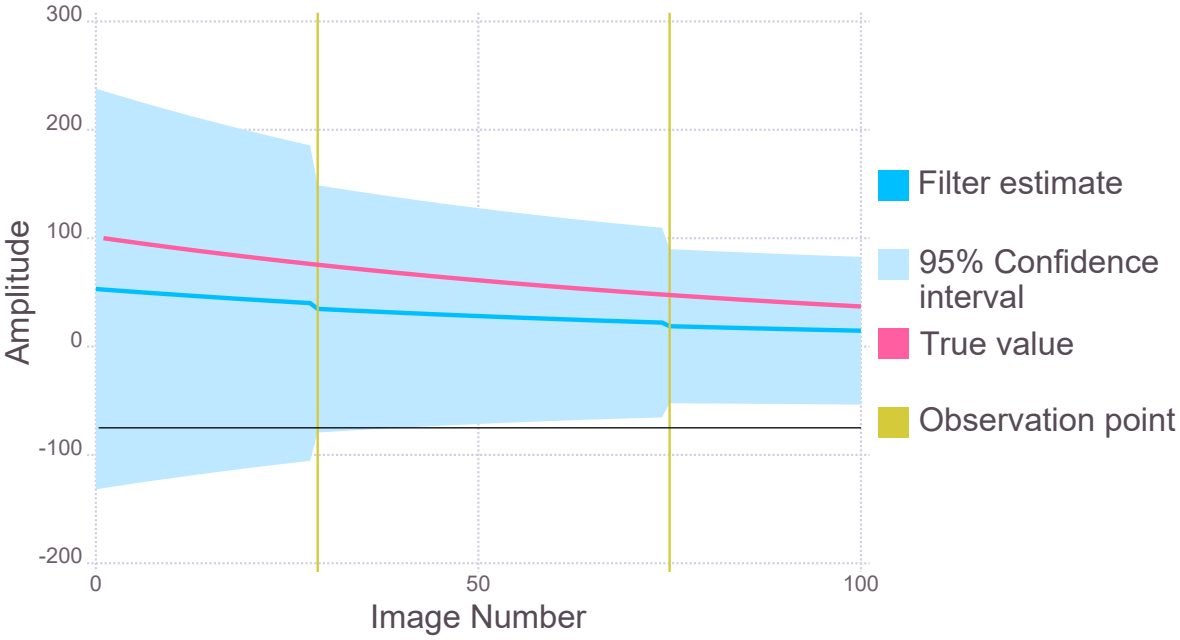
It can be seen that as the number of forward-backward passes increases, the rate of increase (gradient) tends to zero thus showing that the smoothed estimates are converging to the



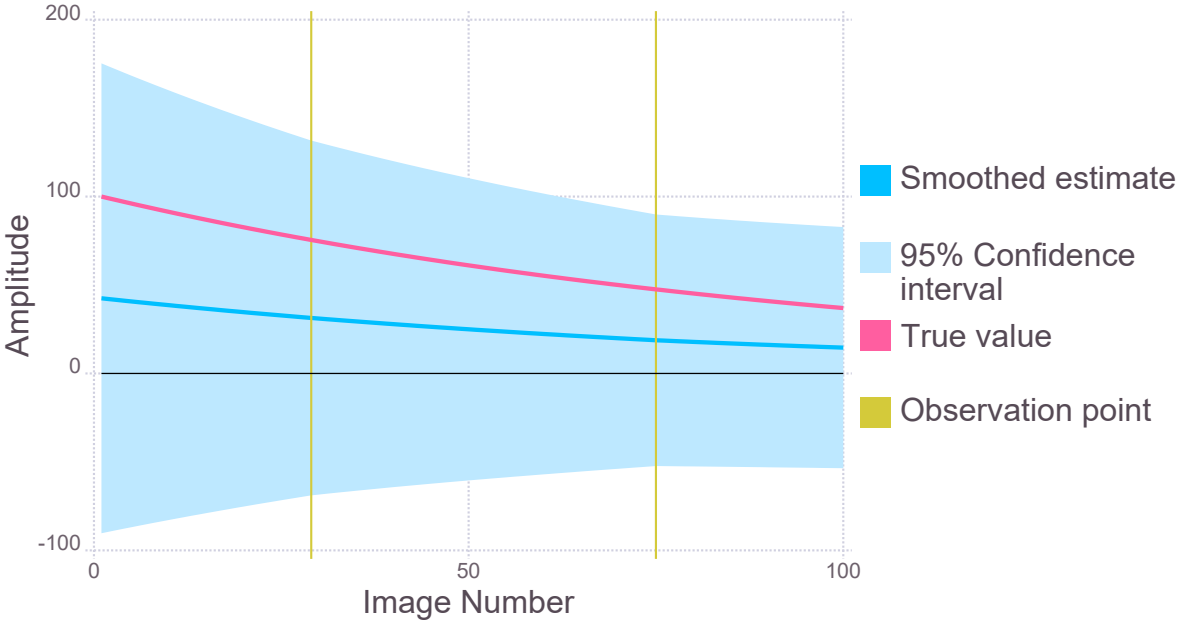
(a) Cycle 1: forward pass.



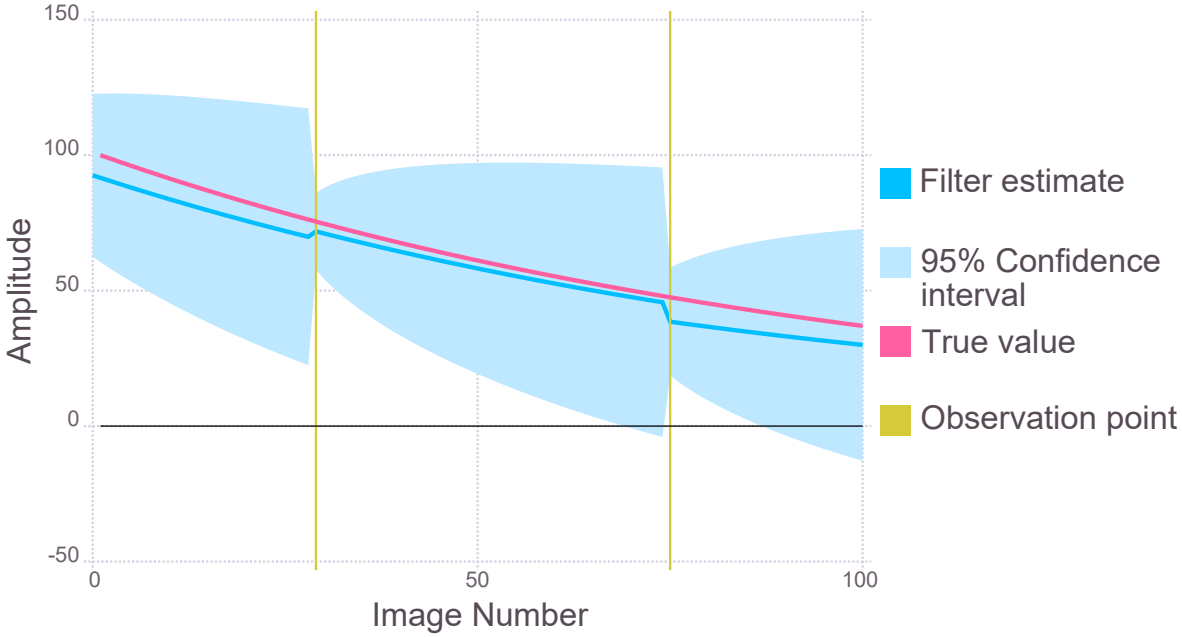
(b) Cycle 1: backward pass.



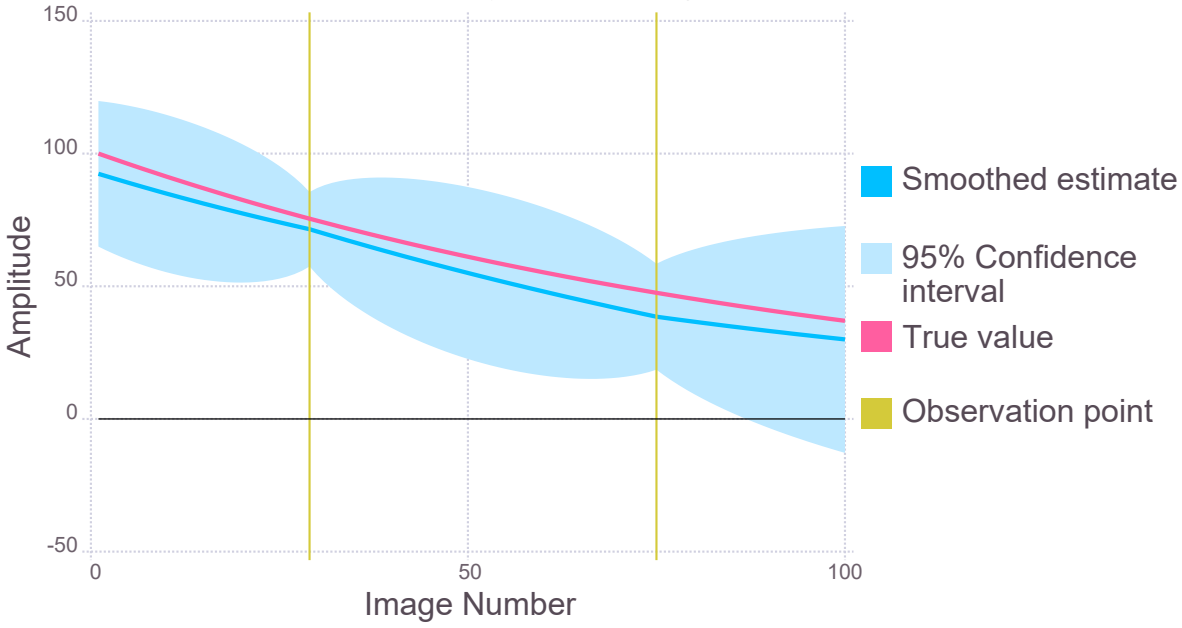
(c) Cycle 2: forward pass.



(d) Cycle 2: backward pass.



(e) Cycle 10: forward pass.



(f) Cycle 10: backward pass.

Figure 4.5: Forward-backward algorithm results for a simulated reflection observed on image 27 and 76 (solid gold lines) of a dataset consisting of 100 images. As the number of cycles increases, the forward-backward estimate (blue solid line) approaches the true value (pink solid line) and the 95% confidence interval decreases (light blue shaded region).

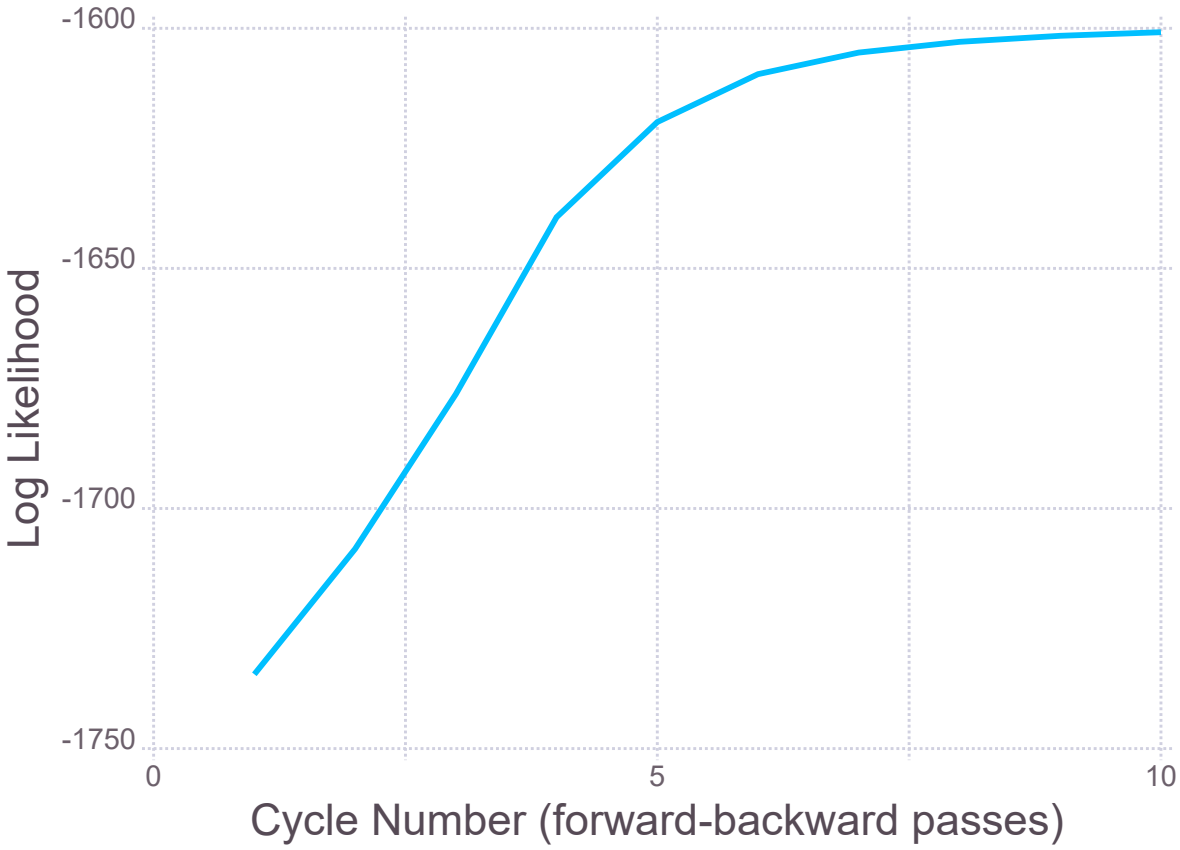
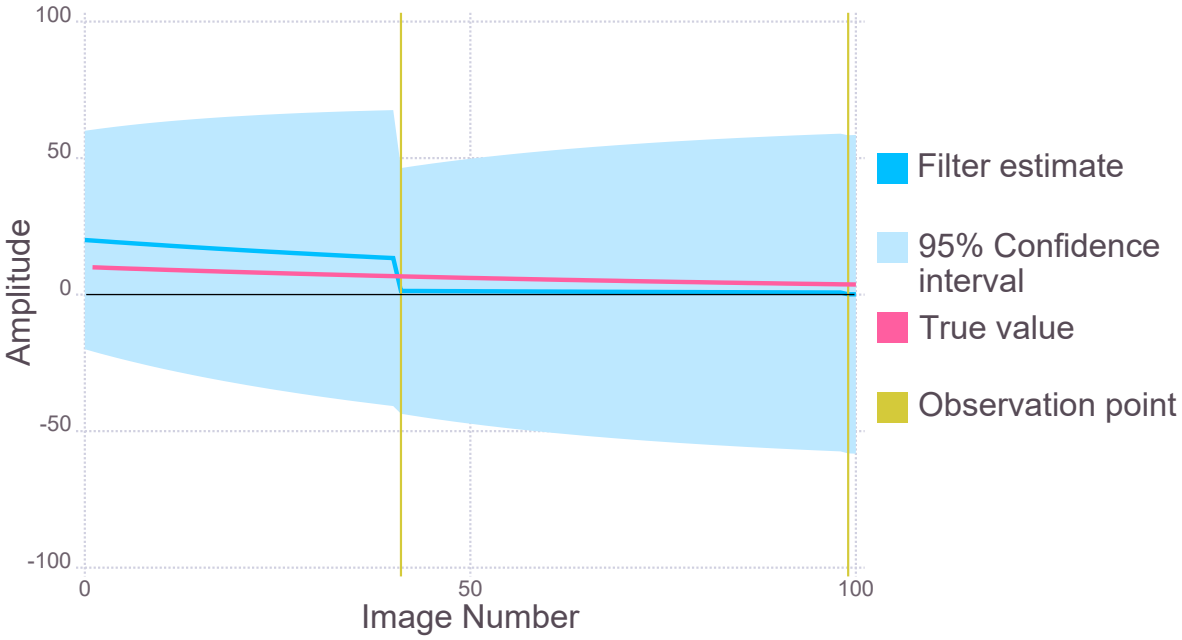


Figure 4.6: Log likelihood values calculated for each smoothing pass. As the cycle number increases the log likelihood starts to plateau meaning that the forward-backward solution is converging on a final solution.

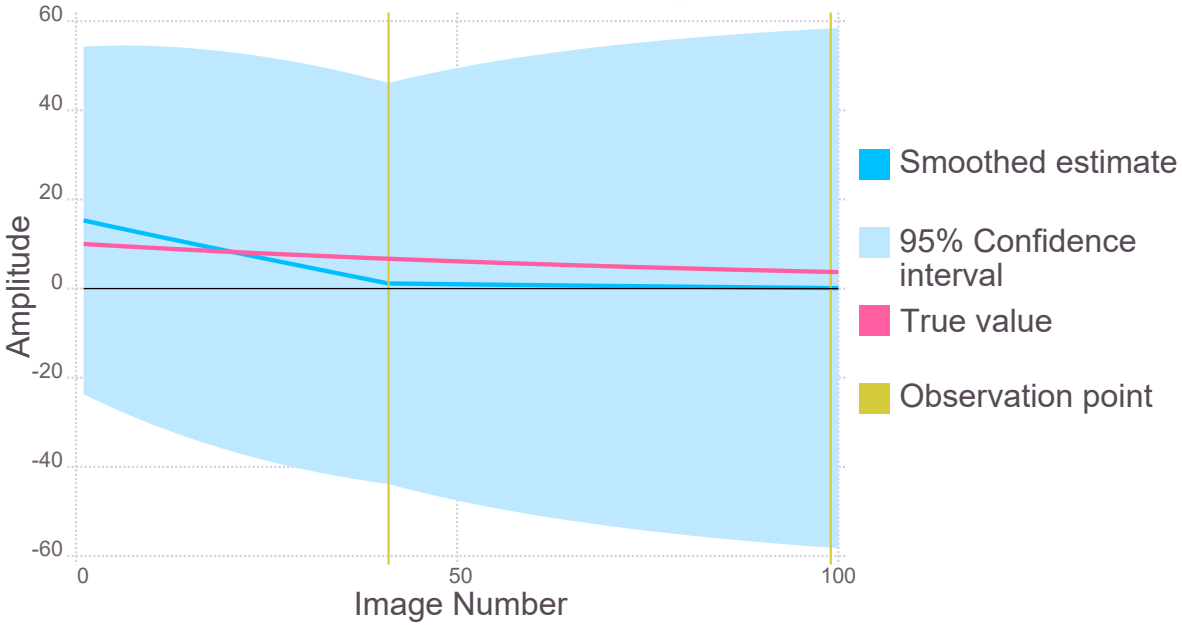
optimal solution for the observed data.

4.5.1 Weak data

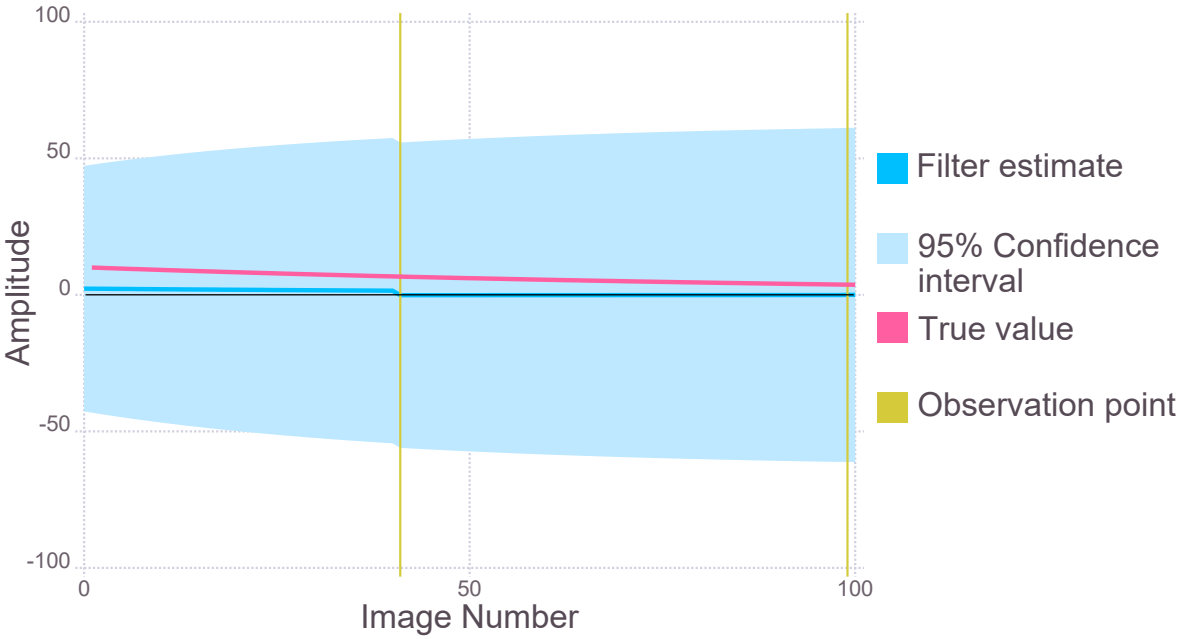
Simulations of the forward-backward algorithm showed that the method works very well for strong reflections. A further simulation was performed to see how effective the algorithm would be for weak reflections. Again the simulation consisted of 100 images where the intensity of a particular reflection was observed twice, this time on the 41st image and the 99th image with simulated Gaussian noise. The true structure factor amplitude of this reflection is initially 10 and again decays by 1% after each image is collected (whether it is observed or not). The same process and observation functions are defined but the estimate supplied to the forward-backward algorithm for the initial amplitude is 20. Additive zero-mean Gaussian noise was applied to the observation model to obtain observed intensities of -1.92 and 2.91 on images 41 and 99 respectively. The results of the forward-backward algorithm for cycles 1 and 8 are shown in Figure 4.7.



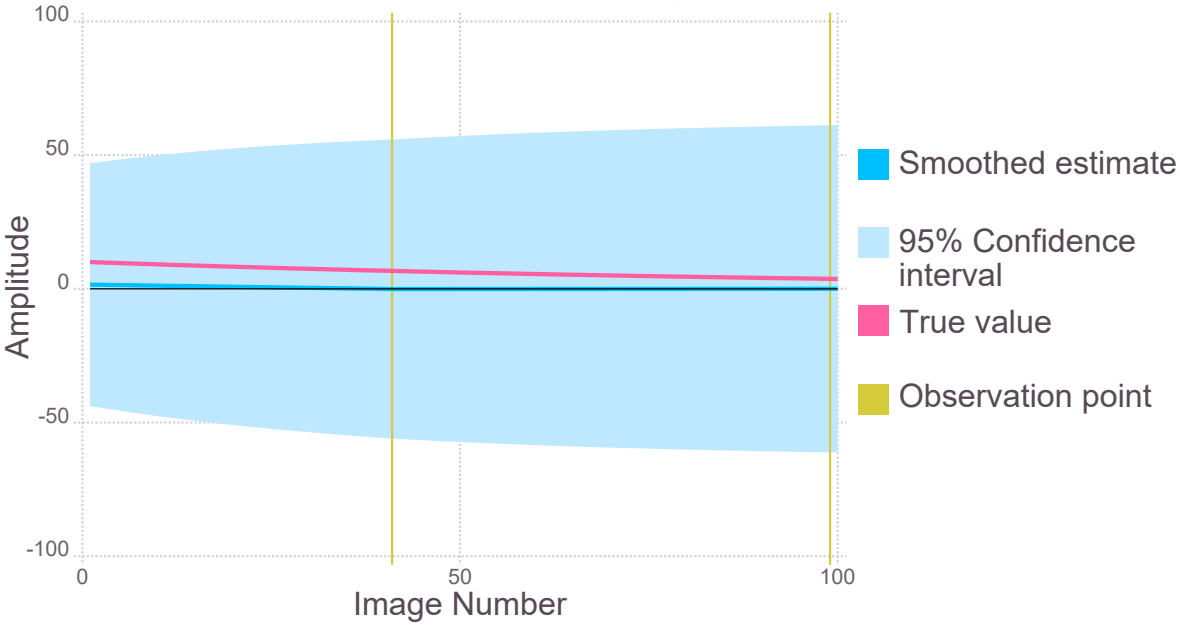
(a) Cycle 1: forward pass.



(b) Cycle 1: backward pass.



(c) Cycle 8: forward pass.



(d) Cycle 8: backward pass.

Figure 4.7: Forward-backward algorithm results for a simulated weak reflection observed on image 41 and 99 (solid gold lines) of a dataset consisting of 100 images. The forward-backward pass for cycle 1 looks like it may lead to a good estimate of the true value. However, by cycle 8 the estimates have converged to zero and the 95% confidence interval has not improved.

The first cycle of the forward-backward algorithm looks promising as the estimate of the initial amplitude is slightly closer to the true value (Figure 4.7b). However, at cycle 8 (Figure 4.7d) it is clear that the forward-backward algorithm is converging to a point where every amplitude estimate is zero, which is not representative of the true amplitude. The log likelihood confirms that the estimates are getting worse as the values decrease with the cycle number (Figure 4.8).

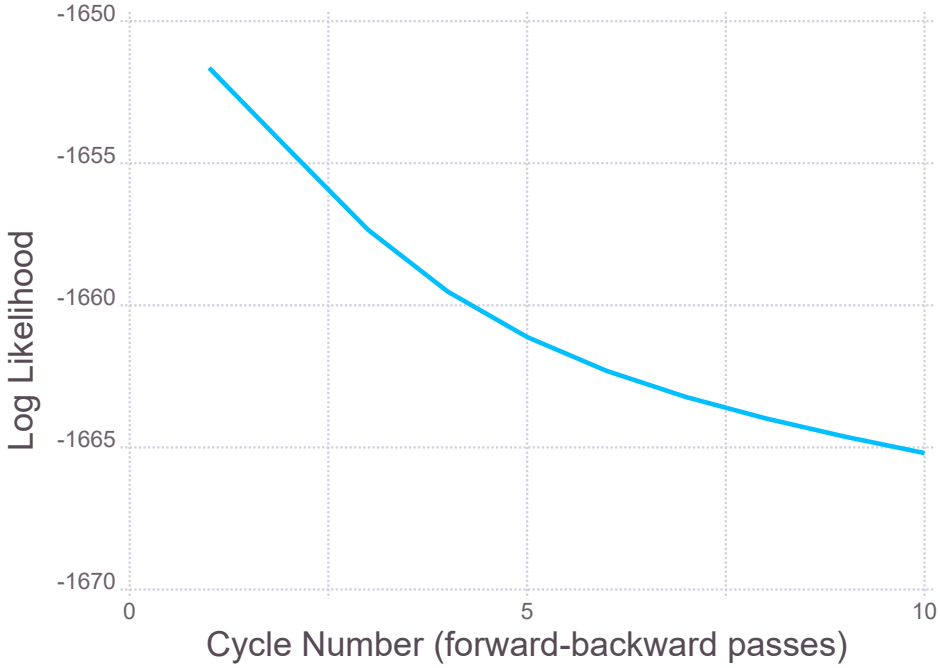


Figure 4.8: Log likelihood values calculated for each smoothing pass. Visually the forward-backward passes seemed to give worse estimates as the number of cycles increased (Figure 4.7). The log likelihood values confirm this as evidenced by the decrease in the values.

To circumvent the problems caused by using the forward-backward algorithm on weak reflections, Bayesian inference is performed on the initial amplitude estimate at the end of each cycle. In particular the expected value of the posterior distribution is used as the initial amplitude estimate. The posterior distribution of interest is

$$P(F_c|F_0) = \frac{P(F_0|F_c) \times P(F_c)}{P(F_0)}, \tag{4.5.1}$$

where $P(F_0|F_c)$ is defined for centric and acentric reflections as

$$P_a(F_0|F_c) = \frac{2F_0}{\varepsilon\sigma_0^2} \exp\left(-\frac{F_0^2 + \boldsymbol{D}^2 F_c^2}{\varepsilon\sigma_0^2}\right) I_0\left(\frac{2F_0 \boldsymbol{D} F_c}{\varepsilon\sigma_0^2}\right), \tag{4.5.2}$$

$$P_c(F_0|F_c) = \left[\frac{2}{\pi\varepsilon\sigma_0^2}\right]^{1/2} \exp\left(-\frac{F_0^2 + \boldsymbol{D}^2 F_c^2}{2\varepsilon\sigma_0^2}\right) \cosh\left(\frac{F_0 \boldsymbol{D} F_c}{\varepsilon\sigma_0^2}\right), \tag{4.5.3}$$

and $P(F_c)$ is also defined for centric and acentric reflections as

$$P_a(F_c) = \frac{2F_c}{\Sigma^2} \exp\left(-\frac{F_c^2}{\Sigma^2}\right), \tag{4.5.4}$$

$$P_c(F_c) = \sqrt{\frac{2}{\pi\Sigma^2}} \exp\left(-\frac{F_c^2}{2\Sigma^2}\right), \tag{4.5.5}$$

where $P(F_c)$ is the Wilson distribution for amplitudes, F_0 and σ_0^2 are the initial amplitude and variance estimates from the forward-backward cycle, and F_c is the amplitude to be calculated. As discussed in section 3.3.2, the denominator of equation 4.5.1 can be given as:

$$P(F_0) = \int_0^\infty P(F_0|F_c) \times P(F_c)\,\mathrm{d}F_c. \tag{4.5.6}$$

This procedure ensures that the initial amplitude value for the next forwards pass is positive. If the final value for the amplitude at the end of a forwards pass is negative, then that value is set to zero for the smoother. Systematic testing is required to determine whether this is a robust solution to the problem

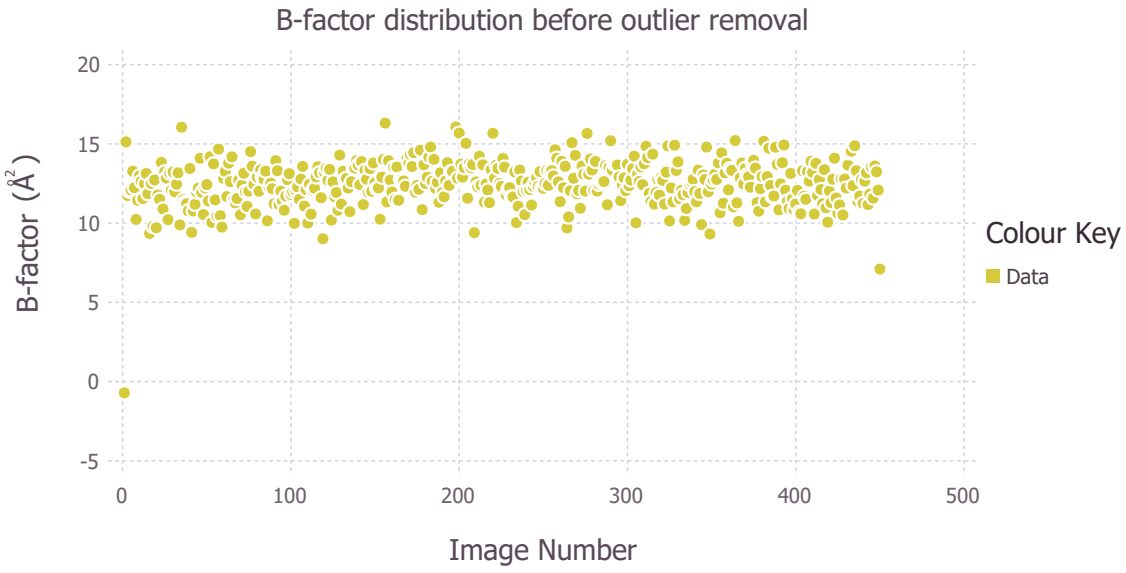
4.6 Protein structure results

4.6.1 Bovine pancreatic insulin

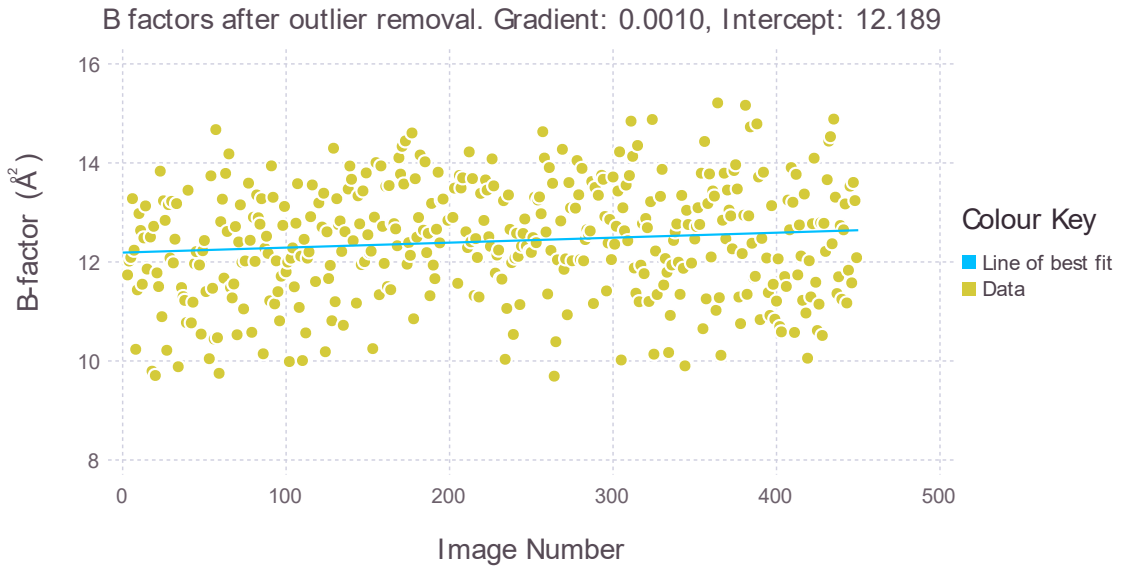
Scale and B factors

Data were collected on a crystal of bovine pancreatic insulin (crystal ID 0259) as described in chapter 2. The atomic composition used to provide expected intensity values, was obtained from the insulin structure with PDB code 2BN3. The B factors could then be calculated for each image according to equation 4.3.18 and are shown in Figure 4.9. It can be seen that there are a couple of points that may be regarded as outliers in Figure 4.9a. To remove the outliers, the mean and standard deviation of the B factors were calculated and any B factor that was more than two standard deviations from the mean was removed. The resulting B factor distribution is plotted in Figure 4.9b.

To ensure that the B-factors exhibited linear behaviour, "damage corrected" B factors, B_{dc}, were calculated by rearranging the linear formula for the B factor increase. If the B-factors



(a)



(b)

Figure 4.9: Calculated B factors for each image in the insulin dataset. (a) Distribution before outlier removal. (b) Distribution after outlier removal. The line of best fit (blue solid line) with gradient, ΔB = 0.001 Å^2 and intercept = 12.189 Å^2, is overlaid on the data. Note the differing y axis scales.

do show linear behaviour then the “damage corrected” B factors should be normally distributed, so the goal is to check for this. The “damage corrected” B factors are determined as

$$B_{dc}^{i} = B_i - \Delta B \times i, \tag{4.6.1}$$

where B_i is the B factor calculated at image i, and ΔB is the gradient of the line fitted to the data. A histogram of the "damage corrected" B factor distribution was then plotted, which should be a Gaussian distribution centred on the intercept of the line in Figure 4.9b. Additionally a QQ plot† was used to ensure that the data were normally distributed (Figure 4.10b). The gradient ΔB was calculated to be 0.001 Å^2.

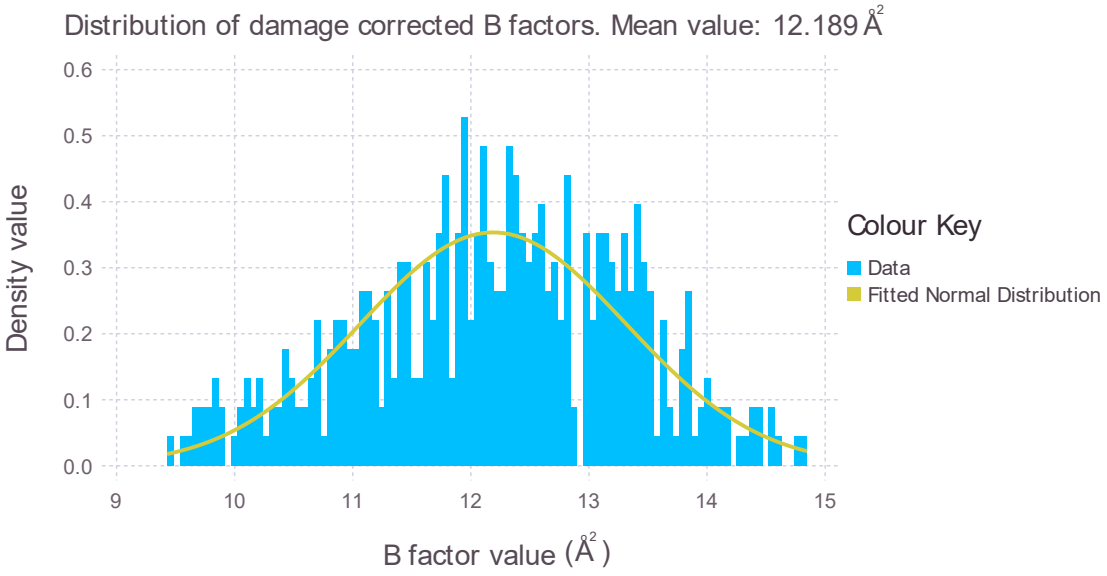
The set of scale factors calculated from the images, $\{s_{images}\}$, are shown in Figure 4.11a. There did not appear to be any outliers from visual inspection, so there was no outlier rejection method performed for the scale factors. The scale factors that corresponded to images that were removed from the B factor outlier rejection analysis were also omitted for consistency. The resulting scale factors, $\{s^*_{images}\}$, are shown in Figure 4.11b.

Figure 4.12 presents the distribution of scale factors as a histogram with a kernel density estimation of the distribution overlaid. The fact that the mode and mean are very similar in value also suggests that the scale factor distribution is Gaussian. However there is no guarantee that the scale factor distribution for a general experiment will be Gaussian and hence there are no checks for normality included in the algorithm for the scale factors. The mean scale factor value of the distribution, s_{mean} = 0.010 was used in the forward-backward algorithm, which is assumed to be constant throughout the experiment. This assumption is not true in the general case, but it should be suitable for the diffraction experiment concerned, for which the insulin crystal was completely immersed in a tophat beam for a 45° rotation. This is because the transmission through the crystal and the illuminated volume should remain constant.

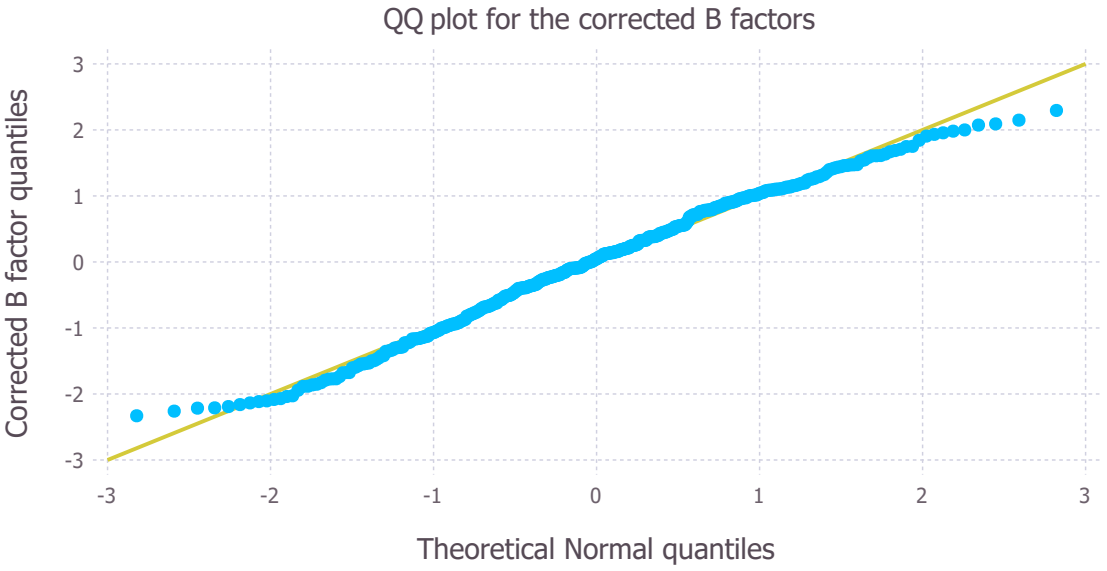
Forward-backward algorithm

The forward-backward algorithm was applied with the following parameters defined:

† A QQ plot graphically determines whether two datasets come from the same distribution. It plots the quantiles of the first dataset against the quantiles of the second. If the two datasets come from the same distribution the points should fall on a 45° reference line which is typically also plotted.

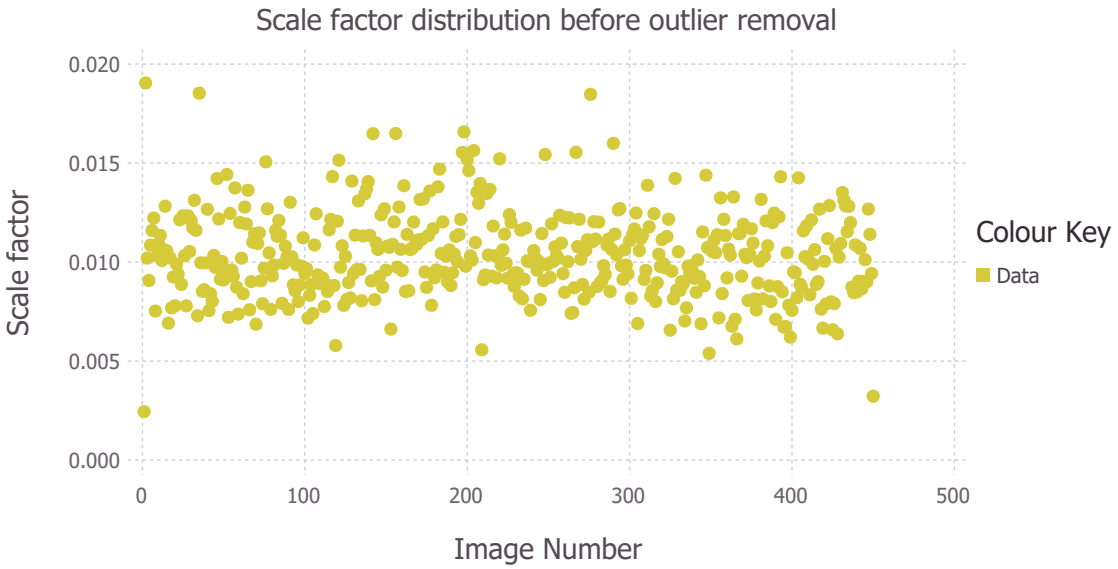


(a)

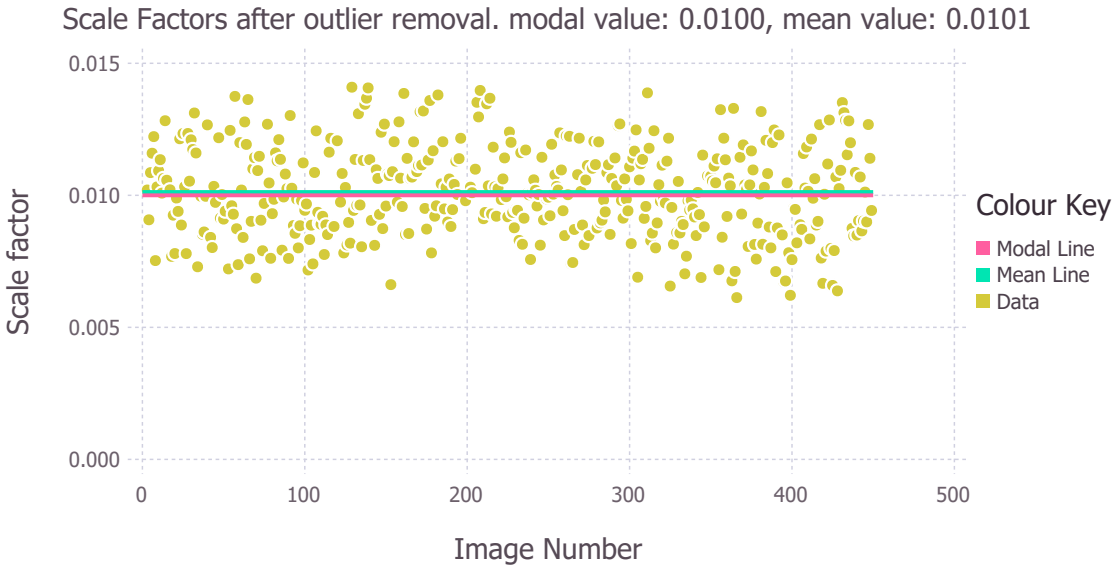


(b)

Figure 4.10: (a) Histogram of damage corrected B factors. The Gaussian shape suggests that a linear assumption for the behaviour of B factors is suitable. (b) QQ plot for the damage corrected B factors. The linearity of the points confirm that the distribution is actually Gaussian.



(a)



(b)

Figure 4.11: Calculated scale factors for each image in the insulin dataset. (a) Distribution before outlier removal. (b) Distribution after outlier removal. The solid green and solid pink lines represent the mean and mode of the distribution respectively. The fact that the mean and mode are close in value suggest that the distribution is Gaussian.

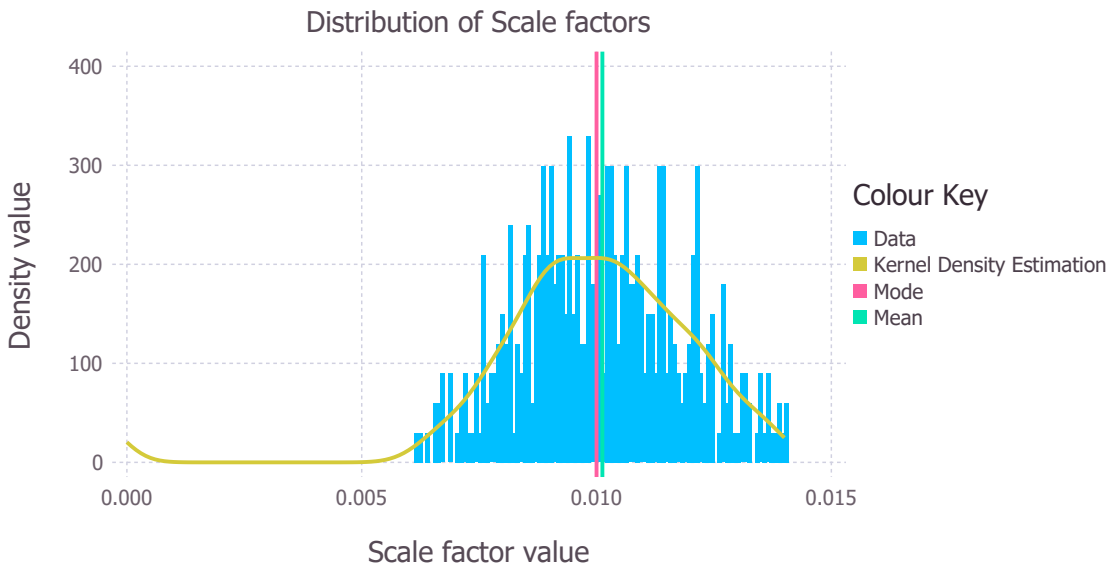


Figure 4.12: Histogram of scale factors with the mean (solid green line), mode (solid pink line) and kernel density estimation (solid gold line) overlaid.

- the minimum and maximum number of forward-backward cycles for an individual reflection was 5 and 200 respectively.
- the forward-backward cycles were regarded to have converged if:
  1. the maximum number of cycles had been reached (200)
  2. the absolute change in log likelihood between consecutive cycles was less than 0.1
  3. the change in initial amplitude value between consecutive cycles was less than 1
- reflections for which the Rician distribution was used for the amplitude process function (equations 4.3.8 and 4.3.9) were defined as reflections where $F_0/\sigma(F_0) < 3$, where F_0 is the initial amplitude estimate.

Amplitude estimates resulting from applying the forward-backward algorithm for 4 reflections are shown in Figure 4.13. It can be seen that for these reflections, the initial amplitude value estimated using CTRUNCATE is within the 95% confidence region as estimated by the forward-backward algorithm. In fact the median percentage error between the amplitude estimates for all reflections from CTRUNCATE and the forward-backward algorithm is 3.77%. This suggests that the constant scale factor assumption (outlined in section 4.3.6 and graphically depicted in Figure 4.11b) used for the simple scaling procedure used for this study is

valid for this experiment.

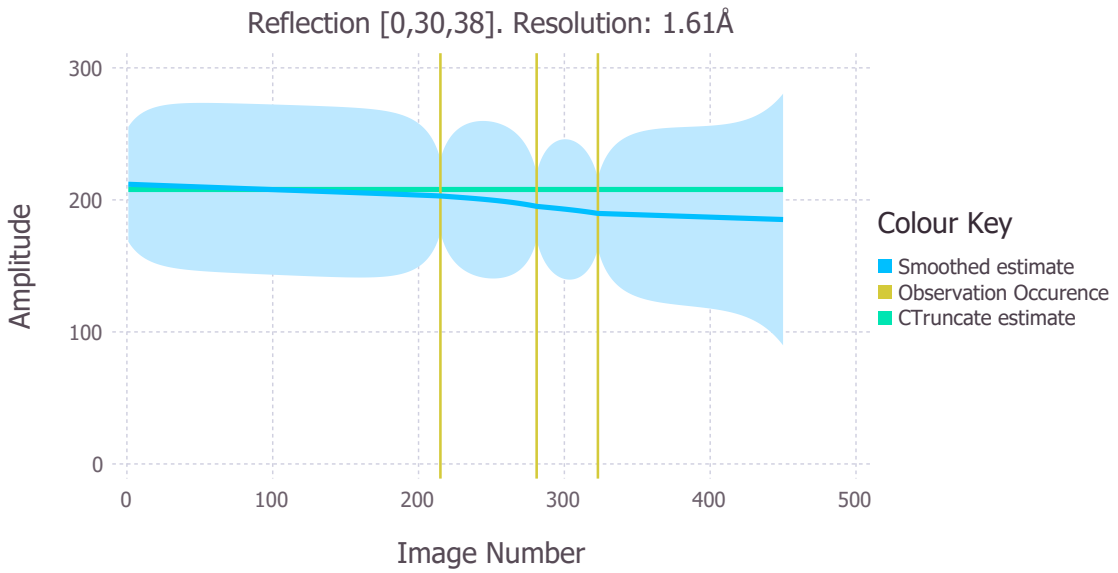
Figures 4.13a and 4.13b exhibit the expected behaviour of an average refection, since both of these reflections decay relatively smoothly as exposure time increases. Figure 4.13c shows a reflection whose amplitude increases as the exposure increases. This demonstrates that the forward-backward algorithm is capable of capturing the different behaviours of various reflections despite the process function describing a monotonic decay of the amplitudes.

Figure 4.13d shows a reflection where the behaviour is very irregular and not very smooth. In this case it is very likely not to be caused by a physical phenomenon and is probably due to incorrect scaling of the data. To prevent this problem, restraints should be imposed during the scaling procedure in a similar manner to those of existing scaling methods to ensure sufficient smoothness of reflection amplitudes (Evans and Murshudov, 2013; Kabsch, 2010a). It should be noted that not all sharp changes in amplitudes relate to noise/incorrect processing. Mechanical or chemical changes of the structure can occur within a few seconds (Allan *et al.*, 2012).

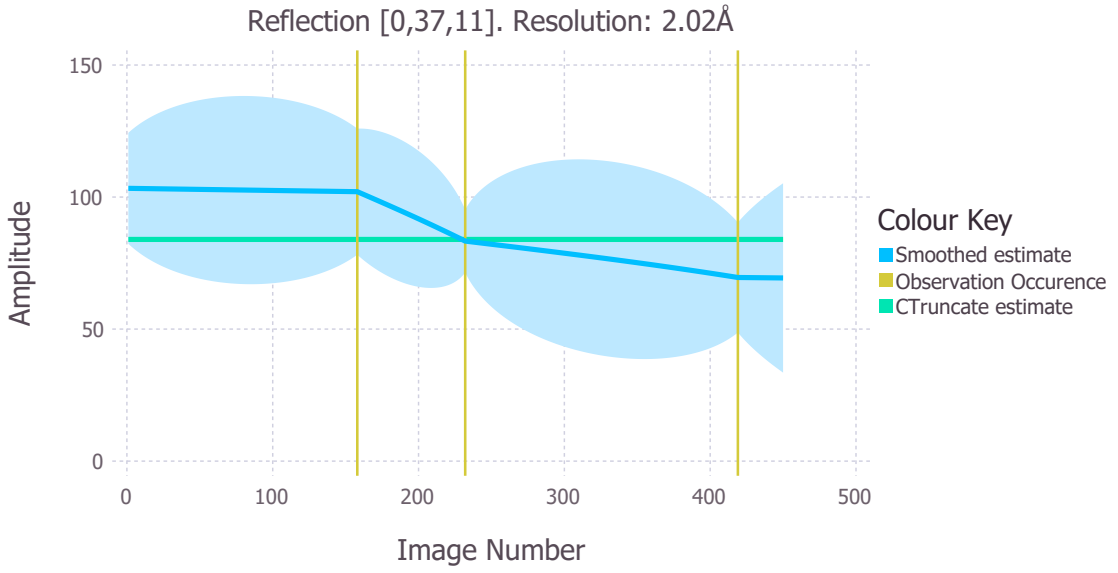
Another undesirable feature of the forward-backward algorithm is the fact that the initial amplitude estimate is solely influenced by the amplitude estimate at the point where the first observation is made. This effect is prominent in Figures 4.13b and 4.13c. Thus the initial amplitude estimate is not influenced by the multiplicity and hence erroneous estimates are more likely to arise if the first observation of a reflection is an outlier. This can be overcome by performing the forward-backward algorithm on each observation separately and merging the entire amplitude curves for equivalent reflections. The relative uncertainties at each point in the data collection experiment will be different because the observations will be collected on different images. This should ensure that the (possibly non-linear) behaviour of the reflection should still be captured by this method.

Refinement results

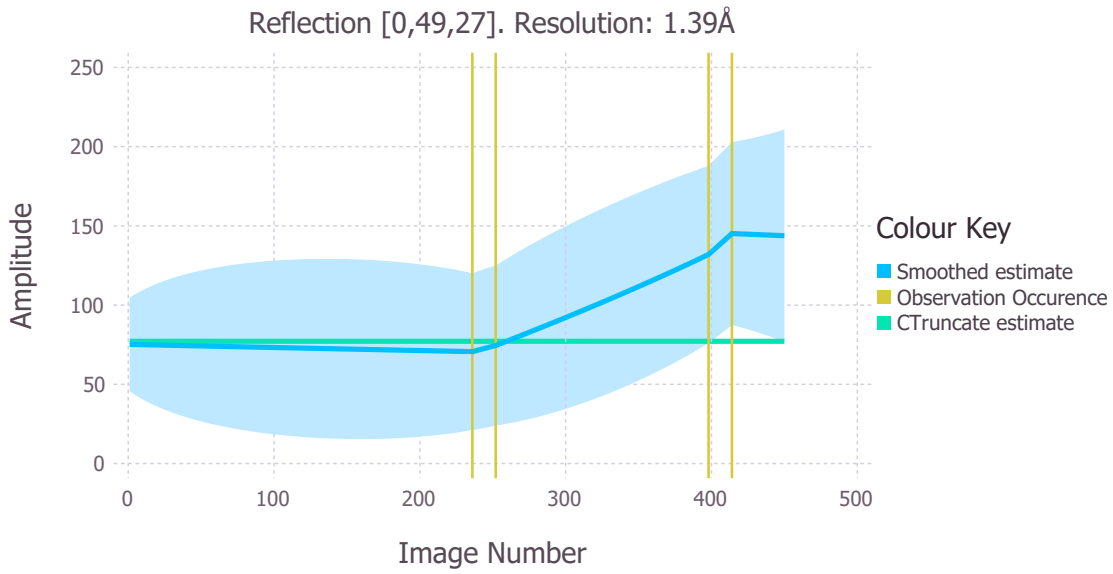
To obtain an electron density map of the structure the initial amplitude estimates resulting from the forward-backward algorithm (FBA) were used along with the deposited insulin model (PDB code 2BN3) and refined with REFMAC (Murshudov *et al.*, 2011) (10 cycles of rigid body refinement followed by 10 cycles of restrained refinement). The same refine-


Reflection [0,30,38]. Resolution: 1.61Å
Amplitude
300
200
100
0
0
100
200
300
400
500
Image Number
Colour Key
Smoothed estimate
Observation Occurence
CTruncate estimate

(a)


Reflection [0,37,11]. Resolution: 2.02Å
Amplitude
150
100
50
0
0
100
200
300
400
500
Image Number
Colour Key
Smoothed estimate
Observation Occurence
CTruncate estimate

(b)



(c)



(d)

Figure 4.13: Amplitude estimates for four different reflections observed in the insulin dataset using the forward-backward algorithm (blue solid line). The estimate produced by CTRUNCATE is shown in green. The estimates all agree within the 95% confidence interval (light blue shaded region) determined by the forward-backward algorithm. (a), (b) and (c) exhibit somewhat smooth changes in the amplitude behaviour, whereas (d) shows very sharp and irregular changes, which are likely to be unphysical.

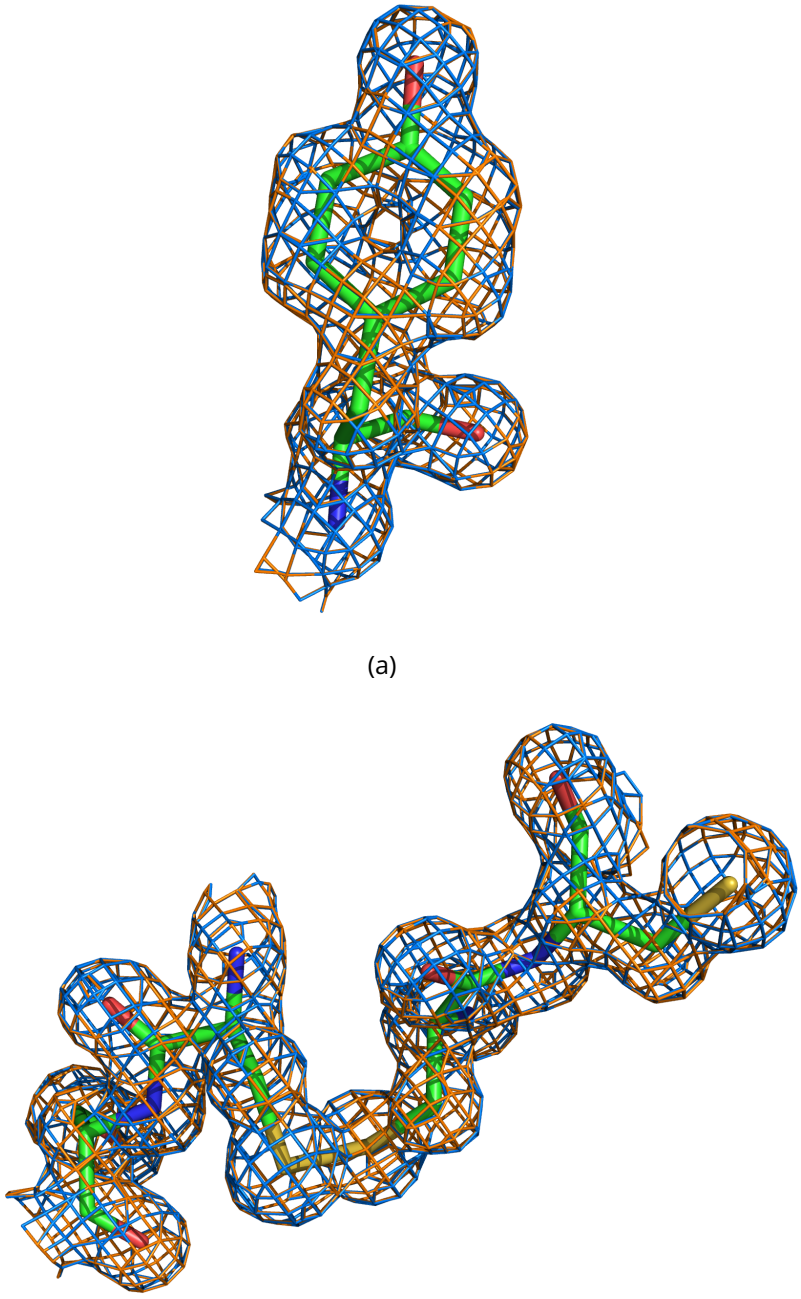
ment procedure was performed with data processed using AIMLESS (Evans and Murshudov, 2013) and CTRUNCATE (ACT pipeline) (Winn *et al.*, 2011) i.e. no processing with the forward-backward algorithm. The resulting electron density maps contoured at the 1σ level (arbitrary choice of σ level) at 1.38 Å for selected residues are shown in Figure 4.14. The maps are practically identical for the two different data reduction pipelines, which is the case for the rest of the structure.

A difference map was also calculated to locate the major differences between the results of the two different data reduction pipelines. The differences were calculated between the resulting amplitudes from the ACT and FBA pipelines, rather than the calculated amplitudes after refinement. Phases for the difference map were obtained from the model generated from the data processed through the ACT pipeline. Figure 4.15 displays the resulting difference map contoured at the 3σ level along with the full insulin structure from which the phases were obtained. The difference electron density is distributed quite uniformly over the unit cell rather than showing large differences overlapping the structure. This supports the result shown above that the two methods give practically equivalent electron density for the structure.

Refinement statistics for both pipelines are shown in Table 4.1. Overall the statistics are quite similar but they are consistently better for the ACT pipeline. It is likely that the FBA results can be improved by merging the amplitude estimates for symmetry related reflections after the algorithm has been applied to each observation individually, as described in section 4.6.1. Another difference between the two pipelines is that the FBA pipeline does not include any outlier rejection whereas AIMLESS does. This is likely to be the reason why there were more reflections at the end of data reduction using the FBA pipeline (16261) compared to that when using the ACT pipeline (16233).

Table 4.1: Final refinement statistics for data processed with the ACT and FBA pipelines

	ACT	FBA
R work	0.165	0.171
R free	0.177	0.182
RMS bond length (Å)	0.029	0.030
RMS Bond Angle (°)	2.493	2.552



(a)

(b)

Figure 4.14: $2F_o$ - F_c electron density maps contoured at the 1σ level at 1.38 Å resolution for the ACT pipeline (blue) and the FBA pipeline (orange) with the insulin structure obtained after refinement with REFMAC with data processed via the ACT pipeline. (a) Tyrosine residue. (b) Disulphide bond. The electron density maps are practically identical and these are representative of the density around the entire structure.

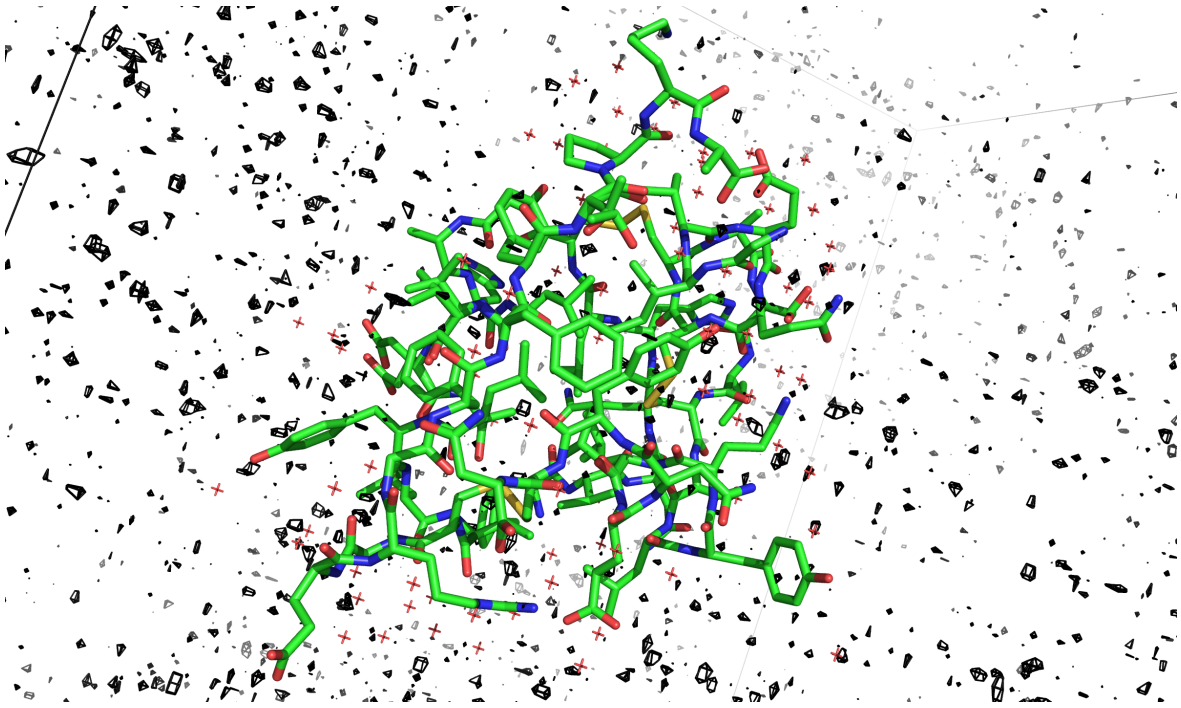


Figure 4.15: Difference electron density map (black mesh) contoured at the 3σ level between the amplitudes resulting from the ACT pipeline and the FBA pipeline using the phases obtained from the model resulting from refinement with the data processed using the ACT pipeline. The insulin structure from which the phases were obtained is also shown as a green stick model. The difference density is uniformly distributed throughout the unit cell and no large differences can be seen overlapping the structure. This suggests that the two methods would result in the same structure as evidenced by the electron density maps in Figure 4.14.

4.6.2 Protein-DNA complex - C.Esp1396I

Scale and B factors

Data were collected from a crystal of the bacterial protein-DNA complex (C.Esp1396I) as described in Bury *et al.* (2015) (Figure 4.16). Notably the crystal ($30\,\mu m \times 30\,\mu m \times 10\,\mu m$) was exposed to a $25\,\mu m$ circular Gaussian profile beam (FWHM dimensions before the $25\,\mu m$ diameter pinhole are $0.212\,mm \times 0.279\,mm$), with the crystal oriented such that the $10\,\mu m$ dimension was aligned parallel to the beam direction. A single dataset consisted of 100 frames with each frame generated from a 1° rotation (100° total wedge). Thus not all of the crystal was immersed in the beam and the rotation led to the X-ray beam path length and illuminated volume through the crystal changing throughout the experiment. The atomic composition used to provide expected intensity values was obtained from the PDB structure 4X4B. The B factors were calculated as described above and are shown in Figure 4.17. Four B factors that were removed in the outlier rejection procedure are clearly visible in Figure 4.17a with B factor values of zero. The initial B factor is much higher for the C.Esp1396I structure

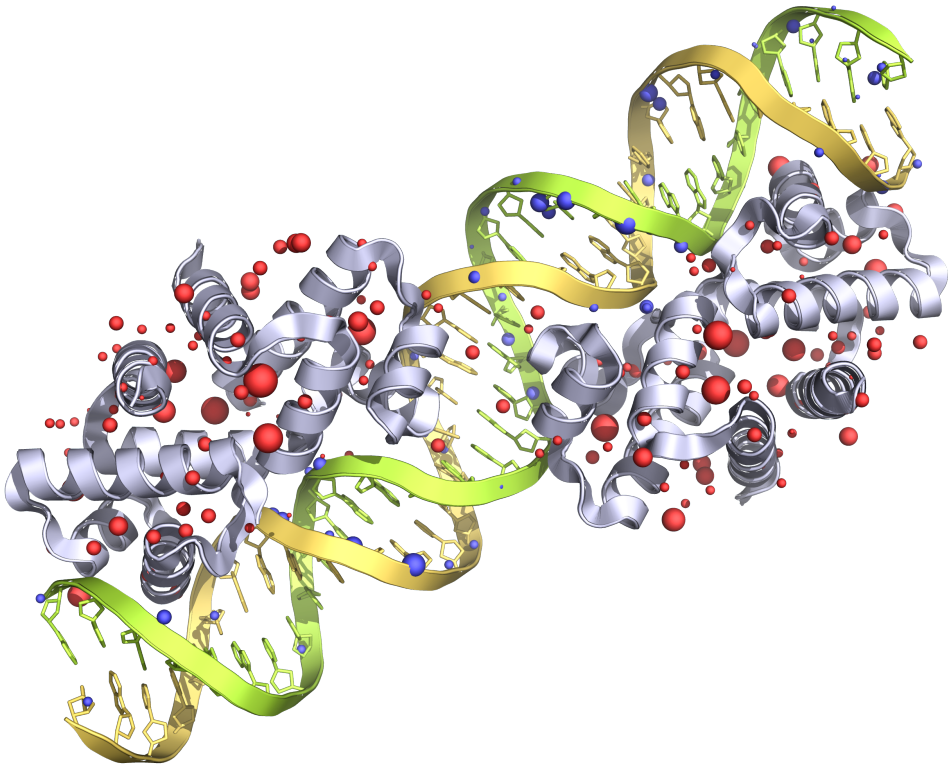
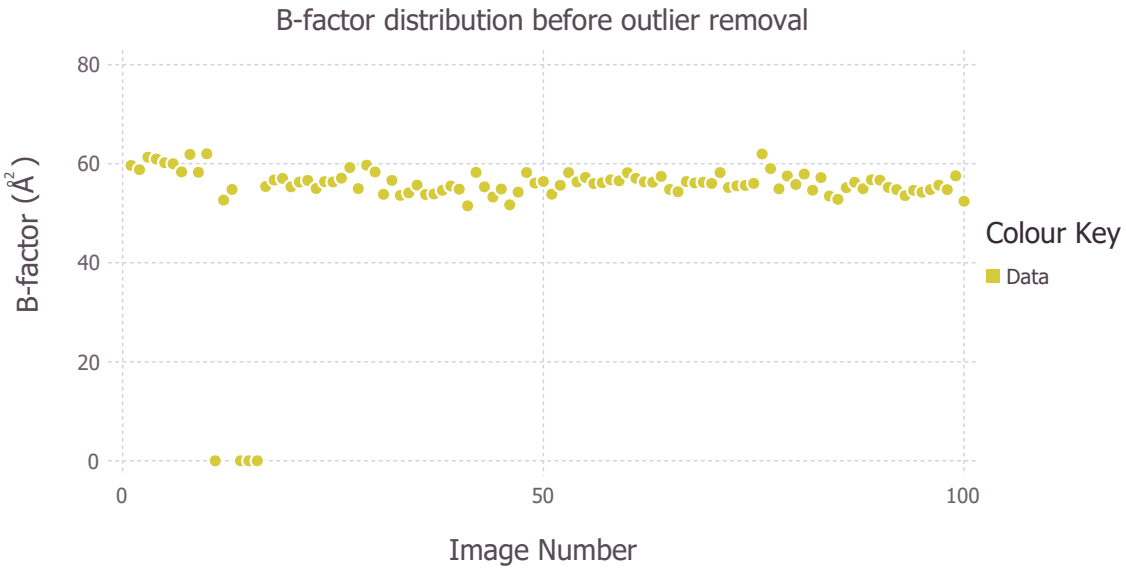


Figure 4.16: Structure of the C.Esp1396I protein-DNA complex. The spheres show sites of specific radiation damage at a dose of 44.6 MGy. The radii of the spheres are proportional to the electron density loss and the spheres closer/further than 2 Å from the DNA strands are coloured blue/red (Bury *et al.*, 2015).

(57.69 Å^2) than it was for the insulin structure (12.19 Å^2). In contrast to the behaviour observed with the insulin dataset, the B factor for the C.Esp1396I structure decreases linearly throughout the dataset (Figure 4.17b). The assumption of a linear behaviour of the B-factor is again justified (Figure 4.18).

The calculated scale factor distribution, $\{s_{images}\}$, is shown in Figure 4.19. As with the insulin dataset, no specific outlier rejection was carried out on the scale factors, so the only ones rejected were those that corresponded to the same images on which the rejected B factors were found, i.e. the four scale factors with a value of 1 in Figure 4.19a. The resulting scale factors, $\{s^*_{images}\}$, calculated from the images (Figure 4.19b) do not exhibit a constant behaviour. This implies that the assumption of a constant scale factor (as was assumed to be the case for insulin) is likely to give incorrect results. The kernel density estimate in Figure 4.20 shows that the scale factor distribution is bimodal and hence a single (constant) value does not represent the distribution adequately. A varying scale factor has not yet been implemented into the forward-backward algorithm and hence the mean value of the calculated scale factor distribution, s_{mean} = 0.0147, was used as the scale factor in the processing.

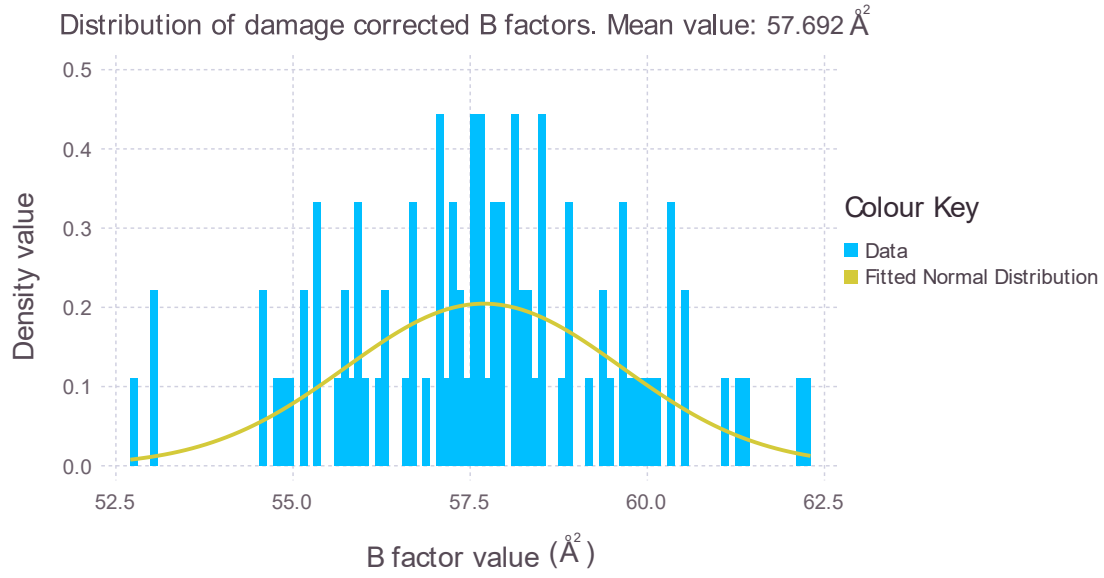


(a)

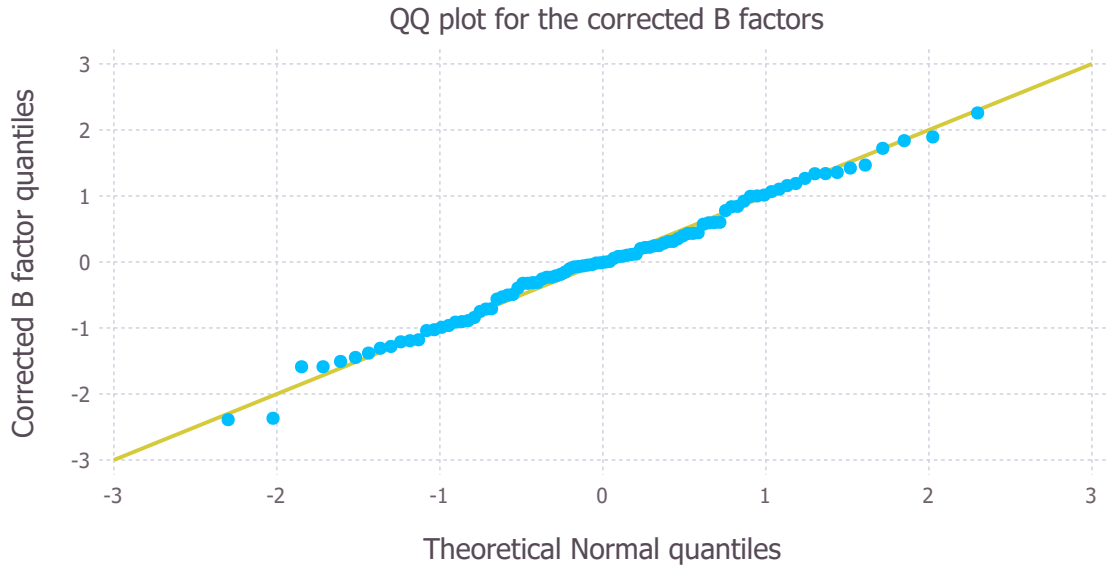


(b)

Figure 4.17: Calculated B factors for each image in the C.Esp1396I dataset. (a) Distribution before outlier removal. (b) Distribution after outlier removal. The line of best fit (blue solid line) with gradient, ΔB = -0.0296 Å^2 and intercept = 57.592 Å^2, is overlaid on the data.



(a)



(b)

Figure 4.18: (a) Histogram of damage corrected B factors for the C.Esp1396I structure. (b) QQ plot for the damage corrected B factors. The linearity of the points in the QQ plot supports the Gaussian approximation in the histogram, suggesting that the B factors change linearly.

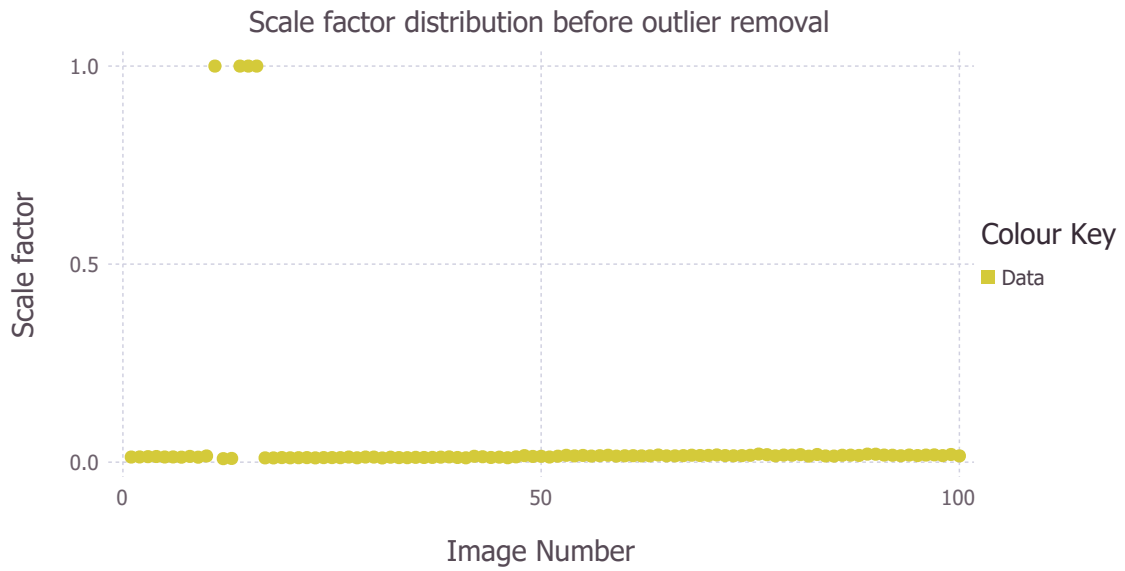
Forward-backward algorithm

The forward-backward algorithm was carried out with the same parameters as defined for the insulin structure. The only difference was that a Gaussian approximation was used as the process function for every reflection. Furthermore the Bayesian inference method described in section 4.5.1 was not performed, because the algorithm suffered numerical issues when calculating the Rician approximation. This is likely to be due to the calculation of the modified Bessel function of the first kind of order zero for arguments with high values. Mathematically this function increases in an exponential manner, especially for large argument values, and ultimately reaches a point where an error is flagged. The Gaussian approximation does not rely on evaluating this function which is why it was used instead for all reflections. Addressing this issue will require the algorithm to check the size of the argument before evaluating it.

The amplitude estimates for two reflections resulting from the processing are shown in Figure 4.21. As was the case with the insulin dataset, some reflections exhibit smooth behaviour (Figure 4.21a), whereas others show sharp changes that are likely to be due to imperfect scale factors (Figure 4.21b). Due to the incorrect scale factor used for the images in the dataset, the CTRUNCATE and the FBA amplitude values do not agree. This is confirmed by the median percentage error between the amplitude estimates for all reflections from CTRUNCATE and the forward-backward algorithm which in this case is 272.60%.

Refinement results

To obtain an electron density map of the structure the initial amplitude estimates resulting from FBA were used along with the deposited C.Esp1396I model (PDB code 3CLC) and refined with REFMAC (Murshudov *et al.*, 2011) (20 cycles of rigid body refinement followed by 20 cycles of restrained refinement). The same procedure was also performed using the amplitudes calculated using the ACT pipeline. The resulting electron density maps contoured at the 1σ level at 2.4 Å for selected residues are shown in Figure 4.22. Once again the maps generally agree structurally but the overlap is less pronounced in this structure than it was



(a)



(b)

Figure 4.19: Calculated scale factors for each image in the C.Esp1396I dataset. (a) Distribution before outlier removal. (b) Distribution after outlier removal. The solid green and solid pink lines represent the mean and mode of the distribution respectively. The scale factor is clearly not constant and shows an increasing trend throughout the experiment. This leads to differences in the mean and modal values of the distribution.

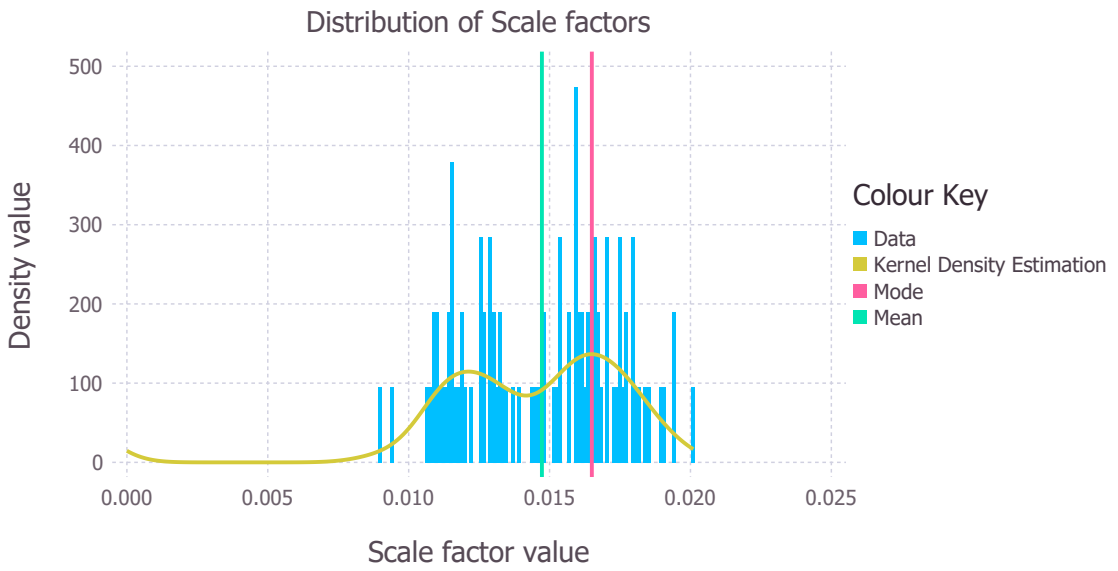


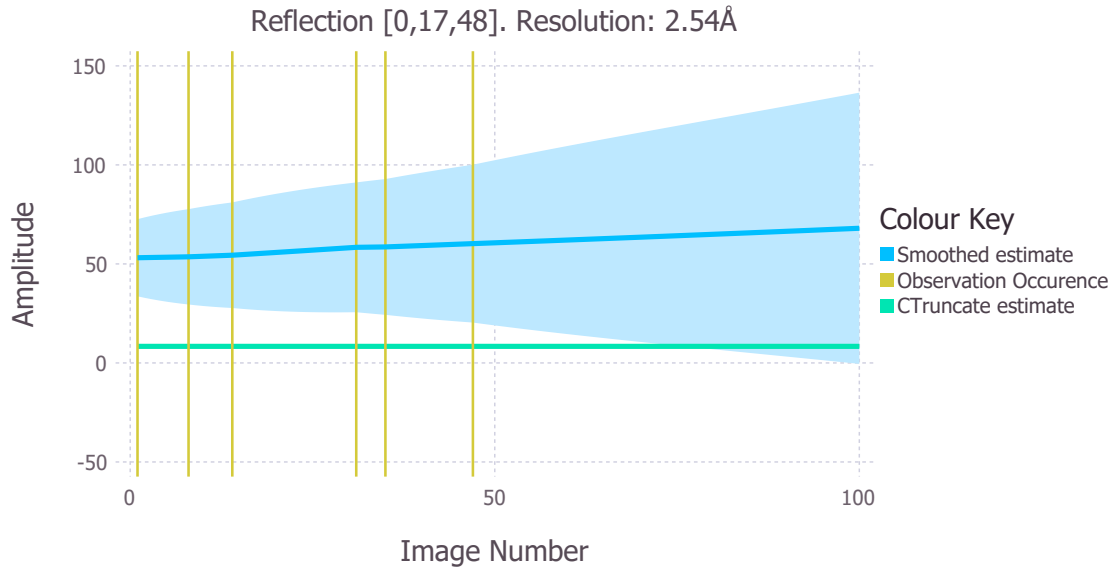
Figure 4.20: Histogram of scale factors with the mean (solid green line), mode (solid pink line) and kernel density estimation (solid gold line) overlaid. The bimodal distribution of the scale factor is clear from the kernel density estimate.

for insulin. This is expected because the amplitude values did not agree as well, largely due to an incorrect scale factor used for the FBA.

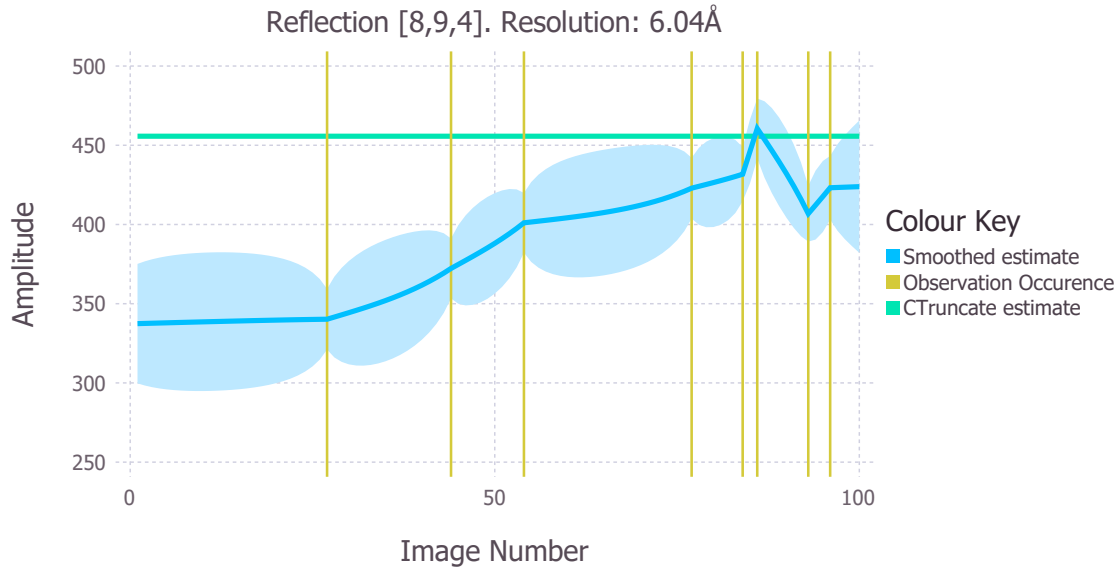
The difference map between the amplitude values calculated from the ACT and FBA pipelines, using phases from the final model after refinement using the ACT amplitudes and the C.Esp1396I structure (PDB 3CLC) as a starting model, is shown in Figure 4.23. The difference density is no longer random and in fact is located around the model. This suggests that the two different pipelines could lead to different structural models.

Again the total number of reflections differ between datasets. The FBA pipeline results in 32944 reflections at the end of data reduction, whereas the ACT pipeline results in 32898 reflections.

Overall refinement statistics using both pipelines are shown in Table 4.2. The R values are slightly better for the ACT pipeline but the RMS values are lower for the FBA pipeline. The statistics obtained using the FBA method are much better than expected considering the scale factor used in the FBA was non-optimal. If the amplitude values resulting from the FBA pipeline are indeed inaccurate then it is likely that the refinement process is significantly "filtering out" the errors that are made during the data reduction stage. It is also possible that the statistical improvement by manually building the model may be more limited with

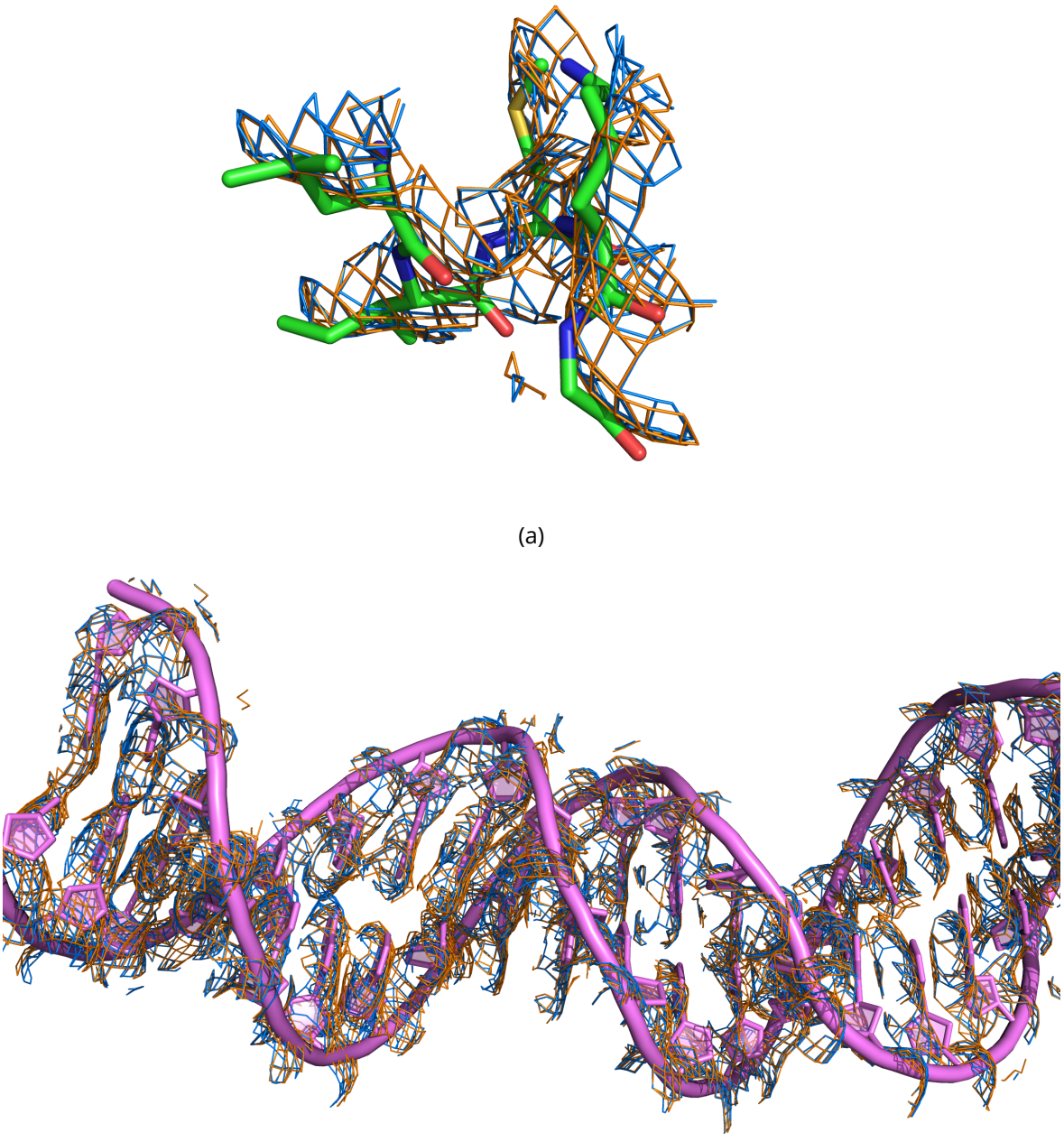


(a)



(b)

Figure 4.21: Amplitude estimates for two different reflections observed in the C.Esp1396I dataset using the forward-backward algorithm (blue solid line). The estimate produced with CTRUNCATE is shown in green. The estimates using the two different pipelines do not agree in both (a) and (b) and this is likely to be due to the incorrect scale factor used for the forward-backward algorithm. (a) In the second half of the experiment, there are no observations of reflection 0,17,48 and hence the uncertainty in the amplitude values propagates and increases. (b) Reflection 8,9,4 exhibits sharp changes in the amplitude, which are likely to be noise.



(a)

(b)

Figure 4.22: $2F_o$ - F_c electron density maps contoured at the 1σ level at 2.42 Å for the ACT pipeline (blue) and the FBA pipeline (orange). The model was obtained after refinement with REFMAC with data processed via the ACT pipeline. (a) Leu-Ile-Met-Lys-Gly residues of the protein from the C.Esp1396I protein-DNA complex. (b) DNA section of the C.Esp1396I protein-DNA complex.

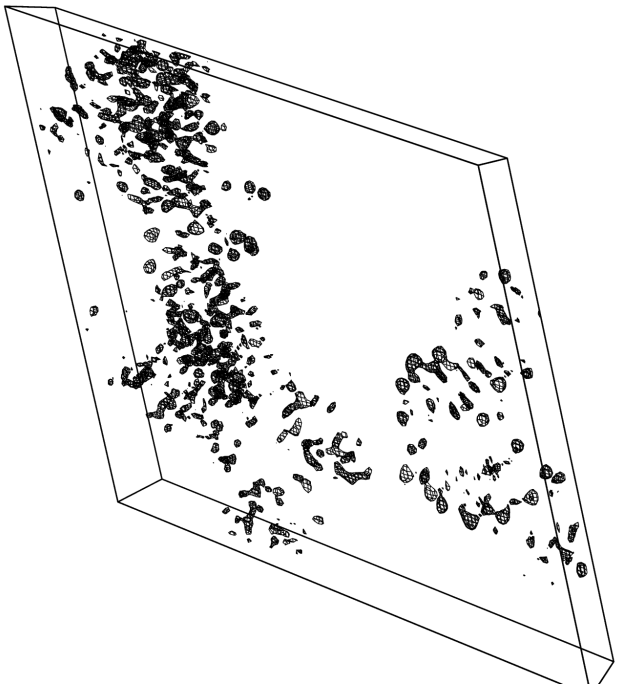


Figure 4.23: Difference electron density map (black mesh) contoured at the 3σ level between the amplitudes resulting from the ACT pipeline and the FBA pipeline using the phases obtained from the model given by refinement with the data processed using the ACT pipeline. The difference density is not random as it was for the insulin structure. Instead most of the difference density is located around the structure, suggesting that the refined models are likely to differ in several regions.

the data for the FBA pipeline compared to that of the ACT pipeline.

Table 4.2: Final refinement statistics for data processed with the ACT and FBA pipelines.

	ACT	FBA
R work	0.257	0.259
R free	0.282	0.288
RMS bond length (Å)	0.014	0.012
RMS Bond Angle (°)	1.937	1.806

4.7 Discussion

4.7.1 Overview of the forward-backward algorithm

The work presented in this chapter introduces a representation of the data collection experiment as a hidden Markov model (HMM). The fundamental idea is that each image obtained in a diffraction experiment is generated from a different crystal. The crystals are related by the fact that the atomic composition of the crystal is the same, and the structural changes due to the X-ray exposure between images are small. The Markov property makes the as-

sumption that the state of one crystal is only dependent on its previous state, not on its entire history. With these assumptions, the Luzzati distribution (Luzzati, 1952; Read, 1990) can then be used to mathematically describe how the crystal state evolves (process function). Furthermore the intensity observations are directly proportional to the square of the structure factor amplitude giving the observation function. With this representation, the UKF and URTSS together (FBA) are used to find the optimal amplitude estimates during the experiment. The log likelihood of the resulting amplitude estimates can then be used to determine whether the FBA has converged for each reflection.

Before FBA is applied, the reflection observation data had to be pre-processed to the required format, which required some data manipulation. The main objective was to allocate each observation to an image. For fully recorded reflections there is no ambiguity but for partial reflections this was achieved by matching the calculated centroids of the reflections to the images. Additionally, intensity and partiality estimates were made for reflection observations that were not fully traversed in the experiment. Further to the error values calculated from the integration software, more uncertainty had to be added to the intensity estimates arising from the collapsing of several partial measurements of an observation to a single image collected at one point in time.

The FBA algorithm was tested on simulated reflection data and the results were extremely good for strong data. The resulting amplitude estimates were close to the true values, and the confidence intervals were sensible, decreasing in width when observations were made and increasing further away from observations. For weak data the FBA estimate did not perform as well, so a correction to the initial amplitude value using Bayesian inference (in a similar manner to the French and Wilson truncation algorithm) was created.

The FBA was successful when applied to real crystallographic data. It generated amplitudes that led to interpretable electron density maps for both insulin and the C.Esp1396I protein-DNA complex. Furthermore, the final model refinement statistics are on a par with those that result from using current data reduction pipeline programs (AIMLESS and CTRUNCATE).

Advantages of the FBA

One of the major advantages of the FBA algorithm is that the error estimates are calculated explicitly at each time point in the data collection experiment. These error estimates are a combination of the uncertainty in the integrated measurement (observation covariance) and the uncertainty of the changes suffered by the crystal (process covariance) propagated through time. Since the amplitudes are found directly by the FBA, it abrogates the need to use existing truncation algorithms including the commonly used French and Wilson algorithm (French and Wilson, 1978). This is an advantage, because for large experimental uncertainties the French and Wilson algorithm produces error estimates that resemble the Wilson distribution, which result in weak measurements having a significant influence on a structural model (Read and McCoy, 2015).

The other major advantage of the FBA is the fact that the amplitude estimates are time/dose resolved, so that several electron density maps can be obtained from a single diffraction dataset. The dose was not explicitly used here because the crystal was irradiated in a standard single axis rotation experiment. This meant that the rotation angle/time are directly proportional to the dose and either could be used as a proxy for it. In a more complex experiment where the crystal is translated, the dose calculated in successive frames does not change monotonically. In this case the order of the images would be explicitly dependent on the calculated dose and the results of a dose calculation program like RADDOSE-3D would be required for this method to work efficiently.

In theory, the set of structure factor amplitudes given at each point in time could lead to slightly different models showing structural changes throughout the experiment (e.g. due to radiation damage). However it is unclear how sensitive downstream processing (such as refinement) is to small amplitude changes in a subset of reflections, and whether the processing will influence the resulting models enough to hinder the observation of altered conformations. For example, if it is the case that the applied restraints in refinement are weighted highly, then subtle structural deviations from the “ideal” conformation may be missed. Further investigation will be required to assess this issue.

An additional, albeit minor, advantage to the FBA is that the framework is very modular by design. The process and observation functions are simply based on the current theoretical

understanding of structure factor statistics and diffraction theory. If the description of the crystal evolution process were to change, then this could simply be incorporated by changing the process function and covariance functions, and it would not require a refactoring of the code. Similarly, if the detector were to operate differently and the theory of how the observations (intensities or not) were to alter, then this would only change the process and observation functions. An example of this is that both the Gaussian and Rician process functions exist in the code as separate functions, and either one can be called very simply. Thus it is also very easy to compare different processes.

Disadvantages of the FBA

The main disadvantage of the FBA is that in its current implementation, it is computationally expensive. The insulin dataset which consists of 450 images took about 16 hours to process through the algorithm, whereas the C.Esp1396I dataset which only had 100 images took around 4 hours to run despite containing around twice the number of reflections. This suggests that the computational time is largely dependent on the total number of images. There are many ways in which the performance can be improved. An obvious one would be to write it in a faster language such as C++, but the code would still need to be written in a more optimal manner. For example (although not the biggest bottleneck), to read reflection information the program runs MTZDUMP and then parses the resulting text to retrieve the information. This can be improved if the binary MTZ file contents are read directly using the MTZLIB. Another way to speed up the code would be to run the FBA in parallel, which is possible because the reflections can be assumed to be independent.

A fundamental feature of the HMM is that the scale factor is simply a parameter concerned with the observation and it is not intrinsic to the crystal. At first this seems at odds with the current scaling assumption that the scale factor contains terms that are intrinsic properties of the crystal e.g. unit-cell volume, and the fact that the radiation damage parameter(s) are refined simultaneously with the other scaling factors (absorption, detector response to X-ray photon, etc.). However, it should be theoretically possible to separate the factors that are affected by the crystal changes and those that are a property of the observation method (e.g. observing intensities on a Pilatus detector). A simple thought experiment to demonstrate this is to consider a crystal that is irradiated by X-rays with no detector present.

The crystal is still going to change due to radiation damage regardless of whether intensity measurements are made. Hence a description of the crystal changes must (in theory) be possible without reference to a scale factor that describes how the intensity observations are made. Therefore it can be inferred that the HMM representation would operate optimally in the ideal scenario where the scale factor is known. This is not that case as "the only information we have is the measured difference between symmetry-related observations" (Evans, 2006).

4.7.2 Improvements and extensions

The algorithm presented has yielded promising results but there are still several improvements that could be made to the software, one of the obvious being to improve the scaling procedure. The current method is very primitive and better methods to obtain the scale and B factors have already been proposed (Popov and Bourenkov, 2003). Furthermore, the program should accommodate methods to handle varying scale and B factors. Non-parametric regression methods such as Gaussian process regression (GPR) could be utilised to do this. GPR makes no assumption about the functional form to describe the evolution of the scale or B factors. An additional benefit is that the Gaussian errors in the regression are also calculated, which allows for the explicit propagation of errors in the algorithm so that a more accurate uncertainty represented by the process and observation covariance values can be obtained. Rather than defining parameters for a certain family of functions (e.g. the gradient and intercept of linear curves) as is the case for parametric regression, non-parametric regression methods usually require the user to loosely define more general properties of a function such as the smoothness and covariance (Rasmussen and Williams, 2006). Restraints would have to be applied during the regression to ensure that the behaviour of resulting amplitude estimates were "sensible". This may require iterative feedback between the amplitude values and the scale and B factors.

Ultimately, the scaling procedure would benefit from utilising the methods that are currently implemented in software programs such as AIMLESS and XSCALE rather than reinventing the wheel. These are very mature algorithms that reflect the current knowledge and best practice in scaling. Therefore an immediate improvement would be to run AIMLESS and output unmerged, scaled intensity values. The resulting B and scale factors can then be used to

construct prior distributions on these parameter values. However, it is the amplitude values that are ultimately of interest, so in this respect, the B and scale parameters can be treated as nuisance variables. This means that marginalisation (integrating out the variables) can be carried out to obtain the most likely distribution for the amplitudes given the information that has already been obtained. The other advantages of doing this are that the current method implemented to extract reflection intensities (described in section 4.4) would effectively be carried out by AIMLESS. Additionally AIMLESS incorporates an outlier rejection algorithm so another one would not necessarily have to be written for the FBA. In the case where AIMLESS is used to scale the data, the FBA algorithm could be used simply as a truncation algorithm giving dose resolved amplitude estimates with sensible error estimates for all reflections.

One of the assumptions that was made in deriving the process function is that the changes in the structure factor amplitude resulting from the average coordinate error could be absorbed into the temperature factor term. This assumption may not hold and investigation into the atomic coordinate error distributions should also be carried out to determine the true effect of it.

An important step in establishing the robustness of this method was applying the FBA to track the behaviour of a simulated reflection. This allowed an understanding of how the algorithm performed with various amplitude changing behaviours. However, before applying the FBA method on real protein crystal data, it should be tested on a simulated dataset with a program like MFLSOM (Holton *et al.*, 2014). This would be beneficial because it is possible to manually vary the extent of the error in the data, which would be crucial for determining which parts of the FBA algorithm need to be tweaked so that it would perform better, and also determine its limitations.

To extend the use of the algorithm, the FBA could be used to merge data from several crystals. First, pairwise cross-covariance matrices between sets of amplitude values resulting from applying the FBA to several different crystals should be calculated. The elements of the resulting matrices can be used to determine whether the data from different crystals can be merged (Garib Murshudov, personal communication).

As mentioned above, one of the major features of the algorithm is that it produces time/dose resolved amplitude estimates, which provides a unique opportunity to perform radiation

damage correction. This means that the behaviour of reflections can be tracked explicitly, particularly for reflections that are observed multiple times. If the assumption is made that reflections in the same resolution bin behave similarly, then the behaviour of reflections that were only observed once can be predicted by determining the “average” behaviour of multiply observed reflections in the same resolution bin. The obstacle with this method that must be overcome is how to average ‘behaviour’ irrespective of the varying scales on which the reflections are observed.

4.7.3 Future work

It should be noted that the main goal of the FBA algorithm is to improve diffraction data for experimental phasing, in particular the uncertainty estimates for weak reflections. Additionally, correction for radiation damage should also improve the phasing signal. Therefore the next steps are to apply the algorithm to SAD data (with the extensions/alterations listed above) to determine whether these improvement gains can be realised.

CHAPTER 5

X-ray Beam Analysis

5.1 Introduction

The ability to accurately calculate the dose absorbed by a sample is vital to the reproducibility of experimental results in radiation damage research and to inter-compare in MX, as well as to develop protocols aimed at advising experiments on how to spread the dose and thus optimise the use of their crystals. As alluded to in the introduction (section 1.4.3), the extent of knowledge on some of the parameters in the diffraction experiment is limited, which results in uncertainty in the calculated dose values. These uncertainties are not quantified and hence RADDOSE-3D does not explicitly calculate errors on the dose values. The uncertainty in some experimental parameters, such as the crystal volume or the unit cell volume, only result in small errors in the aggregate dose metrics (average dose whole crystal, maximum dose, DWD). On the other hand, uncertainty in the beam profile can lead to significant errors. For this reason, a module in RADDOSE-3D was implemented to allow it to simulate MX experiments using experimentally measured X-ray beam profiles, which can be read by the program. However, the measured beam profiles require preprocessing to remove systematic errors resulting from the actual measurement before they can be used in RADDOSE-3D.

The work presented in this chapter describes the preprocessing of experimentally measured beam profiles and explores the errors that can arise at this stage in the dose calculation.

5.2 Processing aperture measurements

Prior to the measurements reported here (July 2013), users at Diamond Light Source (DLS) beamline I02 were provided with the full width half maximum (FWHM) values of the X-ray beam measured by a 10 μm aperture (see below). To obtain a more realistic value of the FHWM it was recommended that 5 μm be subtracted from the supplied FWHM values. This recommendation however was not supported by any systematic experiments or studies.

For this reason investigations were carried out to obtain a better estimate of the true beam profile to so that a more trustworthy value of the FWHM could be provided to users. This was the initial motivation for processing beam measurements made using aperture scans.

5.2.1 Experimental methods

Aperture scans to obtain X-ray beam flux measurements were carried out in collaboration with Dr. Carina Lobley on the DLS I02 beamline. For these scans, a miniap device, a piece of steel with a 10 μm diameter circular hole, is translated across the beam by remote control. The position of the aperture is recorded along with a measure of the current detected in a silicon diode (S3590-09 model purchased from Hamamatsu) on the detector shutter (i_pin) (Owen *et al.*, 2009). The detector current is proportional to the beam flux (photons/second). First the 'centre' x and y position is found such that the diode reading at this position gives the highest current. The aperture is then translated 120 μm in the negative horizontal direction. Measurements of the x and y positions and the diode reading are carried out at 2 μm intervals as the aperture is translated across in the positive horizontal direction for a total of 240 μm (+120 μm from the centre). A similar scan is then carried out in the vertical direction. Examples of data from the measurements are shown in Figure 5.1.

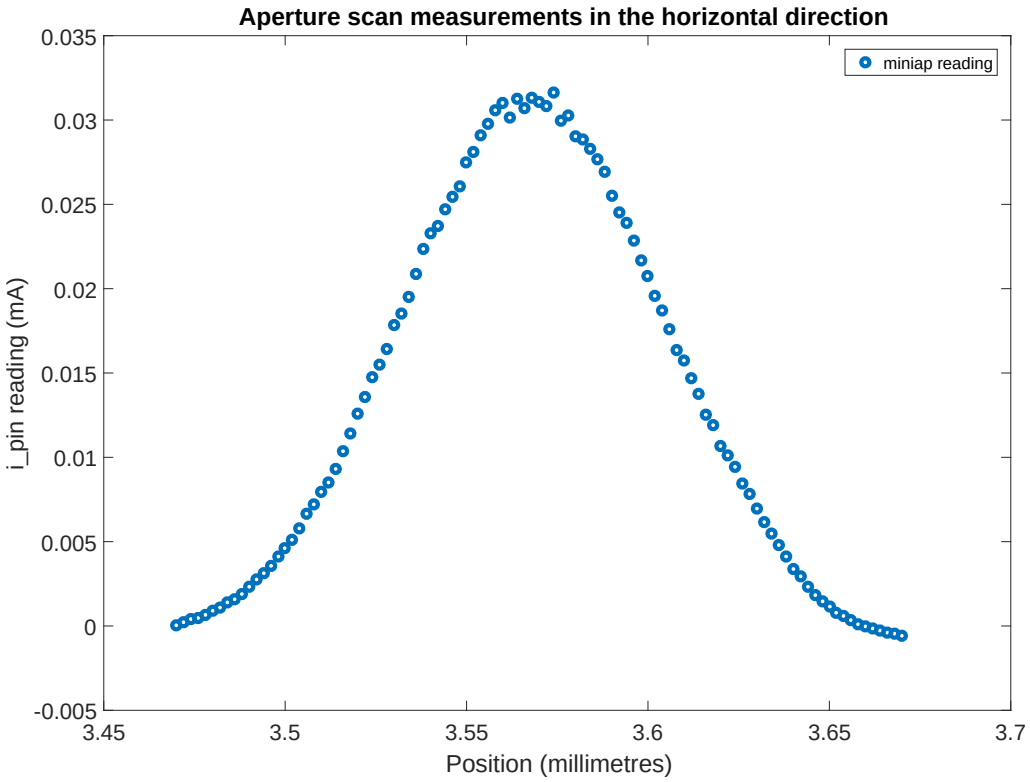
5.2.2 Deconvoluting the X-ray beam measurements

Theoretical introduction

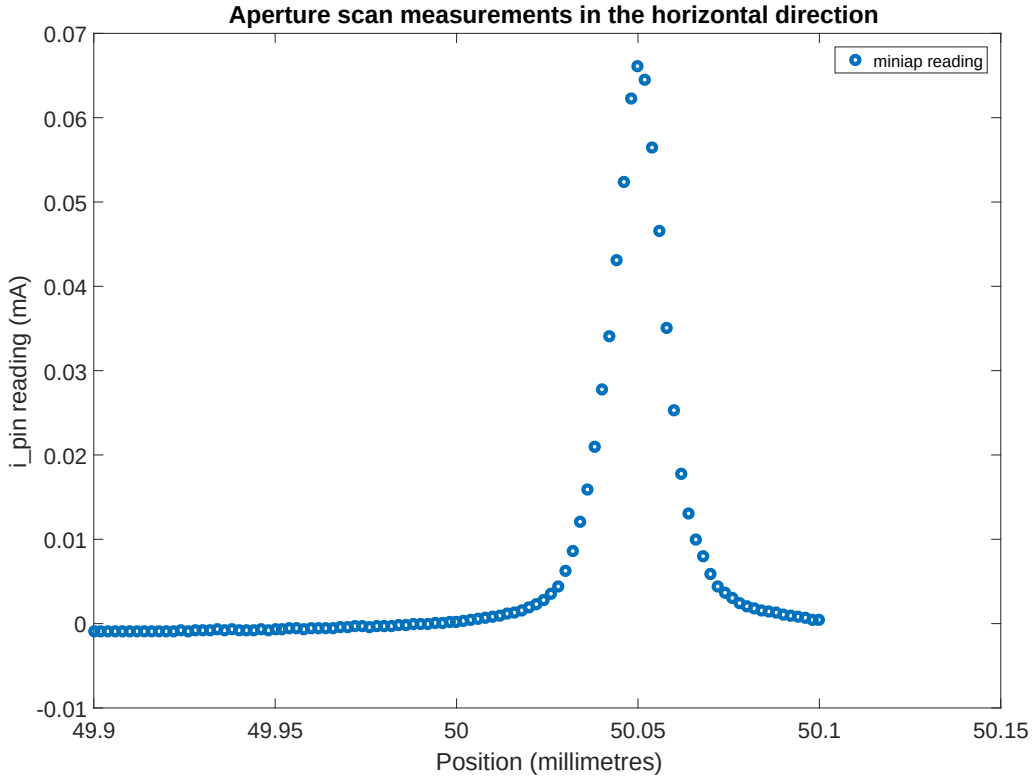
The area of the aperture used on I02 is (10/2)$^2 \times \pi \approx 78.54\,\mu m^2$, so each diode reading at a position (x, y) is the result of the integral of the flux from a 78.54 μm^2 area surrounding the central point (Figure 5.2). Given that the diode measurements were taken at 2 μm intervals, it is possible to deconvolute the measured signal to obtain a truer value of the diode current to a spatial resolution of 2 μm. Only this estimate of the true 2D profile of the current is necessary for the beam profile measurement. This is because RADDOSE-3D additionally requires a total flux estimate, which is distributed across the measured beam profile (a 2D array) according to the current at each spatial position.

To define the problem mathematically, the diode readings can be described as a convolution of the true diode current, f, at a point $\boldsymbol{x}$, and the area of the aperture, g. Explicitly this is

$$[f * g](\boldsymbol{x}) = \int_0^{\boldsymbol{x}} f(\boldsymbol{u}) g(\boldsymbol{x} - \boldsymbol{u}) \mathrm{d}\boldsymbol{u} + n(\boldsymbol{x}), \tag{5.2.1}$$



(a)



(b)

Figure 5.1: Examples of the flux measurement data collected at the Diamond Light Source synchrotron. The horizontal axis represents the aperture position in millimetres and the vertical axis represents the current in the diode (mA). (a) Data collected in the x-direction. (b) Data collected in the y-direction.

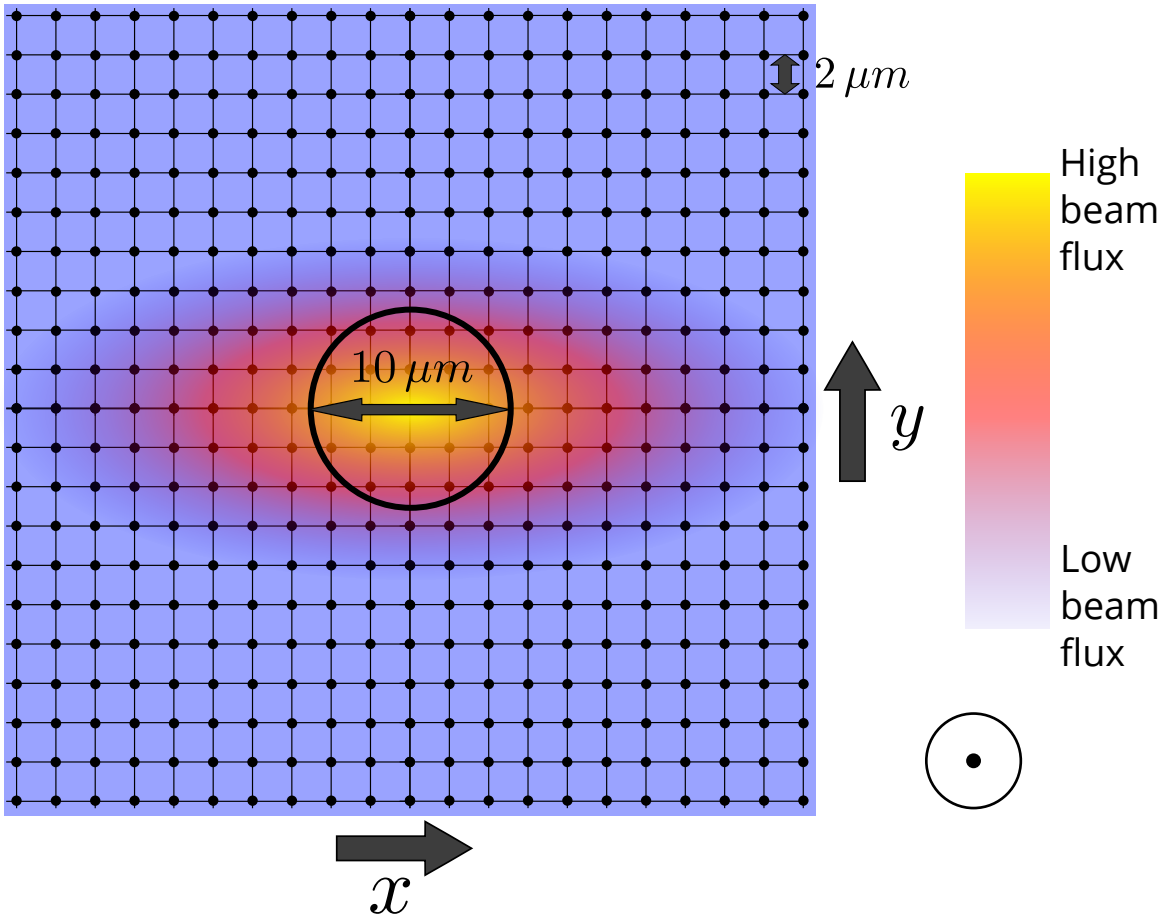


Figure 5.2: A schematic of the X-ray beam and aperture setup as viewed from the detector looking into the beam. Each point represents a spatial position where a beam measurement is taken, hence the distance between any two horizontally or vertically consecutive spots is 2 μm. The circle represents the circular aperture which has a 10 μm diameter. The colour represents the beam flux intensity (not to scale). The beam is smaller in the y direction than it is in the x direction. This diagram shows that a reading at a particular point in space is the integral of all surrounding points within the aperture area. Thus the reading of the total X-ray beam profile that is measured is a convolution of readings from local surrounding space.

where $[f * g]$ is the measured diode reading and $n(\boldsymbol{x})$ represents the additive noise at a particular point in space. The aim of deconvolution in this context is to find the diode current profile, $f(\boldsymbol{x})$, given knowledge of the measured current, $[f * g](\boldsymbol{x})$, and the aperture contribution, $g(\boldsymbol{x})$. Mathematically this aim can be interpreted as finding some $w(x)$ such that:

$$\hat{f}(x) = [w * [f * g]](x), \tag{5.2.2}$$

where $\hat{f}(x)$ is an estimate of $f(x)$ that minimises the mean square error. In general, deconvolution is an ill-posed problem, thus it is likely that a unique solution does not exist (Weisstein, 2016). If it is assumed that the noise has zero mean and is spatially independent (white noise assumption) then the Wiener filter (Wiener, 1949) can be used to deconvolute the signal to find f. The Wiener Filter, W, is defined mathematically in the Fourier domain as:

$$W = \frac{1}{\mathscr{F}(g)} \left[\frac{|\mathscr{F}(g)|^2}{|\mathscr{F}(g)|^2 + \frac{\mathscr{F}(n)}{\mathscr{F}(f)}} \right], \tag{5.2.3}$$

where $\mathscr{F}$ denotes the Fourier transform of its argument:

$$\mathscr{F}[f(x)](k) = \int_{-\infty}^{\infty} f(x) e^{ikx} \mathrm{d}x. \tag{5.2.4}$$

The $\frac{\mathscr{F}(n)}{\mathscr{F}(f)}$ term in equation 5.2.3 is similar to the inverse of a signal to noise ratio in the Fourier domain. If this value is known exactly, then the Wiener filter often works very well. However, more commonly the value of this ratio is unknown, and it is often treated as a constant value throughout the domain.

The Wiener filter is multiplied with the convoluted function in the Fourier domain to give

$$\mathscr{F}[\widehat{f}] = W \mathscr{F}[f * g], \tag{5.2.5}$$

such that $\mathscr{F}[\widehat{f}]$ is the solution found with the minimum mean squared error:

$$\text{error} = E\left[\left(\mathscr{F}[f] - \mathscr{F}[\widehat{f}] \right)^2 \right], \tag{5.2.6}$$

where the true solution is denoted $\mathscr{F}[f]$. Multiplication in the Fourier domain is equivalent to a convolution operation in the real domain, making clear the equivalence between equation 5.2.5 and the statement of the mathematical aim (equation 5.2.2). Parseval's theorem implies that minimising the mean squared error in the Fourier domain is equivalent to min-

imising the mean squared error in the real domain. For further details on the Wiener Filter and its formal derivation the reader is referred to González and Woods (1992) .

Implementing the deconvolution

The deconvolution algorithm was implemented in Matlab R2012a. First the initial diode measurements were read into Matlab and a one-dimensional Gaussian function was fitted to the data in order to obtain the parameter values in both the x and y directions (Figure 5.3). The parameter values were subsequently used for the two-dimensional Gaussian function representing the estimate of the convoluted two dimensional beam profile (Figure 5.4a).

$$g_{2D}(x,y) = A_{max} \exp\left[-\left(\frac{(x-\mu_x)^2}{2\sigma_x^2} + \frac{(y-\mu_y)^2}{2\sigma_y^2}\right)\right], \tag{5.2.7}$$

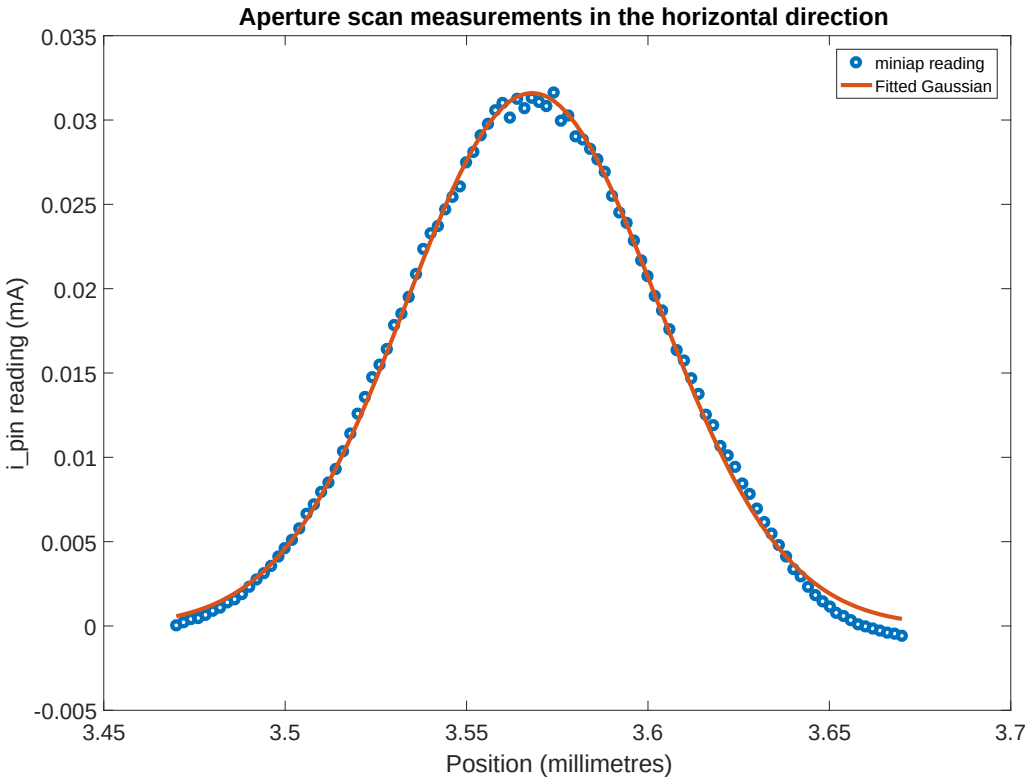
where μ_x, μ_y, σ_x and σ_y are the parameters whose values are obtained from fitting the one-dimensional Gaussian functions

$$g_h(x) = A_x \exp\left[-\frac{(x-\mu_x)^2}{2\sigma_x^2}\right], \tag{5.2.8}$$

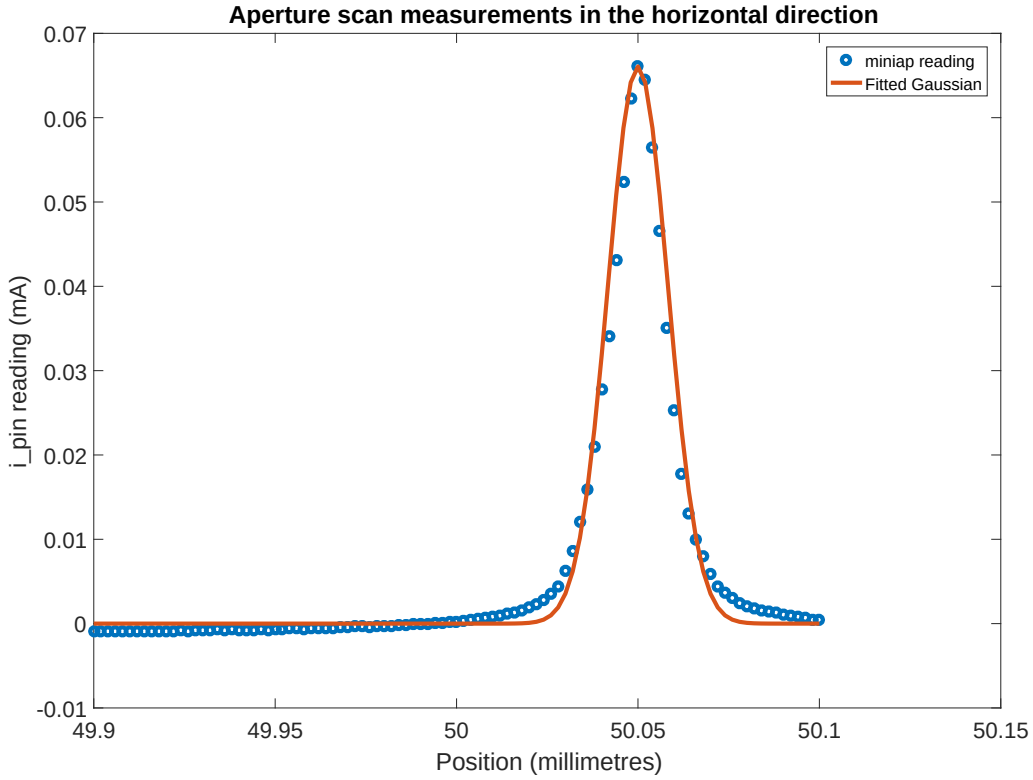
$$g_v(y) = A_y \exp\left[-\frac{(y-\mu_y)^2}{2\sigma_y^2}\right], \tag{5.2.9}$$

and A_{max} is the maximum value of A_x and A_y which are the values of the highest recorded diode current in the x and y directions respectively. If the measurements are taken perfectly (i.e. aperture is scanned exactly across the maximum flux in x and y) then $A_x = A_y$, but due to experimental error this is rarely the case.

To obtain an estimate of the true relative X-ray beam profile, equation 5.2.7 has to be deconvoluted with a matrix that corresponds to the aperture. The aperture matrix is a square matrix created just large enough to contain the area of the circular aperture. It essentially acts as a mask, so elements of the matrix contain the value 1 if the area is inside the aperture, and 0 if the element is outside the aperture. To determine the position of a matrix element with respect to the aperture area, each point in the matrix is given the value of the Euclidean distance away from the central matrix element. This central element is assumed to be at the position of the centre of the aperture. Any points with a Euclidean distance smaller than the aperture radius are determined to lie within the aperture area and hence given a value of 1. However points with a Euclidean distance bigger than the aperture radius



(a)



(b)

Figure 5.3: The original miniap readings from translating the 10 μm diameter aperture are shown as blue circles and the solid orange line is the Gaussian fit to the data. (a) beam profile in the x-direction. (b) beam profile in the y-direction. Data collected on I02, DLS.

are outside and are set to 0.

The $\frac{\mathscr{F}(n)}{\mathscr{F}(f)}$ term in equation 5.2.3 is estimated using the `fminsearch` minimisation function in Matlab. Thus an objective function was written as follows:

1. first an estimate of the signal to noise ratio is passed to the objective function
2. using the estimates a deconvolution of the signal is performed using the two-dimensional beam profile given by equation 5.2.7 and the aperture matrix.
3. a convolution of the deconvoluted signal with the same aperture matrix is performed.
4. the Frobenius norm* of the original convoluted signal subtracted from the newly calculated convoluted signal is calculated and returned by the objective function.

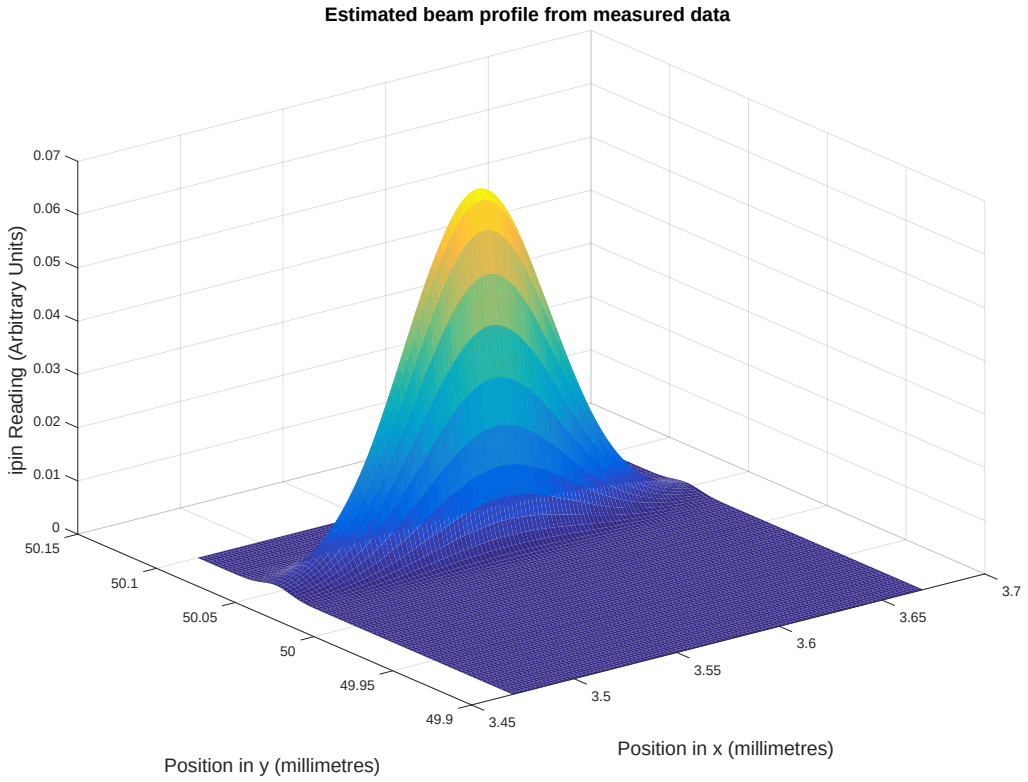
If the two signals give a norm value of zero then they are exactly the same, but otherwise a positive value is returned. The `fminsearch` function returns the value of $\frac{\mathscr{F}(n)}{\mathscr{F}(f)}$ that minimises the objective function. Given the convoluted two-dimensional beam profile, the aperture matrix and $\frac{\mathscr{F}(n)}{\mathscr{F}(f)}$, the beam profile can be deconvoluted using the Matlab function `deconvwnr` which performs a Wiener deconvolution as described in section 5.2.2 (Figure 5.4b).

The deconvoluted 2D beam profile is not smooth, which is typical of deconvolution when the exact noise distribution is unknown. A further 2D Gaussian function is fitted to the deconvoluted data in order to obtain a smooth beam profile (Figure 5.5). This is carried out by taking 1D slices of the 2D beam profile in both the x and the y directions around the slice that gives the maximum readings in the centre, and then obtaining the parameters as outlined above (section 5.2.2).

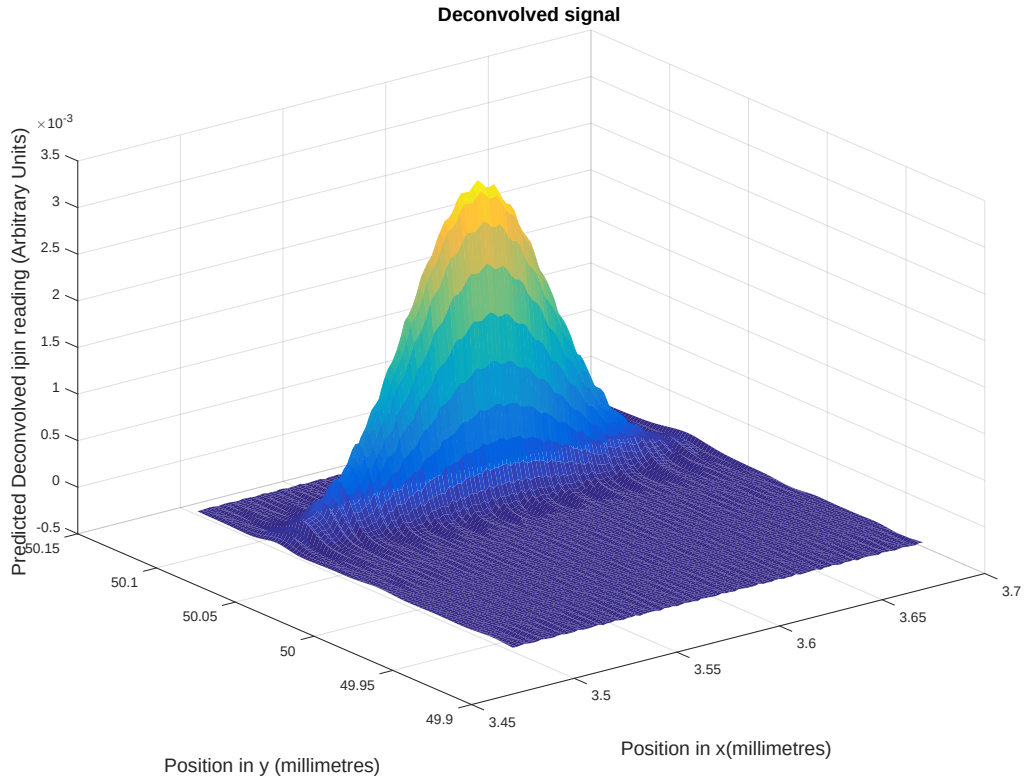
Using the fitted Gaussian model, any properties of the beam can be determined, such as the Full Width at Half Maximum (FWHM) of the beam, which can then be provided to the I02 DLS users. The FWHM is defined mathematically for a Gaussian as

$$FWHM = 2\sigma\sqrt{2\ln(2)}. \tag{5.2.10}$$

*The Frobenius norm, $||A||_F$, of a matrix, A, is defined as $||A||_F \equiv \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2}$ where the a_{ij} represent the elements of the matrix.



(a)



(b)

Figure 5.4: 2D X-ray beam profiles. (a) Gaussian approximation of the measured current profile given by equation 5.2.7 (b) Beam profile after Wiener deconvolution.

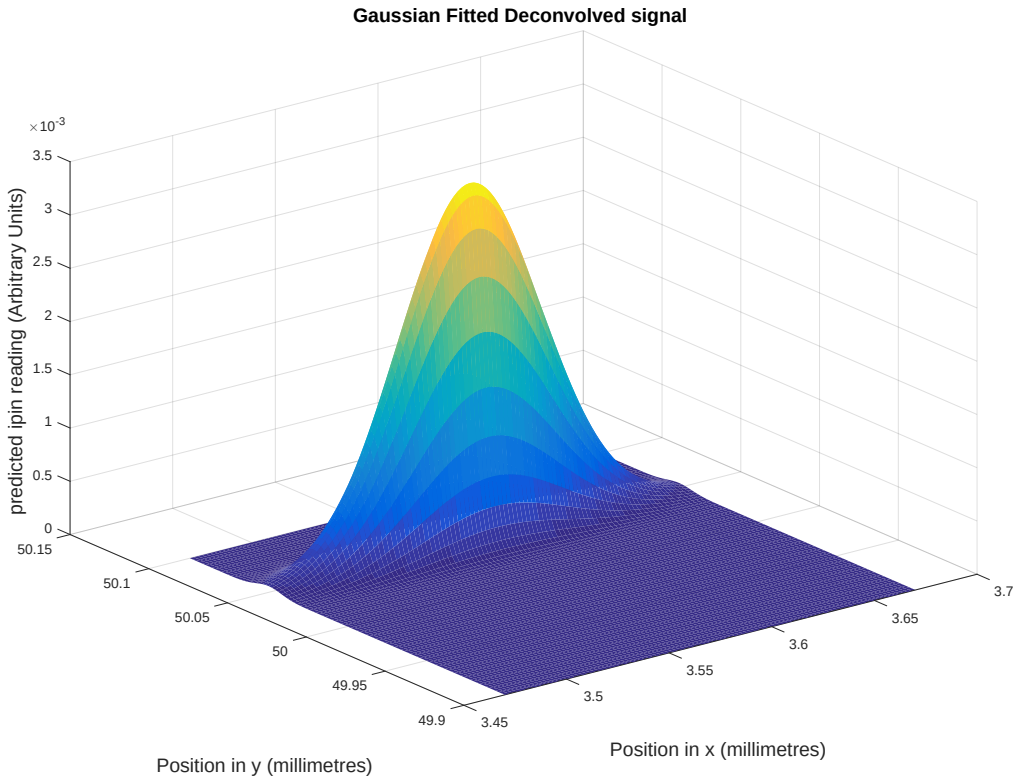


Figure 5.5: Gaussian model fitted to the deconvoluted X-ray beam profile (Figure 5.4b).

Validation of deconvolution approach

No measuring device is available to take measurements at the resolution to which the deconvolution method attempts to spatially resolve the flux readings. Therefore to validate the results, the deconvolution method was used in reverse to predict the FWHM values that would be obtained from the diode readings with a 20 μm diameter aperture. To achieve this the 2D deconvoluted signal (Figure 5.5) was convoluted with an aperture matrix of a different size. 1D slices in the x and y directions were then found such that the flux in the central matrix element was maximum and 1D Gaussian functions were fitted to each slice. FWHM values were then calculated from the 1D Gaussians using equation 5.2.10 and these calculated values could then be compared with the readings taken from actual experimental data obtained using a 20 μm aperture.

5.2.3 Results

FWHM values obtained from the experimentally measured (convoluted) i_pin readings using the 20 μm diameter aperture (referred to as Exp_{20}) were compared with calculated FWHM values derived from the numerical convolution of the deconvoluted profile with the 20 μm aperture (referred to as $Conv_{20}$ - Table 5.1). The calculated FWHM for $Conv_{20}$ are quite accurate, showing less than a 3% error in both the x and y directions when compared to Exp_{20}. This suggests that the deconvolution method used provides reasonable estimates of the 2D X-ray beam profile.

Table 5.1: Comparison of calculated FWHM values with experimentally observed FWHM values. The calculated FWHM values using the 10 μm aperture (in italics) are calculated from the deconvoluted profile. Thus they were not expected to match the experimental values.

	Aperture diameter (μm)	FWHM in x (μm)	% Error in x	FWHM in y (μm)	% Error in y
Exp_{10}	10	81.6	0.49	19.5	4.84
$Deconv_{10}$	*10*	*81.2*		*18.6*	
Exp_{20}	20	81.8	2.57	21.9	1.83
$Conv_{20}$	20	83.9		22.3	

FWHM values obtained from the experimentally measured (convoluted) i_pin readings using the 10 μm diameter aperture (referred to as Exp_{10}) were compared with calculated FWHM values derived from the deconvoluted profile with the 10 μm aperture (referred to as $Deconv_{10}$ - Table 5.1). The errors of the FWHM values between Exp_{10} and $Deconv_{10}$ are also quite small. In theory however, these values do not necessarily have to be close because the FWHM from $Deconv_{10}$ are calculated from the deconvoluted profile, whereas the FWHM from Exp_{10} are calculated from the convoluted beam profile. The fact that these values are close provide evidence that users should not subtract 5 μm from the quoted FWHM values. Furthermore it indicates that deconvoluting the X-ray beam does not significantly change the X-ray beam profile. This is important because it suggests that the measured 2D current profile does not have to be further altered once the profile has been generated from the 1D aperture scans.

5.3 2D X-ray beam profile measurements

The previous section presented work on beam processing performed on measurements of the beam profile obtained from 1D aperture scans. In this section, the analysis is described of measurements of the X-ray beam profile carried out at PETRA III synchrotron, beamline P14, Hamburg, as detailed in section 2.2.3. Specifically, the measurement was made using a scintillator combined with an Allied Vision GC1350C CCD camera, which resulted in a 2D image of the X-ray beam. The processing of these images presents different challenges to those that arise from 1D aperture scan measurements.

The 2D image of the X-ray beam profile was exported as a portable graymap (pgm) file where each pixel contains an integer value between 0 and 255 inclusive (Figure 2.3a). These values were determined by the size of the signal from the scintillator. To spatially resolve the flux over the beam image, RADDOSE-3D calculates the flux, F, for a particular pixel as:

$$F = \frac{p_{ij} T_F}{A_p \sum_{ij} p_{ij}}, \tag{5.3.1}$$

where p_{ij} is the pixel value in the image, T_F is the total measured flux, and A_p is spatial area covered by the pixel. This means that the pixel values can be regarded as weights that correspond to the relative photon flux at that spatial position. In general the pgm file will always contain non-zero values that correspond to background signal. This can be seen in Figures 2.3b and 2.3c where the pixel values never decay to zero at any point in the slices through the beam. RADDOSE-3D interpreted these non-zero values as beam intensity, so some of the flux that should have been in the beam was instead allocated to the background pixels, which resulted in an underestimate of the absorbed dose in the crystal. Therefore it was important to identify the background and remove it from the beam profile measurements before the dose simulation was performed. Several methods for removing the background were explored and the results were compared in order to investigate the effects on the RADDOSE-3D simulation output.

5.3.1 PGM file preprocessing

Original beam profile

The first profile that will be considered is the original pgm file. No preprocessing is performed on the file so that the results from other processed beams can be compared with the unprocessed case. Figure 5.9a shows the profile of this raw beam.

Deconvolution of beam profile

The beam profile measured with the scintillator and CCD camera as described in section 2.2.3 can smear as a result of charge diffusion into adjacent pixels due to CCD pixel subvariations. This response is known as the point spread function (PSF). Thus the image formed in the pgm file is a convolution of the actual X-ray beam profile and the PSF. To deconvolve the beam image, the following procedure was undertaken:

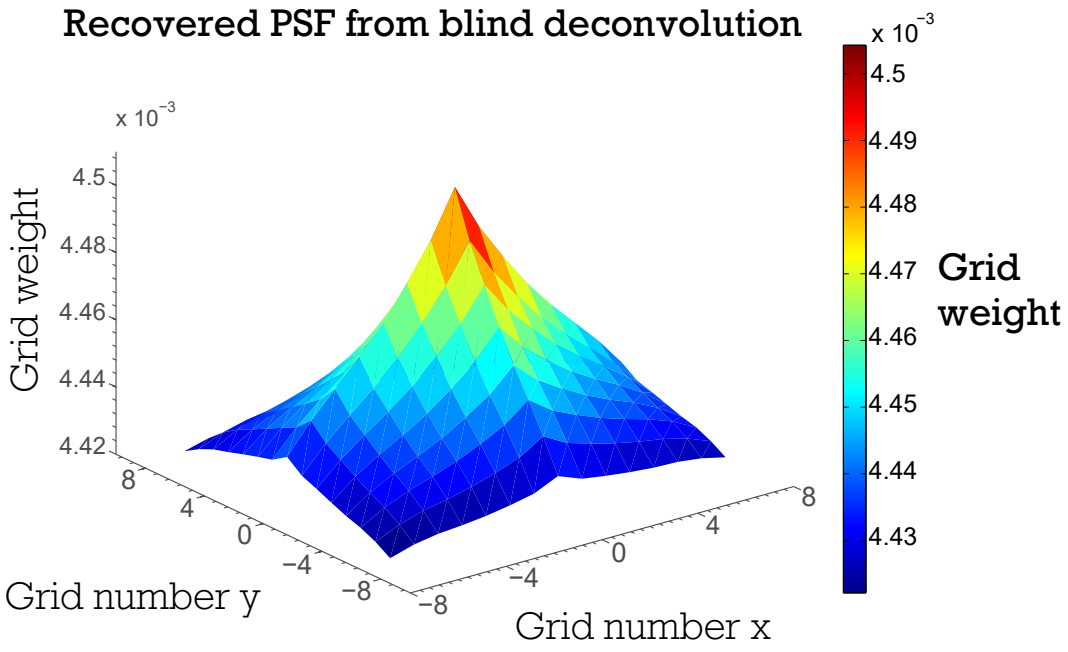
1. An initial blind deconvolution (using the MATLAB `deconvblind` function) was performed to determine an estimate of the PSF (Figure 5.6a). The blind deconvolution requires an initial guess for the PSF, but the PSF restoration in the blind deconvolution is heavily affected by the grid size of the PSF rather than the values of the grid elements. With no knowledge of the PSF (or of the signal to noise power spectrum) an initial guess for the PSF was made as a 7×7 grid of 1's.

2. The restored PSF was then modelled using a Laplacian function of the form

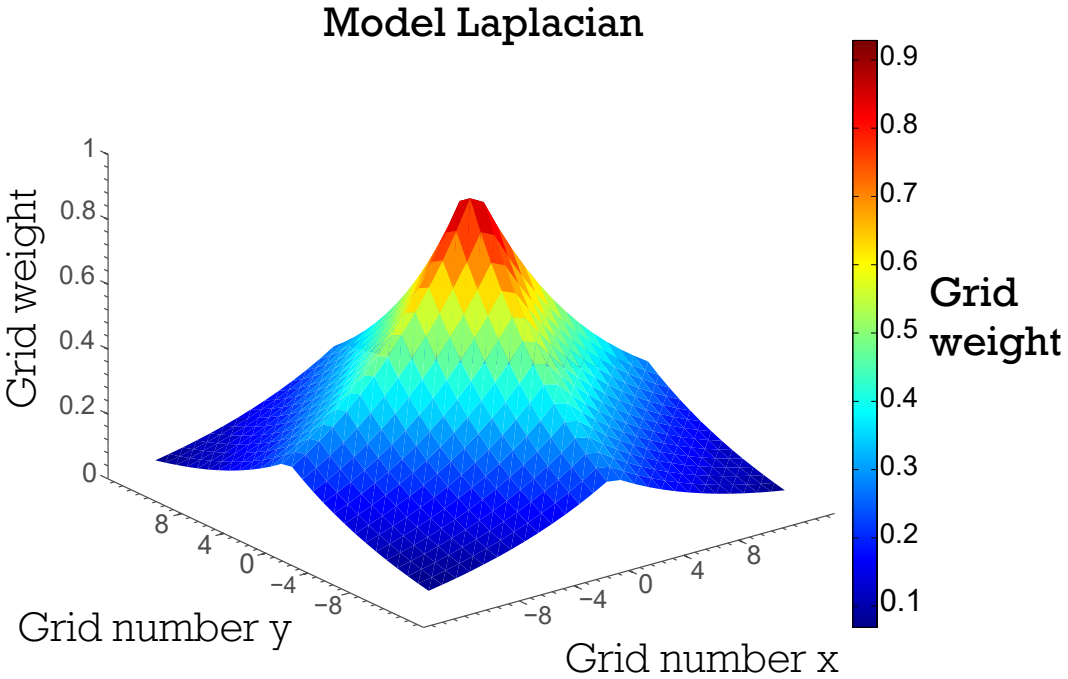
$$f_{lap} = \exp\left(-\frac{|x| + |y|}{b}\right) \tag{5.3.2}$$

   where b is a parameter whose value is to be determined (Figure 5.6b). The Laplacian function was chosen because it modelled the shape of the recovered PSF function very well.

3. The pgm file was then segmented using the MATLAB `activecontour` function, which uses the Chan and Vese region based energy model (Chan and Vese, 2001), to find the spatial area of the beam. Additionally the centroid of the resulting image (the centre of the beam) was determined (Figure 5.7a).



(a)



(b)

Figure 5.6: (a) PSF recovered from performing a blind deconvolution. (b) Laplacian function of the form given in equation (5.3.2) with a 7×7 grid and $b = 10$.

4. A rectangle that bounded the beam (inner red rectangle in Figure 5.7b) was chosen as 145 $\mu m\times$ 140 μm because this contained the beam whilst allowing a buffer for beam area that may have been missed by the segmentation algorithm. These dimensions were stored.

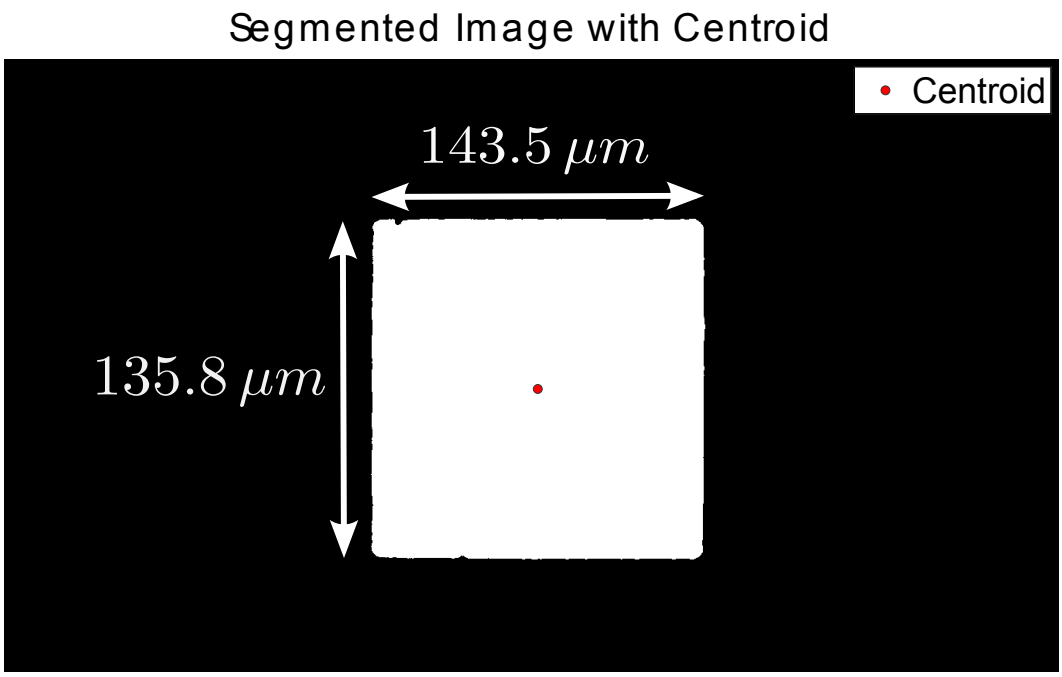
5. The (square) grid size for the PSF, b in equation 5.3.2 and the rectangular dimensions were then used as inputs for an objective function in a minimisation process to determine optimal values for these parameters. The objective function was constructed as follows:

   - A perfect (hypothetical) top hat beam was created inside the inner box (Figure 5.7b).
   - The Laplacian PSF was created using the grid size and b.
   - The top hat beam was convolved with the PSF to create a theoretical (convolved) beam image.
   - the positions of zero pixel values from the theoretical beam image were also set to zero in the original beam image.
   - The height of the theoretical beam image was scaled to the height of the original beam.
   - The matrix 2-norm of the difference between the original beam image with zeroed outer pixel values and the theoretical beam image was calculated and used as the output of the objective function to be minimised (the matrix 2-norm is equal to the square root of the maximum eigenvalue of the matrix multiplied by its conjugate transpose matrix).

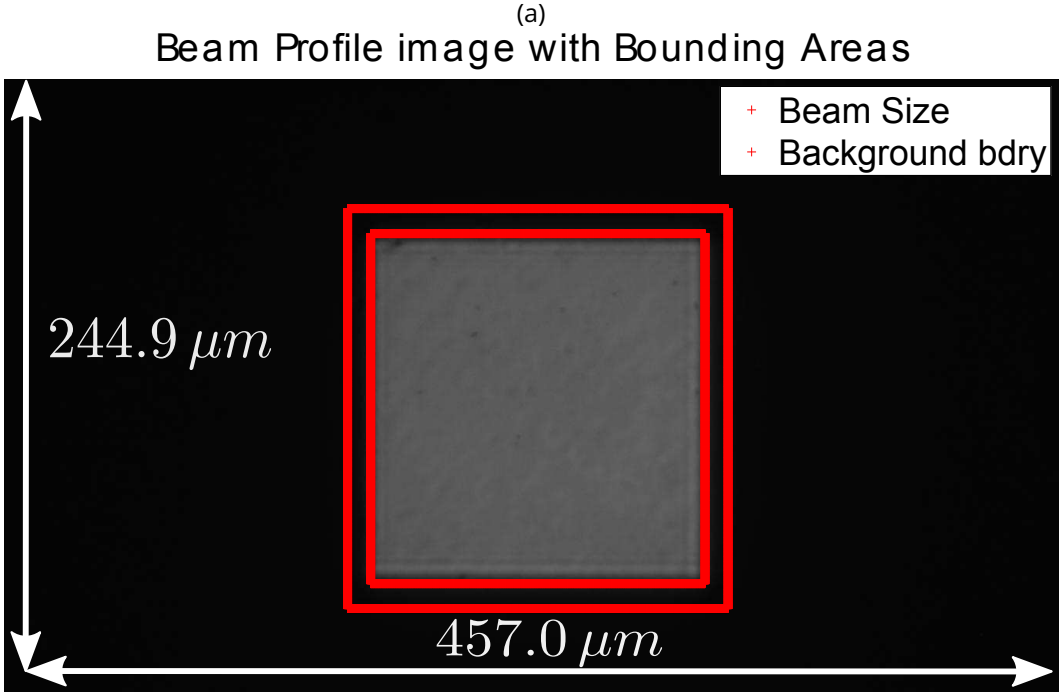
6. Another blind deconvolution was performed as before but this time with the returned grid size from the minimisation procedure as the input. The resulting beam profile was then returned for input into RADDOSE-3D (Figure 5.9b).

Removal of background using deconvolution results

Since the rectangular dimensions of the hypothetical top-hat beam were given as outputs of the minimisation procedure, this shape can be used to distinguish between the X-ray beam



(a)



(b)

Figure 5.7: (a) Segmented image of the beam with its centroid marked as a red circle in the middle. (b) Beam image with the centroid and two bounding boxes overlaid. The inner box is an estimate of the beam size with dimensions 145 $\mu m\times$ 140 μm. These dimensions were chosen according to the coverage of the beam area in the image, as opposed to the size of the aperture from the slit separation (slits were set at 140 $\mu m\times$ 140 μm). The outer box is the boundary,170 $\mu m\times$ 170 μm, where the pixels outside the box are considered part of the background (section 5.3.1).

and background. The two beams described above were also further manipulated by setting any values outside of the rectangle dimensions (inner box) to zero (Figures 5.9c and 5.9d).

Perfect top hat beam

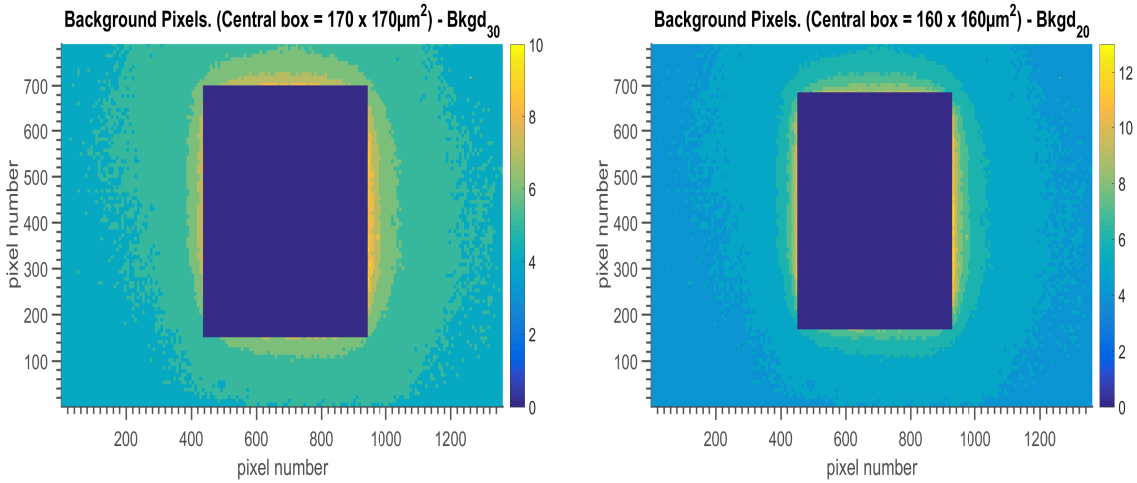
A perfect top hat beam was created within the inner boundary of Figure 5.7b i.e. the value 100 was inserted for pixels that lay inside the inner rectangle, otherwise the pixel values were set to zero (Figure 5.9k).

Beam thresholding

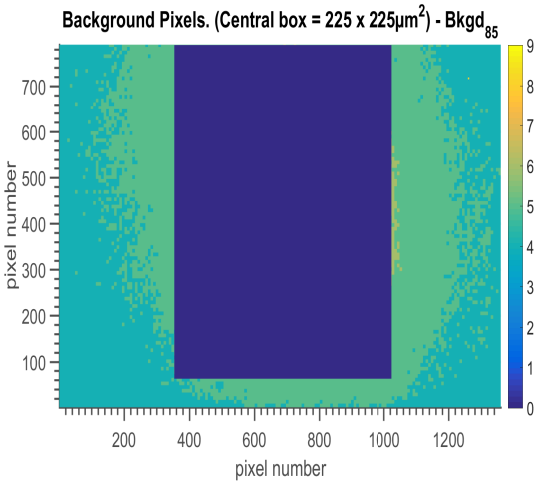
In his D. Phil. thesis, Dr. Oliver Zeldin preprocessed pgm files by setting a threshold value (Zeldin, 2013). The threshold (or dark current) value was determined by taking the average of the pixel values that were "far away" from the main beam centre. However, to my knowledge, there was no systematic way to determine "far away". In the PETRA III MX experiment described here, the slits were adjusted to give an aperture of 140 $\times$ 140 μm^2. To determine how the choice of "far away" affects the threshold to be subtracted, it was decided that background would be considered as any pixel value beyond 30 μm away from the slit edge ($Bkgd_{30}$). The 170 $\times$ 170 μm^2 outer red box in Figure 5.7b is the $Bkgd_{30}$ boundary. Another two beams were created assuming that the background could be considered as any pixel beyond 20 μm from the slit edge ($Bkgd_{20}$ with a 160 $\times$ 160 μm^2 boundary) and 85 μm from the slit edge ($Bkgd_{85}$ with a 225 $\times$ 225 μm^2 boundary). These boundaries are shown in Figure 5.8. The threshold value for subtraction was determined by taking a mean average of the background pixel values and rounding it to the nearest integer (Figures 5.9e, 5.9f and 5.9i). Another way to determine the background was to take the maximum value of the background values and subtract that value from every pixel in the image (Figures 5.9g, 5.9h and 5.9j).

5.3.2 RADDOSE-3D simulation results

All of the beams illustrated in Figure 5.9 were used as input for RADDOSE-3D to simulate the experiment described in section 2.2.3 for insulin crystal ID 0259. Each RADDOSE-3D run produced a set of DWD values; each value corresponded to one of the 271 datasets

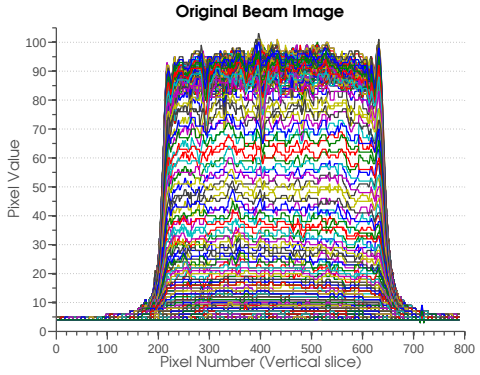


(a) (b)

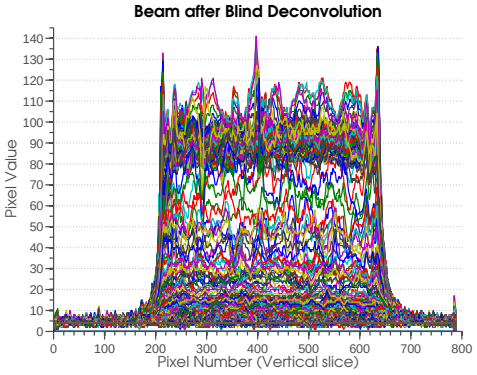


(c)

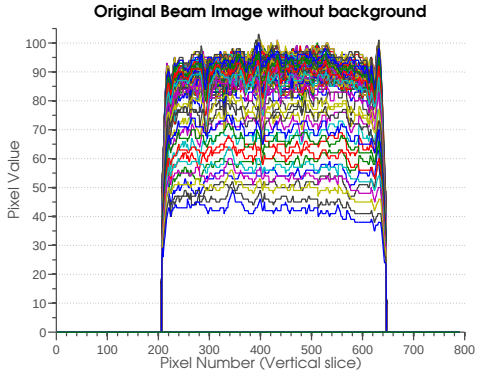
Figure 5.8: Background pixel values for the beam image (beam image dimensions are 457.0 $\mu m \times$ 244.9 μm for all images). The pixel values in the central box for all images are set to zero and the mean and maximum values of the non-zero pixel values are taken for each of the backgrounds. (a) $Bkgd_{30}$: Rounded mean pixel value = 5, maximum pixel value = 10. (b) $Bkgd_{20}$: Rounded mean pixel value = 5, maximum pixel value = 13. (c) $Bkgd_{85}$: Rounded mean pixel value = 4, maximum pixel value = 9.


Original Beam Image
Pixel Value
Pixel Number (Vertical slice)

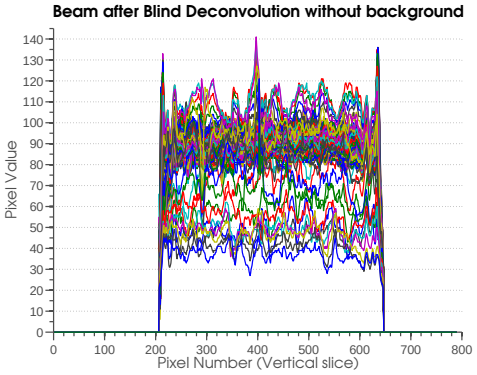
(a)


Beam after Blind Deconvolution
Pixel Value
Pixel Number (Vertical slice)

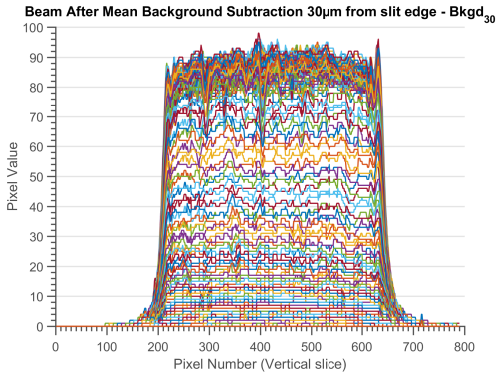
(b)


Original Beam Image without background
Pixel Value
Pixel Number (Vertical slice)

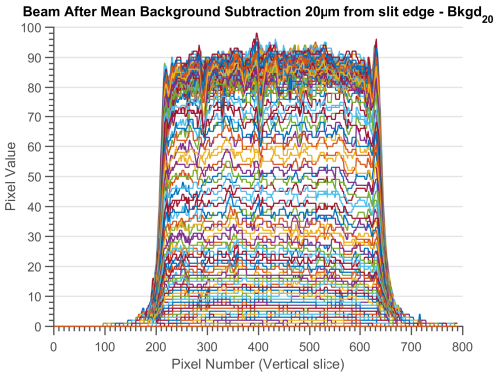
(c)


Beam after Blind Deconvolution without background
Pixel Value
Pixel Number (Vertical slice)

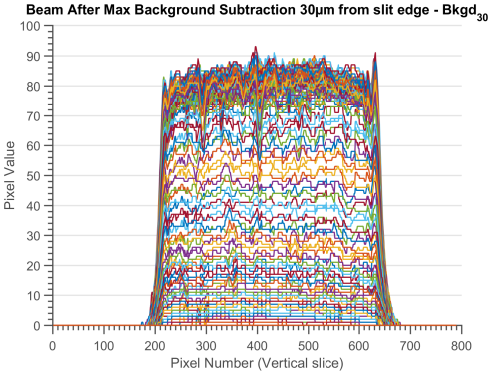
(d)


Beam After Mean Background Subtraction 30µm from slit edge - Bkgd30
Pixel Value
Pixel Number (Vertical slice)

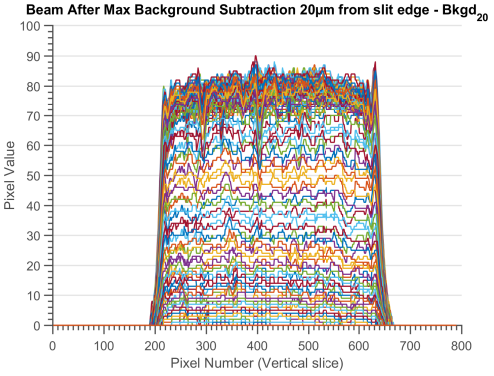
(e)


Beam After Mean Background Subtraction 20µm from slit edge - Bkgd20
Pixel Value
Pixel Number (Vertical slice)

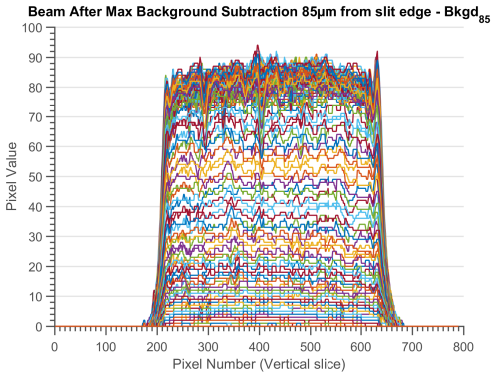
(f)


Beam After Max Background Subtraction 30µm from slit edge - Bkgd30
Pixel Value
Pixel Number (Vertical slice)

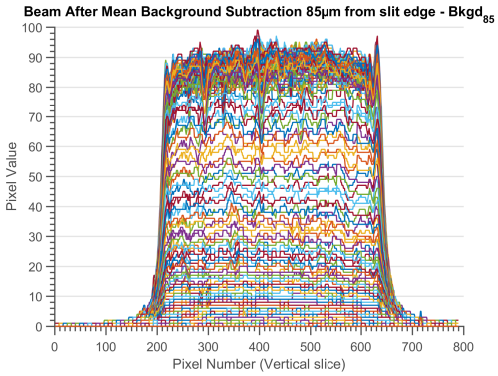
(g)


Beam After Max Background Subtraction 20µm from slit edge - Bkgd20
Pixel Value
Pixel Number (Vertical slice)

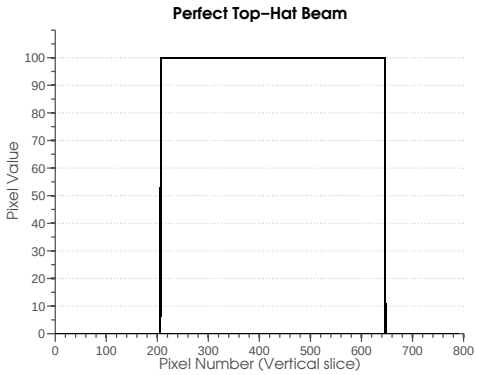
(h)



(i)



(j)



(k)

Figure 5.9: Processed beam profiles used for simulations in RADDOSE-3D. Each figure is a plot with all vertical slices across the corresponding pgm images overlaid on a graph. They are essentially the same as Figure 2.3b except in these images all vertical slices are shown, as opposed to just the central slice.

produced using the rolling window processing method described in section 2.2.4. These doses were plotted against the relative intensities (I_n/I_1) for each dataset and a line of best fit was determined using data where I_n/I_1 > 0.4. $D_{1/2}$ values were then obtained from the line of best fit for each processed beam and plotted against the percentage of pixel values in the image that are equal to 0 as shown in Figure 5.10. The results show that the higher the threshold value (i.e. the more pixels that are considered as background and are set to zero) the higher the $D_{1/2}$ value. This occurs because RADDOSE-3D is run independently from the scaling of the diffraction data and therefore the I_n/I_1 values are the same for each run. As alluded to above, more zero pixel values means that the same total photon flux is distributed over a smaller area (which always contained the crystal), and hence the crystal absorbs more energy which contributes to higher doses. This gives the appearance that the crystal is less radiation sensitive (because the $D_{1/2}$ is higher).

The other major factor in the calculation of the dose values is the calculation of the absorption coefficients. In Figure 5.10, the blue markers correspond to simulations where the absorption coefficients were calculated with RADDOSE version 2 (Paithankar *et al.*, 2009) (denoted RDV2 in Figure 5.10) which gave an absorption coefficient of $4 \times 10^{-4}\ \mu m^{-1}$ (solvent content supplied as 64%). The red markers represent simulations where the absorption coefficients were calculated for an "average protein", $H_{49.8}C_{31.8}N_{8.56}O_{9.54}S_{0.249}$ (denoted *"default"* in Figure 5.10). The absorption coefficient, assuming 50% solvent, is $2.37 \times 10^{-4}\ \mu m^{-1}$ (details of the calculation can be found in Holton and Frankel (2010)). The absorption coefficient of insulin is bigger than that of the average protein because it contains more sulphur than average, and it also contains zinc, which has a relatively high absorption cross section. The difference between the absorption coefficients leads to differences in calculated doses of about a factor of 2.

The results seem to suggest that the two factors investigated here: beam profile and absorption coefficient, are the most important for dose calculations. However, the differences between the profiles of the original beam image, the top hat beam and the deconvoluted beam have a much less pronounced affect on the dose calculations compared to the absorption coefficient. For example, the percentage difference of $D_{1/2}$ between the original beam with no background (denoted $\text{Original}_{\text{NB}}$ in the figure legend of Figure 5.10) and the deconvolved beam with no background is 0.8% (using coefficients calculated using RADDOSE version 2). This shows that deconvolution of the image will not significantly affect the

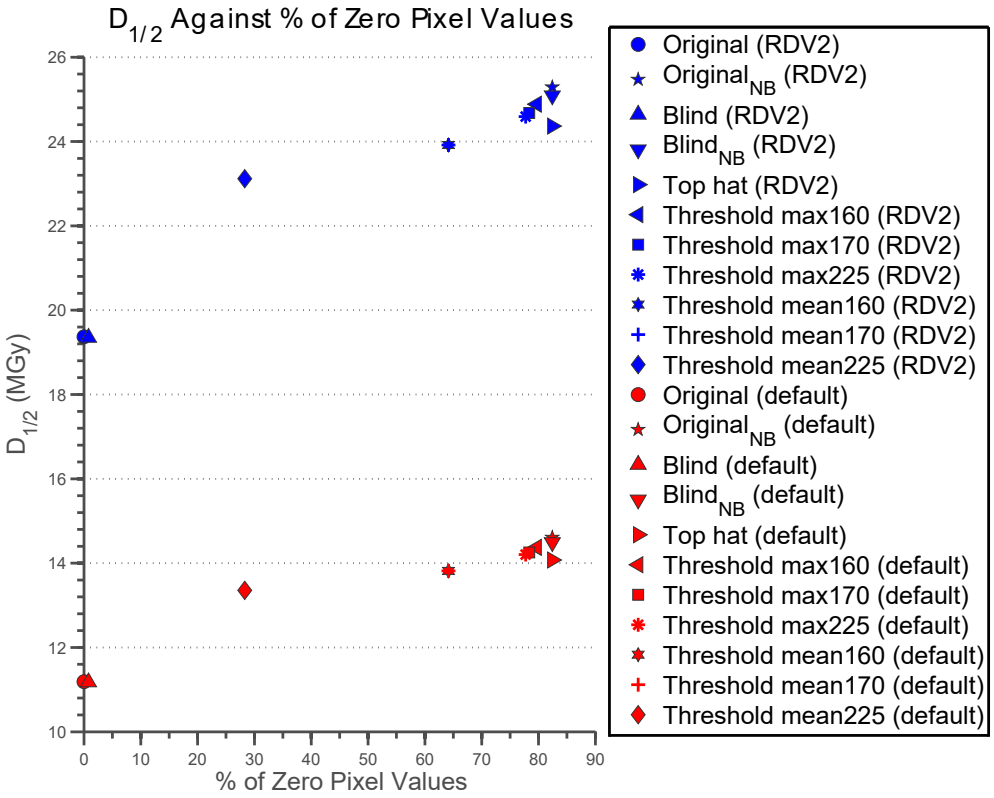


Figure 5.10: $D_{1/2}$ values plotted against the percentage of zero pixel values in the processed pgm beam images. Blue markers correspond to simulations where the absorption coefficients were calculated with RADDOSE version 2 (denoted 'RDV2' in figure legend) and red markers represent simulations where the absorption coefficients were calculated for an "average protein" (denoted 'default' in figure legend). The subscript 'NB' denotes beam profiles where the "background" was removed. The figure shows a trend that the higher the proportion of zero pixel values, the higher the $D_{1/2}$ value. It also shows that the method used to calculate the absorption coefficients can significantly affect the estimated doses, as expected.

dose estimate. This is because the differences in the pixel values are relatively small when compared to the differences one may expect if a Gaussian beam profile is compared with a tophat profile.

Using the threshold method there is a difference in $D_{1/2}$ values ($\approx$ 3.9% - 6.0% depending on the background threshold value) if the maximum pixel value of the background is subtracted from the image or the mean pixel value is subtracted from the image. This is consistent with the observation that the higher the threshold value, the higher the $D_{1/2}$ value. The difference between the maximum threshold value from the 160 $\times$ 160 μm^2 rectangle ($Bkgd_{20}$) and the mean value is 3.89% (using RDV2 calculated absorption coefficients). This result demonstrates that even if the area that is considered background is the same (Figure 5.8), deciding whether to subtract the mean background pixel value or the maximum pixel value will also have a small affect on the resulting $D_{1/2}$ value.

5.4 Processing multiple 1D aperture scan measurements

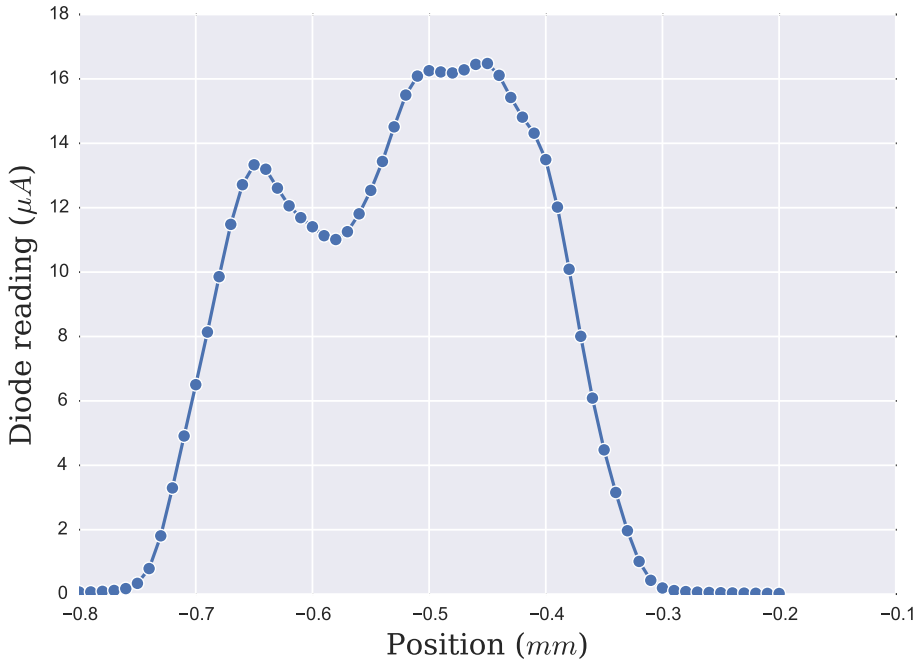
In section 5.2, measurements of the X-ray beam were described which were taken by scanning a 10 μm aperture across the beam horizontally and vertically through the centre of the beam. A 2D beam profile was calculated by fitting a Gaussian function to the separate profiles and substituting those parameter values in the equation of a 2D Gaussian. The implicit assumption made is that the beam profile is similar in every region of space, including the central slices. This is in addition to the explicit assumption that the beam profile is Gaussian. X-ray beam profiles can vary drastically from beamline to beamline even at the same synchrotron, so the assumptions made by the previous method are not generally applicable to all beam profiles. Therefore, a method to generate 2D X-ray beam profiles from multiple 1D aperture scans across the X-ray beam was developed to provide a more generally applicable solution to this problem.

5.4.1 Acquiring the 1D aperture scans

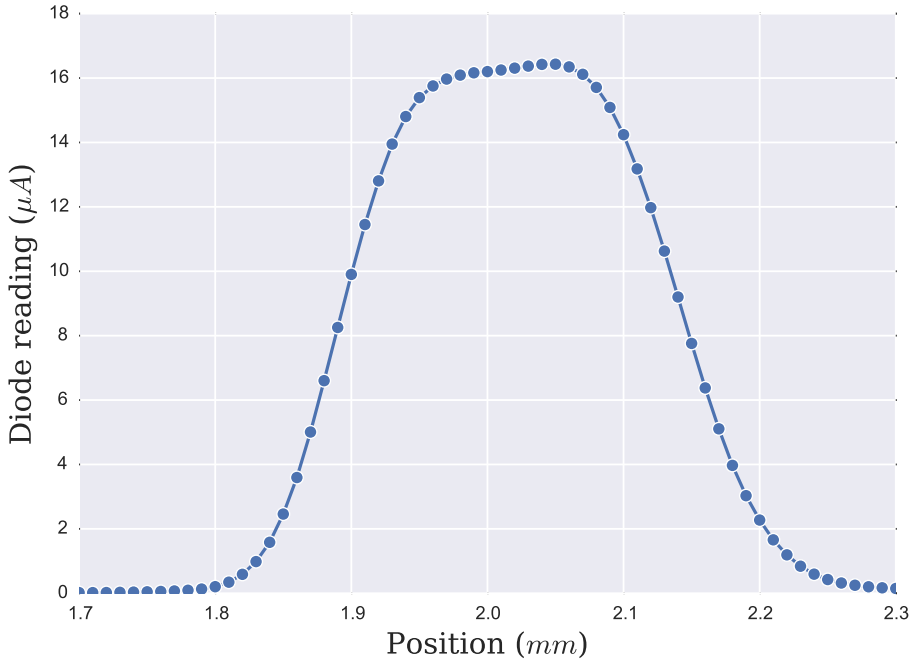
In order to estimate the dose for the SAXS experiments described in chapter 6, aperture scans to obtain X-ray beam flux measurements were carried out in collaboration with Dr. Adam Round and Dr. Martha Brennich on the ESRF BM29 (SAXS) beamline. A 100 μm diameter circular aperture was used to scan across the X-ray beam area and measurements were taken at 10 μm intervals with an OSD1-0 photodiode purchased from Optoelectronics. The scanning was performed 6 times with the collection of 3 horizontal and 3 vertical scans (Figure 5.11).

5.4.2 Creating a 2D beam profile

It is clear from the data shown in Figure 5.11 that fitting a Gaussian shape will not model this beam profile well. Instead a rectangular grid is set up with the edges of the measurement positions in Figures 5.11a and 5.11b used as the boundaries of the grid. The flux at and beyond the grid boundaries is 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



(a)



(b)

Figure 5.11: Diode readings from aperture scans across the beam collected at beamline BM29, ESRF, displayed in Figure 6.9. (a) Vertical scan through the centre of the beam. (b) Horizontal scan through the centre of the beam. 4 other scans (2 vertical, 2 horizontal) were also taken at other locations across the beam.

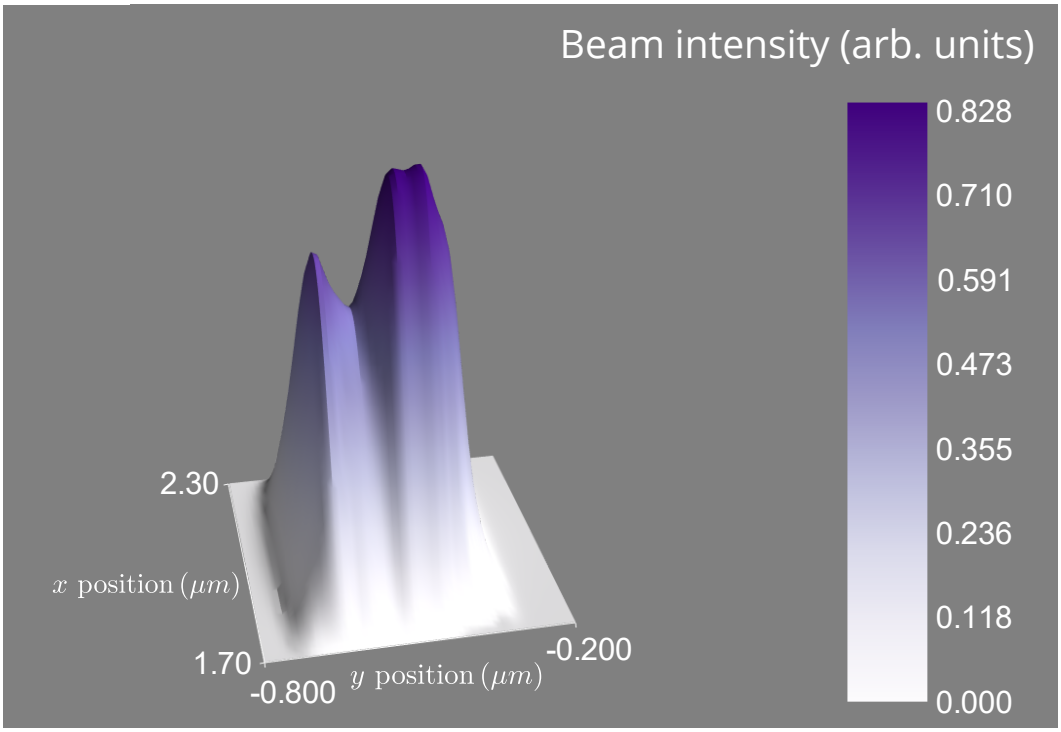
SciPy package in the Python programming language (Jones *et al.*, 2014). The same procedure was performed for the data in the horizontal direction. The results are shown in Figure 5.12. The two interpolated arrays from Figure 5.12 are then averaged to obtain the final beam profile (Figure 5.13).

The final X-ray beam profile in this case is a much truer representation of the overall beam profile than that generated by fitting a Gaussian profile. The main reason is because the measurements are explicitly used as points on the interpolated beam profile grid as opposed to fitting a mathematical function. Secondly, the method uses data from scans taken anywhere in the X-ray beam profile rather than just the central slices. However, the method is not perfect because the final averaging causes a loss of some features. Figure 5.12a exhibits a 'valley' in the beam profile; however after averaging, the valley resembles a 'shoulder' (Figure 5.13a). This could be rectified by applying a weighted averaging scheme where positions that contain measured data in one beam profile array and not the other are up weighted. A more desirable approach would be to use an interpolation routine that could be applied to all measured data in both directions, thus avoiding the need to average at all. It is also important to collect as many aperture scans as possible because the resulting beam profile will be more representative of the true X-ray beam profile, as then more actual values are measured.

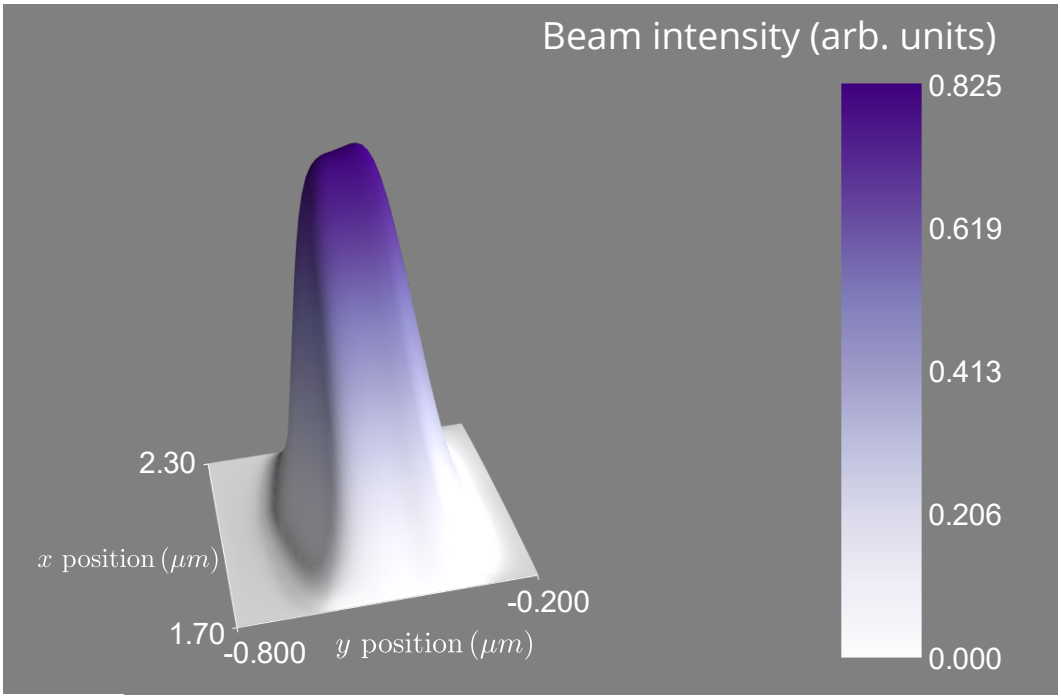
5.5 Discussion

The ability to calculate the dose absorbed by a crystal is vital for the comparison of results from different radiation damage studies and to give guidance to experimenters so MX data collections can be optimised. To ensure that the dose estimates are reliable, it is necessary to accurately parameterise the diffraction experiment. A custom module was written into RADDOSE-3D by Dr. Oliver Zeldin to allow the experiment to be simulated with experimentally measured X-ray beam profiles (Zeldin *et al.*, 2013a). However the experimentally measured beam profiles often need some form of preprocessing before they can be used in the simulation. The type of preprocessing largely depends on the experimental method used to measure the beam profile.

One method to measure the beam profile is to use an aperture to scan across the X-ray

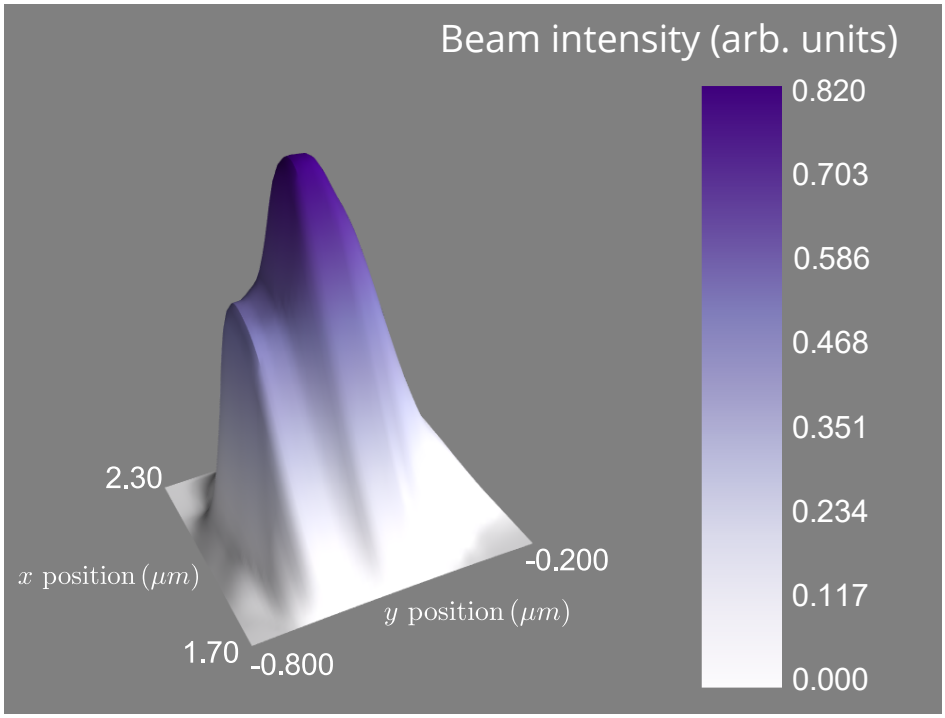


(a)

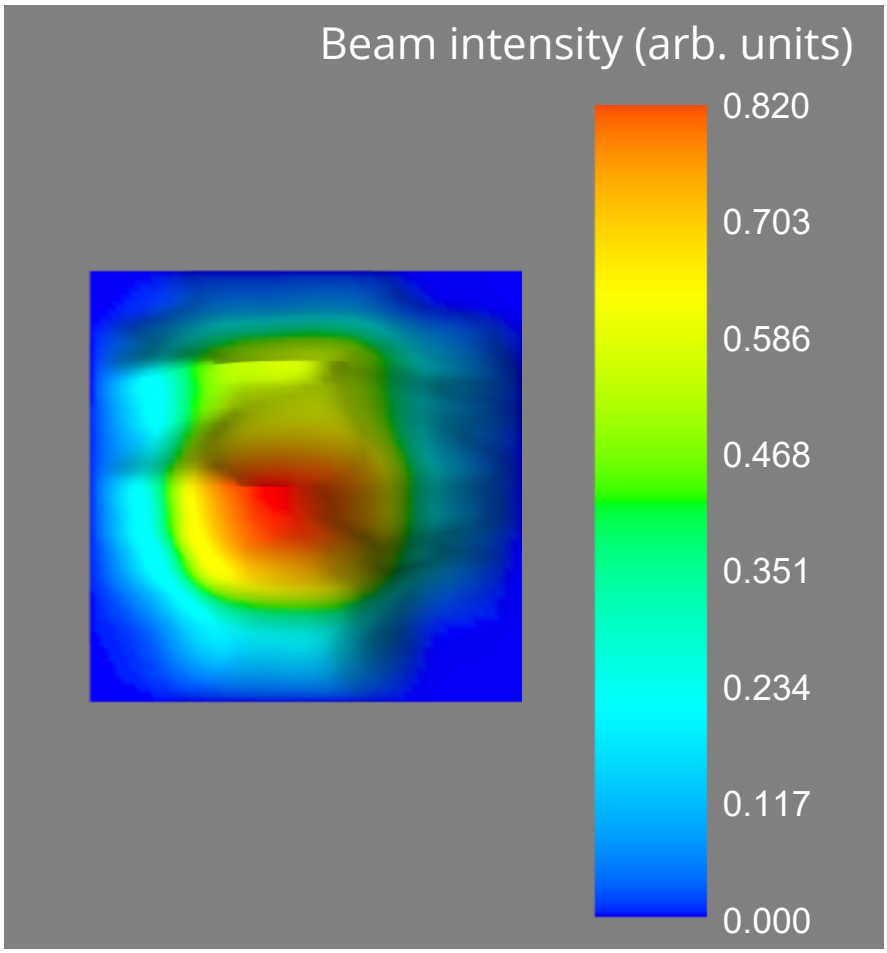


(b)

Figure 5.12: Interpolation of diode readings from the aperture scans. (a) Interpolation of vertical scans. (b) Interpolation of horizontal scans.



(a)



(b)

Figure 5.13: (a) Final X-ray beam profile generated from the average of the two beams in Figure 5.12. (b) Overhead view with a rainbow colour scheme.

beam and measure the current in a silicon diode produced by the X-ray photons at each position. Often a single aperture scan is performed horizontally and again vertically. If the scans resemble Gaussian profiles (Figure 5.1) then it is possible to use the parameter values obtained from 1D Gaussian fits to the data in a 2D Gaussian model representation of the beam. This approach enabled accurate predictions of the data that would be generated using apertures of various sizes, hence validating the method. The model also highlighted another important aspect. It showed that the FWHM values of the true beam are very likely to be close to the FWHMs measured from the diode readings of the aperture scans. The implication of this result for modelling is that deconvoluting (with subsequent smoothing of the noise) the X-ray beam profile from the aperture size does not lead to a significantly different beam profile. However, more experiments with different aperture and beam sizes may need to be performed to verify the generality of this result.

A different method of measuring the beam profile involves taking a 2D image of the beam using a scintillator and a camera. The advantage of this method is that there is no need to explicitly model the 2D beam profile. However, the drawback is that regions of the image that represent background (i.e. areas of zero flux from the X-ray beam) have non-zero pixel values. Several beam processing methods including deconvolution, image segmentation and standard average subtraction, were applied to remove background. It was found that the various methods gave similar results in terms of the calculated half dose ($D_{1/2}$) values i.e. the resulting beam profiles following the processing did not affect the dose values. The biggest factor affecting the dose values is the proportion of the image area that corresponds to background. This suggests that to get an accurate beam profile from a 2D image it is necessary to know the collimation of the X-ray beam so that a suitable background region can be masked. It is important that the scintillator is placed at the sample position because beam divergence will result in a slightly different beam profile if the scintillator is placed in a different plane.

A more reliable (and thus recommended) method to measure the experimental beam profile is to take a series of 1D aperture scans to sample as much of the beam profile region as possible. This bypasses the need to subtract background from an image because flux is measured via the current it produces on the silicon diode. Sampling the 2D space allows the beam to be interpolated between data points. This negates the requirement to explicitly model the beam profile with a 2D Gaussian function for example. In the work presented

here, the interpolation was performed separately on horizontal scans and vertical scans and then averaged. This method still results in a loss of some features as evidenced by lack of the “valley” in the final averaged beam profile, even though it was clearly present in the vertical scans. To avoid the loss of features, in the future it would be desirable to perform the spline interpolation on all data, as opposed to the vertical and horizontal data separately.

CHAPTER 6

Methods to Assess Radiation Damage in SAXS

6.1 Introduction

As outlined in the introduction of this thesis, SAXS is a method to obtain a molecular envelope of a macromolecule in solution. However, during X-ray exposure the sample undergoes radiation induced changes which progress throughout the experiment. Ultimately these changes lead to aggregation of the molecules in the sample which is observed as a progressive dissimilarity of subsequent 1D scattering curves (frames).

This chapter presents a quantitative study of the possible mitigation of radiation damage by addition of various radioprotectant compounds in SAXS experiments. The compounds were chosen for several reasons:

- they have shown good protection ability in MX (Allan *et al.*, 2012) and (Southworth-Davies and Garman, 2007).
- they are commonly used already as radioprotectants in SAXS (Grishaev, 2012).
- the majority of their atomic constituents have relatively small atomic numbers and hence they will not significantly contribute to the dose absorbed by the sample.

6.2 Extending RADDOSE-3D for SAXS

To perform comparative analysis of the extent of radiation damage in SAXS experiments it is necessary to calculate the dose absorbed by the sample. Currently dose estimates in SAXS are calculated using

$$D = \frac{tfE(1-T)}{V\rho}, \tag{6.2.1}$$

where D is the dose, t is the exposure time, f is the X-ray flux, E is the energy of the incident photons, T is the transmission factor, V is the illuminated volume and ρ is the mass density of the sample (Meisburger *et al.*, 2013; Jeffries *et al.*, 2015). To simplify the calculation of the dose, the explicit geometry of the sample is not taken into account. A more accurate calculation can be performed if the SAXS experiment is simulated in three dimensions, providing a spatially resolved dose field. This dose field can then be interpreted with the various dose metrics already developed for MX (Zeldin *et al.*, 2013a, 2012).

This section presents the extensions written into RADDOSE-3D to perform simulations of SAXS experiments for improved dose calculations.

6.2.1 RADDOSE-3D architecture

At its core, RADDOSE-3D takes a description of a crystal and exposes it to a computational beam model via a user specified exposure strategy (wedge) (Zeldin *et al.*, 2013b). The crystal is computationally modelled as a 3D voxel grid, where each grid element stores information about x, y, z coordinates, the dose and the fluence. The dose is calculated using the beam intensity at the leading edge of the voxel and the proportion of the beam that is absorbed by the voxel, which is governed by the absorption coefficient. The beam is described by the intensity profile, photon energy and the total flux. Finally, the wedge specifies the total angular rotation of the sample, exposure time and any rotation offsets or translations. The structure of the program is illustrated in Figure 6.1. A single RADDOSE-3D job performs the MX simulation on a single crystal, but it can be exposed to multiple beams with multiple wedges. Output files giving information about the crystal state are updated after the crystal is exposed to a beam via a single wedge.

At an abstract level, the crystal is defined by several properties that are typically associated with crystals in MX. These include the unit cell volume and the number of molecules in the unit cell. RADDOSE-3D uses this information to determine the composition of the crystal so that an absorption coefficient can be calculated. However, there is no reason that the composition has to be defined by information specific to the crystal, in which case it is more appropriate to refer to what has been termed the "crystal" so far as the "sample". If the sample is a liquid, as is the case in SAXS experiments, then the sample composition should be determined with different inputs. The modular design of RADDOSE-3D enables this type of extension into be easily incorporated to the existing functionality.

6.2.2 Cylindrical sample geometry

In a typical SAXS experiment, liquid samples are generally contained in, or flowed through, a cylindrical capillary during the X-ray exposure. Therefore it is necessary for RADDOSE-3D to be able to model cylindrical sample shapes. RADDOSE-3D had already been extended

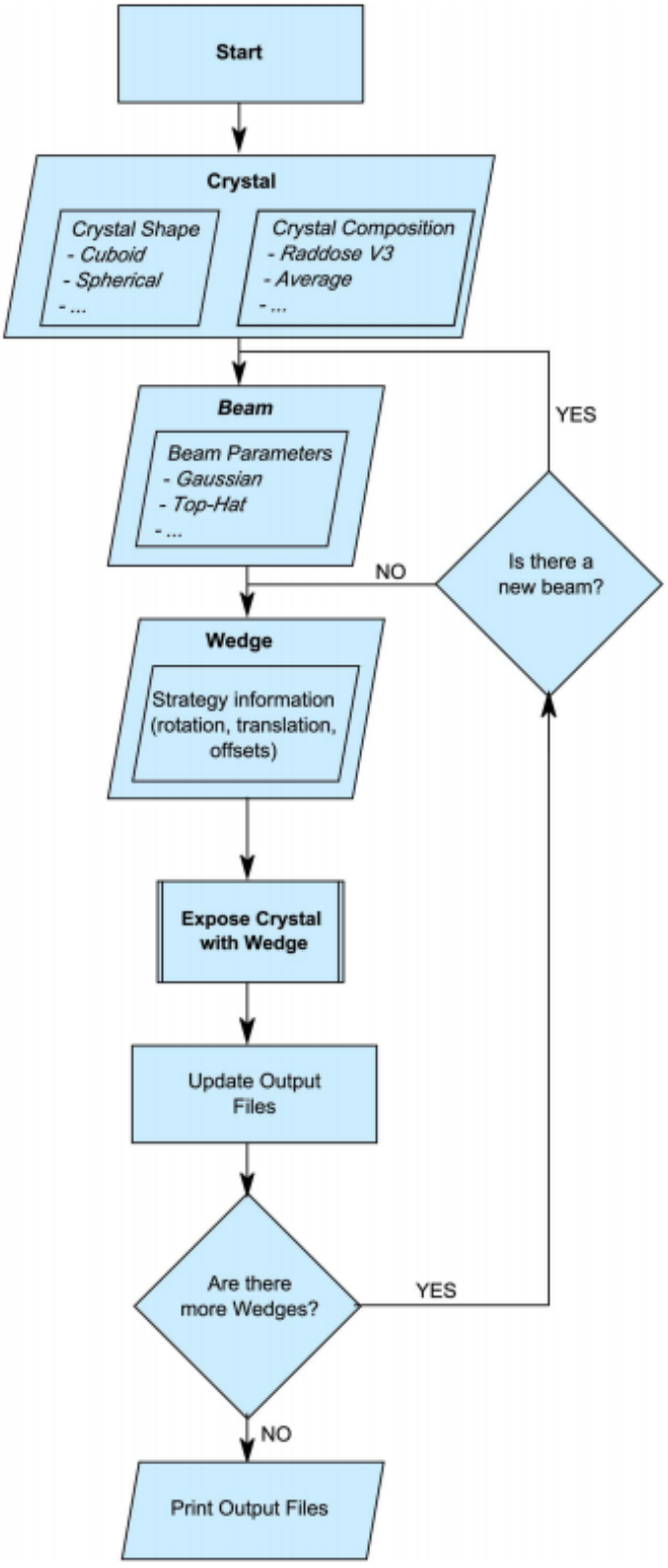


Figure 6.1: Flow chart illustrating the structure of the RADDOSE-3D code. Reproduced from (Zeldin *et al.*, 2013b).

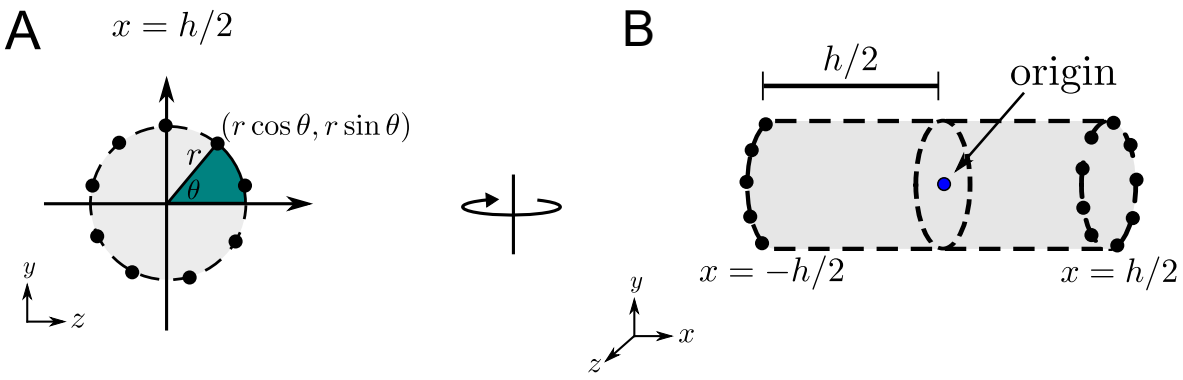


Figure 6.2: Implementation of the cylindrical sample geometry in RADDOSE-3D given user defined diameter, d, and height, h. (A) evenly spaced points around a circle are generated given the radius $r = d/2$ of the circular cross-section. RADDOSE-3D defaults to 32 points. (B) In three dimensions the points 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. The connectivity of these vertices is hard-coded into the RADDOSE-3D source code.

to handle polygonal shapes (Bury *et al.*, 2015), which meant that it was already capable of modelling cylindrical shapes, since any 3D shape can be modelled by a series of polygons, However, the implementation for polyhedral crystals requires the user to define the sample geometry in a non-user friendly manner. A file specifying the geometry of the shape using a set of vertex positions, and their connectivity is required. This is much more complex than defining the diameter of the circular cross section and the length/height of a cylinder. Therefore a module was written that accepts a diameter and height as input and converts this into a polyhedral description (a set of vertices and faces) of a cylinder within RADDOSE-3D.

The cylindrical implementation, which specifies the geometry of the sample alone, not the capillary in which it is contained, is graphically depicted in Figure 6.2 (the effect of the capillary is dealt with separately in section 6.2.4). First the points around a circle are generated using the user defined diameter of the circular cross section. RADDOSE-3D uses 32 points around the circle by default, no matter what dimensions are specified. 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 done using the final coordinate x. Depending on which end a particular point is, 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 this assumes the origin of the system is located at the centre of the cylinder.

Regardless of the dimensions of the cylinder, the connectivity of the vertices remains the

same because the number of vertices and their orientation with respect to one another is constant. Therefore the connectivity has been hard-coded into RADDOSE-3D. It will only ever need to be changed if the number of points is altered. However, this parameter is not exposed to the user and hence it would only change if a developer modified the source code.

The geometry defined to create the cylinder is rotated 90° about the y-axis compared to the usual RADDOSE-3D simulation geometry. This means that the beam would irradiate the sample along the x-axis (or directly into the page looking at Figure 6.2 A). In a typical SAXS experiment the beam direction is along the z-axis defined in Figure 6.2 (i.e. perpendicular to the axial dimension of the cylinder). So whenever a user specifies a SAXS experiment in RADDOSE-3D, the sample is rotated by an additional 90° on the angle which the user specifies as the initial orientation of the sample to the beam (the sample geometry does not have to be specified as cylindrical).

6.2.3 Determining the sample composition

Knowledge of the atomic composition of the sample is necessary to be able to calculate the dose absorbed upon X-ray irradiation. This is because the overall absorption coefficient of the sample, μ_{abs}, is calculated from the individual atomic absorption coefficients, σ_j as

$$\mu_{\text{abs}} = (1/V_c) \sum_{j=1}^{N} \sigma_j, \tag{6.2.2}$$

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). (The previous RADDOSE versions assumed that the absorption coefficient was equal to the attenuation coefficient, $\mu_{\text{att}} i.e. \mu_{\text{abs}} = \mu_{\text{att}}$, so equation 1 in Murray *et al.* (2004) , which is analogous to equation 6.2.2 here, writes 'μ_{att}' instead of 'μ_{abs}'). In MX the crystal composition is calculated from the contents of the unit cell. In SAXS the samples are liquids as opposed to crystals, and hence the notion of a unit cell does not apply. Thus instead, the approach to determine the atomic composition of the sample is to define a volume of liquid and estimate the contents given its protein concentration and buffer composition.

First the molarity of the solution is calculated using the formula

$$\text{Molarity (moles/litre)} = \frac{\text{sample concentration (grams/litre)}}{\text{molecular mass (grams/mole)}}. \tag{6.2.3}$$

The sample concentration is provided by the user in units of grams per litre ($\equiv$ mg/ml). The molecular mass of the molecule is calculated from other parameters provided in the user input. If the sequence file is given for the protein (the sample can also contain DNA and RNA) then the molecular mass can be determined accurately by summing the molecular mass of each residue in the file. Otherwise an average molecular weight is used for each residue (110 Da for protein residues, 339.5 Da for RNA residues and 327 Da for DNA residues) and the user has to specify the type and number of residues for the sample.

Secondly, a suitable volume needs to be specified to calculate the atomic composition. A suitable volume is one that is large enough to contain at least one complete molecule. By default this volume is defined to be (1000 Å)^3 but this can be changed by specifying the length, width and height dimensions of the volume in the input file using the *unit cell* input keyword.

The number of monomers/molecules in the volume can now be calculated by multiplying the molarity, volume (converted to litres) and Avogadro's number (N_A = 6.022 $\times$ 10^{23} mol^{-1}), which is then rounded to the nearest integer. The absorption coefficient can then be computed in the usual way as is described in Paithankar *et al.* (2009). If there is less than 1 molecule in the volume then this is flagged up and the user is advised to increase the volume.

6.2.4 Attenuation of X-ray flux due to capillary

In a typical MX experiment, a crystal is exposed directly to the X-ray beam. In contrast, samples from SAXS experiments are held inside a quartz capillary. This means that the attenuation of the X-ray flux due to the capillary needs to be taken into account 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

$$I/I_0 = \exp\left[-(\mu/\rho)x\right], \tag{6.2.4}$$

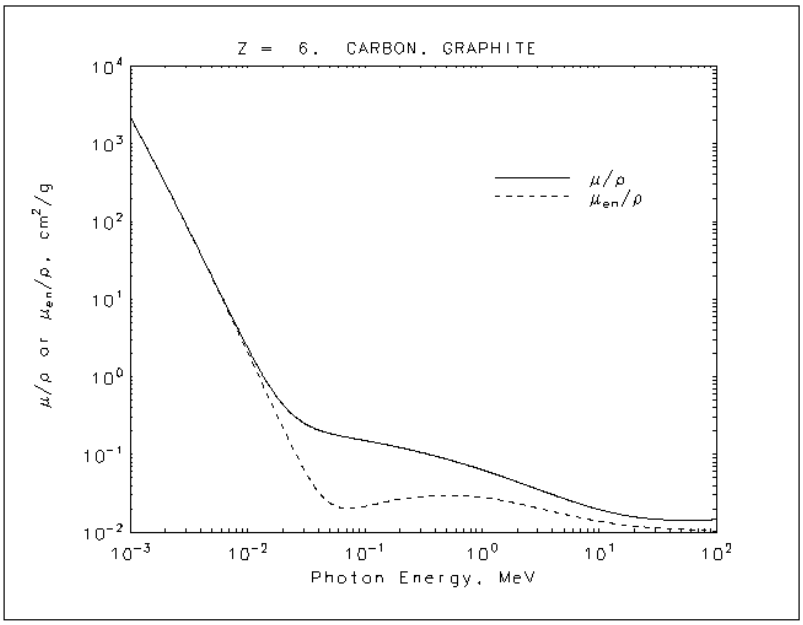
where I is the emergent intensity of the beam after penetrating the material, I_0 is the incident intensity and μ/ρ is defined as the mass attenuation coefficient (Hubbell and Seltzer, 1995). The mass thickness, x, is defined as the mass per unit area and is given by $x = \rho t$ where t is the thickness of the material. The attenuation fraction by the capillary can hence be calculated as $1 - I/I_0$.

RADDOSE-3D requires the user to supply the thickness, density and material composition of the capillary to calculate the attenuation fraction. The mass thickness can be directly calculated using the density and thickness as described above. The mass attenuation coefficient is dependent on the atomic composition of the material as well as the energy of the incident photons. The relevant values can be found online via the National Institute of Science and Technology (NIST) tables (Hubbell and Seltzer, 1996a,b). A section of the webpage for the mass attenuation coefficient table of carbon is shown in Figure 6.3. The tabulated energy values do not explicitly include the typical energies used in crystallograpy and SAXS (around 12 keV), so the mass attenuation coefficient is linearly interpolated between the closest values. Mass attenuation coefficient data for various mixtures including air, borosilicate glass, water, bone and soft tissue are also tabulated in NIST table 4 (Hubbell and Seltzer, 1996b). These mixtures can be used directly by RADDOSE-3D.

When mixtures of atomic species are used (the most commonly used capillary material in SAXS is quartz which has elemental composition SiO_2) the mass attenuation coefficient is obtained using a weighted average given by

$$\mu/\rho = \sum_i w_i(\mu/\rho)_i, \tag{6.2.5}$$

where w_i and $(\mu/\rho)_i$ are the fraction by weight and mass attenuation coefficient of the i^{th} atomic constituent respectively.



Carbon, Graphite
***Z* = 6**

HTML table format

Energy (MeV)	μ/ρ (cm^2/g)	μ_{en}/ρ (cm^2/g)
1.00000E-03	2.211E+03	2.209E+03
1.50000E-03	7.002E+02	6.990E+02
2.00000E-03	3.026E+02	3.016E+02
3.00000E-03	9.033E+01	8.963E+01
4.00000E-03	3.778E+01	3.723E+01
5.00000E-03	1.912E+01	1.866E+01
6.00000E-03	1.095E+01	1.054E+01
8.00000E-03	4.576E+00	4.242E+00
1.00000E-02	2.373E+00	2.078E+00

Carbon, Graphite
***Z* = 6**

ASCII format

Energy (MeV)	μ/ρ (cm^2/g)	μ_{en}/ρ (cm^2/g)
1.00000E-03	2.211E+03	2.209E+03
1.50000E-03	7.002E+02	6.990E+02
2.00000E-03	3.026E+02	3.016E+02
3.00000E-03	9.033E+01	8.963E+01
4.00000E-03	3.778E+01	3.723E+01
5.00000E-03	1.912E+01	1.866E+01
6.00000E-03	1.095E+01	1.054E+01
8.00000E-03	4.576E+00	4.242E+00
1.00000E-02	2.373E+00	2.078E+00
1.50000E-02	8.071E-01	5.627E-01
2.00000E-02	4.420E-01	2.238E-01

Figure 6.3: Section of the X-ray mass attenuation coefficient data for carbon from NIST (Hubbell and Seltzer, 1996a). The mass attenuation coefficient data are tabulated beneath the graph. The exact energy of the X-ray photons used in crystallography and SAXS (typically around 12 keV) is not explicitly tabulated therefore the mass attenuation coefficient is obtained by linear interpolation.

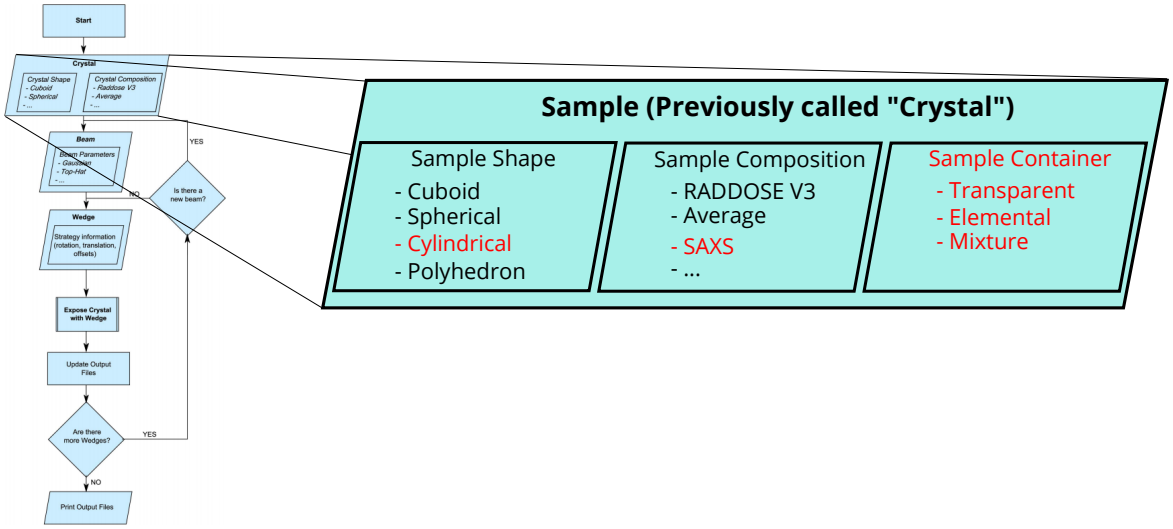


Figure 6.4: RADDOSE-3D flowchart from Figure 6.1 with the updated sample implementation, which extends the previous "crystal" implementation. The text coloured red show the SAXS extensions covered in this section.

6.2.5 Summary of SAXS extensions

Figure 6.4 puts into context how the SAXS extensions fit into the RADDOSE-3D program design. All of the changes have been made in the "crystal" definition which has been referred to as the *sample* in this section.

Figure 6.5 shows explicitly how the extensions described above combine to allow accurate dose calculations for SAXS experiments. The cylindrical sample geometry is defined first from the specified height and diameter of the sample in the capillary. Then the atomic sample composition is defined from the protein concentration and buffer components. Once the container material is specified, the attenuated X-ray beam flux can then be calculated. Finally the SAXS sample is exposed to the attenuated X-ray beam according to the 'wedge' parameters.

6.2.6 Model considerations

The model of the SAXS experiment in RADDOSE-3D makes many implicit assumptions. For instance, with regards 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 are 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

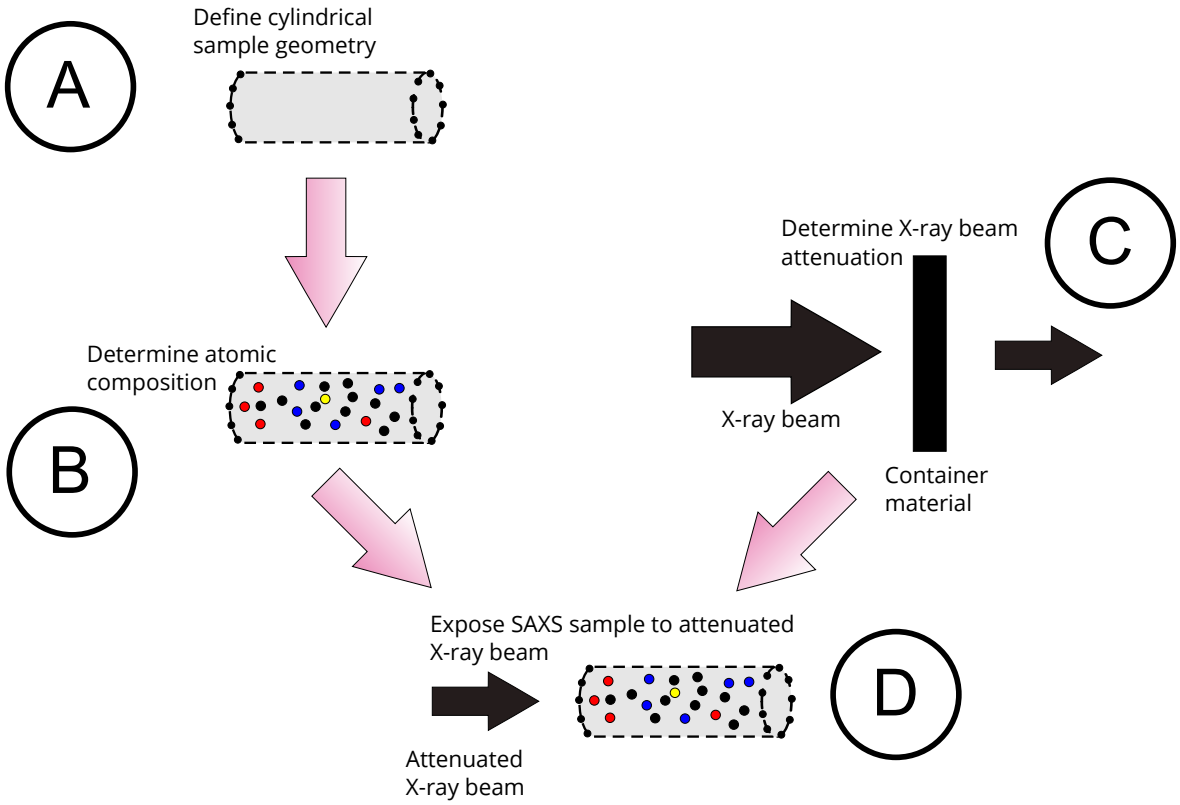


Figure 6.5: Flow diagram showing how the extensions described in this section are combined to enable dose calculation. A: Define the cylindrical geometry of the SAXS sample (section 6.2.2). B: Determine the atomic composition (section 6.2.3). C: Calculate the attenuation of the beam due to the sample container (section 6.2.4). D: Expose the SAXS sample to the attenuated beam according to the defined wedge.

of any rotations or translations.

The sample itself is assumed static, moving as a rigid body when rotated or translated, and completely filling the capillary. This greatly reduces the computational cost when compared to the possibility of modelling the exact 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 is not always valid, especially when the sample is flowed through the capillary. Hopkins and Thorne calculated that for typical experimental parameters (capillary diameters: 1.5-2 mm, flow velocities: 0.5-15 mm s^{-1}) the Reynolds numbers* are in the range 1-25 implying a laminar flow regime (Hopkins and Thorne, 2016). In this case 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). Furthermore, Hopkins and Thorne also calculated that the residence times of the sample in the beam were too short for any appreciable radial diffu-

*The Reynolds number is defined as the ratio of inertial forces to viscous forces and is a quantity used in fluid mechanics to determine properties of fluid flow (Purcell, 1977).

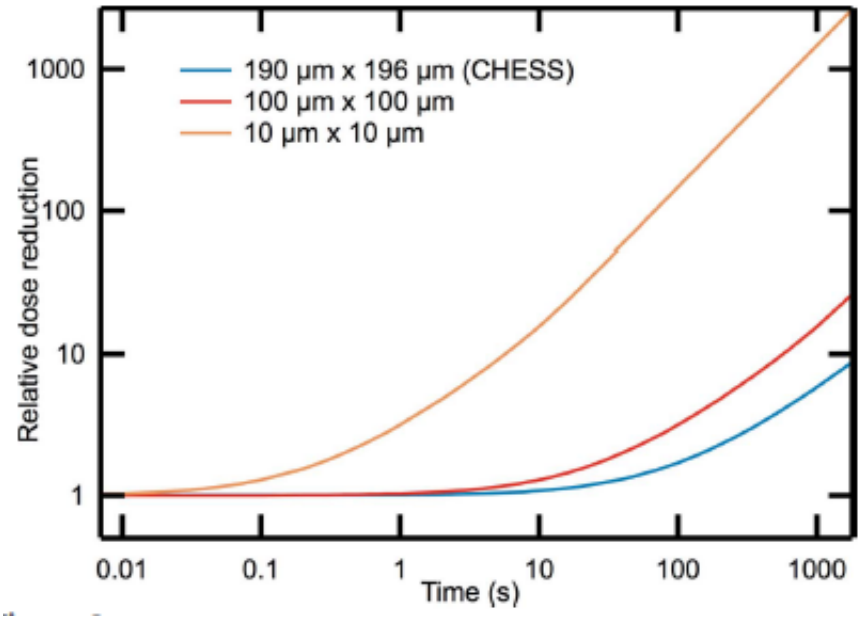


Figure 6.6: Reduction in the dose absorbed by the sample due to diffusive exchange of lysozyme against time for three beam sizes. Extrapolation of these curves for a beam size of 600 $\times$ 600 μm^2 along with 120 second exposure times suggests that the dose reduction due to diffusive turnover is negligible for the experiment reported in this chapter. Figure reproduced from (Hopkins and Thorne, 2016).

sive mixing, hence the flow profile results in radius-dependent residence times in the X-ray beam. Therefore the static assumption here will give misleading dose values if calculated for experiments where the sample is flowed through the beam.

Hopkins and Thorne additionally discuss diffusive turnover as a phenomenon that will affect the dose calculation (Hopkins and Thorne, 2016). 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. In the current work, no account was taken for molecular diffusion. Extrapolation of the plot from (Hopkins and Thorne, 2016) (reproduced in Figure 6.6) shows that diffusion is likely to be negligible for the experiment performed here (section 6.3) where the beam size was 600 $\times$ 600 μm^2 and the total exposure time was 120 seconds.

The SAXS sample can also be manipulated in the same way as a crystal in RADDOSE-3D. Hence helical scanning, translations and rotations of the sample can all be performed. Figure 6.7 is the final dose state of a glucose isomerase sample in an experiment where the sample was rotated 180° for a total of 200 seconds in a 700 $\times$ 700 μm^2 top-hat profile beam with a flux of 1.51 $\times$ 10^{13} ph/s and an incident photon energy of 12.1 keV.

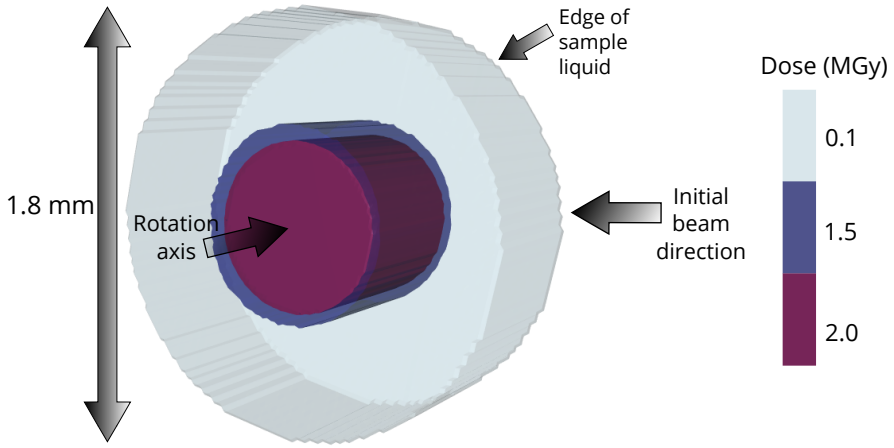


Figure 6.7: Final dose state of a glucose isomerase sample in an experiment where the sample was rotated 180° in a 700 μm × 700 μm top-hat profile beam with a beam flux of 1.51 × 10^{13} ph/s and an incident energy of 12.1 keV for a total exposure time of 200 seconds. The capillary was treated as completely transparent to the X-rays so there was no attenuation. The dose state is only calculated and shown for the sample, not the capillary. The contours correspond to dose iso-surfaces: light blue = 0.1 MGy, dark blue = 1.5 MGy and purple = 2 MGy.

6.3 Experimental methods

This section describes the experimental details for two experiments comparing the efficacy of various radioprotectant compounds. The first of these experiments was carried out in October 2014 whereas the second was carried out in December 2015, both at the ESRF, in Grenoble.

6.3.1 Sample preparation

Crystalline glucose isomerase (GI) used in both experiments was purchased in tetrameric form (1552 residues, 172 kDa) from Hampton Research. GI was chosen because it is a stable, soluble globular protein that has well defined SAXS behaviour and is sufficiently large to scatter well at modest concentrations.

Experiment 1 (Expt 1)

Radioprotectants and GI were dissolved separately in buffer (100 mM HEPES and 10 mM $MgCl_2$ at pH 7.0) to give stock solutions of both at twice the desired radioprotectant and protein concentrations respectively. The protein and scavenger stocks were then mixed thoroughly by pipetting in a 1:1 ratio to create the samples. The radioprotectant stock was

also mixed in a 1:1 ratio with the buffer by the automatic liquid handling on the beamline to create the buffer sample. Both samples were prepared at the beamline immediately before the diffraction experiment in order to minimise protein aggregation before data collection.

The final concentrations of each component in solution were as follows: 0.54 mg/ml GI, 5 mM soluble radioprotectant, 5% v/v glycerol or ethylene glycol. Each sample was prepared in triplicate to allow repeats. The following radioprotectants were investigated: sodium ascorbate, sucrose, sodium nitrate, trehalose, ethylene glycol, (2,2,6,6-tetramethylpiperidin-1-yl)oxyl (TEMPO), glycerol and glycerol + nitrate.

Experiment 2 (Expt 2)

GI was dissolved and dialysed for 24 hours at 277 K with the same buffer components as those used in the first experiment. The final GI concentration, 1 mg/ml, used for all data collection runs was determined using the extinction coefficient given by 45,600 $M^{-1}cm^{-1}$ at 280 nm absorbance. Eight solution additives were tested 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 solutions at four different concentrations: 10 mM, 5 mM, 2 mM and 1 mM, except 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.

6.3.2 SAXS data collection

Experiment 1

Data collection was carried out by Dr. Ed Lowe at the ESRF on beamline BM29. The X-ray photon energy was 12.5 keV (wavelength of 0.9919 Å), with a flux of 4.84×10^{11} photons per second at 100% transmission. The size of the beam was $700 \times 700\,\mu m^2$, however the beam profile was not experimentally measured. The profile was assumed Gaussian for the purposes of dose calculation (section 6.3.3). Using the EMBL sample loading robot, 15 μl of sample was loaded into a 1.8 mm diameter quartz capillary tube with a wall thickness

of 0.03 mm and data were collected at 293 K with flow mode turned off. Frames were collected at 0.5 second intervals on a Pilatus 1M detector for durations of either 30 seconds (60 frames) or 60 seconds (120 frames).

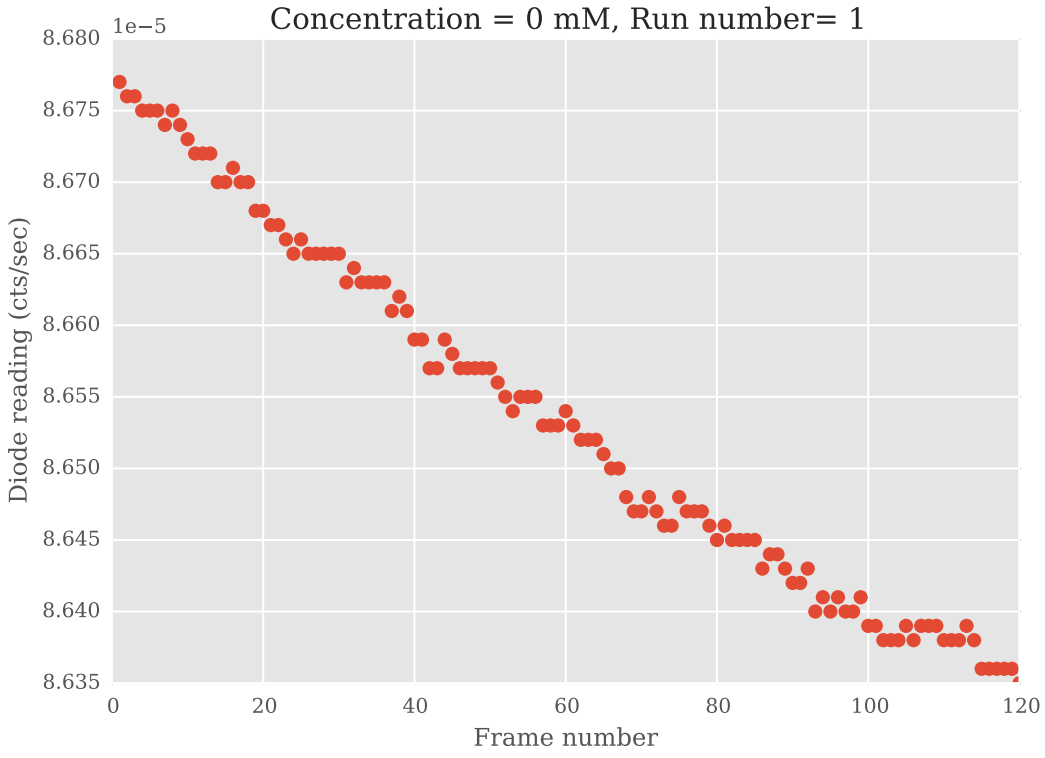
Experiment 2

Data collection was performed at the ESRF by the author on beamline BM29 in collaboration with Adam Round and Martha Brennich. The photon energy used throughout was 12.5 keV and the photon flux was estimated from diode readings which were recorded for every frame using the conversion formula

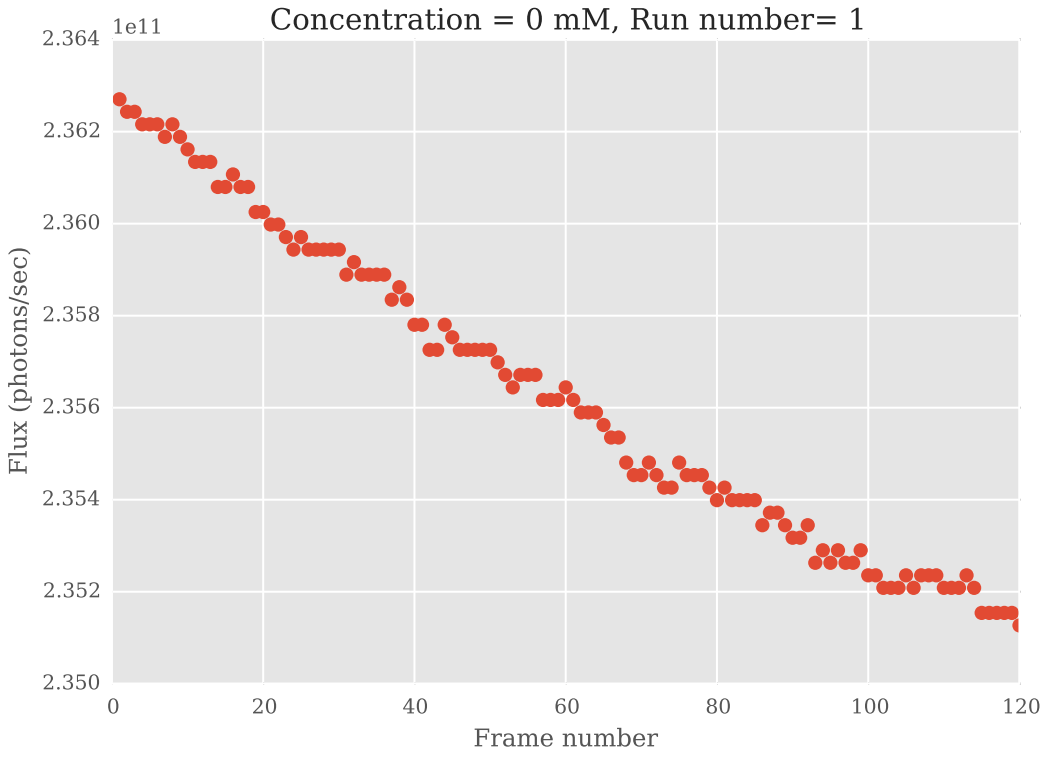
$$\text{flux} = 5.72293 + 2.72295 \times 10^{15} \times \text{d}_\text{r}, \tag{6.3.1}$$

where d_r is the diode reading. The flux obtained using this formula was calibrated with an OSD1-0 photodiode purchased from Optoelectronics, which was a 500 μm thick silicon diode with a 1 mm^2 active area as per Owen *et al.* (2009) . The flux was calculated for each frame because the diode readings constantly changed between frames (Figure 6.8), however the overall change in diode reading was only 0.54%. Despite the small percentage change, account was still taken for this effect in the analysis. The full 2D X-ray beam profile was determined as outlined in section 5.4. The resulting beam is shown in Figure 6.9 as a greyscale image.

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 room temperature using the automated robotic sample changer (Round *et al.*, 2015). For each additive, data were collected at each concentration (given in section 6.3.1), and each of these runs was repeated 3 times. The exposure time for each frame was 1 second, and a total of 120 frames were collected for a single run with the sample kept static. For each radioprotectant concentration a single dataset was collected with only the buffer (no protein) so that a suitable buffer correction (subtraction) could be applied during data analysis.



(a)



(b)

Figure 6.8: Diode readings and flux estimates during the first SAXS repeat for the GI sample with no radioprotectant added (hence concentration = 0 mM). (a) Diode readings for each frame in the experiment. It can clearly be seen that the diode readings decrease throughout the experiment, which is due to the decay of the electron storage ring current. However, the total change during the course of the dataset is only 0.54%. (b) Flux estimates for each frame in the same experiment as (a). As a result of the decreasing diode readings the corresponding flux decreases by 0.54% too.

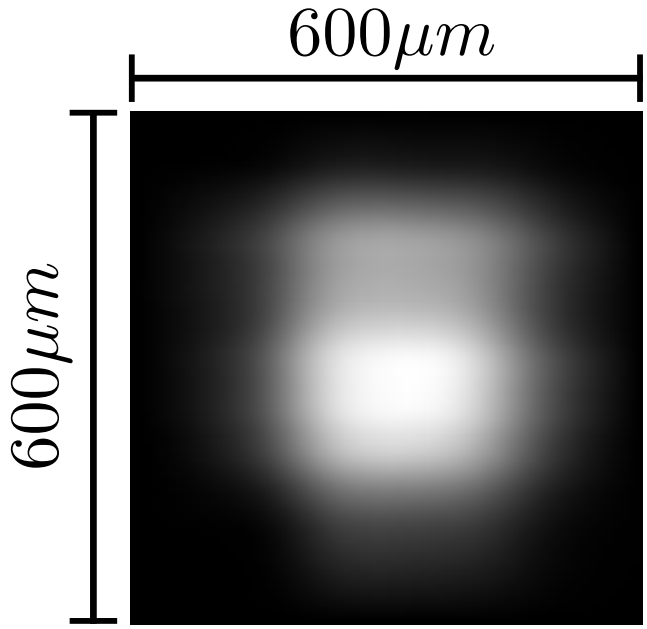


Figure 6.9: A 2D reconstruction of the beam used in the second SAXS experiment (Expt 2) shown as a greyscale image.

6.3.3 Dose calculations

Dose calculations for both experiments were performed using RADDOSE-3D with the modifications described in section 6.2 for modelling SAXS experiments. An example input file for Expt 1 can be seen in Figure 6.10. The beam profile was assumed Gaussian with full width at half maximum of 110 $\mu m \times 200\ \mu m$. For Expt 2, the beam image shown in Figure 6.9 was read directly into RADDOSE-3D and the corresponding flux values recorded for each frame were used to ensure accurate dose estimates. The atomic composition is the same as that defined in Figure 6.10, although the 5 mM concentration of sodium is only defined for the radioprotectants that contain sodium (i.e. sodium ascorbate and sodium nitrate). DTT contains a sulphur atom for which account must be taken also in the solvent concentration. The elemental composition of the quartz capillary was entered into RADDOSE-3D as SiO_2. The capillary thickness was given as 50 μm and a density of 2.648 g/cm^3 was used.

6.4 1D scatter curve similarity analysis

To increase the signal to noise ratio when processing data in SAXS, it is necessary to merge 1D intensity curves from several frames. If the scattering for different frames is from the

```
##############################################################################
#                                Crystal Block                               #
##############################################################################
Crystal
Type Cylinder
Dimensions 1740 1000                  #Diameter and height of sample cylinder
PixelsPerMicron 0.005
AbsCoefCalc SAXS
NumResidues 1552
ProteinHeavyAtoms S 32
ProteinConc 0.54                      #Protein concentration
ContainerMaterialType Elemental
MaterialElements Si 1 O 2             #Elemental composition of quartz
ContainerThickness 50
ContainerDensity 2.648
SolventHeavyConc S 100 Mg 10 Cl 10 Na 5 #Heavy atoms in solvent

##############################################################################
#                             Beam Block                                     #
##############################################################################
Beam
Type Gaussian
Flux 4.84e+11
Energy 12.5
FWHM 110 200
Collimation Rectangular 700 700

##############################################################################
#                             Wedge Block 1                                  #
##############################################################################
Wedge 0 0
ExposureTime 1

##############################################################################
#                             Wedge Block 2                                  #
##############################################################################
Wedge 0 0
ExposureTime 1
```

Figure 6.10: RADDOSE-3D input file used for the dose calculations in Expt 1.

same molecule in the sample, then the frames will be similar i.e. the frames will overlap. However as radiation damage progresses during the experiment, the molecules begin to aggregate, which causes the intensity curve to increase at low q angles and is speculated to decrease at high q angles (Hopkins and Thorne, 2016). On the other hand, fragmentation and molecule repulsion due to protein charging, cause a decrease in scattering at low q angles and an increase at high q angles. Figure 6.11 shows the scattering profile of several SAXS frames collected from the same glucose isomerase sample. As the dose increases, there is a clear decrease in the scattered intensity at low angles suggesting that the sample is undergoing fragmentation. This was also observed with the glucose isomerase samples used in the study conducted by Hopkins and Thorne (Hopkins and Thorne, 2016). Therefore it is necessary to determine the similarity between any two pairs of frames to determine which frames to merge together. Due to the fact that a new method of assessing the similarity was published between performing experiment 1 and 2 (Franke *et al.*, 2015), the analysis performed on the two sets of data and presented here are different. This was necessary because the new version of DATCMP, the software program used to perform the similarity analysis (Petoukhov *et al.*, 2012), does not incorporate some of its old functionality utilised

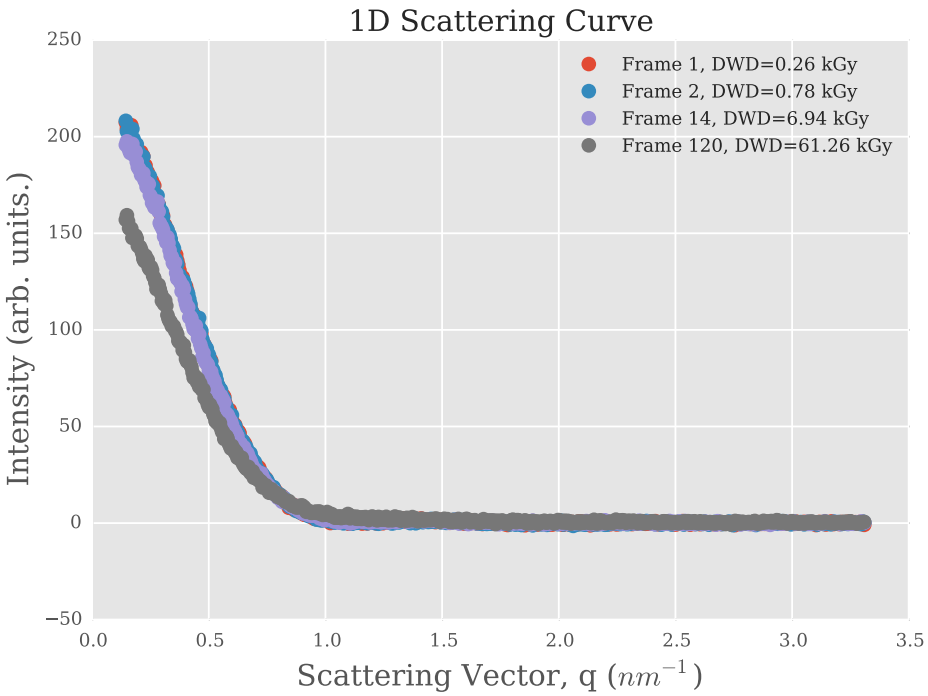


Figure 6.11: 1D scattering curves from the first run of the GI sample with no radioprotectant compounds added. Frames 1 and 2 visually seem to overlap well and the merging analysis determines these frames to be similar. Frame 14 is the first frame found to be dissimilar to frame 1. However, by visual inspection, the dissimilarity is not obvious. Frame 120 was the last frame collected in this run. The dissimilarity of frame 120 from the others is obvious. This suggests that the molecules in the sample have undergone significant radiation induced changes during the experiment.

in Expt 1, and the online manual describing the new features has not yet been updated.

6.4.1 Data analysis - experiment 1

Buffer subtraction was first carried out in ScÅtter (Rambo, R. at DLS, Didcot, UK). DATCROP, a utility program for cropping SAXS data, distributed as part of the ATSAS program suite (Petoukhov *et al.*, 2012), was then used to remove data points from the very low angles around the beam stop and also from the larger angles where the signal-to-noise ratio drops considerably: this was carried out by visual inspection. DATCMP (a program also distributed in the ATSAS suite) was then used to carry out a Scheffe *post hoc* analysis on the resulting 1D scattering data (frames) for each experimental run. This analysis compares the similarity of the first frame to each of the subsequent frames, since the first frame is produced at the lowest dose and so it is assumed to be of the best quality. The result of the Scheffe *post hoc* analysis is a 'fidelity value', which is a value given for each frame to describe its similarity to the first frame. An identical frame is given a fidelity value of 0. Increasing fidelity values correspond to increasing dissimilarity of a particular frame from the first frame. A

plot of the fidelity values against time and diffraction weighted dose (DWD) is shown in Figure 6.11. These data show the changing similarity of frames throughout the experiment, but they do not explicitly provide a quantitative metric to compare the efficacies of the various radioprotectants. Further analysis of these data was required to achieve this.

The curves of fidelity values against both time and DWD visually resemble a logistic relationship (they look like S-shaped curves). Thus 4-parameter logistic (4PL) functions were fitted to the data:

$$F(x) = \frac{a-d}{(1+(x/c)^b)^2} + d, \tag{6.4.1}$$

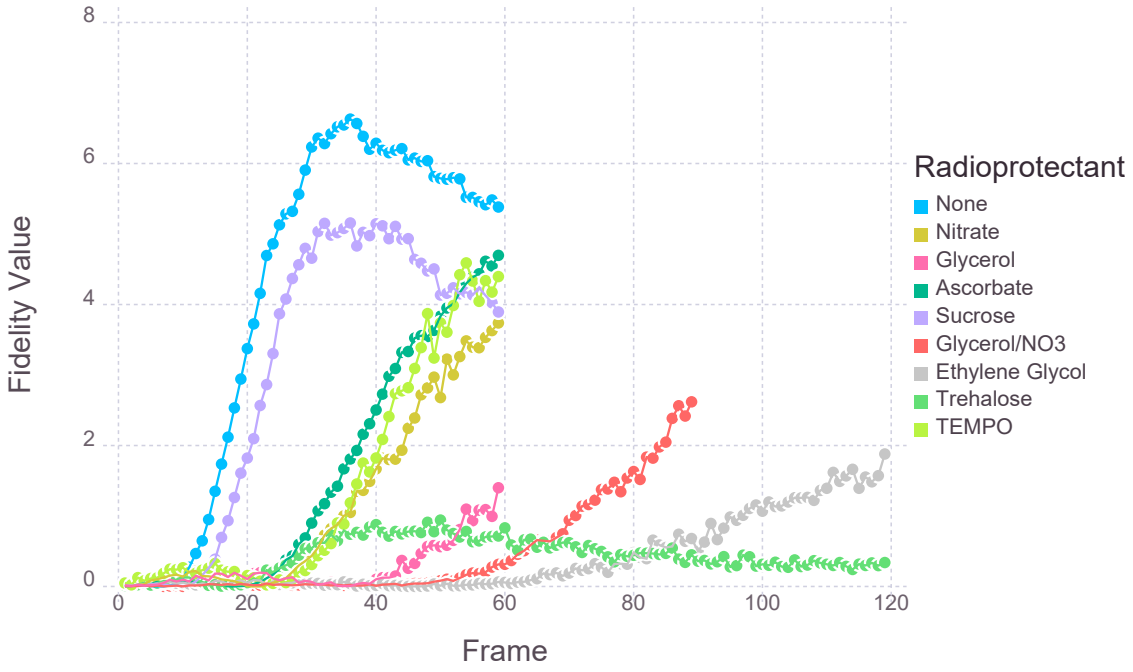
where F is the fidelity value, x is the x-axis coordinate (it could either be time or DWD), and a, b, c and d are the 4 parameters to be determined. Each of the parameters has graphical interpretation in relation to the logistic curve. a and d represent the minimum and maximum values of the curve respectively, b represents the steepness of the slope and c is the location of the point of inflection. The 4PL curve was used instead of the commonly used 3 parameter logistic function (also known as the Hill function) because it allows much more flexibility for the steepness of the curve, which was likely to change with different radioprotectants.

To determine which frames should be merged it is necessary to decide on the threshold similarity criterion. Two criteria were used:

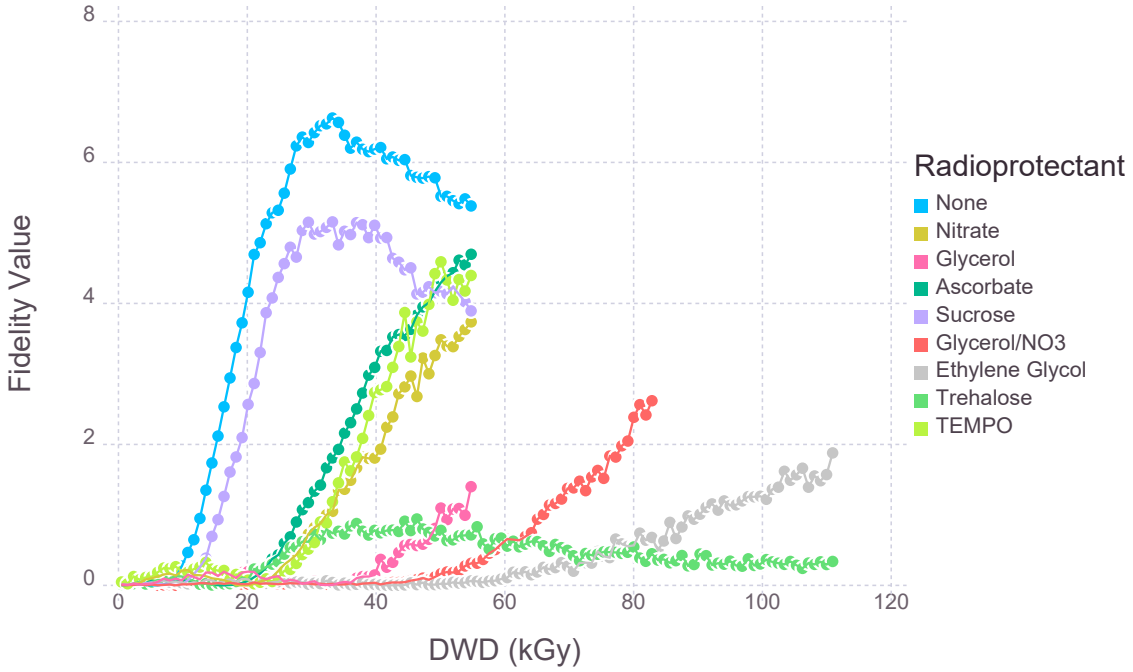
1. the DWD absorbed to reach an arbitrary value, and
2. the DWD absorbed to reach maximum curvature of the fidelity curve.

The first of the two criteria is straightforward to assess. A fidelity value is chosen and then the fitted logistic function can be rearranged to find the DWD value for which the chosen fidelity value is reached.

The second of the criteria is slightly more involved. For frames to be similar, their fidelity values have to be close to zero. Therefore the early frames should be close in value. However when frames start to become dissimilar, the fidelity value increases. In terms of the fitted logistic curve, this corresponds to an increasing gradient as the fidelity values increase. The required value is the total absorbed dose at the point when the increase in the gradient of the fidelity values reaches a maximum. To perform this analysis it was necessary to calculate



(a)



(b)

Figure 6.11: (a) Fidelity value as a function of time (b) Fidelity value as a function of dose. Fidelity values start at 0 for the first frame and increase as the frames become more dissimilar.

the curvature, κ, of the logistic curves, which is given by:

$$\kappa(x) = \frac{\frac{\mathrm{d}^2 F}{\mathrm{d}x^2}}{\left(1 + \frac{\mathrm{d}F}{\mathrm{d}x}\right)^{3/2}}. \tag{6.4.2}$$

This equation was calculated symbolically using the symbolic math toolbox in MATLAB. The maximum of this function was found using the MATLAB function `fminbnd` to find the minimum value of $-\kappa(x)$ i.e. the problem was converted from trying to find the maximum into a problem in which the same solution is found by finding the minimum of the negative of the function. This is a common procedure for optimisation problems.

An implicit assumption with the methods described is that damage is entirely progressive and subsequent frames beyond the threshold are all significantly dissimilar. Mathematically this assumption manifests itself as logistic functions representing the fidelity as a monotonically increasing function of the dose.

6.4.2 Data analysis - experiment 2

In a similar manner to the analysis performed on the results from Expt 1, buffer subtraction and cropping were performed before 1D curve similarity analysis. However these steps were performed with custom written Python scripts. This was done for two reasons: the first was that the pipeline could then be completely scripted (ScÅtter, which was used for buffer subtraction for data processing for Expt 1, is a GUI based program written in Java), and secondly, DATCROP subtracts data from files, so the scripts that were written would require many file handling operations, which are relatively time consuming computationally. Therefore writing the scripts to perform these simple operations (subtraction and cropping) was much more time efficient and allowed for more flexibility.

DATCMP was again used for the similarity analysis. However, as mentioned in section 6.4, the method used in the current version has been changed by the authors of this software since the Expt 1 analysis. The new method is called the Correlation Map (CorMap) test (Franke *et al.*, 2015). Unlike the algorithm used to generate the fidelity values, the CorMap test does not require any estimates of the experimental errors to test similarity.

A full explanation of the method can be found in Franke *et al.* (2015), but the main ideas are

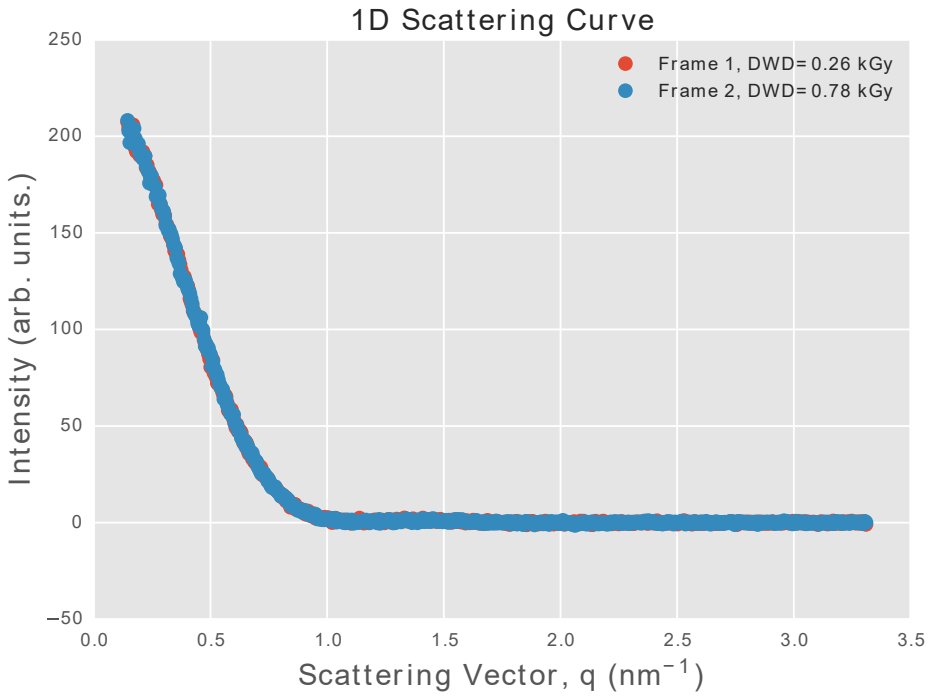
presented here and tested for their suitability for application to our data. Any two frames can be compared using a pairwise correlation, which involves calculating the difference between the two intensity curves. If the two frames are identical up to the noise level, then the difference between them is a random number. Given that the intensity values are assumed to come from a normal distribution (Franke *et al.*, 2015), the difference between them is normally distributed with a mean of zero. Due to the symmetry of the normal distribution, the probability of the difference of the data being either positive or negative is 0.5. The pairwise CorMaps between selected frames from the experiment are shown in Figure 6.12. Notice that for similar frames 1 and 2 (Figure 6.12a), the pairwise CorMap resembles a randomised lattice (Figure 6.12b) as expected for identical data. Conversely, for dissimilar frames 1 and 120 (Figure 6.12c), the pairwise CorMap shows large regions of white and black patches (Figure 6.12d), suggesting that the frames are systematically different. DATCMP performs quantitative analysis that formalises the similarity conclusions that can be drawn from these CorMaps.

The problem can be thought of in an identical manner to a coin toss experiment with the following two conditions:

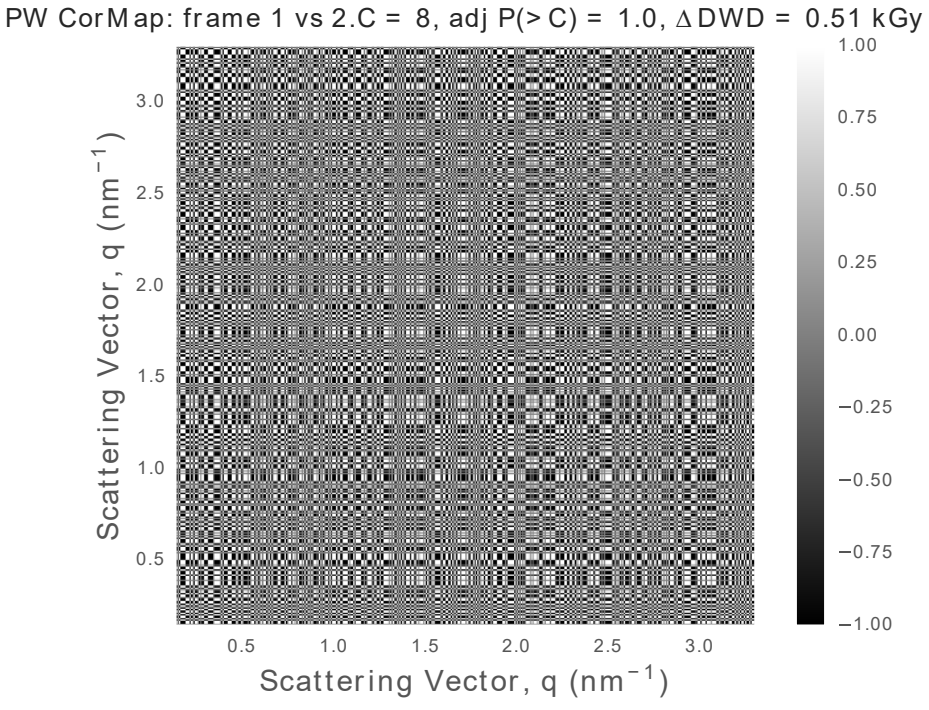
1. the probability of the difference of any two data points being either positive or negative is 0.5,
2. the result of the difference of any two data points with another two are assumed independent,

In the coin toss experiment one can ask: 'what is the probability of observing more than n consecutive heads (or tails)?'. The Schilling distribution (Schilling, 1990) calculates this probability. For the SAXS experiment this is the same as asking for the probability of observing a patch of white (+1) or black (-1) larger than the longest patch of white or black that was actually observed in the pairwise CorMap. More formally, we ask for the probability (P value) of obtaining an edge length larger than C (denoted $P(> C)$) within an n-by-n correlation matrix. This is calculated from the Schilling distribution with parameters n and C. If the p value is smaller than a threshold value α ($\alpha \leq 0.01$ is recommended (Franke *et al.*, 2015)) then the two frames can be considered dissimilar.

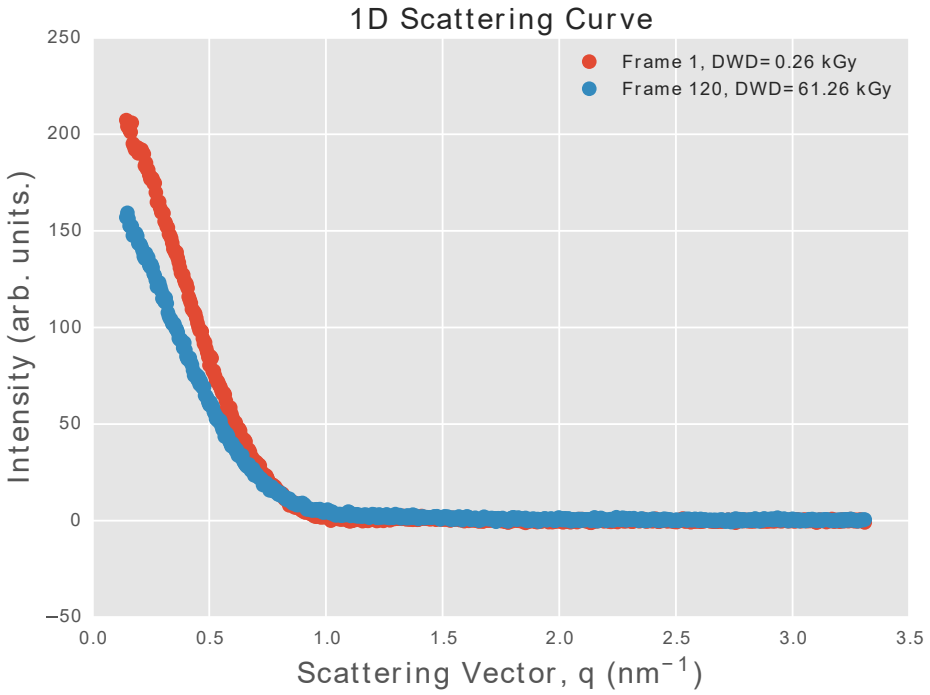
The CorMap test is implemented in the current distribution of DATCMP, although the soft-


1D Scattering Curve
Frame 1, DWD= 0.26 kGy
Frame 2, DWD= 0.78 kGy
Intensity (arb. units.)
Scattering Vector, q (nm^{-1})

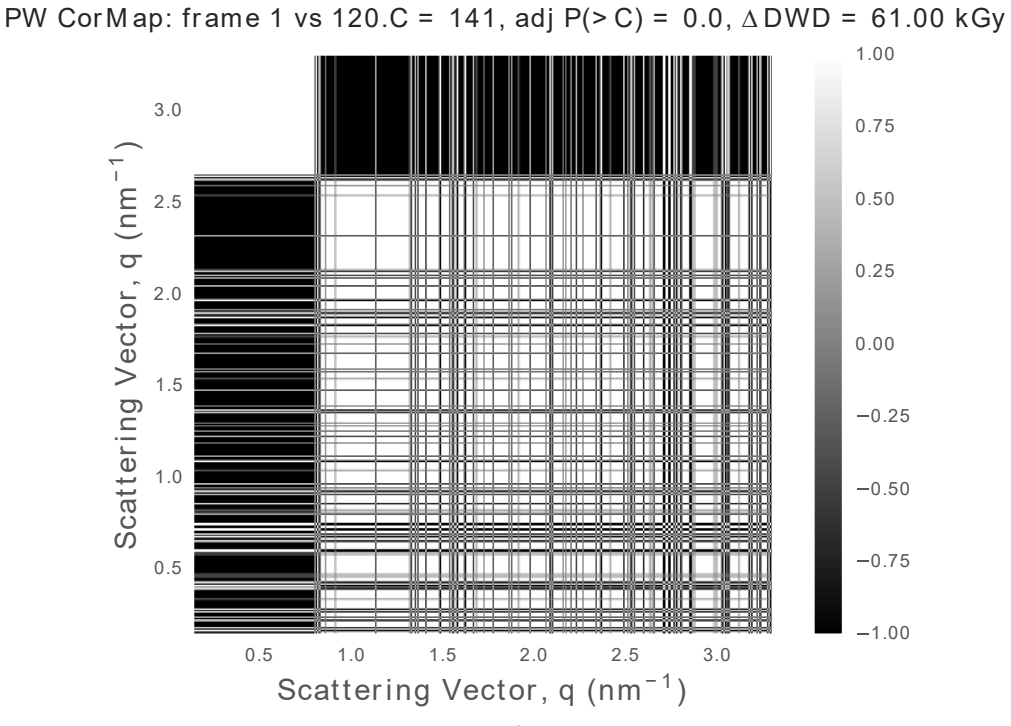
(a)


PW CorMap: frame 1 vs 2.C = 8, adj P(> C) = 1.0, ΔDWD = 0.51 kGy
Scattering Vector, q (nm^{-1})
Scattering Vector, q (nm^{-1})

(b)



(c)



(d)

Figure 6.12: Similarity comparison with selected frames from the first experimental repeat with no radioprotectant added. (a) 1D 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 randomised lattice pattern suggests that the 1D curves are similar. (c) 1D 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.

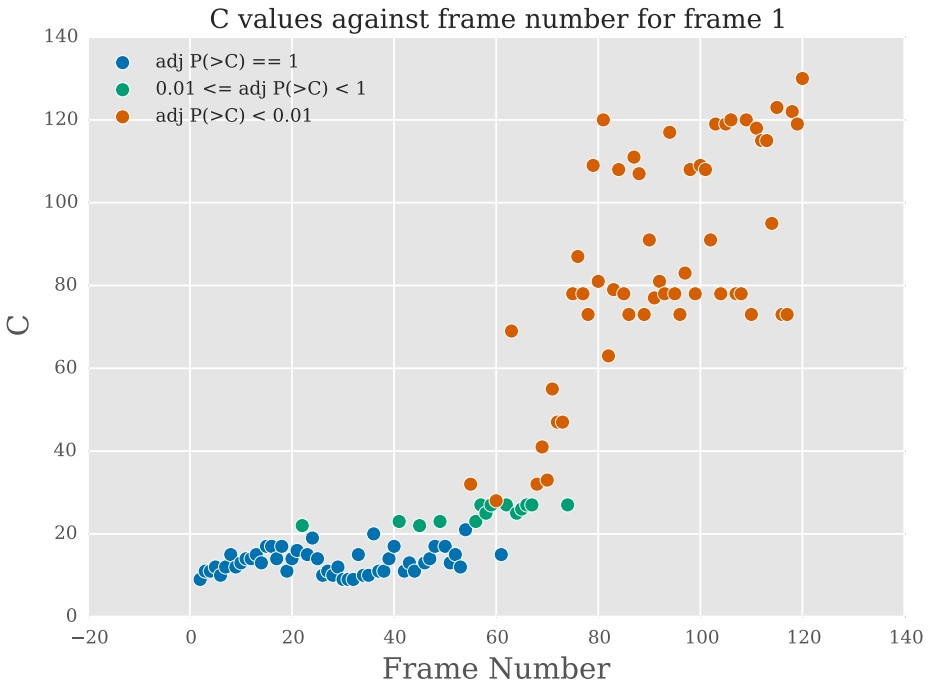


Figure 6.13: Longest observed edge length C against the frame number for pairwise comparisons with frame 1. For similar frames to frame 1, the pairwise CorMaps are more like randomised lattices (Figure 6.12b) and hence C is fairly small. Therefore the chance of observing a longer edge length than C is high ($P_{adj}(> C) = 1$ - blue circles). As frames start becoming more dissimilar, the C values increase and the $P_{adj}(> C)$ values fall. These circles are coloured green. When frames are very dissimilar, C becomes very large (Figure 6.12d) and $P_{adj}(> C) < 0.01$. The circles representing the comparison with these frames are coloured orange. The first dissimilar frame (coloured orange - frame 55) may not necessarily be the point at which frames should stop being merged according to this analysis, because there are circles after frame 55 that are not orange in colour.

ware does not include visualisation tools. Since multiple pairwise tests have to be made from several comparisons, the Bonferroni correction[†] is applied to the $P(> C)$ values resulting in $P_{adj}(> C)$ values. The raw output from the program is essentially a list of $P(> C)$ and $P_{adj}(> C)$ values for all possible pairwise comparisons. The set of $P_{adj}(> C)$ values that result from comparing the first frame to all subsequent frames is shown in Figure 6.13.

Applying this methodology to Expt 2 results

As can be seen in Figure 6.13, the first frame which is calculated to be dissimilar to frame 1 is frame 55 (first orange circle) which has a corresponding dose of 14.73 kGy. Therefore the threshold for radiation damage onset to be significant could be set at that frame. However the frames immediately after 55 do not have $P_{adj}(> C) < 0.01$ (green or blue circles)

[†]If multiple hypotheses are tested then the likelihood of observing a rare event (and thus the likelihood of incorrectly rejecting the null hypothesis) is increased. The Bonferroni correction divides the overall statistical significance level, α, by the number of tests, n, so that the hypotheses are tested individually at a significance level of α/n.

suggesting that these frames are not significantly dissimilar to frame 1. Thus frame 55 may be a noisy outlier and radiation damage may not necessarily be significant at that frame. A more robust check may be to find the first dissimilar frame for which m consecutive frames are dissimilar. This possibility was thus investigated and tested for m = 1, 3, 5, 7 and 10 to determine the value $m = m_0$ for which $m > m_0$ did not significantly change the dose at which frames were determined to be dissimilar. Figure 6.14 shows the result of the test for all of the radioprotectant experiments (note that the dose is used as the threshold for radiation damage instead of frame number). It can be seen that for most radio protectant compounds if m = 1, the spread of the apparent radiation damage onset is generally larger than for m = 3, 5, 7, 10. For m = 3, 5, 7, 10 the values corresponding to the onset of significant radiation damage are practically identical (except for DTT, but this is dealt with in section 6.5.2). Thus for the subsequent radiation damage analysis, the comparisons were performed with m = 3.

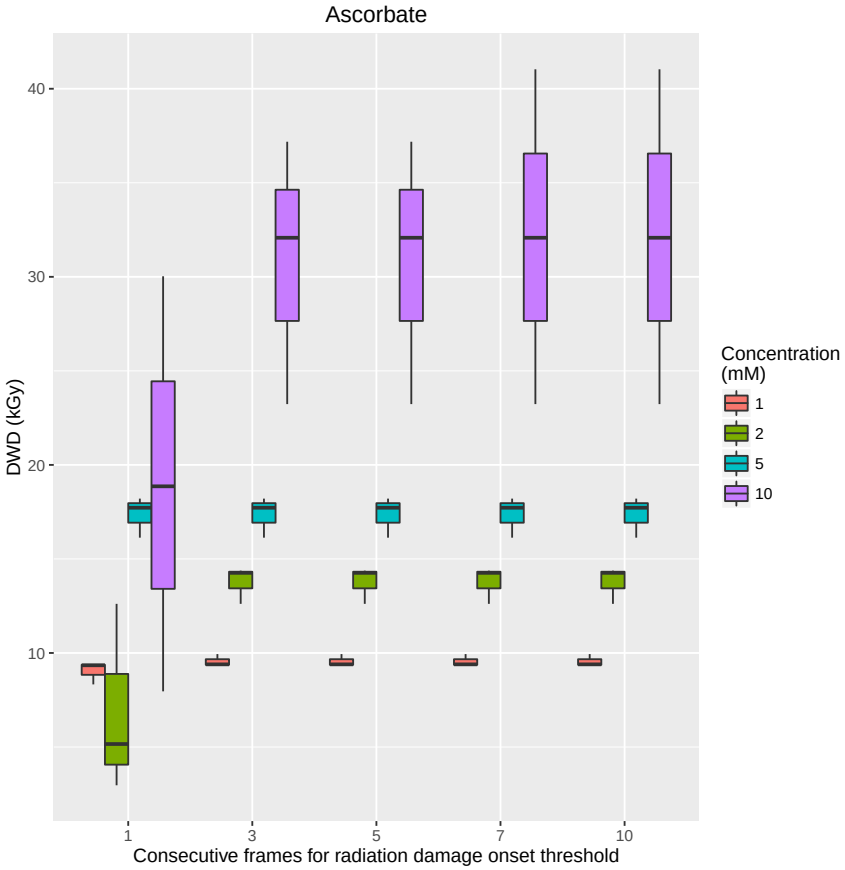
The other parameter that may affect the conclusions of this method for radiation damage analysis is the threshold level α. Therefore three different thresholds were set, α = 0.01, 0.05 and 0.1. Figure 6.15 shows the result of this test with all of the radioprotectant data.

It can be seen that the median values are very similar for the various α values. Given this fact and that α = 0.01 is the recommended (and sufficiently strict) threshold (Franke *et al.*, 2015), α = 0.01 was chosen as the value used for all radioprotectant compounds in the subsequent radiation damage analysis. The advantage of using α = 0.01 is that frames have to be very dissimilar before the $P_{adj}(> C)$ value falls below that value. This means that it is less likely that frames are discarded when they actually are similar (in statistical speak this means there is less chance of a type I error).

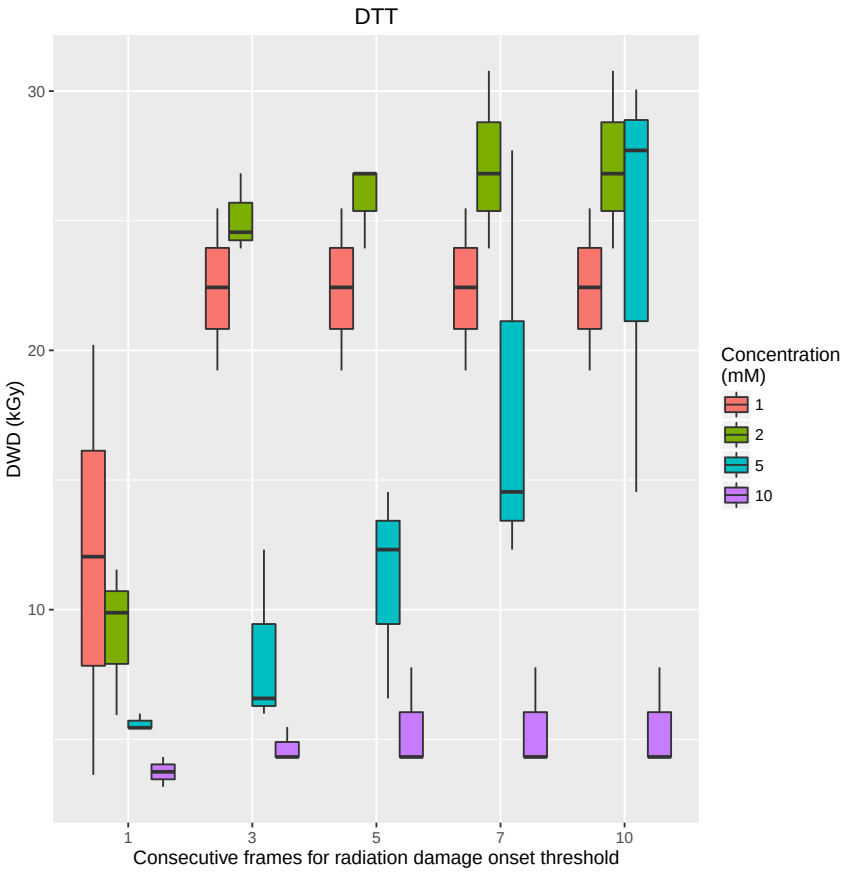
6.5 Results

6.5.1 Experiment 1

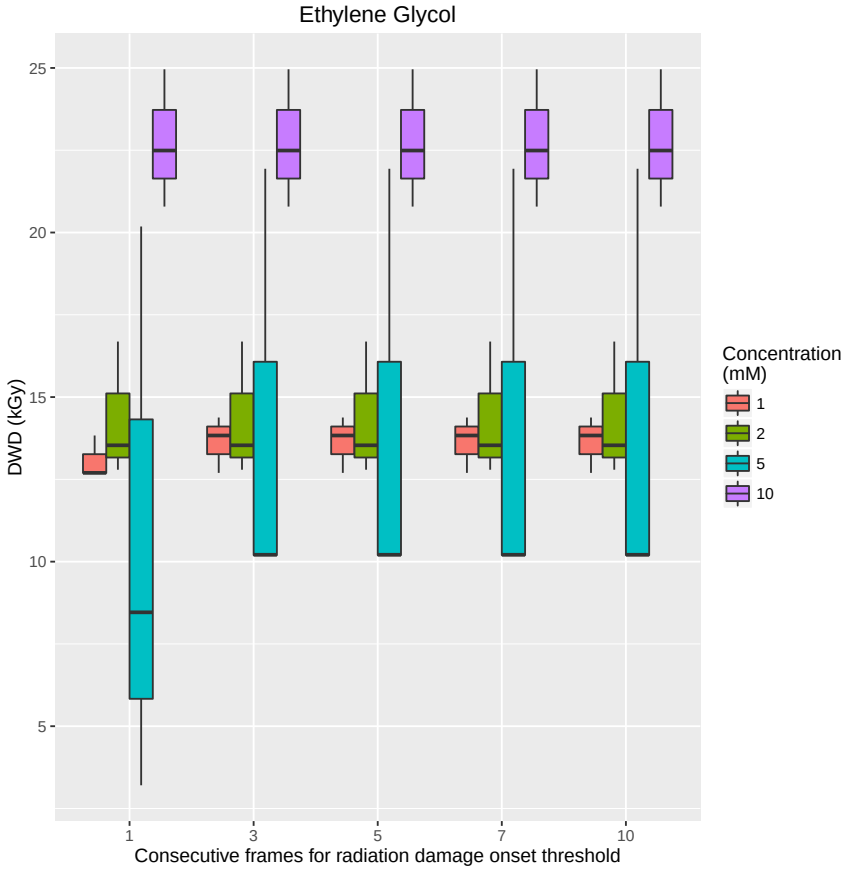
Figure 6.16 shows the results of the logistic curve fits to the fidelity data plotted in Figure 6.11. The five-pointed stars are the points where the fitted logistic curves reach a fidelity value of 0.71. This was an arbitrary value chosen because all samples reached that value. The six-pointed stars are the maximum curvature points of the fitted logistic curves. To com-


Ascorbate
DWD (kGy)
40
30
20
10
1
3
5
7
10
Consecutive frames for radiation damage onset threshold
Concentration (mM)
1
2
5
10

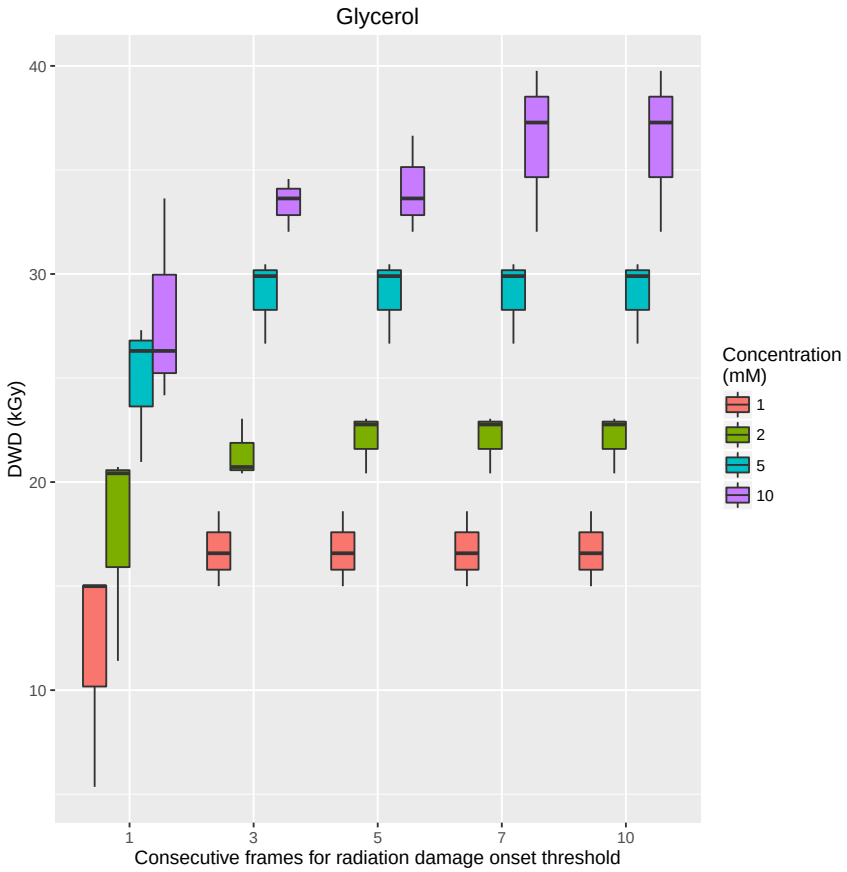
(a)


DTT
DWD (kGy)
30
20
10
1
3
5
7
10
Consecutive frames for radiation damage onset threshold
Concentration (mM)
1
2
5
10

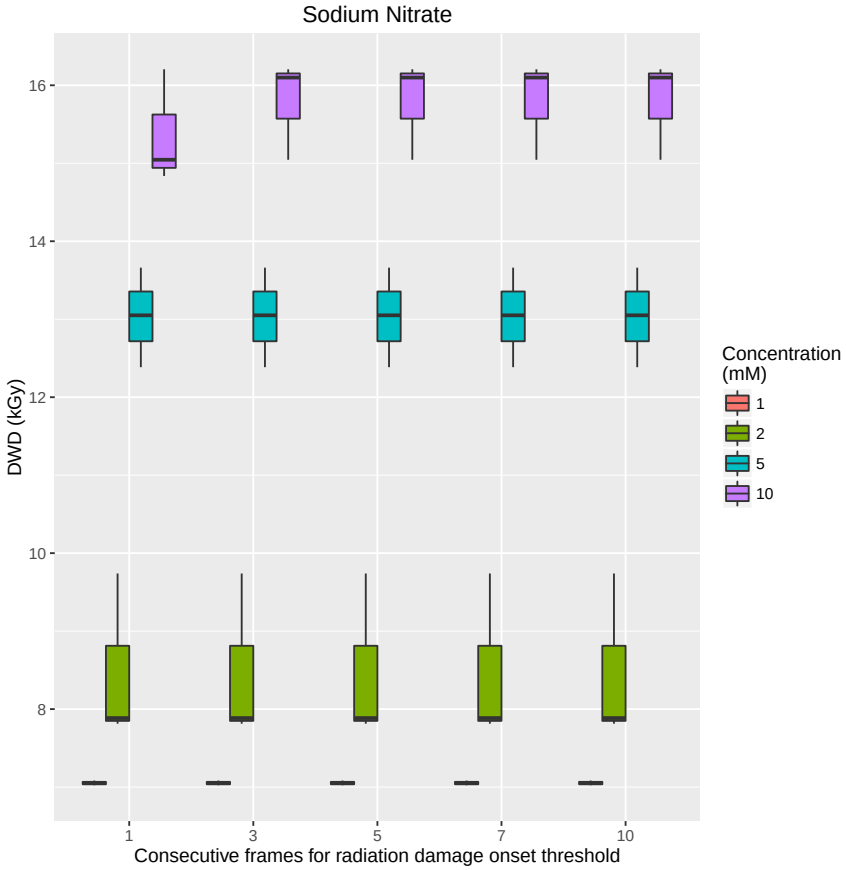
(b)


Ethylene Glycol
DWD (kGy)
25
20
15
10
5
1
3
5
7
10
Consecutive frames for radiation damage onset threshold
Concentration (mM)
1
2
5
10

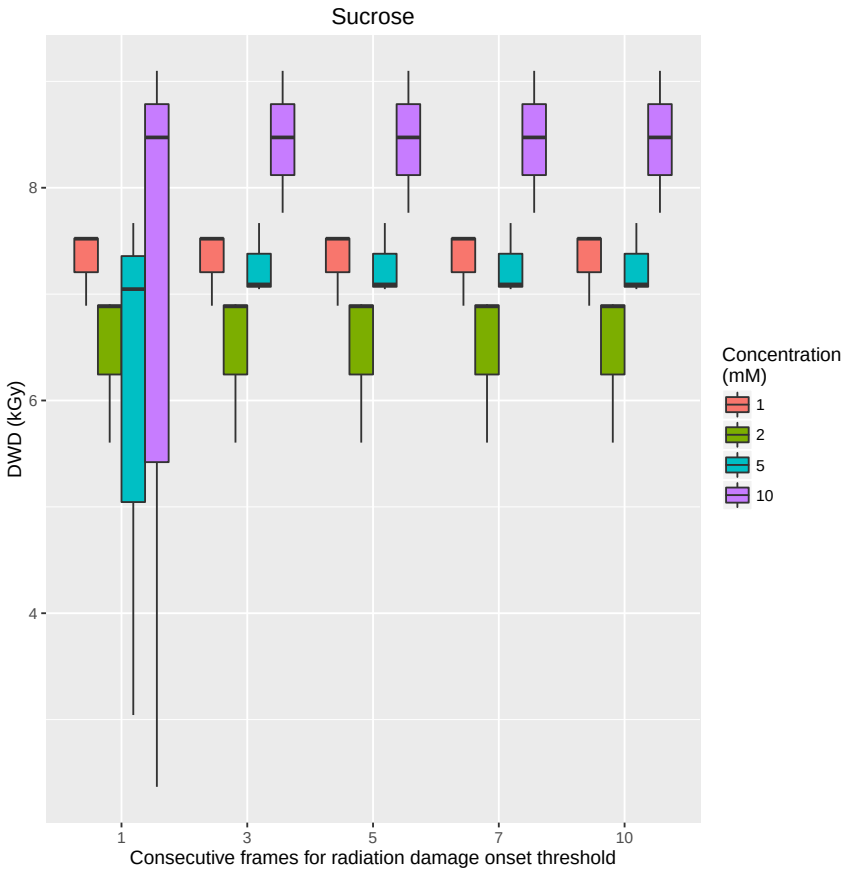
(c)


Glycerol
DWD (kGy)
40
30
20
10
1
3
5
7
10
Consecutive frames for radiation damage onset threshold
Concentration (mM)
1
2
5
10

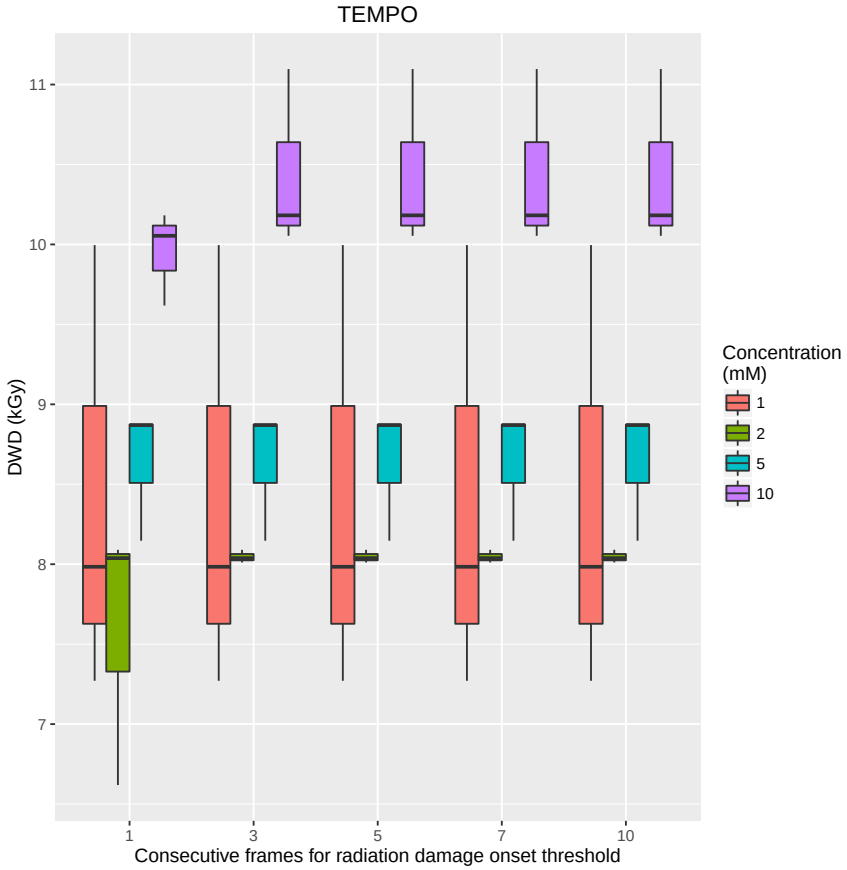
(d)


Sodium Nitrate
DWD (kGy)
16
14
12
10
8
1
3
5
7
10
Consecutive frames for radiation damage onset threshold
Concentration (mM)
1
2
5
10

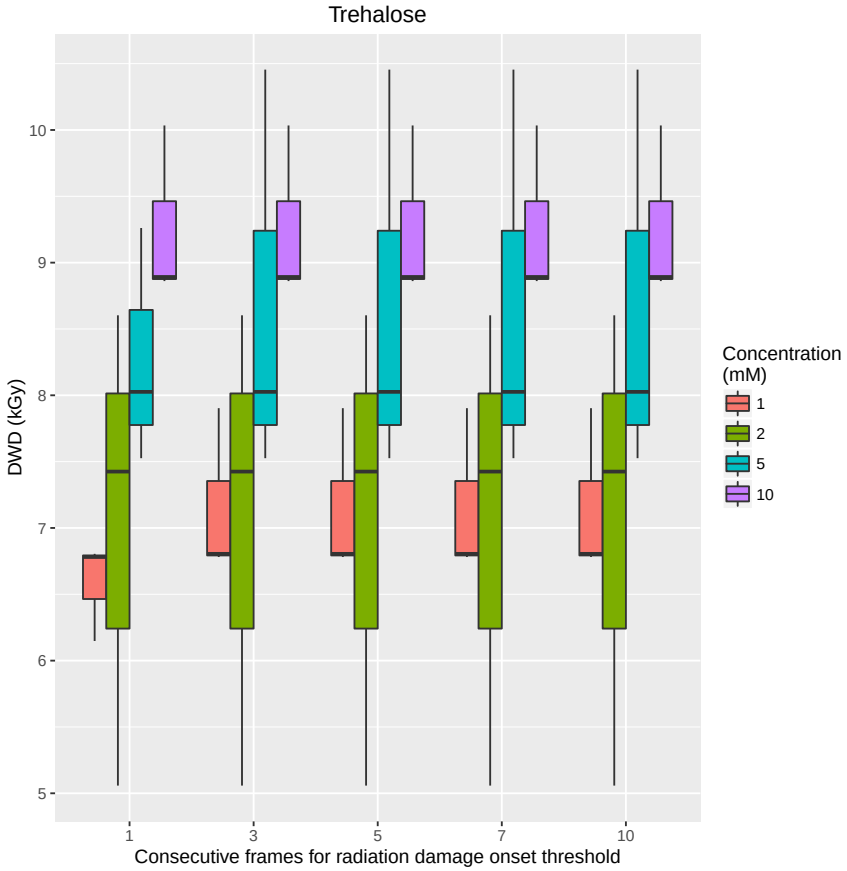
(e)


Sucrose
DWD (kGy)
8
6
4
1
3
5
7
10
Consecutive frames for radiation damage onset threshold
Concentration (mM)
1
2
5
10

(f)

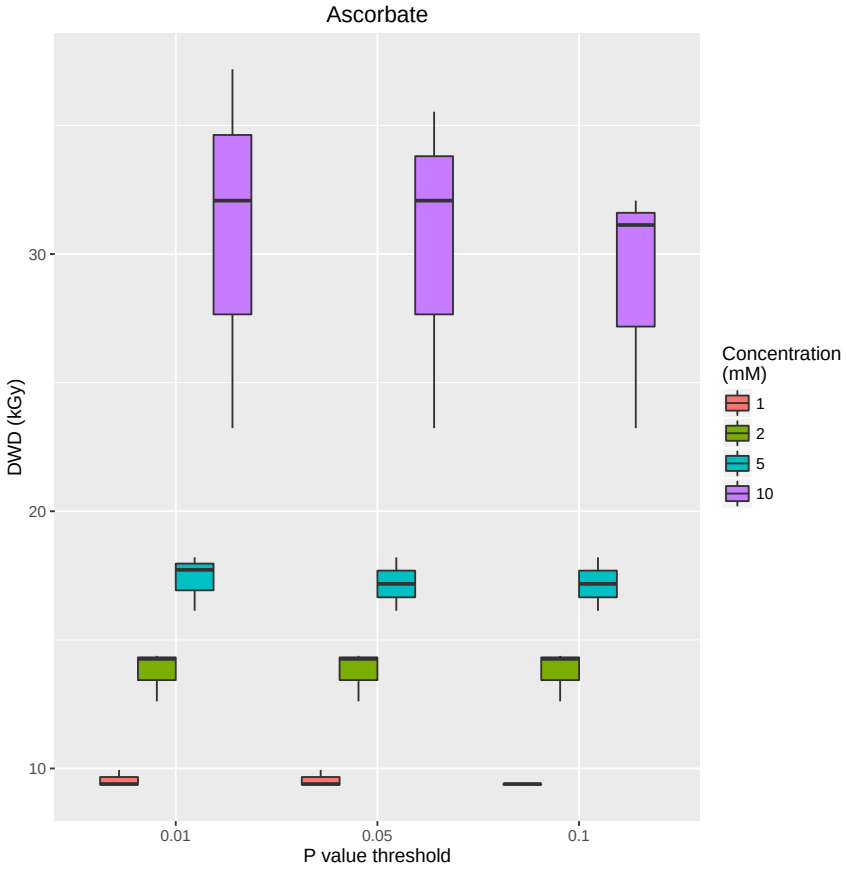


(g)

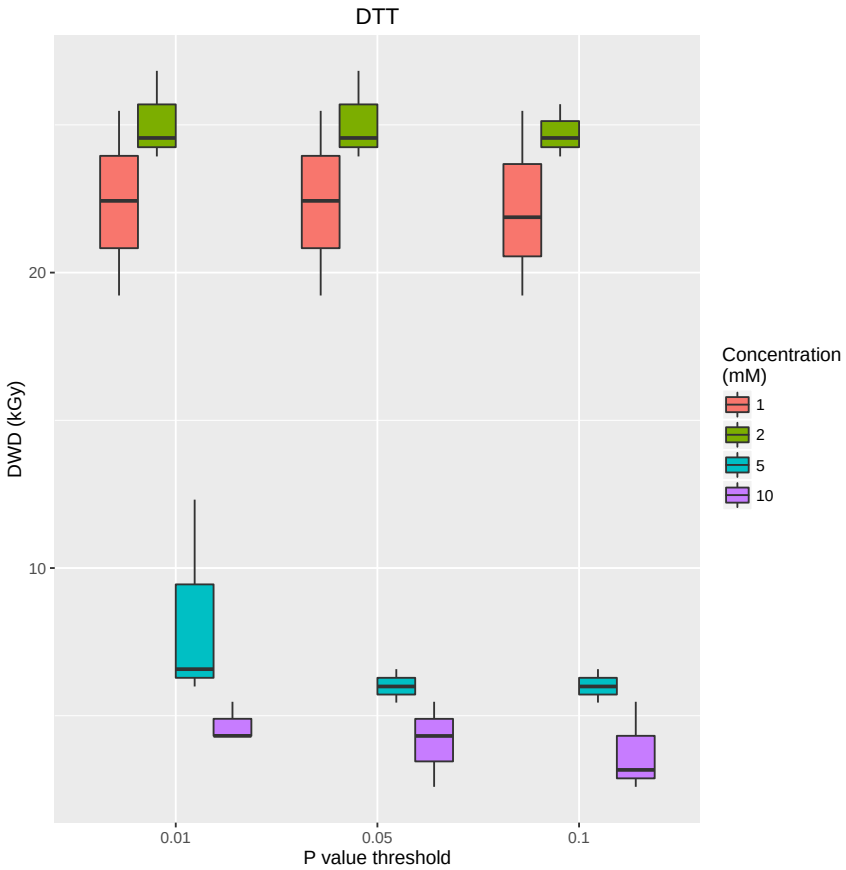


(h)

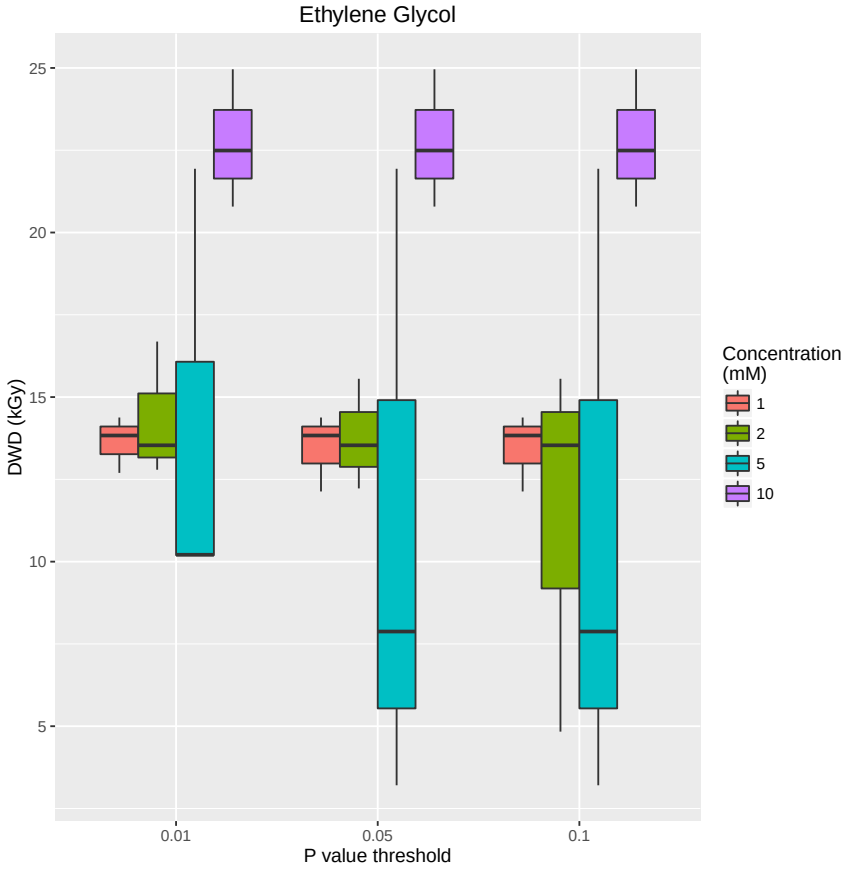
Figure 6.14: Dose at which significant radiation damage is determined to have occurred for different values of m, the number of consecutive dissimilar frames, for the 8 radioprotectants tested in Expt 2.


Ascorbate
DWD (kGy)
30
20
10
0.01
0.05
0.1
P value threshold
Concentration (mM)
1
2
5
10

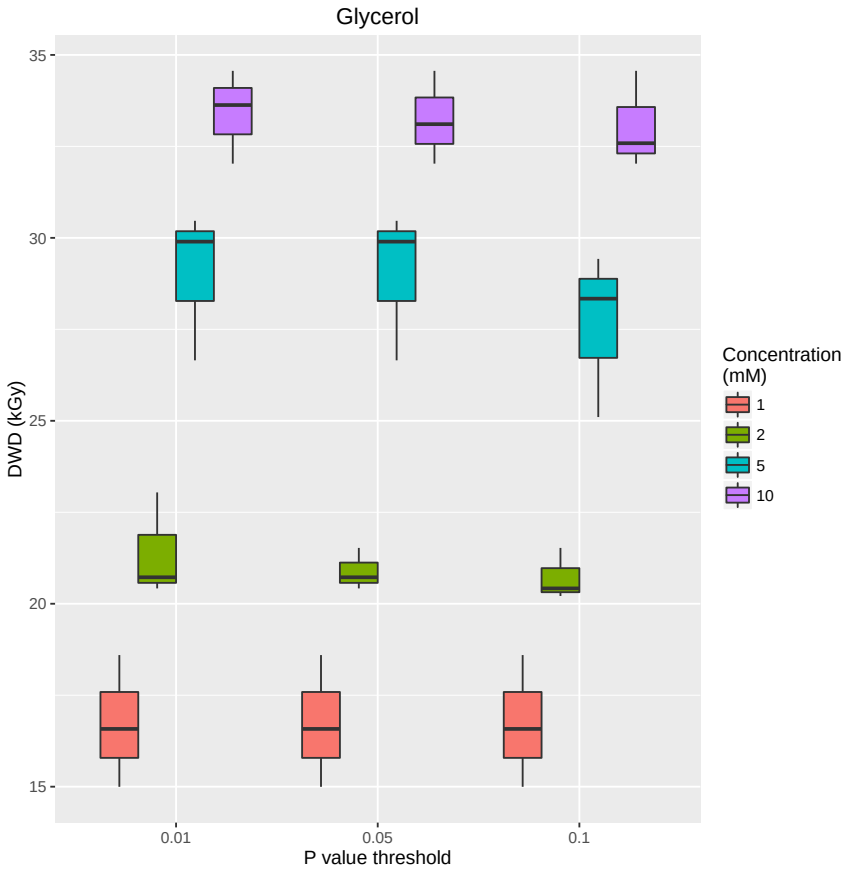
(a)


DTT
DWD (kGy)
20
10
0.01
0.05
0.1
P value threshold
Concentration (mM)
1
2
5
10

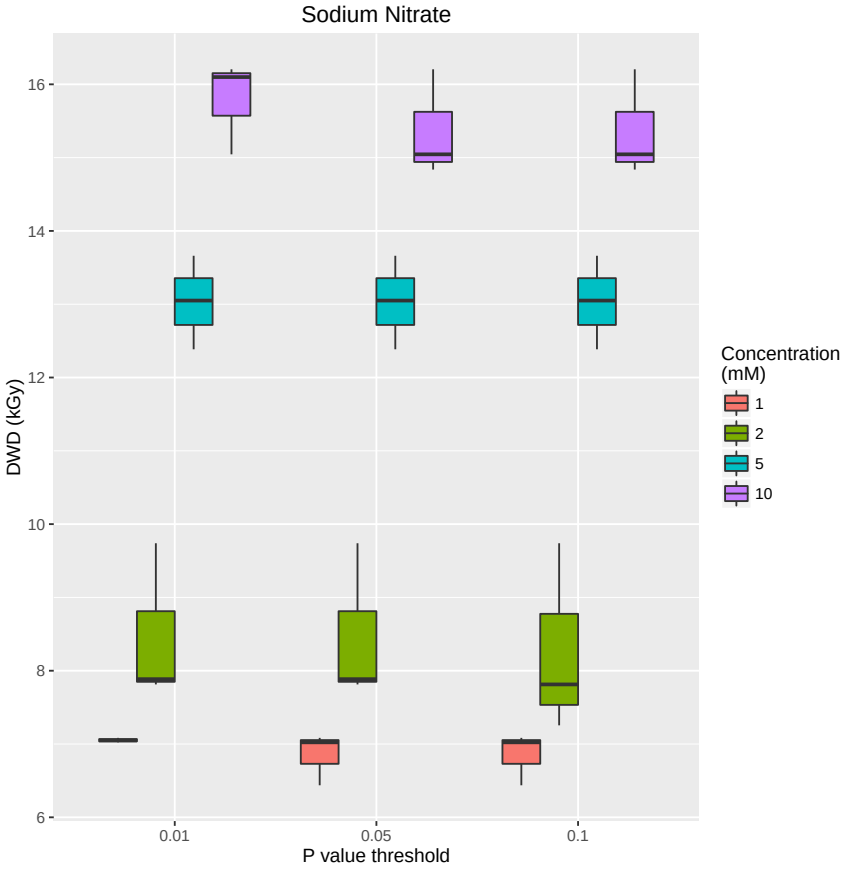
(b)


Ethylene Glycol
DWD (kGy)
25
20
15
10
5
0.01
0.05
0.1
P value threshold
Concentration (mM)
1
2
5
10

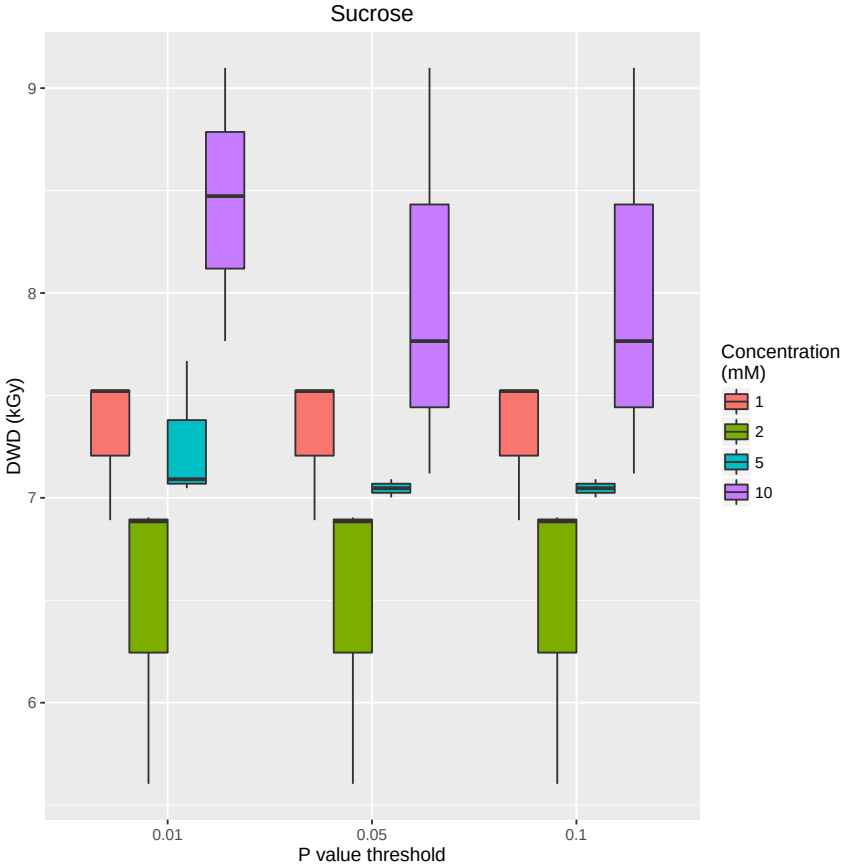
(c)


Glycerol
DWD (kGy)
35
30
25
20
15
0.01
0.05
0.1
P value threshold
Concentration (mM)
1
2
5
10

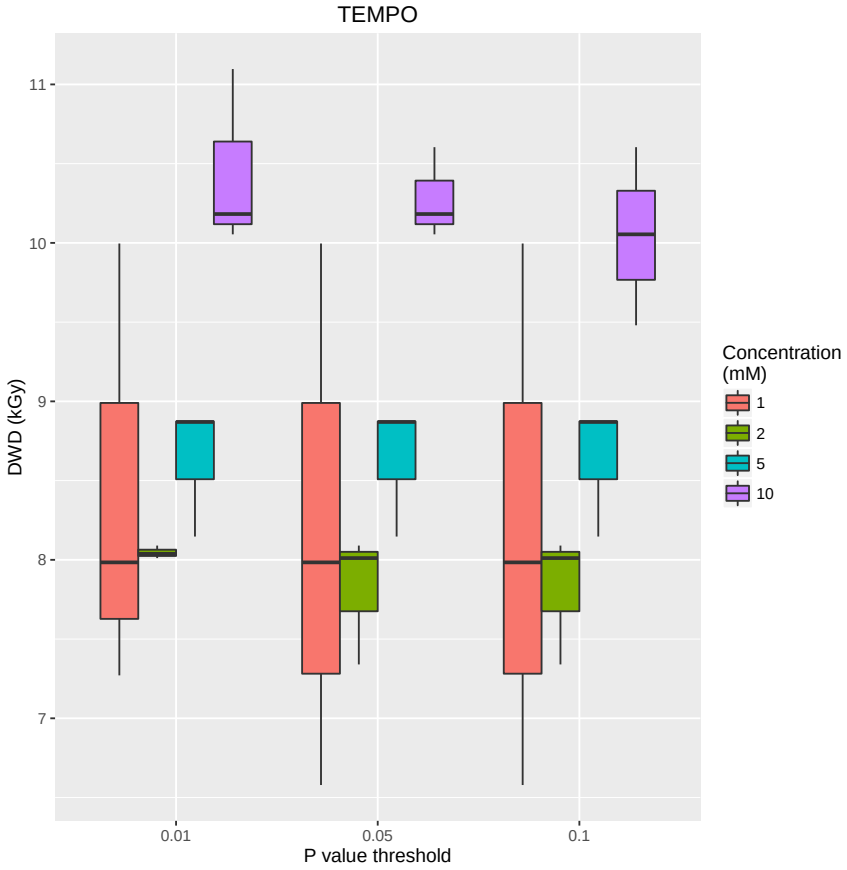
(d)


Sodium Nitrate
DWD (kGy)
16
14
12
10
8
6
0.01
0.05
0.1
P value threshold
Concentration (mM)
1
2
5
10

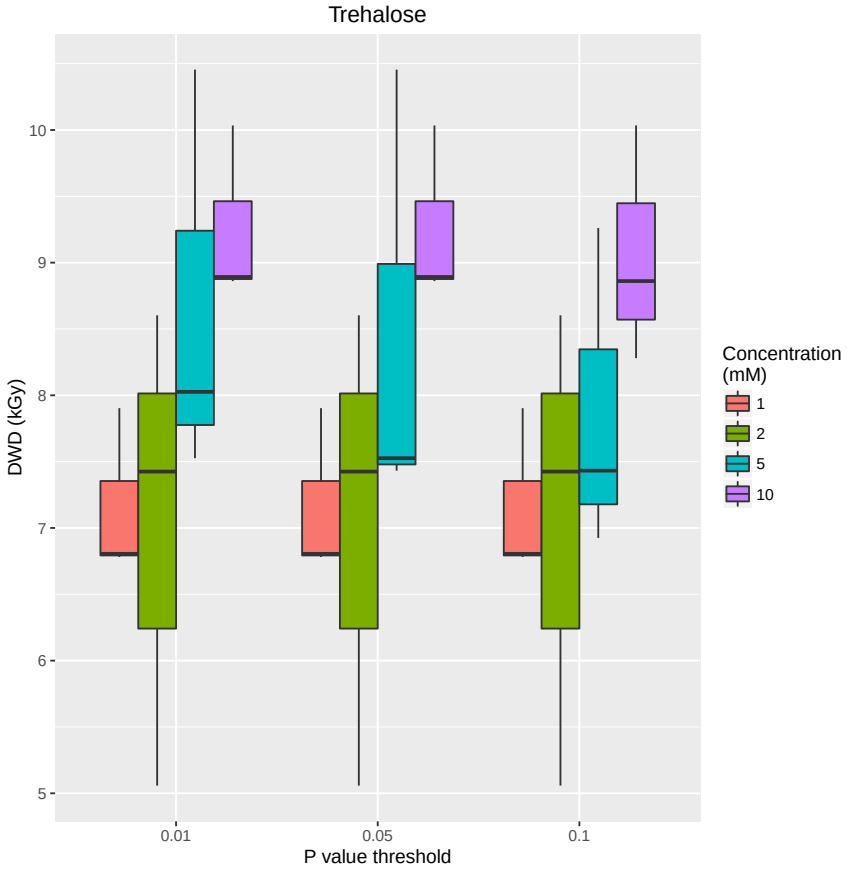
(e)


Sucrose
DWD (kGy)
9
8
7
6
0.01
0.05
0.1
P value threshold
Concentration (mM)
1
2
5
10

(f)



(g)



(h)

Figure 6.15: Dose at which significant radiation damage is determined to have occurred for different values of α, the threshold probability value to determine frame similarity, for the 8 radioprotectants tested in Expt 2.

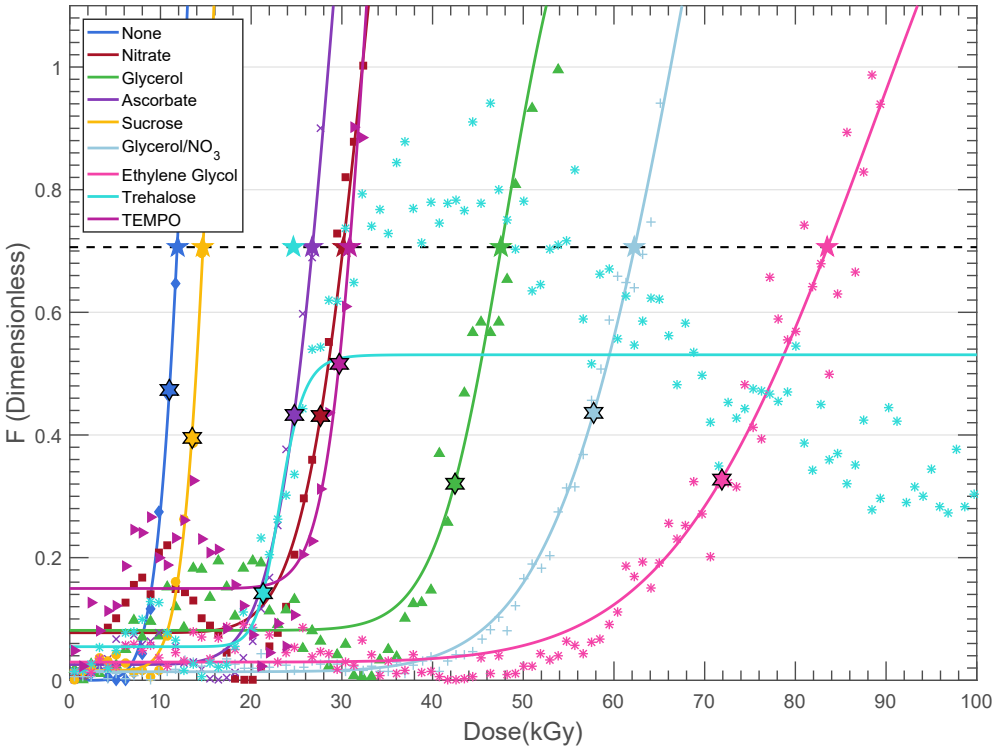
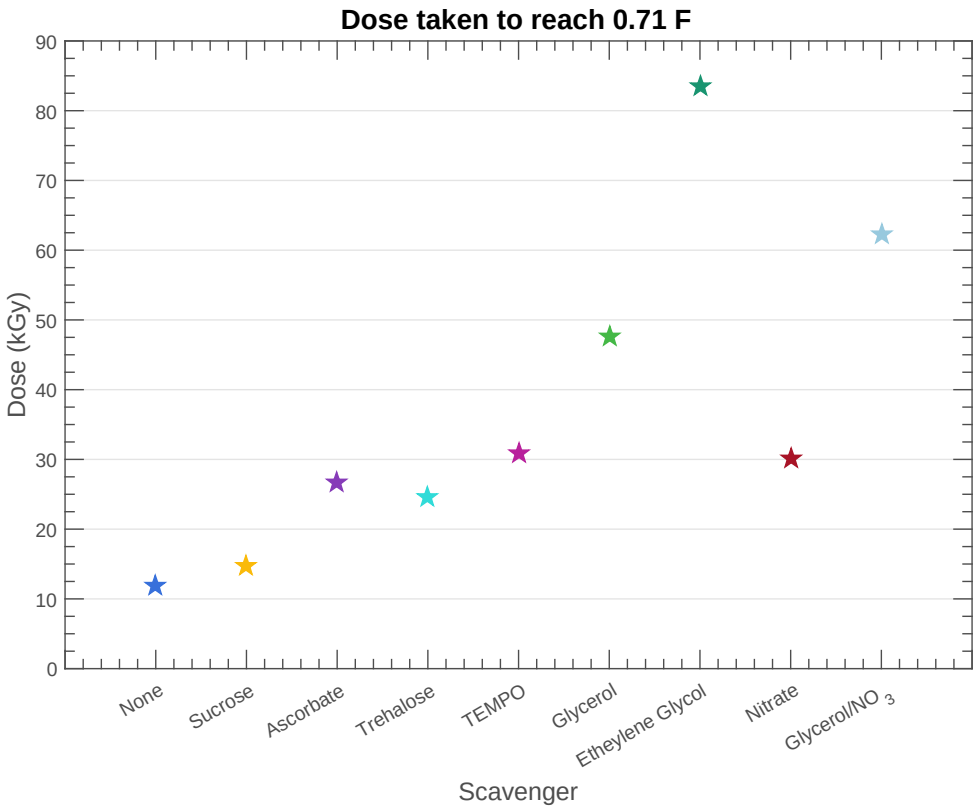


Figure 6.16: Fidelity values against dose for each radioprotectant along with their corresponding fitted logistic curves for the data collected in Expt 1. The five-pointed stars represent the points where the fitted logistic curves reach a fidelity value of 0.71. The six-pointed stars are the maximum curvature points of the fitted curves.

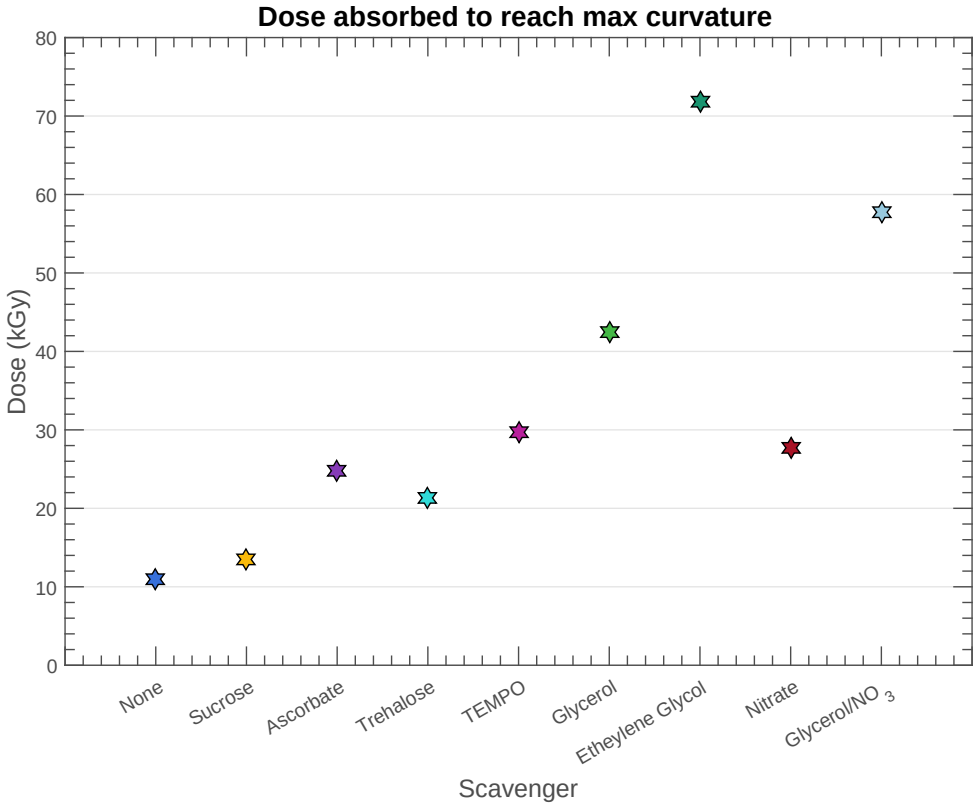
pare the efficacies of each radioprotectant compound, the dose at which these points occur are plotted for each scavenger in Figure 6.17.

Figure 6.17a shows the dose taken to reach a fidelity value of 0.71 for each scavenger. Ethylene glycol is the most effective radioprotectant at 5 mM concentration, followed by the glycerol/sodium nitrate mixture and then glycerol alone. It is clear from these results that adding a radioprotectant allows useable data to be collected for longer because the sample is less sensitive to irradiation.

Similar conclusions are obtained from Figure 6.17b where the dose at which maximum curvature is reached is plotted for each radioprotectant. The results obtained using this metric give an identical relative efficacy ordering to the order seen when using the other metric. i.e. ethylene glycol is most effective, followed by the glycerol/sodium nitrate mixture etc. The main difference between the metrics is that the doses at which the thresholds are determined using the maximum curvature method are consistently lower than those found using the high value threshold.



(a) Expt 1: Dose taken to reach a fidelity value of 0.71 for each scavenger.



(b) Expt 1: Dose at which maximum curvature is reached

Figure 6.17

6.5.2 Experiment 2

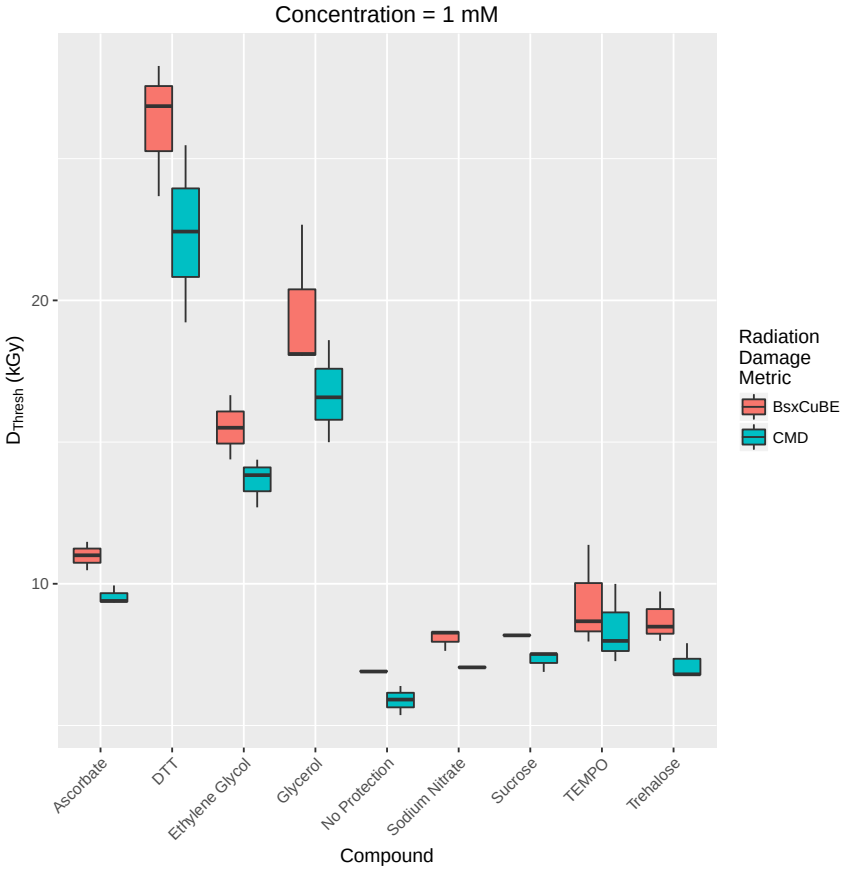
Comparing radiation damage onset metrics

In section 6.4.2 a metric for assessing the frame at which radiation damage had become significant was presented based on the CorMap test. Explicitly, this metric was defined as the point at which three consecutive frames were defined as dissimilar (m = 3) resulting from the CorMap test with threshold α = 0.01. This metric will be referred to as the *CMD* (CorMap Derived) metric.

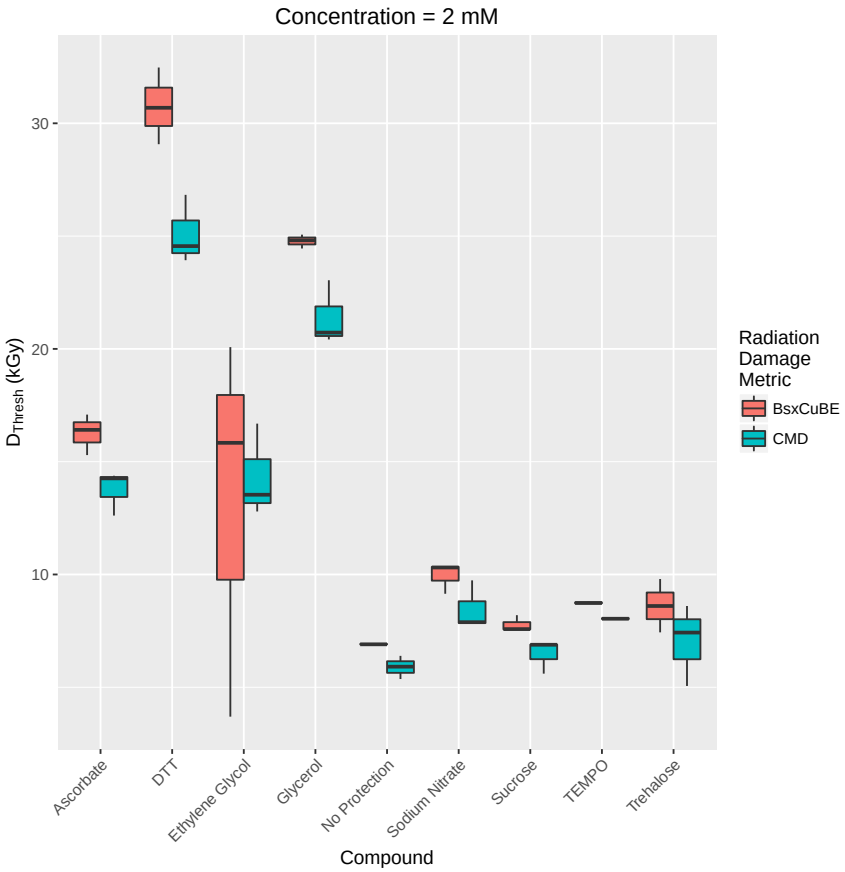
In addition to the *CMD* metric, an automatic data analysis pipeline at beamline BM29 is integrated into the beamline control system, *BsxCuBE*. This pipeline additionally performs analysis of frames and hence gives merging thresholds calculated for radiation damage onset using a metric henceforth denoted the *BsxCuBE* metric.

The results from analysis with both metrics were calculated/recorded and figure 6.18 shows how the two metrics compare for each compound at all concentrations. Generally the two metrics agree on the order of the efficacy of the various radioprotectant compounds, but the *BsxCuBE* metric always suggests that more frames can be merged then the *CMD* metric. Given that the *CMD* metric was developed to avoid individual dissimilar frames prematurely being flagged as the point of significant radiation damage onset, this result is quite surprising. It suggests that the *BsxCuBE* metric employs a very different method to assess the similarity of frames than the *CMD* metric.

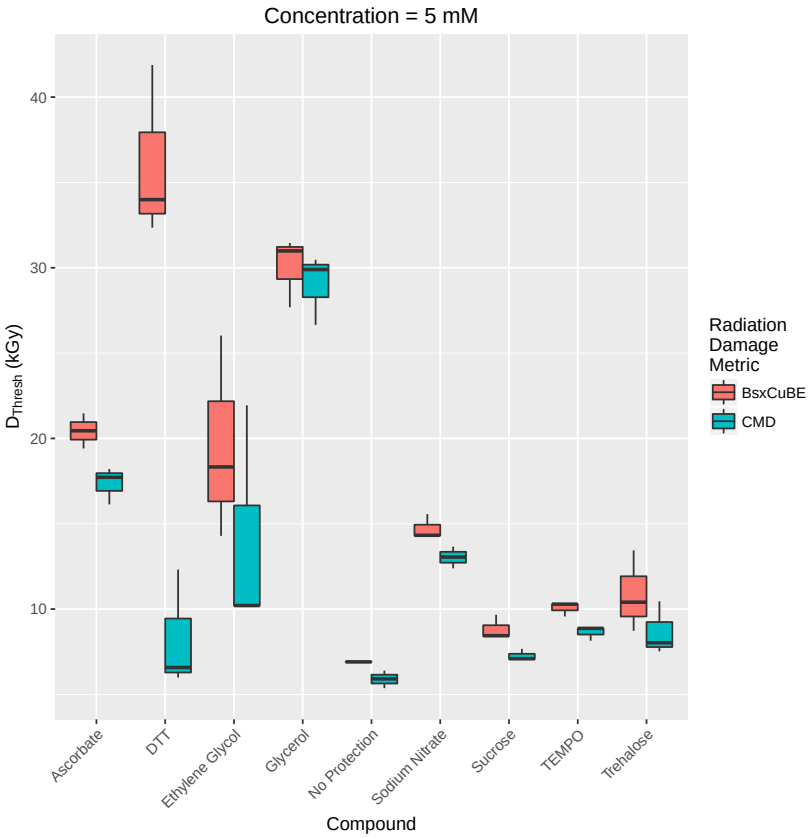
The biggest discrepancy between the two metrics is the result for DTT. The *BsxCuBE* metric predicts a much higher dose tolerance than the *CMD* metric, especially for the 5 mM and 10 mM concentrations. One reason thought to cause this discrepancy in the current work was the choice of m = 3 for the *CMD* metric. Figure 6.14b shows the results of the various values of m for DTT. Unlike the results for the other compounds (Figure 6.14), the chosen value of m could be the problem for 5 mM concentrations. However, this does not necessarily seem to be true for the 10 mM concentration. Thus it was necessary to re-evaluate some of the underlying assumptions of the analysis.


Concentration = 1 mM
D_{Thresh} (kGy)
20
10
Ascorbate
DTT
Ethylene Glycol
Glycerol
No Protection
Sodium Nitrate
Sucrose
TEMPO
Trehalose
Compound
Radiation Damage Metric
BsxCuBE
CMD

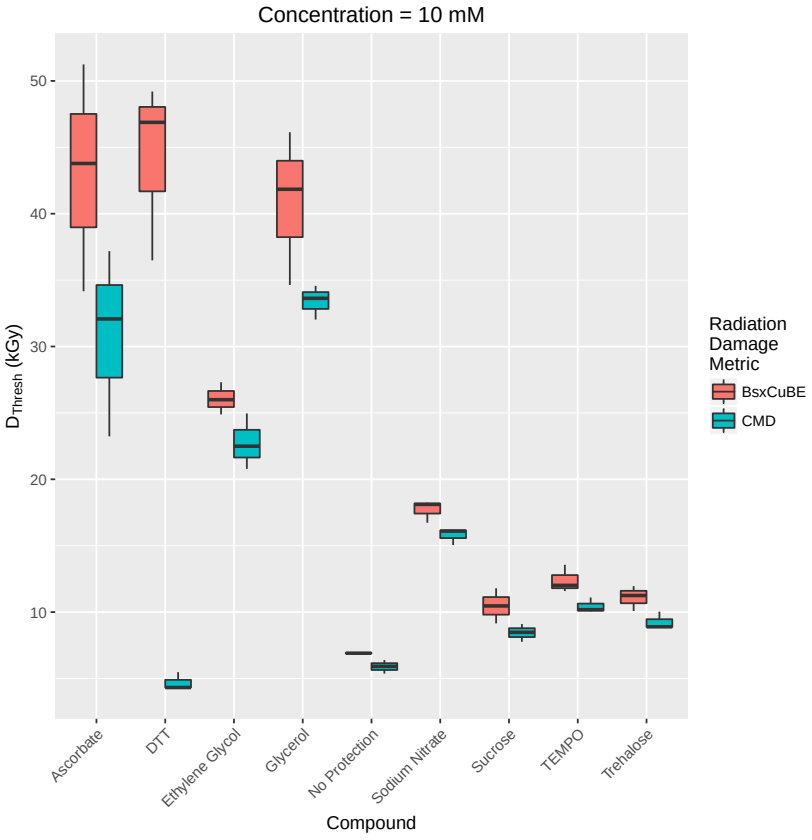
(a)


Concentration = 2 mM
D_{Thresh} (kGy)
30
20
10
Ascorbate
DTT
Ethylene Glycol
Glycerol
No Protection
Sodium Nitrate
Sucrose
TEMPO
Trehalose
Compound
Radiation Damage Metric
BsxCuBE
CMD

(b)



(c)



(d)

Figure 6.18: Expt 2: Dose value at which radiation damage is considered significant for each concentration of radioprotectant used. Each box plot is created from the threshold dose values calculated for three different experimental runs of the same radioprotectant compound at a particular concentration. Pink boxes correspond to the *BsxCuBE* metric, blue boxes correspond to the *CMD* metric

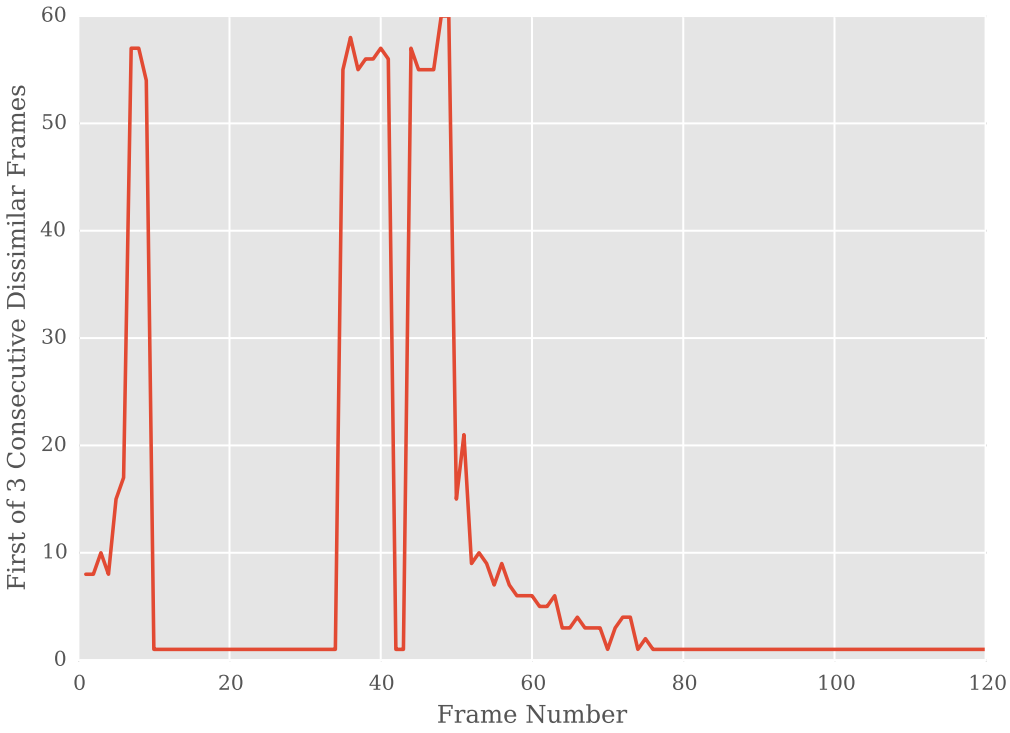


Figure 6.19: Expt 2: Radiation damage threshold frame against reference frame (m = 3) for the first repeat with DTT as the added radioprotectant at a concentration of 10 mM.

Pairwise comparisons with all frames

One assumption that is made in the merging analysis is that because the first frame suffers the least amount of dose, all subsequent frames should be compared to the first frame, the *reference frame*. If instead frames were compared to a different 'reference frame' would the threshold be different? The results shown in Figure 6.19 address this question, where the analysis was performed for the first repeat of GI sample with DTT added at 10 mM concentration. Along the y-axis are the frames which would be considered as the merging limit if the corresponding frame on the x-axis was chosen as the 'reference frame'. It can be seen that the radiation threshold value obtained using the *CMD* metric is highly dependent on the reference frame. This will greatly affect the conclusions that could be drawn from the analysis. From Figure 6.19, if the reference frame is 1, as is the case for the main analysis, then the frame at which radiation damage is considered significant is frame 8 (DWD = 4.32 kGy). However, if the reference frame was frame 7, then the threshold frame would be number 57 (DWD = 32.50 kGy). This value is closer to the value obtained from the *BsxCuBE* metric for that run (DWD = 46.88 kGy).

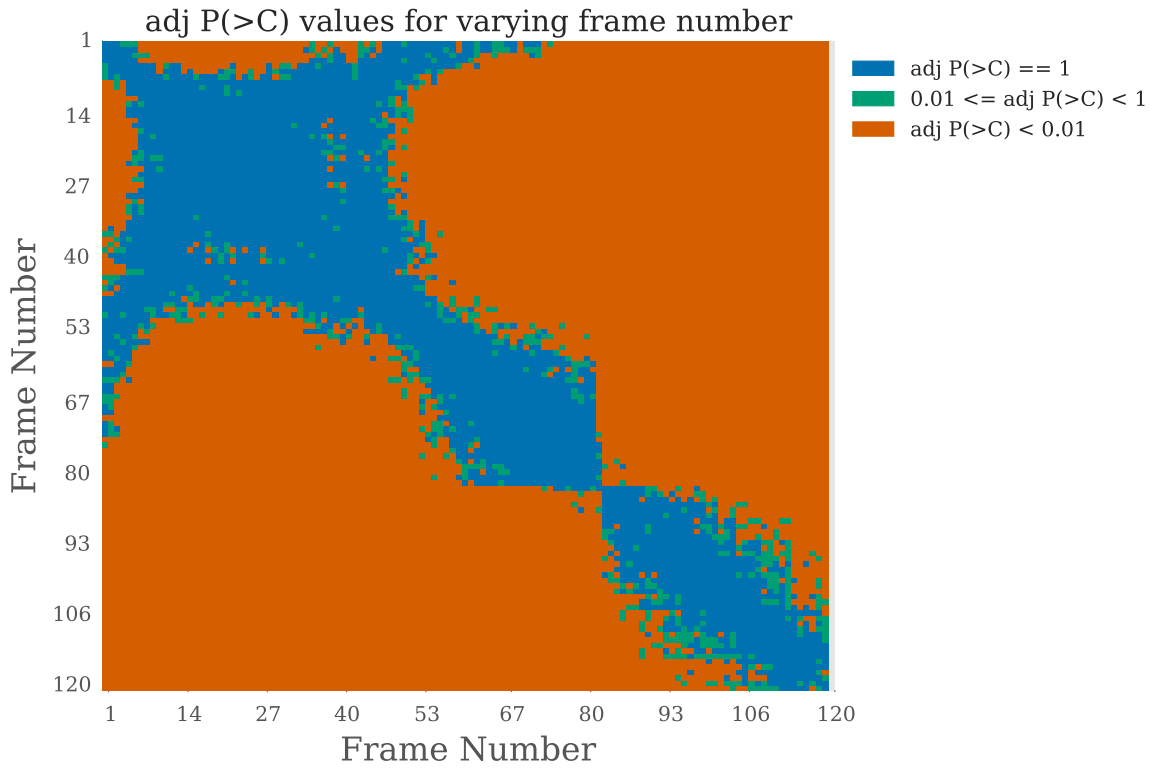


Figure 6.20: Expt 2: Heat map of all possible pairwise frame comparisons for the first repeat with 10 mM concentration DTT added. The y-axis represents the reference frame to which all other frames on the x-axis are compared. Blue - $P_{adj}(> C)$ = 1. Green - 0.01 $\leq P_{adj}(> C) <$ 1. Orange - $P_{adj}(> C) <$ 0.01.

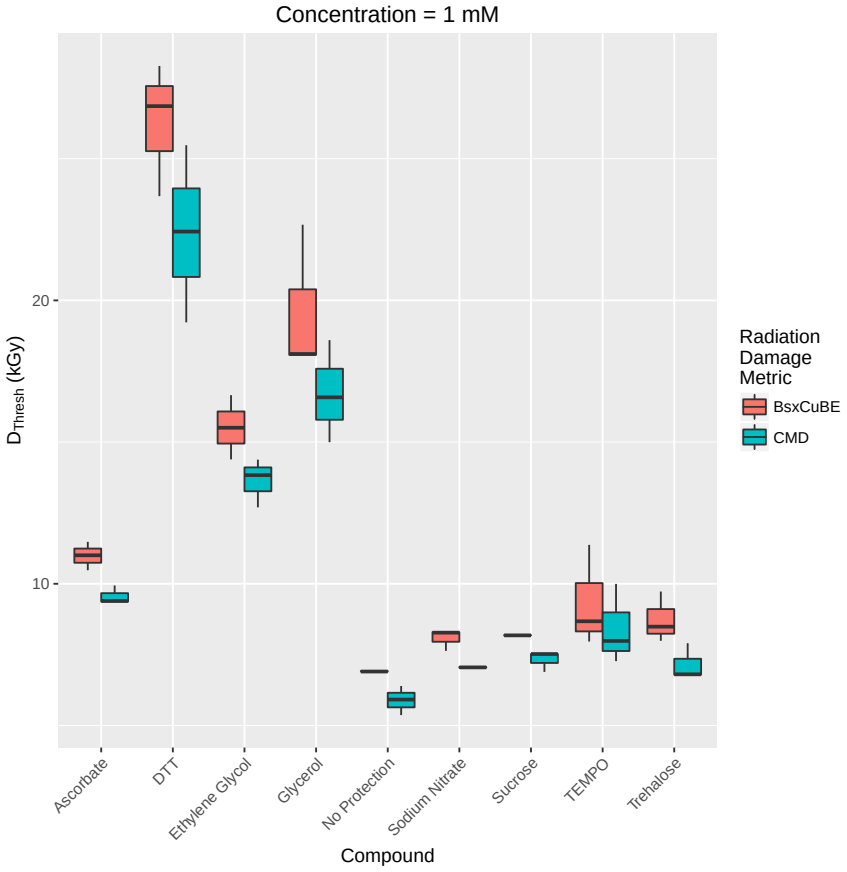
The fact that performing the pairwise comparisons with different reference frames gives varying results suggests that more information can be gained by analysing the results from all possible pairwise comparisons. Figure 6.20 is a heat map showing the results from all possible pairwise frame comparisons. The first row shows the results that were obtained using frame 1 as the reference frame. It shows the *CMD* metric highlighting frames becoming dissimilar very early on in the experiment (frame 8 out of 120). However, there is a large square region of similar frames in the top left of the map suggesting that the best results may be obtained by merging the frames from the larger region. This region notably does not include the first frame. The structure of this map could suggest that there are fast changes occurring in the sample early on in the experiment until a relatively stable molecular conformation is reached. Perhaps merging frames from these different regions may result in molecular envelopes resembling the molecules in different conformations. It may be the case that some of these states are functionally important. Another plausible explanation for these results could be that there were errors at the beginning of the measurements.

Using dose value and frame number as damage thresholds

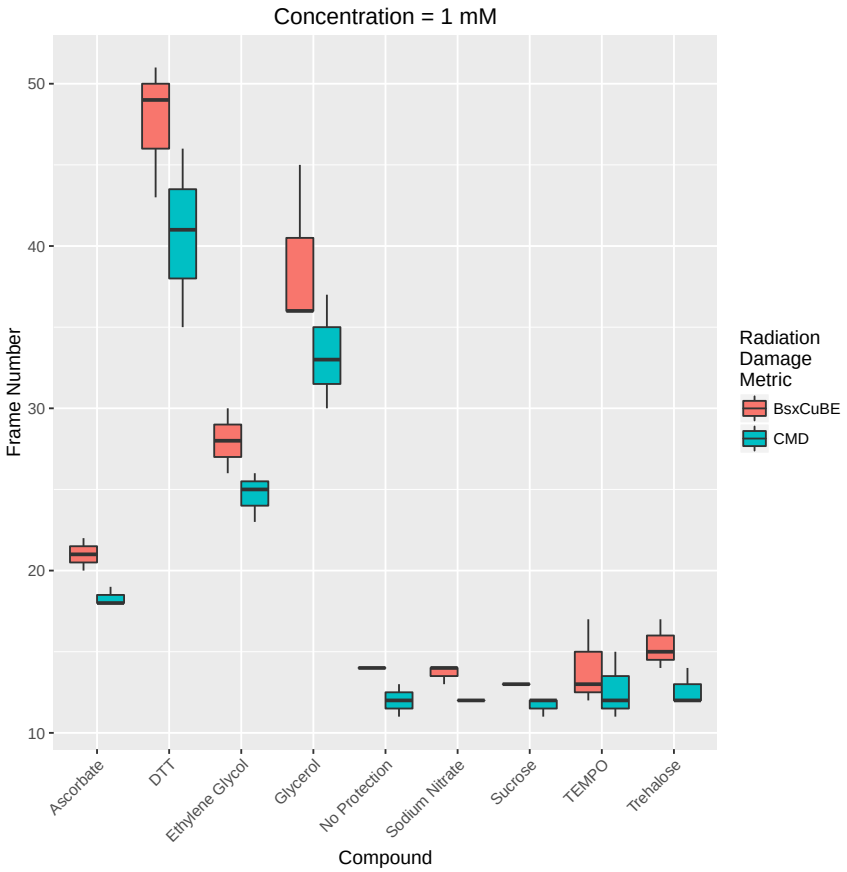
Using the dose as the metric to evaluate the efficacy of the different radioprotectants means that the results are normalised for the energy absorption of the compounds. It is sensible to ask whether this normalisation changes the conclusions that would be drawn if the frame number alone was used as the threshold metric. Figure 6.21 shows the results for the 1 mM and the 10 mM concentrations of radioprotectants with dose value and frame number as the damage thresholds. For the more radiation tolerant samples (ascorbate, ethylene glycol, glycerol and even DTT) there is not much difference in using either the dose or the frame number. However, for the less efficient radioprotectants (sucrose, TEMPO and trehalose) there is a significant difference. If the frame number is used (Figures 6.21b, 6.21d) the order of the relative efficacies of those radioprotectants is not too obvious. However, if the dose value is used instead (Figures 6.21a, 6.21c) then the order becomes much clearer, particularly at 1 mM concentration. The spread of the threshold values for those compounds is also smaller using dose instead of frame number for the 10 mM concentration.

Concentration dependence

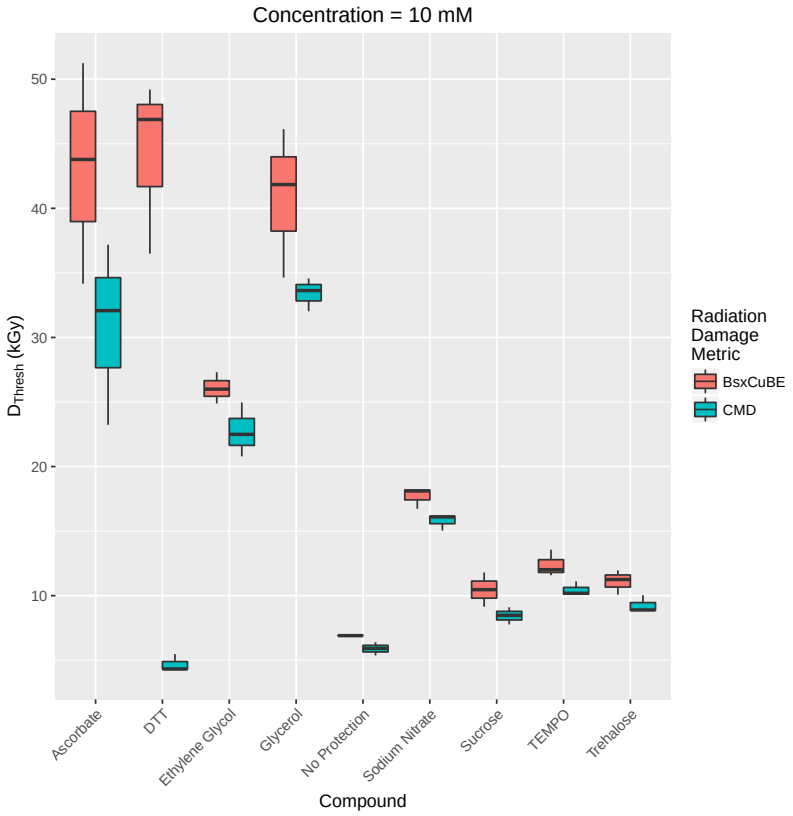
As a result of this work, a new metric, RD onset ratio, was developed to assess the change in radiation tolerance in the sample with the radioprotectant added compared to no protection. This metric is defined as the ratio of the median threshold value with added radioprotectant to the median threshold value with no protection using the *CMD* metric. 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 Figure 6.22. Significant concentration dependence can be observed for several radioprotectants. In particular, ascorbate, glycerol, and sodium nitrate all exhibit a strong positive concentration dependence i.e. the higher the concentration, the better the protection ability. DTT exhibits the opposite behaviour. At low concentrations DTT exhibits 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 RD onset ratio is less than 2. This suggests that these radioprotectants are not very efficient at increasing the dose tolerance of the sample. The RD onset ratio for ethylene


Concentration = 1 mM
D_{Thresh} (kGy)
20
10
Ascorbate
DTT
Ethylene Glycol
Glycerol
No Protection
Sodium Nitrate
Sucrose
TEMPO
Trehalose
Compound
Radiation Damage Metric
BsxCuBE
CMD

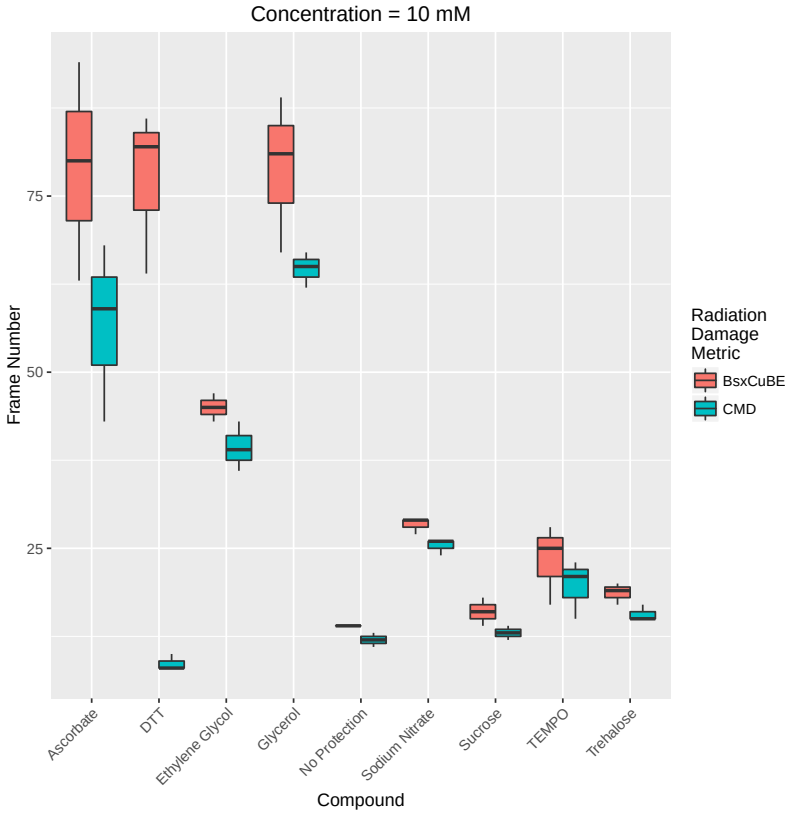
(a)


Concentration = 1 mM
Frame Number
50
40
30
20
10
Ascorbate
DTT
Ethylene Glycol
Glycerol
No Protection
Sodium Nitrate
Sucrose
TEMPO
Trehalose
Compound
Radiation Damage Metric
BsxCuBE
CMD

(b)



(c)



(d)

Figure 6.21: Expt 2: Comparing dose and frame number as the metric by which to track radiation damage in SAXS for all SAXS repeats. Figures (a) and (c) use the dose value as the y-axis whereas (b) and (d) use the frame number. Each box plot is created from the threshold dose values for three different runs of the same radioprotectant compound. Pink boxes correspond to the *BsxCuBE* metric, blue boxes correspond to the *CMD* metric.

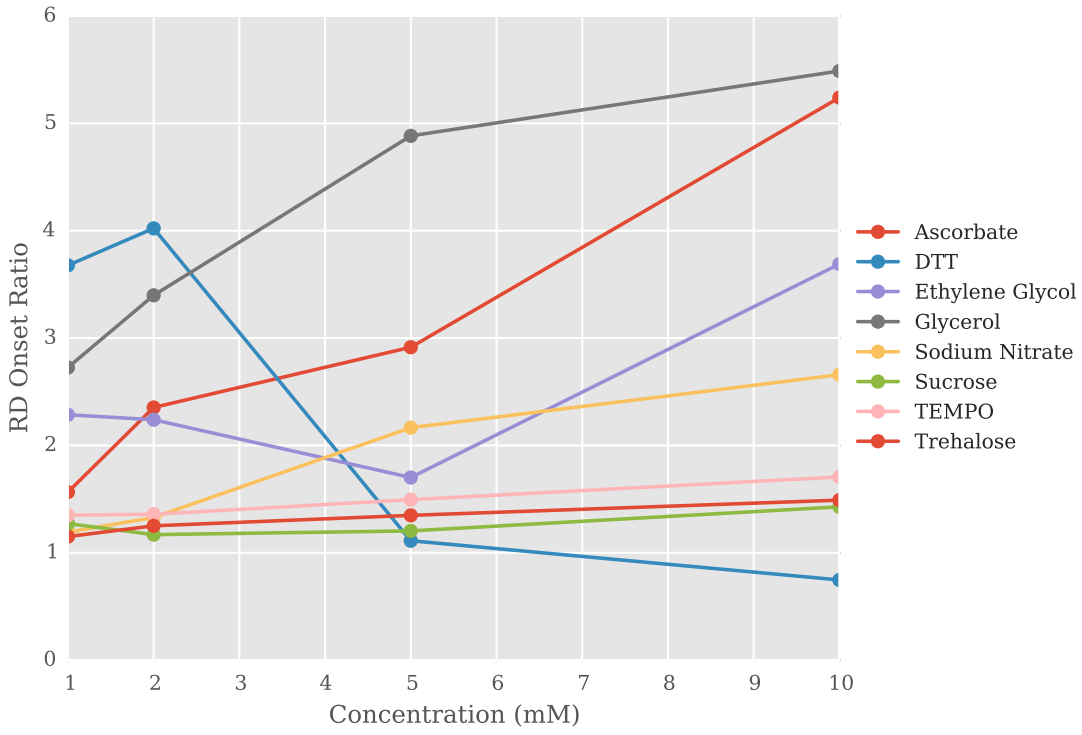


Figure 6.22: Expt 2: RD onset ratio against concentration for the 8 radioprotectants.

glycol decreases as the concentration increases from 1 mM 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.

Relative radioprotectant efficacy

The results shown in Figure 6.22 indicate that the most effective radioprotectant varies depending on the concentration of the compound used in the sample. The *BsxCuBE* and *CMD* metrics show agreement that at low concentrations (1 mM and 2 mM) DTT is the most effective radioprotectant. However, at higher concentrations these metrics disagree. At 5 mM and 10 mM the *BsxCuBE* metric suggests that the most effective radioprotectant is still DTT. On the other hand, the *CMD* metric suggests that the most effective radioprotectant is glycerol. Furthermore, it also suggests that DTT is the least effective radioprotectant. It is important to take into account that using a different reference frame to perform pairwise comparisons can result in different conclusions with the *CMD* metric as shown in Figures 6.19 and 6.20.

The *CMD* metric result also disagrees with the results from Expt 1 where ethylene glycol was found to be the most effective radioprotectant at 5 mM. Furthermore, in Expt 1 ascorbate was the one of the least effective radioprotectants. It only outperformed sucrose and trehalose. Whereas in Expt 2, ascorbate was the second most effective radioprotectant when used at a 5 mM concentration. These results could be caused by the fact that the beam profile was not well characterised in the first experiment as opposed to the second.

An important difference between the results reported for Expt 1 and Expt 2 is that the dose thresholds for Expt 1 are consistently higher than for Expt 2. This is not surprising according to Figure 6.16 because all of the threshold values occur after the fidelity curves have already began to increase steeply away from the original similarity value. Whether the resulting statistics for the structural data can be improved beyond the threshold values given by the *CMD* metric remains to be seen, but this discrepancy may go some way to explaining why the *BsxCuBE* metric gives different results for the DTT data.

6.6 Discussion

RADDOSE-3D extensions to SAXS

The analysis of data from the two SAXS experiments has relied on extensions to RADDOSE-3D to allow it to simulate SAXS data collection. In particular the three major additions were:

1. implementation of a cylindrical sample geometry,
2. determination of the sample composition given a protein concentration,
3. attenuation of the X-ray beam due to a surrounding capillary.

The cylinder geometry (diameter and height) is now an explicit input option for the user to specify in a RADDOSE-3D job. RADDOSE-3D then converts the description of the cylinder provided by the user to a polyhedral geometry description, which is the base crystal implementation. In fact the current cuboid geometry implementation does exactly the same thing. Currently the only shape modelled by RADDOSE-3D that is not converted to a polyhedral representation for the simulation is the sphere.

Some of the RADDOSE-3D code was refactored so that the library of residue information (atomic composition, molecular weight) is now stored as a separate class to the atomic information. This refactoring reduced the complexity of the implementation for RADDOSE-3D to read a FASTA sequence file. The extension for RADDOSE-3D to read any other sequence file format should now be trivial.

RADDOSE-3D now implements a *container* class which defines the material that houses the sample. The attenuation of the beam by the container is considered for all experiments. For SAXS experiments this is usually a quartz capillary, whereas for MX it is assumed to be a container that is *transparent* to X-rays. However, the user can define any container for any experiment, SAXS or MX, giving greater flexibility. An immediate application is for *in situ* RT MX data collection.

As mentioned in section 6.2.1, RADDOSE-3D has been written in a very modular style so that extensions to existing functionality can be easily incorporated. At its core, RADDOSE-3D essentially deals with three main objects: a sample, a beam and a wedge. The work on RADDOSE-3D presented here shows how the "crystal" description was extended to incorporate liquid SAXS samples, which ultimately depends on the input specification. In a similar manner the beam, which currently describes incident X-ray photons, could be extended to represent other beam types (e.g. an electron beam) if its properties are defined differently.

One of the current drawbacks of the extension of RADDOSE-3D to simulate SAXS experiments is that the implementation does not yet support sample flow. This is a common method used in SAXS experiments to mitigate radiation damage (Jeffries *et al.*, 2015). This sample flow results in any given small volume of fluid passing across the X-ray beam as described in section 6.2.6. In RADDOSE-3D this situation reduces to how long a given volume of fluid is exposed to each region of the X-ray beam, which will be dependent on the radial position from the centre of the tube due to the Poiseuille flow velocity profile (Hopkins and Thorne, 2016). In order to implement this method of data collection in RADDOSE-3D, a sufficiently small discretisation of time would be necessary so that the fluid volume does not completely 'pass' the entire beam in a single time step. Care would be needed to avoid the case where the time discretisation is so small that the computational time is unnecessarily long.

Additionally diffusive turnover, as described in section 6.2.6, could be incorporated into the dose calculations. However, beam sizes in SAXS experiments are typically quite large and exposure times are not long enough for the diffusive turnover to cause significant effect (500 μm × 250 μm with maximum exposure time 141 seconds (RT) (Jeffries *et al.*, 2015), 220 μm × 190 μm with 60 seconds exposures (100 K) (Meisburger *et al.*, 2013)). Therefore this would not be a high priority extension.

The formal tests for the SAXS extensions still need to be written before the code is officially released. These include testing that the cylindrical implementation is stable for different inputs and is aligned with the axial length perpendicular to the beam by default. It is also necessary to check that the absorption coefficient calculation works for different types of elemental and mixture compositions. However, the code is currently public on the RADDOSE-3D Github repository in the branch named "SAXS" `https://github.com/GarmanGroup/RADDOSE-3D`. Alpha versions of the SAXS extensions have already been distributed to Dr. Edward Snell (Hauptman-Woodward Institute, Buffalo) and Dr. Adam Round (ESRF, Grenoble) for user testing.

Radioprotectant analysis

Two experiments to compare the efficacy of different radioprotectants for SAXS samples have been carried out and analysed. Although methods already exist to assess the similarity of frames in a pairwise manner, few established methods deal with determination of a threshold for significant radiation damage onset. Establishing a threshold was necessary to compare the efficacy of various radioprotectants. Therefore metrics were developed to establish thresholds for merging analysis so radioprotectant efficacy could be compared using the data that were collected.

For experiment 1 the analysis performed by DATCMP gave a fidelity value which was a measure of the similarity between frame 1 and a subsequent frame. Overall these frames resembled logistic curves, and hence 4PLs were fitted to the data. Two metrics were then developed to give a merging threshold. The first metric was the total dose absorbed to reach an arbitrary fidelity value. This metric is fundamentally flawed because there is no universal value to choose and it would have to change for each comparison experiment. It is also a relative metric i.e. it only compares the efficacy of one radioprotectant with another,

but it does not give any absolute stand-alone information about whether the frames at that fidelity value are damaged or not.

The second metric developed was the dose absorbed before reaching maximum curvature. This metric is absolute in the sense that the threshold represents the point at which the fidelity values begin to deviate maximally from the initial frame(s). A downside to the maximum curvature metric is that if it is being used to ultimately determine which frames to merge, then the value that is determined may be too strict. If the overall curvature of the function is fairly small then it is possible that merging some frames beyond the threshold may still improve data quality because they may not be significantly dissimilar.

Potentially a big problem with the fitting approach overall is that it relies on a parametric function to be fitted to data where there is no guarantee that the data obey that particular form.

More recently, the DATCMP similarity analysis was extended to implement the CorMap test which assesses the similarity of two frames. As this is the default test in DATCMP, it is likely that this is the analysis that the SAXS community will adopt. During the analysis of the SAXS data, there was no open source package available to generate the plots related to the CorMap analysis, including the pairwise CorMap plots themselves (Figures 6.12b and 6.12d). Therefore a library was written to perform all analysis and also to produce the plots. This library is currently available on Github `https://github.com/JonnyCBB/AnalyseSAXSLib/blob/master/CorMapAnalysis.py` but is intended to be released as a distributable Python package.

On its own, 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 therefore may not indicate a true systematic change in the molecules of the sample. It may just be an outlier. Therefore a more robust metric was developed for this work whereby three consecutive dissimilar frames were detected to establish a merging threshold. It was used to determine that there is a concentration dependence on the efficacy of individual radioprotectants. Some radioprotectants such as glycerol, ascorbate and sodium nitrate exhibited a positive concentration dependence, whereas DTT showed a negative concentration dependence. It was also found that glycerol, ascorbate and

ethylene glycol were the most effective scavengers at high concentrations (10 mM). These compounds are all hydroxyl scavengers (k = 1 - 5 $\times$ 10^9 M^{-1} s^{-1} (Garrison, 1987)), a radical species which is known to be mobile at room temperature. On the other hand sodium nitrate, which is an electron scavenger (k = 10^8 - 10^9 M^{-1} s^{-1} (Garrison, 1987; Allan *et al.*, 2012)), generally does not protect as well. DTT was found to be the most effective radioprotectant at the lower concentrations of 1 mM and 2 mM, which were tested.

One aspect that was not considered in the analysis is how the radioprotectants affect the environment of the sample. For example, it is known that DTT reduces disulphide bonds and undergoes oxidation. Glycerol on the other hand, is known to increase the noise and hence reduce the observable signal obtained from the sample (Jeffries *et al.*, 2015). These are considerations that the experimenter must also take into account when deciding on which radioprotectant compound to use.

The heatmap shown in Figure 6.20 displays some exciting possibilities for data analysis. Firstly, it shows that pairwise comparisons with the first frame alone may not necessarily be the best approach to get all of the information possible from SAXS data. Secondly it shows distinct regions of frame similarity. These regions may correspond to distinct and functionally important molecular conformations that would otherwise not be noticed. This finding warrants further analysis to determine how physiologically significant these regions are.

CHAPTER 7

Conclusions

7.1 Radiation damage in MX

Radiation damage is the major cause of failed structure solution in MX. In particular, global radiation damage, which is largely characterised by an overall decay of reflection intensities with increasing X-ray exposure, limits the amount of useful data that can be collected in a diffraction experiment. The work presented in this thesis builds on existing research in MX for quantifying and correcting for global radiation damage.

7.1.1 Extending the DWD metric

In a typical MX experiment, a crystal with a complex geometry is exposed to an X-ray beam with a non-uniform profile via an exposure strategy, which is sometimes non-standard, e.g. helical scanning. Each of these factors can lead to inhomogeneous dose distributions within the crystal. Zeldin *et al.* developed the diffraction weighted dose (DWD), a dose metric designed to account for the dose distribution (Zeldin *et al.*, 2013a). However, the DWD does not account for a loss in diffraction efficiency of a damaged crystal. In this thesis, three dose decay models (DDMs) were assessed for their ability to describe the relative intensity decay for data collected on several insulin crystals. One of these models, *RDE Leal*, which is a normalised form of the DDM proposed by Leal *et al.* (Leal *et al.*, 2012), was used to represent the relative diffraction efficiency (RDE). An η term, which is a function of this RDE, was then added as a weighting term to the standard DWD to account for the effect of the loss in diffraction efficiency of a damaged crystal. The various η forms tested were:

- η = 1. (This is the same as the DWD originally developed in Zeldin *et al.* (2013)),
- η = RDE (Decreasing function),
- η = 1 - RDE (Increasing function).

Inclusion of the increasing and decreasing terms did not result in a decreased variability of the relative intensity as hypothesised by Zeldin *et al.*, but this work does suggest that two metrics should be considered. One metric should represent the damaged state of the crystal, which should monotonically increase with increasing X-ray exposure. This is typically the type of dose metric used in all radiation damage studies to date (maximum dose, which

was reported by RADDOSE v1-3, average dose over the whole crystal and the DWD with η = 1). The other metric would represent the diffraction quality from the crystal using a decreasing η form to weight the DWD. This metric has more complicated behaviour and can decay back towards zero as the X-ray exposure increases. In this case, images that result in similar DWD values should have similar diffraction quality, thus even images that are collected late in the experiment could have better diffraction quality than images collected in the middle of the experiment. This suggests that this metric could be used, in addition to existing methods, to determine which images to merge to obtain good quality, reliable data. To the author's knowledge, this type of metric has not been used previously.

7.1.2 Zero-dose extrapolation

The decay of reflection intensities as a result of global radiation damage can hinder structure solution by swamping the phasing signal in experimental phasing experiments. Scaling algorithms attempt to correct for the overall intensity decay but they do not account for reflection specific intensity changes, which can be non-monotonic. Previous studies have attempted to correct for the specific intensity decay (zero-dose extrapolation), showing that improvement in the phasing signal can be achieved (Diederichs *et al.*, 2003; Diederichs, 2006). However, zero-dose extrapolation is not commonly used (Borek *et al.*, 2007), and is not very effective with low multiplicity data. In this thesis, a zero-dose extrapolation model is presented that uses the Leal *et al.* DDM for the traditional regression based extrapolation. This model was able to capture the non-monotonic behaviour of reflection intensities. The procedure also incorporated several quality checks to ensure the reliability of the regression fits. Furthermore, an additional probabilistic extrapolation routine was developed to perform zero-dose extrapolation for reflections with a small number of observations, thus addressing the problem of performing extrapolation for low multiplicity data. Further investigation of this method is required before it can be used reliably for any crystal. In particular, it would benefit from the development of:

- a method for learning the weight of the scaling factor used for the probabilistic extrapolation,
- an outlier rejection algorithm for poor observation extrapolations,

- a metric to assess the overall quality of the probabilistic extrapolation.

7.1.3 Measuring X-ray beam profiles

As mentioned above, the precise X-ray beam profile is a critical factor that determines the overall dose distribution in a crystal during an MX experiment. The RADDOSE-3D dose calculation software has the capability to use experimentally measured beam profiles to accurately calculate the dose distribution. Various methods for experimentally measuring the X-ray beam profiles exist, but each one requires different preprocessing before it can faithfully represent the actual beam profile. In this thesis, three methods for measuring beam profiles and processing the resulting data were investigated. Aperture scans were performed at DLS beamline I02, where a piece of steel with a circular hole was translated across the X-ray beam, both vertically and horizontally, and measurements of the current from a diode was taken at regular intervals. This only gives 1D data and therefore has to be converted to a 2D beam profile. This was achieved by fitting a 2D Gaussian to the aperture data. It was found that deconvoluting the resulting Gaussian profile with the aperture size did not change the overall beam profile significantly. Another method employed to measure the beam profile at the PETRA III beamline P14, was to use a scintillator combined with an Allied Vision GC1350C CCD camera, resulting in a 2D image of the X-ray beam. The advantage of this method is that it is not necessary to create a 2D profile. The major disadvantage is that the image has background noise, which, if removed incorrectly, was shown to result in large errors in the dose calculation. Various methods were explored to objectively remove the background, with some methods resulting in similar calculated $D_{1/2}$ values, a radiation damage sensitivity metric that represents the absorbed dose for which the relative intensity has decayed to half of its initial value. However it is not objectively clear which method of removing the background is best, so the experimenter should use information that is known about the collimation of the beam to decide where the background is in an image. Finally, the beam profile measurements carried out at ESRF beamline BM29, used several aperture scans, both horizonally and vertically, to obtain several 1D slices of the beam profile. These data were then interpolated, as opposed to being modelled with a regular mathematical function. The benefit of this approach is that the background does not have to be removed because the current actually decays to negligible levels. It is also possible to represent and complex, irregular beam profiles. For this reason it is recommended that beam profile mea-

surements are performed by collecting several vertical and horizontal aperture scans.

7.2 Data reduction

Radiation damage correction has always focussed on attempting to correct the intensities that are observed in the diffraction experiment. These intensities are ultimately generated by the contents of the crystal and so an alternative perspective to correct for radiation damage is to describe (probabilistically) the changes that are occurring to the crystal. Hidden Markov models are mathematical frameworks that track the hidden state of a system according to observations that are generated from it. In this thesis, this framework is applied to MX, where the state of the system is the crystal state represented by the set of structure factor amplitudes and the observations are the intensities. The process and observation functions, along with their respective uncertainties (covariances), are defined as an exponential decay (temperature factor) and the product of the square of the amplitude with a scale factor, respectively. Applying the unscented Kalman filter (forward pass) with the unscented Rauch-Tung-Streibel smoother (backwards pass) is known as the forward-backward algorithm (FBA), which gives time resolved estimates of the amplitude values throughout the experiment. Therefore, it is possible to use the amplitudes at each time point for phasing and refinement to generate a set of electron density maps from a single dataset. The set of amplitudes estimated at the first time point were then used to obtain interpretable electron density maps for insulin and the C.Esp1396I protein-DNA complex by molecular replacement, although the refinement statistics for the latter are not very good and would require manual refinement. The resulting statistics from using this method are comparable to the statistics that result from structure solution using programs in the current data reduction pipeline (AIMLESS and CTRUNCATE). In addition to producing time resolved amplitude estimates, the error estimates are also explicitly calculated in the FBA as a combination of the uncertainty in the measurement and the uncertainty in the (damage) process. This means that the error estimates should be more representative of the expected error, which may lead to more reliable results in experimental phasing, but this is not yet verified and still needs to be investigated.

The HMM representation is not only useful for data reduction. In fact the idea of the HMM

in crystallography was spawned from the idea of modelling radiation damage in structure refinement (Garib Murshudov, personal communication). The time resolved nature of the HMM means that refinement would directly lead to electron density movies. The data reduction pipeline presented in this thesis is a successful proof of concept for the applicability of the HMM to crystallographic data. Conceptually, the extension of the HMM to structure refinement is not too difficult. Essentially the crystal state is no longer a set of structure factor amplitudes, but instead the crystal state is more accurately described by the set of structure factors. The process function would need to be extended to incorporate a model for the change in atomic positions as well as the B factor. In general the functions describing the B factor and the change in atomic positions are likely to be parameterised by the dose, either implicitly or explicitly. However, for standard data collection where the crystal is rotated around a single axis and images are collected at regular intervals, this can be a function of the frame number, as was the case for the model described in this thesis.

7.3 Quantifying radiation damage in SAXS experiments

The work presented in this thesis presents a significant step forward in automating and exploring radiation damage in SAXS experiments. Firstly, RADDOSE-3D has now been extended to allow dose calculations for SAXS experiments, just as easily as can be carried out for MX experiments. Prior to this there was no open source, dose calculation standard for SAXS. Furthermore, a Python library has been written to allow exploratory data analysis of SAXS datasets. It provides a wrapper for executing DATCMP, which performs the similarity analysis described in section 6.4.2, and data visualisation tools. These tools were used to analyse datasets collected on glucose isomerase with different concentrations of radioprotectant compounds. Glucose was found to be the most effective radioprotectant overall. Additionally it was found that DTT exhibits odd behaviour, particularly at high concentrations. One of the most intriguing visualisation tools in the Python library is the heatmap. For DTT it showed regions of frame similarity which cannot be easily determined via other analytical methods. This indicates that it may have implications for deciding which frames experimenters should merge to improve data quality, and possibly distinguish regions of different conformational states. Further work should be done to verify the importance of these regions.

Bibliography

ABERGEL, C. Molecular replacement: tricks and treats. *Acta Crystallographica Section D: Biological Crystallography*, **69**(11):2167–2173, 2013.

ABRAHAMS, S. C. International Union of Crystallography Commission on Crystallographic Apparatus single-crystal radiation damage survey. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, **29**(2):111–116, 1973. doi:10.1107/S056773947300032X.

ABRAHAMS, S. C. AND MARSH, P. Anisotropy in the variation of serially-measured integrated intensities. *Acta Crystallographica Section A: Foundations of Crystallography*, **43**(2):265–269, 1987.

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., *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallographica Section D: Biological Crystallography*, **66**(2):213–221, 2010.

ALLAN, E. G., KANDER, M. C., CARMICHAEL, I., AND GARMAN, E. F. To scavenge or not to scavenge, that is STILL the question. *Journal of Synchrotron Radiation*, **20**(1):23–36, 2012.

AXFORD, D., FOADI, J., HU, N.-J., CHOUDHURY, H., IWATA, S., BEIS, K., EVANS, G., AND ALGUEL, Y. Structure determination of an integral membrane protein at room temperature from crystals in situ. *Acta Crystallographica Section D: Biological Crystallography*, **71**(6):1228–1237, 2015.

AXFORD, D., OWEN, R. L., AISHIMA, J., FOADI, J., MORGAN, A. W., ROBINSON, J. I., NETTLESHIP, J. E., OWENS, R. J., MORAES, I., FRY, E. E., *et al.* In situ macromolecular crystallography using microbeams. *Acta Crystallographica Section D: Biological Crystallography*, **68**(5):592–600, 2012.

BAI, X.-C., MCMULLAN, G., AND SCHERES, S. H. How cryo-EM is revolutionizing structural biology. *Trends in Biochemical Sciences*, **40**(1):49–57, 2015.

BARKER, A. I., SOUTHWORTH-DAVIES, R. J., PAITHANKAR, K. S., CARMICHAEL, I., AND GARMAN, E. F. Room-temperature scavengers for macromolecular crystallography: increased lifetimes and modified dose dependence of the intensity decay. *Journal of Synchrotron Radiation*, **16**(2):205–216, 2009.

BERG, J., TYMOCZKO, J., AND STRYER, L. Biochemistry, Fifth Edition. Biochemistry. W. H. Freeman, 2002. ISBN 9780716746843. URL https://books.google.co.uk/books?id=qs5pAAAAMAAJ.

BEZANSON, J., EDELMAN, A., KARPINSKI, S., AND SHAH, V. B. Julia: A fresh approach to numerical computing. *arXiv preprint arXiv:1411.1607*, 2014.

BEZANSON, J., KARPINSKI, S., SHAH, V. B., AND EDELMAN, A. Julia: A fast dynamic language for technical computing. *arXiv preprint arXiv:1209.5145*, 2012.

BIJVOET, J. M. Structure of optically active compounds in the solid state. *Nature*, **173**:888–891, 1954.

BLAKE, C. C. AND PHILLIPS, D. C. Effects of X-Irradiation on Single Crystals of Myoglobin. *Proceedings of the Symposium on the Biological effects of Ionizing Radiation at the Molecular Level*, pages 183–191, 1962.

BLANCHET, C. E. AND SVERGUN, D. I. Small-angle X-ray scattering on biological macromolecules and nanocomposites in solution. *Annual Review of Physical Chemistry*, **64**:37–54, 2013.

BLESSING, R. H., GUO, D. Y., AND LANGS, D. A. Intensity statistics and normalization. In Direct Methods for Solving Macromolecular Structures, pages 47–71. Springer, 1998.

BLUNDELL, T. L. AND JOHNSON, L. N. Protein Crystallography. Molecular biology. Academic Press, 1976. ISBN 9780121083502. URL https://books.google.co.uk/books?id=o4FkYPmFxKwC.

BOREK, D., GINELL, S. L., CYMBOROWSKI, M., MINOR, W., AND OTWINOWSKI, Z. The many faces of radiation-induced changes. *Journal of Synchrotron Radiation*, **14**(1):24–33, 2007.

BOURENKOV, G. P. AND POPOV, A. N. Optimization of data collection taking radiation damage into account. *Acta Crystallographica Section D: Biological Crystallography*, **66**(4):409–419, 2010.

BOUTET, S., LOMB, L., WILLIAMS, G. J., BARENDS, T. R. M., AQUILA, A., DOAK, R. B., WEIERSTALL, U., DEPONTE, D. P., STEINBRENER, J., SHOEMAN, R. L., MESSERSCHMIDT, M., BARTY, A., WHITE, T. A., KASSEMEYER, S., KIRIAN, R. A., SEIBERT, M. M., MONTANEZ, P. A., KENNEY, C., HERBST, R., HART, P., PINES, J., HALLER, G., GRUNER, S. M., PHILIPP, H. T., TATE, M. W., HROMALIK, M., KOERNER,

L. J., VAN BAKEL, N., MORSE, J., GHONSALVES, W., ARNLUND, D., BOGAN, M. J., CALEMAN, C., FROMME, R., HAMPTON, C. Y., HUNTER, M. S., JOHANSSON, L. C., KATONA, G., KUPITZ, C., LIANG, M., MARTIN, A. V., NASS, K., REDECKE, L., STELLATO, F., TIMNEANU, N., WANG, D., ZATSEPIN, N. A., SCHAFER, D., DEFEVER, J., NEUTZE, R., FROMME, P., SPENCE, J. C. H., CHAPMAN, H. N., AND SCHLICHTING, I. High-resolution protein structure determination by serial femtosecond crystallography. *Science*, **337**(6092):362–364, 2012.

BOWLER, M. W., NURIZZO, D., BARRETT, R., BETEVA, A., BODIN, M., CASEROTTO, H., DELAGENIÈRE, S., DOBIAS, F., FLOT, D., GIRAUD, T., *et al.* MASSIF-1: a beamline dedicated to the fully automatic characterization and data collection from crystals of biological macromolecules. *Journal of Synchrotron Radiation*, **22**(6):1540–1547, 2015.

BROCKHAUSER, S., DI MICHIEL, M., MCGEEHAN, J. E., MCCARTHY, A. A., AND RAVELLI, R. B. G. X-ray tomographic reconstruction of macromolecular samples. *Journal of Applied Crystallography*, **41**(6):1057–1066, 2008.

BROOKS-BARTLETT, J. C. AND GARMAN, E. F. The Nobel Science: One Hundred Years of Crystallography. *Interdisciplinary Science Reviews*, **40**(3):244–264, 2015.

BRÜNGER, A. T. [19] Free R value: Cross-validation in crystallography. *Methods in Enzymology*, **277**:366–396, 1997.

BURMEISTER, W. P. Structural changes in a cryo-cooled protein crystal owing to radiation damage. *Acta Crystallographica Section D: Biological Crystallography*, **56**(3):328–341, 2000.

BURY, C., GARMAN, E. F., GINN, H. M., RAVELLI, R. B., CARMICHAEL, I., KNEALE, G., AND MCGEEHAN, J. E. Radiation damage to nucleoprotein complexes in macromolecular crystallography. *Journal of Synchrotron Radiation*, **22**(2):213–224, 2015.

CHAN, T. F. AND VESE, L. A. Active contours without edges. *IEEE Transactions on Image Processing*, **10**(2):266–277, 2001.

CHAPMAN, H. N., CALEMAN, C., AND TIMNEANU, N. Diffraction before destruction. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **369**(1647):20130313, 2014.

CHAPMAN, H. N., FROMME, P., BARTY, A., WHITE, T. A., KIRIAN, R. A., AQUILA, A., HUNTER, M. S., SCHULZ, J., DEPONTE, D. P., WEIERSTALL, U., DOAK, R. B., MAIA, F. R. N. C., MARTIN, A. V., SCHLICHTING, I., LOMB, L., COPPOLA, N., SHOEMAN, R. L., EPP, S. W., HARTMANN, R., ROLLES, D.,

RUDENKO, A., FOUCAR, L., KIMMEL, N., WEIDENSPOINTNER, G., HOLL, P., LIANG, M., BARTHELMESS, M., CALEMAN, C., BOUTET, S., BOGAN, M. J., KRZYWINSKI, J., BOSTEDT, C., BAJT, S., GUMPRECHT, L., RUDEK, B., ERK, B., SCHMIDT, C., HÖMKE, A., REICH, C., PIETSCHNER, D., STRÜDER, L., HAUSER, G., GORKE, H., ULLRICH, J., HERRMANN, S., SCHALLER, G., SCHOPPER, F., SOLTAU, H., KÜHNEL, K.-U., MESSERSCHMIDT, M., BOZEK, J. D., HAU-RIEGE, S. P., FRANK, M., HAMPTON, C. Y., SIERRA, R. G., STARODUB, D., WILLIAMS, G. J., HAJDU, J., TIMNEANU, N., SEIBERT, M. M., ANDREASSON, J., ROCKER, A., JÖNSSON, O., SVENDA, M., STERN, S., NASS, K., ANDRITSCHKE, R., SCHRÖTER, C.-D., KRASNIQI, F., BOTT, M., SCHMIDT, K. E., WANG, X., GROTJOHANN, I., HOLTON, J. M., BARENDS, T. R. M., NEUTZE, R., MARCHESINI, S., FROMME, R., SCHORB, S., RUPP, D., ADOLPH, M., GORKHOVER, T., ANDERSSON, I., HIRSEMANN, H., POTDEVIN, G., GRAAFSMA, H., NILSSON, B., AND SPENCE, J. C. H. Femtosecond X-ray protein nanocrystallography. *Nature*, **470**(7332):73–77, 2011.

CHEN, V. B., ARENDALL, W. B., HEADD, J. J., KEEDY, D. A., IMMORMINO, R. M., KAPRAL, G. J., MURRAY, L. W., RICHARDSON, J. S., AND RICHARDSON, D. C. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallographica Section D: Biological Crystallography*, **66**(1):12–21, 2010.

CHEN, Z. Bayesian filtering: From Kalman filters to particle filters, and beyond. *Statistics*, **182**(1):1–69, 2003.

COLEMAN, T. F. AND LI, Y. An interior trust region approach for nonlinear minimization subject to bounds. *SIAM Journal on Optimization*, **6**(2):418–445, 1996.

COQUELLE, N., FIORAVANTI, E., WEIK, M., VELLIEUX, F., AND MADERN, D. Activity, stability and structural studies of lactate dehydrogenases adapted to extreme thermal environments. *Journal of Molecular Biology*, **374**(2):547–562, 2007.

COXETER, H. Regular Polytopes. Dover books on advanced mathematics. Dover Publications, 1973. ISBN 9780486614809. URL `https://books.google.co.uk/books?id=iWvXsVInpgMC`.

CRESSIE, N. AND WIKLE, C. Statistics for Spatio-Temporal Data. Wiley, 2015. ISBN 9781119243045. URL `https://books.google.co.uk/books?id=4L_dCgAAQBAJ`.

CRICK, F. H. C. AND MAGDOFF, B. S. The theory of the method of isomorphous replacement for protein crystals. I. *Acta Crystallographica*, **9**(11):901–908, 1956.

CROMER, D. T. AND MANN, J. B. X-ray scattering factors computed from numerical Hartree–Fock wave functions. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, **24**(2):321–324, 1968.

DEN DEKKER, A. J. AND SIJBERS, J. Data distributions in magnetic resonance images: A review. *Physica Medica*, **30**(7):725–741, 2014.

DIEDERICHS, K. Some aspects of quantitative analysis and correction of radiation damage. *Acta Crystallographica Section D: Biological Crystallography*, **62**(1):96–101, 2006.

DIEDERICHS, K., MCSWEENEY, S., AND RAVELLI, R. B. Zero-dose extrapolation as part of macromolecular synchrotron data reduction. *Acta Crystallographica Section D: Biological Crystallography*, **59**(5):903–909, 2003.

DRENTH, J. Principles of Protein X-Ray Crystallography. Springer Advanced Texts in Chemistry. Springer-Verlag GmbH, 1999. ISBN 9780387985879. URL `http://books.google.co.uk/books?id=ABjCdPuly4IC`.

DRENTH, J. Introduction to basic crystallography. In International Tables for Crystallography Volume F: Crystallography of biological macromolecules, pages 45–63. Springer, 2006.

EVANS, P. Scaling and assessment of data quality. *Acta Crystallographica Section D: Biological Crystallography*, **62**(1):72–82, 2006.

EVANS, P. R. An introduction to data reduction: space-group determination, scaling and intensity statistics. *Acta Crystallographica Section D: Biological Crystallography*, **67**(4):282–292, 2011.

EVANS, P. R. AND MURSHUDOV, G. N. How good are my data and what is the resolution? *Acta Crystallographica Section D: Biological Crystallography*, **69**(7):1204–1214, 2013.

FLETTERICK, R. J., SYGUSCH, J., MURRAY, N., MADSEN, N. B., AND JOHNSON, L. N. Low-resolution structure of the glycogen phosphorylase a monomer and comparison with phosphorylase b. *Journal of Molecular Biology*, **103**(1):1–13, 1976.

FRANKE, D., JEFFRIES, C. M., AND SVERGUN, D. I. Correlation Map, a goodness-of-fit test for one-dimensional X-ray scattering spectra. *Nature Methods*, **12**(5):419–422, 2015.

FRENCH, S. AND WILSON, K. On the treatment of negative intensity observations. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, **34**(4):517–525, 1978.

GARMAN, E. Cool data: quantity AND quality. *Acta Crystallographica Section D: Biological Crystallography*, **55**(10):1641–1653, 1999.

GARMAN, E. F. Radiation damage in macromolecular crystallography: what is it and why should we care? *Acta Crystallographica Section D: Biological Crystallography*, **66**(4):339–351, 2010.

GARMAN, E. F. Developments in X-ray Crystallographic Structure Determination of Biological Macromolecules. *Science*, **343**(6175):1102–1108, 2014.

GARMAN, E. F. AND NAVE, C. Radiation damage in protein crystals examined under various conditions by different methods. *Journal of Synchrotron Radiation*, **16**(2):129–132, 2009.

GARMAN, E. F. AND SCHNEIDER, T. R. Macromolecular cryocrystallography. *Journal of Applied Crystallography*, **30**(3):211–237, 1997.

GARRISON, W. M. Reaction mechanisms in the radiolysis of peptides, polypeptides, and proteins. *Chemical Reviews*, **87**(2):381–398, 1987.

GERSTEL, M., DEANE, C. M., AND GARMAN, E. F. Identifying and quantifying radiation damage at the atomic level. *Journal of Synchrotron Radiation*, **22**(2):201–212, 2015.

GONZÁLEZ, R. AND WOODS, R. Digital image processing. Addison-Wesley world student series. Addison-Wesley, 1992. ISBN 9780201508031. URL `http://books.google.co.uk/books?id=C_FRAAAAMAAJ`.

GRÄSLUND, S., NORDLUND, P., WEIGELT, J., BRAY, J., GILEADI, O., KNAPP, S., OPPERMANN, U., ARROWSMITH, C., HUI, R., MING, J., *et al.* Protein production and purification. *Nature Methods*, **5**(2):135–146, 2008.

GRISHAEV, A. Sample Preparation, Data Collection, and Preliminary Data Analysis in Biomolecular Solution X-Ray Scattering. *Current Protocols in Protein Science*, **17**:17–14, 2012.

HEGYI, H. AND GERSTEIN, M. The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *Journal of Molecular Biology*, **288**(1):147–164, 1999.

HENDERSON, R. Cryo-protection of protein crystals against radiation damage in electron and X-ray diffraction. *Proceedings of the Royal Society: Biological Sciences*, pages 6–8, 1990.

HENDRICKSON, W. A. Radiation damage in protein crystallography. *Journal of Molecular Biology*, **106**(3):889–893, 1976. ISSN 00222836. doi:10.1016/0022-2836(76)90271-0.

HENDRICKSON, W. A. Determination of macromolecular structures from anomalous diffraction of synchrotron radiation. *Science*, **254**(5028):51–58, 1991.

HENDRICKSON, W. A., LOVE, W. E., AND KARLE, J. Crystal structure analysis of sea lamprey hemoglobin at 2 Å resolution. *Journal of Molecular Biology*, **74**(3):331–361, 1973.

HOLTON, J. M. A beginner's guide to radiation damage. *Journal of Synchrotron Radiation*, **16**(2):133–142, 2009.

HOLTON, J. M., CLASSEN, S., FRANKEL, K. A., AND TAINER, J. A. The R-factor gap in macromolecular crystallography: an untapped potential for insights on accurate structures. *FEBS Journal*, **281**(18):4046–4060, 2014.

HOLTON, J. M. AND FRANKEL, K. A. The minimum crystal size needed for a complete diffraction data set. *Acta Crystallographica Section D: Biological Crystallography*, **66**(4):393–408, 2010.

HOMER, C., COOPER, L., AND GONZALEZ, A. Energy dependence of site-specific radiation damage in protein crystals. *Journal of Synchrotron Radiation*, **18**(3):338–345, 2011.

HOPE, H. Cryocrystallography of biological macromolecules: a generally applicable method. *Acta Crystallographica Section B: Structural Science*, **44**(1):22–26, 1988.

HOPKINS, J. B. AND THORNE, R. E. Quantifying radiation damage in biomolecular small-angle X-ray scattering. *Journal of Applied Crystallography*, **49**(3):880–890, 2016.

HOWELLS, M. R., BEETZ, T., CHAPMAN, H. N., CUI, C., HOLTON, J. M., JACOBSEN, C. J., KIRZ, J., LIMA, E., MARCHESINI, S., MIAO, H., SAYRE, D., SHAPIRO, D. A., SPENCE, J. C. H., AND STARODUB, D. An assessment of the resolution limitation due to radiation-damage in X-ray diffraction microscopy. *Journal of Electron Spectroscopy and Related Phenomena*, **170**(1):4–12, 2009.

HUBBELL, J. H. AND SELTZER, S. M. Tables of X-ray mass attenuation coefficients and mass energy-absorption coefficients 1 keV to 20 MeV for elements Z= 1 to 92 and 48 additional substances of dosimetric interest. Technical report, National Inst. of Standards and Technology-PL, Gaithersburg, MD (United States). Ionizing Radiation Div., 1995.

HUBBELL, J. H. AND SELTZER, S. M. X-Ray Mass Attenuation Coefficients - Table 3. National Institute of Science and Technology, 1996a. URL `http://physics.nist.gov/PhysRefData/XrayMassCoef/tab3.html`. Accessed: 9 April 2016.

HUBBELL, J. H. AND SELTZER, S. M. X-Ray Mass Attenuation Coefficients - Table 4. National Institute of Science and Technology, 1996b. URL `http://physics.nist.gov/PhysRefData/XrayMassCoef/tab3.html`. Accessed: 9 April 2016.

JEFFRIES, C. M., GRAEWERT, M. A., SVERGUN, D. I., AND BLANCHET, C. E. Limiting radiation damage for high-brilliance biological solution scattering: practical experience at the EMBL P12 beamline PETRAIII. *Journal of Synchrotron Radiation*, **22**(2):273–279, 2015.

JONES, E., OLIPHANT, T., AND PETERSON, P. {SciPy}: open source scientific tools for {Python}. 2014.

JONES, G. D. D., LEA, J. S., SYMONS, M. C. R., AND TAIWO, F. A. Structure and mobility of electron gain and loss centres in proteins. *Nature*, **330**(6150):772–773, 1987.

KABSCH, W. Integration, scaling, space-group assignment and post-refinement. *Acta Crystallographica Section D: Biological Crystallography*, **66**(2):133–144, 2010a.

KABSCH, W. XDS. *Acta Crystallographica Section D: Biological Crystallography*, **66**(2):125–132, 2010b.

KHAN, I., GILLILAN, R., KRIKSUNOV, I., WILLIAMS, R., ZIPFEL, W. R., AND ENGLICH, U. Confocal microscopy on the beamline: novel three-dimensional imaging and sample positioning. *Journal of Applied Crystallography*, **45**(5):936–943, 2012.

KMETKO, J., HUSSEINI, N. S., NAIDES, M., KALININ, Y., AND THORNE, R. E. Quantifying X-ray radiation damage in protein crystals at cryogenic temperatures. *Acta Crystallographica Section D: Biological Crystallography*, **62**(9):1030–1038, 2006.

KMETKO, J., WARKENTIN, M., ENGLICH, U., AND THORNE, R. E. Can radiation damage to protein crystals be reduced using small-molecule compounds? *Acta Crystallographica Section D: Biological Crystallography*, **67**(10):881–893, 2011.

KROJER, T. AND VON DELFT, F. Assessment of radiation damage behaviour in a large collection of empirically optimized datasets highlights the importance of unmeasured complicating effects. *Journal of Synchrotron Radiation*, **18**(3):387–397, 2011.

KUWAMOTO, S., AKIYAMA, S., AND FUJISAWA, T. Radiation damage to a protein solution, detected by synchrotron X-ray small-angle scattering: dose-related considerations and suppression by cryoprotectants. *Journal of Synchrotron Radiation*, **11**(6):462–468, 2004.

LAGARIAS, J. C., REEDS, J. A., WRIGHT, M. H., AND WRIGHT, P. E. Convergence properties of the Nelder–Mead simplex method in low dimensions. *SIAM Journal on Optimization*, **9**(1):112–147, 1998.

LASKOWSKI, R. A., MACARTHUR, M. W., MOSS, D. S., AND THORNTON, J. M. PROCHECK: a program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography*, **26**(2):283–291, 1993.

LEAL, R. M. F., BOURENKOV, G., RUSSI, S., AND POPOV, A. N. A survey of global radiation damage to 15 different protein crystal types at room temperature: a new decay model. *Journal of Synchrotron Radiation*, **20**(1):14–22, 2012.

LESLIE, A. G. W. AND POWELL, H. R. Processing diffraction data with MOSFLM. In Evolving methods for macromolecular crystallography, pages 41–51. Springer, 2007.

LUFT, J. R., WOLFLEY, J. R., SAID, M. I., NAGEL, R. M., LAURICELLA, A. M., SMITH, J. L., THAYER, M. H., VEATCH, C. K., SNELL, E. H., MALKOWSKI, M. G., *et al.* Efficient optimization of crystallization conditions by manipulation of drop volume ratio and temperature. *Protein Science*, **16**(4):715–722, 2007.

LUZZATI, V. Traitement statistique des erreurs dans la determination des structures cristallines. *Acta Crystallographica*, **5**(6):802–810, 1952.

MCCOY, A. J. Liking likelihood. *Acta Crystallographica Section D: Biological Crystallography*, **60**(12):2169–2183, 2004.

McCoy, A. J. Solving structures of protein complexes by molecular replacement with Phaser. *Acta Crystallographica Section D: Biological Crystallography*, **63**(1):32–41, 2007.

McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C., and Read, R. J. Phaser crystallographic software. *Journal of Applied Crystallography*, **40**(4):658–674, 2007.

Meents, A., Gutmann, S., Wagner, A., and Schulze-Briese, C. Origin and temperature dependence of radiation damage in biological samples at cryogenic temperatures. *Proceedings of the National Academy of Sciences of the United States of America*, **107**(3):1094–1099, 2010.

Meisburger, S. P., Warkentin, M., Chen, H., Hopkins, J. B., Gillilan, R. E., Pollack, L., and Thorne, R. E. Breaking the radiation damage limit with cryo-SAXS. *Biophysical Journal*, **104**(1):227–236, 2013.

Milne, J. L. S., Borgnia, M. J., Bartesaghi, A., Tran, E. E. H., Earl, L. A., Schauder, D. M., Lengyel, J., Pierson, J., Patwardhan, A., and Subramaniam, S. Cryo-electron microscopy–a primer for the non-microscopist. *FEBS Journal*, **280**(1):28–45, 2013.

Mitchell, E., Kuhn, P., and Garman, E. Demystifying the synchrotron trip: a first time user's guide. *Structure*, **7**(5):R111–R122, 1999.

Moré, J. J. The Levenberg-Marquardt algorithm: implementation and theory. In Numerical analysis, pages 105–116. Springer, 1978.

Murray, J. and Garman, E. Investigation of possible free-radical scavengers and metrics for radiation damage in protein cryocrystallography. *Journal of Synchrotron Radiation*, **9**(6):347–354, 2002.

Murray, J. W., Garman, E. F., and Ravelli, R. B. G. X-ray absorption by macromolecular crystals: the effects of wavelength and crystal composition on absorbed dose. *Journal of Applied Crystallography*, **37**(4):513–522, 2004.

Murshudov, G. N., Skubák, P., Lebedev, A. A., Pannu, N. S., Steiner, R. A., Nicholls, R. A., Winn, M. D., Long, F., and Vagin, A. A. REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallographica Section D: Biological Crystallography*, **67**(4):355–367, 2011.

MURSHUDOV, G. N., VAGIN, A. A., AND DODSON, E. J. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallographica Section D: Biological Crystallography*, **53**(3):240–255, 1997.

NASS, K., FOUCAR, L., BARENDS, T. R., HARTMANN, E., BOTHA, S., SHOEMAN, R. L., DOAK, R. B., ALONSO-MORI, R., AQUILA, A., BAJT, S., BEAN, R., BEYERLEIN, K. R., BUBLITZ, M., DRACHMANN, N., GREGERSEN, J., JÖNSSON, H. O., KABSCH, W., KASSEMEYER, S., KOGLIN, J. E., KRUMREY, M., MATTLE, D., M., M., NISSEN, P., REINHARD, L., SITSEL, O., SOKARAS, D., WILLIAMS, G. J., HAU-RIEGE, S., TIMNEANU, N., CALEMAN, C., CHAPMAN, H. N., BOUTET, S., AND SCHLICHTING, I. Indications of radiation damage in ferredoxin microcrystals using high-intensity X-FEL beams. *Journal of Synchrotron Radiation*, **22**(2), 2015.

NAVE, C. Radiation damage in protein crystallography. *Radiation Physics and Chemistry*, **45**(3):483–490, 1995.

NAVE, C. AND GARMAN, E. F. Towards an understanding of radiation damage in cryocooled macromolecular crystals. *Journal of Synchrotron Radiation*, **12**(3):257–260, 2005.

O'NEILL, P., STEVENS, D. L., AND GARMAN, E. Physical and chemical considerations of damage induced in protein crystals by synchrotron radiation: a radiation chemical perspective. *Journal of Synchrotron Radiation*, **9**(6):329–332, 2002.

OTWINOWSKI, Z., BOREK, D., MAJEWSKI, W., AND MINOR, W. Multiparametric scaling of diffraction intensities. *Acta Crystallographica Section A: Foundations of Crystallography*, **59**(3):228–234, 2003.

OWEN, R. L., AXFORD, D., NETTLESHIP, J. E., OWENS, R. J., ROBINSON, J. I., MORGAN, A. W., DORE, A. S., LEBON, G., TATE, C. G., FRY, E. E., *et al.* Outrunning free radicals in room-temperature macromolecular crystallography. *Acta Crystallographica Section D: Biological Crystallography*, **68**(7):810–818, 2012.

OWEN, R. L., HOLTON, J. M., SCHULZE-BRIESE, C., AND GARMAN, E. F. Determination of X-ray flux using silicon pin diodes. *Journal of Synchrotron Radiation*, **16**(2):143–151, 2009.

OWEN, R. L., PATERSON, N., AXFORD, D., AISHIMA, J., SCHULZE-BRIESE, C., REN, J., FRY, E. E., STUART, D. I., AND EVANS, G. Exploiting fast detectors to enter a new dimension in room-temperature crystallography. *Acta Crystallographica Section D: Biological Crystallography*, **70**(5):1248–1256, 2014.

OWEN, R. L., RUDIÑO-PIÑERA, E., AND GARMAN, E. F. Experimental determination of the radiation dose limit for cryocooled protein crystals. *Proceedings of the National Academy of Sciences of the United States of America*, **103**(13):4912–4917, 2006.

OWEN, R. L., YORKE, B. A., GOWDY, J. A., AND PEARSON, A. R. Revealing low-dose radiation damage using single-crystal spectroscopy. *Journal of Synchrotron Radiation*, **18**(3):367–373, 2011.

PAITHANKAR, K. S. AND GARMAN, E. F. Know your dose: RADDOSE. *Acta Crystallographica Section D: Biological Crystallography*, **66**(4):381–388, 2010.

PAITHANKAR, K. S., OWEN, R. L., AND GARMAN, E. F. Absorbed dose calculations for macromolecular crystals: improvements to RADDOSE. *Journal of Synchrotron Radiation*, **16**(2):152–162, 2009.

PANNU, N. S. AND READ, R. J. Improved structure refinement through maximum likelihood. *Acta Crystallographica Section A: Foundations of Crystallography*, **52**(5):659–668, 1996.

PERUTZ, M. F. Isomorphous replacement and phase determination in non-centrosymmetric space groups. *Acta Crystallographica*, **9**(11):867–873, 1956.

PETOUKHOV, M. V., FRANKE, D., SHKUMATOV, A. V., TRIA, G., KIKHNEY, A. G., GAJDA, M., GORBA, C., MERTENS, H. D., KONAREV, P. V., AND SVERGUN, D. I. New developments in the ATSAS program package for small-angle scattering data analysis. *Journal of Applied Crystallography*, **45**(2):342–350, 2012.

PHILIPS, R. How big is the "average" protein? [online]. http://book.bionumbers.org/how-big-is-the-average-protein/, 2015. Accessed: 30 Dec. 2015.

PITTERI, M. AND ZANZOTTO, G. On the definition and classification of Bravais lattices. *Acta Crystallographica Section A: Foundations of Crystallography*, **52**(6):830–838, 1996.

POLLACK, L. SAXS studies of ion-nucleic acid interactions. *Annual Review of Biophysics*, **40**:225–242, 2011.

POPOV, A. N. AND BOURENKOV, G. P. Choice of data-collection parameters based on statistic modelling. *Acta Crystallographica Section D: Biological Crystallography*, **59**(7):1145–1153, 2003.

PURCELL, E. M. Life at low Reynolds number. *American Journal of Physics*, **45**(1):3–11, 1977.

RAI, M. AND PADH, H. Expression systems for production of heterologous proteins. *Current Science-Bangalore-*, **80**(9):1121–1128, 2001.

RAMAGOPAL, U. A., DAUTER, Z., THIRUMURUHAN, R., FEDOROV, E., AND ALMO, S. C. Radiation-induced site-specific damage of mercury derivatives: phasing and implications. *Acta Crystallographica Section D: Biological Crystallography*, **61**(9):1289–1298, 2005.

RASMUSSEN, C. E. AND WILLIAMS, C. K. I. Gaussian Processes for Machine Learning. Adaptive computation and machine learning. MIT Press, 2006. ISBN 9780262182539. URL `https://books.google.co.uk/books?id=GhoSngEACAAJ`.

RAVELLI, R. B. AND GARMAN, E. F. Radiation damage in macromolecular cryocrystallography. *Current Opinion in Structural Biology*, **16**(5):624–629, 2006.

RAVELLI, R. B. AND MCSWEENEY, S. M. The "fingerprint" that X-rays can leave on structures. *Structure*, **8**(3):315–328, 2000.

RAVELLI, R. B., THEVENEAU, P., MCSWEENEY, S., AND CAFFREY, M. Unit-cell volume change as a metric of radiation damage in crystals of macromolecules. *Journal of Synchrotron Radiation*, **9**(6):355–360, 2002.

READ, R. J. Structure-factor probabilities for related structures. *Acta Crystallographica Section A: Foundations of Crystallography*, **46**(11):900–912, 1990.

READ, R. J. Detecting outliers in non-redundant diffraction data. *Acta Crystallographica Section D: Biological Crystallography*, **55**(10):1759–1764, 1999.

READ, R. J. AND MCCOY, A. J. A log-likelihood-gain intensity target for crystallographic phasing in the presence of experimental error. *Acta Crystallographica Section D: Biological Crystallography*, **72**(3):375–387, 2015.

ROSSMANN, M. G. Processing oscillation diffraction data for very large unit cells with an automatic convolution technique and profile fitting. *Journal of Applied Crystallography*, **12**(2):225–238, 1979.

ROSSMANN, M. G., LESLIE, A. G. W., ABDEL-MEGUID, S. S., AND TSUKIHARA, T. Processing and post-refinement of oscillation camera data. *Journal of Applied Crystallography*, **12**(6):570–581, 1979.

ROUND, A., FELISAZ, F., FODINGER, L., GOBBO, A., HUET, J., VILLARD, C., BLANCHET, C. E., PERNOT, P., MCSWEENEY, S., ROESSLE, M., *et al.* BioSAXS Sample Changer: a robotic sample changer for rapid and reliable high-throughput X-ray solution scattering experiments. *Acta Crystallographica Section D: Biological Crystallography*, **71**(1):67–75, 2015.

SÄRKKÄ, S. Unscented Rauch–Tung–Striebel Smoother. *Automatic Control, IEEE Transactions on*, **53**(3):845–849, 2008.

SÄRKKÄ, S. Bayesian Filtering and Smoothing. Cambridge University Press, 2013. ISBN 110703065X. URL `https://books.google.com/books?id=5VlsAAAAQBAJ&pgis=1`.

SCHILLING, M. F. The longest run of heads. *College Math. J*, **21**(3):196–207, 1990.

SLIZ, P., HARRISON, S. C., AND ROSENBAUM, G. How does radiation damage in protein crystals depend on X-ray dose? *Structure*, **11**(1):13–19, 2003.

SOUTHWORTH-DAVIES, R. J. AND GARMAN, E. F. Radioprotectant screening for cryocrystallography. *Journal of Synchrotron Radiation*, **14**(1):73–83, 2007.

SOUTHWORTH-DAVIES, R. J., MEDINA, M. A., CARMICHAEL, I., AND GARMAN, E. F. Observation of decreased radiation damage at higher dose rates in room temperature protein crystallography. *Structure*, **15**(12):1531–1541, 2007.

STARR, C., EVERS, C., AND STARR, L. Biology: Concepts and Applications without Physiology. Cengage Learning, 2010. ISBN 9781133008682. URL `https://books.google.co.uk/books?id=dXQIAAAAQBAJ`.

STEWART, J. M. AND KARLE, J. The calculation of ε associated with normalized structure factors, E. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, **32**(6):1005–1007, 1976.

SVENSSON, O., MALBET-MONACO, S., POPOV, A., NURIZZO, D., AND BOWLER, M. W. Fully automatic characterization and data collection from crystals of biological macromolecules. *Acta Crystallographica Section D: Biological Crystallography*, **71**(8):1757–1767, 2015.

SYGUSCH, J. AND ALLAIRE, M. Sequential radiation damage in protein crystallography. *Acta Crystallographica Section A: Foundations of Crystallography*, **44**(4):443–448, 1988.

TAYLOR, G. The phase problem. *Acta Crystallographica Section D: Biological Crystallography*, **59**(11):1881–1890, 2003.

TAYLOR, G. L. Introduction to phasing. *Acta Crystallographica Section D: Biological Crystallography*, **66**(4):325–338, 2010.

TENG, T.-Y. Mounting of crystals for macromolecular crystallography in a free-standing thin film. *Journal of Applied Crystallography*, **23**(5):387–391, 1990.

TENG, T.-Y. AND MOFFAT, K. Radiation damage of protein crystals at cryogenic temperatures between 40 K and 150 K. *Journal of Synchrotron Radiation*, **9**(4):198–201, 2002.

TERWILLIGER, T. C. AND BERENDZEN, J. Bayesian weighting for macromolecular crystallographic refinement. *Acta Crystallographica Section D: Biological Crystallography*, **52**(4):743–748, 1996.

WAN, E. A. AND VAN DER MERWE, R. The unscented Kalman filter for nonlinear estimation. In Adaptive Systems for Signal Processing, Communications, and Control Symposium 2000. AS-SPCC. The IEEE 2000, pages 153–158. Ieee, 2000.

WAN, E. A. AND VAN DER MERWE, R. The Unscented Kalman Filter. In Kalman Filtering and Neural Networks, chapter 7, pages 221–280. John Wiley & Sons, Inc., 2002. ISBN 9780471221548. doi:10.1002/0471221546.ch7.

WARKENTIN, M. AND THORNE, R. E. Glass transition in thaumatin crystals revealed through temperature-dependent radiation-sensitivity measurements. *Acta Crystallographica Section D: Biological Crystallography*, **66**(10):1092–1100, 2010.

WATERMAN, D. G., WINTER, G., GILDEA, R. J., PARKHURST, J. M., BREWSTER, A. S., SAUTER, N. K., AND EVANS, G. Diffraction-geometry refinement in the DIALS framework. *Acta Crystallographica Section D: Structural Biology*, **72**(4):558–575, 2016.

WATERMAN, D. G., WINTER, G., PARKHURST, J. M., FUENTES-MONTERO, L., HATTNE, J., BREWSTER, A., SAUTER, N. K., AND EVANS, G. The DIALS framework for integration software. *CCP4 Newsletter on Protein Crystallography*, **49**:16–19, 2013.

WEIK, M. AND COLLETIER, J.-P. Temperature-dependent macromolecular X-ray crystallography. *Acta Crystallographica Section D: Biological Crystallography*, **66**(4):437–446, 2010.

WEIK, M., RAVELLI, R. B. G., KRYGER, G., MCSWEENEY, S., RAVES, M. L., HAREL, M., GROS, P., SILMAN, I., KROON, J., AND SUSSMAN, J. L. Specific chemical and structural damage to proteins produced by synchrotron radiation. *Proceedings of the National Academy of Sciences of the United States of America*, **97**(2):623–628, 2000.

WEIK, M., RAVELLI, R. B. G., SILMAN, I., SUSSMAN, J. L., GROS, P., AND KROON, J. Specific protein dynamics near the solvent glass transition assayed by radiation-induced structural changes. *Protein Science*, **10**(10):1953–1961, 2001.

WEISSTEIN, E. W. Deconvolution [online]. http://mathworld.wolfram.com/Deconvolution.html, 2016. Accessed 6 May 2016.

WIENER, N. Extrapolation, interpolation, and smoothing of stationary time series, volume 2. MIT press Cambridge, MA, 1949.

WILSON, A. J. C. The probability distribution of X-ray intensities. *Acta Crystallographica*, **2**(5):318–321, 1949.

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., MCCOY, A., *et al.* Overview of the CCP4 suite and current developments. *Acta Crystallographica Section D: Biological Crystallography*, **67**(4):235–242, 2011.

WOOLFSON, M. An improvement of the 'heavy-atom' method of solving crystal structures. *Acta Crystallographica*, **9**(10):804–810, 1956.

YORKE, B. A., BEDDARD, G. S., OWEN, R. L., AND PEARSON, A. R. Time-resolved crystallography using the Hadamard transform. *Nature Methods*, **11**(11):1131–1134, 2014.

ZELDIN, O. B. Methods Development for Structural Biology. DPhil thesis, University of Oxford, UK, 2013.

ZELDIN, O. B., BROCKHAUSER, S., BREMRIDGE, J., HOLTON, J. M., AND GARMAN, E. F. Predicting the X-ray lifetime of protein crystals. *Proceedings of the National Academy of Sciences of the United States of America*, **110**(51):20551–20556, 2013a.

ZELDIN, O. B., GERSTEL, M., AND GARMAN, E. F. Optimizing the spatial distribution of dose in X-ray macromolecular crystallography. *Journal of Synchrotron Radiation*, **20**(1):49–57, 2012.

ZELDIN, O. B., GERSTEL, M., AND GARMAN, E. F. *RADDOSE-3D*: time- and space-resolved modelling of dose in macromolecular crystallography. *Journal of Applied Crystallography*, **46**(4):1225–1230, 2013b.